Genetically guided drug development

María Gordillo Marañón

Thesis submitted for the degree of Doctor of Philosophy at University College London

Institute of Cardiovascular Science

School of Life and Medical Science

University College London

I, María Gordillo Marañón, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in this thesis.

Abstract

Background

Attrition is a major issue in drug development with less than 5% of drug development programmes yielding licensed drugs. Retrospective studies have suggested that human genomic data could be used to help prioritise drug development programmes and reduce the risk of clinical-stage failure. The investment of pharmaceutical companies in healthcare genomic initiatives has been incentivised largely by studies showing that genetically-supported targets would succeed twice as often as those without genetic support, and comparative studies revealing that the effect of licensed drugs on biomarkers and disease endpoints coincide with the observed associations of variants in the genes encoding the corresponding target. However, historically, genome-wide association studies (GWAS) of human diseases and pharmaceutical research and development have largely proceeded independently. Knowledge of the overlap between existing GWAS and current or historical drug development programmes is important to maximise the utility of existing data for repurposing opportunities and mechanism-based adverse effect prediction. Additionally, for novel target identification, questions remain about what type of genomic data is most informative and what methods are most robust. Mendelian randomisation (MR), a genetic epidemiology approach for causal inference, has been used to assess the causal nature of exposures on outcomes. Its application has recently been extended to the evaluation of drug targets against disease ('drug target MR'). However, very limited validation of the parameters used in drug target MR studies exists across multiple target loci and diseases.

Aim

To investigate the extent to which the spectrum of human diseases has been addressed by genetic analyses, or by drug development, and the degree to which these efforts overlap. To evaluate the genetic support for approved drug target-indication pairs from GWAS and drug target MR applications.

Methods

Human disease information from the Disease Ontology and drug data from ChEMBL version 25 were used. Genetic associations with diseases and clinical endpoints were sourced from the GWAS Catalog and UK Biobank (through Neale Lab), and genetic associations for circulating protein levels measured by the SomaLogic v4 proteomic platform from the Fenland study and UCLEB Consortium.

I calculated the disease coverage, overlap and divergence of human genetic studies and pharmaceutical research and development. I provided a revised estimate of the value of genetic evidence for drug target-indication pairs in progressing in clinical-phase drug development, and investigated different approaches to assign genetic associations identified by GWAS to causal genes. I evaluated the drug target MR framework with a curated 'truth' set of drug target-indication pairs for which genetic associations with the circulating levels of the protein target and the intended indication were available. I applied the drug target MR framework using genetic associations with blood lipids (LDL-cholesterol, HDL-cholesterol and triglycerides) to prioritise drug targets for the treatment and prevention of coronary heart disease.

Results

Only 9% (953 out of 10,901) of human diseases have been studied by GWAS. Of these, only 369 correspond to diseases with an approved treatment and/or a treatment under clinical

or preclinical development, leaving 584 diseases that have been the subject of investigation in GWAS, but which have yet to be investigated in drug development. For those indications that are or have been the subject of clinical phase drug development and have been studied by GWAS, I found that drug target-indication pairings with genetic support are twice more likely to get approved than those without genetic support (2.18; 95%CI: 1.86; 2.51). The evaluation of the drug target MR framework with the subset of target-indication pairings of approved drugs with available genetic associations with the circulating protein levels recapitulated the mechanism of action of up to 13% (16 out of 121) of the drug target gene – indication pairings and returned results in the unanticipated direction of effect for 11% (14 out of 121) of the pairings explored. The systematic application of the biomarker-weighted drug target MR using blood lipid levels robustly identified 30 targets that should be prioritised for the prevention or treatment of coronary heart disease.

Conclusion

I identified points of convergence or divergence between genomic research and drug development efforts in the sample space of all the human drug targets and diseases, and demonstrated the utility of GWAS data for drug target identification and validation through the mapping of genetic associations to causal genes and the application of the drug target MR framework. The work of this thesis informs prioritisation strategies in drug development and future research so the investment and impact of human genetic studies can be maximised.

Impact statement

It has been suggested that human genomics may help increase the efficiency of drug development by generating evidence for drug target identification and validation. Pharmaceutical companies have shown growing interest in the use of human genomic data, however, the detailed analysis of disease coverage, overlap and divergence of human genetic studies and pharmaceutical research and development in Chapter 4, shows that less than 10% of human diseases have been studied by genome-wide association studies (GWAS), indicating that further efforts are needed to explore the genetic predisposition of the remaining diseases, and more importantly, the genetic contribution for those >9,000 diseases without an approved or investigational drug. In addition, the analysis described in Chapter 5 provides further evidence of the additional value of genetic evidence for drug target-indication pairings in progressing in the drug development pipeline. Genetic support could help prioritise medicines for cardiovascular disease or repurposing approved drugs. The findings from this chapter encourage the research community and pharmaceutical industry to align efforts and perform genetic studies in cardiovascular diseases or other therapeutic areas without an approved or investigational drug.

Large population-based cohort studies, and particularly biobanks, have emerged as a powerful resource to advance biomedical research. The linkage of human genetic data to medical records, clinical biomarkers and molecular traits, such as circulating protein levels, represents an unique opportunity to exploit genomic data and inform drug target identification and validation. Different techniques in genetic epidemiology are used to infer the effect of a drug on a target in a particular disease. This thesis evaluated in Chapter 6 the use of drug target Mendelian Randomisation using circulating protein levels to estimate the effect of perturbating a target in a particular disease using a set of licensed drug target – indication pairings. The

methods and findings from the work in Chapter 6 are intended in large part to pave the way for further studies exploring the application of drug target Mendelian randomisation with protein level data for drug target validation and identification.

While Mendelian randomisation methods using molecular traits become better understood, traditional clinically-validated biomarkers are used to infer the effect of perturbing a drug target in a particular disease. Chapter 7 prioritises a set of 30 targets that might elicit beneficial effects in the prevention or treatment of coronary heart disease using blood lipid data as the exposure.

The full integration of genome-wide association studies in the drug development pipeline is still very much a work-in-progress. There are several drugs that have been prioritised based on population-level genetic data showing promising therapeutic benefit. Academic research and clinical trials of these candidates are ongoing at the time of writing this thesis. This thesis anticipates that the mining of data from genome-wide association studies will help address the attrition problem in the pharmaceutical industry.

Acknowledgements

I would like to thank my PhD supervisors, Professor Aroon Hingorani, Dr Chris Finan and Dr Floriaan Schmidt. I am extremely grateful for their invaluable advice, continuous support and patience. Their expertise, plentiful experience and wisdom have encouraged me throughout my PhD research and daily life.

My gratitude extends to the research group, colleagues and collaborators for their kind help and support. I am particularly grateful to my peers, who shared the adventure of the PhD and helped me navigate through the difficult times.

I am also truly fortunate to have been mentored by Dr Christine Reynet, and I owe her a debt of gratitude for her advice, encouragement and kindness.

I would also like to thank the British Heart Foundation for the studentship that allowed me to conduct this thesis.

Lastly, words cannot express how grateful I am for the unstinting support from my family. To my brother, Jaime, without his optimism and encouragement in the past few years, it would have been difficult for me to complete this thesis. To my parents, Lucia and Javier, every accomplishment in my career has been possible thanks to the educational opportunities they provided. For this reason, and many more, I dedicate this thesis to them.

Funding

This work was supported by the British Heart Foundation 4-year studentship in Cardiovascular Research at University College London (FS/17/70/33482).

Related academic work

Publications arising from this thesis

1. **M Gordillo-Marañón,** *et al.,* Validation of lipid-related therapeutic targets for coronary heart disease prevention using human genetics, Nature Communications, 2021, https://doi.org/10.1038/s41467-021-25731-z

2. A Henry, **M Gordillo-Marañón,** *et al.,* Therapeutic targets for heart failure identified using proteomics and Mendelian randomisation, Circulation 2022;145:1205–1217, https://doi.org/10.1161/CIRCULATIONAHA.121.056663

3. AF Schmidt, N B Hunt, **M Gordillo-Marañón,** *et al.*. Cholesteryl ester transfer protein (CETP) as a drug target for cardiovascular disease. Nature Communications 12, 2021, https://doi.org/10.1038/s41467-021-25703-3

4. AF Schmidt, C Finan, **M Gordillo-Marañón**, *et al*., Genetic drug target validation using Mendelian randomisation, Nature Communications, 2020, https://doi.org/10.1038/s41467-020-16969-0

5. AF Schmidt, R Joshi , **M Gordillo-Marañón,** *et al*., Biomedical consequences of elevated cholesterol-containing lipoproteins and apolipoproteins on cardiovascular and non-cardiovascular disease and related traits. (*Accepted at Communications Medicine)* 2022, https://doi.org/10.1101/2022.03.11.22272251

6. **M Gordillo-Marañón,** *et al.,* Disease coverage, overlap and divergence of genome-wide association studies and pharmaceutical research and development. (*Manuscript under internal revision*)

Oral presentations

I have delivered an oral presentation related to the contents of this PhD at the following conferences and meetings:

- Institute of Cardiovascular Science International Women's Day, March 2021
- European Medicines Agency, April 2021
- Target Validation Using Genomics and Bioinformatics, December 2019

Poster presentations

I have delivered poster presentations related to the contents of this PhD at the following conferences and meetings:

- American Society of Human Genetics 2019
- International Genetic Epidemiology Society 2020

Contents

List of Abbreviations

AD	Alzheimer's disease
AF	Atrial fibrillation
ATC	Anatomical therapeutic chemical
BNF	British National Formulary
bp	Base pair
CHD	Coronary heart disease
CI	Confidence interval
CT	Clinical trial
DO	Disease Ontology
eQTL	Expression quantitative trait loci
FEV	Forced expiratory volume in the first second
GLGC	Global lipid genetic consortium
GWAS	Genome-wide association study
GWASdb	GWAS database
HDL-C	High-density lipoprotein cholesterol
ICD	International Classification of Diseases
ICD-10	International Classification of Diseases, tenth revision
InSIDE	Instrument Strength Independent of Direct Effect
IV	Instrumental variable
IVW	Inverse-variane weighted
kbp	Kilo base pair
LD	Linkage disequilibrium
LDL-C	Low-density lipoprotein cholesterol
MAF	Minor allele frequency
Mbp	Megabase pair
MeSH	Medical subject headings
mQTL	Metabolite-levels quantitative trait loci
MR	Mendelian randomisation
MRC	Medical research council
MVMR	Multivariable mendelian randomisation

NMR	Nuclear magnetic resonance
OR	Odds ratio
pQTL	Protein quantitative trait loci
QTL	Quantitative trait loci
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
SNP	Single nucleotide polymorphism
TG	Triglycerides
UCLEB	University college london-edinburgh-bristol Consortium
UMLS	Unified medical language system
xMHC	Extended major histocompatibility complex

1 | Introduction

This introductory chapter will provide an overview of the current state of drug development, the potential of human genetic studies to address the high attrition rates and increase efficacy in clinical development, and the application of Mendelian randomisation for drug target identification, validation and prioritisation.

1.1. The current state of drug development

Most successful medicines target proteins. Therefore, the challenge in drug development is to identify disease relevant proteins and design compounds that can modify their function to treat disease. However, less than 5% of drug development programmes yield licensed drugs[1,2]. Reasons for failure include the compound failing to show benefits compared to another treatment or placebo (lack of therapeutic efficacy, ~60% of failures), safety concerns (~17% of failures), or strategic reasons, for example, when a pharmaceutical company ceases the development due to market competition or financial constraints (~20% of failures)[3].

The vast majority of failures arising due to lack of efficacy occur at a late stage in the development pipeline, in phase II or phase III randomised clinical trials[3,4]. Many of these drugs may have been strong pre-clinical candidates indicating that early experiments in cells and animals are poor predictors of human efficacy. In addition, early-phase clinical trials (phase I), which evaluate dose safety and tolerability, are not designed to determine if the drug target plays a relevant role in a disease. Phase I studies are usually performed in small cohorts of healthy volunteers over a short period of time to help evaluate pharmacokinetics and dose range, as well as to measure any commonly observed adverse effects rather than to confirm or test target validity[5].

Late-phase failures raise ethical concerns (e.g. thousands of patients being exposed to ineffective or potential harmful drugs) and have financial implications, because a phase III trial requires an enormous investment in addition to costs already incurred to progress a compound to that stage. The average cost of introducing a drug into the market is estimated in $985.3 million[6] and in some cases even several billion dollars[7]. Clearly, the current situation is not sustainable and demands improved methodologies that can provide robust evidence of target efficacy in early stages of the drug development process. A key requirement of any new method

would be to enable early, reliable insight on the likelihood of success of any target and disease indication combination to remove those pairings unlikely to be successful from the drug development pipeline prior to clinical phase trials, thus reducing overall development cost.

1.2. The potential of genome-wide association studies in drug development

Genome-wide association studies (GWAS) in patients and populations test relationships between natural sequence variation (genotype) and disease risk factors, biomarkers and clinical endpoints using population-based cohort and case-control designs[8]. In the last 13 years, over 5,687 GWAS have been completed in approximately 4,083 traits[9]. The rise of genome-wide association studies has been enabled by the significant reduction of genotyping costs and the substantial investment in sequencing, genotyping or molecular phenotyping of large cohort studies (e.g., University College London-Edinburgh-Bristol Consortium; UCLEB[10]) and national biobanks which are connected to routinely collected primary and secondary care health records (e.g., UK Biobank[11] and FinGenn[12]). Many of these comprise molecular traits such as proteomics and metabolomics measures in addition to genetic data. Some examples of the largest biobanks with 'omic-' data are shown in Table 1.1. Future initiatives that will incorporate genomic data linked to medical history include the All of Us program in the USA[13] or the planned Three Million African Genomes[14].

Several public repositories exist that systematically catalogue, curate and store GWAS summary statistics. For example, the latest update of the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) GWAS catalog[9] contains over 5,000 publications of published GWAS done in different human populations, and almost

400,000 associations between SNP and traits. Another GWAS repository is the GWAS database (GWASdb), developed at the University of Hong Kong, and combines GWAS results with functional annotations[15]. Similarly, the Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP) collects information about significant associations in genetic studies, including methylation and expression quantitative trait loci (eQTL) analysis[16]. Other relevant resource is provided by Neale Lab[17] which has released to the public summary statistics from genome-wide association studies for approximately 2,000 phenotypes measured in ~337,000 participants of the UK Biobank.

Table 1.1. Examples of the population-based biobanks

Study (Location)	**Study type**	**Number of participants**	**Omic data available (samples)**
UK Biobank[11] (UK)	Biobank	500,000	• Genotype data (500,000) • Whole exome sequencing data (200,000) • Whole genome sequencing data (200,000) • Proteomics (53,000) • Metabolomics – 249 molecules measured by Nuclear magnetic resonance (120,000) • 34 clinical biomarkers (500,000)
FinnGen[12] (Finland)	Biobank	476,400	• Genotype data (365,000)
The Estonian Biobank (Estonia)[18]	Biobank	200,000	• Genotype data (200,000) • Whole exome sequencing data (2,500) • Whole genome sequencing data (3,000) • Metabolomics NMR - 120 molecules (11,000) • 42 Clinical biomarkers (2,700) • Proteomics (~1,000) • Transcriptomics (~1,000)
BioBank Japan[19,20] (Japan)	Biobank	260,000	• Genotype data (~220,000) • Whole genome sequencing data (~218,000) • Metabolomics - 39 molecules measured by capillary electrophoresis mass spectrometry (500)

The design of genome-wide association studies has shown potential as a novel resource for drug development. Retrospective analyses of successful drugs whose indications have been studied by GWAS have shown that selecting drugs targets where genetic associations have been found near or in the gene encoding the target could double the success rate in clinical development[21,22]. Further, several analyses have been completed that demonstrate GWAS have rediscovered 39 drug targets, including 8 targets for cardiovascular drugs (Table 1.2). Moreover, GWAS have potentially uncovered numerous repurposing opportunities (Fig. 1.1). In an effort to streamline drug development from GWAS data, Finan *et al.*, 2017 defined the druggable genome[23], the set of genes whose protein products are already drugged or have a greater probability of encoding a protein amenable to targeting with a pharmaceutical. The most recent definition comprises 4,863 genes and incorporates potential targets for monoclonal antibodies.

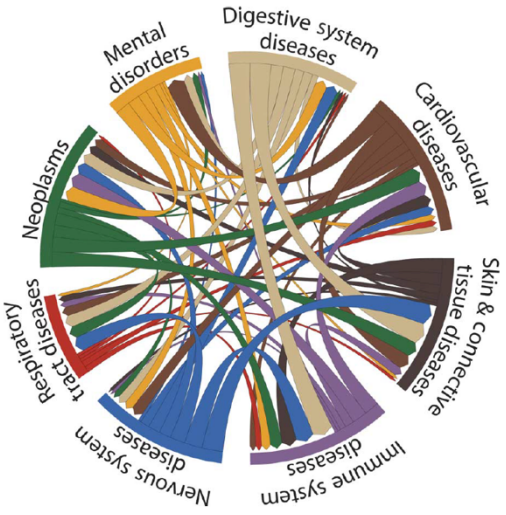


Figure 1.1 Potential repurposing opportunities uncovered by GWAS. The disease categories on the circumference are MeSH root disease terms. The directional chords represent a connection from an indication class of drug to a GWAS phenotype. This connection is determined by a drug target gene occurring within 50 kilo base pair (kbp) of a GWAS association. The width of the chords is proportional to the number of genes connecting two therapeutic classes. Figure adapted from Finan *et al.*, 2017[23].

As described by Hingorani *et al*., 2019[24], GWAS overcome many design flaws inherent in preclinical experiments in isolated systems (cells, tissues, isolated organs) and animal models as they are performed in the organism of interest (the human), have a low false discovery rate and have the capability to interrogate every potential drug target in the condition of interest. Yet, the main challenge in GWAS interpretation is the identification of the true 'causal' gene driving the association, given that the majority of genetic associations found through GWAS are located in non-coding regions and that may include several genes. Recently, several statistical tools have been developed, including coloc[25], moloc[26], CaMMEL[27] and SMR[28], that aim to co-localise genetic associations with mRNA (or protein) expression and disease endpoints to help assign the responsible gene in an linkage disequilibrium (LD) interval. Despite all the proposed methodologies, assigning variants to genes based on genomic proximity has been described as the most reliable approach to map causal genes[29,30]. Still, very few discovery GWAS have identified the gene(s) driving the association so at present, it is not clear which method is optimal.

Nevertheless, while GWAS alone can potentially inform drug target identification and validation, deciding whether to design an inhibitor or activator of the target cannot be readily inferred simply from identification of the locus.

Table 1.2. Examples of GWAS 'rediscoveries' of licensed drug targets

GWAS Phenotype	Associated Gene	Compound
Total/LDL cholesterol	3-hydroxy-3-methylglutaryl-CoA reductase (*HMGCR)*	Lovastatin, Pravastatin, Simvastatin
Diastolic blood pressure	CACNA1D calcium voltage-gated channel subunit alpha1 D *(CACNA1D)*	Amlodipine
Large artery stroke	Plasminogen (*PLG*)	Alteplase
Heart rate	Acetylcholinesterase (*ACHE*)	Neostigmine Methylsulfate
	Cholinergic receptor muscarinic 2 (*CHRM2*)	Tolterodine Tartrate
Type 2 diabetes	Potassium inwardly-rectifying channel subfamily J member 11 (*KCNJ11)*	Glimepiride, Glipizide, Glyburide, Nateglinide, Repaglinide
	ATP binding cassette subfamily C member 8 (*ABCC8*)	Glimepiride, Glipizide, Glyburide, Nateglinide, Repaglinide
	Peroxisome proliferator-activated receptor gamma (*PPARG)*	Pioglitazone

1.3. Nature's randomised trials: Mendelian randomisation

Mendel's second law (the 'law of independent assortment') states that the segregation of alleles at a locus during conception is mutually independent and independent of other factors, and thus, genotypes of individuals are obtained by the random allocation of alleles during meiosis when DNA is passed from parents to offspring (Mendelian randomisation; MR). If an allele of a genetic variant results in an increase or decrease in disease risk or biomarker level, then Mendel's second law is analogous to the randomisation of an active drug or placebo in a randomised controlled trial. Therefore, genetic variation can be used to mimic randomised clinical trials without requiring the time-consuming and costly development of a drug compound[31] (Fig. 1.2).

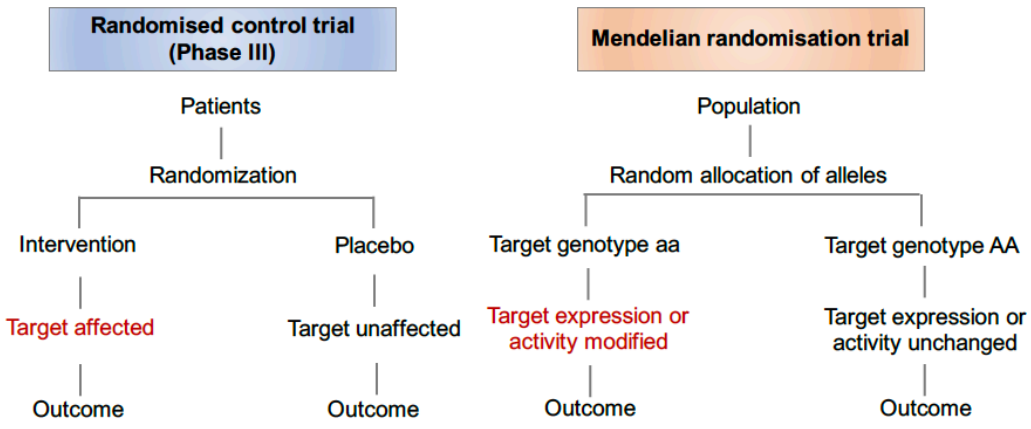


Figure 1.2. Mendelian randomisation trials as a nature phase III randomised clinical trial. Expected outcome from hypothetical randomised control trial and from Mendelian randomisation analysis, if the target is causal in the development of the disease. 'AA' and 'aa' refer to alleles of the gene encoding the target of a drug (only homozygous individuals are shown). In this example, genotypes are also associated with low or high risk of developing the particular disease. Figure adapted from Hingorani & Humphries, 2005[31].

Where the variable of interest in an MR analysis is a disease biomarker, rather than a specific drug target, established MR approaches have utilised selected variants in or near multiple genes that have been identified in GWAS of biomarker levels from throughout the genome. Collectively, these variants are known as a genetic instrument. An MR analysis will assess the effects of the genetic instrument on a biomarker level and the effects of the genetic instrument with respect to the disease risk to determine if the biomarker exposure is causal in the disease outcome. The resulting estimate will determine how much an increase or decrease in the biomarker impacts the increase or decrease in disease risk. This is referred to as MR analysis for biomarker validation or 'genome-wide biomarker MR'[32].

For illustration, genome-wide biomarker MR studies have further validated the causal role of low-density lipoprotein cholesterol (LDL-C) on coronary heart disease (CHD), which was first established in observational studies and eventually confirmed as causal by randomised controlled trials of LDL-C lowering statin drugs[33,34], PCSK9 inhibitors[35] and ezetimibe, which targets NPC1L1[36,37] (Table 1.3).

Table 1.3. Causal odds ratios (95% CI) for coronary heart disease per standard deviation increase in each lipid fraction. All the studies used variants from the Global Lipid Genetic Consortium (GLGC) to instrument causal effects of the three lipid subfractions on CHD from the CardiogramPlusC4D Consortium. (*) Derived from Table 3 of Do *et al*., 2013[38].

Method	LDL-C OR (95%CI)	HDL-C OR (95%CI)	Triglycerides OR (95%CI)	Ref.
Regression-based method*	1.46 (1.37, 1.57) $n_{variants}$ = 185	0.96 (0.89, 1.03) $n_{variants}$ = 185	1.43 (1.28, 1.61) $n_{variants}$ = 185	38
Multivariable IVW MR	1.48 (1.36, 1.61) $n_{variants}$ = 185	0.93 (0.85, 1.02) $n_{variants}$ = 185	1.16 (1.04, 1.29) $n_{variants}$ = 185	39
Restricted allelele score	1.92 (1.68, 2.19) $n_{variants}$ = 19	0.91 (0.42, 1.98) $n_{variants}$ = 19	1.61 (1.00, 2.59) $n_{variants}$ = 19	40

Genome-wide biomarker MR has also been applied to non-LDL lipid subfractions such as high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG), for which a causal role in CHD risk remains controversial. Non-randomised observational studies have reported a risk increasing association between TG and CHD[40], an association that has recently been suggested to be causal by genome-wide biomarker MR studies (Table 1.3). The role of TG in CHD is currently under investigation in clinical trials of evinacumab, an ANGPTL3 inhibitor predicted to reduce CHD risk by lowering TG levels[41]. In contrast, and despite suggestive but inconclusive MR estimates (Table 1.3), causality of the HDL-C and CHD association remains controversial. Despite several attempts to raise HDL-C by inhibiting CETP, a key enzyme in HDL-C metabolism, none of the CETP inhibitors[43–46] have been approved yet, questioning HDL-C role in CHD and leading to confusion as to therapeutic targeting of HDL-C metabolism

is likely to be fruitful. Furthermore, only the CETP inhibitor anacetrapib showed a reduction in cardiovascular events in phase III clinical trials[46], suggesting between-compound heterogeneity. Therefore, the anticipated CHD effect may depend on the method of intervening on downstream lipid biomarkers (i.e. which proteins are targeted by drugs). To explore this situation in which the therapeutic response varies between different interventions on a biomarker and to comprehensively evaluate a drug effect on a specific target protein regardless of heterogeneity in downstream pathways, Schmidt *et al.,* 2020[32] proposed a drug target MR approach.

1.4. Mendelian randomisation for drug target validation

It has been shown that variants in a gene encoding a specific drug target, that alter the target's expression or function, can be used as a tool to anticipate the effect of drug action on the same target. This application of Mendelian randomisation is known as 'drug target MR'[47]. In contrast to a genome-wide biomarker MR, where the variants comprising the genetic instrument are selected from across the genome, in a drug target MR analysis, variants are selected from the gene of interest or neighbouring genomic region because these variants are most likely to associate with the expression or function of the encoded protein (acting in *cis*). Whereas genome-wide biomarker MR helps infer the causal relevance of a biomarker for a disease, a drug target MR helps infer whether and, in certain cases in what direction, a drug that acts on the encoded protein, whether an antagonist, agonist, activator or inhibitor, will alter disease risk (Table 1.4).

Table 1.4. Main conceptual differences between *genome-wide biomarker* and *drug target* MR approaches.

	Biomarker MR	**Drug target MR**
Aim	Causal effect of a biomarker	Causal relevance of a drug target
SNP selection	Genome-wide	Locus specific
Ideal exposure	Clinically relevant quantitative trait	mRNA or protein expression of the encoded gene
MR methods	IVW, MR-Egger and other (see later section)	Methods accounting for residual genetic correlation to maximise power

Further evidence on the validity of this approach is that the licensed LDL-C lowering targets have also been rediscovered by drug target MR approaches. Polymorphisms in *NPC1L1,* the gene that encodes the target of ezetimibe, are associated both with lower LDL-C levels and decreased CHD risk (OR: 0.95, 95% CI: 0.92,0.99)[48]. The effect of instrumenting LDL-C on CHD using LDL-lowering variants in *HMGCR* is 0.81 (95% CI: 0.72, 0.90) and 0.81 (95% CI: 0.74, 0.89) when using variants in *PCSK9*[49], consistent with the effect of statins and PCSK9 inhibitors in clinical trials[33,35]. Furthermore, a drug target MR of *CETP* on CHD, using variants in the *CETP* gene weighted by their effect on HDL-C, indicates protection from disease (odds ratio: 0.87; 95%CI: 0.84, 0.90)[32]. The finding is consistent with the effect of allocation to the CETP-inhibitor anacetrapib in a placebo-controlled trial (0.93; 95%CI: 0.86, 0.99) and is compatible with the view that targeting *CETP* is an effective therapeutic approach to prevent CHD (Fig. 1.3)[46].

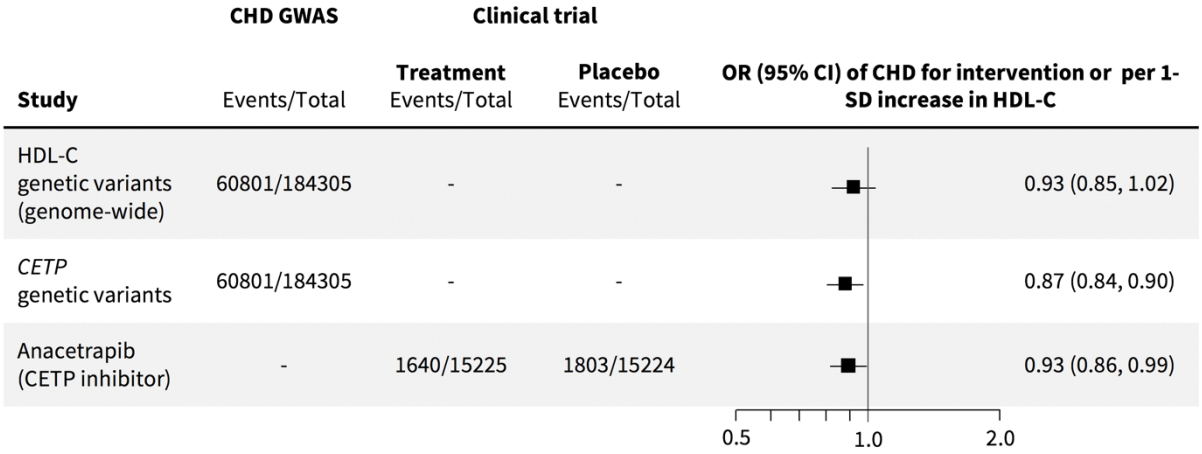


Figure 1.3. HDL-C, CETP inhibitor and CHD: genome-wide biomarker vs drug target MR. Forest plot of the HDL-C biomarker MR estimate (Holmes *et al.*, 2015[39]), drug target MR estimate of CETP level and function using HDL-C as a proxy (Schmidt *et al.*, 2020[32]) and odds ratio of anacetrapib clinical trial (HPS3/TIMI55–REVEAL Collaborative Group, 2017[46]). OR = odds ratio; CI = confidence interval; SD = standard deviation.

In addition to drug target validation, drug target Mendelian randomisation has also been employed to anticipate the outcome of a phase II/III randomised clinical trial[49] and identify potential drug repurposing opportunities. For example, it has been demonstrated that the increased risk of type 2 diabetes associated with statin treatment is an effect of HMG-CoA inhibition[48], whereas the blood pressure raising effect of torcetrapib, a CETP inhibitor, was an off- target effect and unrelated to CETP inhibition[50]. Further applications include drug repositioning. For instance, tocilizumab, a monoclonal antibody that blocks the interleukin-6 receptor originally licensed to treat rheumatoid arthritis, was later suggested as a potential therapeutic agent for the treatment of coronary heart diseases based on the causal role of the target in the development of the disease[51]. Inhibition of the same target may also be effective in abdominal aortic aneurysms[52], atrial fibrillation[53], and inflammatory bowel disease[54] but might increase the risk of asthma[53]. This illustrates the concept that drugs targeting a single protein may affect multiple disease outcomes.

1.5. Aim and objectives

As introduced earlier in the chapter, mapping disease loci identified by genome-wide association studies (GWAS) to the genes encoding the protein targets of licensed drugs has suggested that i) GWAS could provide a useful tool for systematic identification of new drug targets for human disease, ii) drug targets genetically-validated by GWAS are more likely to succeed. However, the extent to which GWAS are exploited and used to inform drug development is unknown. Furthermore, deciding whether to design an inhibitor or activator (agonist or antagonist for receptor targets) of the target cannot be readily inferred simply from identification of the locus. To help infer the correct mechanism of action for a new drug, I propose the *cis*-Mendelian Randomisation (MR) approach ('drug target MR'). By using protein expression levels (protein quantitative trait loci; pQTL) as a potential proxy for protein function, a drug target MR analysis assesses the effects of variants in a single gene on its pQTL with respect to disease risk. The inference determines whether and by how much an increase or decrease in the protein impacts disease risk, suggesting a plausible mechanism of action for the drug. However, as discussed in the following chapter, multiple parameters determine MR performance including linkage disequilibrium (LD) or strength of the association with the exposure.

I hypothesise that by using publicly available GWAS data combined with drug information and in-house genetic and proteomic data, I will be able to investigate the performance of large scale drug target MR analysis. By using these parameters I could (a) better predict the efficacy of preclinical candidates (b) uncover repurposing opportunities (c) predict mechanism-based side effects of licensed and drugs in development and (d) evaluate the therapeutic potential of novel druggable genes in cardiovascular, among other disease

outcomes. The hypothesis outlined here will be accomplished by working on the following aims:

1. Investigating the extent to which the spectrum of human diseases has been addressed by genome-wide association studies, or by drug development, and the degree to which these efforts overlap to inform genetically guided pharmaceutical research.

2. Evaluating the genetic evidence from GWAS on drug target-indication progression along the drug development process and providing an updated estimate of the probability of success for drug target-indication pairing given genetic support.

3. Validating the drug-target MR approach using a 'truth' set of approved drugs for which available GWAS data on circulating protein levels (pQTL) of the target are available and the intended indication has also been studied by GWAS to investigate if the 'pQTL-weighted drug target MR' framework recapitulates their mechanism of action.

4. Consolidating a 'biomarker-weighted drug target MR' approach to systematically prioritise and validate drug targets where circulating protein levels have not been measured directly, and genetic associations with a clinical biomarker downstream to the protein are available and could be used as a proxy for protein concentration or activity.

1.6. References

1. Munos, B. Lessons from 60 years of pharmaceutical innovation. *Nat Rev Drug Discov* **8**, 959–968 (2009).
2. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9**, 203–214 (2010).
3. Hwang, T. J. *et al.* Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA Intern Med* **176**, 1826–1833 (2016).
4. Harrison, R. K. Phase II and phase III failures: 2013-2015. *Nature Reviews Drug Discovery* **15**, 817–818 (2016).
5. Naci, H. & Ioannidis, J. P. A. How good is 'evidence' from clinical studies of drug effects and why might such evidence fail in the prediction of the clinical utility of drugs? *Annu. Rev. Pharmacol. Toxicol.* **55**, 169–189 (2015).
6. Wouters, O. J., McKee, M. & Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **323**, 844–853 (2020).
7. Herper, Matthew, H., Matthew. The Cost Of Creating A New Drug Now $5 Billion, Pushing Big Pharma To Change. https://www.forbes.com/sites/matthewherper/2013/08/11/how-the-staggering-cost-of-inventing-new-drugs-is-shaping-the-future-of-medicine/#448fe61d13c3 (2013).
8. Witte, J. S. Genome-wide association studies and beyond. *Annu Rev Public Health* **31**, 9-20 4 p following 20 (2010).
9. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

10. Shah, T. *et al.* Population Genomics of Cardiometabolic Traits: Design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PLOS ONE* **8**, e71345 (2013).
11. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779 (2015).
12. FinnGen-tutkimushanke vie suomalaiset löytöretkelle genomitietoon. *FinnGen* https://www.finngen.fi/fi/finngen_tutkimushanke_vie_suomalaiset_loytoretkelle_genomitietoon.
13. The "All of Us" Research Program. *New England Journal of Medicine* **381**, 668–676 (2019).
14. Wonkam, A. Sequence three million genomes across Africa. *Nature* **590**, 209–211 (2021).
15. Li, M. J. *et al.* GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* **44**, D869–D876 (2016).
16. Leslie, R., O'Donnell, C. J. & Johnson, A. D. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**, i185-194 (2014).
17. UK Biobank. *Neale lab* http://www.nealelab.is/uk-biobank.
18. The Estonian Biobank. *EIT Health Scandinavia* https://www.eithealth-scandinavia.eu/biobanks/the-estonian-biobank/.
19. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2–S8 (2017).
20. BioBank Japan. https://biobankjp.org/en/index.html#01.

21. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856–860 (2015).
22. King, E. A., Davis, J. W. & Degner, J. F. 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**, e1008489 (2019).
23. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, (2017).
24. Hingorani, A. D. *et al.* Improving the odds of drug development success through human genomics: modelling study. *Sci Rep* **9**, 18911 (2019).
25. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
26. Giambartolomei, C. *et al.* A Bayesian Framework for Multiple Trait Colo-calization from Summary Association Statistics. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty147.
27. A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease | bioRxiv. https://www.biorxiv.org/content/early/2017/12/01/219428.
28. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
29. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, e3–e3 (2019).
30. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* 1–6 (2021) doi:10.1038/s41586-021-03446-x.
31. Hingorani, A. & Humphries, S. Nature's randomised trials. *Lancet* **366**, 1906–1908 (2005).

32. Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nature Communications* **11**, 3255 (2020).
33. Collaborators, C. T. T. (CTT). The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *The Lancet* **380**, 581–590 (2012).
34. Collins, R. *et al.* Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet* **388**, 2532–2561 (2016).
35. Schmidt, A. F. *et al.* PCSK9 monoclonal antibodies for the primary and secondary prevention of cardiovascular disease. *Cochrane Database Syst Rev* **4**, CD011748 (2017).
36. Bohula, E. A. *et al.* Prevention of Stroke with the Addition of Ezetimibe to Statin Therapy in Patients With Acute Coronary Syndrome in IMPROVE-IT (Improved Reduction of Outcomes: Vytorin Efficacy International Trial). *Circulation* **136**, 2440–2450 (2017).
37. Cannon, C. P. *et al.* Ezetimibe Added to Statin Therapy after Acute Coronary Syndromes. *N. Engl. J. Med.* **372**, 2387–2397 (2015).
38. Do, R. *et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature Genetics* **45**, 1345–1352 (2013).
39. Burgess, S., Freitag, D. F., Khan, H., Gorman, D. N. & Thompson, S. G. Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. *PLoS ONE* **9**, e108891 (2014).
40. Holmes, M. V. *et al.* Mendelian randomization of blood lipids for coronary heart disease. *Eur. Heart J.* **36**, 539–550 (2015).
41. Emerging Risk Factors Collaboration *et al.* Major lipids, apolipoproteins, and risk of vascular disease. *JAMA* **302**, 1993–2000 (2009).

42. Ahmad Zahid *et al.* Inhibition of Angiopoietin-Like Protein 3 With a Monoclonal Antibody Reduces Triglycerides in Hypertriglyceridemia. *Circulation* **140**, 470–486 (2019).
43. Barter, P. J. *et al.* Effects of torcetrapib in patients at high risk for coronary events. *N. Engl. J. Med.* **357**, 2109–2122 (2007).
44. Schwartz, G. G. *et al.* Effects of Dalcetrapib in Patients with a Recent Acute Coronary Syndrome. *New England Journal of Medicine* **367**, 2089–2099 (2012).
45. Lincoff, A. M. *et al.* Evacetrapib and Cardiovascular Outcomes in High-Risk Vascular Disease. *N. Engl. J. Med.* **376**, 1933–1942 (2017).
46. HPS3/TIMI55–REVEAL Collaborative Group *et al.* Effects of Anacetrapib in Patients with Atherosclerotic Vascular Disease. *N. Engl. J. Med.* **377**, 1217–1227 (2017).
47. Ference, B. A., Majeed, F., Penumetcha, R., Flack, J. M. & Brook, R. D. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2 × 2 factorial Mendelian randomization study. *J. Am. Coll. Cardiol.* **65**, 1552–1561 (2015).
48. Ference, B. A. *et al.* Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *New England Journal of Medicine* **375**, 2144–2153 (2016).
49. Talmud, P. J. & Holmes, M. V. Deciphering the Causal Role of sPLA2s and Lp-PLA2 in Coronary Heart Disease. *Arterioscler. Thromb. Vasc. Biol.* **35**, 2281–2289 (2015).
50. Sofat, R. *et al.* Separating the mechanism-based and off-target actions of cholesteryl ester transfer protein inhibitors with CETP gene polymorphisms. *Circulation* **121**, 52–62 (2010).
51. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium *et al.* The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet* **379**, 1214–1224 (2012).

52. Harrison, S. C. *et al.* Interleukin-6 receptor pathways in abdominal aortic aneurysm. *Eur Heart J* **34**, 3707–3716 (2013).
53. Rosa, M. *et al.* A Mendelian randomization study of IL6 signaling in cardiovascular diseases, immune-related disorders and longevity. *npj Genomic Medicine* **4**, 1–10 (2019).
54. Parisinos, C. A. *et al.* Variation in Interleukin 6 Receptor Gene Associates With Risk of Crohn's Disease and Ulcerative Colitis. *Gastroenterology* **155**, 303-306.e2 (2018).

2 | Review of Mendelian Randomisation methods, considerations and applications

Several Mendelian randomisation (MR) methods have been developed to assess causality using data from genetic studies, each of them with distinct strengths and limitations. In this chapter, I will describe the standard Mendelian randomisation model, discuss the instrumental variables assumptions and review the Mendelian randomisation approaches commonly used in the field of genetic epidemiology, and frequently applied in *genome-wide biomarker* MR. In a subsequent section, I describe methods relevant to *drug target* MR.

2.1. Canonical Mendelian randomisation model and assumptions

The canonical Mendelian randomisation model assumes that for each individual *i (i* = 1,..., *N)*, *J* genetic variants *Gij* (*j* = 1, . . . , *J*), a modifiable exposure (X_i), an outcome variable (Y_i), and unknown confounders (U_i). Suppose the *exposure* is defined as a linear function of *J* genetic variants, the unknown confounders and an independent error term (ϵ_i^X), with the coefficient β_{Xj} representing the effects of each genetic variant *j* on the exposure:

$$X_i = \sum_{j=1}^{J} \beta_{Xj} G_{ij} + U_i + \epsilon_i^X \qquad (1)$$

Suppose the *outcome* is defined as a linear function of *J* genetic variants, the exposure, the confounders and an independent error term (ϵ_i^Y). The coefficient α_j represents the direct effect of each genetic variant on the outcome, and $\mu\beta_{Xj}$ the indirect effect via the exposure:

$$Y_i = \sum_{j=i}^{J} \alpha_j G_{ij} + \mu X_i + U_i + \epsilon_i^Y \qquad (2)$$

To be a valid instrumental variable (IV), the genetic variant G_j must hold to the following assumptions:

i) 'Relevance' assumption. The genetic variants must be associated with the exposure of interest (X). This assumption implies $\beta_{Xj} \neq 0$.

ii) ‘Exchangeability’ assumption. There should be no unmeasured confounders of the associations between the genetic variant (G_j) and outcome (Y). This assumption could be violated in the presence of genetic confounding such as population stratification, when there is a systematic difference in allele frequencies between subpopulations in a sample due to different ancestry, and cryptic relatedness when there is unknown or undocumented familial relationships among individuals in the sample[1,2]. Both scenarios should have been controlled for during the genetic association study stage, and thus, they should not impact the MR inference.

iii) ‘Exclusion restriction’ assumption. The variants should affect the outcome only through their effect on the risk factor of interest. This assumption implies $\alpha_j = 0$. It is also known as the ‘no-horizontal pleiotropy’ assumption, where pleiotropy is defined as a situation in which a genetic variant influences multiple traits. If the variant influences multiple traits in the same biological pathway as the exposure it is referred to as ‘vertical pleiotropy’, if it influences multiple traits in independent pathways it is referred to as ‘horizontal pleiotropy’. Whereas horizontal pleiotropy compromises causal inference in a MR analysis, vertical pleiotropy does not.

The model described above, including the instrumental variable assumptions and coefficients from equations (1) and (2) are illustrated in Figure 2.1.

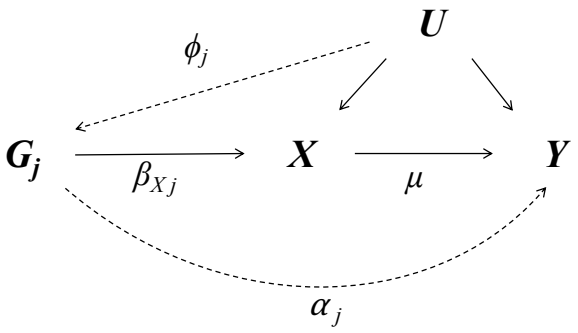


Figure 2.1. Canonical Mendelian randomisation model. Diagram of the model assumed for genetic variant G_j, showing the effect on the exposure X (β_X), the indirect effect on the outcome Y through confounders (ϕ_j), the direct effect on the outcome Y (α_j) and the causal effect of exposure X on the outcome Y (μ). Solid lines indicate instrumental variable assumptions and dashed lines ways these assumptions could be violated.

2.2. Comparison of Mendelian randomisation methods

Based on the data source, two different Mendelian randomisation settings can be defined: one-sample and two-sample MR. One-sample MR is performed when the genetic associations with the exposure and the outcome are from the same population and requires access to individual participant data. This scenario is sensitive to 'winner's curse' bias which can overestimate true causal effects in overlapping samples[3]. Furthermore, it is also subject to 'weak instrument' bias which depends on the strength of the genetic instrument, and arises if the chance difference in confounders explains more of the variation in the outcome than the association of the genetic instrument with the exposure[4]. The increasing availability of genetic summary data allows the evaluation of causality using genetic associations from independent studies (two-sample MR) under the assumption that the associations are derived from the same underlying population and adjusted for the same covariates[5]. By using separate studies, the statistical power in the two-sample MR scenario increases due to the possibility of obtaining more precise estimates of the genetic associations with the outcome[6]. Several Mendelian

randomisation methods have been developed to evaluate causality in both settings, the most commonly used will be briefly described in the following section and summarised in Table 2.1.

2.2.1. Wald method

The simplest Mendelian randomisation method is the Wald method or ratio estimator method in which a single variant is used in the genetic instrument[7]. In this case, the outcome is defined as:

$$Y = (\alpha_j + \mu \beta_{Xj}) G_j + \epsilon'_j{}^{Y} \quad (3)$$

The Wald ratio is estimated as the coefficient from regressing the outcome (3) on the genetic variant ($\hat{\beta}_{Yj}$) divided by the coefficient from regression of the exposure on the variant ($\hat{\beta}_{Xj}$) :

$$\hat{\beta}_{Wald} = \frac{\hat{\beta}_{Yj}}{\hat{\beta}_{Xj}} \quad (4)$$

If the genetic variant is a valid instrumental variable, then $\alpha_j = 0$ and the casual effect of the exposure on the outcome $(\alpha_j + \mu \beta_{Xj}) / \beta_{Xj} = \mu$. This estimate can be interpreted as a μ change in the outcome for one unit increase in the exposure.

2.2.2. Inverse-variance weighted (IVW) method

When multiple *uncorrelated* genetic variants satisfy the IV assumptions, an optimisation of the Wald ratio allows to include all of them in a single analysis to maximise the power to detect a causal effect[8]. The estimate is a weighted average of the ratio estimates for *J* genetic variants (inverse-variance weighted estimate):

$$\hat{\beta}_{IVW\ uncorr} = \frac{\sum_{j=i}^{J} \omega_j \ \hat{\beta}_{Wald\,j}}{\sum_{j=i}^{J} \omega_j} \tag{5}$$

where the weights (ω_j) are derived from the first-order term of the delta expansion of the variance[9] and represent the inverse-variance of the ratio estimates:

$$\omega_j = \frac{\hat{\beta}_{Xj}^{2}}{se\left(\hat{\beta}_{Yj}\right)^{2}} \tag{6}$$

If the association of the genetic variant G_j with the exposure is $\hat{\beta}_{Xj}$ with standard error $se\ (\hat{\beta}_{Xj})$, and with the outcome is $\hat{\beta}_{Yj}$ with standard error $se\ (\hat{\beta}_{Yj})$, then the causal estimate derived from expanding the formula for the weights (6) into the equation (5) is:

$$\hat{\beta}_{IVW\ uncorr} = \frac{\sum_{j=i}^{J} \hat{\beta}_{Xj}^{2}\, se\left(\hat{\beta}_{Yj}\right)^{-2}\, \hat{\beta}_{Wald\,j}}{\sum_{j=i}^{J} \hat{\beta}_{Xj}^{2}\, se\left(\hat{\beta}_{Yj}\right)^{-2}} = \frac{\sum_{j=i}^{J} \hat{\beta}_{Xj}\, se\left(\hat{\beta}_{Yj}\right)^{-2}\, \hat{\beta}_{Yj}}{\sum_{j=i}^{J} \hat{\beta}_{Xj}^{2}\, se\left(\hat{\beta}_{Yj}\right)^{-2}} \tag{7}$$

The IVW estimate is equivalent to the two-stage least square estimate with summary data, where the exposure is regressed on the genetic instrument in a first stage regression and the outcome is regressed on the fitted values[10]. When the genetic variants are *correlated*, the method can be extended to account for their correlation using a weighting matrix (Ω) where $\rho_{j1\,j2}$ is the correlation coefficient between variants *j1* and *j2*[11] :

$$\Omega_{j1\,j2} = se\left(\hat{\beta}_{Yj1}\right) se\left(\hat{\beta}_{Yj2}\right) \rho_{j1\,j2} \tag{8}$$

with $\hat{\beta}_X$ and $\hat{\beta}_Y$ as the genetic associations vectors for the exposure and outcome respectively, and ^{T} the transpose vector, the IVW estimate accounting for correlation is defined as:

$$\hat{\beta}_{IVW\,corr} = \frac{\Omega\ \hat{\beta}_X{}^T\ \hat{\beta}_Y}{\Omega^{-1}\ \hat{\beta}_X{}^T\ \hat{\beta}_X} \tag{9}$$

If all the genetic variants are valid IVs, the IVW estimator provides the most precise estimates across all the MR methods.

2.2.3. Principal component analysis - IVW method

If too many correlated variants are included in the IVW model, even accounting for the correlation can lead to numerical instabilities and inflated Type 1 error rates[12]. These issues can occur due to inconsistencies in the data (i.e. rounding of association estimates) and near-singular correlation matrices, which result in the model failing, misleading estimates and/or over-precision in the causal estimates.

A method based on principal component analysis has been developed to allow the inclusion of multiple correlated variants under the assumption that all the variants are estimated in the same sample size[12]. In this situation, the IVW model uses a weighted version of the genetic correlation matrix,

$$\Psi_{j1\,j2} = \hat{\beta}_{Xj1}\,\hat{\beta}_{Xj2}\,se\left(\hat{\beta}_{Yj1}\right)^{-1} se\left(\hat{\beta}_{Yj2}\right)^{-1}\rho_{j1\,j2} \tag{10}$$

Then, the first principal component is a linear combination of the variants explaining the largest proportion of variance in the exposure. This method implies the choice of a threshold of variance to define the number of principal components in the weighting correlation matrix. The causal effect is estimated using the IVW method with the transformed vectors of genetic associations and the transformed correlation matrix as indicated in the following expression, where W_K is the matrix constructed for the first K principal components:

$$\hat{\beta}_{PCA\,IVW} = \frac{\left(W_K{}^T\,\Omega\,W_K\right)\left(W_K{}^T\,\hat{\beta}_X{}^T\right)\left(W_K{}^T\,\hat{\beta}_Y\right)}{\left(W_K{}^T\,\hat{\beta}_X\right)\left(W_K{}^T\,\Omega\,W_K\right)^{-1}\left(W_K{}^T\,\hat{\beta}_X{}^T\right)} \tag{11}$$

This method is suitable for highly correlated variants (e.g. fine-mapped genetic data), and results in estimates more robust than the ones derived from methods that LD prune instead, however these are less precise[12].

2.2.4. Methods with invalid genetic instruments

The methods described in the subsequent sections aim to estimate the causal effect when genetic variants are 'invalid' instruments due to the presence of horizontal pleiotropy.

2.2.4.1. Median-based method

The median-based method provides a consistent estimate of the causal effect even if up to 50% of the variants in the instrument are invalid ('majority valid' assumption)[13]. There are three different modalities: the simple, the weighted and the penalized weighted median estimator. The simple median estimator is the median of the Wald ratio of the variants. To account for variability in the precision of the individual estimates, the weighted median estimator uses the inverse of the variances of the ratio estimates as the weights. Being ω_j the weight for the j-th ordered ratio, the weighted-median estimator is the median of a distribution having estimate $\hat{\beta}_j$ as its ρ_j – th percentile:

$$\rho_j = 100 \left(\sum_{k=1}^{j} \omega_K - \omega_j/2 \right) \tag{12}$$

The penalized weighted median estimator down-weights the contribution of genetic variants with outlying Wald ratios. The Cochran's Q statistics (Q) is used to quantify the heterogeneity[14]:

$$Q = \sum_j Q_j = \sum_j \omega_j \left(\hat{\beta}_j - \hat{\beta}_{IVW} \right) \tag{13}$$

Cochran's Q statistics follows a chi-squared distribution with $J - 1$ degrees of freedom under the assumption that all variants are valid IVs and show the same causal effect (i.e., the *j*-th contribution to Q, Q_j, is approximately chi-squared distributed on 1 degree of freedom). In the penalized method, outlying variants are down-weighted by multiplying the inverse-variance weights by the one-sided upper p value on a chi-squared distribution corresponding to Q_j, multiplied by 20 (or by 1 if the p value > 0.05).

This method is robust to outliers, as the median of the distribution is not affected by the magnitude of the ratio estimates. However, it is sensitive to changes in the selection of variants when constructing the genetic instrument.

2.2.4.2. Mode-based method

The mode-based method obtains the mode of the ratio estimates if the true causal effect is the value taken for the largest number of genetic variants ('plurality valid' assumption)[15]. Since in finite samples the mode does not exist, this method generates a normal density for each genetic variant centred around the ratio estimate. The spread of the density depends on a bandwidth parameter and, in the case of the weighted mode estimator, the precision of the ratio estimate. The causal estimate is the maximum point of a smoothed density function constructed by adding the normal densities of all variants.

Similar to the median-based method, the mode-based estimator is robust to pleiotropic outliers, however the causal estimates are influenced by the selection of variants. In addition, the mode-based estimator requires the choice of a value for the bandwidth parameter.

2.2.4.3. MR-Egger regression method

MR-Egger regression provides consistent causal estimates even in the presence of invalid instruments under the assumption that the association of each variant with the exposure is independent of the strength of the pleiotropic effects α_j ('Instrument Strength Independent of Direct Effect (InSIDE)' assumption)[16,17]. This model requires all the genetic associations with the exposure orientated in the positive direction and uses the inverse-variance of the ratio estimates as the weights in the regression. A non-zero intercept term (β_{0E}) is allowed in the linear regression which can be interpreted as the average pleiotropic effect of all J genetic variants.

$$\hat{\beta}_Y = \beta_{0E} + \beta_{1E}\hat{\beta}_{Xj} + \epsilon_j^Y \quad (14)$$

If the average pleiotropic effect is zero, referred to as 'balanced horizontal pleiotropy', then the MR Egger estimate β_{1E} will equal the IVW estimate. If there is directional horizontal pleiotropy or the InSIDE assumption is violated, the intercept term will differ from zero indicating that the IVW estimate is biased. As this estimator is a modification of the IVW method, it can also be extended to account for the correlation between genetic variants using a weighting correlation matrix.

Under the InSIDE assumption, the MR Egger method estimates a consistent causal effect even when all the genetic variants are invalid IVs. However, it is sensitive to outliers and provides less precise causal estimates due to the variability between the genetic associations with the exposure.

2.2.4.4. Multivariable Mendelian randomisation method

The multivariable Mendelian randomisation method (MVMR) is an extension of the IVW and MR-Egger estimators that uses genetic variants associated with multiple exposures to estimate the causal relevance of each exposure in a single model (Fig. 2.2)[10,18]. To be included in the instrument, a genetic variant must adhere to the following rules:

i. It is associated with at least one of the exposures.

ii. It is not associated with a confounder of any of the exposure–outcome associations.

iii. It is conditionally independent of the outcome given the exposure and confounders.

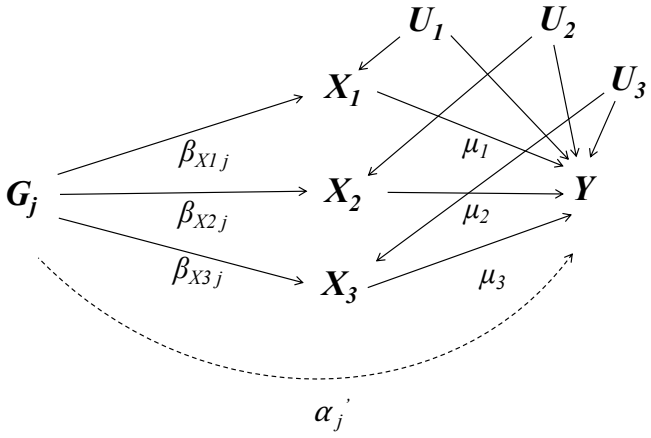


Figure 2.2. Diagram of the multivariable model. Model assumed for genetic variant G_j, showing the effect on three exposures X_1 ($\beta_{X1\,j}$), X_2 ($\beta_{X2\,j}$), X_3 ($\beta_{X3\,j}$), the direct effect on the outcome Y (α'_j) and the three causal effect of exposures on the outcome Y (μ_1, μ_2, μ_3). Solid lines indicate instrumental variable assumptions and dashed lines ways these assumptions could be violated.

In a three exposure scenario, the multivariable IVW method estimates the causal effects using a multivariable weighted linear regression of the genetic association estimates, with the intercept set to zero and the inverse variance weights $se\left(\hat{\beta}_{Yj}\right)^{-2}$:

$$\hat{\beta}_Y = \mu_{1MI}\hat{\beta}_{X1j} + \mu_{2MI}\hat{\beta}_{X2j} + \mu_{3MI}\hat{\beta}_{X3j} + \epsilon_{MIj} \quad (15)$$

It can also be extended to multivariable MR Egger by allowing for a intercept term (μ_{0ME}):

$$\hat{\beta}_Y = \mu_{0ME} + \mu_{1ME}\hat{\beta}_{X1j} + \mu_{2ME}\hat{\beta}_{X2j} + \mu_{3ME}\hat{\beta}_{X3j} + \epsilon_{MEj} \quad (16)$$

Since these methods are based on the univariable IVW estimator, they can account for correlation between genetic variants using a weighting correlation matrix.

The multivariable MR method accounts for measured pleiotropy (and unmeasured pleiotropy in the case of MVMR MR Egger) by evaluating the causal effect of multiple exposures in a single regression analysis even if none of the genetic variants are uniquely associated with one of the exposures. The multivariable extension of the IVW and MR Egger methods is sensitive to 'weak instrument bias' due to the inclusion of multiple variants not strongly associated with the exposures in the model, and the precision in the estimates is affected when using highly correlated exposures.

Table 2.1. Commonly used Mendelian Randomisation approaches with large number of genetic variants using summary data.

Method	Assumption	Potentials	Limitations	Ref.
Inverse-variance weighted (IVW)	All variants are valid IVs	•Allows for correlated variants • Provides precise estimates	• Biased in the presence of directional pleiotropy	8–11
PCA IVW	Associations are estimated in the same sample size	• Allows for highly correlated variants • Robust to variable selection	• Biased in the presence of directional pleiotropy • Less precise than IVW	12
MR Egger	InSIDE assumption	• Allows for correlated variants • Reliable when all variants are invaild IVs	• Sensitive to outliers • Imprecise estimates	16,17
Median-based	'Majority valid' assumption	• Robust to outliers	• Sensitive to the choice of genetic variants	13
Mode-based	'Plurality valid' assumption	• Robust to outliers	• Sensitive to the choice of genetic variants and bandwidth parameter • Generally conservative	15
Multivariable IVW	Any association with the outcome is via the measured exposures	• Allows for correlated variants • Accounts for measured pleiotropy	• Susceptible to 'weak instrument' bias • Sensitive to highly correlated exposures	18
Multivariable MR Egger	InSIDE assumption must hold for all measured exposures	• Allows for correlated variants • Accounts for measured and unmeasured pleiotropy	• Susceptible to 'weak instrument' bias • Sensitive to highly correlated exposures • Imprecise estimates	18

2.2.5. Other methods

Several additional Mendelian randomisation approaches have been developed to overcome some of the limitations of the methods described in the previous section. However, most of them have not been as commonly used in applied examples as the methods described earlier.

For instance, the **contamination mixture method** provides a consistent estimates under the 'plurality valid' assumption by constructing a likelihood function based on the ratio estimates and assuming that the values estimated by invalid instruments are normally distributed around zero with a large standard deviation[19]. While apparently robust to outliers, this method is particularly sensitive to the choice of the standard deviation parameter.

The **MR-Pleiotropy Residual Sum and Outlier (MR-PRESSO**) performs in a IVW framework by removing genetic variants based on a heterogeneity measurement until all the variants have similar estimates[20]. It inherits the precision of the IVW method but it is more time-consuming than other methods and unstable when multiple variants are pleiotropic.

One of the most recent methods is the **multivariable MR approach based on Bayesian model averaging (MR-BMA)** which is optimised for analyses with high-dimensional sets of potential risk factors[21]. It performs a Bayesian variable selection step before the weighted regression model and computes the marginal inclusion probability for each exposure (i.e. the sum of the posterior probabilities over all models where the exposure is present). While it allows the selection of causal risk factors from a large set of variables, it is influenced by the choice of parameters and assumes that the proportion of true causal exposures compared with all potential exposures is small. The developers also highlighted that the causal estimates should not be interpreted absolutely and rather be used to compare exposures or to interpret direction of effects.

2.3. Detecting and accounting for heterogeneity in Mendelian randomisation

In section 2.2.4.1, the Cochran's Q statistic was introduced to assess heterogeneity and detect pleiotropy based on the assumption that valid IVs should follow, asymptotically, a chi-squared distribution, with degrees of freedom (df) equal to the number of genetic variants minus 1[14,22]. If a genetic variant shows excessive heterogeneity, this could indicate the violation of the 'no-horizontal pleiotropy' assumption. For example, genetic variants in or near *APOE* gene are associated with LDL-C as well as very strongly associated with Alzheimer's disease (AD). In MR studies using variants across the genome to estimate the lipid effect on AD risk, SNPs in this locus showed large heterogeneity and they were excluded based on their established pleiotropic effect on AD risk[23]. However, if there is heterogeneity due to pleiotropy but the InSIDE assumption holds and the pleiotropy is balanced, then the IVW estimator under a random-effects model instead of the fixed-effects model can be used to account for the additional uncertainty due to pleiotropy[22]. When the InSIDE assumption holds, but there is directional pleiotropy, the MR Egger method can be used to estimate the mean pleiotropic effect and provide a reliable causal estimate, as described in section 2.2.4.3.

It is possible to test for residual heterogeneity in the MR-Egger model using an extended version of the Cochran's Q statistic, known as Rücker's Q' statistic[24,25]. The Rücker model-selection framework (Fig. 2.3) uses both statistical values to inform the selection of fixed-effect IVW, random-effects IVW , fixed-effect MR-Egger, random-effects MR-Egger models based on their goodness of fit. This hierarchical framework involves the following steps:

i) An initial IVW analysis under a fixed-effect model is performed and the Cochran's Q statistic (Q) calculated.

ii) A random-effects IVW model is preferred over the fixed-effect model if Q reveals sufficient heterogeneity at significance level δ (e.g. 0.05) with respect to a chi-squared distribution with degrees of freedom equal to the number of genetic variants minus 1.

iii) A fixed-effect MR-Egger analysis is performed and the Rücker's Q' statistic (Q') calculated. If the difference Q-Q' is significant at level δ with respect to a chi-squared distribution with degrees of freedom equal to the number of genetic variants minus 2, this model is selected.

iv) A random-effects MR-Egger model is selected if Q' still reveals sufficient heterogeneity at significance level δ with respect to a chi-squared distribution with degrees of freedom equal to the number of genetic variants minus 2.

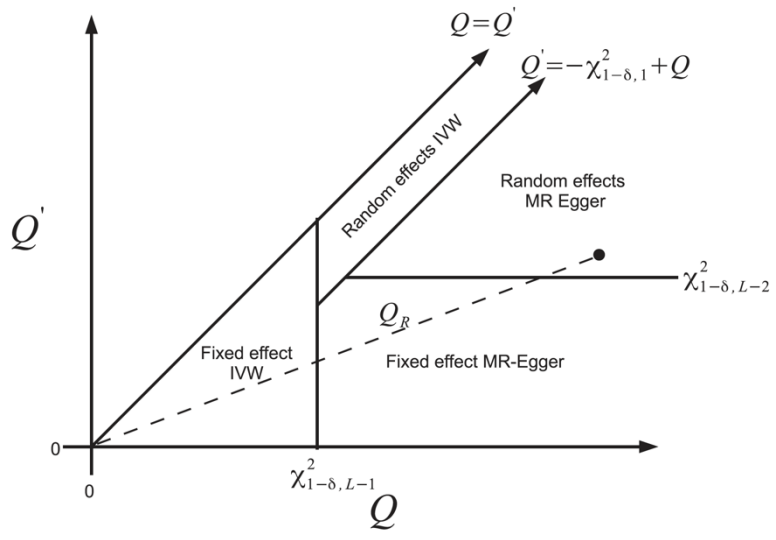


Figure 2.3. Illustration of the Rücker model-selection framework. The two dimensional space is defined by Q, Q', *L* genetic variants and a significance threshold for detecting pleiotropy δ. From Bowden *et al.*, 2018[22].

The Rücker model-selection framework is an automatic statistical method that favours the IVW model and recommends the MR–Egger model only when there is an improvement of the goodness of fit of the data when this approach is used. While it is a systematic and fast approach to choose between competing MR models, the uncertainty about the optimal model still remains. A Bayesian model averaging framework has been developed to account for model uncertainty in posterior causal estimates[26], however, it is sensitive to the choice of priors and more computationally expensive. Another suggested approach is the mixture-of-experts machine learning framework[27] (MR-MoE 1.0) which is trained using random forest decision trees, however, it can lead to high type 1 error rates as has been observed in other data driven approaches[28].

Other statistical measurements besides Cochran's Q statistic and Rücker's Q' statistic have been suggested to detect outliers in regression models[29], and their rationale is that variants with excessive contribution to the model can be identified based on their effect on the regression ('leverage'). Variants with high leverage can influence the regression model and provide misleading causal estimates.

2.4. Molecular traits in drug target Mendelian randomisation

Many of the traits studied by GWAS are diseases, clinically relevant biomarkers and quantitative phenotypes such as expression quantitative trait loci (eQTL), metabolite-levels quantitative trait loci (mQTL) or protein-levels quantitative trait loci (pQTL). Numerous publicly available GWAS summary estimates of eQTLs are available, for example, GTEX[30] with a total of 11,688 samples and 53 tissues across 714 donors, or eQTLGen[31] with 31,684 blood samples from healthy individuals. Recently, GWAS of circulating proteins (pQTLs) have become available such as the Interval study (~3,000 proteins)[32] and the SCALLOP Consortium (~1,000 proteins)[33]. These data provide estimates for a substantial proportion of the encoded human proteome, the latest assays from SomaLogic[34] cover ~7,000 proteins (SomaLogic 7k panel) including some potential cardiovascular targets such as CETP.

Crucially, the summary estimates from many of these studies are publicly available, with novel MR techniques able to use these summary level data as inputs for analysis. While the increase in GWAS sample sizes has boosted the power of MR studies in binary traits, genetic associations with molecular quantitative trait loci, particularly pQTLs, provide a valuable resource for drug target MR analyses as proteins are the targets of most drugs. In the absence of pQTL or protein activity data, eQTL associations can be used to weight instruments in a drug target MR analysis, where the major caveat is deciding on the relevant tissue for a particular disease. Therefore, the raw material now exists for large scale drug target validation analyses.

2.4.1. Additional considerations for defining genetic instrumental variables using molecular exposures

In addition to the assumptions discussed in section 2.1, each Mendelian randomisation setting requires careful selection of the parameters that define the genetic instrument. Since Mendelian randomisation for drug target validation is framed as a *cis*-focused analysis (i.e. the exposure of interest is the protein encoded by a specific gene or a proxy of the protein's function or level), and explores the effect of modifying a particular protein target pharmacologically, the instrument selection is different from MR for validating the causal relevance of other exposures (e.g. disease biomarkers such as blood lipids). Furthermore, it comprises additional challenges and choices, for instance, defining the locus of interest; selecting and accounting for linkage disequilibrium between genetic variants; and selecting the exposure used to weight the effect of the genetic instruments on the disease risk[35].

One consideration that applies to both genome-wide biomarker and drug target MR settings is the *p* value threshold for genetic associations used to identify potential instruments. Yet, there is no consensus concerning the optimal threshold. The thresholds employed vary from very conservative cut-offs (e.g. *p* value $\leq 5 \times 10^{-8}$) to less stringent thresholds (e.g. *p* value $\leq 10^{-5}$). The latter often results in improved performance, particularly in the *cis*-MR setting[36,35], and could be justified if there are strong priors and/or the burden of multiple testing is reduced compared to a GWAS where the *p* value threshold is typically 5×10^{-8}. While the statistical power is maximised using methods that include multiple genetic instruments, they usually involve a first LD clumping step to remove highly correlated variants. Despite some evidence showing that high LD thresholds lead to numerical instabilities[35,12], an agreement on the choice of a general LD threshold has also not been reached yet. An extra complexity arises when using multiple correlated variants, since the modelling of the remaining pairwise LD requires the

selection of a LD-reference panel. Resources such as UK Biobank where individual level data is available for thousands of samples, are likely to improve the accuracy of the modelling compared to previous studies based on 1000 genomes populations[37]. Such resources also provide more precise allele frequencies, where a minor allele frequency (MAF) threshold of 0.01 is usually used to define common variants.

Several intermediate traits, such as lipid blood levels, have been previously used to inform drug target validation. However, since over 90% of drug targets are proteins[38], weighting by protein levels or activity in a disease-relevant tissue would provide the most informative *cis*-MR analysis for drug target validation. Since some drugs are designed to target circulating proteins (e.g. PCSK9 inhibitors), and these can now be measured by high throughput proteomics technologies, opportunities for *cis*-MR analysis are increasing. Figure 2.4 illustrates the protein-weighted MR model.

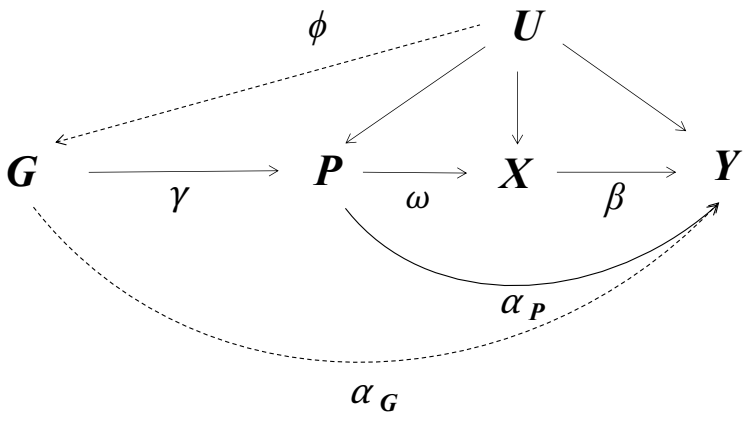


Figure 2.4. Protein weighted MR model. Diagram of the model assumed for genetic variant *G*, showing the direct effect on the protein *P*, the indirect effect on *X* (γ), the indirect effect on the outcome *Y* through confounders (ϕ), the direct effect on the outcome *Y* (α) and the causal effect of exposure *X* on the outcome *Y* (β). Solid lines indicate instrumental variable assumptions and dashed lines ways these assumptions could be violated. Adapted from Schmidt *et al.,* 2020[35].

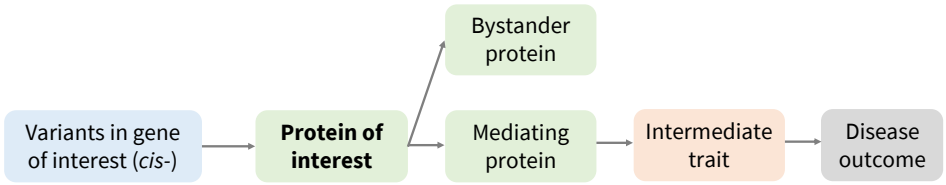
By selecting genetic variants in-and-around the gene encoding the protein of interest, the *cis*-MR analysis of proteins is less prone to violation of the horizontal pleiotropy assumption. The rationale for this was presented by Schmidt *et al*., 2020[35] and illustrated in Figure 2.5. In the first scenario (Fig. 2.5a), the protein of interest is instrumented by using genetic variants in its encoding gene (*cis*-). In this example, the genetic variants associate with multiple proteins on the same biological pathway, where the protein instrumented is upstream of all of the other proteins in the causal pathway. It illustrates how valid instruments for *cis*-MR can also have, and indeed would oftentimes be expected to also have, *trans*- effects. In the second scenario (Fig. 2.5b), the protein of interest is instrumented by using genetic variants in the other genes that are associated with the level of the protein of interest (*trans*-). The genetic variants associate with multiple proteins on the same biological pathway, where the protein instrumented is in the causal pathway. The effect on the outcome is still through the instrumented protein and thus, the *trans*-MR analysis provides the correct inference. In the third scenario (Fig. 2.5c), the protein of interested is also instrumented by using genetic variants in the other genes (*trans*-). However, the genetic variants associate with multiple proteins on different biological pathways, where the protein instrumented is not in the causal pathway. Here, the association of the *trans*-variants with the instrumented protein is due to horizontal pleiotropy and any inference about a causal association of the protein of interest with the disease outcome is erroneous.

While a *cis*-MR approach reduces the potential for misleading inferences due to horizontal pleiotropy, defining the locus of interest and the size of the surrounding *cis*-genetic region are additional challenges that can impact MR performance, as neighbouring genes can lead to pleiotropy effects due to LD. Again, defining a standard region that is generalisable to all the genes in the genome and is able to capture accurately all variants involved in expression,

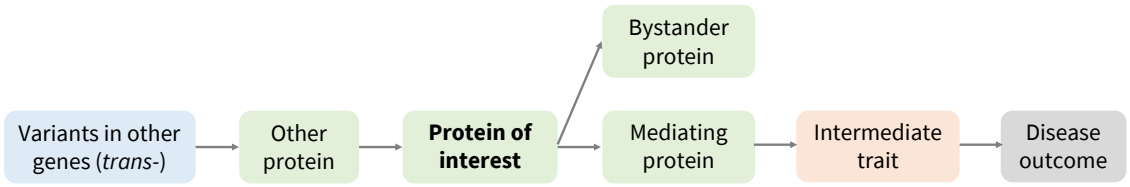
regulation and function is not possible, and pairwise grid-searches for each exposure and outcome have been proposed to select the optimal region size for each gene[35].

While tissue-relevant protein QTL (pQTL) data is not currently available, many of the circulating proteins measured by existing proteomics platforms are the actual targets for many approved or developmental therapeutics (e.g. from ~ 2036 druggable proteins in SomaLogic v4 or 973 in O-link[35]). Previous drug target MR analyses weighting blood protein levels of F10, of interleukin-12 subunit beta (IL12B) and plasminogen (PLG) have shown that MR with proteomics data has potential for genetic target validation through direct assay of the efficacy target. For example, the drug target MR analysis of circulating F10 recapitulated the mechanism of action of F10 inhibitors in stroke prevention in patients with atrial fibrillation[35,39]. Similarly, higher circulating concentration of IL12B and PLG were associated with higher risk of Crohn's disease and lower risk of ischaemic stroke, respectively[35]. Both drug target MR analyses rediscovered the mechanism of action of the approved drugs for these indications[40,41].

a. Vertical pleiotropy scenario when using *cis*-genetic variants



b. Vertical pleiotropy scenario when using *trans*-genetic variants



c. Horizontal pleiotropy scenario when using *trans*-genetic variants

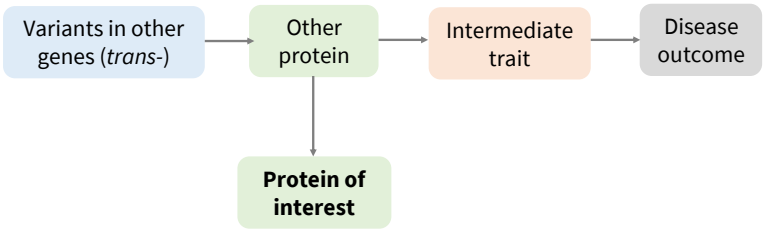


Figure 2.5. Paradoxical scenarios in protein Mendelian Randomisation. a. An example of protein MR using genetic variants in the encoding gene (*cis*-). The genetic variants associate with multiple proteins on the same biological pathway, where the protein instrumented is upstream of all of the other proteins in the causal pathway. b. An example of protein MR using genetic variants in another gene (*trans*-). The genetic variants associate with multiple proteins on the same biological pathway, and the protein instrumented is in the causal pathway. The effect on the outcome is still through the instrumented protein and thus, the trans-MR analysis provides the correct inference. c. An example of protein MR using genetic variants in another gene (*trans*-). The genetic variants associate with multiple proteins on different biological pathways, but the protein instrumented is not in the causal pathway. Here, the association of the *trans* variants with the instrumented protein is due to horizontal pleiotropy and any

inference that there is a causal association of the protein of interest with the disease outcome is erroneous. Figure adapted from Schmidt *et al.,* 2020[35].

Even in the absence of relevant pQTL data, a protein can remain the inferential target in a *cis*-MR setting by weighting the analysis using an intermediate trait positioned downstream between the protein and the disease. In such circumstances, the intermediate biomarker is known to be altered by the perturbation of the protein of interest. For example, GWAS on blood lipids levels have been used to genetically validate drug targets such as PCSK9[42] for CHD prevention. Later, the causal effect anticipated by the *cis*-MR analysis using LDL-C as an intermediate phenotype was confirmed using PCSK9 pQTL measurements when protein level data became available[35].

All the parameters discussed in this section, in addition to the general MR and method-specific assumptions, should be carefully scrutinised before constructing the genetic instrument. The setting (i.e. biomarker or drug target MR) as well as the exposure type (i.e. pQTL, eQTL or intermediate traits) should guide the choice of these parameters.

2.5. References

1. Nitsch, D. *et al.* Limits to Causal Inference based on Mendelian Randomisation: A Comparison with Randomized Controlled Trials. *Am J Epidemiol* **163**, 397–403 (2006).
2. Swanson, S. A., Tiemeier, H., Ikram, M. A. & Hernán, M. A. Nature as a trialist? Deconstructing the analogy between Mendelian Randomisation and randomized trials. *Epidemiology* **28**, 653–659 (2017).
3. Taylor, A. E. *et al.* Mendelian randomisation in health research: using appropriate genetic variants and avoiding biased estimates. *Econ Hum Biol* **13**, 99–106 (2014).
4. Burgess, S. & Thompson, S. G. Avoiding bias from weak instruments in Mendelian randomisation studies. *Int J Epidemiol* **40**, 755–764 (2011).
5. Aromataris, E. *et al.* Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. *Int J Evid Based Healthc* **13**, 132–140 (2015).
6. Burgess, S., Scott, R. A., Timpson, N. J., Davey Smith, G. & Thompson, S. G. Using published data in Mendelian randomisation: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol* **30**, 543–552 (2015).
7. Wald, A. The Fitting of Straight Lines if Both Variables are Subject to Error. *Ann. Math. Statist.* **11**, 284–300 (1940).
8. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomisation analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
9. Thomas, D. C., Lawlor, D. A. & Thompson, J. R. Re: Estimation of bias in nongenetic observational studies using 'Mendelian triangulation' by Bautista et al. *Ann Epidemiol* **17**, 511–513 (2007).

10. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomisation: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).
11. Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple instrumental variables in Mendelian randomisation: comparison of allele score and summarized data methods. *Stat Med* **35**, 1880–1906 (2016).
12. Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B. B. & Hopewell, J. C. Mendelian randomisation with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. *Genetic Epidemiology* **41**, 714–725.
13. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomisation with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304–314 (2016).
14. Greco M, F. D., Minelli, C., Sheehan, N. A. & Thompson, J. R. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat Med* **34**, 2926–2940 (2015).
15. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomisation via the zero modal pleiotropy assumption. *Int J Epidemiol* **46**, 1985–1998 (2017).
16. Kolesár, M., Chetty, R., Friedman, J., Glaeser, E. & Imbens, G. W. Identification and Inference With Many Invalid Instruments. *Journal of Business & Economic Statistics* **33**, 474–484 (2015).
17. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomisation with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512–525 (2015).

18. Rees, J. M. B., Wood, A. M. & Burgess, S. Extending the MR-Egger method for multivariable Mendelian randomisation to correct for both measured and unmeasured pleiotropy. *Stat Med* **36**, 4705–4718 (2017).
19. Burgess, S., Foley, C. N., Allara, E., Staley, J. R. & Howson, J. M. M. A robust and efficient method for Mendelian randomisation with hundreds of genetic variants. *Nature Communications* **11**, 1–11 (2020).
20. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomisation between complex traits and diseases. *Nature Genetics* **50**, 693–698 (2018).
21. Zuber, V., Colijn, J. M., Klaver, C. & Burgess, S. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomisation. *Nature Communications* **11**, 1–11 (2020).
22. Bowden, J. *et al.* Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomisation via the Radial plot and Radial regression. *International Journal of Epidemiology* **47**, 1264–1278 (2018).
23. Williams, D. M., Finan, C., Schmidt, A. F., Burgess, S. & Hingorani, A. D. Lipid lowering and Alzheimer disease risk: A mendelian randomisation study. *Annals of Neurology* **87**, 30–39 (2020).
24. Rücker, G., Schwarzer, G., Carpenter, J. R., Binder, H. & Schumacher, M. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics* **12**, 122–142 (2011).
25. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomisation. *Stat Med* **36**, 1783–1802 (2017).
26. Thompson, J. R. *et al.* Mendelian randomisation incorporating uncertainty about pleiotropy. *Stat Med* **36**, 4627–4645 (2017).

27. Hemani, G. *et al.* Automating Mendelian randomisation through machine learning to construct a putative causal map of the human phenome. *bioRxiv* 173682 (2017) doi:10.1101/173682.
28. N, I., B, K., Jb, Z. & W, H. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods* **13**, 577–580 (2016).
29. Yu, C. & Yao, W. Robust linear regression: A review and comparison. *Communications in Statistics - Simulation and Computation* **46**, 6261–6282 (2017).
30. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).
31. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018) doi:10.1101/447367.
32. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
33. The SCALLOP Consortium. https://www.olink.com/scallop/. *Olink*.
34. Somalogic platform. http://somalogic.com/.
35. Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nature Communications* **11**, 3255 (2020).
36. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
37. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
38. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940 (2018).
39. López-López, J. A. *et al.* Oral anticoagulants for prevention of stroke in atrial fibrillation: systematic review, network meta-analysis, and cost effectiveness analysis. *BMJ* **359**, j5058 (2017).

40. Sandborn, W. J. *et al.* Ustekinumab Induction and Maintenance Therapy in Refractory Crohn's Disease. *New England Journal of Medicine* **367**, 1519–1528 (2012).
41. IST-3 collaborative group *et al.* The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. *Lancet* **379**, 2352–2363 (2012).
42. Schmidt, A. F. *et al.* PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study. *Lancet Diabetes Endocrinol* **5**, 97–105 (2017).

3 | Methods: An overview

This chapter provides an overview of the datasets, exposure and outcome phenotype measures used throughout this thesis. Detailed methods are discussed further in the relevant results chapters.

3.1. Data sources

3.1.1. Human diseases

To estimate the disease coverage, overlap and divergence of human genetic studies and pharmaceutical research and development (Chapter 4), the total number of human diseases was estimated using information from the following disease classification systems and ontologies as of the 30th November 2021: ICD-10, ICD-11, Human Disease Ontology (DO)[1], Medical Subject Headings (MeSH)[2], Human Phenotype Ontology[3], Clinical Classification Software[4], PheWAS Catalog[5], SNOMED CT[6]. For the MeSH terminology, MeSH terms falling within the categories C (diseases) or F (Psychiatry and Psychology) were selected. Since the number of terms in the DO is updated regularly, the rationale described in previous studies[7] was followed and a figure of 10,901 (i.e., disease terms in the DO as of 30 November 2021) was proposed as a reasonable estimate of the number of common human diseases with genetic susceptibility. To facilitate further mappings to estimate the overlap between all human diseases, disease studied by genome-wide association studies (GWAS) and diseases investigated in pharmaceutical research and development, disease terms were mapped to Unified Medical Language System (UMLS)[8] concepts using the UMLS2020AA. The UMLS was selected as the anchoring coding system as it integrates several medical vocabularies to enable interoperability between data sources and facilitate the link between terms from different ontologies. Further details on the UMLS system are provided in section 3.3 and in the succeeding results chapters.

3.1.2. Genetic association data

Data from GWAS were used throughout the thesis to estimate the disease coverage, overlap and divergence of human genetic studies and pharmaceutical research and development (Chapter 4), to estimate the genetic support of approved drug target-indication pairings (Chapter 5) and as the exposure and outcome data in Mendelian Randomisation analyses (Chapter 6 and 7).

Several public repositories exist that systematically catalogue, curate and store GWAS summary statistics. In this thesis, the European Bioinformatics Institute (EMBL-EBI) GWAS Catalog v1.0.2[9] was used to extract diseases studied by GWAS and download summary statistics for biomarkers and diseases used in the drug target MR analyses. The collection of traits in the GWAS Catalog was enriched by adding summary statistics of GWAS performed in the UK Biobank and available through Neale data (GWAS Round 2, Results shared 1st August 2018[10]), and summary statistics from the University College London–Edinburgh-Bristol (UCLEB) Consortium[11].

Genetic associations with protein quantitative trait locus (pQTL) were used as the exposure data in the drug target MR analyses performed in Chapter 6. GWAS data on pQTL was accessed through an established collaboration with Claudia Langenberg's group at the Medical Research Council (MRC) Epidemiology Unit in Cambridge, and included 10,078 samples who were participants in the Fenland study assayed using the SomaLogic proteomic platform (SomaLogic v4 panel). This technology utilises short single-stranded oligonucleotides ('SOMAmers') that bind with high affinity and specificity to a variety of proteins and enable the quantification of their levels. The SomaLogic v4 platform included 5,284 SOMAmers. Following the company advice, 373 SOMAmers were excluded due to lack of specificity or incorrect SOMAmer– protein mapping.

In. addition, genetic associations with protein activity was correlated to pQTL data in Chapter 6 to illustrate the potential of pQTL-weighted drug target MR approach when GWAS data on protein activity or function is not available. Genetic associations with Butyrylcholinesterase (BCHE) were sourced from a published GWAS[12] and those with coagulation factor VII activity data were obtained from the UCLEB Consortium.

Lastly, in Chapter 7, genetic associations with lipid subfractions were sourced from a meta-analysis of GWAS summary statistics of metabolic measures by the UCLEB Consortium and Kettunen *et al.,*[13] utilizing Nuclear magnetic resonance (NMR) spectroscopy.

3.1.3. Drug, target and indication data

ChEMBL[14] is a manually curated database that compiles data about drugs or drug-like small molecules, their targets and associated indications, and provides detailed information about their molecular structure, mechanism of action and bioactivity profile. Compound, target and drug indication data (where relevant) were extracted from ChEMBL version. 25 (v25)[14]. ChEMBL includes compounds under both preclinical (phase 0) and clinical development (phases 1-3), and licenced (phase 4). Information in ChEMBL is itself based on several resources including United States Adopted Name (USAN) applications, ClinicalTrials.gov; the FDA Orange Book database, the British National Formulary, and the ATC classification for compounds with a license. Additional information on intended indications is sourced from DailyMed and the ATC classification.

Since proteins are the major category of drug targets (the main focus of this thesis), drug targets were mapped to their corresponding UniProt identifiers, and thence to gene identifiers in Ensembl version 95 (GRCh37) (see section 3.2). Compounds flagged as withdrawn (n=239)

or non-human targets (n=262) were excluded from all the analyses performed throughout this thesis.

From the drug repurposing perspective, the development and improvement of databases that integrate data from clinical trials is crucial. Not only do successful trials provide valuable information, but also studies that fail due to safety reasons or inadequate efficacy can be relevant for clinical practice, drug discovery or repositioning. The clinicaltrials.gov database compiles information from interventional studies and displays a summary of the study, including the number of participants, outcomes measured and adverse effects. This database was used in Chapter 7 to examine if known lipid-related trial outcomes and adverse events were identified via biomarker-weighted drug target MR for drugs and clinical candidates.

3.2. The druggable genome

The set of genes encoding proteins that are already drugged or have a greater probability of being amenable to targeting with a pharmaceutical druggable genome is known as the druggable genome. It was first described in 2002 by Hopkins and Groom[15] and updated by Finan *et al.*, in 2017[16]. At the time of this thesis, the definition comprises 4,729 human genes and encompasses potential targets for monoclonal antibodies.

In this thesis, the druggable genome was used to generate a sample space bounded by all druggable genes and all human diseases, diseases in clinical and preclinical development (Chapter 4). In addition, it was used to map drug targets to the encoding gene in Ensembl version 95 (GRCh37), which facilitated the extraction of genomic coordinates for the investigation of the support of genetic evidence from genome-wide association studies for approved drug targets (Chapter 5) and drug target MR analyses (Chapter 6 and 7).

3.3. Standardisation of GWAS and indication data

Ontologies are increasingly used in research and clinical settings. Essentially, they are repositories of standardised vocabulary that provide standard terms and identifiers for conditions and relations to enable data integration across multiple systems[17]. Several databases incorporate terms from a variety of ontologies to index the different phenotypes, diseases, molecules and pathways associated to an entry. The Medical Subject Headings (MeSH)[18] and the Unified Medical Language System (UMLS)[8] are biomedical ontologies and organise the knowledge in hierarchies with the purpose of generating a standard terminology for use in healthcare systems and research. The Experimental Factor Ontology (EFO) provides a systematic description of diseases, chemical compounds and other experimental variables available in EBI databases, and it is currently being used to unify the phenotypes of association studies collected in the GWAS Catalog[19].

In this thesis, the UMLS version 2020AA, which contained approximately 4.28 million concepts (CUIs) and 15.5 million unique concept names (AUIs), was used as the anchoring coding system for existing diseases, traits studied by GWAS and drug indications. This system was selected because it integrates several medical vocabularies and enables interoperability between data sources by facilitating the link between terms from different ontologies, which are not consistently used across GWAS and drug databases. The phenotypes available in the GWAS Catalog were mapped to UMLS terms through a combination of several approaches including manual curation (details provided in Chapter 4). The set of traits sourced from the UK Biobank were provided using International Classification of Diseases 10th revision (ICD-10) codes, which allowed for a direct mapping to the UMLS as the ICD-10 system is one of the multiple vocabularies included. Similarly, MeSH terms are provided for drug indications in ChEMBL v25, which were latter mapped to UMLS terms.

3.4. Statistical analysis methods

3.4.1. Mendelian Randomisation analyses

Drug target Mendelian Randomisation (MR) analyses were performed using different strategies for drug target gene and instrument selection based on the specific research question. Specific MR methods are described in detail in the succeeding results chapters in the respective methods sections.

As an overview, in Chapter 6 and Chapter 7, the Rücker model-selection framework was used to decide between competing inverse-variance weighted (IVW) fixed-effects, IVW random-effects, MR-Egger fixed effects or MR-Egger random-effects models[20]. While IVW models assume an absence of directional horizontal pleiotropy, Egger models allow for possible directional pleiotropy at the cost of power. The Rücker model-selection framework was chosen as it provides a systematic, fast and data-driven approach to choose between competing MR models. Details and differences between models were described in Chapter 2.2. In addition, genetic variants with large heterogeneity or leverage were removed to avoid outliers to influence the regression model and result in misleading causal estimates. See Chapter 2.3 for approaches to detecting and accounting for heterogeneity in Mendelian randomisation.

In addition, all the Mendelian randomisation analyses performed accounted for residual correlation between variants by using a linkage disequilibrium (LD) reference dataset derived from UK Biobank. LD reference matrices were created by extracting a random subset of 5,000 unrelated individuals of European ancestry from UK Biobank. Details on the quality control steps performed can be found in the succeeding results chapters.

Additionally, a drug target multivariable MR analysis was conducted in Chapter 7 to account for potential pleiotropic effects of target perturbation via other pathways. Further details on the MVMR are provided in the section 2.2.4.4 and 7.3.4.

3.4.2. Other statistical analyses

Chapter 5 estimates a series of probabilities related to the added value of genetic support in the probability of success or failure of a drug target-indication pair in drug development. Information on the proportion of successful and unsuccessful drug target gene -indication pairs and the proportion of drug development programmes with and without genetic support was sourced and 2x2 tables generated for each phase of development progression and overall. Details can be found in the succeeding results chapters.

In addition, to assess the possibility of false positive results during the biomarker-weighted drug target MR analyses (Chapter 7), the empirical *p* value distribution of the MR findings was compared against the continuous uniform distribution using the Kolmogorov-Smirnov goodness-of-fit test (two-sided). Under the null hypothesis of no association, *p* values follow a continuous uniform distribution between 0 and 1[21].

All the analysis were performed in Python 3.7.6 and 3.7.7, R Studio 3.6.1., locally or in High Performance Computing (HPC) environments (e.g., Myriad, CS Cluster and eMedlab[22,23]). Visualisations were generated using Python 3.7.7. The code used is available in GitLab (https://gitlab.com/mgordi).

3.5. References

1. Schriml, L. M. *et al.* The Human Disease Ontology 2022 update. *Nucleic Acids Research* **50**, D1255–D1261 (2022).
2. Medical Subject Headings - Home Page. https://www.nlm.nih.gov/mesh/meshhome.html.
3. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Research* **49**, D1207–D1217 (2021).
4. Clinical Classifications Software (CCS) for ICD-10-PCS (beta version). https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp.
5. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102–1111 (2013).
6. SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html.
7. Hingorani, A. D. *et al.* Improving the odds of drug development success through human genomics: modelling study. *Sci Rep* **9**, 18911 (2019).
8. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-270 (2004).
9. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
10. UK Biobank. *Neale lab* http://www.nealelab.is/uk-biobank.
11. Shah, T. *et al.* Population genomics of cardiometabolic traits: design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PloS one* **8**, e71345 (2013).

12. Benyamin, B. *et al.* GWAS of butyrylcholinesterase activity identifies four novel loci, independent effects within BCHE and secondary associations with metabolic risk factors. *Human Molecular Genetics* **20**, 4504–4514 (2011).
13. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122 (2016).
14. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940 (2018).
15. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery* **1**, 727–730 (2002).
16. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, (2017).
17. Hoehndorf, R., Schofield, P. N. & Gkoutos, G. V. The role of ontologies in biological and biomedical research: a functional perspective. *Brief Bioinform* **16**, 1069–1080 (2015).
18. Rogers, F. B. Medical subject headings. *Bull Med Libr Assoc* **51**, 114–116 (1963).
19. Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).
20. Bowden, J. *et al.* Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *International Journal of Epidemiology* **47**, 1264–1278 (2018).
21. Storey, J. D. A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 479–498 (2002).
22. UCL. Research Computing Platforms. *Research IT Services* https://www.ucl.ac.uk/research-it-services/services/research-computing-platforms (2018).

23. UCL. emedlab. *Research IT Services* https://www.ucl.ac.uk/research-it-services/emedlab (2018).

4 | Disease coverage, overlap and divergence of human genetic studies and pharmaceutical research and development

4.1. Abstract

Human genomics may help increase the efficiency of drug development by generating evidence for drug target identification and validation. However, the extent to which the spectrum of human diseases has been addressed by genetic analyses, or by drug development, and the degree to which these efforts overlap remains unclear. In this chapter different data sources are harmonised and integrated to create a sample space of all the human drug targets and diseases and identify points of convergence or divergence of genomics and drug development efforts. Approximately 9% (953 out of 10,901) of human diseases have been studied by genome-wide association studies (GWAS). Of these, only 369 correspond to diseases with an approved treatment and/or a treatment under clinical or preclinical development, leaving 584 diseases that have been the subject of investigation in GWAS, but which have yet to be investigated in drug development. This chapter illustrates how different regions of the drug target-disease space can be used to identify opportunities for genetic studies, either to help prioritise conditions with unmet clinical need, to expand the indications for licensed drugs or to identify repurposing opportunities for clinical candidates that failed in their originally intended indication.

4.2. Introduction

Pre-clinical, cell and animal model-based approaches for drug target identification and validation have been poorly predictive of human efficacy, contributing to the high failure rate in clinical phase drug development[1–3] due to lack of therapeutic efficacy or unanticipated mechanism-based adverse effects[4,5].

Human genomics may help improve drug development efficiency by helping to map drug targets to diseases more accurately and systematically through genome-wide association studies (GWAS) (target identification); and by using DNA sequence variants in a gene encoding a drug target, that influence its expression or function, to anticipate the full range of beneficial and harmful mechanism-based effects of a drug acting on the encoded protein (target validation), using drug target Mendelian randomisation[6–10]. Several lines of empirical evidence support this concept: (1) Many GWAS have rediscovered established drug targets for the corresponding diseases[11–13]; (2) Target-disease pairings with genetic support are enriched among successful drug development programmes[14–16]; (3) Comparative studies have shown that the effect of licensed drugs on biomarkers and disease endpoints coincide with the observed associations of variants in the genes encoding the corresponding target[17–19]; and (4) Several drugs have now been successfully developed or repurposed on the basis of human genetic evidence (e.g., maraviroc for treatment of HIV infection[20,21]; PCSK9 inhibitors for hypercholesterolaemia and coronary disease prevention[18,22] and tocilizumab for treatment of SARS-CoV-2 infection[23,24]).

For this reason, the pharmaceutical industry has shown growing interest in the use of human genomic data to help prioritise drug development programmes and reduce the risk of clinical-stage failure. For example, joint-pharma partnerships have provided substantial investment for sequencing, genotyping or molecular phenotyping of large national biobanks

which are connected to routinely collected primary and secondary care health records (e.g., in the UK[25] and Finland[26]). Some have engaged in partnerships with healthcare providers (e.g., Regeneron with Geisinger Healthcare in the US). Others with consumer genetic testing companies (e.g., GSK with 23andMe[27]). Several pharmaceutical companies have also invested in Open Targets, a partnership with the European Bioinformatics Institute and the Welcome Trust Sanger Institute that seeks to harness summary level genetic association data from GWAS to inform therapeutic hypotheses[13].

However, until recently, genetic studies of human diseases and pharmaceutical research and development have largely proceeded independently. Thus, the extent to which the causes of human disease have been addressed by genetic analyses, or by drug development, and the degree to which these efforts overlap, has not been investigated systematically. Filling this gap in knowledge will have several applications. First, a survey of this type would help understand where future drug development programmes could be directed if they are seeking to exploit existing genetic evidence. Conversely, such an effort could help prioritise new, large-scale GWAS or sequencing studies to help stimulate drug development for diseases currently without effective treatments. Third, it could help quantify opportunities to expand the indications for licensed drugs or identify repurposing opportunities for the many safe drugs that failed in clinical trials because of lack of efficacy in the originally intended indication. To address this gap, disparate sources of data were connected to evaluate disease coverage and overlap of genomic and pharmaceutical research and development.

4.3. Methods

4.3.1. Human diseases

To estimate the total number of human diseases, information from the following disease classification systems and ontologies was retrieved on the 30th November 2021: ICD-10, ICD-11, Human Disease Ontology (DO)[28], Medical Subject Headings (MeSH)[29,30], Human Phenotype Ontology[31], Clinical Classification Software[32], PheWAS Catalog[33], SNOMED CT[34]. The websites from where these data were sourced are specified in Table 4.1. MeSH terms falling within the categories C (diseases) or F (Psychiatry and Psychology) were selected. As of 30 November 2021, the DO had 10,901 disease terms. Since the number of terms in the DO is updated regularly, the rationale described in previous studies[35] was followed and a figure of 10,901 was proposed as a reasonable estimate of the number of common human diseases with genetic susceptibility. Diseases with an approved treatment and/or a treatment under clinical or preclinical development were sourced from ChEMBL version 25 (v25)[36], which provided standardised indication terms based on MeSH. To facilitate further mappings and estimate the coverage, overlap and divergence of human genetic studies and diseases investigated in pharmaceutical research and development, disease terms from DO and ChEMBL v25 were mapped to Unified Medical Language System (UMLS)[37] concepts using the UMLS2020AA. The UMLS was selected as the anchoring coding system as it integrates several medical vocabularies to enable interoperability between data sources and facilitate the link between terms from different ontologies.

Table 4.1. The number of disease terms within widely used classification systems and ontologies as of 30 November 2021.

Coding Scheme	Type	Number of terms	Data source
ICD-10	Disease classification	8,196	https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html
ICD-11	Disease classification	12,096	https://icd.who.int/dev11/downloads/
Human Disease Ontology	Ontology	10,901	https://github.com/DiseaseOntology/HumanDiseaseOntology/blob/main/RELEASES.md#2021-releases
Medical Subject Headings	Ontology	5,785	https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html
Human Phenotype Ontology	Ontology	14,547	https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html
Clinical Classification Software	Disease groups	259	http://www.ahrq.gov/research/data/hcup/icd10usrgd.html
PheWAS Catalog	Disease groups	1,670	https://phewascatalog.org/
SNOMED CT	Clinical terminology	349,385	https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

4.3.2. Drug and target data

Compound, target and drug indication data were extracted from ChEMBL version 25 (v25)[36]. ChEMBL includes compounds under both preclinical and clinical development. Information in ChEMBL is itself based on several resources including United States Adopted Name (USAN) applications, ClinicalTrials.gov; the FDA Orange Book database, the British National Formulary, and the ATC classification for compounds with a license. Additional information on intended indications was sourced from DailyMed and the ATC classification. Since proteins are the major category of drug targets, drug targets were mapped to the corresponding UniProt identifiers, and thence to gene identifiers in Ensembl version 95

(GRCh37) through the updated druggable genome[11]. Compounds flagged as withdrawn (n=239) or directed to non-human targets (n=262) were excluded from the analysis.

4.3.3. GWAS data

The collection of traits studied by GWAS were obtained from a public central repository (GWAS Catalog v1.0.2[38]) and from UK Biobank through Neale data (GWAS Round 2, Results shared 1st August 2018[39]). These included 2,452 unique traits and 633 clinical diagnoses, respectively. To filter human diseases from the 2,452 traits in the GWAS Catalog, terms were mapped to UMLS concepts using several complementary approaches. One thousand eight traits were mapped to 1,364 UMLS concepts using MetaMap[40], 225 traits were mapped to 227 UMLS concepts using direct string matching, 14 traits were mapped to 16 UMLS concepts using the UMLS and 35 traits were mapped to 75 UMLS concepts using cross-mapping between ontologies in DisGeNET[41], and 1,099 traits were manually mapped to 967 terms using the UMLS Methasaurus. The 633 ICD-10 diagnosis in Neale data were automatically mapped to UMLS concepts using the UMLS2020AA. In total, 983 unique diseases were identified and manually curated. The diseases were mapped to disease areas according to ICD10 chapters. Diseases classified in the chapters: ‘Animal diseases’, ‘Findings, not elsewhere classified’ and ‘Pregnancy, childbirth and the puerperium’ were excluded, resulting in a total of 953 unique disease terms.

4.4. Results

4.4.1. Protein-coding genes and genes encoding drug targets

To generate a sample space bounded by all human protein-coding genes and all human diseases, estimates of the total number of protein-coding genes were obtained. From this, the subset of protein coding genes considered to be most amenable to targeting by drugs, a subset of the protein-coding genome known as the 'druggable genome'[11], was identified. At the time of analysis, the total number of protein-coding genes in the human genome was estimated in 19,955[42]; of which 4,729 were estimated to be amenable to targeting by small molecule drugs or bio-therapeutics. Of all human genes encoding druggable targets, 672 (14.2%) are already the gene targets of approved drugs, 1,113 (23.5%) are the targets of drugs in clinical development, 278 (5.8%) are gene targets of drugs in preclinical development and 3,604 (76.2%) are currently 'undrugged' (Fig. 4.1). Data on drugs in preclinical development may be incomplete as information on many withdrawn targets is not publicly available.

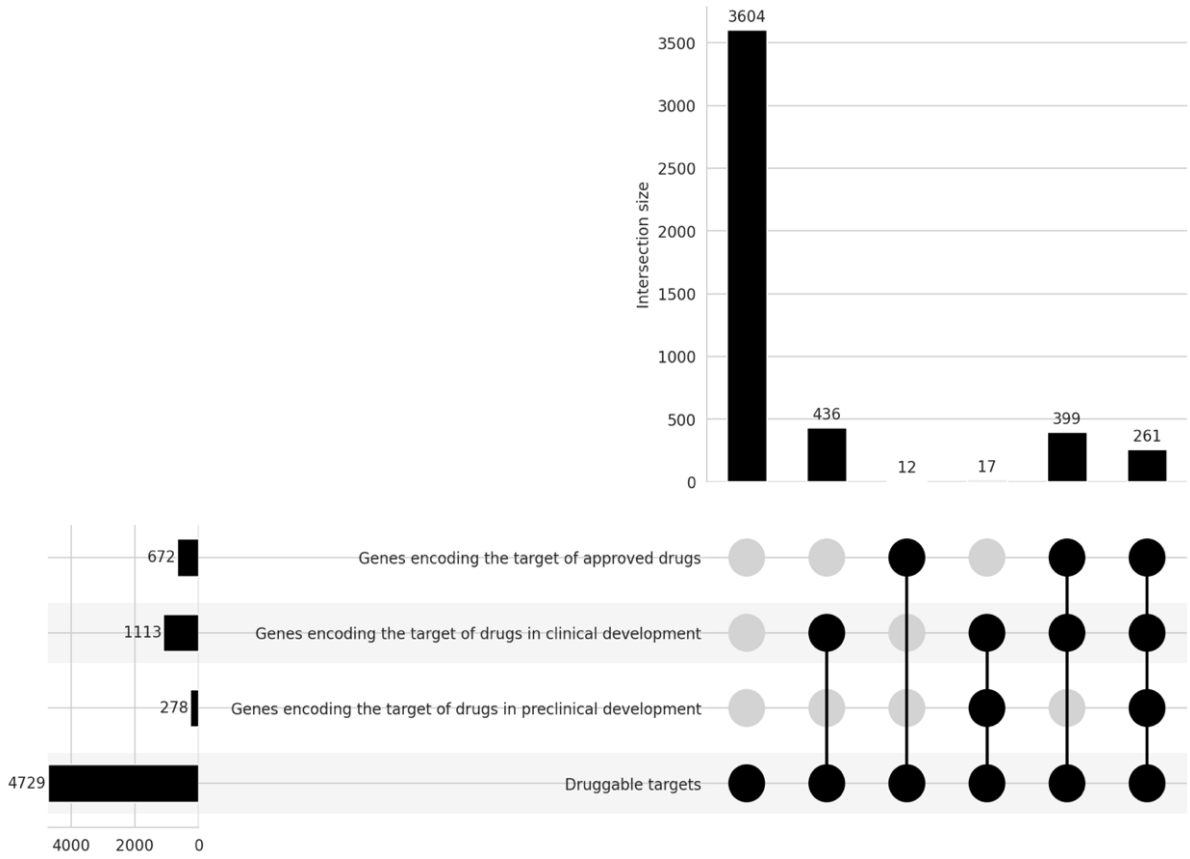


Figure 4.1. Total count of genes encoding druggable targets, with subsets and overlaps of genes encoding the targets of approved drugs, and drugs in clinical or preclinical development.

4.4.2. Human diseases evaluated in drug development and in GWAS

Producing a stable, exact figure for the total number of human diseases (the 'disease-ome') is challenging due to the hierarchical nature of biomedical vocabularies, duplications and descriptive terms beyond diagnoses present in clinical terminologies and disease classification systems. In 2019, a figure of 10,000 was proposed as a reasonable estimate of the number of common human diseases with genetic susceptibility[35]. In this analysis, an updated figure of 10,901 diseases was used which corresponded to number of terms in the DO as of 30 November 2021. The DO was selected as it is updated regularly and would provide the most up-to-date figure. Separate estimates could be derived for monogenic diseases (~7,000[43]), for

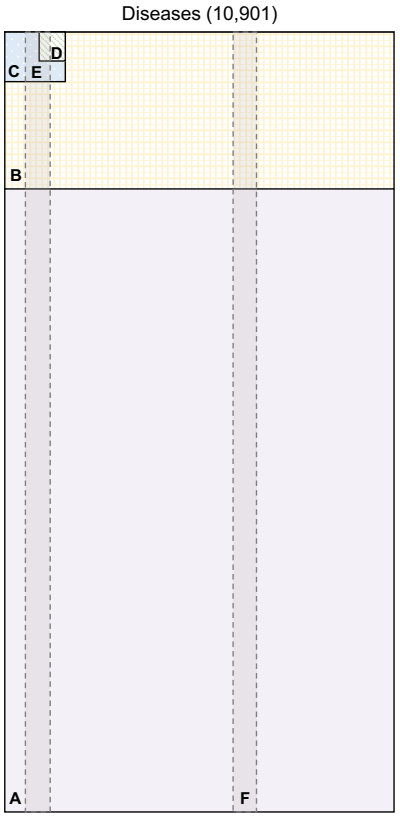
which loss-of-function variants have correctly predicted the safety and phenotypic effect of pharmacological inhibition[44]. However, the analysis of predicted loss-of-function variants requires very large sample sizes due to their low frequency in the population, and thus, a figure of the common polygenic human diseases (which are the ones subjected to GWAS) was used.

By sourcing data from the open-access drug database ChEMBL v25[36], it was found that only 1,370 diseases (12.6% of the total number of human diseases listed in DO) have an approved treatment and/or a treatment under clinical or preclinical development. This comprises 463 diseases that are the indication of approved drugs, 1,242 diseases that are or have been the indication of drugs in clinical development and 217 diseases that are or have been indications for drugs in preclinical development.

Equally, estimating the proportion of diseases covered by genome-wide association studies is difficult because some diseases could have been studied through a validated clinical biomarker (e.g., LDL cholesterol for coronary heart disease) as well as directly with the disease endpoint. There may also be inconsistencies in annotation of clinical end points to a coding system (e.g., non-small cell lung cancer and non-small cell lung carcinoma have different codes in the unified medical language system, UMLS). Nevertheless, with these caveats, 953 diseases covered by GWAS (8.7% of the total number of common human diseases) were identified based on the mapping and manual curation of phenotype terms in the GWAS Catalog[38] and UK Biobank through Neale data[39] to UMLS concepts. However, it was found that only 369 of the 1,370 diseases with an approved treatment and/or a treatment under clinical or preclinical development had also been investigated by GWAS (Fig. 4.2a and Fig. 4.3) leaving 584 diseases that have been the subject of investigation in GWAS, but which have yet to be investigated in drug development. However, this intersection of GWAS and drug development efforts varied by disease area (Fig. 4.4). For example, 48 (34.0%; confidence interval: 26.2 - 41.9%) out of

the 141 diseases of the circulatory system with an approved treatment and/or with a treatment under clinical or preclinical development had been studied in a GWAS, while for endocrine, nutritional or metabolic diseases this figure was 14.4% (27 out of 188; confidence interval: 9.3 - 19.4%).

a



A - Full sample space of target-indication pairs

B - Druggable genome

C - Targets-disease pairs under clinical investigation

D - Approved target-disease pair

E - Targets-disease pairs studied by GWAS where the disease has been studied by drug development

F - Targets-disease pairs studied by GWAS where the disease has not been studied by drug development

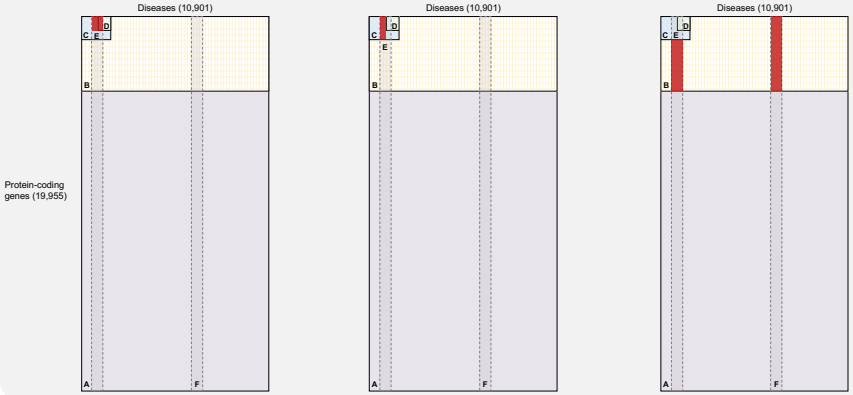
b

EXPLOITING EXISTING OPPORTUNITIES

Indication expansion
Intersection between targets of approved drugs and diseases studied by GWAS

Repurposing
Intersection between targets of drugs under clinical investigation and diseases studied by GWAS

***De novo* drug development**
Intersection between druggable targets other than those for approved and investigational drugs and diseases studied by GWAS



c

CREATING NEW OPPORTUNITIES

New disease indications with genetic support

- Expanding the scope of GWAS
- GWAS in routine healthcare systems (e.g., EHR datasets)

Increasing the universe of therapeutic targets

- Protein targets
- mRNA targets (liver/tissue constraints)
- Gene editing (liver/tissue constraints)

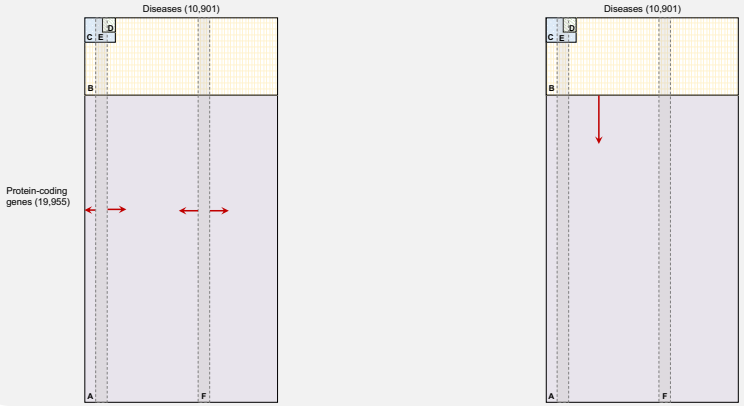


Figure 4.2. Illustration of the sample space and subsets of human proteins and diseases. The complete sample set (A) is bounded by the total number of protein coding genes and the sum total of common, complex human diseases. The subset of all potentially druggable target-disease indication pairings is indicated by subset B, the drug target-disease indication pairings studied in clinical phase drug development by subset C, and the target-disease indication pairings of approved drugs by subset D. The vertical lines represent diseases studied by GWAS on the assumption that GWAS interrogate all genes in the human genome (subset E and F). The presence of two GWAS subsets is to illustrate the point that only a subset of diseases studied in GWAS have also been the subject of drug development (E). See text for further explanation.

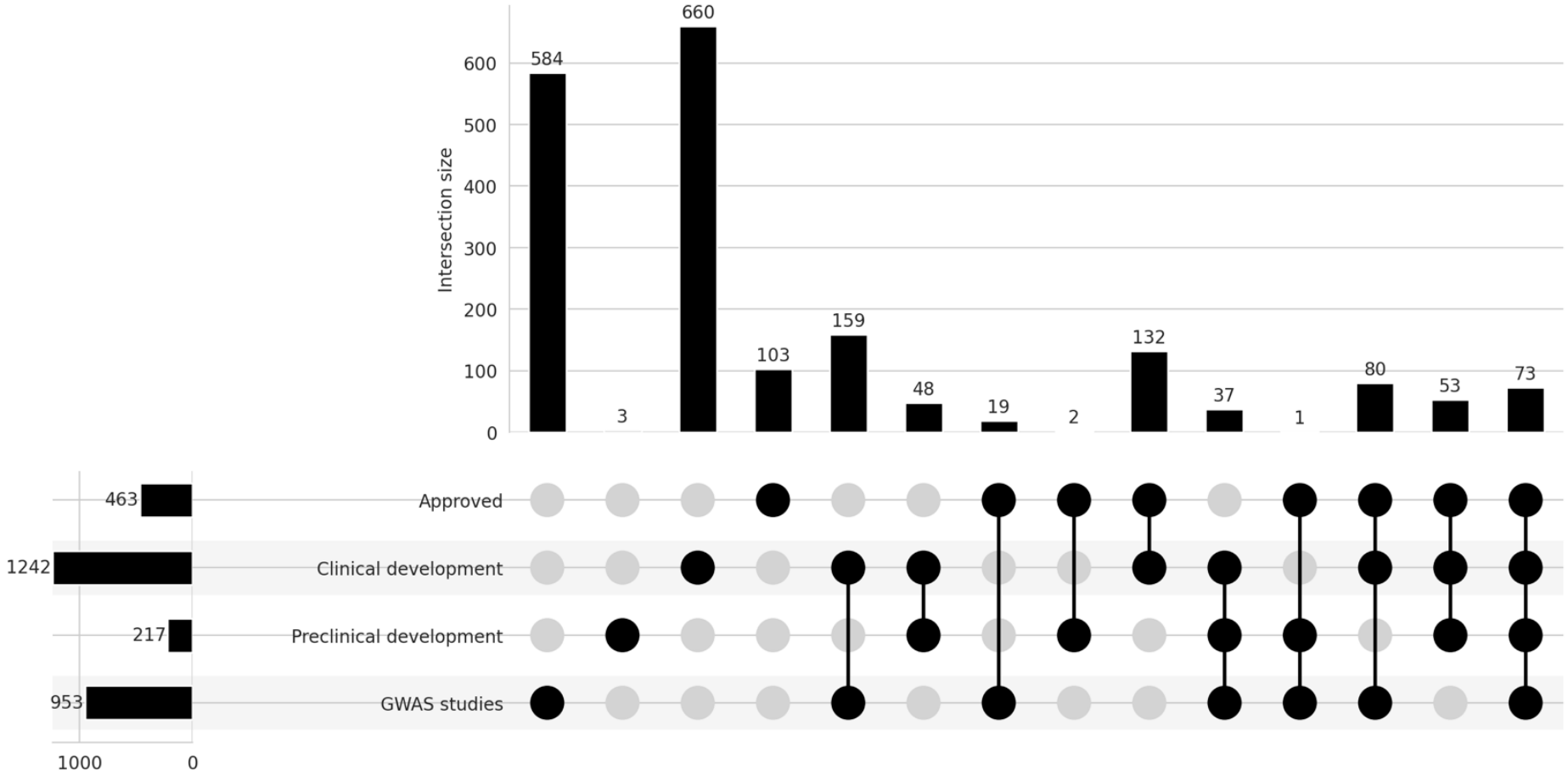


Figure 4.3. Intersection between diseases with a current approved treatment, with a treatment under clinical development, with a treatment under preclinical development, or investigated by GWAS. Data sources: ChEMBL v25 (approved, clinical and preclinical development), GWAS Catalog (GWAS studies) and UK Biobank through Neale data (GWAS studies).

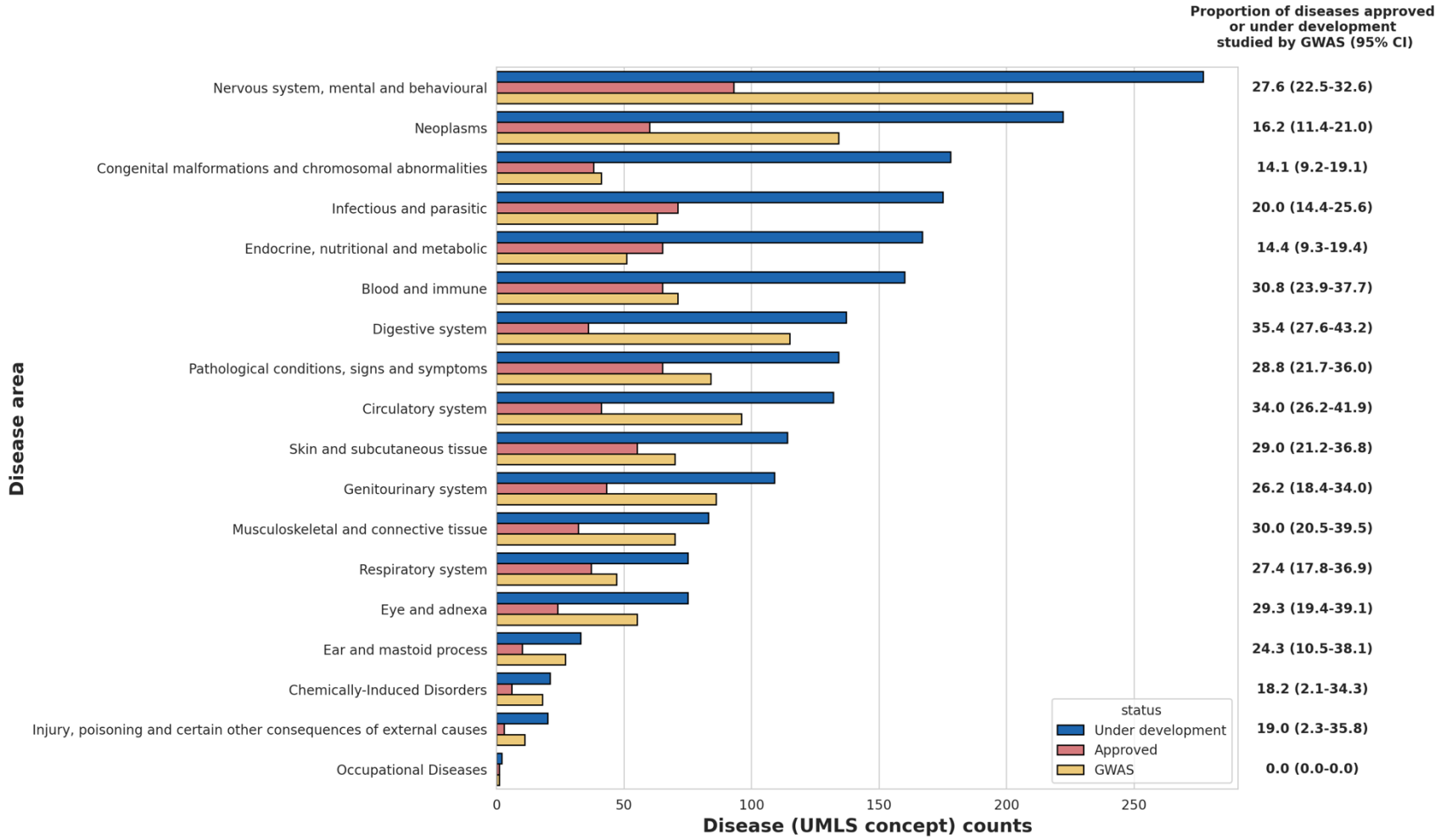


Figure 4.4. Diseases with an approved treatment, a treatment under investigation and studied by GWAS by disease area (ICD-10 chapter). Total numbers: 463 diseases that are the indication for an approved drug, 1,248 diseases with a drug investigated in clinical or preclinical studies and 953 diseases studied by GWAS.

4.4.3. Important subcategories of drug target-disease indication pairings

Based on the previous mappings, sample spaces based on different sub-categories of drug target-disease indication pairings were generated to help inform future genomic and drug development efforts.

Sample space bounded by all protein coding genes and diseases

As a denominator, a sample space bounded by 19,955 protein coding genes and 10,901 diseases was generated, which produces ~217 million protein-disease indication pairings (217,529,455; labelled A in Fig. 4.2a).

Sample space bounded by the druggable genome and all human diseases

Since not all proteins are readily targeted by small molecule drugs or monoclonal antibody or peptide therapeutics, the sample space more relevant to drug development is bounded by 4,729 genes encoding druggable targets[11] and the 10,901 human diseases, which produces ~52 million (51,550,829) drug target-disease indication pairings that might be the subject of drug development. This space is labelled B in Figure 4.2a.

Sample space bounded by target-indication pairings under clinical investigation

Having defined these key denominator values, the number of drug target-disease indication pairs that are or have been the subject of clinical phase drug development was investigated. This space, labelled 'C' in Figure 4.2a, is bounded by 1,113 genes encoding the targets of drugs (Fig. 4.1) and 1,242 diseases that have been the investigated in clinical phase drug development (Fig. 4.3), giving around 1.4 million (1,382,346) target-indication parings. It should be noted that although this sample space encompasses ~1.4 million drug target-disease indication pairings, it represents only about 2.5% of the ~52 million drug target-

indication pairings that could be studied (sample space B), and 0.6% of the ~217 million protein-disease pairings (sample space A). Moreover, of the ~1.3 million the number of drug target-disease indication pairings, only 29,326 (2.1%) have actually been explored. Further, coverage of targets and disease areas is uneven with some disease and targets being intensively investigated and others less so or not at all (Fig. 4.5).

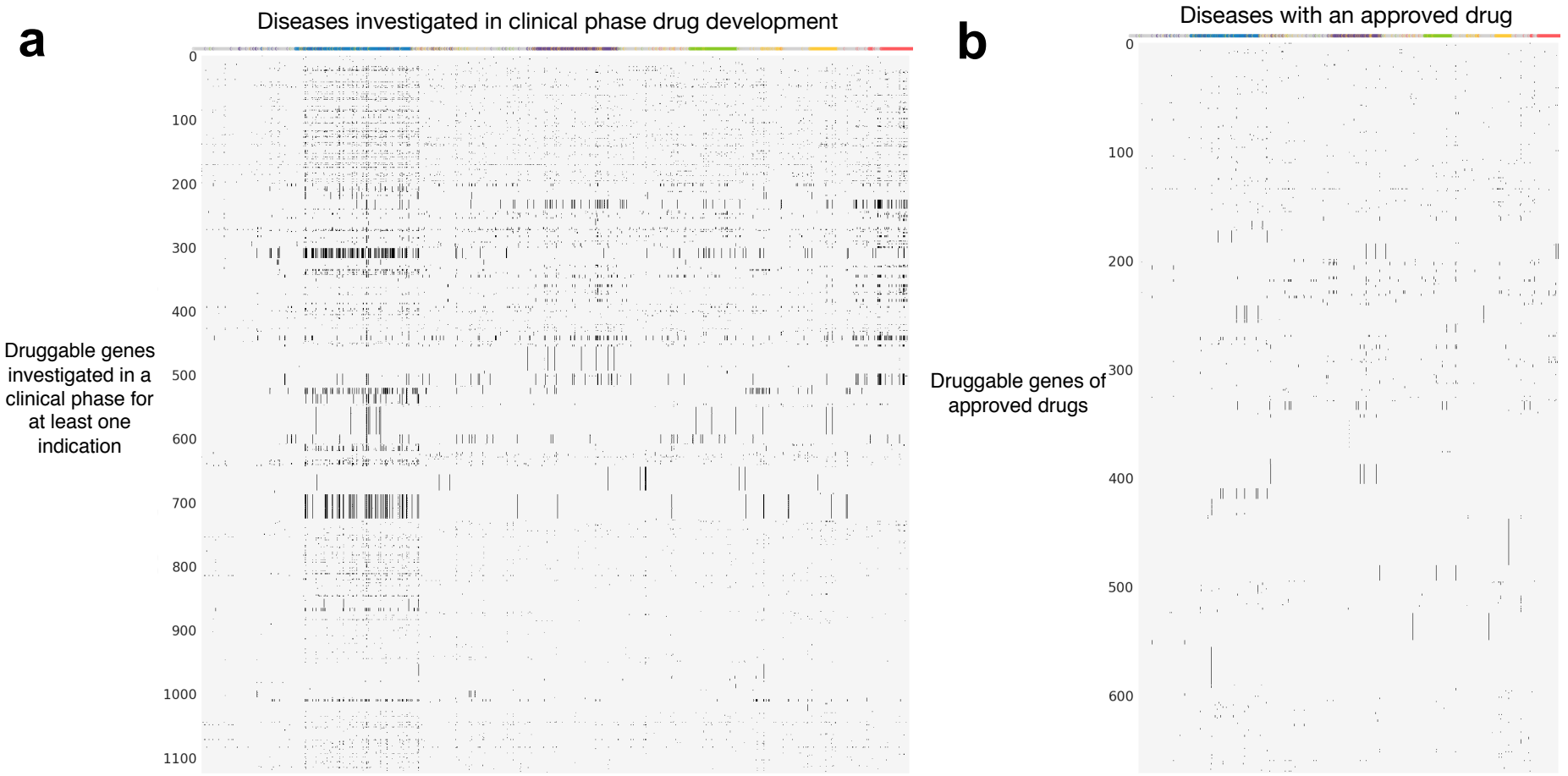


Figure 4.5. Sample space bounded by target indication pairs under clinical investigation (a) or by target-indication pairings for approved drugs (b). (a) The Y-axis includes the 1,125 druggable genes investigated in a clinical phase for at least one indication and the X-axis the 1,242 diseases that have been the tackled in clinical phase drug development. (b) The Y-axis includes the 672 druggable genes of approved drugs and the X-axis the 463 disease indications. The colours in the X-axis indicate five major group of diseases: neoplasms (blue), nervous system diseases (purple), cardiovascular diseases (green), endocrine, nutritional and metabolic (orange), psychiatry and psychology disorders (pink), and others (grey).

Sample space bounded by target-indication pairings for approved drugs

I identified 672 targets of approved drugs (Fig. 4.1) for 463 disease indications (Fig. 4.3), giving a sample space (labelled D in Fig. 4.2a) of just under 312,000 target indication pairs (312,261). Again, the number drug target-disease indication hypotheses that have actually been explored and led to approval within this bounded space is ~ 1% (n=3,154) of the maximum space available at the time of analysis. As for target-indication pairings investigated in clinical development, the coverage of targets and indications of approved drugs is uneven. Some diseases (e.g., hypertension) have a large number of targets for approved drugs (e.g., there are 24 approved drug targets for the treatment of hypertension), whereas others (e.g., Iridocyclitis) have treatments directed at a single target (Fig. 4.6). The median number of drug targets per approved indication is two. Similarly, several drug targets have been approved for multiple indications, including different disease areas. For example the glucocorticoid receptor is employed for the treatment of up to 87 diseases, including disorders of the blood, immune, circulatory, respiratory systems and different cancers (Fig. 4.7). Others have only been licensed for a single disease (e.g., Fig. 4.7).

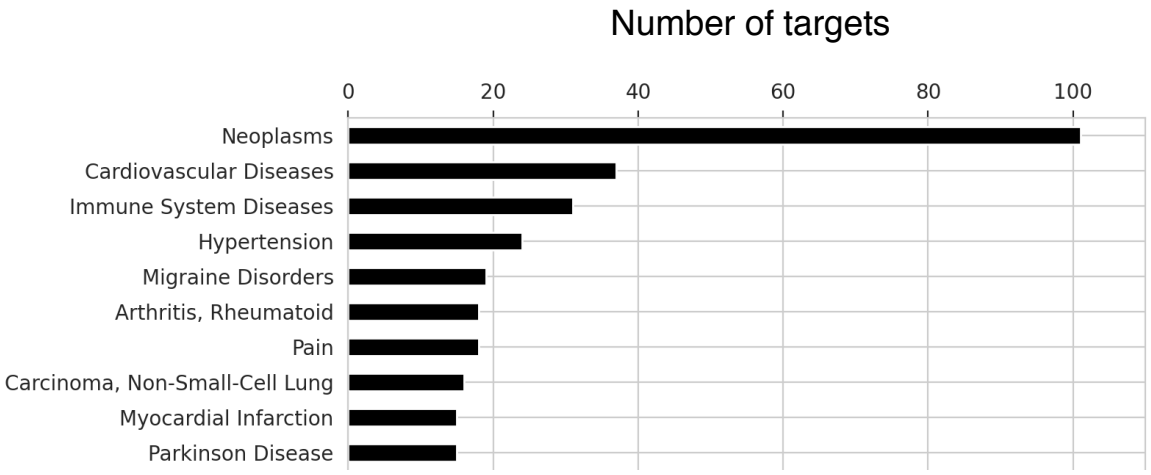


Figure 4.6. Number of drug targets by disease (top 10 diseases).

Figure 4.7. Disease indications by drug target. The dashed line separates the drug targets with the most approved indications from a random subset of ten drug targets with a single approved indication.

Diseases and targets evaluated in GWAS and drug development

I identified 672 targets of currently approved drugs (14% of all druggable targets) employed in the treatment of 463 diseases (4% of all 10,901 diseases). Of these diseases, 173 have also been studied in GWAS. It is through this intersection that it has been possible to show that GWAS have frequently rediscovered established drug targets for the corresponding diseases[11–13]. The 1,125 targets of drugs that are or have been the subject of clinical investigation (which includes the targets of approved drugs), have been or are being evaluated for the treatment of 1,370 diseases. Prior research has shown that drugs which the target-indication pairing has genetic support have higher rates of approval. However, of the 1,242 disease indications being evaluated in clinical development, only 349 has been the subject of a GWAS. High failure rates in clinical phase drug development have heightened interest in therapeutic repurposing of drugs that failed in their originally intended indication for lack of efficacy. Previous modelling studies have suggested that any given drug target might be useful in the treatment of multiple diseases[35]. There are well-established examples of this. Beta-adrenoceptor antagonists are used in the treatment of hypertension, coronary heart disease, heart failure, portal hypertension and migraine. SGLT2-inhibitors developed for diabetes also reduce risk of heart failure with preserved ejection fraction, coronary and renal disease and can also treat obesity. Since GWAS can be used as a source of evidence for drug target identification, one route to expanding the indications of licensed drugs or those in development, or to repurpose investigational drugs that fail in their intended indication, would be to systematically interrogate the association of variants in the genes encoding the targets of these drugs in GWAS data. Since GWAS have already investigated 953 diseases there is already a large dataset that could be utilised for this purpose. For example, the interleukin-6 receptor is the target of an approved drug (tocilizumab) used in the treatment of rheumatoid arthritis. However, the gene encoding this receptor has also been identified using GWAS of coronary

heart disease, abdominal aortic aneurysm and atrial fibrillation, suggesting a number of indication expansion opportunities[19,45,46]. Another example is the interleukin-23 receptor inhibitor ustekinumab, which was originally intended to treat psoriasis, and after identifying a GWAS signal for Crohn's disease was investigated for such indication and eventually approved in 2017[47–49].

Creating new opportunities for genetic drug target validation

The development of a sample space of druggable targets and disease indications illustrates how new opportunities for genetic drug target validation can be exploited.

One way would be by increasing the range of druggable targets (space B in Fig 2a). This is becoming possible through technological developments. These include: 1) the growing use of monoclonal antibodies and the development of cyclic peptides as therapeutics for protein targets that lack a binding pocket amenable to targeting by conventional small molecule therapeutics[47–51], 2) the targeting of RNAs rather than proteins using RNA silencing approaches and the emergence of CRISPR-Case 9 based gene editing in cases for proteins that remain difficult to drug[52–54].

A complementary approach, necessary to map the expanded range of druggable targets to the correct diseases is to increase the range of diseases that have been studies in GWAS. This is becoming possible by the greater deployment of genetic studies within large national biobanks linked to health care data[25,55–57], and even in healthcare systems[58,59].

4.5. Discussion

4.5.1. Summary

Previous research has shown that human genetic evidence could support drug development[11,14,15,35]. However, the extent to which the genomic efforts, specifically GWAS, align with ongoing drug development efforts and unmet need has not been explored in detail. The current analysis shows: 1) Only a small fraction of the 10,901 diseases curated in the human DO have been investigated in drug development (13%; 1,370 out of 10,901) or GWAS (9%; 953 out of 10,901); 2) of disease being pursued in clinical phase drug development, only 27% (369 out of 1,370) has been the subject of a GWAS; 3) even for the 349 diseases that are the subject of ongoing clinical phase drug development and have been covered by GWAS, it remains uncertain how many specific target-indication pairings have genetic support. The construction of a sample space of disease and targets including subsets of target-disease pairings that have been covered by GWAS (which interrogate all possible targets by design) and clinical phase drug development can help generate insights into how these efforts can be utilised in concert.

For example, the intersection between targets of approved drugs and diseases studied by GWAS can help identifying new indications for existing approved drugs. On the other hand, the intersection between targets of drugs under clinical investigation and diseases studied by GWAS can lead to potential repurposing opportunities of drugs that proved safe but lacked efficacy for the originally intended indication, or for indication expansion of approved drugs. Both indication expansion and repurposing are attractive alternatives to the *de novo* drug development, mainly because such compounds have been proven to engage well-characterised targets and the medicines have proven safe in clinical trials, which leads to a reduction of the costs and development timelines[60]. In addition, the sample space of human targets and diseases

could also inform *de novo* drug development for druggable targets and disease indication pairings that have yet to be investigated.

4.5.2. Research in context

There are groups of targets that could especially benefit from having genetic support. For example, identifying soluble or secreted protein targets with genetic evidence for a particular disease represent an attractive venture since such proteins are readily targeted by monoclonal antibodies or peptides, which typically exhibit higher selectivity and reduced development timelines compared to small molecules[61]. Information on the set of human secreted proteins (the human 'secretome'[9]) is available in the public domain, and researchers and the pharmaceutical industry could use these resources to identify high priority putative circulating protein targets. In addition to therapeutics that exert their action at the protein level, novel therapies based on RNA silencing or interference provide a solution to downregulate protein targets that are resistant to small or large molecule therapeutics[52]. While this technique is challenged by the effective delivery of the RNA into the target tissue, existing technologies support efficient targeting of the liver with RNA-based therapeutics[62]. Therefore, genetically-supported targets with an elevated gene expression in liver may be prioritised for RNA silencing therapy.

Furthermore, the sample space of human protein targets and diseases can be used to inform new drug development programmes and research (Fig. 2c). For example, only 9% of the human diseases have been investigated in a GWAS, and over 8,000 diseases exist without an approved treatment or under clinical investigation. Prioritising diseases for genomic analysis with a view to generating critical evidence for drug development is one of the numerous applications of the current analysis. Large biobanks with genetic data linked to routinely

collected primary and secondary care health records provide an opportunity to investigate targets with genetic support in conditions with unmet medical needs or to increase the power in diseases where a GWAS is available but the number of cases were not sufficient to reliably identify genetic associations. Furthermore, increasing population representativeness in genetic studies may also be important (since approximately 86% of the genetic studies have been performed in Europeans[63]) to evaluate if the findings are transferable across ancestries and to ensure fairness in the application of human genomics.

Part of this analysis was based on the *druggable genome* but this concept is an evolving entity. While it is currently defined as the set of *proteins* with potential to be modulated by a drug-like small molecule or monoclonal antibody, novel therapeutic modalities, such as RNA silencing or gene editing, hold the promise of modifying the function of any protein targeting any gene in the genome. This is likely to expand the range of potential druggable targets[64–66]. Lastly, in addition to the advances in molecular therapeutics, several companies have shown growing interest in the use of artificial intelligence for target identification and drug discovery. The application of data-driven approaches and computer modelling have solved protein structures and revealed previously unknown protein motifs, turning undruggable protein targets into druggable ones[67].

4.5.3. Strengths and limitations

The results presented in this chapter represent the first systematic survey of the coverage, overlap and divergence of human genetic studies and diseases investigated in pharmaceutical research and development. One of the strengths of this analysis is that the data used were available in the public domain which facilitates the revisiting of the estimates in the future. Another is that the analysis was stratified to show how the overlap between diseases with an

approved treatment, a treatment under clinical development and studied by GWAS differs also at the level of individual disease. Moreover, standardisation of terms across data sources was challenged by the use of different coding systems in the drug and GWAS databases and the lack of a direct mapping across terminologies. By using the UMLS as an anchoring ontology to standardise the diseases across data sources and including a step of manual curation of the disease terms and areas, the error due to inaccurate mapping cross-databases was reduced.

There are several limitations to the analysis described. First, information on drugs in preclinical or clinical development may be incomplete or not available in the public domain, which may lead to an underestimation of the number of diseases studied in drug development, particularly for the preclinical candidates which did not progress to clinical trials. Regarding the number of diseases investigated by GWAS, some diseases could have been studied through a validated clinical biomarker which may not have been captured by this approach. To minimise this error, where possible, GWAS of biomarkers in the GWAS Catalog were manually curated and linked to diseases.

4.6. Conclusion

The analysis described in this chapter shows the divergence between diseases studied by GWAS and those investigated by the pharmaceutical industry. Only 369 of the 1,370 diseases with an approved treatment and/or a treatment under clinical or preclinical development have also been investigated by GWAS. Further efforts are needed to explore the genetic predisposition of the remaining diseases, and more importantly, the genetic contribution for those >9,000 diseases without an approved or investigational drug, based on ChEMBL v25 database. Nevertheless, almost 1,000 diseases have been investigated by GWAS which provides opportunities to investigate the additional value of genetic support in drug development and evaluate the genetic evidence of drug target-indication pairings using genetic epidemiology methodologies such as Mendelian Randomisation. These two applications of GWAS will be described in the following chapters.

4.7. References

1. Macleod, M. R. *et al.* Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. *PLOS Biology* **13**, e1002273 (2015).
2. Perel, P. *et al.* Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ* **334**, 197 (2007).
3. Worp, H. B. van der *et al.* Can Animal Models of Disease Reliably Inform Human Studies? *PLOS Medicine* **7**, e1000245 (2010).
4. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat Biotechnol* **32**, 40–51 (2014).
5. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov* **11**, 191–200 (2012).
6. Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nature Communications* **11**, 3255 (2020).
7. Gill, D. *et al.* Mendelian randomization for studying the effects of perturbing drug targets. *Wellcome Open Res* **6**, 16 (2021).
8. Walker, V. M., Davey Smith, G., Davies, N. M. & Martin, R. M. Mendelian randomization: a novel approach for the prediction of adverse drug events and drug repurposing opportunities. *International Journal of Epidemiology* **46**, 2078–2089 (2017).
9. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov* **12**, 581–594 (2013).
10. Hingorani, A. & Humphries, S. Nature's randomised trials. *The Lancet* **366**, 1906–1908 (2005).
11. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, (2017).

12. Forgetta, V. *et al.* An effector index to predict target genes at GWAS loci. *Hum Genet* (2022) doi:10.1007/s00439-022-02434-z.
13. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research* **49**, D1311–D1320 (2021).
14. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856–860 (2015).
15. King, E. A., Davis, J. W. & Degner, J. F. 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**, e1008489 (2019).
16. Ochoa, D. *et al.* Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nature Reviews Drug Discovery* **21**, 551–551 (2022).
17. Gordillo-Marañón, M. *et al.* Validation of lipid-related therapeutic targets for coronary heart disease prevention using human genetics. *Nat Commun* **12**, 6120 (2021).
18. Schmidt, A. F. *et al.* PCSK9 monoclonal antibodies for the primary and secondary prevention of cardiovascular disease. *Cochrane Database Syst Rev* **4**, CD011748 (2017).
19. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium *et al.* The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet* **379**, 1214–1224 (2012).
20. Dean, L. Maraviroc Therapy and CCR5 Genotype. in *Medical Genetics Summaries* (eds. Pratt, V. M. et al.) (National Center for Biotechnology Information (US), 2012).
21. Gu, W.-G. & Chen, X.-Q. Targeting CCR5 for anti-HIV research. *Eur J Clin Microbiol Infect Dis* **33**, 1881–1887 (2014).
22. Lopez, D. Inhibition of PCSK9 as a novel strategy for the treatment of hypercholesterolemia. *Drug News Perspect.* **21**, 323–330 (2008).

23. Salama, C. *et al.* Tocilizumab in Patients Hospitalized with Covid-19 Pneumonia. *New England Journal of Medicine* **384**, 20–30 (2021).
24. Bovijn, J., Lindgren, C. M. & Holmes, M. V. Genetic variants mimicking therapeutic inhibition of IL-6 receptor signaling and risk of COVID-19. *Lancet Rheumatol* **2**, e658–e659 (2020).
25. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779 (2015).
26. FinnGen-tutkimushanke vie suomalaiset löytöretkelle genomitietoon. *FinnGen* https://www.finngen.fi/fi/finngen_tutkimushanke_vie_suomalaiset_loytoretkelle_genomitietoon.
27. GSK and 23andMe sign agreement to leverage genetic insights for the development of novel medicines | GSK. https://www.gsk.com/en-gb/media/press-releases/gsk-and-23andme-sign-agreement-to-leverage-genetic-insights-for-the-development-of-novel-medicines/.
28. Schriml, L. M. *et al.* The Human Disease Ontology 2022 update. *Nucleic Acids Research* **50**, D1255–D1261 (2022).
29. Rogers, F. B. Medical subject headings. *Bull Med Libr Assoc* **51**, 114–116 (1963).
30. Medical Subject Headings - Home Page. https://www.nlm.nih.gov/mesh/meshhome.html.
31. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Research* **49**, D1207–D1217 (2021).
32. Clinical Classifications Software (CCS) for ICD-10-PCS (beta version). https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp.

33. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102–1111 (2013).
34. SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html.
35. Hingorani, A. D. *et al.* Improving the odds of drug development success through human genomics: modelling study. *Sci Rep* **9**, 18911 (2019).
36. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940 (2018).
37. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-270 (2004).
38. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
39. UK Biobank. *Neale lab* http://www.nealelab.is/uk-biobank.
40. MetaMap. https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html.
41. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **48**, D845–D855 (2020).
42. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766–D773 (2019).
43. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514–D517 (2005).
44. Minikel, E. V. *et al.* Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, 459–464 (2020).

45. Harrison, S. C. *et al.* Interleukin-6 receptor pathways in abdominal aortic aneurysm. *Eur Heart J* **34**, 3707–3716 (2013).

46. Rosa, M. *et al.* A Mendelian randomization study of IL6 signaling in cardiovascular diseases, immune-related disorders and longevity. *npj Genom. Med.* **4**, 1–10 (2019).

47. Simon, E. G., Ghosh, S., Iacucci, M. & Moran, G. W. Ustekinumab for the treatment of Crohn's disease: can it find its niche? *Therap Adv Gastroenterol* **9**, 26–36 (2016).

48. Duerr, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).

49. Wang, K. *et al.* Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *Am J Hum Genet* **84**, 399–405 (2009).

50. Wang, L. *et al.* Therapeutic peptides: current applications and future directions. *Sig Transduct Target Ther* **7**, 1–27 (2022).

51. Joo, S. H. Cyclic Peptides as Therapeutic Agents and Biochemical Tools. *Biomol Ther (Seoul)* **20**, 19–26 (2012).

52. Kim, D. H. & Rossi, J. J. Strategies for silencing human disease using RNA interference. *Nat Rev Genet* **8**, 173–184 (2007).

53. Bumcrot, D., Manoharan, M., Koteliansky, V. & Sah, D. W. Y. RNAi therapeutics: a potential new class of pharmaceutical drugs. *Nat Chem Biol* **2**, 711–719 (2006).

54. Sahin, U., Karikó, K. & Türeci, Ö. mRNA-based therapeutics — developing a new class of drugs. *Nat Rev Drug Discov* **13**, 759–780 (2014).

55. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* **44**, 1137–1147 (2015).

56. Hansen, T. F. *et al.* DBDS Genomic Cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors. *BMJ Open* **9**, e028401 (2019).

57. Kurki, M. I. *et al.* FinnGen: Unique genetic insights from combining isolated population and national health register data. 2022.03.03.22271360 Preprint at https://doi.org/10.1101/2022.03.03.22271360 (2022).
58. Ayatollahi, H., Hosseini, S. F. & Hemmat, M. Integrating Genetic Data into Electronic Health Records: Medical Geneticists' Perspectives. *Healthc Inform Res* **25**, 289–296 (2019).
59. Lau-Min, K. S. *et al.* Real-world integration of genomic data into the electronic health record: the PennChart Genomics Initiative. *Genet Med* **23**, 603–605 (2021).
60. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* **18**, 41–58 (2019).
61. Shepard, H. M., Phillips, G. L., Thanos, C. D. & Feldmann, M. Developments in therapy with monoclonal antibodies and related proteins. *Clin Med (Lond)* **17**, 220–232 (2017).
62. Holm, A., Løvendorf, M. B. & Kauppinen, S. Development of siRNA Therapeutics for the Treatment of Liver Diseases. in *Design and Delivery of SiRNA Therapeutics* (eds. Ditzel, H. J., Tuttolomondo, M. & Kauppinen, S.) 57–75 (Springer US, 2021). doi:10.1007/978-1-0716-1298-9_5.
63. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat Med* 1–8 (2022) doi:10.1038/s41591-021-01672-4.
64. Zaafar, D., Elemary, T., Hady, Y. A. & Essawy, A. RNA-targeting Therapy: A Promising Approach to Reach Non-Druggable Targets. *Biomedical and Pharmacology Journal* **14**, 1781–1790 (2021).
65. Fellmann, C., Gowen, B. G., Lin, P.-C., Doudna, J. A. & Corn, J. E. Cornerstones of CRISPR-Cas in drug discovery and therapy. *Nat Rev Drug Discov* **16**, 89–100 (2017).
66. Schneider, M. *et al.* The PROTACtable genome. *Nat Rev Drug Discov* **20**, 789–797 (2021).

67. He, H., Liu, B., Luo, H., Zhang, T. & Jiang, J. Big data and artificial intelligence discover novel drugs targeting proteins without 3D structure and overcome the undruggable targets. *Stroke Vasc Neurol* **5**, (2020).

5 | The support of genetic evidence from genome-wide association studies for approved drug targets

5.1. Abstract

In the previous chapter it was shown that only 27% of the diseases with an approved drug or a drug under clinical investigation have been studied by genome-wide association studies (GWAS). Despite the limited GWAS data on existing indications, previous studies that mapped genetic associations identified by GWAS to the genes encoding the protein targets of approved drugs have suggested that GWAS could provide a useful tool for systematic identification of new drug targets for human disease. In this chapter, I use a 'truth' set of approved drug target gene – indication pairings to investigate how different *p* value thresholds and physical proximity of the causal gene to the association signal, identified in a GWAS of the intended indication, influence genetic rediscoveries of known drug targets. By expanding the set to compounds in clinical development, I provide an updated estimate of the probability of phase progression for drug target gene - indication pairings given genetic support. The findings showed that the use of stringent *p* value threshold to select significant associations may lead to an oversight of true genetic associations, and relaxing the *p* value threshold to 5×10^{-6} increased the percent of rediscoveries by 32% on average. Moreover, in up to 43% of the genetic association - drug target gene - indication combinations, the target gene was within the five closest genes. Lastly, I provide additional evidence on the value of GWAS for target identification, by showing that the odds to get approved for a target-indication pairing with genetic support is almost three times greater than the odds for a pairing without genetic support, an increase over previous estimates.

5.2. Introduction

Previously, mapping disease loci identified by genome-wide association studies (GWAS) to the genes encoding the protein targets of approved drugs has suggested that GWAS could provide a useful tool for systematic identification of new drug targets for human disease[1]. In fact, after mapping genetic variants to potential causal genes, Nelson *et al.*[2], showed that selecting genetically supported targets could double the success rate in clinical development. This result was then replicated by King *et al*[3]. These studies rely on assigning genetic associations from GWAS data to a causal gene, which remains a challenge in GWAS interpretation because association signals from variants in high linkage disequilibrium (LD) may span multiple genes. Several gold-standard datasets have been used to explore the best approach to assign GWAS signals to genes. These 'truth' sets include genes whose perturbation causes a Mendelian form of a common disease[4], the set of expression and protein QTLs[5], curated metabolite QTLs[6], manually curated examples from the literature[7], and approved drug target-indication pairings where the indication has been studied by GWAS[1,7]. Numerous approaches have been suggested to assign GWAS signals to genes, such as co-localisation[8], or machine-learning techniques[7]. Yet, physical proximity remains the simplest and most widely used proxy to map association signals to genes[6,9]. Although examples exist where the closest gene is not the putative causal gene[10,11], several studies using set of genes with well validated causal relationships to disease have revealed the closest gene to a GWAS signal to be the causal gene in about two-thirds of cases[6], and have shown that the relative distance to the gene is the best single predictor of causal genes[7].

In this chapter, I evaluate the genetic support from GWAS on drug target -indication progression along the drug development process and investigate how often the closest gene is the causal gene when evaluating genetic associations using different threshold *p* values. To do

so, I first create a 'truth' set of approved drug target gene - indication pairs available for rediscovery by GWAS of the corresponding diseases, under the assumption that there is a 1:1 relationship between the drug target gene and its encoded protein. Such dataset was anticipated to include a larger number of drug target gene-indication pairs compared to previous datasets generated by Nelson *et al*., (19,085 target-indication pairs)[2] and King *et al*. (21,934 target-indication pairs)[3]. Second, I evaluate the utility of different *p* value thresholds and physical proximity of the causal gene to the association signal for target identification. Third, I provide an updated estimate of the probability of success for drug target-indication pairings given genetic support. Lastly, I discuss strengths and limitations of using GWAS data to reduce the high attrition rate in drug development due to lack of efficacy.

5.3. Methods

5.3.1. Drug data

Drug data were extracted from ChEMBL version 25 (v25)[12], which included compounds under preclinical (phase 0) or clinical development (phase 1-3), and licensed (phase 4). Information in ChEMBL is itself based on several resources including United States Adopted Name (USAN) applications, ClinicalTrials.gov; the FDA Orange Book database, the British National Formulary, and the ATC classification for compounds with a license. Additional information on intended indications is sourced from DailyMed and the ATC classification. The corresponding drug targets were mapped to UniProt identifiers, and to gene identifiers in Ensembl version 95 (GRCh37) through the updated druggable genome[1](see Chapter 3.2), and the standardised indications in Medical Subject Headings (MeSH) used in ChEMBL v25 were mapped to Unified Medical Language System (UMLS)[13] concepts using the UMLS2020AA to facilitate further mappings (see Chapter 3.1.1.). Compounds flagged as withdrawn, not intended for human use or whose target is encoded by a gene in the extended major histocompatibility complex (xMHC) region (chr6: 28477797- 33448354, GRCh37), were excluded from the analysis. For each drug target gene-indication pairs, the maximum development phase was selected for any drug.

5.3.2. GWAS data

Genetic associations were obtained from the public central repository (GWAS Catalog v1.0.2) and from UK Biobank through Neale data (GWAS Round 2, Results shared 1st August 2018). Genetic associations from UK Biobank were filtered for a p value $\leq 1\times10^{-5}$ to match the minimum significance threshold required by the GWAS Catalog[14]. The GWAS Catalog

included 6,021 MeSH terms from 3,374 publications that were mapped to UMLS concepts. The UK Biobank Neale dataset covered 633 ICD10 main diagnosis that were mapped to 633 UMLS concepts to facilitate the mapping to drug indications. Because one of the aims of this analysis was to compared the updated to previous estimates, the approach described by Nelson *et al.,* 2015[2] was used, which restricted the GWAS data to those indications that have been reasonably well studied by genetic approaches. Therefore, the initial GWAS dataset was further restricted to indications with at least five genetic associations reported and to genetic associations reaching genome-wide significance for the analysis of the probability of success and phase progression given genetic support. In addition to the argument provided by Nelson *et al*.[2], such restriction of the data would also imply that the sample size used in the GWAS was large enough to detect significant association. Disease categories were mapped using a standard list based on the MeSH subcategories (Category C – diseases and Category F - F – Psychiatry and Psychology) and ICD10 chapters. A list of the GWAS traits evaluated is shown in Appendix 5.A.

5.3.3. Linking GWAS associations to drug targets

Two approaches were used to map association signals to drug target genes: absolute distance and relative distance. Using absolute distance, a drug target gene-indication pair was considered to have genetic support if a genetic association with the intended indication was present within the gene boundaries plus or minus 5 kbp. Using the relative distance, a drug target gene-indication pair was considered to have genetic support if the target gene was the closest protein-coding gene to the association signal according to their base pair distance. For each approved drug target-indication pair, genetic associations that overlap within a 1 Mega base pair (Mbp) window upstream and downstream the target gene were extracted. Such

distance has been recently suggested as the cut-off for *cis*- vs *trans*- signals based on empirical evidence from molecular traits, under the assumption that *cis*- signals are acting through the gene in closest proximity[5]. Variants located within the gene were given a distance of 0 bp. For each genetic association - drug target gene - indication combination, the relative distance according to base pair distance from the target gene to the GWAS significant SNP was calculated, using all the genes in the genome excluding the xMHC region (57,392 genes) or limiting the ranking to protein-coding genes excluding the xMHC region (20,147 genes).

5.3.4. Estimating $P(S^+|G^+)$ from $P(G^+|S^+)$

While the interest in drug development is on the probability of success given genetic support $P(S+|G+)$, I only had access to the inverse (i.e., the probability of genetic support given approval $P(G+|S+)$). However, it is possible to derive $P(S+|G+)$ from $P(G+|S+)$ using Bayes' Rule together with information on the proportion of successful and unsuccessful drug target gene - indication pairs and the proportion of drug development programmes with and without genetic support. To do so, information on the number of drug target gene - indication pairs per maximum phase of indication ('no success', S-) was extracted, and the successful pairs (S+) derived by subtracting unsuccessful pairs to the total number of pairs. For drug target-indication pairs with genetic support, the two metrics described in the previous section 5.3.3. were used to define genetic evidence. This information was then used to generate 2x2 tables as follows for each phase of development progression and overall. An example with real data is shown below:

Phase I to phase II	Success (S+)	No success (S-)	Total
Genetic support (G+)	444	69	513
No genetic support (G-)	14,328	3,224	17,552
Total	14,772	3,293	18,065

The probability of genetic support given success $P(G+|S+)$ is given by:

$$P(G+|S+) = \frac{P(G+ \cap S+)}{P(S+)} \Rightarrow P(G+ \cap S+) = P(S+) \cdot P(G+|S+)$$

The probability of success given genetic support $P(S+|G+)$ is given by:

$$P(S+|G+) = \frac{P(S+ \cap G+)}{P(G+)} \Rightarrow P(S+ \cap G+) = P(G+) \cdot P(S+|G+)$$

Since $P(G+ \cap S+) = P(S+ \cap G+)$:

$$P(S+) \cdot P(G+|S+) = P(G+) \cdot P(S+|G+)$$

Thus,

$$P(S+|G+) = \frac{P(S+) \cdot P(G+|S+)}{P(G+)}$$

From the example table, values for $P(S+), P(G+|S+)$ and $P(G+)$ are $\left(\frac{14,772}{18,065}\right)$, $\left(\frac{444}{14,772}\right)$ and $\left(\frac{513}{18,065}\right)$ respectively, thus

$$P(S+|G+) = \frac{\left(\frac{14,772}{18,065}\right) \cdot \left(\frac{444}{14,772}\right)}{\left(\frac{513}{18,065}\right)} = 0.87$$

This process was repeated for all phases of progression using data summarised in the 2 x2 tables in Appendix 5.B.

The following probabilities were also estimated based on the contingency tables: $P(S^+|G^+)$ (Positive predictive value), $P(S^+|G^-)$ (False omission rate), $P(S^-|G^-)$ (Negative predictive value), $P(S^-|G^+)$ (False discovery rate), $P(G^+|\ S^+)$ (Recall rate), $P(G^+|S^-)$ (False positive rate), $P(G^-|\ S^+)$ (False negative rate), $P(G^-|\ S^-)$ (True negative rate), and the following odds: $odds(S^+|G^+)$, $odds(S^+|G^-)$. Subsequently, the following ratios were estimated: $P(S^+|G^+)/P(S^+|G^-)$, $P(S^-|\ G^-)/P(S^-|G^+)$, positive likelihood ratio, negative likelihood ratio, and the diagnostic odds ratio. The calculations are illustrated in Figure 5.1. Confidence intervals were computed using the 'riskratio.boot' function in the 'epitools' R package, and the 'epi.tests' function in the 'epiR' R package.

Total target-indication pairs	**Success (S+)**	**No success (S-)**
Genetic support (G+)	a	b
No genetic support (G-)	c	d

Probabilities		**Likelihood ratios**	**Odds ratio**
P(S+|G+) =	Positive predictive value = a/a+b	$\frac{P(S+\|G+)}{P(S+\|G-)}$	
P(S+|G-) =	False omission rate = c/c+d		
P(S-|G-) =	Negative predictive value = d/c+d	$\frac{P(S-\|G-)}{P(S-\|G+)}$	
P(S-|G+) =	False discovery rate = b/a+b		
P(G+|S+) =	Recall rate = a/a+c	$\frac{P(G+\|S+)}{P(G+\|S-)}$ (Positive likelihood ratio)	$\frac{\text{Positive likelihood ratio}}{\text{Negative likelihood ratio}}$ (Diagnostic odds ratio)
P(G+|S-) =	False positive rate = b/b+d		
P(G-|S+) =	False negative rate = c/c+d	$\frac{P(G-\|S+)}{P(G-\|S-)}$ (Negative likelihood ratio)	
P(G-|S-) =	True negative rate = d/b+d		

Figure 5.1. Probabilities, likelihoods and odds ratios estimated to evaluate the impact of genetic support in drug target gene-indication progression.

5.4. Results

5.4.1. GWAS rediscoveries of approved drug target-indication pairs

Determining if an approved drug target-indication pair has been rediscovered by genetic associations with the intended indication is directly influenced by the definition of genetic evidence. Therefore, I first evaluated the impact of defining genetic evidence using different *p* value thresholds and physical proximity to map genetic associations to causal genes. To do so, a 'truth' set of approved drug target-indication pairs was created by sourcing data on approved drugs, their targets and intended indications from ChEMBL v25[12]. Following the approach of Nelson *et al.* 2015[2], drugs with nonhuman (e.g., antimicrobial drugs which target a protein in the pathogen) or extended major histocompatibility complex (xMHC) targets were excluded. In total, ChEMBL included 371 indications (UMLS concepts) and 898 approved drugs which target proteins encoded by 665 drug target genes (Fig. 5.2). The total number of unique drug target gene-indications pairs was 3,118.

Genetic associations were obtained from a public central repository (GWAS Catalog v1.0.2) and from UK Biobank through Neale dataset. Overall, 213 approved indications were covered in genetic studies (GWAS Catalog plus UK Biobank, Appendix 5.A), which represented 2,338 unique target-indication pairs (Fig. 5.2).

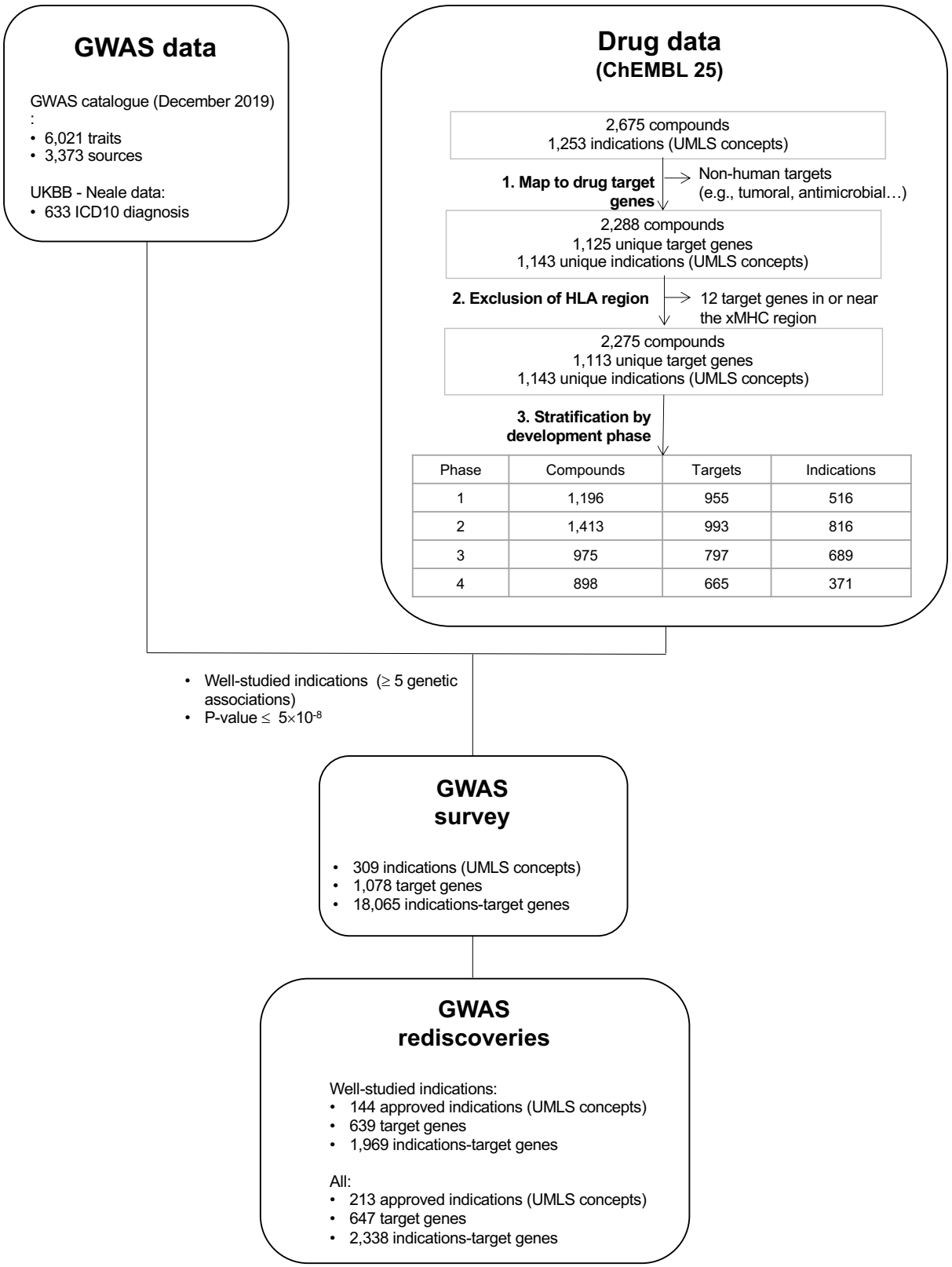


Figure 5.2. Summary of data sources and mappings between them. Summary of each data source and the key filtering and processing steps applied to create the final set of gene-trait and drug target–indication combinations investigated in this study. GWAS Catalog sources correspond to unique PubMed ID.

To explore if the absolute distance could help identifying the causal gene responsible for an association signal, and under the assumption that the drug target gene should be the causal gene in the region, the distance from the associated variant to the gene of interest was calculated. It was found that, as the flanking region expanded, the number of drug target gene-indications rediscovered increased at the cost of increasing the median number of protein-coding genes between the target gene and the genetic association (Fig. 5.3). For example, genetic associations could be found within 1Mbp for 27% of the drug target gene – indication pairs (p value ≤ 1×10^{-5}), being the drug target gene within the closest six genes for most drug target gene – indication pairs explored. Noticeably, the percentage of drug target gene – indication pairs rediscovered did not reach 100%. This is explained by genetic associations located in a chromosome other than the chromosome containing the gene encoding the drug target.

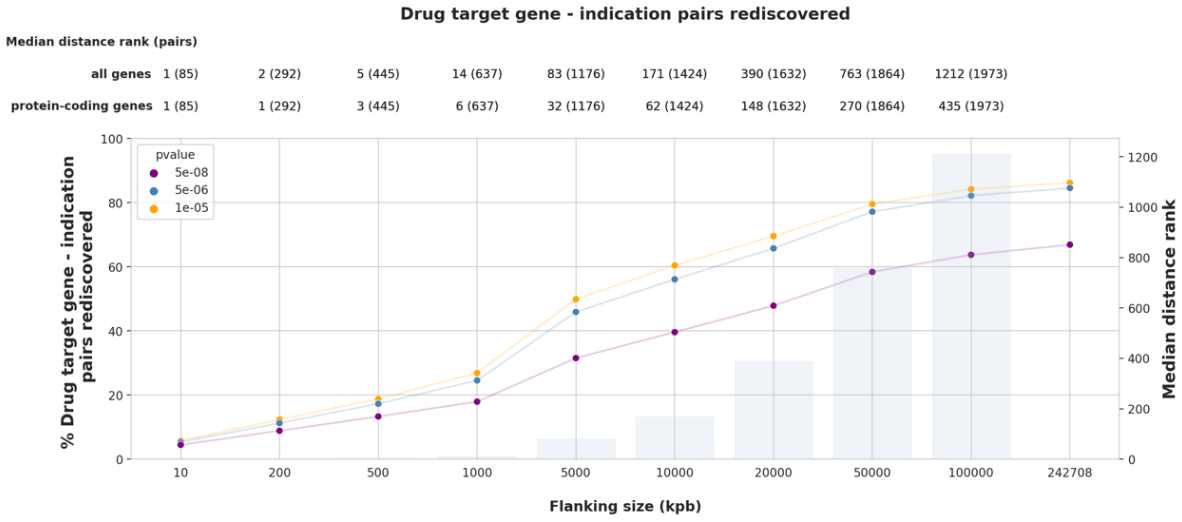


Figure 5.3. Analyses of drug target genes-indications pairs rediscovered by GWAS associations. SNP associations with the intended indication were mapped to the gene encoding the drug target allowing for different flanking regions: 242Mbp (whole chromosome 2, largest chromosome), 100Mbp, 50 Mbp, 20 Mbp, 10Mbp, 5Mbp,1 Mbp and 500kbp, 100kbp, 200kbp and 10kbp, for three significance thresholds: 5×10^{-8}, 5×10^{-6}, 1×10^{-5}. Total: 2,338 approved drug target gene – indication pairings, of which 2,023 had associations in the same chromosome as the drug target gene.

Subsequently, the gene distance rank (or relative distance) was defined using all or protein-coding genes in the region to investigate how often the closest gene is the causal gene when a significant genetic association with the intended indication lies in or near the gene of interest. Table 5.1 shows the percentage of drug target gene - indication pairs rediscovered by the relative distance using different *p* value thresholds, with an illustration of the calculation of the different measures in Figure 5.4. It was found that when filtering for genome-wide significant associations, the closest protein-coding gene was the drug target gene in 20.5% (95% CI: 18.9; 22.1) of the 2,441 genetic association - drug target gene - indication combinations explored (31.6% of the target genes rediscovered), with an enrichment of GWAS signals within 250 kbp upstream the target gene (Fig. 5.5). Moreover, in 42.8% of cases the drug target gene was within the five closest protein-coding genes. The percentage decreases when including all the genes in the region, as shown by the decrease to 16.4% (95% CI: 14.9; 17.9) of rediscoveries for the total genetic association - drug target gene - indication combinations when filtering for genome-wide significant associations. Lastly, it was also observed that relaxing the *p* value threshold used to filter significant genetic association did not have a substantial impact on the number of drug target gene -indication pairs rediscovered.

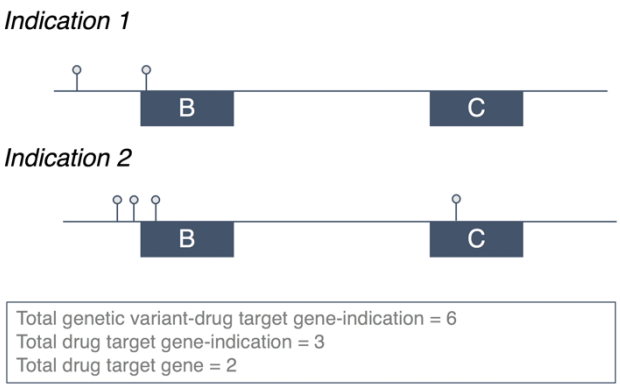


Figure 5.4. Illustration of the calculation of the measures genetic variant-drug target gene-indication, drug target gene-indication and drug target gene used for the rediscoveries estimation. The boxes represent drug target genes B and C. Genetic variants are represented with lollipops.

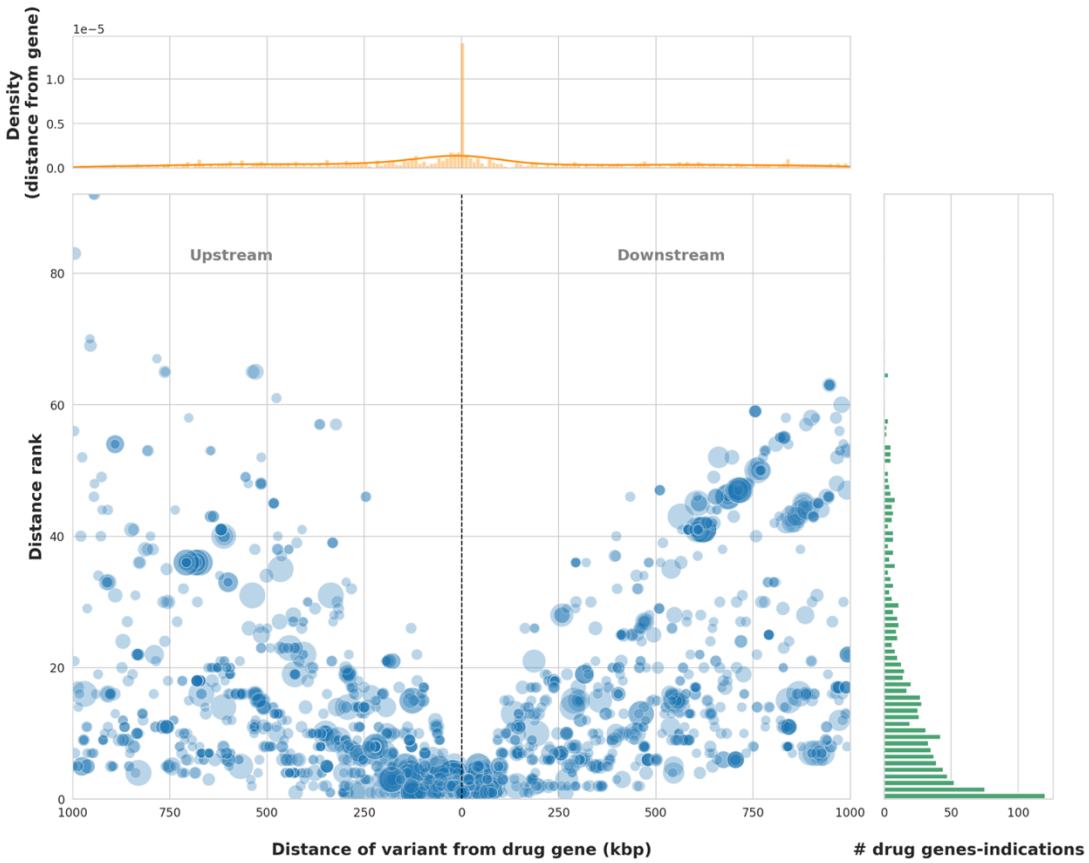


Figure 5.5. Absolute and relative distance of drugged target genes to GWAS SNPs (p value ≤ 5×10^{-8}). Each point in the scatterplot represents a GWAS signal located within 1Mbp of a drugged gene, where the GWAS trait represents the intended indication. The position on the x axis indicates the absolute distance of the SNP to the drugged target gene. Position in the y axis indicates the number of protein-coding genes in the interval that are closer to the signal than the drugged target gene, excluding those in the xMHC region. The top panel indicates the signal density for all such SNPs, and the side panel provides the counts of drug target gene-indication pairs rediscovered using a gene distance rank = 1. Total number of drug target gene-indication pairs overlapping genetic associations (p value ≤ 5×10^{-8}) with the intended indication: 425.

Table 5.1. Rediscoveries by relative distance. Percentage and 95% CI of rediscoveries at the level of genetic variant (i.e., the denominator is the total genetic variant-drug target gene-indication combinations), drug target gene-indication pair (i.e., the denominator is the total drug target gene-indication pairs), and drug target gene (i.e., the denominator is the total drug target gene investigated, regardless of the indication).

Gene distance rank	Gene distance rank including all the genes			Gene distance rank including protein-coding genes		
	Percentage of rediscoveries at the level of genetic variant (%)	Percentage of rediscoveries at the level of drug target gene-indication pair (%)	Percentage of rediscoveries at the level of drug target gene (%)	Percentage of rediscoveries at the level of genetic variant (%)	Percentage of rediscoveries at the level of drug target gene-indication pair (%)	Percentage of rediscoveries at the level of drug target gene (%)
SNPs with p value ≤ 1×10^{-5}						
1	15.3 (14.1; 16.5)	20.7 (17.68; 23.8)	26.3 (21.7; 30.9)	19.2 (17.8; 20.6)	25.8 (22.4; 29.2)	31.5 (26.7; 36.3)
2	3.8 (3.1; 4.5)	10.5 (8.1; 12.9)	14.3 (10.7; 17.9)	8.4 (7.4; 9.4)	14.2 (11.5; 16.9)	18.3 (14.3; 22.3)
3 - 5	10.0 (9.0; 11.0)	16.8 (13.9; 19.7)	20.7 (16.5; 24.9)	13.3 (12.1; 14.5)	27.0 (23.6; 30.4)	33.7 (28.8; 38.6)
6-10	8.2 (7.3; 9.1)	18.7 (15.7; 21.7)	24.9 (20.4; 29.4)	18.4 (17.1; 19.7)	32.9 (29.3; 36.5)	42.4 (37.3; 47.5)
>10	62.8 (61.1; 64.5)	81.3 (78.3; 84.3)	86.5 (83.0; 90.0)	40.7 (39.0; 42.4)	56.1 (52.2; 60.0)	63.2 (58.2; 68.2)
Total	3265	637	357	3265	637	357
SNPs with p value ≤ 5×10^{-6}						
1	15.4 (14.1; 16.7)	20.8 (17.5; 24.1)	25.2 (20.6; 29.8)	19.4 (18.0; 20.8)	25.9 (22.3; 29.5)	30.2 (25.4; 35.0)
2	3.9 (3.2; 4.6)	11.1 (8.6; 13.6)	14.5 (10.8; 18.2)	8.6 (7.6; 9.6)	14.9 (12.0; 17.8)	18.3 (14.2; 22.4)
3 - 5	10.1 (9.0; 11.2)	17.2 (14.1; 20.3)	19.7 (15.5; 23.9)	13.3 (12.1; 14.5)	27.5 (23.9; 31.1)	32.8 (27.8; 37.8)
6-10	8.2 (7.2; 8.2)	19.2 (16.0; 22.4)	24.3 (19.8; 28.8)	18.3 (16.9; 19.7)	33.2 (29.4; 37.0)	40.7 (35.5; 45.9)
>10	62.4 (60.7; 64.1)	82.0 (78.9; 85.1)	86.4 (82.8; 90.0)	40.5 (38.8; 42.2)	56.2 (52.2; 60.2)	63.1 (58.0; 68.2)
Total	3118	583	345	3118	583	345
SNPs with p value ≤ 5×10^{-8} (Genome-wide significant)						
1	16.4 (14.9; 17.9)	24.5 (20.4; 28.6)	28.9 (23.5; 34.3)	20.5 (18.9; 22.1)	28.2 (23.9; 32.5)	31.6 (26.0; 37.2)
2	3.9 (3.1; 4.7)	12.0 (8.9; 15.1)	15.0 (10.7; 19.3)	9.0 (7.9; 10.1)	17.6 (14.0; 21.2)	20.7 (15.8; 25.6)
3 - 5	10.8 (9.6; 12.0)	19.8 (16.0; 23.6)	21.1 (16.2; 26.0)	13.3 (12.0; 14.6)	26.6 (22.4; 30.8)	29.7 (24.2; 35.2)
6-10	8.0 (6.9; 9.1)	20.5 (16.7; 24.3)	24.4 (19.2; 29.6)	17.5 (16.0; 19.0)	33.9 (29.4; 38.4)	41.4 (35.5; 47.3)
>10	60.9 (59.0; 62.8)	80.2 (76.4; 84.0)	86.5 (82.4; 90.6)	39.7 (37.8; 41.6)	55.3 (50.6; 60.0)	62.8 (57.0; 68.6)
Total	2441	425	266	2441	425	266

5.4.2. Probability of success and phase progression given genetic support

To provide a revised estimate of the probability of drug development progression given the drug target has genetic support in the intended indication, the dataset was expanded to drugs in clinical development (phase I, II, III clinical trials).

Compounds at various stages of clinical development, their indications, maximum development phase and targets were extracted from ChEMBL v25. Of a total of 2,675 compounds, 2,275 were known to modulate the target encoded by 1,113 non-xMHC genes for 1,143 UMLS indications. Of the 1,113 non-xMHC genes, 604 encoded single protein targets, 668 encoded a protein belonging to a protein family or protein complex, eight encoded proteins involved in a selectivity group (i.e. pair of proteins for which selectivity has been assessed), two encoded a part of a chimeric protein, and four encoded proteins involved in protein-protein interactions. Data were aggregated at drug target gene-indication level to avoid duplications due to shared mechanism of action between compounds and to account for multiple genes involved in a single target scenario (i.e. protein complex). This yielded a total of 32,022 drug target gene-indication pairs.

To allow comparisons with previous estimates, the approach described by Nelson *et al.,* 2015[2] was followed and the summary results were filtered for those GWAS traits that contained at least 5 genome-wide significant associations (p value $\leq 5\times10^{-8}$), yielding 3,403 traits that had been reasonably investigated by GWAS.

To investigate the association of genetic support for progression or approval of drug target-indications, the overlap between the 3,403 GWAS traits and the 1,143 indications (UMLS concepts) reported in ChEMBL v25 was used (Fig. 5.2). This returned a total of 309 unique indications for 1,078 unique drug target genes (18,065 target genes-indications pairs, in

contrast to 4,184 pairs by Nelson *et al*. 2015[2]), of which 1,969 target gene-indication pairs corresponded to 144 unique indications and 639 unique targets encoding genes for approved drugs. To determine genetic support, two previously published definitions of genetic evidence were used and compared: i) if a genetic association with the intended indication was present within the gene boundaries plus or minus 5 kbp[2] or ii) if the target gene was the closest protein-coding gene according to their base pair distance (relative distance or gene distance rank)[1].

Of the 1,969 approved target gene-indication pairs that overlapped genetic associations with the intended indication, 123 (6.2%; 95% CI: 5.2; 7.3) and 150 (7.6%; 95%CI: 6.4; 8.8) approved drug target-indications pairs were supported by one or more genetic associations, when defining genetic support based on absolute distance or relative distance, respectively. Moreover, variability was found among indication areas, with chemically-induced disorders (e.g. alcoholism) and circulatory system diseases showing the highest degree of genetic support (Fig. 5.6), and neoplasms and diseases of the genitourinary system the lowest evidence. Such low genetic support for drug targets in neoplasms could be explained by cancer drugs targeting proteins that are overexpressed in the tumour due to somatic mutations, and therefore, genetic associations from GWAS which are based on germline variation may not be as useful in these diseases. As expected, the percentage of drug target gene - indication pairs with genetic evidence increased with phase progression (Table 5.2). For instance, when using the relative distance (gene distance rank = 1) to define genetic evidence, 2.9% of drug target gene – indication pairs had genetic support at phase I compared to the 7.6% at approved phase.

Table 5.2. Drug target gene - indication pairs with genetic support by maximum phase of indication and source of genetic evidence

Maximum phase of indication	Number of drug target gene – indication pairs	Number of drug target gene – indication pairs with genetic support	Percentage	Source of genetic evidence
Phase I	3293	69	2.10	absolute distance
Phase II	8279	174	2.10	absolute distance
Phase III	4524	147	3.25	absolute distance
Approved	1969	123	6.25	absolute distance
Total	**18065**	**513**	**2.84**	**absolute distance**
Phase I	3293	95	2.88	relative distance
Phase II	8279	227	2.74	relative distance
Phase III	4524	185	4.09	relative distance
Approved	1969	150	7.62	relative distance
Total	**18065**	**657**	**3.64**	**relative distance**

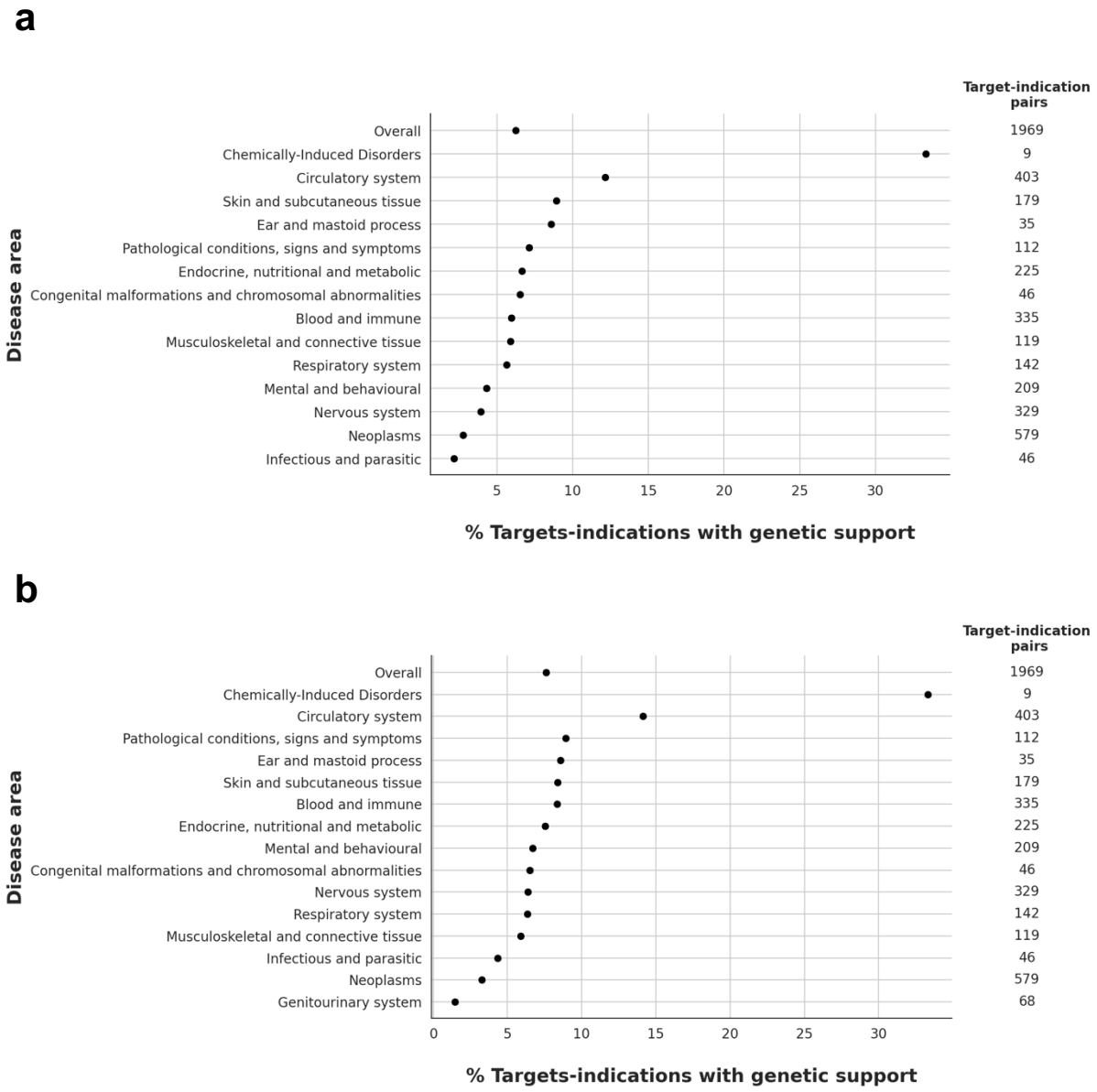


Figure 5.6. Proportion of approved target-indications pairs by disease area with genetic support. Genetic support defined as a genetic association is present within 5kbp window from the gene (a) or the target gene being the closest gene to a genetic association with the intended indication (b). Disease are is defined by ICD10 chapter.

Subsequently, I calculated the probabilities, likelihoods and odds ratios as described in Figure 5.1 to estimate the rate of success and phase progression given genetic support using contingency tables (Appendix 5.C). It was found that the probability of a drug target gene - indication pair with genetic support (gene distance rank = 1) progressing from phase I to approval was 2.18 (95%CI: 1.86; 2.51) times the probability of progressing without genetic evidence. For drug target gene – indication pairs that did not progress in the pipeline, it was found that the lack of genetic support, estimated as the probability of no progression without genetic support P(S-|G-) divided by the probability of no progression with genetic support P(S-|G+), had the greatest impact from phase II to phase III (1.40, 95%CI: 1.28; 1.56; Appendix 5.C), as expected since phase II trials aim to evaluate clinical efficacy. Lastly, the ratio of the probability of a drug target gene – indication pair progressing in the drug development pipeline with genetic support P(S+|G+) was compared to the probability of the drug progressing without genetic support P(S+|G-) to that previously reported by Nelson *et al.,* 2015[2] (Table 5.3). These findings suggest that selecting genetically supported targets could increase the success rate in clinical development and, after the comparison with previous studies, the estimates in the current analysis which are based on a larger dataset indicate that the increase may be greater than two-fold.

Table 5.3. Comparison of the relative value of genetic support for the probability that a target indication-pair progresses along the drug development pipeline with estimates by Nelson *et al.* 2015[2].

	$\frac{P(S+\|G+)}{P(S+\|G-)}$		
	Data source for genetic associations: GWAS Catalog (Genetic support based on absolute distance)	**Data source for genetic associations: GWAS Catalog (Genetic support based on relative distance)**	**Data source for genetic associations: GWASdb (Nelson et al. 2015)**
Phase I to approval	2.3 (1.9; 2.7)	2.2 (1.9; 2.5)	1.8 (1.3; 2.3)
Phase III to approval	1.5 (1.3; 1.8)	1.5 (1.3; 1.7)	1.0 (0.8; 1.2)
Phase II to Phase III	1.4 (1.3; 1.5)	1.4 (1.3; 1.5)	1.4 (1.2; 1.7)
Phase I to Phase II	1.1 (1.0; 1.1)	1.0 (1.0; 1.1)	1.2 (1.1; 1.3)
Total number of target-indication pairings evaluated	18,065	18,065	4,184

5.5. Discussion

5.5.1. Summary

The growing interest and investment in human genomics to inform drug development demands solid evidence of the potential value of GWAS and GWAS-based approaches for target identification and validation. In the previous chapter it was shown that, despite the increasing evidence from the published literature on the usefulness of human genetic data in drug target discovery, prioritisation and validation, a large proportion of human diseases have not yet been studied by GWAS and still an enormous sample space of genes – human diseases can be interrogated through GWAS to generate evidence on novel drug target gene – indication pairs, repurposing opportunities and expansion of new indications for drugs already approved. The analysis presented in this chapter further supports previous statements on the potential of genetic evidence in drug development by providing a revised estimate of the likelihood of progression in the drug development pipeline and a detailed analysis of the characteristics of the approved drug target-indication pairs rediscovered by GWAS.

In this chapter I investigated the number of drug target gene - indications pairs rediscovered by genetics by interrogating publicly available GWAS data from studies based research-based case ascertainment (GWAS Catalog) and routine electronic health records (UK Biobank). The findings show that, as the flanking region expanded, the number of target-indications rediscovered increased at the cost of increasing the median number of protein-coding genes between the target gene and the genetic association. Using a stringent *p* value threshold to select significant associations may lead to an oversight of true genetic associations, and relaxing the *p* value threshold to 5×10^{-6} increased the percent of rediscoveries by 32% on average. Moreover, in 21% of the associations-target-indications pairs explored the closest protein-coding gene was the target gene, which represents 32% of the total drug target genes

available to be rediscovered. Further, in up to 43% of the genetic association – drug target gene - indication combinations the target gene was within the five closest genes.

Using a 'truth' set of drug target-indication pairings, I provided further evidence that pairings with genetic support are twice more likely to get approved than those without genetic support (2.18; 95%CI: 1.86; 2.51). I found that the probability of progression given genetic support increases along the clinical phases (Table 5.3) and that the lack of genetic support had the greatest impact from phase II to phase III (P(S-|G-) / P(S-|G+) = 1.40, 95%CI: 1.28; 1.56), where drugs are typically tested for clinical efficacy.

5.5.2. Research in context

The results presented here are compared to existing knowledge in this area. Similar to the findings presented here, a study by Mountjoy *et al*., 2021[7] and funded by OpenTargets which evaluated different genomic features in a model to predict causal protein-coding genes at GWAS loci reported that the mean distance was the most predictive feature, where the distance relative to other genes is more important than the absolute distance. The present study also showed that the relative distance (gene distance rank = 1) rediscovered more drug target-indication pairs than the use of the absolute distance (i.e., absolute distance). In addition, it was found that in 27% of the drug target gene – indication pairs, genetic associations with the indication were within 1 Mbp from the drug target gene, and that increasing the genomic distance led to a change in the curve from exponential to logarithmic (Fig. 5.1) suggesting that expanding the region around the drug target gene would not substantially increase the number of rediscoveries but rather increase the median number of protein-coding genes between the target gene and the genetic association. In fact, in a recent publication Fauman *et al.,*2022[5] have estimated a distance cut-off of 944 kbp (95%CI 767-1,161) separating the *cis* and *trans*

regimes, which in line with the results from this chapter, suggests that approaches for mapping genetic associations to genes based on distance should be restricted to a maximum of 1 Mbp.

Previous work by Nelson *et al*., 2015[2] and King *et al.,* 2019[3] showed that targets with genetic evidence were more likely to be successful in clinical development. Here, using two approaches for genetic evidence and a larger dataset (18,065 drug target-indications pairs), I further confirmed that the probability of a target-indication pair with genetic support progressing from phase I to approval to the probability of progressing without genetic evidence is greater than two-fold.

5.5.3. Strengths and limitations

This study has several strengths. First, the 'truth' dataset included 32,022 drug target-indications pairs as the initial set to estimate the value of genetic support in phase progression. This presents almost 10,000 more pairings compared to the target-indication pairs reported by King *et al*., 2019[3] (21,934) and that used by Nelson *et al.*, 2015[2] (19,085). Second, it focused not only on GWAS data from studies based on research-based case ascertainment, but also included genetic associations from electronic health records (UK Biobank). Third, two metrics based on absolute distance were used to define genetic evidence, where assigning the closest gene as the causal gene has been previously described as the simplest and, in many cases, the most accurate way to assign genetic association to causal genes. Fourth, different probabilities, odds and ratios often used to evaluate the performance of diagnostic tests, such as the positive predictive value and the false discovery rate, were computed to provide multiple metrics of the value of genetic support by clinical phase.

Some limitations of this study are noteworthy. First, while this study covered the largest dataset of drug target gene - indication pairs to date, there were limitations due to i) the genetic mapping, as the ability to identify causal genes was based on proximity-based rather than co-localisation approaches ; and ii) the indication mapping, as it may exclude drug target gene - indication pairs where the intended indication has not been studied by GWAS but has available data on clinically-validated biomarkers. However, the latter issue should have been captured to some extent by the annotation of related terms in the GWAS Catalog. Moreover, certain indications may have been studied by GWAS but could not be included in this study because the summary statistics were not deposited in the GWAS Catalog. Even for those studies included in the analysis, genetic associations may have been missed due to sample sizes not being large enough to detect all the responsible genes; or due to incomplete genomic coverage by the genotyping array. There are several reasons for drug discontinuation besides lack of efficacy, including safety concerns, strategic decisions or the compound failing to show extra benefits compared to another treatment. The limited data in the public domain on drug failures and their reason for discontinuation makes it difficult to account for this variable in the current analysis. However, it is expected that the inclusion of such drug candidates will not inflate the inference made on the value of genetic support on phase progression. A 1:1 relationship was assumed between the gene and the encoded protein that is targeted by a drug. Such assumption may not always prevail as some genes encode multiple proteins due to, for example, post-transcriptional modifications. Lastly, another potential source of bias is that genetic evidence from GWAS may already be used to inform drug development. However, in line with the argument presented by Nelson *et al*., 2015[2] and due to the long timelines in drug development (on average 10 years), the impact of this bias would not inflate the estimate but rather underestimate the value of genetic support as it would increase the number of drugs with genetic support in the early phases of the development process.

5.6. Conclusion

As described in Chapter 4, only a small fraction of drug indications have been investigated by GWAS. However, the analysis performed in this chapter provides further evidence that drug target – indications pairings with genetic support from GWAS are more likely to progress in the drug development pipeline. In the next chapter, I will investigate how GWAS data can be leveraged using drug target Mendelian Randomisation to further support target validation by inferring the correct mechanism of action for a new drug.

5.7. References

1. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, (2017).
2. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856–860 (2015).
3. King, E. A., Davis, J. W. & Degner, J. F. 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**, e1008489 (2019).
4. Forgetta, V. *et al.* An effector index to predict target genes at GWAS loci. *Hum Genet* (2022) doi:10.1007/s00439-022-02434-z.
5. Fauman, E. B. & Hyde, C. An optimal variant to gene distance window derived from an empirical definition of cis and trans protein QTLs. 2022.03.07.483314 Preprint at https://doi.org/10.1101/2022.03.07.483314 (2022).
6. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, e3–e3 (2019).
7. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* 1–7 (2021) doi:10.1038/s41588-021-00945-5.
8. Giambartolomei, C. *et al.* A Bayesian Framework for Multiple Trait Colo-calization from Summary Association Statistics. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty147.
9. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* 1–6 (2021) doi:10.1038/s41586-021-03446-x.
10. Porcu, E. *et al.* Mendelian randomisation integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat Commun* **10**, 3300 (2019).

11. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
12. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940 (2018).
13. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-270 (2004).
14. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896–D901 (2017).

5.8. Appendices

Appendix 5.A. Indications of approved drugs studied by GWAS.

UMLS concept	UMLS concept ID
Carcinoma, Non-Small-Cell Lung	C0007131
Depressive Disorder	C0011581
Acute Pain	C0184567
Schizophrenia	C0036341
Bipolar Disorder	C0005586
Aggression	C0001807
Tendinopathy	C1568272
Rhinitis, Allergic, Seasonal	C0018621
Asthma	C0004096
Rhinitis, Allergic	C2607914
Urticaria	C0042109
Cardiovascular Diseases	C0007222
Migraine Disorders	C0149931
Hodgkin Disease	C0019829
Psychotic Disorders	C0033975
Autistic Disorder	C0004352
Depressive Disorder, Major	C1269683
Anxiety	C0003467
Back Pain	C0004604
Hypertension	C0020538
Arrhythmias, Cardiac	C0003811
Nephrotic Syndrome	C0027726
Stroke	C0038454
Heart Failure	C0018801
Liver Cirrhosis	C0023890
Dyslipidemias	C0242339
Venous Thromboembolism	C1861172
Atrial Fibrillation	C0004238
Thrombosis	C0040053
Venous Thrombosis	C0042487
Pulmonary Embolism	C0034065
Immune System Diseases	C0021053
Polycythemia Vera	C0032463
Neoplasms	C0027651

UMLS concept	UMLS concept ID
Melanoma	C0025202
Epilepsies, Partial	C0014547
Osteoporosis	C0029456
Osteoporosis, Postmenopausal	C0029458
Multiple Myeloma	C0026764
Tobacco Use Disorder	C0040336
Primary Ovarian Insufficiency	C0085215
Osteoporotic Fractures	C0521170
Osteitis Deformans	C0029401
Acne Vulgaris	C0001144
Arthritis, Juvenile	C3495559
Prostatic Hyperplasia	C2937421
Arthritis, Rheumatoid	C0003873
Colitis, Ulcerative	C0009324
Depression	C0011570
Epilepsy	C0014544
Neuroendocrine Tumors	C0206754
Neoplasm Metastasis	C0027627
Thyroid Neoplasms	C0040136
Diarrhea	C0011991
Cystic Fibrosis	C0010674
Waldenstrom Macroglobulinemia	C0024419
Multiple Sclerosis, Chronic Progressive	C0393665
Granulomatosis with Polyangiitis	C3495801
Microscopic Polyangiitis	C2347126
Multiple Sclerosis	C0026769
Psoriasis	C0033860
Precursor Cell Lymphoblastic Leukemia-Lymphoma	C1961102
Crohn Disease	C0010346
Leukemia, Myeloid, Acute	C0023467
Prostatic Neoplasms, Castration-Resistant	C3658267
Melanosis	C0025209
Lentigo	C0023321
Arthritis, Psoriatic	C0003872

UMLS concept	UMLS concept ID
Leukemia, Myelogenous, Chronic, BCR-ABL Positive	C0023473
Prostatic Neoplasms	C0033578
Carcinoma, Renal Cell	C0007134
Gastroesophageal Reflux	C0017168
Lymphoma, Non-Hodgkin	C0024305
Leukemia, Lymphocytic, Chronic, B-Cell	C0023434
Lymphoma, Follicular	C0024301
Parkinson Disease	C0030567
Myelodysplastic Syndromes	C3463824
Pruritus	C0033774
Sinusitis	C0037199
Sepsis	C0243026
Restless Legs Syndrome	C0035258
Tourette Syndrome	C0040517
Diabetes Mellitus, Type 2	C0011860
Conduct Disorder	C0149654
Angioedema	C0002994
Heart Arrest	C0018790
Glaucoma	C0017601
Hemorrhage	C0019080
Pain	C0030193
Liver Cirrhosis, Biliary	C0023892
Infection	C3714514
HIV Infections	C0019693
Turner Syndrome	C0041408
Renal Insufficiency, Chronic	C0403447
Diabetes Mellitus	C0011849
Coronary Artery Disease	C1956346
Angina Pectoris	C0002962
Hyperlipidemias	C0020473
Myocardial Infarction	C0027051
Hypercholesterolemia	C0020443
Nausea	C0027497
Varicose Veins	C0042345
Glaucoma, Open-Angle	C0017612
Ocular Hypertension	C0028840
Hepatitis C, Chronic	C0524910
Kidney Failure, Chronic	C0022661

UMLS concept	UMLS concept ID
Peritoneal Neoplasms	C0031149
Ovarian Neoplasms	C0919267
Familial Primary Pulmonary Hypertension	C0340543
Virus Diseases	C0042769
Hepatitis B	C0019163
Ventricular Dysfunction, Left	C0242698
Parkinson Disease, Secondary	C0030569
Erectile Dysfunction	C0242350
Diabetic Nephropathies	C0011881
Dementia	C0497327
Diabetic Retinopathy	C0011884
Pancreatic Neoplasms	C0030297
Precursor B-Cell Lymphoblastic Leukemia-Lymphoma	C0023485
Anemia	C0002871
Kidney Diseases	C0022658
Wet Macular Degeneration	C2237660
Postpartum Hemorrhage	C0032797
Alzheimer Disease	C0002395
Myasthenia Gravis	C0026896
Macular Edema	C0271051
Retinal Neovascularization	C0035320
Colorectal Neoplasms	C0009404
Ventricular Fibrillation	C0042510
Inflammation	C0021368
Leukemia	C0023418
Gout	C0018099
Hyperuricemia	C0740394
Common Cold	C0009443
Familial Mediterranean Fever	C0031069
Epilepsy, Absence	C0014553
Acute Coronary Syndrome	C0948089
Pulmonary Disease, Chronic Obstructive	C0024117
Peripheral Arterial Disease	C1704436
Bronchitis, Chronic	C0008677
Urinary Bladder Neoplasms	C0005695
Intraocular Pressure	C0021888
Rosacea	C0035854
Pharyngitis	C0031350

UMLS concept	UMLS concept ID
Breast Neoplasms	C1458155
Sleep Initiation and Maintenance Disorders	C0021603
Osteoarthritis	C0029408
Anxiety Disorders	C0003469
Panic Disorder	C0030319
Dupuytren Contracture	C0013312
Obesity	C0028754
Dermatitis, Atopic	C0011615
Rheumatic Diseases	C0035435
Polymyositis	C0085655
Pemphigus	C0030807
Stevens-Johnson Syndrome	C0038325
Chronic Pain	C0150055
Sarcoidosis	C0036202
Dermatomyositis	C0011633
Gaucher Disease	C0017205
Irritable Bowel Syndrome	C0022104
Vomiting	C0042963
Skin Diseases	C0037274
Hypersensitivity	C0020517
Glaucoma, Angle-Closure	C0017605
Tachycardia, Paroxysmal	C0039236
Chorioretinitis	C0008513
Intracranial Arteriosclerosis	C0007771
Anemia, Hemolytic	C0002878
Neuralgia	C0027796
Alcoholism	C0001973
Uterine Cervical Neoplasms	C0007873
Adrenal Insufficiency	C0001623
Keratitis	C0022568
Tuberculosis, Meningeal	C0041318
Otitis Externa	C0029878
Herpes Labialis	C0019345
Hypothyroidism	C0020676

UMLS concept	UMLS concept ID
Idiopathic Pulmonary Fibrosis	C1800706
Attention Deficit Disorder with Hyperactivity	C1263846
Influenza, Human	C0021400
Dysmenorrhea	C0013390
Stomach Ulcer	C0038358
Spondylitis, Ankylosing	C0038013
Esophagitis	C0014868
Heartburn	C0018834
Duodenal Ulcer	C0013295
Anemia, Iron-Deficiency	C0162316
Lupus Erythematosus, Systemic	C0024141
Narcolepsy	C0027404
Seasonal Affective Disorder	C0085159
Sleep Wake Disorders	C4042891
Thrombocytopenia	C0040034
Glioblastoma	C0017636
Huntington Disease	C0020179
Dermatitis Herpetiformis	C0011608
Substance-Related Disorders	C0236969
Lipodystrophy	C0023787
Cough	C0010200
Tuberculosis, Pulmonary	C0041327
Hypertension, Pulmonary	C0020542
Pneumonia	C0032285
Muscular Dystrophy, Duchenne	C0013264
Uveitis, Anterior	C0042165
Dermatitis, Seborrheic	C0036508
Neurotic Disorders	C0027932
Eye Diseases	C0015397
Diabetes Mellitus, Type 1	C0011854
Colonic Neoplasms	C0009375
Thrombocythemia, Essential	C0040028
Urinary Tract Infections	C0042029
Shock, Septic	C0036983

Appendix 5.B. Contingency tables to estimate the impact of genetic evidence on progressing in the drug development pipeline, for drug target gene - indication pairs where the indication had at least 5 genetic associations reaching genome-wide significance.

	Absolute distance[1]		**Relative distance[2]**	
	Success (S+)	No success (S-)	Success (S+)	No success (S-)
Phase I to phase II				
Genetic support (G+)	444	69	562	95
No genetic support (G-)	14328	3224	14210	3198
Phase II to phase III				
Genetic support (G+)	270	174	335	227
No genetic support (G-)	6223	8105	6158	8052
Phase III to Approval				
Genetic support (G+)	123	147	150	185
No genetic support (G-)	1846	4377	1819	4339
Phase I to Approval				
Genetic support (G+)	123	390	150	507
No genetic support (G-)	1846	15706	1819	15589

[1]*Absolute distance: genetic association with the intended indication present within the gene boundaries plus or minus 5 kbp*

[2]*Relative distance: target gene is the closest protein-coding gene according to their base pair distance (i.e., gene distance rank = 1)*

Appendix 5.C. Probabilities, likelihoods and odds ratios to estimate the rate of success and phase progression given genetic support based on the 2x2 tables in Appendix 5.B.

	Genetic support based on the absolute distance				**Genetic support based on the relative distance (distance rank = 1)**			
	Phase I to II	**Phase II to III**	**Phase III to approval**	**Phase I to approval**	**Phase I to II**	**Phase II to III**	**Phase III to approval**	**Phase I to approval**
P(S+|G+) =Positive predictive value	0.87 (0.83; 0.89)	0.61 (0.56; 0.65)	0.46 (0.40; 0.52)	0.24 (0.20; 0.28)	0.86 (0.83; 0.88)	0.60 (0.56; 0.64)	0.45 (0.39; 0.50)	0.23 (0.20; 0.26)
P(S+|G-) = False omission rate	0.82 (0.81; 0.82)	0.43 (0.43; 0.44)	0.30 (0.29; 0.32)	0.11 (0.10; 0.11)	0.82 (0.81; 0.82)	0.43 (0.43; 0.44)	0.30 (0.29; 0.31)	0.10 (0.10; 0.11)
P(S-|G-) = Negative predictive value	0.18 (0.18; 0.19)	0.57 (0.56; 0.57)	0.70 (0.69; 0.71)	0.89 (0.89; 0.90)	0.18 (0.18; 0.19)	0.57 (0.56; 0.57)	0.70 (0.69; 0.72)	0.90 (0.89; 0.90)
P(S-|G+) = False discovery rate	0.13 (0.11; 0.17)	0.39 (0.35; 0.44)	0.54 (0.48; 0.60)	0.76 (0.72; 0.80)	0.14 (0.12; 0.17)	0.40 (0.36; 0.45)	0.55 (0.50; 0.61)	0.77 (0.74; 0.80)
P(G+|S+) = Recall rate	0.03 (0.03; 0.03)	0.04 (0.04; 0.05)	0.06 (0.05; 0.07)	0.06 (0.05; 0.07)	0.04 (0.04; 0.04)	0.05 (0.05; 0.06)	0.08 (0.06; 0.09)	0.08 (0.06; 0.09)
P(G+|S-) = False positive rate	0.02 (0.02; 0.03)	0.02 (0.02; 0.02)	0.03 (0.01; 0.05)	0.02 (0.02; 0.03)	0.03 (0.02; 0.04)	0.03 (0.02; 0.03)	0.04 (0.02; 0.05)	0.03 (0.03; 0.03)
P(G-|S+) = False negative rate	0.97 (0.97; 0.97)	0.96 (0.95; 0.96)	0.94 (0.93; 0.95)	0.94 (0.93; 0.95)	0.96 (0.96; 0.97)	0.95 (0.94; 0.95)	0.92 (0.91; 0.94)	0.92 (0.91; 0.94)
P(G-|S-) = True negative rate	0.98 (0.97; 0.98)	0.98 (0.97; 0.98)	0.97 (0.96; 0.97)	0.98 (0.97; 0.98)	0.97 (0.96; 0.98)	0.97 (0.97; 0.98)	0.96 (0.95; 0.98)	0.97 (0.97; 0.97)
odds(S+|G+)	6.43 (6.18; 6.69)	1.55 (1.36; 1.74)	0.84 (0.59; 1.08)	0.32 (0.11; 0.52)	5.92 (5.70; 6.13)	1.48 (1.33; 1.65)	0.81 (0.60; 1.03)	0.30 (0.11; 0.48)
odds(S+|G-)	4.44 (4.40; 4.48)	0.77 (0.73; 0.80)	0.42 (0.37; 0.48)	0.12 (0.07; 0.17)	4.44 (4.41; 4.48)	0.76 (0.73; 0.80)	0.42 (0.36; 0.47)	0.12 (0.07; 0.16)
P(S+|G+)/P(S+|G-)	1.06 (1.02; 1.1)	1.40 (1.29; 1.51)	1.54 (1.33; 1.75)	2.28 (1.91; 2.65)	1.05 (1.01; 1.08)	1.38 (1.28; 1.47)	1.52 (1.33; 1.70)	2.18 (1.86; 2.51)
P(S-|G-)/P(S-|G+)	1.37 (1.11; 1.74)	1.44 (1.29; 1.64)	1.29 (1.16; 1.45)	1.18 (1.12; 1.24)	1.27 (1.07; 1.56)	1.40 (1.28; 1.56)	1.28 (1.16; 1.41)	1.16 (1.11; 1.21)
Positive likelihood ratio	1.43 (1.11; 1.84)	1.98 (1.64; 2.39)	1.92 (1.52; 2.43)	2.58 (2.11; 3.14)	1.32 (1.06; 1.66)	1.88 (1.60; 2.22)	1.86 (1.51; 2.29)	2.42 (2.03; 2.88)
Negative likelihood ratio	0.99 (0.99; 1.00)	0.98 (0.97; 0.98)	0.97 (0.96; 0.98)	0.96 (0.95; 0.97)	0.99 (0.98; 0.99)	0.98 (0.97; 0.98)	0.96 (0.95; 0.96)	0.95 (0.94; 0.97)
Odds ratio	1.45 (1.12; 1.87)	2.02 (1.67; 2.45)	1.98 (1.55; 2.54)	2.68 (2.18; 3.30)	1.33 (1.07; 1.66)	1.93 (1.63; 2.29)	1.93 (1.54; 2.41)	2.54 (2.10; 3.10)

6 | The support of genetic evidence from drug target Mendelian Randomisation for approved drug targets

6.1. Abstract

This chapter describes the application of the drug target MR to a set of approved drug target gene - indication pairs to evaluate if the framework recapitulates the known mechanism of action in terms of effect direction and significance. The analysis focuses on approved drugs where protein quantitative trait locus (pQTL) data could be used to instrument the effect on the drug target. A 'truth' set of 160 licensed drug target gene-indication pairs with an available pQTL genome-wide association study (GWAS) as well as an available GWAS on the intended indication was defined. The pQTL GWAS data was based on the SomaLogic assay which utilises short single-stranded oligonucleotides ('SOMAmers') that bind with high affinity and specificity to a variety of proteins and enable the quantification of protein levels. A total of 320 drug target gene – SOMAmer – GWAS trait combinations was obtained after mapping SOMAmers binding drug target proteins to the encoding genes and to the GWAS trait corresponding to the approved indication. The application of the drug target MR approach consistently (p value ≤ 0.05 in over 50% of the models) rediscovered the mechanism of action of only a small proportion (16 out of 121 in the most conservative analysis) of the drug target gene – SOMAmer – trait combinations explored using the standard set of drug target MR parameters. While most of the pairings could not be evaluated due to lack of genetic associations with the exposure that meet the requirements to be in the genetic instrument, 11% (14 out of 121 in the most conservative analysis) were in the unanticipated direction of effect

based on the known drug mechanism. Such pairings are also discussed in the chapter to help inform future pQTL-weighted drug target MR. The findings suggest that a set of gold standard parameters for the optimal performance of drug target MR cannot be defined yet, and the selection of exposure data and MR parameters should be tailored to the drug target-indication of interest. Situations where there is a discordance between the drug target tissue, protein effect tissue and assay tissue could yield misleading results. Therefore, given the large proportion of results in the unanticipated direction of effect, expert knowledge is essential to interpret findings and minimise the risk of naïve interpretations, particularly when using the pQTL-weighted drug target MR approach for discovery of novel drug target mechanisms or prediction of adverse events.

6.2. Introduction

Genome-wide association studies (GWAS) can potentially inform drug target prioritisation through the identification of genetic variants in drug target genes that also impact disease risk. However, deciding whether to design an inhibitor or activator (blocker or agonist for receptor targets) of the potential drug target cannot be readily inferred simply from identification of the locus. The *cis*-Mendelian Randomisation (MR) approach ('drug target MR')[1] has been proposed to help infer the correct mechanism of action for a new drug. In an ideal scenario, a drug target MR analysis would assess the effect of modulating protein activity or function with respect to disease risk using genetic instruments in the encoding gene. This is because the vast majority of successful drugs achieve their effect by binding to and modifying the activity of a protein[2]. For example, small-molecule inhibitors inhibit catalytic sites of enzymes and antagonists bind in well-defined pockets that block receptor function, while agonists or activators, which are often more challenging to develop, increase enzyme activity or activate receptors. Therefore, the inference from a drug target MR analysis using such data would determine whether and by how much an increase or decrease in the protein function or activity impacts disease risk, suggesting a plausible mechanism of action for the drug.

However, GWAS on protein *activity* are limited, expensive and unscalable. Recently, GWAS of circulating protein concentration (pQTLs) have become available such as the INTERVAL study[3] (~3,000 proteins) and the SCALLOP Consortium[4] (~1,000 proteins). These data provide estimates for a substantial proportion of the human proteome, with the latest assays from SomaLogic covering ~7,000 proteins (SomaLogic v4.1 panel). If protein expression (pQTL) acts as a potential proxy for protein function or activity, then the new technologies for large-scale proteomics analysis could be used to inform drug target validation using drug target MR.

Even under the assumption that pQTLs are a valid proxy for protein function or activity, the performance of drug target MR is influenced by multiple parameters, some of them intrinsic to the MR methodology (e.g., modelling of the correlation between genetic instruments due to the linkage disequilibrium or the strength of the genetic associations used to instrument the exposure), and others specific to each drug target gene – outcome pairing, for example, expected protein abundance in plasma, number of protein-coding variants in the gene or protein subunits.

In this chapter, I evaluate the drug target MR framework using pQTL data by:

1. Generating pQTL data of 4,911 circulating protein levels in 2,253 participants from the UCLEB Consortium[5], to contribute to a meta-analysis through an established collaboration with Claudia Langenberg's group at the MRC Epidemiology Unit in Cambridge, who had access to a further 10,708 samples assayed on the SomaLogic 5k panel[6].

2. Comparing genetic associations with circulating levels (pQTL) and activity data for the same protein for two use cases to illustrate the potential of pQTL-weighted drug target MR approach when GWAS data on protein activity or function is not available.

3. Assessing the performance of the drug target MR framework as a GWAS-based approach to predict the effect of modulating a target in a particular disease using a 'truth' set of licensed drug target-indication pairs with available GWAS data on pQTL and the intended indication.

4. Lastly, for the drug target gene – indication pairings with results consistently in the unanticipated direction of effect to the expected, investigating potential reasons why

the drug target MR framework using pQTL data did not recapitulate the known mechanism of action.

6.3. Methods

6.3.1. pQTL data

To improve on the publicly available data (Aim 1) I performed a discovery GWAS of the SomaLogic v4 platform within a subset of the UCLEB Consortium (2,253 European samples) against the variants on the human DrugDev array[6]. This genotyping array combines a genome-wide variant backbone with enriched variant coverage in genes encoding druggable proteins. The design ensures capture of variation in the druggable genome therefore, it is an ideal platform to conduct association studies for drug target selection and validation. The SomaLogic assay utilises short single-stranded oligonucleotides ('SOMAmers') that bind with high affinity and specificity to a variety of proteins and enable the quantification of protein levels. The SomaLogic v4 platform included 5,284 SOMAmers. Following the company advice, 373 SOMAmers were excluded due to lack of specificity or incorrect mapping, leading to a reduced set of 4,911 SOMAmers. SomaLogic provided a mapping file between the SomaLogic sequence identifier (SOMAmer) and the target identifier using UniProt identifiers (ID). Such Uniprot IDs were mapped to the gene encoding the protein using Ensembl version 95 (GRCh37), which also returned gene identifiers. Of note, the same protein could be targeted by different SOMAmers because they target different isoforms of the same protein or because they bind to different epitopes, therefore, a 1:1 relationship between SOMAmer:Protein-Gene was not always observed in the dataset. Information on whether the measured proteins were located outside the cell membrane or were not secreted proteins (i.e., not present in secretion pathways or do not contain signal sequencies) was sourced, as the latter were not anticipated to be functionally circulating unless the carrying cell or a product breakdown could be found in blood.

The GWAS was performed on the rank inverse normalised residuals derived from the regression of the relative SOMAmers abundances on age, sex and the first ten principal components to account for inter and intra-sample variability. The genetic data excluded multiallelic variants, poorly imputed variants, and variants with minor allele count (MAC) less than 2. Then, a univariable linear regression model on all autosomes using an additive genetic model was performed using SNPTEST v2.5.4[7].

The GWAS was intended to contribute to a larger meta-analysis of several cohorts measured by SomaLogic v4 panel, including 10,708 samples who were participants in the Fenland study. However, due to delays in data analysis in other cohorts, the results from the meta-analysis were not available at the time of the drug target MR analysis (Aim 3). Instead, the genetic associations identified in the Fenland study were used as they were estimated in a larger sample size compared to the UCLEB study, and represents the largest pQTL discovery GWAS at the time of analysis. Access to such data was possible thanks to an established collaboration with Claudia Langenberg's group at the MRC Epidemiology Unit in Cambridge.

6.3.2. GWAS data on protein activity

To compare genetic associations with circulating protein levels (pQTL) to genetic associations with activity data and illustrate the potential of pQTL-weighted drug target MR approach when GWAS data on protein activity or function is not available (Aim 2), genetic associations with protein activity for two proteins were used. Genetic associations (p value $\leq 5\times10^{-8}$) with Butyrylcholinesterase (BCHE) were sourced from a published GWAS of 8,971 individuals[8]. Genetic associations with coagulation factor VII activity data (p value $\leq 5\times10^{-8}$)

were obtained from the UCLEB Consortium (8,700 participants). Variants in *cis* - were extracted based on the gene boundaries (GRCh37) plus a flanking region of 1 Mbp.

6.3.3. GWAS data on drug indication

To assess the performance of the drug target MR framework using a 'truth' set of licensed drug target-indications pairs (Aim 3), genetic associations with the intended indications were obtained from the public central repository (GWAS Catalog v1.0.2) and from UK Biobank through Neale data (GWAS Round 2, Results shared 1st August 2018). Genetic associations from UK Biobank were filtered for a p value $\leq 1\times10^{-5}$ to match the minimum significance threshold required by the GWAS Catalog[9]. The GWAS Catalog included 6,021 MeSH terms from 3,374 publications that were mapped to UMLS concepts. The UK Biobank Neale dataset covered 633 ICD10 main diagnosis that were mapped to 633 UMLS concepts. The list of traits was expanded and manually curated by a clinical expert to include GWAS data on clinically-relevant disease biomarkers based on quantitative traits measured in UK Biobank, UCLEB or available through the GWAS Catalog. A summary of the traits and data sources used as the outcome in the drug target MR analyses is shown in Appendix 6.A.

6.3.4. Drug data

To assess the performance of the drug target MR framework using a 'truth' set of licensed drug target-indications pairs, drug data were extracted from ChEMBL version 25 (v25)[10]. Information in ChEMBL is itself based on several resources including United States Adopted Name (USAN) applications, ClinicalTrials.gov; the FDA Orange Book database, the British National Formulary, and the ATC classification, with their intended indications sourced from

DailyMed and the ATC classification. The UniProt identifiers for the corresponding drug targets available through ChEMBL v25 were mapped to gene identifiers in Ensembl version 95 (GRCh37) through the updated druggable genome[6]. Of note, some drug target multiple single proteins, protein families or complexes, therefore a 1:1 relationship between a target:protein-gene is not always observed. The standardised indications in Medical Subject Headings (MeSH) used in ChEMBL v25 were mapped to Unified Medical Language System (UMLS) concepts to facilitate further mappings (see Chapter 3.1.1). Compounds flagged as withdrawn, not intended for human use or whose target is encoded by a gene in the extended major histocompatibility complex (xMHC) region (chr6: 28477797- 33448354, GRCh37), were excluded from the analysis. For each drug target gene – indication pairs, the latest phase in development was selected for any drug and those pairs with a maximum phase of development equal to 4 (licensed) were selected. This yielded in 665 unique drug target genes and 371 unique indications.

6.3.5. Drug target Mendelian Randomisation

Drug target MR analyses were performed using different parameter combinations for each SOMAmer - drug target gene - trait (768 tests/pair): *p* value threshold (1×10^{-8}, 1×10^{-6}, 1×10^{-4}, 1×10^{-2}), LD pruning (r^2: 0.2, 0.4, 0.6, 0.8), flanking region (bp: 2500, 10000, 50000, 100000), flanking region location (upstream, downstream or both) and minor allele frequencies (0.01, 0.05). To account for residual correlation between variants in the MR analyses, a generalised least squares framework with a LD reference dataset derived from UK Biobank was applied. LD reference matrices were created by extracting a random subset of 5,000 unrelated individuals of European ancestry from UK Biobank[11] using the same random seed of 1. Variants with a MAF < 0.001, and imputation quality < 0.3 were excluded. To ensure that

SNPs with lower MAF have higher confidence, variants were removed if MAF < 0.005 and genotype probability < 0.9; MAF < 0.01 and genotype probability < 0.8; MAF < 0.03 and genotype probability < 0.6. A model-selection framework was used to decide between competing inverse-variance weighted (IVW) fixed-effects, IVW random-effects, MR-Egger fixed effects or MR-Egger random-effects models[14]. While IVW models assume an absence of directional horizontal pleiotropy, Egger models allow for possible directional pleiotropy at the cost of power. After removing variants with large heterogeneity (p value < 0.001 for Cochran's Q test) or leverage, this model selection framework was re-applied and used the final model. Results were presented as mean difference (MD) or odds ratio (OR) with 95% confidence interval (95% CI).

6.4. Results

6.4.1. GWAS on plasma protein circulating levels (pQTL)

To improve on the publicly available data, a discovery GWAS of the SomaLogic v4 platform (4,911 SOMAmers, 4,631 UniProt identifiers) was performed within a subset of the UCLEB Consortium (2,253 participants) against the contents of the human DrugDev array[6]. As a quality control step, for those SOMAmers also included in the INTERVAL study, the Pearson's correlation between the effect sizes was calculated for those variants reported as 'sentinel variants' (variants with the lowest *p* value in the region) by Sun *et al.,*[3]. A total of 1,128 SOMAmers were included in the comparison exercise (382 with genetic associations in *cis*-, 841 with genetic associations in *trans*-). A strong correlation (ρ) was observed between the effect sizes (Fig. 6.1), which was higher for *cis*- signals (ρ = 0.96), compared to *trans*-signals (ρ = 0.75).

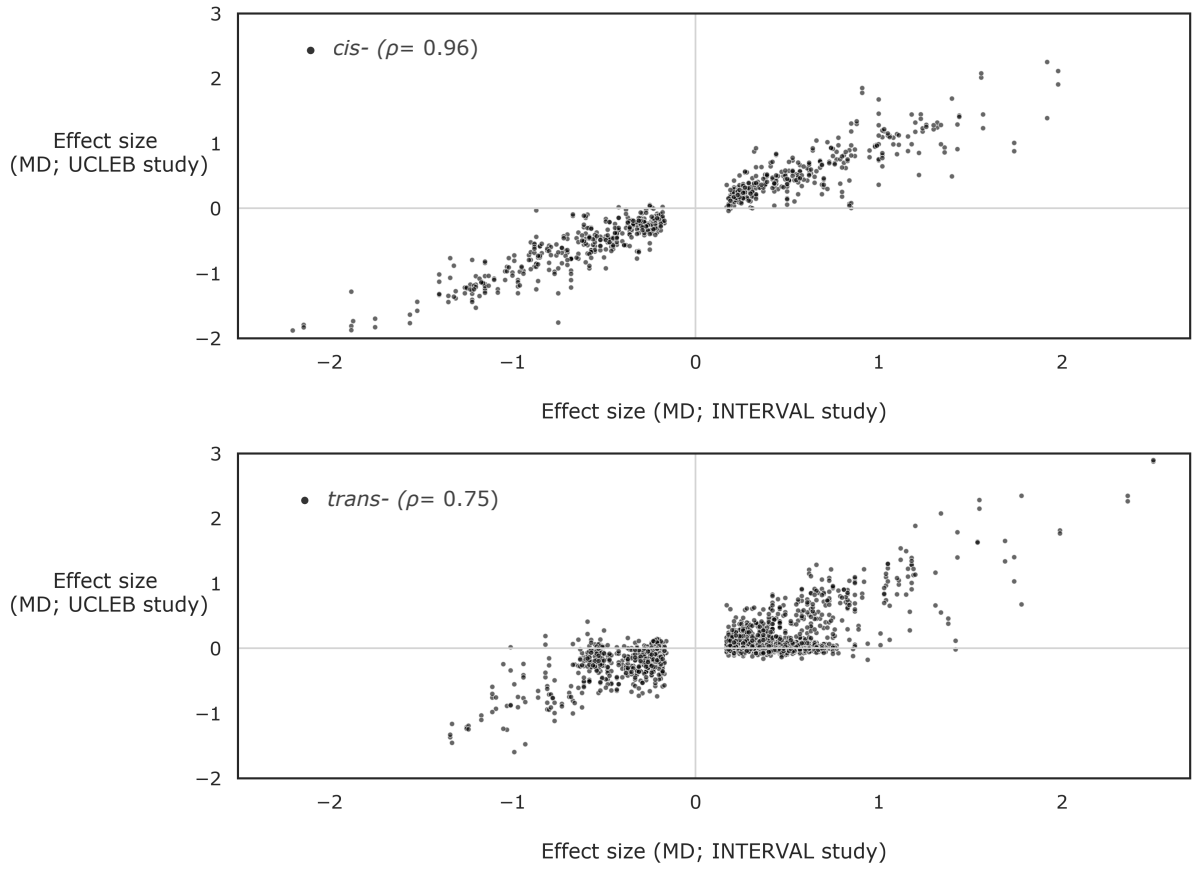


Figure 6.1. Scatterplot of pQTL effect size estimates from INTERVAL versus UCLEB subset, showing genetic associations in *cis*- (top panel) and *trans*- (bottom panel). ρ is Pearson's correlation coefficient. The x-axis shows the per allele effect on pQTL expressed as mean difference (MD).

6.4.2. Correlation between protein activity and circulating protein levels

In drug target MR, genetic associations with the activity or function of the protein target represent the ideal exposure to instrument the therapeutic effect of modulating such target in a particular disease since most drugs impact protein activity or function. However, GWAS on protein activity or function are only available to very few proteins, and thus genetic associations with protein levels (known as *pQTL*) have been proposed instead to evaluate drug targets, under the assumption that the protein levels are a proxy of activity or function. Several examples

support this, such as the drug target Mendelian randomisation of CETP or PCSK9 protein concentration which replicated on-target effects previously reported in clinical trials[1,12]. To evaluate this, a comparison was performed using two proteins where genetic associations were available for both activity and protein level. The analysis was restricted to genetic associations in *cis*- since the aim of this comparison was to evaluate the utility of genetic variants associated with protein levels in drug target MR analyses, which utilises genetic instruments in-and-around the gene encoding the protein of interest (see Chapter 2.4.1. for details). Genetic associations for the butyrylcholinesterase (BCHE) and coagulation factor VII *protein levels* were measured by the SomaLogic v4 platform in the Fenland cohort (10,708 participants). Genetic associations with *protein activity* for BCHE were sourced from a published GWAS of 8,971 individuals[8], and for the coagulation factor VII from the UCLEB Consortium (8,700 participants).

For both BCHE and coagulation factor VII there was a strong correlation between genetic associations with activity and level for variants acting in *cis*-. The Pearson's correlation coefficient for the BCHE using genetic variants in *cis*- (n variants = 373) was $\rho = 0.99$ (Fig. 6.2A). The correlation for, coagulation factor VII, was slightly lower ($\rho = 0.96$, n *cis*- variants = 56), as shown in Figure 6.2B.

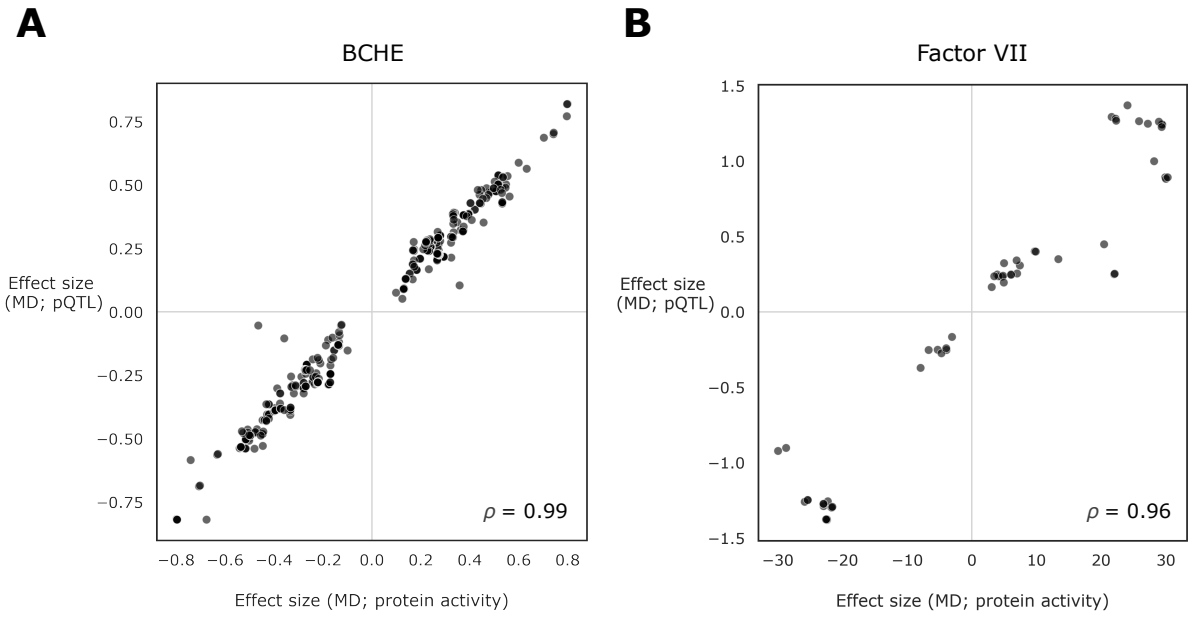


Figure 6.2. Scatterplot of effect size estimates of genetic associations with protein activity versus genetic associations with protein levels for butyrylcholinesterase BCHE (A) and coagulation factor VII (B). ρ is Pearson's correlation coefficient. The x-axis shows the per allele effect on protein activity and the y-axis the per allele effect on pQTL expressed as mean difference (MD).

6.4.3. Drug target MR rediscoveries of approved mechanism of actions

In the previous section it has been described that, in two examples where comparisons were possible, genetic associations with protein level and activity were highly correlated for variants acting in *cis*-. Although only two proteins were investigated due to the lack of available data, such observations provided further evidence to that already in the literature[12–15] that supports the use of genetic associations in or near a gene encoding a drug target protein that alter the protein's expression as a tool to anticipate the phenotypic effect of drug action on the same target. Based on the literature and as described in the previous section, variation in circulating plasma protein concentration (pQTL) was used as a proxy for protein activity to instrument the effect of perturbing a particular drug target. The aim was to investigate if the drug target MR framework rediscovered the mechanism of action of licensed drugs where

pQTL associations were available for the target protein and the intended indication or a clinically-relevant disease biomarker had been studied in GWAS.

Of the 665 genes that encode a human proteins targeted via a licensed drug (Fig. 6.3), 205 had pQTL data available for the encoded protein. The SomaLogic assay utilises short single-stranded oligonucleotides ('SOMAmers') that bind with high affinity and specificity to a variety of proteins and enable the quantification of protein levels. The SOMAmer were mapped to target identifiers by the company using UniProt identifiers (ID). Such Uniprot IDs were mapped to the gene encoding the protein using Ensembl version 95 (GRCh37). Of note, the same protein could be targeted by different SOMAmers because they target different isoforms of the same protein or because they bind to different epitopes, therefore, a 1:1 relationship between SOMAmer:Protein-Gene was not always observed in the dataset.

After curating the overlap between drug indications and genetic studies with available summary statistics in the GWAS Catalog, the final dataset comprised 320 SOMAmer-drug target gene-trait pairs (188 SOMAmer-drug target gene-indication pairs) for 48 indications, 71 drug target genes and 112 drugs (Fig. 6.3). Data was aggregated to account for targets measured by multiple SOMAmers. This was done because the protein encoded by a drug target gene could be measured by more than one SOMAmer through different aptamers (i.e., binding to different domains). By aggregating the data at the SOMAmer-drug target gene-trait, protein differences in binding across SOMAmers could be taken into account in the analysis. Of the 71 target genes, 24 (34%) encoded proteins located outside the cell membrane, while 47 (66%) encoded proteins not secreted (i.e. are not present in secretion pathways or do not contain signal sequencies, and thus not anticipated to be functionally circulating unless the carrying cell or a product breakdown could be found in blood).

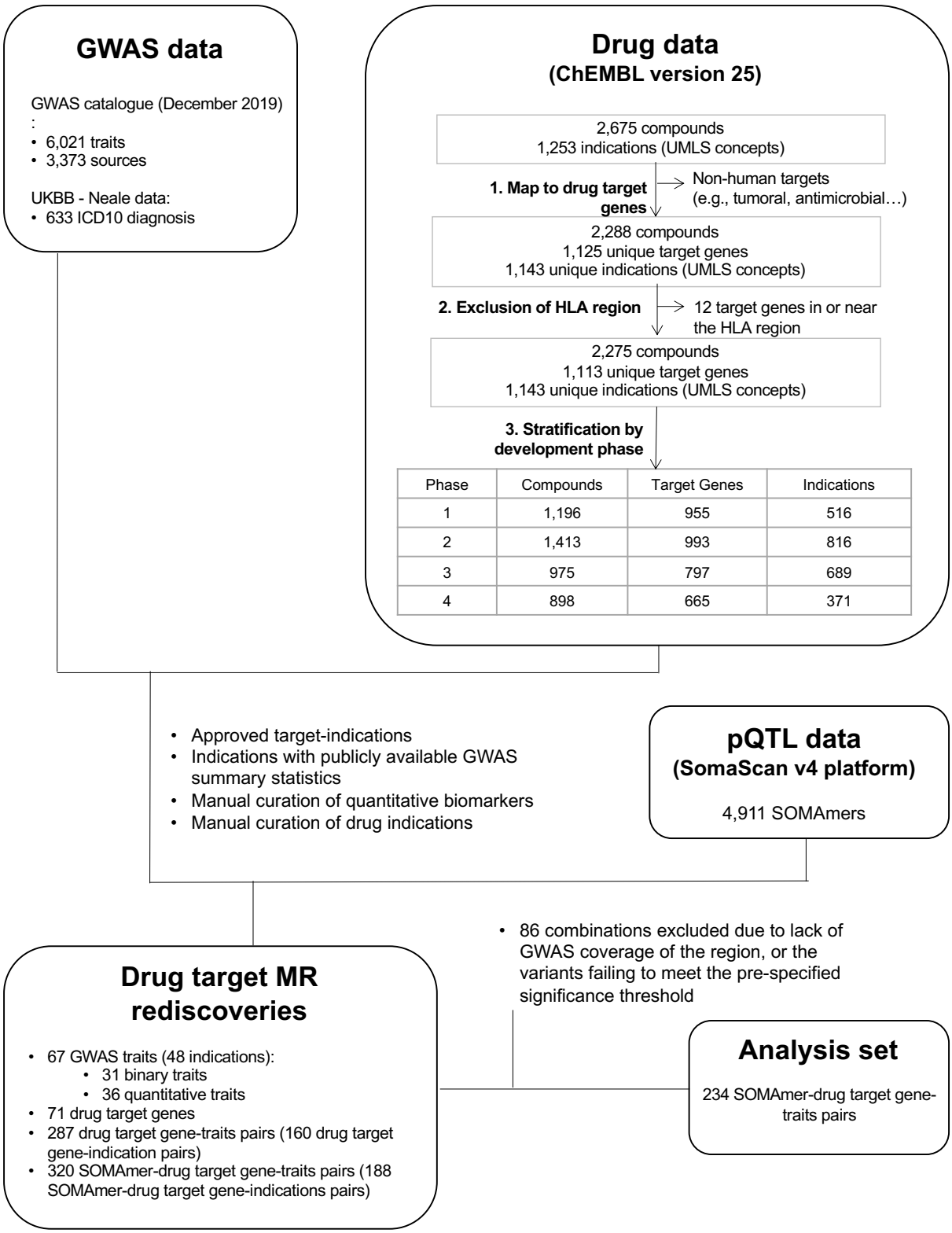


Figure 6.3. Summary of data sources and mappings between them. Summary of each data source and the key filtering and processing steps applied to create the final set of gene-trait and drug target–indication combinations investigated in this study. GWAS Catalog sources correspond to unique PubMed ID.

Drug target MR analyses were performed using different parameter combinations for each SOMAmer-drug target gene-trait (768 tests/pair): *p* value threshold for inclusion of variants (1×10^{-8}, 1×10^{-6}, 1×10^{-4}, 1×10^{-2}), LD clumping threshold r^2 (0.2, 0.4, 0.6, 0.8), flanking region size (bp: 2500, 10000, 50000, 100000), flanking region location (upstream, downstream or both) minor allele frequencies (0.01, 0.05), and automatic removal of potential pleiotropic variants based on leverage and Q-statistics (see Methods). Eighty-six out of the 320 SOMAmer-drug target gene-trait combinations could not be analysed in any of the models due to lack of GWAS coverage of the region, or the variants failing to meet the pre-specified significance threshold. The results of the drug target MR analyses are presented based on four different scenarios: i) all the parameter combinations are taken into account even if a particular parameter combination could not run (*main analysis*), ii) only parameter combinations that yield results are considered (*sensitivity analysis 1*), iii) only credible parameter combinations (p value $\leq 1\times10^{-4}$; $r^2 \leq 0.4$; flanking region $\leq$ 50 kbp both upstream and downstream, with automatic outlier removal) that increase the accuracy of the results while holding the MR assumptions based on previous studies[1,16,17] are taken into account (*sensitivity analysis 2*), iv) only credible parameter combinations that yield results are considered (*sensitivity analysis 3*).

The percentage of non-significant estimates and significant estimates in the expected or unexpected direction of effect (depending on the drug target mechanism) for each of the SOMAmer-drug target genes-trait pairs for the sensitivity analysis 3 is shown in Figures 6.4 and 6.5 for binary and quantitative traits respectively, where binary traits represent the intended indication and qualitative traits a clinically relevant biomarker of the disease. The percentages for the main, sensitivity analysis 1 and 2 are shown with similar figures in Appendix 6.B. In addition, Appendix 6.C includes a qualitative evaluation of three cases where both expected and unexpected results were observed in the sensitivity analysis 3.

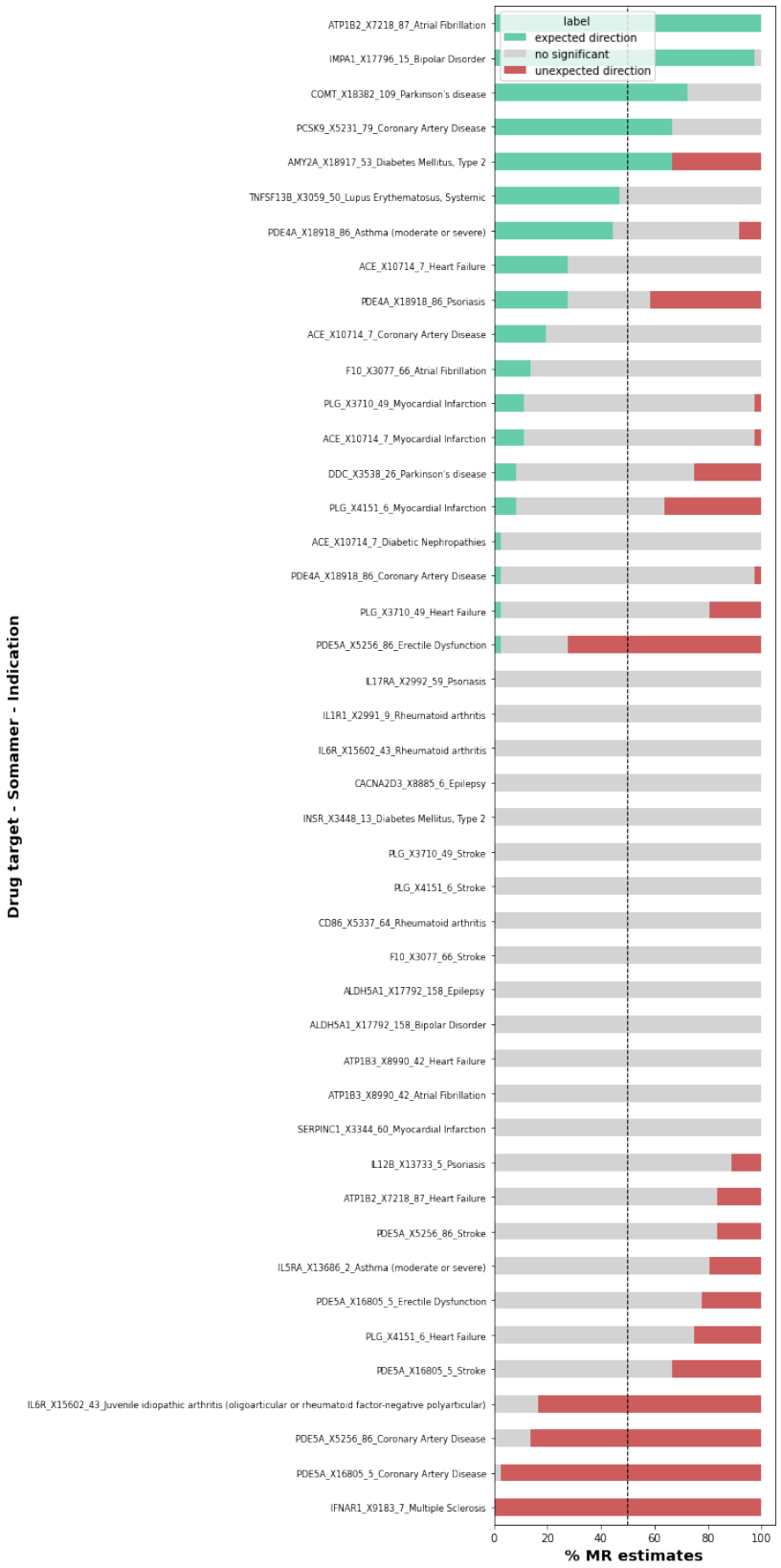


Figure 6.4. Results of the sensitivity analysis 3 (only credible parameter combinations that yielded results). Percentage of significant MR estimates in the expected (green), unexpected (red) direction of effect or non-significant estimates (grey) for SOMAmer-drug target gene-trait pairs where binary traits were used as the outcome.

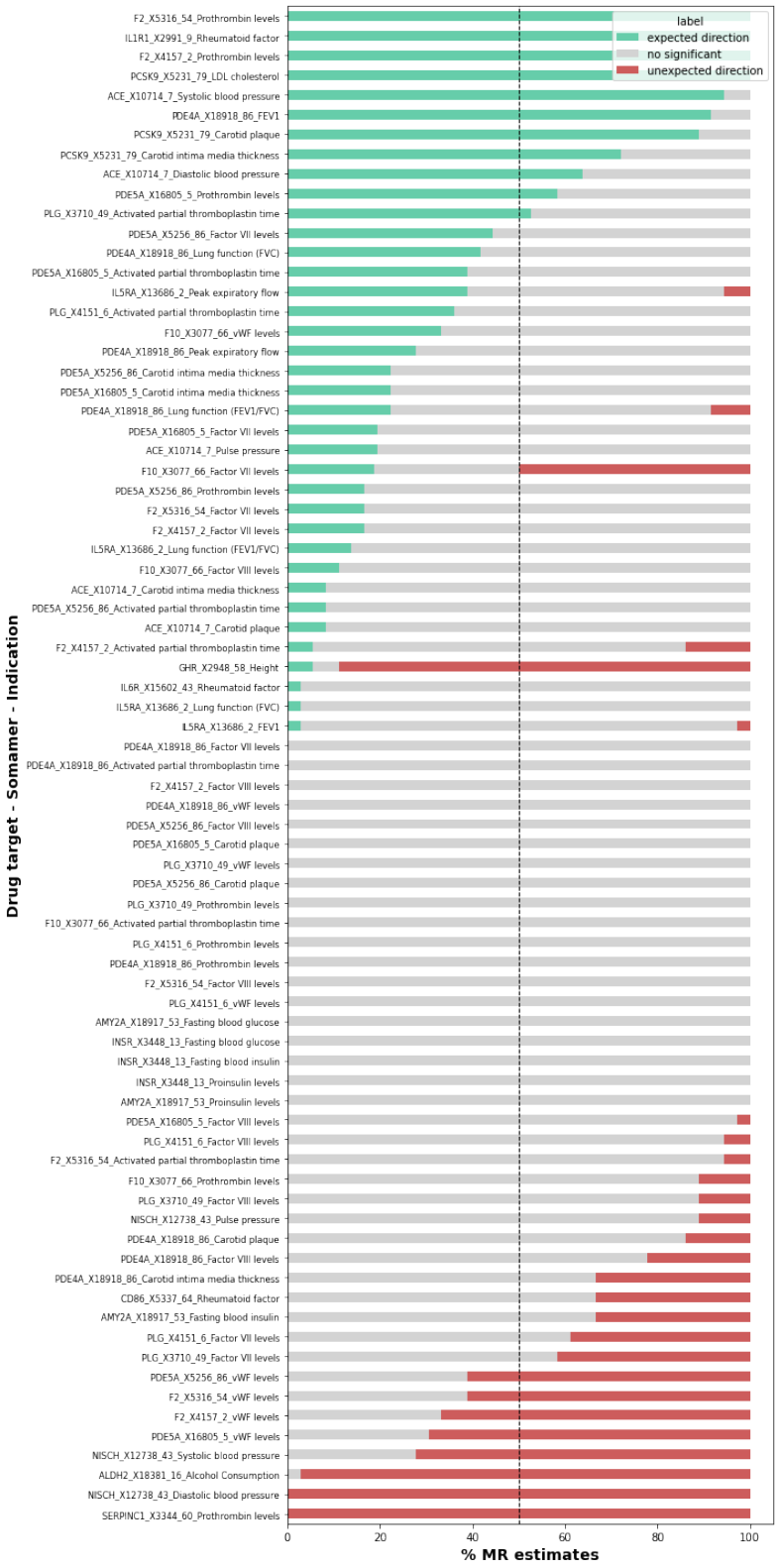


Figure 6.5. Results of the sensitivity analysis 3 (only credible parameter combinations that yielded results). Percentage of significant MR estimates in the expected (green), unexpected (red) direction of effect or non-significant estimates (grey) for SOMAmer-drug target gene-trait pairs where quantitative traits were used as the outcome.

The number of SOMAmer-drug target gene-trait pairs consistently (p value ≤ 0.05 in over 50% of the models) in the expected or unexpected of direction of effect under the different scenarios are shown in Table 6.1. The SOMAmer-drug target gene-trait pairs with more significant results in the anticipated direction of effect vs unanticipated were 82 in the main analysis and in the sensitivity analysis 1, 49 in the sensitivity analysis 2 and 46 in the sensitivity analysis 3, however the number of significant results in the anticipated direction of effect did not reach the 50%. For example, for the 768 tests performed for the drug target IL6R and the GWAS trait rheumatoid factor (biomarker for the intended indication rheumatoid arthritis), 255 were significant in the anticipated direction of effect and 513 were not significant, although 467 of the 513 were in the anticipated direction of effect. Seventy seven SOMAmer-drug target gene-trait pairs were more times in the unexpected direction of effect vs expected in the main analysis and sensitivity analysis 1, which decreased to 49 in the sensitivity analysis 2 and 40 in the sensitivity analysis 3.

Table 6.1. SOMAmer-drug target gene-trait pairs consistently in the expected or unexpected direction of effect under different parameter combinations.

SOMAmer-drug target gene-trait pairs	**Main analysis**	**Sensitivity Analysis 1**	**Sensitivity Analysis 2**	**Sensitivity Analysis 3**
Consistently in the expected direction of effect (>50%)	19	27	15	16
Consistently in the unexpected direction of effect (>50%)	12	26	9	14
Total	234	234	234	121

Main analysis: all the parameter combinations are taken into account even if a particular parameter combination could not run; Sensitivity analysis 1: only parameter combinations that yield results are considered; Sensitivity analysis 2: only credible parameter combinations (p value ≤ 1×10^{-4}; r^2 ≤0.4; flanking region ≤50kbp both upstream and downstream, with automatic outlier removal) that increase the accuracy of the results while holding the MR assumptions based on previous studies[1,16,17] *are taken into account; Sensitivity analysis 3: only credible parameter combinations that yield results are considered.*

Independent pQTL data was sourced from Ahola-Olli *et al.* 2017[18], Yao *et al.* 2018[19], and Folkersen *et al.* 2020[20] to replicate the findings based on SomaLogic v4 platform. Out of the 71 drug target proteins, only IL6R, DPP4 and TNF were available in the replication datasets. However, only genetic associations with TNF could be analysed (Appendix 6.C) due to the lack of GWAS coverage of the region or the variants failing to meet the pre-specified significance threshold for IL6R and DPP4. The discovery analysis of TNF and its intended indications did not return significant results using the SomaLogic v4 platform, and thus, the results across platforms could not be compared.

6.4.4. Case review of drug target gene-indications pairs in the unexpected direction of effect

The results presented in the previous section included several SOMAmer-drug target gene-trait pairs that were in the unexpected direction of effect in >50% of the scenarios explored. In an attempt to better understand the reasons and the potential limitations and inform future drug target MR analyses with pQTL data as the exposure, this section provides a review of drug target gene-indications pairs consistently in the unexpected direction of effect. Many reasons exist that may explain why a drug target MR analysis does not recapitulate the mechanism of action of a known drug target in a particular disease indication, and, instead, returns results in the opposite direction to the known drug targeting mechanism. These include technical errors, inaccuracies in disease definitions or biological mechanisms. In the previous section, between 9 and 26 SOMAmer-drug target gene-trait pair had MR associations in the opposite direction to that expected in 50% or more of the analyses, where 14 were consistently in the unexpected direction of effect in the most stringent scenario (sensitivity analysis 3: only credible parameter combinations that yield results are considered). After ensuring that these

findings were not due to a technical error such as using the wrong effect allele in the MR analysis, each drug target – indication pair was reviewed and compared with observations across GWAS traits to determine plausible explanation for the unexpected results. Based on the review, three distinct groups could be identified (groups 1, 2, 3) which are presented below, and the remaining pairs that did not fall under such categories are discussed in Appendix 6.C. The groups were defined as follows: group 1 includes examples of pairings where the effect in the target tissue may not be captured by plasma levels of the circulatory protein, group 2 includes examples of drugs targeting a protein family rather than a single protein, and group 3 includes examples of drug target proteins with both secreted or membrane bond forms for which the measured circulating protein might not reflect the level of the membrane bound form due to extracellular components.

The effect in the target tissue is not captured by plasma levels

In addition to the assumption described in Chapter 6.4.2. regarding the use of pQTL as a proxy for protein activity, the analysis presented in this chapter also assumed that the protein levels observed in plasma are representative of the protein levels in the effector tissue (i.e., where the drug exerts its action). This assumption is robust when the protein of interest is secreted and exerts its action within the circulation or at the cell surface. However, many proteins present in the circulation and measured using the SomaLogic platform are likely present due to the cell turnover or damage and thus may not be representative of their status in the effector tissue. The following example illustrates this possibility.

Moxonidine, an agonist of nischarin (NISCH; Imidazoline-1 receptor), is used to treat hypertension. Diastolic, systolic and pulse blood pressure were used as the outcomes in the drug target MR analysis as clinically relevant biomarkers. If the drug target MR framework recapitulated the agonist mechanism of action, one would have anticipated a negative

association between levels of NISCH and blood pressure. However, the MR estimates for diastolic and systolic blood pressure were consistently opposite direction to the anticipated direction of effect, from the known blood pressure lowering effect of monoxidine. One potential explanation could be that the effect allele was not correctly defined in the blood pressure GWAS, however this option was discarded as the drug target MR of another antihypertensive target, ACE, rediscovered consistently the mechanism of action of the drug for both systolic and diastolic blood pressure. Then, under the assumption that there are no technical errors in the SOMAmer measurements, the next potential explanation of the unexpected results could be that, since moxonidine acts primarily as an antihypertensive drug in the central nervous system[21], circulating NISCH may not adequately reflect the drug effect in the intended tissue.

Drugs targeting a protein family

While most drugs target single proteins, some perturb multiple proteins of a complex, or the target is indicated as a protein complex or family because the actual effector protein in the family is not known. This section describes two examples of drug target gene-indications pairs involving a protein family or complex where the drug target MR framework did not estimate the direction anticipated based on the mechanism of action of the drug.

Peginterferon beta-1a is a drug used to treat multiple sclerosis by activating the type I interferon receptor (IFNAR), composed of IFNAR1 and IFNAR2. The ligand type 1 interferon is thought to bind first to the high-affinity IFNAR2 subunit, and the ligand binding to the low-affinity IFNAR1 subunit induces signal transduction[22]. In this chapter, the drug target MR approach was applied only to IFNAR1 since the other component of the receptor IFNAR2 was not measured in the SomaLogic v4 panel. Since the drug activates the IFNAR receptor, in a drug target MR analysis one would expect a negative association between circulating levels of

IFNAR1 and disease status, where lower levels of the protein are associated with higher disease risk. However, the results of the analysis presented in this chapter were consistently in the unexpected direction of effect (effect size > 0). This may indicate that the mechanism of action of Peginterferon beta-1a acts primarily through IFNAR2 which binds with high affinity, and that levels of IFNAR1 are increased to compensate the low quantities or activity of IFNAR2 in the membrane. This hypothesis could be tested when pQTL data on IFNAR2 become available.

Dipyridamole and pentoxifylline are non-selective inhibitors of the phosphodiesterase protein family (PDEs) used to prevent and treat thrombosis and coronary artery disease. The phosphodiesterase protein family includes 21 members[23] of which PDE1A, PDE2A, PDE4A, PDE4D, PDE5A, PDE7A and PDE9 circulating levels were measured by the SomaLogic v4 platform. In addition to the non-specific drugs, avanafil and tadalafil are inhibitors of the PDE5 protein in particular and are used to treat erectile dysfunction for their vasodilator effect. The example illustrated in this paragraph is centred on PDE5A, which was measured by two independent SOMAmers (X5256_86 and X16805_5). A discordant direction of effect was observed consistently (p value ≤ 0.05 in over 50% of the models) in three independent traits (coronary artery disease, erectile dysfunction, vWF levels) out of the ten traits evaluated (coronary artery disease, stroke, vWF level, factor VII levels, factor VIII levels, prothrombin levels, carotid plaque, carotid intima media thickness, activated partial thromboplastin time, erectile dysfunction). Further, for coronary artery disease and vWF levels both PDE5A SOMAmers yielded the same conclusion (Fig. 6.6) which suggests that the mechanism underlying this unexpected behaviour may be related to the drug target itself or the pQTL rather than due to technical oversights or errors in the outcome source data. One hypothesis may be that the therapeutic effect of PDE5A inhibitors occurs at a tissue level and not in circulating plasma, where the target is found as it is not secreted to the circulatory system (likely scenario in erectile dysfunction in particular). Regarding thrombosis and coronary artery disease where

the inhibitors are non-selective, an alternative hypothesis could be that PDE5A is not one of the effective target of PDE inhibitors, and that the therapeutic action is driven by another member in the protein family. PDE1A, PDE2A, PDE4A, PDE4D, PDE7A and PDE9 were also studied in the context of coronary artery disease and thrombosis. However, the analysis of PDE1A, PDE2A, PDE4D, PDE7A and PDE9A could not be conducted due to the lack of strong genetic associations with the pQTL levels, while the sensitivity analysis 3 of PDE4A returned more significant results in the unexpected direction of effect than in the expected for traits related to thrombosis and cardiovascular disease (439 non-significant tests, 62 in the unexpected direction of effect and 3 in the expected direction of effect), although not robustly (i.e., p value > 0.05 in more than 50% of the models).

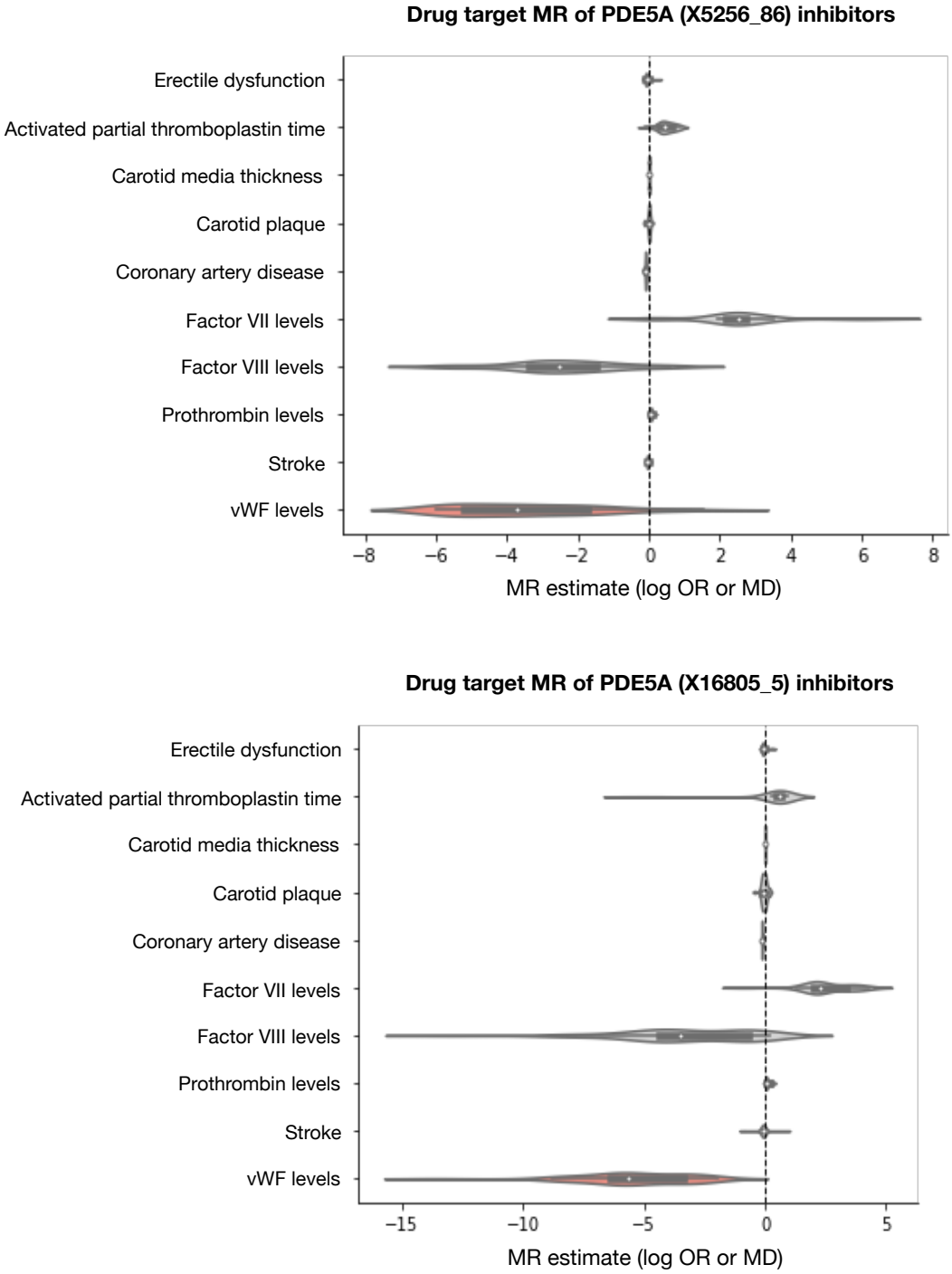


Figure 6.6. Distribution of MR estimates for PDE5A (exposure) by SOMAmer and outcome investigated. The intended indications (erectile dysfunction, thrombosis, and stroke) were instrumented using genetic associations with the binary trait as well as a clinically relevant biomarker (carotid media thickness and carotid plaque for coronary artery disease, and activated partial thromboplastin time, factor VII, factor VIII, prothrombin and vWF levels for stroke). Results from the drug target MR analysis are presented as mean difference (MD) for quantitative outcomes or log odds ratio (OR) for binary outcomes.

Intracellular vs extracellular proteins

Proteins with both secreted or membrane bond forms pose an additional challenge since the level of the measured circulating protein might not reflect the level of the membrane bound form. In fact, higher abundancies in plasma might indicate lower abundance in the plasma membrane. If the membrane bound form is critical for the biological function, this might explain the opposite direction of effect as illustrated in the following examples.

Tocilizumab is an anti-IL6-receptor antibody used to treat rheumatoid arthritis. IL-6 is highly expressed in patients with rheumatoid arthritis and plays a critical role in perpetuating inflammation. Its receptor (IL-6R) can either be membrane bound, which promotes the 'classical' IL6 signalling, or in soluble form in plasma (sIL-6R) which increases the circulating half-life of IL-6 and therefore, negatively regulates classical IL-6 signalling[24]. To treat rheumatoid arthritis, tocilizumab blocks the 'classical' IL6 signalling, which leads to an increase of circulating IL6 and the soluble form of the receptor (sIL6R). In this chapter, a drug target MR analysis was performed using as the exposure genetic associations with the IL6R and juvenile idiopathic arthritis (oligoarticular or rheumatoid factor-negative polyarticular) as the outcome. Based on the mechanism of action of the drug (*blocker*) one would expect a positive relationship between IL6R levels and the disease outcome, however, the results were consistently in the unanticipated direction of effect (effect size < 0 and p value ≤ 0.05). One hypothesis could be that the pQTL measured by the SomaLogic platform corresponded to sIL-6R, whose circulating levels show an inverse relationship with the membrane bound form. In fact, such hypothesis would be supported by evidence from a previously published *cis*-MR analysis, which also showed an inverse association between sIL-6R and rheumatoid arthritis[25].

Somatropin is a growth hormone (GH) replacement therapy used to treat growth hormone deficiency, also known as pituitary dwarfism. Genetic associations with the receptor of the growth hormone (GHR) was used to instrument the exposure and height was used as a biomarker of the intended indication. Therefore, assuming that the levels of the GHR correlate positively with the levels of the GH, a negative association was anticipated for the drug-target MR analysis. However, the estimates were consistently in the unexpected direction of effect. One explanation could be that the lack of growth hormone in such disease may lead to a negative feedback loop where GHR production is increased to compensate for the low GH levels. If high GHR levels are observed in pituitary dwarfism, then the relationship would not be negative, but instead positive. An alternative hypothesis may involve the growth hormone binding protein (GHBP), which in humans is derived from the cleavage of the extracellular domain of the GHR[26]. Soluble GHBP might compete for growth hormone binding and thus is a negative regulator of growth hormone signalling. If the SOMAmer measured by the proteomics platform corresponded to soluble GHBP rather than membrane GHR, a positive association would be expected since high levels of GHBP could indicate low levels of GHR.

6.5. Discussion

6.5.1. Summary

This chapter described the results of a discovery GWAS of 4,631 circulating protein levels measured by the SomaLogic v4 platform for 2,253 individuals of the UCLEB Consortium. The quality of the GWAS was assessed by comparing the effect sizes for a set of variants previously reported by Sun *et al*[3], which showed a strong correlation measured by the Pearson's correlation coefficient (ρ) for both *cis*- (ρ = 0.96) and *trans*- signals (ρ = 0.75). Subsequently, the work in this chapter evaluated the correlation between protein activity and circulating protein levels. This analysis was performed to evaluate the common assumption of pQTL-weighted drug target MR which states that protein levels are a proxy of activity or function and therefore can be used to instrument the therapeutic effect of modulating a drug target in a particular disease. At the time of the analysis, genetic associations with protein activity were only available for the butyrylcholinesterase (BCHE) and coagulation factor VII, whose protein levels had been also measured by the SomaLogic v4 platform in the Fenland cohort (10,708 participants). A strong correlation measured by the Pearson's correlation coefficient was observed for variants in *cis*- for both BCHE (ρ = 0.99) and coagulation factor VII, was slightly lower (ρ = 0.96).

After illustrating the potential of genetic associations with pQTL as a proxy of protein activity for two case studies, a 'truth' set of 160 licensed drug target gene-indication pairs with available GWAS data on pQTL and the intended indication was analysed to investigate if the pQTL-weighted drug target MR framework recapitulated the known mechanism of action in terms of effect direction and significance.

It was found that out of the 665 genes that encode the protein targeted by an approved drug, 71 had available pQTL data (measured by high-affinity and specific oligonucleotides called 'SOMAmers') and GWAS data on the intended indication, which allowed for the evaluation of the drug target MR methodology on a set of 160 drug target gene - indication pairs. Such pairs were mapped to GWAS phenotypes related to the intended indication and to the measured protein target through the SOMAmers, which returned a total of 320 SOMAmer-drug target gene-traits pairs. The application of the drug target MR framework recapitulated the mechanism of action of several drug target gene – indication pairings (range: 15-27 drug target gene-SOMAmer-trait) under different models, which ranged from all possible parameter combinations to those combinations with a credible set of parameters in terms of strength of the association with the exposure, degree of linkage disequilibrium and the extent of the flanking region around the target gene. The set of validated drug targets that consistently showed the expected direction of effect in the drug target MR approach, could be explored against other outcomes beyond the intended indication to identify opportunities for indication expansion and validate (or anticipate) on-target adverse effects through a drug target MR - phenome-wide association study (PheWAS).

On the other hand, it was found that between 38- 50% of the drug target gene-SOMAmer-trait combinations analysed that returned significant MR estimates were consistently in the unexpected direction of effect based on their reported mechanism of action (range 9/24-26/53 drug target gene-SOMAmer-trait). Although potential explanations for this were already discussed in section 6.4.4., this analysis relied on the accuracy of the proteomic platform and the summary statistics of the intended indication. This potential source of bias together with biologically plausible mechanisms may explain some of the unanticipated findings, however, further research is needed to validate the results using additional data sources. Nevertheless,

these findings suggest that results from drug target MR should be interpreted cautiously and informed by biological knowledge.

Noticeably, the drug target MR analysis of multiple drug target gene - SOMAmer - trait combinations did not return significant results in the main or sensitivity analyses, and even in some cases, the analyses could not be performed due to the lack of instruments. While there may be many reasons for this, the low affinity of the SOMAmer with the target protein and the lack of power in the indication GWAS may explain a fraction of the non-conclusive results.

6.5.2. Research in context

Drug target MR analyses that utilise genetic associations with circulating protein levels to study the effects of perturbing drug targets assume that protein levels are a proxy of protein activity or function. To formally evaluate this, a comparison was performed using two proteins where genetic associations were available for both activity and protein level. Due to the lack of available GWAS data on protein activity, a more extensive evaluation including more proteins could not be conducted. However, the correlation observed for BCHE and coagulation factor VII, together with previous studies of pQTL-weighted drug target MR[1,12], suggested that drug target MR using pQTL could be a valid alternative approach when GWAS data on activity or function is not available.

Under such assumption, the analysis presented in this chapter evaluated for the first time at the time of analysis the performance of the drug target MR framework using a 'truth' set of drug target gene-indication pairings, where circulating levels of the target protein have been measured by a high-throughput proteomic platform and the indication has been studied by GWAS.

Previously, a Mendelian randomisation study on 1,002 proteins and 225 phenotypes identified four drug target gene – approved indication pairs for which the MR recapitulated the mechanism of action and two drug target gene – approved indication pairs for which the MR approach returned results in the unexpected direction of effect out of 73 pairs with potential to be rediscovered[27]. The analysis described in this chapter used a larger set of pQTL data which allowed for the evaluation of more drug target gene – approved indication pairs. Such increase in the sample size showed an increase in the number of pairs 'rediscovered' by the drug target MR framework which ranged between 11-13% (i.e., 27/234 in the sensitivity analysis 1 and 16/121 in the sensitivity analysis 3) in the current analysis compared to the 5% 'rediscovered' by Zheng *et al*.,[27]. The target gene – indication pairs in the expected direction of effect identified by Zheng *et al*.,[27], included the PCSK9 for hypercholesterolemia and hyperlipidaemia, ACE for hypertension, IL12B for psoriatic arthritis and psoriasis, and TNFRSF11A for osteoporosis. In the analysis presented in this chapter, PCSK9 and ACE consistently showed a concordant and significant direction of effect under all the models explored, while TNFRSF11A showed a concordant direction of effect when using heel bone mineral density as the outcome, however the association was not statistically significant. In this chapter, the association between IL12B and psoriasis was in the unexpected direction of effect in some of the scenarios, although most of the combinations analysed did not yield significant results. Out of the two drug target gene - indication pairs found by Zheng *et al*.,[27], in the unexpected direction of effect, IL6R was also identified in the current analysis while PROC was not analysed as it is not recorded as the target of an approved drug in ChEMBL v25. In their work, Zheng *et al*.,[27], in line with the observations outlined in the case study section above, indicated that for IL6R the alleles associated with higher soluble protein levels have been shown to also lead to lower intracellular pathway activation[28], suggesting consistency of direction with the therapeutic approach.

In addition to PCKS9 and ACE, the analysis in this chapter further identified nine targets that consistently showed a concordant direction of effect under all the models: AMY2A and type 2 diabetes Mellitus; ATP1B2 and atrial fibrillation; COMT and Parkinson's disease; F2 and prothrombin levels; IL1R1 and rheumatoid factor; IMPA1 and bipolar disorder; PDE4A and forced expiratory volume in the first second (FEV1); PDE5A and prothrombin levels; PLG and activated partial thromboplastin time. The findings for IL1R1 and PLG are in line with a previous study which presented the drug target MR framework using a set of selected positive controls, which also included PCSK9[1]. In addition to the findings from Zheng *et al.*[27], described in the previous paragraph, genetic associations with ACE pQTL data have been used to instrument the effect of modifying ACE circulating levels on different outcomes[29], including susceptibility to SARS-CoV-2 infection or COVID-19 severity[30], and drug target MR analyses on the intended indication have been conducted using expression QTL (eQTL)[31]. The genetic validation performed in this chapter of ACE as a drug target for hypertension provides supportive evidence of the validity of pQTL data to instrument the effect of the drug in past and future drug target MR studies. For approved drug targets, such as ACE, such genetic validation on the intended indication should be always conducted, where possible, before exploring new outcomes. Similarly, phosphodiesterases have been previously studied using both eQTL and pQTL data on different outcomes[32], however, so far, these have not included the intended indication. Previous mendelian randomisation studies on coagulation factors and the intended indication (venous thrombosis) have been published using intermediate traits such as activated thromboplastin time as the exposure[33], however, drug target MR analyses using F2 pQTL data have not previously been reported. Moreover, to my knowledge, AMY2A, COMT, ATP1B2 have not been previously studied in drug target MR analyses of the intended indication using pQTL or eQTL data.

The application of the drug target MR approach in a systematic manner requires further evaluation as suggested by the large proportion of drug target-indication pairs with results in the unanticipated direction of effect based on their mechanism of action. Complementary techniques such as co-localization[34] could be used to source additional evidence by exploring if the association observed in the drug target MR analysis is not attributable to genetic confounding through a variant in linkage disequilibrium[35].

This analysis also showed that many combinations of drug target gene - SOMAmer - traits could not be evaluated because of the lack of genetic instruments which could be explained by the sample size and the limited power to detect significant associations. This situation is likely to improve thanks to the commercialisation of cost-effective high-throughput technologies for protein measurement and the linkage of biobank data to electronic health records. For example, the genetic associations identified by deCODE genetics using the SomaLogic 5K platform in 35,559 Icelanders[36], or the promising UK Biobank Pharma Proteomics Project[37] which aims to measure circulating concentrations of up to 1,500 plasma proteins in approximately 53,000 UK Biobank participants using the Olink technology. In addition, the number of proteins covered by the proteomics platform is increasing, with the latest SomaLogic and Olink assays measuring up to 7,000[38] and approximately 3,000 proteins[39], respectively.

6.5.3. Strengths and limitations

The results presented in this chapter represent the first (at the time of analysis) systematic evaluation of the drug target MR framework using a 'truth' set of licensed drug target-indications pairs. One of the strengths of this analysis was the large number of approved drug targets for which measured protein levels and GWAS on the intended indication were available.

While only 11% (71/665) of all the genes that encode an approved drug target could be evaluated due to the lack of exposure or outcome GWAS data, this analysis could recapitulate the mechanism of action of approved drug target-indication pairs and inform the future direction of MR analysis for drug target identification and validation, as based on the findings presented in this chapter, a set of optimised parameters have not been identified yet and the performance of the drug target MR framework is defined by each drug target gene – SOMAmer – trait combination. This study also benefited from several sensitivity analyses which returned a set of well-validated target-indications. However, some drug target gene – SOMAmer – trait combinations which had their mechanism of action rediscovered in the sensitivity analysis 1 may have not reached significance in the more stringent analysis (sensitivity analyses 2 and 3) because the genetic associations with the exposure did not meet the criteria of 'strong genetic instruments'. Lastly, an extensive review of the drug target gene – SOMAmer – trait combinations was performed to hypothesize about potential explanations of the unexpected findings of the drug target MR analyses.

Some limitations of this analysis are noteworthy. First, certain indications may have been studied by GWAS but were not included in this study because the summary statistics were not deposited in the GWAS Catalog. Even for those traits available through the GWAS Catalog, the summary statistics may be incomplete and lack essential information for the MR analyses, such as effect sizes or effect/reference alleles. Secondly, it was assumed that protein expression levels (pQTL) can be used as a proxy of protein function. While two examples are provided at the beginning of the chapter which support such assumption, this has not been studied in detail due to the lack of GWAS data on protein activity. Moreover, protein levels corresponded to circulating protein in plasma, however, many proteins are not secreted or circulating in plasma, and therefore, their presence in the blood tissue could rather indicate physiological conditions. Since the function of these proteins should take place in a different tissue, it is unclear if the

levels in plasma recapitulate those in the drug effector tissue, or, on the contrary, they are unrelated to their function and should not be used to infer the effect of modifying such protein by a drug. For example, a membrane-bound protein detached from the plasma membrane or a protein inactivated by a post-translation process could still be detected by the proteomics platform if the part of the protein that is detected by the SOMAmer remains unchanged. Lastly, the lack of units for the pQTL, which is a major limitation of aptamer-based technologies, did not allow for a comparison with the effect size observed with the drug treatment, or limited the comparison across targets that for instance are targeted by drugs for the same indication. While this limitation does not have an impact on the current analysis, it may when applying the drug target MR framework with pQTL data as the one used here for target discovery.

6.6. Conclusion

The analysis presented in this chapter showed that the ability of the drug target MR framework to rediscover the mechanism of action of approved drug target-indication varies on a case-by-case basis. Only between 11-13% (i.e., 27/234 in the sensitivity analysis 1 and 16/121 in the sensitivity analysis 3) of the drug target gene – SOMAmer – trait combinations analysed rediscovered the mechanism of action of the drug. Therefore, the findings suggest that a set of gold standard parameters for the optimal performance of drug target MR cannot be defined yet, and the selection of parameters should be tailored to the drug target-indication of interest. Nonetheless, this analysis identified a set of targets genetically validated for the intended indication that could be investigated using a drug target MR - PheWAS approach. One of the major limitations of the drug target MR framework using pQTL data for systematic drug target identification and validation is the lack of the exposure data (e.g., only 11% of the targets for an approved drug had pQTL data available). In the next chapter, an application of the *biomarker-weighted* drug target MR approach will be presented as an alternative to pQTL-weighted drug target MR when pQTL data is not available to investigate the research question.

6.7. References

1. Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nature Communications* **11**, 3255 (2020).
2. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery* **1**, 727–730 (2002).
3. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
4. The SCALLOP Consortium. https://www.olink.com/scallop/. *Olink*.
5. Shah, T. *et al.* Population Genomics of Cardiometabolic Traits: Design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PLOS ONE* **8**, e71345 (2013).
6. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, (2017).
7. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913 (2007).
8. Benyamin, B. *et al.* GWAS of butyrylcholinesterase activity identifies four novel loci, independent effects within BCHE and secondary associations with metabolic risk factors. *Human Molecular Genetics* **20**, 4504–4514 (2011).
9. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896–D901 (2017).
10. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940 (2018).

11. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779 (2015).
12. Schmidt, A. F. *et al.* Cholesteryl ester transfer protein (CETP) as a drug target for cardiovascular disease. *Nat Commun* **12**, 5640 (2021).
13. Würtz, P. *et al.* Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase. *J. Am. Coll. Cardiol.* **67**, 1200–1210 (2016).
14. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium *et al.* The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet* **379**, 1214–1224 (2012).
15. Schmidt, A. F. *et al.* PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study. *Lancet Diabetes Endocrinol* **5**, 97–105 (2017).
16. Gordillo-Marañón, M. *et al.* Validation of lipid-related therapeutic targets for coronary heart disease prevention using human genetics. *bioRxiv* 2020.11.11.377747 (2020) doi:10.1101/2020.11.11.377747.
17. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
18. Ahola-Olli, A. V. *et al.* Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am J Hum Genet* **100**, 40–50 (2017).
19. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat Commun* **9**, 3268 (2018).
20. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat Metab* **2**, 1135–1148 (2020).

21. Ernsberger, P. Pharmacology of Moxonidine: An I1-Imidazoline Receptor Agonist. *Journal of Cardiovascular Pharmacology* **35**, S27 (2000).
22. Dhagat, U. *et al.* Cytokine Receptors and Their Ligands. in *Encyclopedia of Cell Biology* (eds. Bradshaw, R. A. & Stahl, P. D.) 22–36 (Academic Press, 2016). doi:10.1016/B978-0-12-394447-4.30002-5.
23. Conti, M. & Beavo, J. Biochemistry and Physiology of Cyclic Nucleotide Phosphodiesterases: Essential Components in Cyclic Nucleotide Signaling. *Annual Review of Biochemistry* **76**, 481–511 (2007).
24. Hunter, C. A. & Jones, S. A. IL-6 as a keystone cytokine in health and disease. *Nat Immunol* **16**, 448–457 (2015).
25. Rosa, M. *et al.* A Mendelian randomization study of IL6 signaling in cardiovascular diseases, immune-related disorders and longevity. *npj Genomic Medicine* **4**, 1–10 (2019).
26. Fisker, S. Physiology and pathophysiology of growth hormone-binding protein: methodological and clinical aspects. *Growth Horm IGF Res* **16**, 1–28 (2006).
27. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature Genetics* **52**, 1122–1131 (2020).
28. Ferreira, R. C. *et al.* Functional IL6R 358Ala allele impairs classical IL-6 receptor signaling and influences risk of diverse inflammatory diseases. *PLoS Genet* **9**, e1003444 (2013).
29. Gill, D. *et al.* ACE inhibition and cardiometabolic risk factors, lung ACE2 and TMPRSS2 gene expression, and plasma ACE2 levels: a Mendelian randomization study. *Royal Society Open Science* **7**, 200958.
30. Butler-Laporte, G. *et al.* The effect of angiotensin-converting enzyme levels on COVID-19 susceptibility and severity: a Mendelian randomization study. *Int J Epidemiol* **50**, 75–86 (2020).

31. Chauquet, S. *et al.* Association of Antihypertensive Drug Target Genes With Psychiatric Disorders: A Mendelian Randomization Study. *JAMA Psychiatry* **78**, 623–631 (2021).
32. Gaziano, L. *et al.* Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. *Nat Med* **27**, 668–676 (2021).
33. Yuan, S. *et al.* Genetically Proxied Inhibition of Coagulation Factors and Risk of Cardiovascular Disease: A Mendelian Randomization Study. *J Am Heart Assoc* **10**, e019644 (2021).
34. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
35. Gill, D. & Burgess, S. The evolution of mendelian randomization for investigating drug effects. *PLoS Med* **19**, e1003898 (2022).
36. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* **53**, 1712–1721 (2021).
37. UK Biobank launches one of the largest scientific studies. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/uk-biobank-launches-one-of-the-largest-scientific-studies.
38. The SomaScan Platform - Our Science - Platform. *SomaLogic* https://somalogic.com/somascan-platform/.
39. Olink Explore 3072. *Olink* https://www.olink.com/products-services/explore/.
40. Dahlbäck, B. & Villoutreix, B. O. Regulation of Blood Coagulation by the Protein C Anticoagulant Pathway. *Arteriosclerosis, Thrombosis, and Vascular Biology* **25**, 1311–1320 (2005).
41. Jin, S. *et al.* Brain ethanol metabolism by astrocytic ALDH2 drives the behavioural effects of ethanol intoxication. *Nat Metab* **3**, 337–351 (2021).

6.8. Appendices

Appendix 6.A. Data sources

GWAS trait	GWAS trait (CUI)	Drug indication	Drug indication (CUI)	N cases	N controls	Pubmed ID
Diastolic blood pressure	C0428883	Hypertension	C0020538	29136	-	19430479
Systolic blood pressure	C0488055	Hypertension	C0020538	29136	-	19430479
Pulse pressure	C0232108	Hypertension	C0020538	146562	-	27618448
Primary biliary cholangitis	C0023892	Liver Cirrhosis, Biliary	C0023892	64164	561055	26394269
Psoriasis	C0033860	Psoriasis	C0033860	10588	22806	23143594
Stroke	C0038454	Stroke	C0038454	40585	406111	29531354
Factor VIII levels	C0015506	Thrombosis	C0040053	8700	-	UCLEB
Factor VII levels	C0015502	Thrombosis	C0040053	8700	-	UCLEB
vWF levels	C0042971	Thrombosis	C0040053	9007	-	UCLEB
Prothrombin levels	C0033706	Thrombosis	C0040053	10000	-	Fenland
Activated partial thromboplastin time	C1318441	Thrombosis	C0040053	2406	-	UCLEB
Coronary Artery Disease	C1956346	Cardiovascular Diseases	C0007222	60801	123504	26343387
Coronary Artery Disease	C1956346	Coronary Artery Disease	C0010054	60801	123504	26343387
Coronary Artery Disease	C1956346	Coronary Artery Disease	C1956346	60801	123504	26343387
Carotid intima media thickness	C1960466	Cardiovascular Diseases	C0007222	60000	-	UCLEB
Carotid intima media thickness	C1960466	Coronary Artery Disease	C0010054	60000	-	UCLEB
Carotid intima media thickness	C1960466	Coronary Artery Disease	C1956346	60000	-	UCLEB
Carotid plaque	Plaque	Cardiovascular Diseases	C0007222	48434	-	UCLEB
Carotid plaque	Plaque	Coronary Artery Disease	C0010054	48434	-	UCLEB

GWAS trait	GWAS trait (CUI)	Drug indication	Drug indication (CUI)	N cases	N controls	Pubmed ID
Carotid plaque	Plaque	Coronary Artery Disease	C1956346	48434	-	UCLEB
FEV1	C0429706	Asthma	C0004096	321047	-	30804560
FEV1	C0429706	Pulmonary Disease, Chronic Obstructive	C0024117	321047	-	30804560
Lung function (FEV1/FVC)	C0429745	Asthma	C0004096	321047	-	30804560
Lung function (FEV1/FVC)	C0429745	Pulmonary Disease, Chronic Obstructive	C0024117	321047	-	30804560
Lung function (FVC)	C0580371	Asthma	C0004096	321047	-	30804560
Lung function (FVC)	C0580371	Pulmonary Disease, Chronic Obstructive	C0024117	321047	-	30804560
Peak expiratory flow	C0030735	Asthma	C0004096	321047	-	30804560
Peak expiratory flow	C0030735	Pulmonary Disease, Chronic Obstructive	C0024117	321047	-	30804560
Asthma (moderate or severe)	C0004096	Asthma	C0004096	88486	447859	30552067
Heart Failure	C0018801	Heart Failure	C0018801	47309	930014	31919418
Atrial Fibrillation	C0004238	Atrial Fibrillation	C0004238	60620	970216	30061737
Dermatitis, Atopic	C0011615	Dermatitis, Atopic	C0011615	21399	95464	26482879
Multiple Sclerosis	C0026769	Multiple Sclerosis	C0026769	14498	24091	24076602
Alzheimers disease (late onset)	C0002395	Alzheimer Disease	C0002395	24087	55058	30617256
Parkinsons disease	C0030567	Parkinson Disease	C0030567	15056	12637	31701892
Bipolar Disorder	C0005586	Bipolar Disorder	C0005586	7647	27303	27329760
Schizophrenia	C0036341	Schizophrenia	C0036341	35476	46839	25056061
Lupus Erythematosus, Systemic	C0024141	Lupus Erythematosus, Systemic	C0024141	6748	11516	28714469

GWAS trait	GWAS trait (CUI)	Drug indication	Drug indication (CUI)	N cases	N controls	Pubmed ID
Alcohol Consumption	C0001948	Alcoholism	C0001973	480842	-	31358974
Fractures	C0016658	Osteoporosis	C0029456	53184	373611	30598549
Heel bone mineral density	C0005938	Osteoporosis	C0029456	426824	-	30598549
Ferritin levels	C0015879	Anemia, Iron-Deficiency	C0162316	4948	-	UCLEB
Chronic Kidney Insufficiency	C0403447	Kidney Failure, Chronic	C0022661	12315	227987	31152163
Chronic Kidney Insufficiency	C0403447	Renal Insufficiency, Chronic	C0403447	12315	227987	31152163
Blood urea nitrogen levels	C0005845	Renal Insufficiency, Chronic	C0403447	416178	-	31152163
Estimated glomerular filtration rate	C3811844	Renal Insufficiency, Chronic	C0403447	567460	-	31152163
Epilepsy	C0014544	Epilepsy	C0014544	15212	29677	30531953
LDL cholesterol	C0023824	Hypercholesterolemia	C0020443	188577	-	24097068
LDL cholesterol	C0023824	Hyperlipidemias	C0020473	188577	-	24097068
LDL cholesterol	C0023824	Dyslipidemias	C0242339	188577	-	24097068
LDL cholesterol	C0023824	Cardiovascular Diseases	C0007222	188577	-	24097068
LDL cholesterol	C0023824	Coronary Artery Disease	C0010054	188577	-	24097068
LDL cholesterol	C0023824	Coronary Artery Disease	C1956346	188577	-	24097068
Diabetes Mellitus, Type 2	C0011860	Diabetes Mellitus, Type 2	C0011860	74124	898130	30297969
Proinsulin levels	C0033362	Diabetes Mellitus, Type 2	C0011860	10701	-	21873549
Fasting blood glucose	C0428568	Diabetes Mellitus, Type 2	C0011860	58074	-	22581228
Fasting blood insulin	C2676369	Diabetes Mellitus, Type 2	C0011860	51750	-	22581228
Myocardial Infarction	C0027051	Myocardial Infarction	C0027051	40149	126310	26343387
Rheumatoid arthritis	C0003873	Arthritis, Rheumatoid	C0003873	19234	60565	24390342

GWAS trait	GWAS trait (CUI)	Drug indication	Drug indication (CUI)	N cases	N controls	Pubmed ID
Juvenile idiopathic arthritis (Oligoarticular juvenile idiopathic arthritis)	C3495559	Arthritis, Juvenile	C3495559	2816	13056	23603761
Factor VIII levels	C0015506	Venous Thrombosis	C0042487	8700	-	UCLEB
Factor VII levels	C0015502	Venous Thrombosis	C0042487	8700	-	UCLEB
vWF levels	C0042971	Venous Thrombosis	C0042487	9007	-	UCLEB
Prothrombin levels	C0033706	Venous Thrombosis	C0042487	10000	-	Fenland
Activated partial thromboplastin time	C1318441	Venous Thrombosis	C0042487	2406	-	UCLEB
Diabetes Mellitus, Type 2	C0011860	Diabetes Mellitus	C0011849	74124	898130	30297969
Erectile Dysfunction	C0242350	Erectile Dysfunction	C0242350	6175	217630	30583798
Gout	C0018099	Gout	C0018099	13179	750634	31578528
Urate levels	C0455272	Gout	C0018099	288649	-	31578528
Urate levels	C0455272	Hyperuricemia	C0740394	288649	-	31578528
Diabetic Nephropathies	C0011881	Diabetic Nephropathies	C0011881	5447	4717	29703844
Triglycerides	C0041004	Coronary Artery Disease	C0010054	188577	-	24097068
Triglycerides	C0041004	Coronary Artery Disease	C1956346	188577	-	24097068
Proinsulin levels	C0033362	Diabetes Mellitus	C0011849	10701	-	21873549
Fasting blood glucose	C0428568	Diabetes Mellitus	C0011849	58074	-	22581228
Fasting blood insulin	C2676369	Diabetes Mellitus	C0011849	51750	-	22581228
Diabetes Mellitus, Type 1	C0011854	Diabetes Mellitus, Type 1	C0011854	6683	12173	25751624
Obesity	C0028754	Obesity	C0028754	32858	65840	23563607
Body mass index	C0005893	Obesity	C0028754	806834	-	30239722

GWAS trait	GWAS trait (CUI)	Drug indication	Drug indication (CUI)	N cases	N controls	Pubmed ID
Waist-to-hip ratio	C0205682	Obesity	C0028754	694649	-	30239722
Tumor necrosis factor alpha levels	C1168005	Inflammation	C0021368	3454	-	27989323
Tumor necrosis factor beta levels	C0024320	Inflammation	C0021368	1559	-	27989323
Vascular endothelial growth factor levels	C0078058	Inflammation	C0021368	7118	-	27989323
Spondylitis, Ankylosing	C0038013	Spondylitis, Ankylosing	C0038013	10619	15145	23749187
Colitis, Ulcerative	C0009324	Colitis, Ulcerative	C0009324	6968	20464	26192919
Crohn Disease	C0010346	Crohn Disease	C0010346	5956	14927	26192919
Height	C0005890	Deficiency of growth hormone	C0013338	347086	-	30124842

Appendix 6.B. Drug target MR analyses

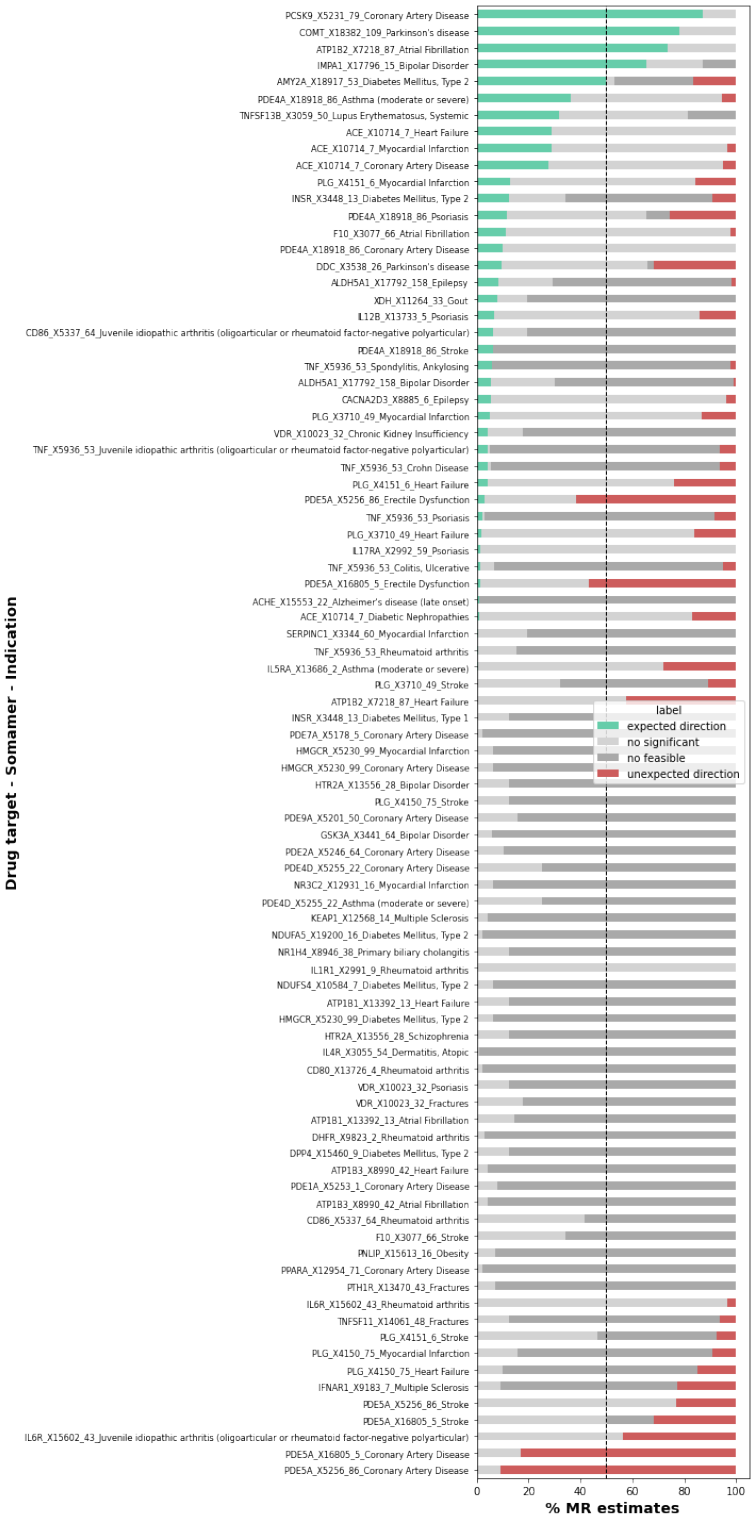


Figure 6.B1. Results of the main analysis (all parameter combinations): Percentage of significant MR estimates in the expected (green), unexpected (red) direction of effect, non-significant estimates (light grey) or non-feasible tests (dark grey) for SOMAmer-drug target gene-trait pairs where binary traits were used as the outcome.

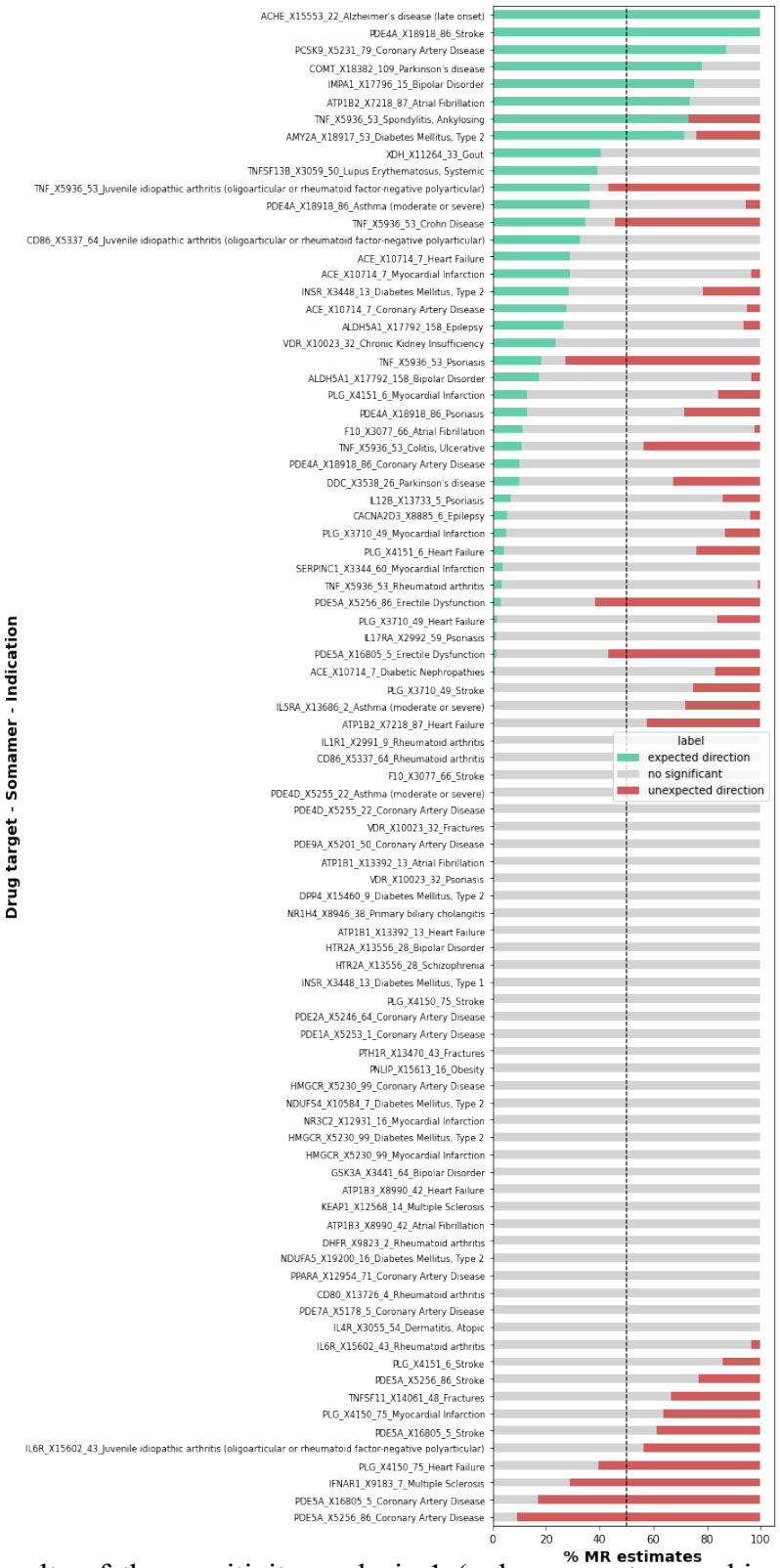


Figure 6.B2. Results of the sensitivity analysis 1 (only parameter combinations that yield results): Percentage of significant MR estimates in the expected (green), unexpected (red) direction of effect or non-significant estimates (grey) for SOMAmer-drug target gene-trait pairs where binary traits were used as the outcome.

Figure 6.B3. Results of the sensitivity analysis 2 (only credible parameter combinations): Percentage of significant MR estimates in the expected (green), unexpected (red) direction of effect, non-significant estimates (light grey) or non-feasible tests (dark grey) for SOMAmer-drug target gene-trait pairs where binary traits were used as the outcome.

Figure 6.B4. Results of the main analysis (all parameter combinations): Percentage of significant MR estimates in the expected (green), unexpected (red) direction of effect, non-significant estimates (light grey) or non-feasible tests (dark grey) for SOMAmer - drug target gene - trait pairs where quantitative traits were used as the outcome.

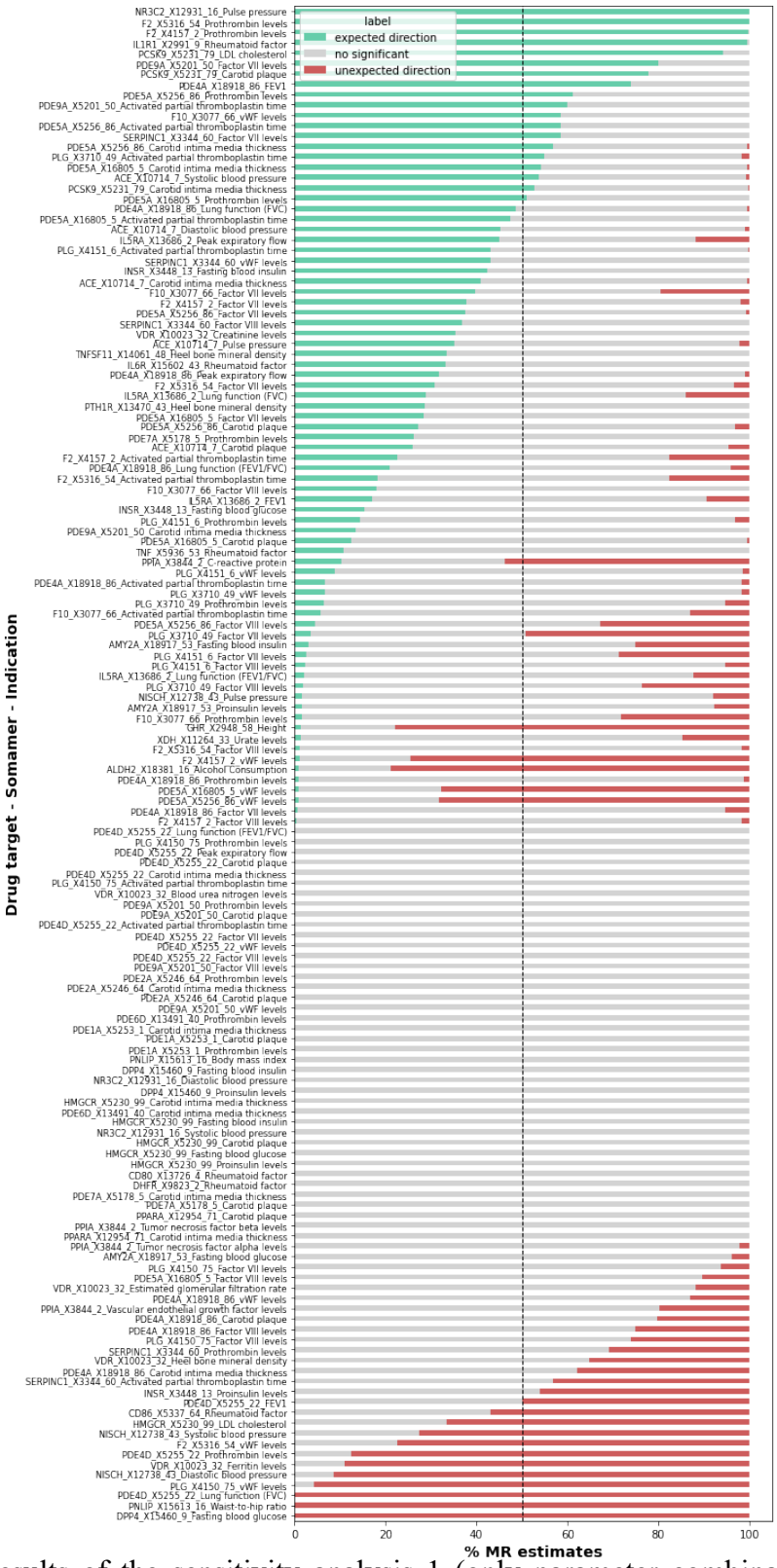


Figure 6.B5. Results of the sensitivity analysis 1 (only parameter combinations that yield results). Percentage of significant MR estimates in the expected (green), unexpected (red) direction of effect or non-significant estimates (grey) for SOMAmer-drug target gene-trait pairs where quantitative traits were used as the outcome.

Figure 6.B6. Results of the sensitivity analysis 2 (only credible parameter combinations). Percentage of significant MR estimates in the expected (green), unexpected (red) direction of effect, non-significant estimates (light grey) or non-feasible tests (dark grey) for SOMAmer-drug target gene-trait pairs where quantitative traits were used as the outcome.

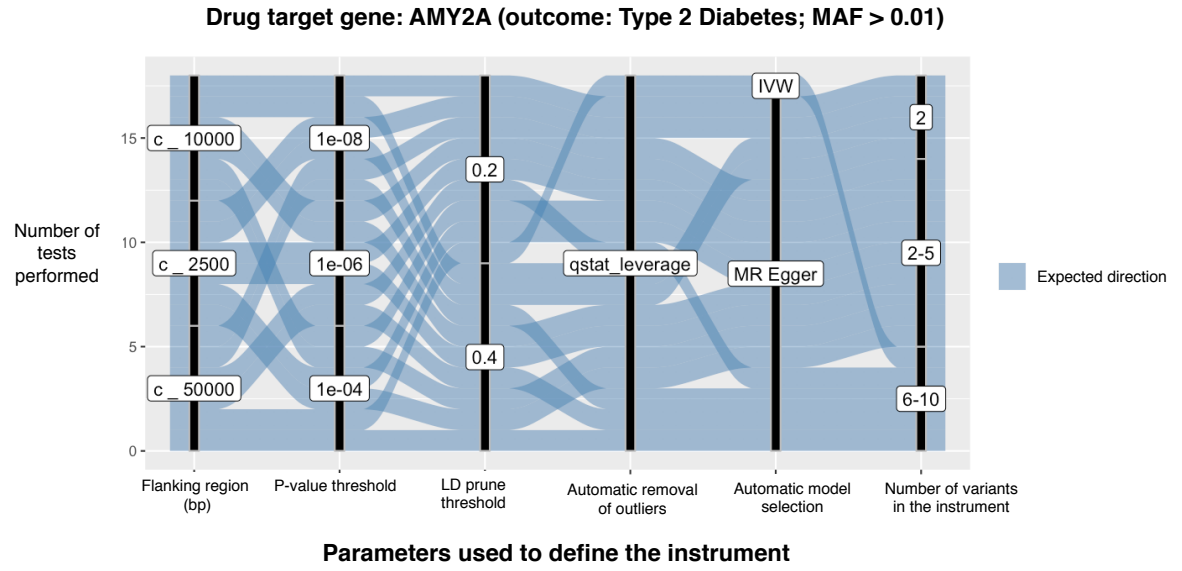
Appendix 6.C. Case review

Combinations with results both in the expected and unexpected direction of effect

Some drug target gene – SOMAmer – trait combinations showed both the expected and the unexpected direction of effect in the analysis with the most stringent parameters (sensitivity analysis 3). This section includes a qualitative evaluation of three cases were the modification of one or more parameter led to a change in the direction of effect.

The first example is AMY2A, the drug target of ACARBOSE and used to treat Type 2 Diabetes. In the sensitivity analysis 3, all the results were significant, with 24 in the expected direction of effect and 12 in the unexpected direction of effect (Figure 6.4). To explore the nature of the change, a plot was built to visualise each of the MR tests performed and to identify common characteristics for those that were in the unexpected direction of effect. When using a minor allele frequency (MAF) threshold of 0.01 to select genetic instruments, all the MR estimates were in the expected direction of effect (Fig. 6.B7.A). When the MAF was increased to 0.05, fewer variants were selected resulting in MR tests of a single variant or only 2 variants in the unexpected direction of effect (Fig. 6.B7.B). In fact, the shift in the direction of effect was caused by the genetic variant 1:104158889 (GRCh 37), which is located 1.1kbp upstream of the encoding gene in a non-coding region. Literature or database references of the variant were not found. This example illustrates how building a genetic instrument with a single variant can lead to spurious findings, particularly, if the genetic variant has not been documented before for its effect on the nearby gene. Therefore, leveraging multiple genetic variants in and around the gene reduces the potential bias due to invalid instrumental variables.

A



B

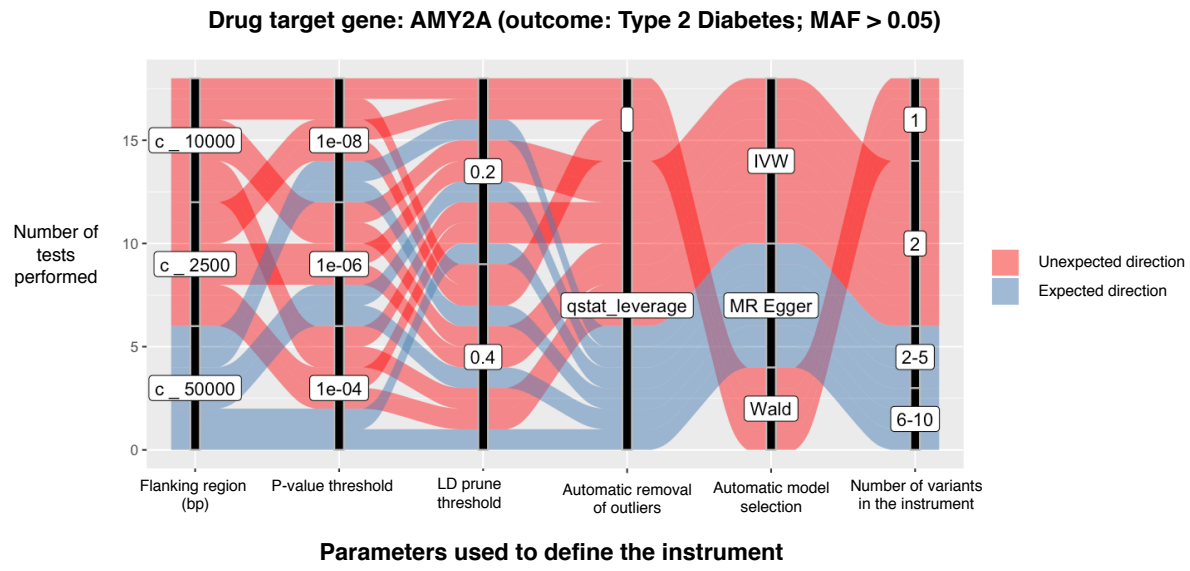
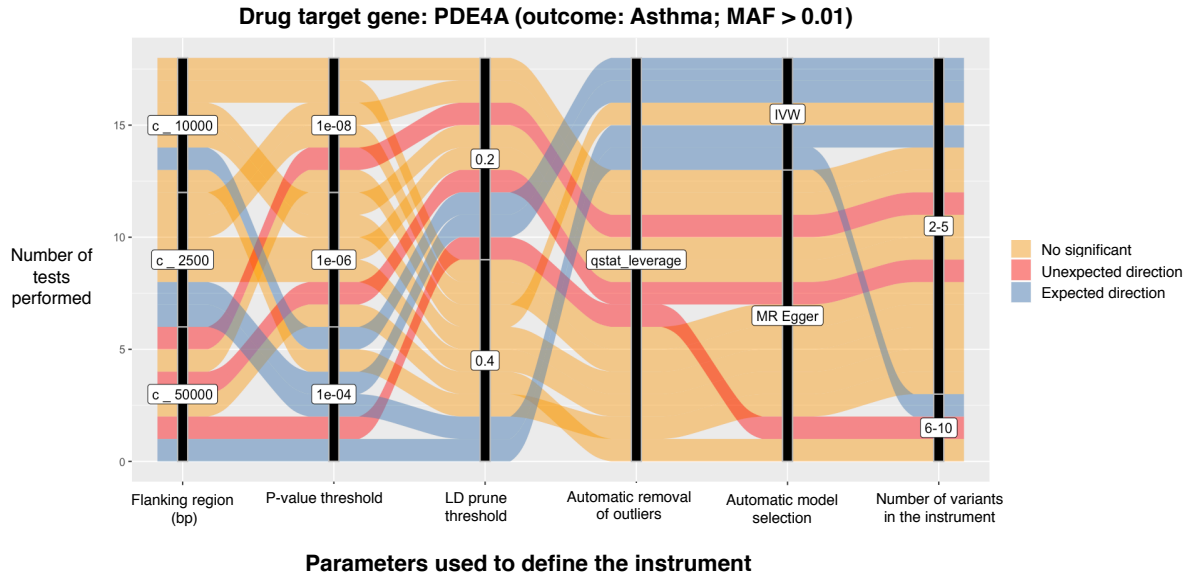


Figure 6.B7. Mendelian Randomisation tests performed under sensitivity analysis 3 for the drug target AMY2A and the indication Type 2 Diabetes. Panel A shows the approach for instrument selection for variants with a MAF > 0.01, and panel B the approach for variants with MAF > 0.05.

The second example involves PDE4A, the drug target of THEOPHYLLINE, AMINOPHYLLINE and ROFLUMILAST for the treatment of asthma. In the sensitivity analysis 3, 16 tests were significant in the expected direction of effect, 3 were significant in the unexpected direction of effect and 17 were not significant (Figure 6.4). The visualisation of each of the MR tests for the drug target-indication pair showed that increasing the MAF from 0.01 to 0.05 leads to the exclusion of a genetic variant (19_10568883_C_G) that forced the slope of the association to go in the opposite direction of effect (Fig. 6.B8). This is illustrated further in Figure 6.B9, where the estimates from two MR tests, with or without 19_10568883_C_G, showed that its inclusion in the genetic instrument led to the selection of the MR Egger over the IVW method. Although this genetic variant (rs145530718) is described as intron variant in Ensembl, the large effect size on both the exposure and outcome and the opposite direction compared to the other variants in the instrument suggest a potential pleiotropic effect through a different pathway.

A



B

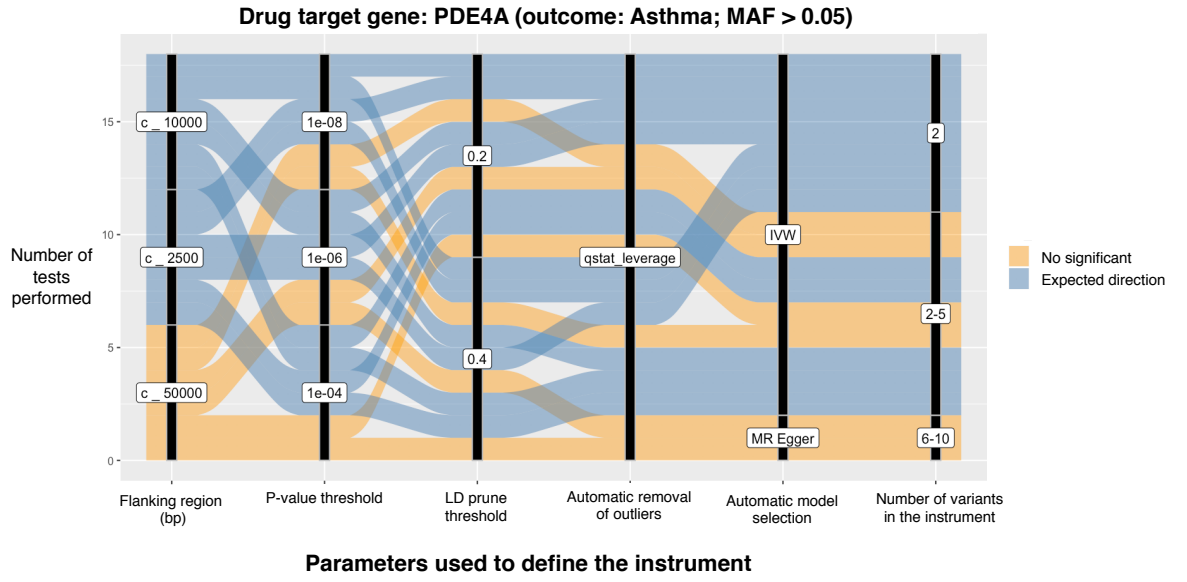


Figure 6.B8. Mendelian Randomisation tests performed under the sensitivity analysis 3 for the drug target PDE4A and the indication Asthma. Panel A shows the approach for instrument selection for variants with a MAF > 0.01, and panel B the approach for variants with MAF > 0.05.

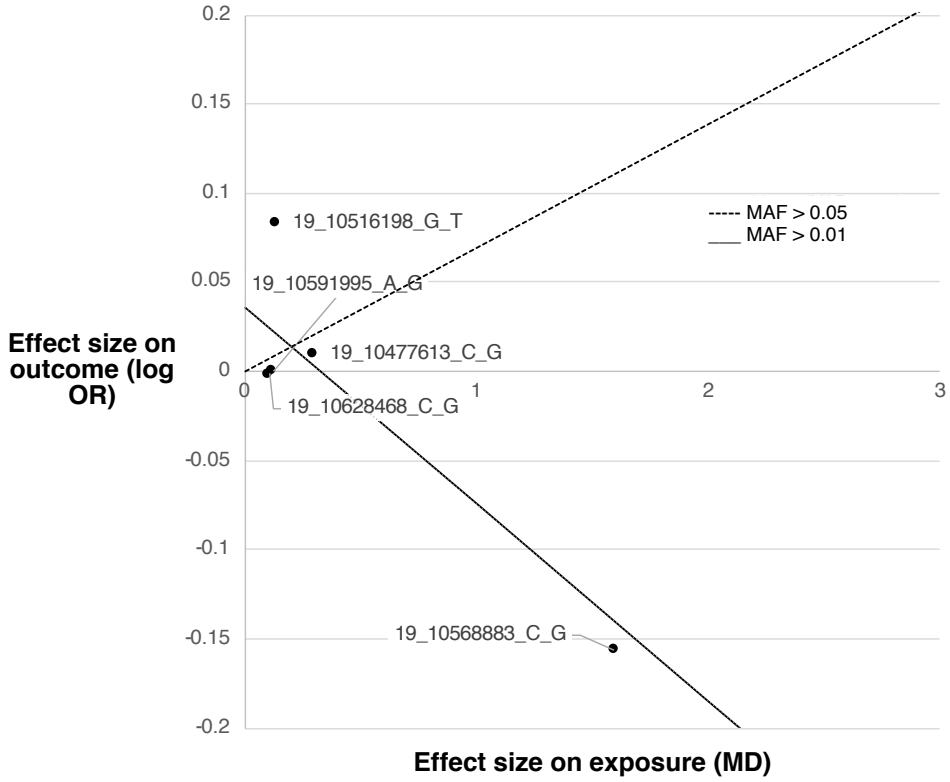


Figure 6.B9. Illustration of the impact of including heterogeneous variants in the genetic instrument. This is a real example of one of the MR tests performed for the drug target PDE4A and asthma.

The third example involves COMT, the drug target of ENTACAPONE and TOLCAPONE, used to treat the symptoms of Parkinson's disease. This example is particularly interesting as the drug is intended to treat the symptoms of the disease but the analysis in the current chapter also showed a significant association with the disease. It also illustrates the importance of selecting strong genetic associations to build the genetic instrument and to ensure that the first assumption of the MR framework ('Relevance assumption') holds. In the sensitivity analysis 3, 26 tests were significant in the expected direction of effect, and 10 were not significant (Figure 6.4). Figure 6.B10 shows that the power of the MR analysis to detect a causal association decreases when the *p* value threshold for selecting genetic variants is relaxed from 1×10^{-6} to 1×10^{-4}.

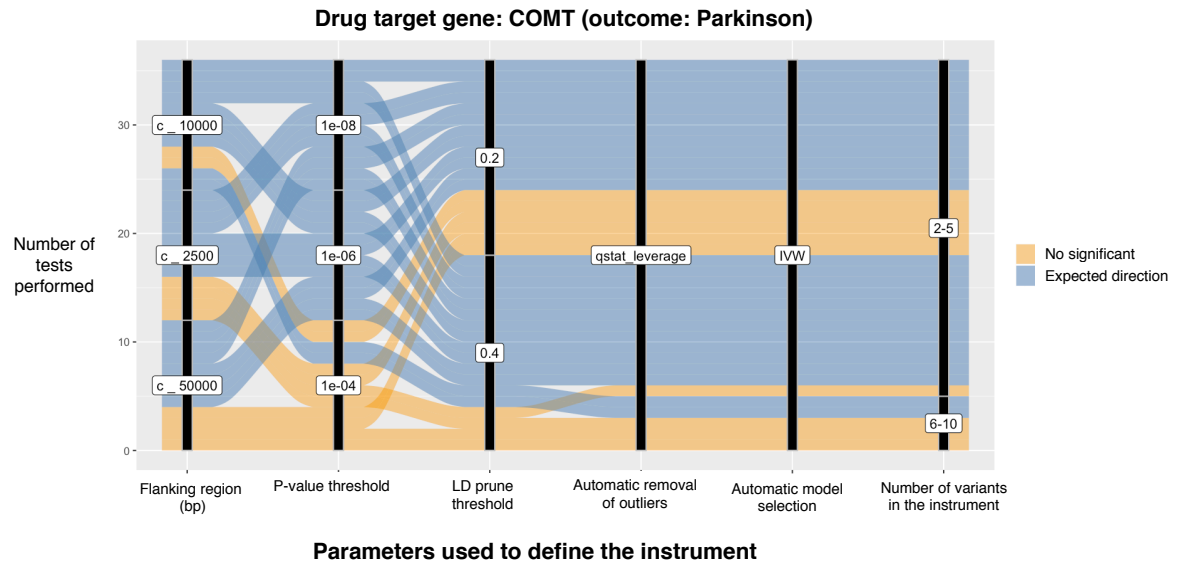


Figure 6.B10. Mendelian Randomisation tests performed under the sensitivity analysis 3 for the drug target COMT and the indication Parkinson.

Combinations consistently in the unexpected direction of effect

Drug target gene – SOMAmer – trait combinations consistently in the unexpected direction of effect in the analysis with the most stringent parameters (sensitivity analysis 3), for which a plausible mechanism to explain the discordant association could not be found or if suggested, did not fall within the three categories presented in the results section.

Argatroban and bivalirudin are direct thrombin (factor II) inhibitors used to treat or prevent thrombosis. vWF is a carrier protein of factor VIII and once activated, generates a complex with factor IXa and activates factor X, which in a complex with FVa converts prothrombin (factor II) to thrombin (factor IIa). Fibrinogen is then converted to a fibrin clot by thrombin, which can lead to thrombosis. High prothrombin levels inhibit the inactivation of factor VIIIa, while thrombin downregulates factor VIIIa levels through the Protein C anticoagulant pathway[40]. Thus, one of the potential explanations for the unexpected direction of effect observed for vWF and factor VIII levels may be that the measurement by the

SomaLogic platform did not only capture inactivated factor II (prothrombin) but also factor IIa (thrombin), and the association observed in the MR analysis recapitulated the negative feedback between thrombin and factor VIII/vWF levels.

Dalteparin sodium, danaparoid sodium, enoxaparin sodium and tinzaparin sodium are activators of SERPINC1 (antithrombin III). Antithrombin inactivates thrombin, factor IXa and factor Xa to impede clot formation. Eight parameter combinations were explored in the drug target MR analyses (sensitivity analysis 2), which always returned a single variant intronic to construct the instrument (GChr37: 1_173910084_C_T; rs146832357). The lack of coverage in the exposure together with the lack of information on the only available variant challenge the interpretation of the results in the unexpected direction of effect.

The aldehyde dehydrogenase 2 (ALDH2) is the drug target of disulfiram, a single protein inhibitor used to treat alcoholism. The drug target MR analysis of blood circulating ALDH2 levels on alcohol consumption showed a discordant direction of effect in >50% of the scenarios explored. Although ALDH2 is detected in all tissues, it is not actively secreted, and its presence in plasma may not be a results of the normal homeostasis of the protein lifecycle, and therefore the effector tissue of disulfiram. In fact, brain ethanol metabolism by the ALDH2 in astrocytes has recently been suggested as the contributor to the behavioural effects associated with ethanol intoxication[41].

7 | Biomarker-weighted drug target Mendelian Randomisation: applications in cardiovascular disease treatment and prevention

The work from this chapter has been published in *Nature Communications*[1]

7.1. Abstract

Biomarker-weighted drug target Mendelian randomisation (MR) studies use DNA sequence variants in or near a gene encoding a drug target, that alter the target's expression or function, as a tool to anticipate the effect of drug action on the same target through the association with a downstream biomarker. Here I applied biomarker-weighted drug target MR to prioritise drug targets for their causal relevance for coronary heart disease (CHD). The targets were further prioritised using independent replication and by sourcing data from the British National Formulary and clinicaltrials.gov. Out of the 341 drug targets identified through their association with blood lipids (HDL-C, LDL-C and triglycerides), 30 targets that might elicit beneficial effects in the prevention or treatment of CHD were robustly prioritised, including NPC1L1 and PCSK9, the targets of drugs used in CHD prevention. In this chapter I also discuss how this approach can be generalised to other targets, disease biomarkers and endpoints to help prioritise and validate targets during the drug development process.

7.2. Introduction

A well-established role of Mendelian randomisation (MR) analysis is to use genetic variants (mostly identified from GWAS) as instrumental variables to identify which disease biomarkers (e.g. blood lipids such as low- and high-density lipoprotein cholesterol and triglycerides) may be causally related to disease endpoints (e.g. coronary heart disease; CHD)[2,3]. It has also been shown that variants in a gene encoding a specific drug target (acting in *cis*), that alter the target's expression or function, can be used as a tool to anticipate the effect of drug action on the same target, which is known as 'drug target MR'[4] and has extensively been described in the previous chapters. Both 'genome-wide biomarker' and 'drug target MR' approaches were described in Chapter 1.4, with the main conceptual differences detailed in Table 1.4. In summary, whereas 'genome-wide biomarker MR' helps infer the causal relevance of a biomarker for a disease, a 'drug target MR' helps infer whether and, in certain cases in what direction, a drug that acts on the encoded protein, whether an antagonist, agonist, activator or inhibitor, will alter disease risk.

Different subtypes of 'drug target MR' analyses are used based on the exposure data, each with their unique strengths and limitations. In Chapter 6, the 'pQTL-weighted drug target MR' framework was applied and evaluated using genetic associations with circulating protein levels to instrument the effect of perturbing the drug target on the approved indication. Such approach uses the most accurate exposure, in principle, for drug target characterisation, because the vast majority of successful drugs achieve their activity by binding to and modifying the level, function or activity of a protein[5]. However, genetic associations with circulating protein levels have not become available until recently, and only include a subset of the proteome, and its usefulness and systematic application in drug development is still under investigation. On the other hand, 'biomarker-weighted drug target MR' has been extensively used as described

in the Chapter 1 using CETP and coronary heart disease for illustration, with references to studies showing that a drug target MR of *CETP* on CHD, using variants in the *CETP* gene weighted by their effect on HDL-C, indicates protection from disease (odds ratio (OR): 0.87; 95%CI: 0.84, 0.90)[4], which is consistent with the effect of allocation to the CETP-inhibitor anacetrapib in a placebo-controlled trial (OR:0.93; 95%CI: 0.86, 0.99) and compatible with the view that targeting *CETP* is an effective therapeutic approach to prevent CHD[6]. Importantly, as discussed in detail by Schmidt *et al.,*[4], drug target MR analyses which use genetic associations with biomarkers downstream to the protein such as HDL-C, use this effect as a proxy for protein concentration or activity (where this has not been measured directly), and do not provide evidence on whether the biomarker used for the weighting itself mediates disease. Rather, they inform on the validity of the drug target for a disease, regardless of the mediating pathway. Biomarker-weighted drug target MR analyses are particularly relevant when genetic associations with the drug target protein levels or activity have not been measured directly, or if available, do not represent strong instruments. Instead, genetic associations with a downstream biomarker in or near the gene encoding the drug target could be used as a proxy for protein concentration or activity.

The already published 'biomarker-weighted drug target MR' analyses suggest that unexploited drug targets might exist for the prevention or treatment of CHD that could be identified through their association with blood lipids even though such analyses do not presume that the effect on CHD is mediated through these lipids.

In this chapter, I applied drug target MR on a set of druggable proteins identified through genetic associations with circulating blood lipids and assessed their causal relevance for CHD. Summary statistics from GWAS of blood lipids and CHD were used to select genes associated with blood lipids that encode druggable targets and the effects of these drug targets on CHD

were tested using ‘drug target MR’ in two independent datasets. Subsequently, data from clinicaltrials.gov and the British National Formulary (BNF) was sourced for drugs in clinical phase development and licensed medicines, respectively, to identify agents that might be pursued rapidly in clinical phase testing for treatment or prevention of CHD. Because of interest in this area, though not the focus of the work, I also evaluated potential mediators of these effects using multivariable MR (MVMR). Finally, I discussed how this approach might be generalised to other drug targets and clinical endpoints, providing a route to translating findings from GWAS into new drug development.

7.3. Methods

7.3.1. Data sources

To determine the causal role and replicate previously reported results on the causal effect of LDL-C, HDL-C and TG on CHD, summary-level genetic estimates were obtained from the Global Lipids Genetics Consortium (188,577 individuals)[7] and from CardiogramPlusC4D (60,801 cases and 123,504 controls)[8].

Independent replication data were sourced using lipids exposure data from a GWAS meta-analysis of metabolic measures by the University College London–Edinburgh-Bristol (UCLEB) Consortium[9] and Kettunen *et al.,*[10] utilizing NMR spectroscopy measured lipids (joint sample size up to 33,029). Independent CHD data was obtained from a publicly available GWAS of 34,541 cases and 261,984 controls in UK Biobank[11].

Individual-level data from a random subset of 5,000 unrelated individuals of European ancestry from UK Biobank was used to generate the LD reference matrices as described in the Instrument selection section.

7.3.2. Drug target gene selection

To estimate the causal effect of modulating the level of each lipid sub-fraction via a druggable gene on CHD, genetic variants associated with LDL-C, HDL-C and/or TG with a *p* value $\leq 1x10^{-6}$ were selected. Druggable genes overlapping a 50 kbp region around the selected variants were extracted, resulting in 341 associated drug target genes (149 for LDL-C, 180 for HDL-C and 154 for TG). The set of genes in the 'druggable genome' were identified[12] (see Chapter 3.2), and identifiers were updated to Ensembl version 95 (GRCh37), used in this

analysis. All of these IDs were also present in Ensembl 95 (GRCh37), used in this analysis. Because only genetic associations with the druggable genome were scanned for, protein-coding genes that were the 'true' causal gene but not yet druggable would be missed and the association mis-assigned. To mitigate this and provide information about potential effects through non-druggable genes, the minimum distance from the variant to the druggable gene is provided in Appendix 7.A, where variants located within a gene were given a distance of 0bp, together with a gene distance rank value according to their base pair distance (including all protein-coding genes), and a column indicating if the druggable gene had been prioritised by GLGC in previous studies[7].

7.3.3. Instrument selection

For the biomarker or genome-wide MR analyses, a *p* value threshold of $1x10^{-6}$ was used to select exposure variants associated with LDL-C, HDL-C and/or TG. For *cis*- or drug target MR analyses, variants within the 341 selected genes (±50 kbp) were selected based on a *p* value $\leq 1x10^{-4}$. In both settings, variants were filtered on a MAF > 0.01 and LD clumped to an $r^2 < 0.4$. These parameters showed the most consistent estimates in a grid-search in the discovery data using the positive control examples: PCSK9, NPC1L1, HMGCR and CETP (Fig. 7.1). To account for residual correlation between variants in the MR analyses, a generalised least squares framework with a LD reference dataset derived from UK Biobank was applied[13] (see Chapter 2.3. for details on the framework). LD reference matrices were created by extracting a random subset of 5,000 unrelated individuals of European ancestry from UK Biobank. Variants with a MAF < 0.001, and imputation quality < 0.3 were excluded. To ensure that SNPs with lower MAF have higher confidence, variants were removed if MAF < 0.005 and genotype

probability < 0.9; MAF < 0.01 and genotype probability < 0.8; MAF < 0.03 and genotype probability < 0.6.

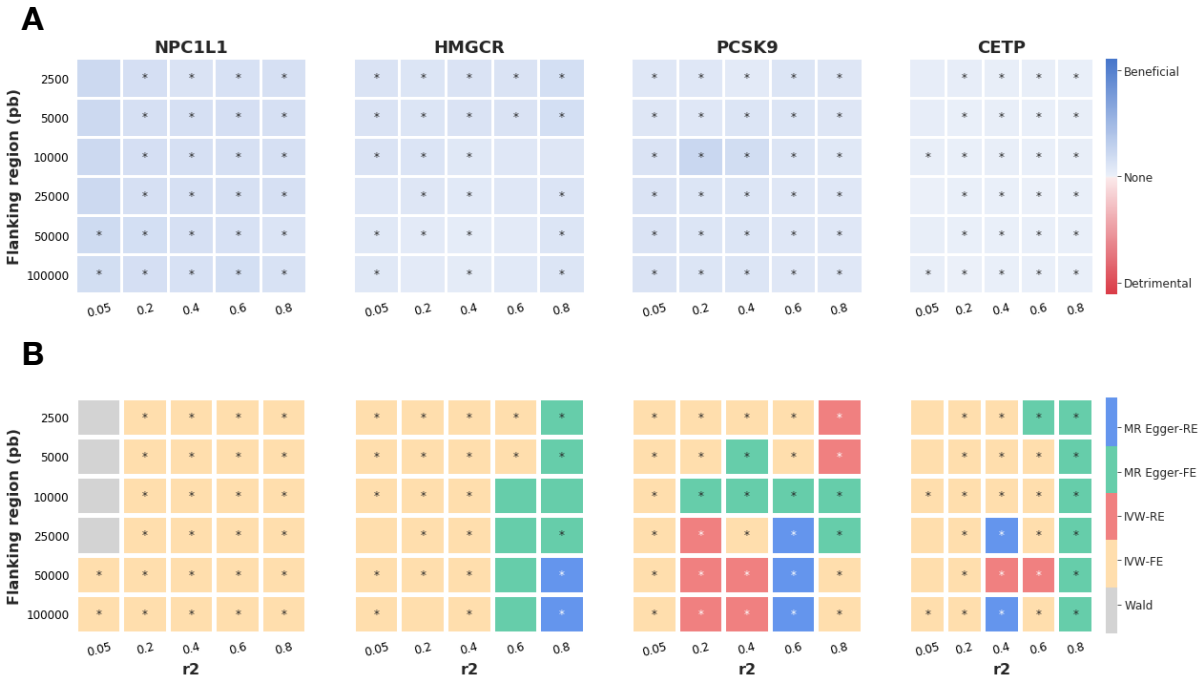


Figure 7.1. Drug target MR of positive control examples. Grid search of LD threshold and region around the gene encoding a druggable target using genetic associations with LDL-C and HDL-C from the Global Lipid Genetic Consortium (GLGC) with CHD events from the CardiogramPlusC4D Consortium. MR estimates (A) and preferred model (B) for three licensed LDL-lowering drug targets and HDL-lowering CETP using lipid data from GLGC and CHD data from CardiogramPlusC4D in the discovery analysis. Models explored: MR Egger-RE (random effects), MR Egger-RE (fixed effects), inverse variance weighted (IVW)-RE (random effects), IVW-FE (fixed effects), Wald ratio. In panel A, blue indicates a beneficial effect on CHD risk, and red a detrimental effect per SD difference with respect to the indicated lipid sub-fraction. Significant estimates are indicated with an asterisk (*).

7.3.4. Mendelian Randomisation analysis

A model-selection framework was used to decide between competing inverse-variance weighted (IVW) fixed-effects, IVW random-effects, MR-Egger fixed effects or MR-Egger random-effects models[14]. While IVW models assume an absence of directional horizontal pleiotropy, Egger models allow for possible directional pleiotropy at the cost of power. See Chapter 2.3 for details on the model-selection framework, IVW and Egger models. After removing variants with large heterogeneity (p value < 0.001 for Cochran's Q test) or leverage, this model selection framework was re-applied and the final model used. The influence of parameter selection in the drug target MR performance was explored in a grid-search of several r^2 and gene boundaries combinations using the positive control examples PCSK9, NPC1L1, HMGCR and CETP, where the lipid perturbation is the intended indication. To assess the possibility of false positive results, the empirical p value distribution of the discovery MR findings was compared against the continuous uniform distribution using the Kolmogorov-Smirnov goodness-of-fit test (two-sided). Under the null hypothesis of no association, p values follow a continuous uniform distribution between 0 and 1[15].

Additionally, a drug target multivariable MR analysis was conducted using genetic associations with the three lipid sub-fractions and CHD risk in a single regression model, to identify likely mediating lipids in the causal pathway of CHD. For details on multivariable MR analysis, see Chapter 2.2.4.4.

Results were presented as mean difference (MD) or odds ratio (OR) with 95% confidence interval (95%CI) coded towards the canonical drug target effect direction; i.e., towards lower LDL-C and triglyceride concentration, and higher HDL-C concentration.

7.3.5. Drug indications and adverse effects

To evaluate if the drug target MR analyses rediscovered known drug indications, adverse effects or predicted repurposing opportunities, drug information and clinical trial data was extracted for the set of 341 druggable targets. Drug target genes were mapped to UniProt identifiers and indications and clinical phase for compounds that bind the target were extracted from the ChEMBL database (version 25)[16]. Drug indications and lipid adverse effects data for licensed drugs were extracted from the British National Formulary (BNF) website in July, 2019.

To further examine the effects of the drugs and clinical candidates that are known to act through binding to the 341 druggable targets, relevant clinical trial data were downloaded from the clinicaltrials.gov registry. Compound name and synonyms were extracted from ChEMBL database (version 25)[16] and used to identify clinical trials with matching interventions. In case of non-exact matches, the results were inspected manually to ensure that only relevant trial records were used in the analysis. Lipid-related trial outcomes and adverse events were identified by searching the relevant fields within the trial records with the keywords: lipo*, lipid*, ldl*, hdl*, cholest* and triglyceride*. For adverse events, the search was limited to the trial arm in which the drug of interest was administered (as opposed to placebo or active control used in the study) and only adverse events that affected at least one study participant were included.

7.4. Results

7.4.1. Biomarker-weighted univariable drug target Mendelian Randomisation

Drug target MR was used to determine the effect on CHD of perturbing druggable proteins that influence one or more of the three lipid fractions. First, genes previously shown to encode druggable proteins were selected in regions around variants associated with one or more of the major circulating lipid subfractions applying a p value $\leq 1x10^{-6}$. This identified 341 genes; 149 for an association with LDL-C, 180 for HDL-C and 154 for TG[12]. One hundred forty genes (41%) were associated with a single lipid subfraction, 171 (50%) were associated with two subfractions and 30 (9%) were associated with all three subfractions (Fig. 7.2, Appendix 7.A). Subsequently, a drug target MR analysis was performed on CHD accounting for genetic correlation between variants. In the absence of direct measures of the encoded protein, the effect of genetic drug target perturbation was proxied through the downstream effect on one or more of the three lipid sub-fractions. Here genetic associations with LDL-C, HDL-C, and TG were used as a proxy for drug target effects on CHD, which does not provide direct evidence on whether the drug target itself affects CHD through the leveraged lipid weight; this mediation question is subsequently explored using multivariable MR.

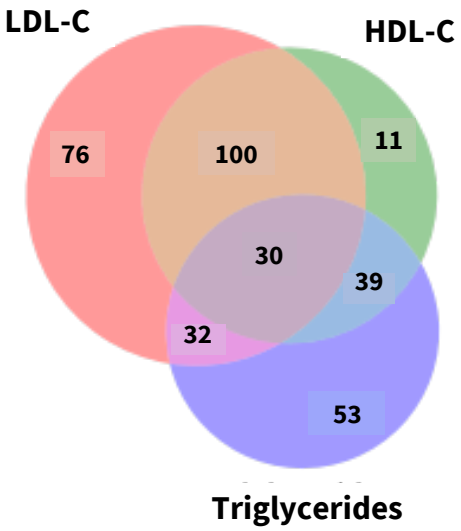


Figure 7.2. Overlap between genes encoding druggable targets associated with the major lipid subfractions. The Venn diagram shows genes exhibiting overlapping or exclusive associations with LDL-C, HDL-C and/or TG.

Of the 341 drug targets, 165 could be associated with CHD, with 131 of these estimates being consistent with a protective effect when instrumented for a reduction in LDL-C or TG and/or elevation in HDL-C (Fig. 7.3, Appendix 7.B). When weighted by LDL-C, eighty-seven targets showed a significant effect on CHD after orientating towards an increasing LDL-C direction, with the first and third quartiles (Q) of the CHD OR of 1.93 and 3.32. Similarly, the Q1 and Q3 after orientating the OR towards an increasing HDL-C direction were 0.22 and 0.53 for the 49 significant HDL-C instrumented targets, and for the 49 significant TG instrumented targets these were 1.95 and 4.35, respectively.

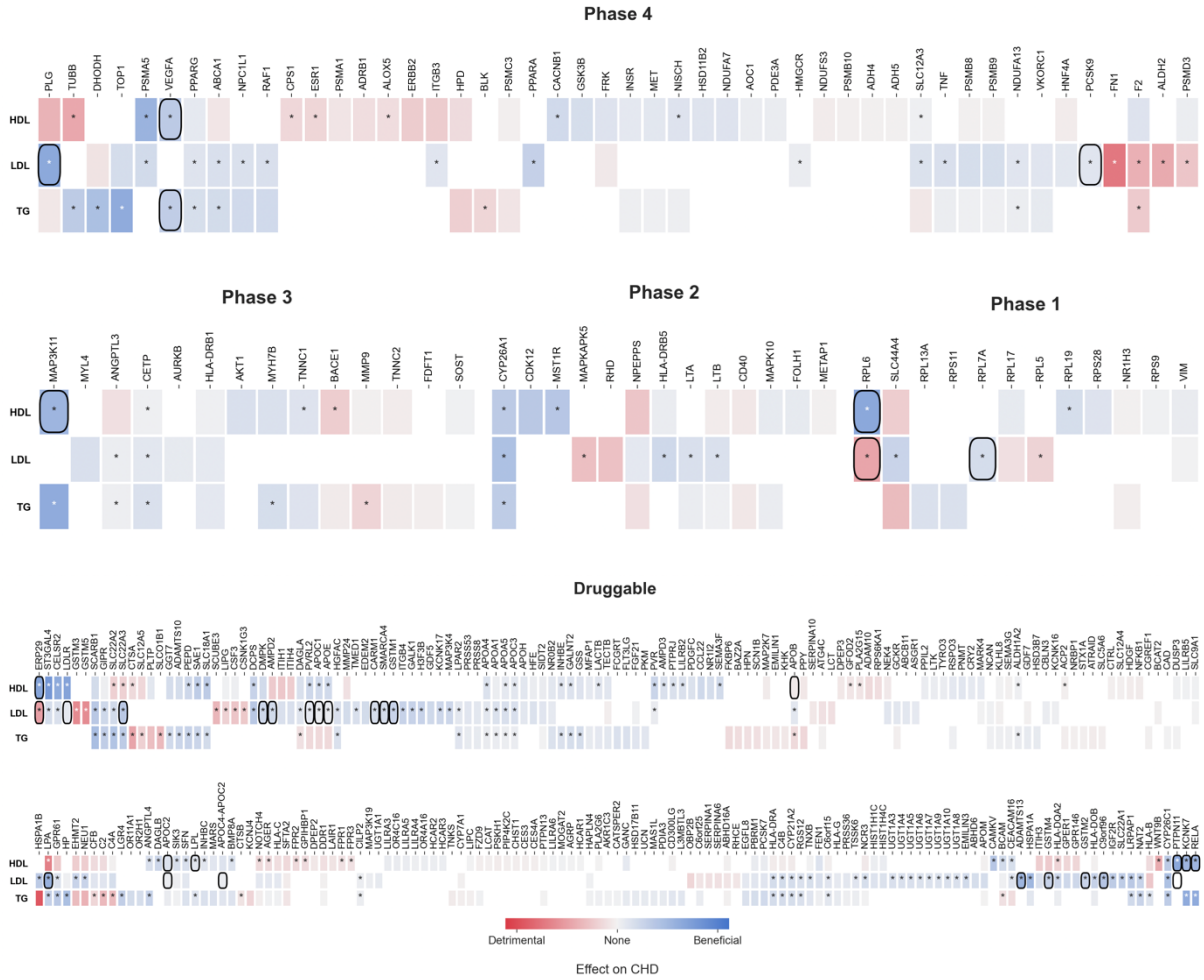


Figure 7.3. Discovery drug target MR estimates on CHD. Analyses were performed using genetic associations with LDL-C, HDL-C and TG from the Global Lipid Genetic Consortium (GLGC) with CHD events from the CardiogramPlusC4D Consortium. Drug targets are grouped by maximum clinical phase according to ChEMBL v25 database. Blue indicates a beneficial effect on CHD risk, and red a detrimental effect per SD difference with respect to the indicated lipid sub-fraction. Significant estimates are indicated with an asterisk (*).

To assess the potential for false positive results, the distribution of the exposure-specific *p* values was tested against the uniform distribution expected under the null hypothesis[15]. The Kolmogorov-Smirnov (KS) goodness-of-fit test was not consistent with the hypothesis that the observed findings could be readily explained by multiple testing (Fig. 7.4).

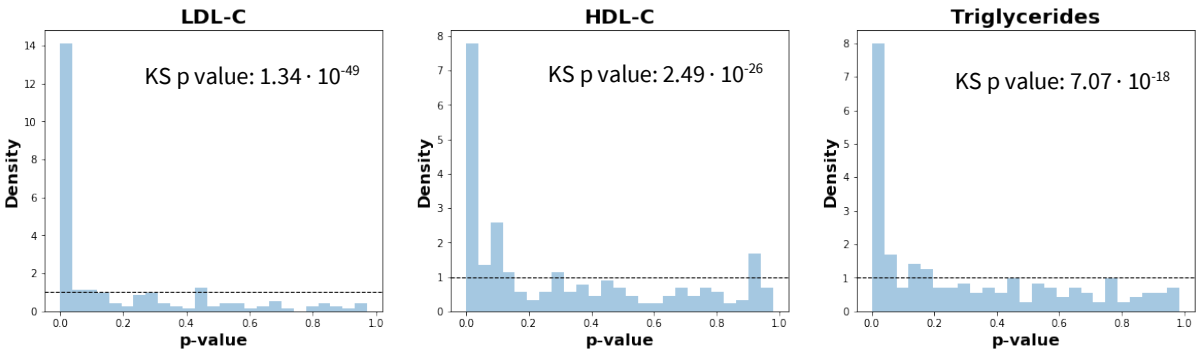


Figure 7.4. Density distribution of the *p* values in the discovery analysis by exposure. Kolmogorov-Smirnov (KS) goodness-of-fit test (two-sided) against the continuous uniform distribution of *p* values (black dashed line) expected under the null-hypothesis of no association between any of the targets and coronary heart disease, when the effect is instrument via LDL-C, HDL-C and TG effects.

7.4.2. Rediscoveries of indications and on-target adverse effects

To investigate if the drug target MR analysis rediscovered the mechanism of action of drugs with a license for lipid modification or compounds with a different indication but with reported lipid-related effects, compounds with reported lipid indications or adverse effects were extracted from the BNF website, which comprises prescribing information for all UK licensed drugs. Out of the 341 druggable genes included in the analysis, five encoded the targets of drugs with a lipid-modifying indication (PCSK9, PPARG, PPARA, NPC1L1, HMGCR) of which NPC1L1, HMGCR and PCSK9 are targets of drugs used in CHD prevention; and 6 encoded a protein target of a drug with reported lipid-related adverse effects (ADRB1, TNF, ESR1, FRK, BLK and DHODH) (Appendix 7.C). To include outcome and side effect data of candidates in clinical phase development, the 341 drug targets were mapped to compound data available in the clinicaltrials.gov database. This database differentiates between endpoints monitored throughout the trial ('outcomes'), and unanticipated harmful episodes during the

study that may be on-target or off-target effects of the trial agent ('adverse events'). Of the 341 drug targets, 23 had reported lipid related outcomes and 40 had reported lipid-related adverse events (Appendix 7.C).

The pool of druggable targets that were modelled using higher LDL-C as a proxy for the pharmacological action on a drug target included 14 targets of clinically used drugs, three of which were licensed for CHD treatment by lowering LDL-C (HMGCR, PCSK9 and NPC1L1). The non-CHD indications of clinically used drugs included dyslipidemias (PPARA), type 2 diabetes (PPARG and NDUFA13), autoimmune diseases (TNF), neoplasms (RAF1 and PSMA5), circulatory disorders (ABCA1, PLG, ITGB3 and F2) and alcohol-dependency (ALDH2) (Table 7.1). With the exception of F2, instrumenting the target action through a higher LDL-C effect was associated with a higher CHD risk. Two drug targets were for compounds already in phase 3 trials for CHD prevention (ANGPTL3 and CETP). Lastly, three targets were in phase 2 trials of compounds developed for other indications (CYP26A1, LTA and LTB). The remaining 82 of the 101 targets had not yet been drugged by compounds in clinical phase development.

When using higher HDL-C as a proxy for pharmacological action, MR of four drug targets with compounds approved for non-CHD indications showed a directionally beneficial effect on CHD (VEGFA, PSMA5, CACNB1 and NISCH), suggesting potential for indication expansion (Table 7.1). Three were targets for drugs approved for non-CHD indications but which showed a potentially detrimental effect direction on CHD when instrumented through increasing HDL-C concentration (ESR1, ALOX5, TUBB). Both CYP26A1 and CETP were associated with lower CHD risk when the effect on CHD was instrumented through an elevation of HDL-C. The remaining 65 of the 74 targets have not yet been drugged by compounds in clinical phase development.

Lastly, the set of druggable targets with compounds developed for non-CHD indications that were modelled using higher TG as a proxy for the pharmacological action on the target included PPARG, DHODH, VEGFA, TOP1, TUBB, NDUFA13, ABCA1, BLK, and F2 (Table 7.1). Of these, instrumenting the CHD effect through higher TG via drug action on BLK or F2 increased CHD risk. For the remaining targets, which included CETP, ANGPTL3 and CYP26A1, instrumenting the target effect through lowering TG levels decreased the risk of CHD, while the remaining 52 of the 64 targets have not been drugged by licensed compounds or clinical candidates yet.

Table 7.1. Univariable drug target MR estimates for drug targets approved for indications other than lipid-lowering.

Drug target gene	LDLC (OR, 95% CI)	HDL -C (OR, 95% CI)	Triglycerides (OR, 95% CI)	Mechanism of action and indication
ESR1	-	2.11 (1.13, 3.93)*	-	AGONIST: Neoplasms, Hypogonadism, Menorrhagia, Primary Ovarian Insufficiency, Acne Vulgaris, Postmenopausal Osteoporosis ANTAGONIST: Breast Neoplasms, Neoplasms MODULATOR: Infertility, Dyspareunia, Breast Neoplasms, Postmenopausal Osteoporosis
TNF	2.03 (1.05, 3.93)*	-	1.21 (0.78, 1.9)	INHIBITOR: Ankylosing Spondylitis, Crohn Disease, Psoriasis, Rheumatoid Arthritis, Colitis, Ulcerative, Psoriatic Arthritis, Immune System Diseases, Juvenile Arthritis
BLK	-	-	0.46 (0.31, 0.7)*	INHIBITOR: Precursor Cell Lymphoblastic Leukemia-Lymphoma, Neoplasms
DHODH	0.66 (0.44, 1.0)	-	7.42 (2.32, 23.71)*	INHIBITOR: Rheumatoid Arthritis, Immune System Diseases, Multiple Sclerosis
PPARG	1.67 (1.04, 2.68)*	0.71 (0.35, 1.48)	2.18 (1.14, 4.15)*	AGONIST: Type 2 Diabetes Mellitus, Diabetes Mellitus, Colitis, Ulcerative, Cardiovascular Diseases
PPARA	3.77 (1.44, 9.85)*	-	-	AGONIST: Cardiovascular Diseases, Hypercholesterolemia, Dyslipidemias
NDUFA13	1.63 (1.13, 2.35)*	-	1.18 (1.0, 1.39)*†	INHIBITOR: Diabetes Mellitus, Type 2 Diabetes Mellitus
ALDH2	0.14 (0.07, 0.29)*	-	-	INHIBITOR: Ectoparasitic Infestations, Alcoholism
NISCH	-	0.57 (0.35, 0.93)*	1.16 (0.31, 4.34)	AGONIST: Hypertension
ABCA1	2.05 (1.34, 3.15)*	1.41 (0.66, 3.0)	2.4 (1.29, 4.49)*	INHIBITOR: Cardiovascular Diseases
F2	0.17 (0.05, 0.59)*	0.57 (0.13, 2.43)	0.35 (0.13, 0.94)*	INHIBITOR: Venous Thrombosis, Thrombosis, Unstable Angina, Thrombocytopenia, Atrial Fibrillation, Embolism, Stroke
TUBB	-	7.56 (1.18, 48.38)*	4.46 (2.13, 9.36)*	INHIBITOR: Breast Neoplasms, Neoplasms, Hodgkin Disease, Large-Cell Anaplastic Lymphoma, Non-Small-Cell Lung Carcinoma, Gout, Familial Mediterranean Fever
VEGFA	-	0.22 (0.15, 0.3)*	4.16 (2.45, 7.08)*†	ANTAGONIST: Retinal Neovascularization INHIBITOR: Diabetic Retinopathy, Retinal Neovascularization, Wet Macular Degeneration, Macular Edema, Colorectal Neoplasms, Neoplasms, Glioblastoma, Renal Cell Carcinoma, Non-Small-Cell Lung Carcinoma, Uterine Cervical Neoplasms
RAF1	2.06 (1.48, 2.86)*	-	2.63 (0.79, 8.83)	INHIBITOR: Neoplasms
PSMA5	2.47 (1.8, 3.39)*†	0.08 (0.02, 0.29)*	-	INHIBITOR: Multiple Myeloma, Neoplasms, Mantle-Cell Lymphoma
ALOX5	-	1.74 (1.18, 2.58)*	-	INHIBITOR: Asthma, Ulcerative Colitis, Rheumatoid Arthritis, Juvenile Arthritis
CACNB1	-	0.38 (0.2, 0.72)*	-	BLOCKER: Cardiovascular Diseases MODULATOR: Fibromyalgia, Seizures, Epilepsy, Neuralgia, Restless Legs Syndrome, Postherpetic Neuralgia
PLG	18.35 (5.47, 61.6)*	5.48 (0.07, 456.86)	0.75 (0.18, 3.14)	ACTIVATOR: Thrombosis, Pulmonary Embolism, Stroke, Myocardial Infarction, Heart Failure, Hepatic Veno-Occlusive Disease INHIBITOR: Hemorrhage, Menorrhagia
ITGB3	1.64 (1.06, 2.52)*	2.79 (0.81, 9.62)	-	INHIBITOR: Thrombosis, Unstable Angina
TOP1	2.3 (0.15, 35.62)	-	16.72 (4.19, 66.8)*	INHIBITOR: Neoplasms

*These drug targets showed lipid records in clinicaltrials.gov and/or the British National Formulary (BNF). * indicates significance in the discovery analysis; † indicates significance in both original and validation study and concordant direction of effect. OR = odds ratio of CHD per 1-standard deviation increase in LDL-C, HDL-C or triglycerides; CI = confidence interval.*

7.4.3. Independent validation of the drug target MR estimates

To help verify the MR findings and reduce the multiple testing burden, an independent two sample drug target MR analysis was conducted using summary statistics from a GWAS of blood lipids measured using an NMR spectroscopy platform[10,17], and genetic associations with CHD risk derived from UK Biobank[11]. The validation analysis identified 47 significant MR estimates (p value < 0.05), of which 39/47 (83%) showed a concordant direction of effect with the initial analysis (Table 7.2) corresponding to 30 drug targets. Replicated targets included the licensed LDL-lowering drug targets PCSK9 and NPC1L1 (Appendix 7.C). While the majority of the replicated drug targets were anticipated to decrease CHD risk when instrumenting their effect through LDL-C concentration based on the univariable results, 9 of the drug targets analysed were significantly associated with lower CHD when the drug target effects were modelled through HDL-C and/or TG (Fig. 7.5).

Table 7.2. Replication of drug target MR findings.

Source of data				
	Lipids measures		**Disease endpoints**	
Discovery	Clinical chemistry *(GLGC, N= 188,578*)		Research-based case ascertainment (*CardiogramPlusC4D, N= 184,305 cases*)	
Replication	Nuclear magnetic resonance (NMR) spectroscopy *(Kettunen et al, 2016, UCLEB Meta-analysis, N=33,029)*		Routine Electronic Health Records *(UK Biobank, N=34,541 cases)*	
Direction of effect				
	LDL-C	**HDL-C**	**Triglycerides**	**Overall**
Concordant	21	6	12	39
Discordant	4	0	4	8

The discovery and replication analyses used different data sources for both exposure and outcome. 145 replication MR analyses were performed in which the gene boundaries included genetic associations exceeding the pre-specified significance threshold (p value ≤ $1x10^{-4}$).

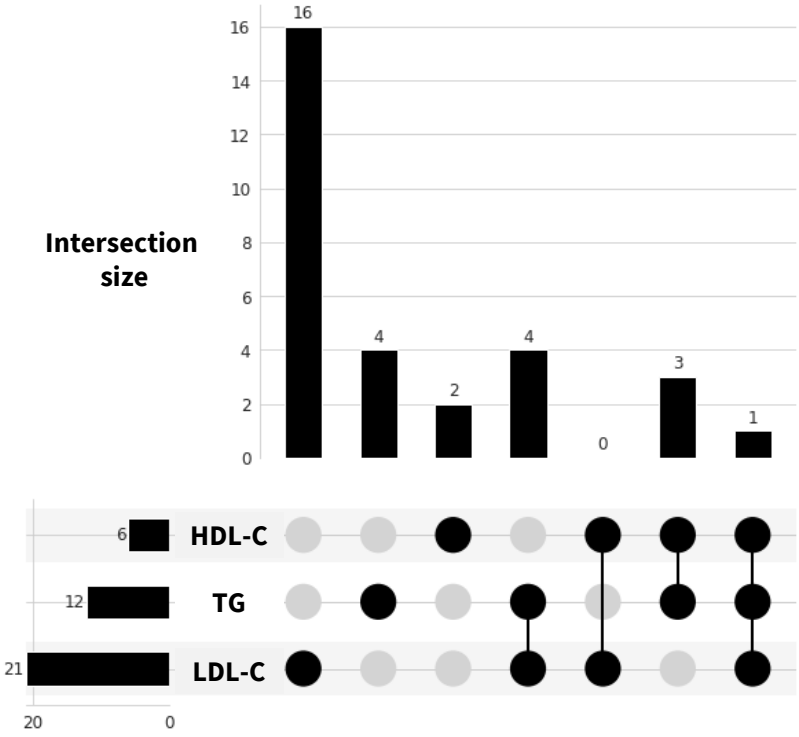


Figure 7.5. The sets of assigned genes associated with LDL-C, HDL-C, TG that encode druggable targets. Genes encoding druggable targets were included if they demonstrated concordant direction of effect in the discovery and validation studies on CHD showing a causal effect of one or more lipid sub-fractions.

7.4.4. Discriminating independent lipid effects using MVMR

After considering each lipid sub-fraction as a single measure on disease risk in the univariable drug target MR analyses, a multivariable drug target MR (MVMR) analysis was performed including LDL-C, HDL-C and TG in a single model to account for potential pleiotropic effects of target perturbation via the other lipid sub-fractions and, in contrast to the previous univariable drug target MR, attempt to directly identify any potential lipid mediating pathway. Twenty-six of the replicated targets had sufficient data (3 or more variants) for the multivariable analysis. This analysis identified a single likely lipid fraction for 12 targets (SLC12A3, APOB, APOA1, PVRL2, APOE, APOC1, CELSR2, GPR61, PCSK9 and

CEACAM16 through LDL-C; LPL through HDL-C; and ALDH1A2 through TG) (Appendix 7.D). It was found that SMARCA4 and APOA5 likely affected CHD through LDL-C and TG, and that RPL7A likely affected CHD through LDL-C and HDL-C pathways. Due to the limited number of variants in VEGFA, CILP2, NDUFA13 and ANGPTL4, multivariable MR analysis could not distinguish the lipid fraction through which CHD was likely affected. Additionally, the presence of horizontal pleiotropy in the MVMR analysis based on heterogeneity tests suggested that PCSK9, LPL, APOC1, APOE, PVRL2, APOB, APOC3, CETP, APOA1 and CELSR2 may affect CHD through additional pathways beyond the lipid sub-fractions LDL-C, HDL-C and TG included in the current model.

7.5. Discussion

7.5.1. Summary

In this chapter, 'biomarker-weighted drug target MR' was used to evaluate the effect of perturbing targets encoded by druggable genes in CHD prevention. By combining publicly available GWAS datasets on blood lipids and coronary heart disease and applying MR approaches with drug information and clinical data, I have genetically validated and prioritised drug targets for CHD prevention. While, as introduced in Chapter 1 and later investigated in Chapter 6, the ideal exposure in a MR analysis for drug target validation are protein activity or levels, restricting the study to those targets with available protein data would have led to a significant reduction in the number of drug targets evaluated. In fact, only 39% (i.e., 133/341) of the 341 druggable genes identified in this analysis had the levels of the encoded protein measured by the largest proteomic platform available (SomaLogic v4). Furthermore, as discussed in Chapter 6, pQTL-weighted drug target MR may be inaccurate in some scenarios, and well-studied alternatives such as 'biomarker-weighted drug target MR' represent an opportunity to evaluate drug targets on a large scale. Therefore, one of the aims of the analysis presented in this chapter was to illustrate that the lack of pQTL data should not be a limitation to perform drug target MR analyses when genetic associations with a downstream biomarker are available.

One hundred thirty one drug target genes associated with CHD risk were identified from a set of 341 druggable genes overlapping associations with one or more of the major blood lipid fractions. The set of targets included NPC1L1, HMGCR and PCSK9, which are known targets of LDL-lowering drugs whose efficacy in CHD prevention has been proven in clinical trials. An independent replication study was performed both to corroborate the targets and the direction of the effects. The findings were replicated in independent datasets (UCLEB

Consortium and UK Biobank) in which lipids were measured using a different platform (NMR spectroscopy in UCLEB) and the disease endpoints ascertained by linkage to routinely recorded health data (UK Biobank). The validation study replicated 83% (39/47) of the initial estimates, including the mechanism of action of current lipid-modifying drug targets PCSK9 and NPC1L1 and the suggested mechanism of action of compounds under investigation for lipid modification through TG or HDL-C, such as CETP inhibitors[18,19].

It is essential to highlight that, while the drug target analysis uses genetic associations with these lipid sub-fractions as weights, the inference throughout has been on the therapeutic relevance of perturbing the proteins encoded by the corresponding genes which are the main category of molecular target for drug action. The genetic associations with the corresponding lipids are merely used as a proxy for protein activity and/or concentration, serving to orientate the MR effects in the direction of a therapeutic effect. They do not provide comprehensive evidence on the pathway through which perturbation of such targets causally affects CHD. Nevertheless, multivariable MR does provide insight on the potential relevance of lipid pathways in mediating the effects of drug target perturbation. In general, results that do not meet the significance threshold should not be over-interpreted as proof of absence of effect[20]. This may be exacerbated here by potential weak instrument bias, which will be expected to attenuate results towards the no-effect direction.

7.5.2. Research in context

In addition to the known lipid-modifying drug targets PCSK9 and NPC1L1, the set of 30 replicated drug targets also included lipoprotein lipase (LPL), a target that could potentially decrease CHD risk based on the univariable MR findings, with an effect through HDL-C further endorsed by the multivariable MR analyses (Fig. 7.6). In contrast to current lipid-

lowering drug targets which are specifically expressed in the liver, LPL shows highest specific expression in adipose tissue which suggests tissues beyond the liver may be relevant to target lipid metabolism. Several pharmacological attempts have been pursued to target LPL[21,22], and gene therapy has also been applied to treat LPL deficiency by introducing extra copies of the functional enzyme in patients with hypertriglyceridemia[23]. The approval of gene therapy interventions and the known indirect activation of LPL by drugs targeting other proteins, such as fibrates[24] and metformin[25], suggest that the previous failure of compounds targeting LPL in initial trials may have been idiosyncratic. LPL activity is also modulated by another protein in the replicated dataset, apolipoprotein A5 (ApoA5), which is exclusively expressed in liver tissue. The multivariable MR suggest that ApoA5 (partially) affects CHD through LDL-C and TG-mediated pathways. Regardless of the mediating lipid or lipids, the genetic findings in relation to both LPL and ApoA5 are consistent and point to this as an important potentially targetable pathway in atherosclerosis, supporting prior work[26].

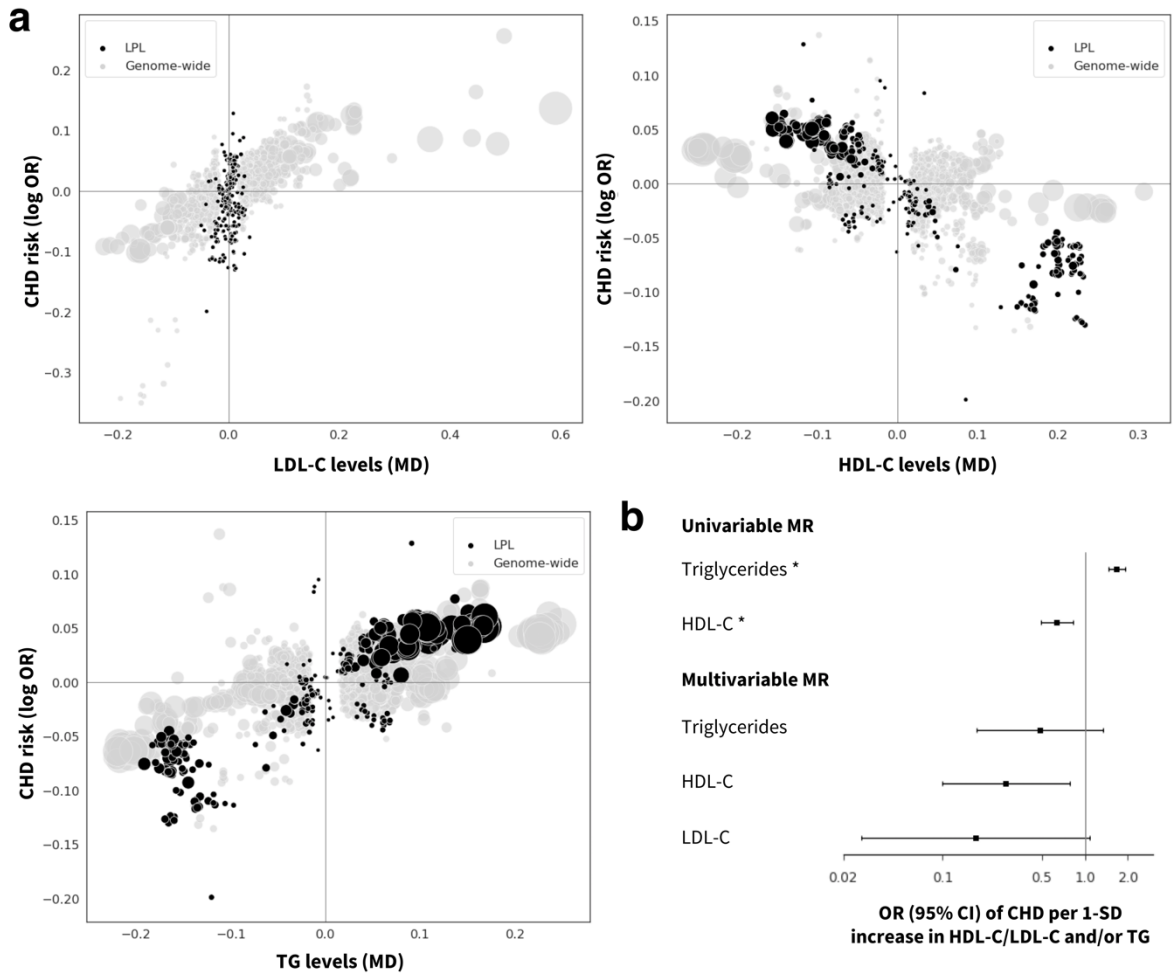


Figure 7.6. Prioritised target: lipoprotein lipase (LPL). a. Genetic associations at the locus (± 50 kbp) in black vs genome-wide associations (grey, *p* value < $1x10^{-6}$ based on two-sided z-tests). The x-axis shows the per allele effect on the corresponding lipid expressed as mean difference (MD) from GLGC and the y-axis indicates the per allele effect on CHD expressed as log odds ratios (OR) from CardiogramPlusC4D. The marker size indicates the significance of the association with the lipid sub-fraction (*p* value). b. Univariable and multivariable (drug target) *cis*-MR results presented as OR and 95% confidence intervals with lipid exposure (n=188,577 individuals) and CHD outcome (n= 60,801 cases and 123,504 controls). An asterisk (*) indicates the MR estimates as being replicated.

This chapter describes and applies an approach to move from GWAS signals to drug targets and disease indications through 'biomarker-weighted drug target MR'. Its potential has been illustrated using genetic association data on lipids and CHD data, but the approach could also be applied in other settings where there are GWAS of diseases and biomarkers thought to be potentially affected by the drug target.

7.5.3. Strengths and limitations

Some limitations of this study are noteworthy. First, only genes regarded as encoding druggable proteins were included, which currently comprise approximately 25% of all protein coding genes[12]. As knowledge advances, additional proteins will become druggable, and alternative therapeutic strategies such as antisense oligonucleotides and gene therapy may extend the range of mechanisms that can be targeted. The approach described here is in fact agnostic to therapeutic modality and could be adapted accordingly. Second, variants were assigned to druggable genes based on genomic proximity, which may be as reliable as other approaches in mapping causal genes[33–35]. However, simple genomic proximity might result in misleading assignment of the causal gene in a region containing multiple genes in high LD (e.g. *PVRL2*, *APOC1* and *APOE* are all located in a region of LD in Chr19:45349432-45422606, GRCh37). In an effort to account for this, all the druggable genes (± 50 kbp) that overlap one of the genetic variants associated with LDL-C, HDL-C or TG were included in the analysis, and information on proximity of the variant to the gene, a gene distance rank value (in base pairs), and previous gene prioritisation data by the Global Lipids Genetics Consortium (GLGC)[7] is also provided to inform scenarios in which the causal gene may be a non-druggable gene but reside in the same region (Appendix 7.A).

Cis-MR was used to evaluate the relevance of each drug target to CHD, which is less prone to violation of the horizontal pleiotropy assumption than MR analyses with *trans* instruments[4], which also require direct measurement of the protein of interest. However, *cis*-MR also requires some decisions to be made regarding instrument selection: defining the locus of interest, the significance threshold for the association with the exposure and the LD threshold to prune correlated instruments. Since an agreement on the choice of a general LD threshold and flanking region has yet to be reached, a window of 50 kbp and LD threshold of 0.4 were used, which showed the most consistent estimates in a grid-search in the discovery data using the four positive control examples: PCSK9, NPC1L1, HMGCR and CETP. Based on previous studies showing that using less stringent *p* value thresholds often results in improved performance in *cis*-MR settings, the threshold below genome-wide significance was relaxed to select the genetic associations to instrument the exposure; and accounted for LD correlation by pruning and LD modelling during the MR analysis[4,36].

Multiple testing in the MR analyses was addressed in a number of complementary ways. To assess the potential for false positive results, the distribution of the exposure-specific *p* values was tested against the uniform distribution expected under the null hypothesis[15]. The Kolmogorov-Smirnov (KS) goodness-of-fit test indicated that the number of extreme *p* values obtained would be highly unlikely under the null hypothesis, suggesting that they are unlikely to represent false positives. Subsequently, the findings were validated with independent data sources and a second drug target MR was conducted, although several drug target genes could not be evaluated in the validation analysis because the gene boundaries did not include genetic associations exceeding the pre-specified significance threshold (p value $\leq 1x10^{-4}$), likely related to the 'modest' sample size of the NMR replication data (N=33,029). By drawing inference on replicated data, the multiple testing burden was considerably reduced

(0.05^2=0.0025), which when applied to 98 drug targets retained after replication would suggest up to one result being a false positive.

Beyond univariable MR analyses, I attempted to further validate the findings with a multivariable extension of the inverse-variance weighted (IVW) and MR Egger methods, however, in some cases imprecise estimates were obtained in line with previous studies which attributed this to the inclusion of highly correlated exposures in the model[37].

The effect directions of the replicated drug targets were compared to results from clinical trials using data from the clinicaltrials.gov registry. However, the lack of precision in annotation of events associated with lipid perturbations (e.g. hyperlipidaemia) in this dataset hinders the assignment of reported lipid abnormalities to a particular lipid sub-fraction. Moreover, the proportion of clinical trials with reported results in clinicaltrials.gov is less than 54.2%[38], suggesting that additional drug candidates with lipid effects might have been investigated but were not included in this analysis because of the lack of accessible data. Furthermore, the analysis relied on mapping clinical trial interventions to compounds known to act through binding to the targets of interest, which could potentially miss clinical trials of compounds annotated with less synonyms (such as research codes for compounds used by individual trial sponsors).

7.6. Conclusion

In Chapter 6, genetic variants in or near a drug target gene that have been associated with the circulating levels of the encoded protein were used to evaluate the performance of the drug target MR framework. Such analysis showed that measured levels are yet not available for several drug target proteins, and even when these have been measured, the genetic associations do not represent valid instruments or the drug target MR framework does not yield the anticipated result. As an alternative to the pQTL-weighted drug target MR, the drug target MR using genetic variants in and around the gene encoding the target protein associated with a downstream biomarker could be used as a proxy for protein concentration or activity, without implying a mediation effect between the biomarker and the disease. As an example, biomarker-weighted drug target MR was applied to a set of 341 drug targets identified through their association with blood lipids (HDL-C, LDL-C and triglycerides), to evaluate their causal relevance for coronary heart disease (CHD). Thirty of these targets were further prioritised including NPC1L1 and PCSK9, the targets of drugs used in CHD prevention. When used as a screening tool, the biomarker-weighted drug target MR could help reduce the high failure rate problem in drug discovery by genetically validating targets in the earlier phases of the drug development pipeline.

7.7. References

1. Gordillo-Marañón, M. *et al.* Validation of lipid-related therapeutic targets for coronary heart disease prevention using human genetics. *Nat Commun* **12**, 6120 (2021).
2. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomisation: using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**, 1133–1163 (2008).
3. Davey Smith, G. & Ebrahim, S. 'Mendelian randomisation': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1–22 (2003).
4. Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nature Communications* **11**, 3255 (2020).
5. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery* **1**, 727–730 (2002).
6. HPS3/TIMI55–REVEAL Collaborative Group *et al.* Effects of Anacetrapib in Patients with Atherosclerotic Vascular Disease. *N. Engl. J. Med.* **377**, 1217–1227 (2017).
7. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274 (2013).
8. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
9. Shah, T. *et al.* Population genomics of cardiometabolic traits: design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PloS one* **8**, e71345 (2013).
10. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122 (2016).

11. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation Research* **122**, 433–443 (2018).
12. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, (2017).
13. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. Bowden, J. *et al.* Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomisation via the Radial plot and Radial regression. *International Journal of Epidemiology* **47**, 1264–1278 (2018).
15. Storey, J. D. A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 479–498 (2002).
16. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940 (2018).
17. Shah, T. *et al.* Population Genomics of Cardiometabolic Traits: Design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PLOS ONE* **8**, e71345 (2013).
18. Hovingh, G. K. *et al.* Cholesterol ester transfer protein inhibition by TA-8995 in patients with mild dyslipidaemia (TULIP): a randomised, double-blind, placebo-controlled phase 2 trial. *Lancet* **386**, 452–460 (2015).
19. Dewey, F. E. *et al.* Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *New England Journal of Medicine* **377**, 211–221 (2017).
20. Alderson, P. Absence of evidence is not evidence of absence. *BMJ* **328**, 476–477 (2004).
21. Tsutsumi, K. *et al.* The novel compound NO-1886 increases lipoprotein lipase activity with resulting elevation of high density lipoprotein cholesterol, and long-term

administration inhibits atherogenesis in the coronary arteries of rats with experimental atherosclerosis. *The Journal of clinical investigation* (1993) doi:10.1172/JCI116582.

22. Yin, W. & Tsutsumi, K. Lipoprotein Lipase Activator NO-1886. *Cardiovascular Drug Reviews* **21**, 133–142 (2003).
23. Gaudet, D., Méthot, J. & Kastelein, J. Gene therapy for lipoprotein lipase deficiency. *Curr. Opin. Lipidol.* **23**, 310–320 (2012).
24. Schoonjans, K. *et al.* PPARalpha and PPARgamma activators direct a distinct tissue-specific transcriptional response via a PPRE in the lipoprotein lipase gene. *EMBO J.* **15**, 5336–5348 (1996).
25. Ohira, M. *et al.* Effect of metformin on serum lipoprotein lipase mass levels and LDL particle size in type 2 diabetes mellitus patients. *Diabetes Research and Clinical Practice* **78**, 34–41 (2007).
26. Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factors Collaboration *et al.* Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. *Lancet* **375**, 1634–1639 (2010).
27. Kinney, J. W. *et al.* Inflammation as a central mechanism in Alzheimer's disease. *Alzheimers Dement (N Y)* **4**, 575–590 (2018).
28. Haddick, P. C. G. *et al.* A common variant of IL-6R is associated with elevated IL-6 pathway activity in Alzheimer's disease brains. *J Alzheimers Dis* **56**, 1037–1054 (2017).
29. Khandaker, G. M., Zammit, S., Burgess, S., Lewis, G. & Jones, P. B. Association between a functional interleukin 6 receptor genetic variant and risk of depression and psychosis in a population-based birth cohort. *Brain Behav. Immun.* **69**, 264–272 (2018).
30. Thibord, F. *et al.* A Genome Wide Association Study on plasma FV levels identified PLXDC2 as a new modifier of the coagulation process. *Journal of Thrombosis and Haemostasis* **17**, 1808–1814 (2019).

31. Weitz, J. I. & Fredenburgh, J. C. Factors XI and XII as Targets for New Anticoagulants. *Front Med (Lausanne)* **4**, (2017).
32. Tillman, B. & Gailani, D. Inhibition of factor XI and factor XII for Prevention of Thrombosis Induced by Artificial Surfaces. *Semin Thromb Hemost* **44**, 60–69 (2018).
33. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, e3–e3 (2019).
34. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomisation studies. *Hum Mol Genet* **27**, R195–R208 (2018).
35. Zheng, J. *et al.* Phenome-wide Mendelian randomisation mapping the influence of the plasma proteome on complex diseases. *Nature Genetics* **52**, 1122–1131 (2020).
36. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
37. Rees, J. M. B., Wood, A. M. & Burgess, S. Extending the MR-Egger method for multivariable Mendelian randomisation to correct for both measured and unmeasured pleiotropy. *Stat Med* **36**, 4705–4718 (2017).
38. Zwierzyna, M., Davies, M., Hingorani, A. D. & Hunter, J. Clinical trial design and dissemination: comprehensive analysis of clinicaltrials.gov and PubMed data since 2005. *BMJ* **361**, (2018).

7.8. Appendices

Appendix 7.A. Proximity to GWAS SNP, protein-coding distance rank and previous evidence of druggable genes near genetic associations with LDL-C, HDL-C and TG from GLGC.

Druggable gene	Genomic coordinates (GRCh37)	Min distance (pb) to LDL-C assoc. (distance rank)	Min distance (pb) to HDL-C assoc. (distance rank)	Min distance (pb) to TG assoc. (distance rank)	Prioritised by GLGC
RHD	1:25598884-25656936	31340 (4)	-	-	No
RHCE	1:25688740-25756683	0 (1)	-	-	No
RPS6KA1	1:26856252-26901521	-	0 (1)	-	No
SFN	1:27189633-27190947	-	45265 (6)	45265 (6)	No
NR0B2	1:27237980-27240457	44738 (4)	44738 (4)	44738 (4)	No
SLC9A1	1:27425306-27493472	-27702 (1)	-	-	No
BMP8A	1:39957318-39991607	-	-7238 (2)	-7238 (2)	No
PCSK9	1:55505221-55530525	0 (1)	-	-	Yes
ANGPTL3	1:63063158-63071830	-49762 (2)	-	-7878 (2)	Yes
ATG4C	1:63249806-63331184	0 (1)	-	0 (1)	No
RPL5	1:93297582-93307481	-27758 (2)	-	-	No
CELSR2	1:109792641-109818372	0 (1)	0 (1)	-	No
PSMA5	1:109941653-109969062	-19276 (2)	-	-	No
GPR61	1:110082494-110091028	-20869 (3)	-	-	No
AMPD2	1:110158726-110174673	-49687 (4)	-	-	No
GSTM4	1:110198703-110208118	-29513 (3)	-	-	No
GSTM2	1:110210644-110252171	44118 (4)	-	-	No
GSTM1	1:110230436-110251661	44628 (5)	-	-	No
GSTM5	1:110254864-110318050	0 (2)	-	-	No
GSTM3	1:110276554-110284384	11905 (3)	-	-	No
CSF1	1:110452864-110473614	-	5297 (1)	-	No
HDGF	1:156711899-156736717	-	-11248 (4)	-	No
GALNT2	1:230193536-230417870	-	0 (1)	0 (1)	Yes
GDF7	2:20866424-20873418	13808 (2)	-	-	No
APOB	2:21224301-21266945	29601 (1)	-28051 (1)	-28051 (1)	Yes
EMILIN1	2:27301435-27309271	-	-	33623 (7)	No
KHK	2:27309615-27323640	-	-	19254 (5)	No
CGREF1	2:27321757-27341995	-	-	899 (1)	No
SLC5A6	2:27422455-27435826	-	-	7370 (3)	No
ATRAID	2:27434895-27440046	-	-	3150 (2)	No
CAD	2:27440258-27466811	-	-	30737 (4)	No
UCN	2:27530268-27531313	-	-	-32720 (5)	No
NRBP1	2:27650657-27665126	-	-	-10332 (2)	No
GCKR	2:27719709-27746554	0 (1)	-	32018 (4)	Yes
MAP3K19	2:135722061-135805038	0 (1)	-	-	No
LCT	2:136545410-136594750	0 (1)	-	-	No
ABCB11	2:169779448-169887832	0 (1)	-	-	Yes
CPS1	2:211342406-211543831	-	0 (1)	-	Yes
FN1	2:216225163-216300895	0 (1)	-	-	Yes
UGT1A8	2:234526291-234681956	0 (9)	-	-	No
UGT1A10	2:234545100-234681951	0 (8)	-	-	No
UGT1A9	2:234580499-234681946	0 (7)	-	-	No
UGT1A7	2:234590584-234681945	0 (4)	-	-	No
UGT1A6	2:234600253-234681946	0 (6)	-	-	No
UGT1A5	2:234621638-234681945	0 (5)	-	-	No
UGT1A4	2:234627424-234681945	0 (3)	-	-	No
UGT1A3	2:234637754-234681945	0 (2)	-	-	No
UGT1A1	2:234668894-234681945	0 (1)	-	-	Yes
PPARG	3:12328867-12475855	0 (1)	0 (1)	13487 (1)	No
RAF1	3:12625100-12705725	0 (1)	-	-	Yes
CAMKV	3:49895421-49907655	-	0 (1)	-	No
MST1R	3:49924435-49941299	-	30215 (4)	-	No
SEMA3F	3:50192478-50226508	-	-20081 (2)	-	No
SEMA3G	3:52467069-52479101	-	-35398 (4)	-	No
TNNC1	3:52485118-52488086	-	4621 (2)	-	No
NISCH	3:52489134-52527087	-	0 (1)	-	No
PBRM1	3:52579368-52719933	-	0 (1)	-	No
NEK4	3:52744800-52804965	-	39569 (8)	-	No
ITIH1	3:52811603-52826078	-	18456 (4)	-	No
ITIH3	3:52828784-52843025	-	1509 (1)	-	No

Druggable gene	Genomic coordinates (GRCh37)	Min distance (pb) to LDL-C assoc. (distance rank)	Min distance (pb) to HDL-C assoc. (distance rank)	Min distance (pb) to TG assoc. (distance rank)	Prioritised by GLGC
ITIH4	3:52846991-52865495	-	-2457 (2)	-	No
ABHD6	3:58223233-58281420	12447 (3)	-	-	No
NR1I2	3:119499331-119537332	-	4965 (2)	-	No
GSK3B	3:119540170-119813264	-	0 (1)	-	Yes
RGS12	4:3294755-3441640	0 (1)	-	31499 (3)	No
HGFAC	4:3443614-3451211	-8729 (2)	-	21928 (2)	No
LRPAP1	4:3508103-3534286	-34964 (4)	-	-34964 (4)	Yes
MAPK10	4:86936276-87515284	-	-	0 (1)	No
PTPN13	4:87515468-87736324	-	32193 (3)	-41692 (2)	No
KLHL8	4:88081255-88161466	-	-28170 (2)	49590 (3)	Yes
HSD17B11	4:88257762-88312538	-	-	-46706 (2)	No
METAP1	4:99916771-99983964	-	30841 (3)	-	No
ADH5	4:99992132-100009952	-	4853 (1)	-	Yes
ADH4	4:100044808-100078949	-	-30003 (2)	-	No
NFKB1	4:103422486-103538459	-	-49200 (2)	-	No
PDGFC	4:157681606-157892546	-	0 (1)	-	No
HMGCR	5:74632154-74657929	-49992 (2)	-	-	Yes
CSNK1G3	5:122847793-122952739	0 (1)	-	-	Yes
HIST1H1C	6:26055968-26056699	36442 (6)	-	-	No
HFE	6:26087509-26098571	0 (1)	-	-	Yes
HIST1H4C	6:26104104-26104518	-10963 (2)	-	-	No
OR11A1	6:29393281-29424848	-	-	17853 (3)	No
OR2H1	6:29424958-29432105	-	-	10596 (1)	No
MAS1L	6:29454474-29455738	-	-	-11773 (2)	No
HLA-G	6:29794744-29798902	-	-	21684 (1)	No
TUBB	6:30687978-30693203	-	-	15752 (3)	No
DDR1	6:30844198-30867933	-	-	-45501 (1)	No
SFTA2	6:30899130-30899952	-	-	20172 (2)	No
C6orf15	6:31079000-31080336	49371 (8)	25077 (6)	32878 (7)	No
HLA-C	6:31236526-31239907	32354 (1)	-	25632 (1)	No
HLA-B	6:31321649-31324965	29139 (2)	29139 (2)	-49933 (2)	No
LTA	6:31539831-31542101	-27032 (6)	-	930 (2)	No
TNF	6:31543344-31546113	-30545 (7)	-	-313 (1)	No
LTB	6:31548302-31550299	-35503 (9)	-	45839 (12)	No
NCR3	6:31556672-31560762	-43873 (11)	41081 (10)	35376 (7)	No
APOM	6:31620193-31625987	-	-18350 (4)	39207 (17)	No
ABHD16A	6:31654726-31671221	-	-35150 (11)	0 (1)	No
C6orf25	6:31686371-31694491	-	-	-21177 (8)	No
HSPA1A	6:31783291-31785723	-	-	22713 (5)	No
HSPA1B	6:31795512-31798031	48981 (9)	-	39246 (7)	No
NEU1	6:31825436-31830683	17537 (4)	-	20671 (5)	No
SLC44A4	6:31830969-31846823	1397 (2)	-	4531 (2)	No
EHMT2	6:31847536-31865464	0 (1)	-	0 (1)	No
C2	6:31865562-31913449	15565 (7)	-	-14208 (3)	No
CFB	6:31895475-31919861	9153 (4)	-	-44121 (8)	No
C4A	6:31949801-31970458	-20787 (8)	-	-20787 (8)	No
C4B	6:31982539-32003195	-35079 (10)	-	4264 (3)	No
CYP21A2	6:32006042-32009447	36828 (3)	-	0 (1)	No
TNXB	6:32008931-32083111	0 (1)	-	-1472 (2)	No
EGFL8	6:32132360-32136058	-	-	48287 (8)	No
AGER	6:32148745-32152101	-	37740 (4)	32244 (4)	No
NOTCH4	6:32162620-32191844	-	0 (1)	0 (1)	No
HLA-DRA	6:32407619-32412823	-48188 (3)	-44404 (3)	0 (1)	No
HLA-DRB5	6:32485120-32498064	-	-40922 (2)	-	No
HLA-DRB1	6:32546546-32557625	42432 (3)	18358 (1)	42432 (3)	No
HLA-DQA2	6:32709119-32714992	-37979 (2)	-39746 (2)	-24862 (1)	No
HLA-DOB	6:32780540-32784825	0 (1)	-	-25250 (2)	No
PSMB8	6:32808494-32812480	-23874 (4)	-	-	No
PSMB9	6:32811913-32827362	-27293 (5)	-	-	No
SCUBE3	6:35182190-35220856	-49116 (2)	-20049 (1)	-	No
KCNK17	6:39266777-39282329	-15940 (1)	-	-	Yes
KCNK16	6:39282474-39290744	-31637 (2)	-	-	No
VEGFA	6:43737921-43754224	-	10327 (1)	10327 (1)	Yes
FRK	6:116252312-116381921	0 (1)	-	-	Yes
RSPO3	6:127439749-127518910	-	2217 (1)	-47328 (1)	Yes
L3MBTL3	6:130334844-130462594	-	-	0 (1)	No
ESR1	6:151977826-152450754	-	0 (1)	-	No
IGF2R	6:160390131-160534539	8609 (2)	-	-	No
SLC22A1	6:160542821-160579750	0 (1)	-	-	No
SLC22A2	6:160592093-160698670	19974 (1)	69 (1)	-	No
SLC22A3	6:160769300-160876014	-46295 (2)	-32178 (1)	0 (1)	No
LPA	6:160952515-161087407	0 (1)	0 (1)	-45381 (2)	Yes

Druggable gene	Genomic coordinates (GRCh37)	Min distance (pb) to LDL-C assoc. (distance rank)	Min distance (pb) to HDL-C assoc. (distance rank)	Min distance (pb) to TG assoc. (distance rank)	Prioritised by GLGC
PLG	6:161123270-161174347	-40471 (2)	-40809 (2)	-	No
MAP3K4	6:161412759-161538417	-21972 (1)	-	-	No
GPR146	7:1084212-1098897	-	-435 (2)	-	Yes
GPER1	7:1121844-1133451	-	-38067 (3)	-	No
DAGLB	7:6448757-6523821	-	-13853 (2)	-	Yes
NPC1L1	7:44552134-44580914	41372 (4)	-	-	Yes
FKBP6	7:72742167-72772634	-	-	0 (1)	No
FZD9	7:72848109-72850450	-	5980 (2)	32656 (2)	No
STX1A	7:73113536-73134002	-	-	-53 (2)	No
MET	7:116312444-116438440	-	-	0 (1)	Yes
AOC1	7:150521715-150558592	-	0 (1)	-	No
TNKS	8:9413424-9639856	-	-	0 (1)	No
BLK	8:11351510-11422113	-	-	0 (1)	No
FDFT1	8:11653082-11696818	-	-	-36672 (4)	No
CTSB	8:11700033-11726957	-	-	-10805 (2)	No
NAT2	8:18248755-18258728	13710 (1)	-	9752 (1)	Yes
LPL	8:19759228-19824769	-	46744 (1)	46744 (1)	Yes
SLC18A	8:20002366-20040717	-	-46760 (1)	-41092 (1)	Yes
CYP7A1	8:59402737-59412795	-4276 (1)	-	-49203 (2)	Yes
GPIHBP1	8:144295068-144299044	-	-49193 (2)	-	Yes
ABCA1	9:107543283-107690518	0 (1)	0 (1)	0 (1)	Yes
OBP2B	9:136080664-136084630	47720 (1)	-	-	No
RPL7A	9:136215069-136218281	-24059 (3)	-	-	No
C9orf96	9:136243117-136271220	0 (1)	-	-	No
ADAMTS13	9:136279478-136324508	1740 (2)	-	-	No
AKR1C3	10:5077546-5149878	-	-	46395 (3)	No
VIM	10:17270258-17279592	-9968 (1)	-	-	No
ALOX5	10:45869661-45941561	-	0 (1)	-	No
CYP26C1	10:94821021-94828454	-	-15356 (2)	-15356 (2)	No
CYP26A1	10:94833232-94837647	-	-27567 (3)	-27567 (3)	Yes
TECTB	10:114043493-114064793	-	0 (1)	0 (1)	No
ADRB1	10:115803806-115806667	-	-11019 (1)	-	No
AMPD3	11:10329860-10529126	-	0 (1)	-	Yes
PSMA1	11:14515329-14665181	-	-10866 (2)	-	No
LGR4	11:27387508-27494322	-	-	31783 (2)	No
CHST1	11:45670427-45687172	-	47215 (1)	-	No
CRY2	11:45868669-45904798	-	-28960 (2)	-	No
F2	11:46740730-46761056	-	42729 (2)	-18509 (3)	No
ACP2	11:47260853-47270457	-	40910 (3)	-34365 (4)	No
NR1H3	11:47269851-47290396	-	20971 (2)	-43363 (6)	No
PSMC3	11:47440320-47447993	-	1551 (1)	-	No
NDUFS3	11:47586888-47606114	-	-3767 (2)	-	No
PTPRJ	11:48002113-48189670	-	0 (1)	-	No
FOLH1	11:49168187-49230222	-	0 (1)	-	No
OR4A16	11:55110627-55111707	-	12013 (2)	-	No
OR4C16	11:55339604-55340536	-	-15296 (2)	-	No
DAGLA	11:61447905-61514473	49826 (6)	49826 (6)	49826 (6)	No
FEN1	11:61560109-61564716	15044 (3)	15044 (3)	15044 (3)	No
KCNK7	11:65360326-65363467	-	27850 (4)	-	No
MAP3K11	11:65365226-65382853	-	8464 (2)	-	No
RELA	11:65421067-65430565	-	-29750 (5)	-	No
MOGAT2	11:75428864-75444003	-	11018 (1)	-	No
APOA5	11:116660083-116663136	-4483 (2)	-4483 (2)	-43681 (3)	No
APOA4	11:116691419-116694022	-35819 (4)	-35819 (4)	-35819 (4)	No
APOC3	11:116700422-116703788	-44822 (5)	-44822 (5)	-44822 (5)	No
APOA1	11:116706467-116708666	-42616 (6)	-38922 (6)	-42616 (6)	Yes
SIK3	11:116714118-116969153	0 (1)	0 (1)	-46573 (7)	No
SIDT2	11:117049449-117068160	7406 (3)	-3252 (2)	-23075 (2)	No
PCSK7	11:117075053-117103241	0 (1)	-28856 (4)	-48679 (4)	No
BACE1	11:117156402-117186975	-	0 (1)	-4990 (2)	No
DCPS	11:126173647-126215644	26052 (2)	12356 (2)	-	No
ST3GAL4	11:126225535-126310239	0 (1)	0 (1)	-	Yes
PDE3A	12:20522179-20837315	-	-48421 (1)	-	Yes
SLCO1B1	12:21284136-21392180	-	-	0 (1)	No
BAZ2A	12:56989380-57030600	-	-	-38161 (2)	No
INHBC	12:57828543-57844611	-	0 (1)	0 (1)	No
INHBE	12:57846106-57853063	-	-2057 (2)	-2057 (2)	No
MARS	12:57869228-57911352	-	-25179 (6)	-25179 (6)	No
PIP4K2C	12:57984957-57997198	-	15035 (5)	-	No
ALDH2	12:112204691-112247782	37809 (2)	-	-	No
MAPKAPK5	12:112279782-112334343	5500 (1)	-	-	No
ERP29	12:112451120-112461255	-10389 (2)	25563 (2)	-	No

Druggable gene	Genomic coordinates (GRCh37)	Min distance (pb) to LDL-C assoc. (distance rank)	Min distance (pb) to HDL-C assoc. (distance rank)	Min distance (pb) to TG assoc. (distance rank)	Prioritised by GLGC
RPL6	12:112842994-112856642	-29579 (2)	49773 (2)	-	No
PTPN11	12:112856155-112947717	0 (1)	0 (1)	-	No
HPD	12:122277433-122301502	-	-	-28342 (4)	No
HCAR1	12:123104824-123215390	-	0 (2)	-	No
HCAR2	12:123185840-123187890	-	12878 (3)	-	No
HCAR3	12:123199303-123201439	-	0 (1)	-	No
SCARB1	12:125261402-125367214	-	0 (1)	-	Yes
CBLN3	14:24895738-24900160	-12108 (2)	-	-	No
SERPINA10	14:94749650-94759608	35884 (2)	-	-	No
SERPINA6	14:94770585-94789731	5761 (1)	-	-	No
SERPINA1	14:94843084-94857030	-47592 (3)	-	-	No
AKT1	14:105235686-105262088	-	15121 (2)	-	No
LTK	15:41795836-41806085	-	23145 (3)	-	No
TYRO3	15:41849873-41871536	-	-20643 (2)	-	No
GANC	15:42565431-42645864	-	37923 (3)	37923 (3)	No
CATSPER2	15:43920701-43960316	-	-	-26883 (4)	No
PDIA3	15:44038590-44065477	-	-	-22173 (2)	No
MFAP1	15:44096690-44117000	-	-	35817 (3)	No
ALDH1A2	15:58245622-58790065	-	0 (1)	0 (1)	No
LIPC	15:58702768-58861151	-	-21963 (2)	-22125 (2)	Yes
ADAM10	15:58887403-59042177	-	-34294 (2)	-	No
LACTB†	15:63413999-63434260	-	-498 (1)	-498 (1)	Yes
PKM	15:72491370-72524164	-	-	42451 (4)	No
HSD3B7	16:30996519-31000473	-	-	47606 (8)	No
PRSS53	16:31094746-31100949	-	-	41044 (7)	No
VKORC1	16:31102163-31107301	-	-	34692 (5)	No
PRSS8	16:31142756-31147083	-	-	-763 (2)	No
PRSS36	16:31150246-31161415	-	-	-8253 (3)	No
SLC12A3	16:56899119-56949762	43263 (4)	43263 (4)	37253 (4)	No
CETP	16:56995762-57017757	-2737 (1)	-2737 (1)	-8747 (1)	Yes
CCL22	16:57392684-57400102	-	-38750 (2)	-	No
CES3	16:66995140-67009051	-	42996 (3)	-	No
CES4A	16:67022492-67043661	-	8386 (1)	-	No
HSD11B2	16:67464555-67471456	-	-45403 (4)	-	No
AGRP	16:67516474-67517716	-	37623 (2)	-	No
GFOD2	16:67708434-67753324	-	5454 (2)	-	No
PSKH1	16:67927175-67963581	-	-42556 (7)	-	No
CTRL	16:67961543-67966317	-	27326 (6)	-	No
PSMB10	16:67968405-67970990	-	22653 (4)	-	No
LCAT	16:67973653-67978034	-	46961 (6)	-	Yes
SLC12A4	16:67977377-68003504	-	21491 (4)	-	No
DPEP3	16:68009566-68014732	-	10263 (3)	-	No
DPEP2	16:68021297-68034489	-	0 (2)	-	No
PLA2G15	16:68279207-68294961	-	0 (1)	-	No
DHODH	16:72042487-72058954	-34262 (1)	-	49139 (6)	No
HP	16:72088491-72094954	0 (2)	-	13139 (3)	No
ASGR1	17:7076750-7082883	6040 (2)	-	-	No
AURKB	17:8108056-8113918	47231 (5)	-	-	No
CACNB1	17:37329709-37353956	-	35453 (4)	-	No
RPL19	17:37356536-37360980	-	28429 (3)	-	No
CDK12	17:37617764-37721160	-	18114 (1)	-	No
PNMT	17:37824234-37826728	-	31950 (4)	-	No
ERBB2	17:37844167-37886679	-	0 (1)	-	No
PSMD3	17:38137050-38154213	-	-15057 (2)	-	No
CSF3	17:38171614-38174066	-	-49621 (6)	-	No
SOST	17:41831099-41836156	-	-	42010 (4)	No
DUSP3	17:41843489-41856356	-	-	21810 (3)	No
CD300LG	17:41924516-41940997	-	0 (1)	0 (1)	No
PPY	17:42018172-42019836	-	-39416 (5)	-	No
WNT9B	17:44910567-44964096	-	46625 (4)	-	No
MYL4	17:45277812-45301045	12133 (1)	-	-	No
ITGB3	17:45331212-45421658	3457 (2)	-	-	No
NPEPPS	17:45600308-45700642	49954 (3)	-	49954 (3)	No
APOH	17:64208151-64252643	0 (1)	-	-	No
ITGB4	17:73717408-73753899	28292 (4)	-	-	No
GALK1	17:73747675-73761792	20399 (3)	-	-	No
H3F3B	17:73772515-73781974	217 (2)	-	-	No
RPL17	18:47014851-47018906	-	10670 (1)	-	No
LIPG	18:47087069-47119272	37732 (1)	40585 (1)	-	Yes
INSR	19:7112266-7294045	-	-	0 (1)	Yes
MAP2K7	19:7968728-7979363	-	0 (1)	-	No
NDUFA7	19:8373490-8386280	-	46916 (6)	-	No

Druggable gene	Genomic coordinates (GRCh37)	Min distance (pb) to LDL-C assoc. (distance rank)	Min distance (pb) to HDL-C assoc. (distance rank)	Min distance (pb) to TG assoc. (distance rank)	Prioritised by GLGC
RPS28	19:8386042-8388224	-	44972 (5)	-	No
ANGPTL4	19:8428173-8439257	-	2522 (1)	-	Yes
ADAMTS10	19:8645126-8675620	-	-34232 (3)	-34232 (3)	No
TMED1	19:10943114-10946994	44833 (4)	-	-	No
CARM1	19:10982189-11033453	0 (1)	-	-	No
SMARCA4	19:11071598-11176071	0 (1)	-	-	No
LDLR	19:11200038-11244492	0 (1)	-	-	Yes
NCAN	19:19322782-19363042	20713 (4)	-	20713 (4)	No
HAPLN4	19:19366450-19373605	10150 (3)	-	10150 (3)	No
TSSK6	19:19623227-19626838	30794 (6)	-	30794 (6)	No
NDUFA13	19:19626545-19644285	13347 (4)	-	13347 (4)	No
CILP2	19:19649057-19657468	164 (1)	-	164 (1)	Yes
LPAR2	19:19734477-19739739	-12501 (2)	-	-12501 (2)	No
PEPD	19:33877856-34012700	-	0 (1)	0 (1)	Yes
SCN1B	19:35521588-35531352	-	-	25392 (2)	No
HPN	19:35531410-35557475	-	-	0 (1)	No
PVR	19:45147098-45166850	28463 (3)	-	-	No
CEACAM16	19:45202421-45213986	-7108 (1)	33641 (3)	33641 (3)	No
BCAM	19:45312328-45324673	4541 (1)	4541 (1)	46165 (4)	No
PVRL2	19:45349432-45392485	3134 (2)	3134 (2)	3134 (2)	No
APOE	19:45409011-45412650	-13392 (3)	-13392 (3)	-13392 (3)	Yes
APOC1	19:45417504-45422606	-21885 (4)	-21885 (4)	-21885 (4)	No
APOC4-APOC2	19:45445495-45452822	-49876 (5)	-49876 (5)	-49876 (5)	No
APOC2	19:45449243-45452822	-47577 (7)	-49899 (7)	37606 (5)	No
MARK4	19:45582546-45808541	0 (2)	-	-	No
GIPR	19:46171502-46186982	20828 (4)	-	-	No
DMPK	19:46272975-46285810	-31643 (4)	-	-	No
SAE1	19:47616531-47713886	-	-26636 (2)	-26636 (2)	No
FGF21	19:49258816-49261587	-44542 (7)	-	-	No
BCAT2	19:49298319-49314286	-48080 (7)	-	-	No
FLT3LG	19:49977464-49989488	-	-	38675 (7)	No
RPL13A	19:49990811-49995565	-	-	32598 (6)	No
RPS11	19:49999622-50002946	-	-	25217 (4)	No
FCGRT	19:50010073-50029590	-	-	0 (1)	No
FPR1	19:52248425-52307363	-	15725 (2)	-	No
FPR2	19:52255279-52273779	-	49309 (4)	-	No
FPR3	19:52298416-52329442	-	0 (1)	-	No
RPS9	19:54704610-54752862	-	39907 (5)	-	No
LILRA6	19:54720737-54746649	-	46120 (6)	-	No
LILRB5	19:54754263-54761164	-	31605 (4)	-	No
LILRB2	19:54777675-54785039	-	42441 (5)	-	No
LILRA3	19:54799854-54809952	-	17528 (3)	-	Yes
LILRA5	19:54818353-54824409	-	3071 (1)	-	No
LILRA4	19:54844456-54850421	-	-16976 (2)	-	No
LAIR1	19:54865362-54882165	-	-37882 (4)	-	No
GGT7	20:33432523-33460663	-	-23173 (2)	-	No
GSS	20:33516236-33543620	-	0 (1)	0 (1)	No
MYH7B	20:33563206-33590240	-	-33440 (3)	-37799 (3)	No
EDEM2	20:33703167-33865928	-	0 (1)	-	No
PROCR	20:33759876-33765165	-	12818 (2)	-	No
MMP24	20:33814457-33864801	-	-36474 (3)	-	No
GDF5	20:34021145-34042568	3933 (2)	-	-	No
TOP1	20:39657458-39753127	0 (1)	-	-	Yes
EMILIN3	20:39988606-39995467	-20655 (2)	-	-	No
HNF4A	20:42984340-43061485	0 (1)	0 (1)	-	Yes
TNNC2	20:44451853-44462384	-	-24435 (4)	-24435 (4)	No
CTSA	20:44518783-44527459	-	4796 (2)	20734 (3)	No
PLTP	20:44527399-44540794	-	0 (1)	7399 (1)	Yes
MMP9	20:44637547-44645200	-	0 (1)	0 (1)	No
SLC12A5	20:44650356-44688784	-	-5391 (2)	-5391 (2)	No
CD40	20:44746911-44758502	-	-12540 (1)	-	No
PPIL2	22:22006559-22054304	-	-23667 (5)	-	No
PLA2G6	22:38507502-38601697	-	-	0 (1)	Yes
KCNJ4	22:38822332-38851205	-	-	46846 (4)	No
PPARA	22:46546424-46639653	0 (1)	-	-	Yes

For each druggable gene included in the analysis, the minimum distance from the gene to the variant (variants located within a gene were given a distance of 0bp and distance to variants upstream the gene are indicated with a negative value), a gene distance rank value according to their base pair distance, and indicated the druggable genes prioritized by GLGC are provided. OR = odds ratio per 1-SD increase in LDL-C/HDL-C or triglycerides; CI = confidence interval.

Appendix 7.B. Univariable drug target MR estimates in the discovery analysis.

Druggable gene	Genomic coordinates	LDL-C (OR, 95% CI)	HDL-C (OR, 95% CI)	Triglycerides (OR, 95% CI)
ABCA1	chr9:107543283-107690518	2.05 (1.34, 3.15)*†	1.41 (0.66, 3.0)	2.4 (1.29, 4.49)*
ABCB11	chr2:169779448-169887832	1.51 (0.7, 3.25)	-	-
ABHD16A	chr6:31654726-31671221	0.7 (0.14, 3.54)	1.06 (0.21, 5.4)	0.94 (0.33, 2.68)
ABHD6	chr3:58223233-58281420	2.25 (0.87, 5.87)	-	-
ACP2	chr11:47260853-47270457	-	1.1 (1.0, 1.2)*	0.86 (0.58, 1.26)
ADAM10	chr15:58887403-59042177	-	1.87 (0.94, 3.75)	-
ADAMTS10	chr19:8645126-8675620	-	0.43 (0.16, 1.17)	3.1 (1.21, 7.98)*
ADAMTS13	chr9:136279478-136324508	11.18 (4.37, 28.59)*†	-	-
ADH4	chr4:100044808-100078949	-	1.05 (0.39, 2.85)	-
ADH5	chr4:99992132-100009952	-	1.05 (0.39, 2.85)	-
ADRB1	chr10:115803806-115806667	-	1.67 (0.58, 4.8)	-
AGER	chr6:32148745-32152101	1.47 (0.71, 3.06)	1.82 (1.08, 3.04)*	1.07 (0.15, 7.64)
AGRP	chr16:67516474-67517716	-	0.98 (0.66, 1.45)	-
AKR1C3	chr10:5077546-5149878	-	-	1.05 (0.49, 2.25)
AKT1	chr14:105235686-105262088	-	0.49 (0.18, 1.36)	-
ALDH1A2	chr15:58245622-58790065	-	0.89 (0.81, 0.99)*	1.28 (1.07, 1.54)*†
ALDH2	chr12:112204691-112247782	0.14 (0.07, 0.29)*	-	-
ALOX5	chr10:45869661-45941561	-	1.74 (1.18, 2.58)*	-
AMPD2	chr1:110158726-110174673	2.51 (1.53, 4.11)*	2.97 (0.88, 10.06)	-
AMPD3	chr11:10329860-10529126	-	0.5 (0.27, 0.92)*	-
ANGPTL3	chr1:63063158-63071830	1.21 (1.11, 1.33)*	1.61 (0.52, 5.01)	1.16 (1.08, 1.25)*
ANGPTL4	chr19:8428173-8439257	-	0.48 (0.28, 0.83)*†	3.38 (1.02, 11.22)*†
AOC1	chr7:150521715-150558592	-	0.81 (0.46, 1.41)	-
APOA1	chr11:116706467-116708666	1.88 (1.49, 2.36)*†	0.84 (0.63, 1.11)	1.25 (1.12, 1.4)*†
APOA4	chr11:116691419-116694022	1.51 (1.23, 1.86)*†	0.53 (0.38, 0.74)*	1.27 (1.14, 1.43)*†
APOA5	chr11:116660083-116663136	2.05 (1.4, 3.02)*†	0.72 (0.6, 0.87)*†	1.21 (1.12, 1.31)*†
APOB	chr2:21224301-21266945	1.5 (1.18, 1.9)*†	1.23 (0.72, 2.12)	0.53 (0.29, 0.98)*†
APOC1	chr19:45417504-45422606	1.31 (1.22, 1.41)*†	0.39 (0.25, 0.59)*	0.51 (0.17, 1.47)
APOC2	chr19:45449243-45452822	1.2 (0.87, 1.66)	0.55 (0.26, 1.14)	1.29 (0.31, 5.39)
APOC3	chr11:116700422-116703788	2.04 (1.72, 2.42)*†	0.67 (0.58, 0.78)*	1.26 (1.12, 1.41)*†
APOC4-APOC2	chr19:45445495-45452822	1.18 (0.86, 1.63)	0.54 (0.26, 1.12)	1.66 (0.9, 3.07)
APOE	chr19:45409011-45412650	1.3 (1.2, 1.41)*†	0.39 (0.26, 0.59)*	0.5 (0.17, 1.45)
APOH	chr17:64208151-64252643	1.52 (0.76, 3.02)	0.66 (0.29, 1.53)	-
APOM	chr6:31620193-31625987	2.32 (0.82, 6.58)	1.06 (0.21, 5.4)	0.96 (0.36, 2.6)
ASGR1	chr17:7076750-7082883	1.39 (0.55, 3.49)	-	-
ATG4C	chr1:63249806-63331184	0.64 (0.31, 1.33)	-	0.94 (0.55, 1.61)
ATRAID	chr2:27434895-27440046	-	-	0.85 (0.71, 1.02)
AURKB	chr17:8108056-8113918	1.35 (0.64, 2.84)	-	-
BACE1	chr11:117156402-117186975	-	2.72 (1.68, 4.39)*	0.82 (0.04, 16.33)
BAZ2A	chr12:56989380-57030600	-	-	0.56 (0.2, 1.63)
BCAM	chr19:45312328-45324673	1.09 (0.71, 1.69)	0.4 (0.18, 0.87)*	0.63 (0.4, 0.97)*
BCAT2	chr19:49298319-49314286	0.94 (0.48, 1.83)	-	-
BLK	chr8:11351510-11422113	-	-	0.46 (0.31, 0.7)*
BMP8A	chr1:39957318-39991607	-	0.52 (0.4, 0.69)*	1.8 (0.6, 5.46)
C2	chr6:31865562-31913449	1.62 (0.93, 2.8)	1.85 (0.66, 5.17)	0.21 (0.07, 0.6)*
C4A	chr6:31949801-31970458	2.26 (0.88, 5.85)	1.04 (0.21, 5.12)	0.22 (0.08, 0.65)*
C4B	chr6:31982539-32003195	2.41 (1.6, 3.63)*	1.23 (0.33, 4.63)	2.28 (1.11, 4.68)*
C6orf15	chr6:31079000-31080336	1.62 (1.23, 2.14)*	1.08 (0.43, 2.72)	1.5 (1.0, 2.25)*
C6orf25	chr6:31686371-31694491	0.7 (0.14, 3.54)	-	1.26 (0.43, 3.69)
C9orf96	chr9:136243117-136271220	5.77 (2.71, 12.31)*†	-	-
CACNB1	chr17:37329709-37353956	-	0.38 (0.2, 0.72)*	-
CAD	chr2:27440258-27466811	-	-	1.01 (0.85, 1.19)
CAMKV	chr3:49895421-49907655	-	0.18 (0.1, 0.31)*	-
CARM1	chr19:10982189-11033453	2.27 (1.68, 3.05)*†	-	-
CATSPER2	chr15:43920701-43960316	-	-	1.05 (0.46, 2.37)
CBLN3	chr14:24895738-24900160	1.25 (0.61, 2.54)	-	-
CCL22	chr16:57392684-57400102	-	0.37 (0.11, 1.22)	-
CD300LG	chr17:41924516-41940997	-	1.04 (0.57, 1.91)	1.58 (0.58, 4.35)
CD40	chr20:44746911-44758502	-	1.32 (0.46, 3.77)	0.75 (0.2, 2.72)
CDK12	chr17:37617764-37721160	-	0.19 (0.0, 37.54)	-
CEACAM16	chr19:45202421-45213986	1.66 (1.31, 2.11)*†	0.46 (0.27, 0.79)*	0.56 (0.25, 1.27)
CELSR2	chr1:109792641-109818372	1.97 (1.78, 2.18)*†	0.06 (0.04, 0.09)*	-
CES3	chr16:66995140-67009051	-	1.15 (0.6, 2.18)	-
CES4A	chr16:67022492-67043661	-	1.15 (0.6, 2.18)	-

Druggable gene	Genomic coordinates	LDL-C (OR, 95% CI)	HDL-C (OR, 95% CI)	Triglycerides (OR, 95% CI)
CETP	chr16:56995762-57017757	1.49 (1.29, 1.72)*	0.91 (0.87, 0.95)*†	1.98 (1.63, 2.4)*†
CFB	chr6:31895475-31919861	1.61 (0.94, 2.77)	1.85 (0.66, 5.16)	0.46 (0.24, 0.91)*
CGREF1	chr2:27321757-27341995	-	-	0.94 (0.66, 1.36)
CHST1	chr11:45670427-45687172	-	1.15 (0.45, 2.94)	-
CILP2	chr19:19649057-19657468	1.19 (1.01, 1.39)*	-	1.18 (1.0, 1.39)*†
CPS1	chr2:211342406-211543831	-	2.05 (1.1, 3.82)*	-
CRY2	chr11:45868669-45904798	-	0.71 (0.48, 1.04)	-
CSF1	chr1:110452864-110473614	-	0.57 (0.28, 1.15)	-
CSF3	chr17:38171614-38174066	0.3 (0.12, 0.74)*	0.87 (0.4, 1.89)	-
CSNK1G3	chr5:122847793-122952739	0.33 (0.2, 0.55)*	-	-
CTRL	chr16:67961543-67966317	-	1.11 (0.91, 1.35)	-
CTSA	chr20:44518783-44527459	-	1.89 (1.3, 2.75)*	0.12 (0.05, 0.28)*
CTSB	chr8:11700033-11726957	-	-	0.65 (0.44, 0.98)*
CYP21A2	chr6:32006042-32009447	2.34 (1.49, 3.66)*	1.23 (0.33, 4.63)	2.22 (1.03, 4.81)*
CYP26A1	chr10:94833232-94837647	7.25 (4.25, 12.37)*	0.22 (0.09, 0.51)*	4.35 (2.79, 6.79)*
CYP26C1	chr10:94821021-94828454	7.25 (4.23, 12.42)*	0.22 (0.09, 0.51)*	4.36 (2.79, 6.83)*
CYP7A1	chr8:59402737-59412795	0.95 (0.47, 1.91)	-	0.86 (0.33, 2.28)
DAGLA	chr11:61447905-61514473	1.67 (1.21, 2.29)*	0.49 (0.1, 2.47)	0.63 (0.53, 0.75)*
DAGLB	chr7:6448757-6523821	-	0.35 (0.22, 0.54)*	-
DCPS	chr11:126173647-126215644	4.96 (1.89, 13.06)*	0.29 (0.13, 0.66)*	-
DDR1	chr6:30844198-30867933	0.95 (0.5, 1.8)	1.62 (0.42, 6.27)	0.99 (0.37, 2.67)
DHODH	chr16:72042487-72058954	0.66 (0.44, 1.0)	-	7.42 (2.32, 23.71)*
DMPK	chr19:46272975-46285810	2.78 (1.64, 4.73)*	0.4 (0.11, 1.45)	-
DPEP2	chr16:68021297-68034489	-	1.36 (0.98, 1.89)	-
DPEP3	chr16:68009566-68014732	-	1.36 (0.98, 1.9)	-
DUSP3	chr17:41843489-41856356	-	0.93 (0.25, 3.55)	1.03 (0.39, 2.69)
EDEM2	chr20:33703167-33865928	2.37 (0.76, 7.37)	0.0 (0.0, 0.0)*	-
EGFL8	chr6:32132360-32136058	1.47 (0.71, 3.06)	-	1.96 (0.97, 3.97)
EHMT2	chr6:31847536-31865464	3.32 (1.25, 8.83)*	2.66 (0.73, 9.69)	0.23 (0.03, 1.56)
EMILIN1	chr2:27301435-27309271	-	-	0.73 (0.34, 1.57)
EMILIN3	chr20:39988606-39995467	2.51 (1.29, 4.86)*	-	-
ERBB2	chr17:37844167-37886679	-	2.82 (0.25, 31.53)	-
ERP29	chr12:112451120-112461255	0.11 (0.06, 0.18)*	0.04 (0.02, 0.13)*	-
ESR1	chr6:151977826-152450754	-	2.11 (1.13, 3.93)*	-
F2	chr11:46740730-46761056	0.17 (0.05, 0.59)*	0.57 (0.13, 2.43)	0.35 (0.13, 0.94)*
FCGRT	chr19:50010073-50029590	-	-	1.95 (0.75, 5.07)
FDFT1	chr8:11653082-11696818	-	-	0.88 (0.44, 1.73)
FEN1	chr11:61560109-61564716	2.02 (0.99, 4.12)	0.54 (0.2, 1.47)	1.13 (0.71, 1.8)
FGF21	chr19:49258816-49261587	1.06 (0.79, 1.43)	-	1.62 (0.46, 5.78)
FKBP6	chr7:72742167-72772634	-	-	0.52 (0.2, 1.34)
FLT3LG	chr19:49977464-49989488	-	-	1.95 (0.75, 5.07)
FN1	chr2:216225163-216300895	0.04 (0.01, 0.22)*	-	-
FOLH1	chr11:49168187-49230222	-	0.76 (0.36, 1.61)	-
FPR1	chr19:52248425-52307363	-	1.58 (1.11, 2.24)*	-
FPR2	chr19:52255279-52273779	-	1.89 (1.24, 2.87)*	-
FPR3	chr19:52298416-52329442	-	1.58 (1.1, 2.26)*	-
FRK	chr6:116252312-116381921	0.76 (0.48, 1.21)	0.57 (0.17, 1.94)	-
FZD9	chr7:72848109-72850450	-	1.13 (0.52, 2.46)	1.08 (0.72, 1.61)
GALK1	chr17:73747675-73761792	3.38 (1.46, 7.85)*	-	-
GALNT2	chr1:230193536-230417870	-	0.56 (0.42, 0.74)*	2.19 (1.48, 3.25)*
GANC	chr15:42565431-42645864	-	0.72 (0.31, 1.67)	1.32 (0.65, 2.71)
GCKR	chr2:27719709-27746554	1.49 (0.88, 2.51)	-	0.9 (0.45, 1.83)
GDF5	chr20:34021145-34042568	2.71 (0.77, 9.5)	-	-
GDF7	chr2:20866424-20873418	0.96 (0.6, 1.53)	-	1.33 (0.63, 2.8)
GFOD2	chr16:67708434-67753324	-	1.42 (1.01, 2.0)*	-
GGT7	chr20:33432523-33460663	-	0.43 (0.15, 1.23)	2.61 (1.1, 6.18)*
GIPR	chr19:46171502-46186982	2.72 (1.06, 7.02)*	0.89 (0.37, 2.12)	4.17 (1.16, 15.02)*
GPER1	chr7:1121844-1133451	2.81 (0.85, 9.29)	2.12 (0.84, 5.34)	-
GPIHBP1	chr8:144295068-144299044	-	1.82 (1.23, 2.71)*	-
GPR146	chr7:1084212-1098897	2.81 (0.85, 9.29)	2.12 (0.85, 5.33)	-
GPR61	chr1:110082494-110091028	1.97 (1.56, 2.5)*†	3.02 (0.77, 11.91)	5.14 (1.43, 18.48)*
GSK3B	chr3:119540170-119813264	-	0.45 (0.16, 1.25)	-
GSK3B	chr3:119540170-119813264	-	0.45 (0.16, 1.25)	-
GSS	chr20:33516236-33543620	-	0.57 (0.2, 1.65)	2.25 (1.05, 4.85)*
GSTM1	chr1:110230436-110251661	3.28 (1.83, 5.87)*	-	-
GSTM2	chr1:110210644-110252171	3.28 (1.83, 5.89)*	-	-
GSTM3	chr1:110276554-110284384	0.06 (0.0, 0.79)*	-	-
GSTM4	chr1:110198703-110208118	3.41 (1.54, 7.53)*	2.97 (0.88, 10.06)	-
GSTM5	chr1:110254864-110318050	0.06 (0.0, 0.84)*	-	-

Druggable gene	Genomic coordinates	LDL-C (OR, 95% CI)	HDL-C (OR, 95% CI)	Triglycerides (OR, 95% CI)
H3F3B	chr17:73772515-73781974	3.38 (1.46, 7.85)*	-	-
HAPLN4	chr19:19366450-19373605	1.06 (0.91, 1.22)	-	1.07 (0.93, 1.25)
HCAR1	chr12:123104824-123215390	-	0.99 (0.68, 1.44)	-
HCAR2	chr12:123185840-123187890	-	0.85 (0.52, 1.38)	-
HCAR3	chr12:123199303-123201439	-	0.85 (0.52, 1.39)	-
HDGF	chr1:156711899-156736717	-	1.18 (0.44, 3.19)	-
HFE	chr6:26087509-26098571	1.92 (0.77, 4.83)	0.81 (0.18, 3.61)	-
HGFAC	chr4:3443614-3451211	2.07 (1.35, 3.16)*	-	1.97 (1.34, 2.89)*
HIST1H1C	chr6:26055968-26056699	1.83 (0.79, 4.25)	0.84 (0.18, 4.01)	-
HIST1H4C	chr6:26104104-26104518	1.92 (0.77, 4.83)	0.81 (0.18, 3.61)	-
HLA-B	chr6:31321649-31324965	0.31 (0.08, 1.16)	1.81 (0.84, 3.88)	1.8 (1.46, 2.22)*
HLA-C	chr6:31236526-31239907	1.47 (0.06, 35.1)	1.73 (0.91, 3.28)	1.06 (0.8, 1.4)
HLA-DOB	chr6:32780540-32784825	3.57 (2.2, 5.78)*	1.06 (0.31, 3.59)	1.31 (0.81, 2.11)
HLA-DQA2	chr6:32709119-32714992	2.95 (2.41, 3.6)*	2.23 (1.19, 4.19)*	1.17 (0.82, 1.68)
HLA-DRA	chr6:32407619-32412823	2.29 (1.55, 3.37)*	1.2 (0.78, 1.85)	2.34 (1.41, 3.86)*
HLA-DRB1	chr6:32546546-32557625	1.37 (0.97, 1.93)	0.92 (0.6, 1.41)	1.44 (0.63, 3.32)
HLA-DRB5	chr6:32485120-32498064	2.75 (1.13, 6.7)*	0.8 (0.12, 5.14)	-
HLA-G	chr6:29794744-29798902	1.39 (0.44, 4.38)	-	1.33 (0.41, 4.3)
HMGCR	chr5:74632154-74657929	1.22 (1.03, 1.45)*	-	-
HNF4A	chr20:42984340-43061485	1.51 (0.67, 3.38)	1.18 (0.87, 1.6)	-
HP	chr16:72088491-72094954	1.1 (0.88, 1.38)	-	7.42 (2.32, 23.71)*
HPD	chr12:122277433-122301502	-	1.81 (0.68, 4.83)	0.38 (0.12, 1.16)
HPN	chr19:35531410-35557475	-	-	0.61 (0.24, 1.51)
HSD11B2	chr16:67464555-67471456	-	0.58 (0.31, 1.1)	-
HSD17B11	chr4:88257762-88312538	-	-	1.36 (0.9, 2.05)
HSD3B7	chr16:30996519-31000473	-	-	1.32 (0.63, 2.78)
HSPA1A	chr6:31783291-31785723	10.64 (2.34, 48.34)*	-	1.5 (0.44, 5.09)
HSPA1B	chr6:31795512-31798031	4.6 (1.61, 13.1)*	-	0.02 (0.0, 2.7)
IGF2R	chr6:160390131-160534539	4.56 (2.73, 7.61)*	-	224064.11 (5.34, 9394669504.1)*
INHBC	chr12:57828543-57844611	-	0.64 (0.41, 1.0)*	1.95 (0.99, 3.84)
INHBE	chr12:57846106-57853063	-	0.58 (0.35, 0.97)*	1.95 (1.0, 3.81)*
INSR	chr19:7112266-7294045	-	0.69 (0.47, 1.04)	1.15 (0.83, 1.62)
ITGB3	chr17:45331212-45421658	1.64 (1.06, 2.52)*	2.79 (0.81, 9.62)	-
ITGB4	chr17:73717408-73753899	3.38 (1.46, 7.85)*	-	-
ITIH1	chr3:52811603-52826078	2.07 (0.59, 7.23)	3.11 (0.85, 11.31)	-
ITIH3	chr3:52828784-52843025	2.07 (0.59, 7.23)	2.75 (0.81, 9.38)	-
ITIH4	chr3:52846991-52865495	2.07 (0.59, 7.23)	2.75 (0.81, 9.38)	-
KCNJ4	chr22:38822332-38851205	-	-	0.45 (0.16, 1.3)
KCNK16	chr6:39282474-39290744	1.13 (0.42, 3.04)	-	-
KCNK17	chr6:39266777-39282329	2.89 (1.46, 5.74)*	-	-
KCNK7	chr11:65360326-65363467	-	0.09 (0.03, 0.3)*	16.79 (4.99, 56.51)*
KHK	chr2:27309615-27323640	-	-	0.74 (0.41, 1.34)
KLHL8	chr4:88081255-88161466	0.96 (0.31, 2.93)	0.98 (0.47, 2.05)	1.15 (0.95, 1.41)
L3MBTL3	chr6:130334844-130462594	-	-	1.54 (0.57, 4.14)
LACTB	chr15:63413999-63434260	-	0.64 (0.48, 0.86)*	1.43 (0.87, 2.36)
LAIR1	chr19:54865362-54882165	-	1.53 (0.87, 2.69)	-
LCAT	chr16:67973653-67978034	-	1.11 (0.92, 1.34)	-
LCT	chr2:136545410-136594750	0.61 (0.23, 1.58)	-	-
LDLR	chr19:11200038-11244492	1.37 (0.98, 1.93)	0.04 (0.01, 0.34)*	-
LGR4	chr11:27387508-27494322	-	-	4.07 (2.03, 8.16)*
LILRA3	chr19:54799854-54809952	-	0.76 (0.44, 1.33)	-
LILRA4	chr19:54844456-54850421	-	0.87 (0.44, 1.69)	-
LILRA5	chr19:54818353-54824409	-	0.8 (0.45, 1.43)	-
LILRA6	chr19:54720737-54746649	-	0.96 (0.65, 1.41)	-
LILRB2	chr19:54777675-54785039	-	0.51 (0.27, 0.96)*	-
LILRB5	chr19:54754263-54761164	-	0.96 (0.65, 1.41)	-
LIPC	chr15:58702768-58861151	-	0.99 (0.89, 1.11)	0.91 (0.44, 1.86)
LIPG	chr18:47087069-47119272	0.41 (0.25, 0.67)*	0.95 (0.82, 1.1)	-
LPA	chr6:160952515-161087407	13.5 (7.17, 25.42)*	21.73 (4.69, 100.69)*	3.77 (1.78, 7.99)*
LPAR2	chr19:19734477-19739739	1.48 (1.22, 1.8)*	-	1.62 (1.39, 1.9)*
LPL	chr8:19759228-19824769	-	0.63 (0.49, 0.82)*†	1.68 (1.46, 1.92)*†
LRPAP1	chr4:3508103-3534286	5.03 (1.74, 14.56)*	-	2.59 (1.6, 4.2)*
LTA	chr6:31539831-31542101	2.03 (1.04, 3.97)*	-	1.22 (0.78, 1.9)
LTA	chr6:31539831-31542101	2.03 (1.04, 3.97)*	-	1.22 (0.78, 1.9)
LTB	chr6:31548302-31550299	2.01 (1.03, 3.93)*	1.2 (0.44, 3.25)	1.11 (0.76, 1.62)
LTK	chr15:41795836-41806085	-	0.8 (0.49, 1.29)	-
MAP2K7	chr19:7968728-7979363	-	1.24 (0.51, 3.02)	0.73 (0.2, 2.71)
MAP3K11	chr11:65365226-65382853	-	0.09 (0.03, 0.3)*	16.79 (4.99, 56.51)*

Druggable gene	Genomic coordinates	LDL-C (OR, 95% CI)	HDL-C (OR, 95% CI)	Triglycerides (OR, 95% CI)
MAP3K19	chr2:135722061-135805038	1.21 (0.53, 2.8)	-	-
MAP3K4	chr6:161412759-161538417	2.89 (1.75, 4.78)*	-	-
MAPK10	chr4:86936276-87515284	-	0.75 (0.47, 1.19)	1.22 (0.53, 2.77)
MAPKAPK5	chr12:112279782-112334343	0.26 (0.12, 0.57)*	-	-
MARK4	chr19:45582546-45808541	1.02 (0.73, 1.41)	0.72 (0.18, 2.87)	-
MARS	chr12:57869228-57911352	-	0.63 (0.39, 1.03)	1.95 (1.0, 3.83)
MAS1L	chr6:29454474-29455738	-	-	1.32 (0.62, 2.83)
MET	chr7:116312444-116438440	-	0.66 (0.43, 1.01)	1.09 (0.42, 2.87)
METAP1	chr4:99916771-99983964	-	1.05 (0.39, 2.85)	-
MFAP1	chr15:44096690-44117000	-	0.72 (0.24, 2.17)	1.33 (0.56, 3.19)
MMP24	chr20:33814457-33864801	2.37 (0.76, 7.37)	1.56 (0.93, 2.63)	-
MMP9	chr20:44637547-44645200	-	1.07 (0.14, 8.55)	0.46 (0.24, 0.85)*
MOGAT2	chr11:75428864-75444003	-	0.96 (0.43, 2.14)	-
MST1R	chr3:49924435-49941299	-	0.18 (0.11, 0.31)*	-
MYH7B	chr20:33563206-33590240	-	0.5 (0.19, 1.3)	2.25 (1.05, 4.85)*
MYL4	chr17:45277812-45301045	2.05 (0.77, 5.49)	-	-
NAT2	chr8:18248755-18258728	3.37 (2.07, 5.49)*	-	3.33 (1.94, 5.71)*
NCAN	chr19:19322782-19363042	1.04 (0.9, 1.2)	-	1.1 (0.86, 1.39)
NCR3	chr6:31556672-31560762	2.15 (1.09, 4.25)*	1.3 (0.46, 3.71)	1.03 (0.69, 1.55)
NDUFA13	chr19:19626545-19644285	1.63 (1.13, 2.35)*	-	1.18 (1.0, 1.39)*†
NDUFA7	chr19:8373490-8386280	-	0.57 (0.3, 1.07)	-
NDUFS3	chr11:47586888-47606114	-	1.18 (0.83, 1.69)	-
NEK4	chr3:52744800-52804965	1.58 (0.48, 5.24)	1.28 (0.43, 3.77)	-
NEU1	chr6:31825436-31830683	3.32 (1.24, 8.84)*	2.66 (0.73, 9.69)	0.18 (0.02, 1.59)
NFKB1	chr4:103422486-103538459	-	1.25 (0.5, 3.08)	-
NISCH	chr3:52489134-52527087	-	0.57 (0.35, 0.93)*	1.16 (0.31, 4.34)
NOTCH4	chr6:32162620-32191844	1.47 (0.7, 3.07)	1.9 (1.17, 3.1)*	0.87 (0.67, 1.13)
NPC1L1	chr7:44552134-44580914	2.01 (1.48, 2.73)*†	-	2.56 (0.75, 8.68)
NPEPPS	chr17:45600308-45700642	1.43 (0.67, 3.07)	3.28 (0.96, 11.23)	0.6 (0.22, 1.63)
NR0B2	chr1:27237980-27240457	1.38 (0.6, 3.18)	0.67 (0.27, 1.64)	1.91 (0.71, 5.17)
NR1H3	chr11:47269851-47290396	-	1.07 (0.97, 1.18)	0.86 (0.58, 1.27)
NR1I2	chr3:119499331-119537332	-	0.43 (0.17, 1.08)	-
NRBP1	chr2:27650657-27665126	-	-	0.89 (0.72, 1.1)
OBP2B	chr9:136080664-136084630	0.47 (0.06, 3.46)	-	-
OR11A1	chr6:29393281-29424848	-	-	2.42 (0.53, 11.09)
OR2H1	chr6:29424958-29432105	-	-	2.24 (0.55, 9.16)
OR4A16	chr11:55110627-55111707	-	0.83 (0.36, 1.93)	-
OR4C16	chr11:55339604-55340536	-	0.76 (0.32, 1.84)	-
PBRM1	chr3:52579368-52719933	1.58 (0.48, 5.24)	0.77 (0.41, 1.44)	2.21 (0.66, 7.34)
PCSK7	chr11:117075053-117103241	1.62 (0.85, 3.11)	0.78 (0.51, 1.21)	2.57 (0.09, 73.1)
PCSK9	chr1:55505221-55530525	1.6 (1.45, 1.77)*†	-	-
PDE3A	chr12:20522179-20837315	-	0.79 (0.4, 1.58)	-
PDGFC	chr4:157681606-157892546	-	0.52 (0.19, 1.39)	-
PDIA3	chr15:44038590-44065477	-	-	1.33 (0.56, 3.19)
PEPD	chr19:33877856-34012700	-	0.36 (0.25, 0.51)*	3.41 (1.92, 6.08)*
PIP4K2C	chr12:57984957-57997198	-	1.14 (0.27, 4.85)	-
PKM	chr15:72491370-72524164	-	-	1.54 (0.59, 4.03)
PLA2G15	chr16:68279207-68294961	-	1.45 (1.06, 1.97)*	-
PLA2G6	chr22:38507502-38601697	-	1.01 (0.57, 1.78)	1.09 (0.4, 3.0)
PLG	chr6:161123270-161174347	18.35 (5.47, 61.6)*	5.48 (0.07, 456.86)	0.75 (0.18, 3.14)
PLTP	chr20:44527399-44540794	-	0.67 (0.1, 4.53)	0.25 (0.02, 2.61)
PNMT	chr17:37824234-37826728	-	0.64 (0.36, 1.15)	-
PPARA	chr22:46546424-46639653	3.77 (1.44, 9.85)*	-	-
PPARG	chr3:12328867-12475855	1.67 (1.04, 2.68)*	0.71 (0.35, 1.48)	2.18 (1.14, 4.15)*
PPIL2	chr22:22006559-22054304	-	0.82 (0.47, 1.43)	-
PPY	chr17:42018172-42019836	-	1.34 (0.73, 2.47)	0.48 (0.13, 1.75)
PROCR	chr20:33759876-33765165	-	0.0 (0.0, 0.0)*	-
PRSS36	chr16:31150246-31161415	1.52 (0.43, 5.39)	-	1.16 (0.4, 3.32)
PRSS53	chr16:31094746-31100949	1.62 (0.45, 5.86)	-	1.32 (0.52, 3.35)
PRSS8	chr16:31142756-31147083	1.52 (0.43, 5.39)	-	1.32 (0.52, 3.36)
PSKH1	chr16:67927175-67963581	-	1.18 (0.92, 1.52)	-
PSMA1	chr11:14515329-14665181	-	1.43 (0.5, 4.11)	-
PSMA5	chr1:109941653-109969062	2.47 (1.8, 3.39)*†	0.08 (0.02, 0.29)*	-
PSMB10	chr16:67968405-67970990	-	1.11 (0.91, 1.35)	-
PSMB8	chr6:32808494-32812480	2.23 (0.99, 4.98)	1.06 (0.31, 3.59)	1.09 (0.6, 1.97)
PSMC3	chr11:47440320-47447993	-	1.12 (0.9, 1.4)	0.68 (0.18, 2.53)
PSMD3	chr17:38137050-38154213	0.3 (0.12, 0.74)*	0.85 (0.55, 1.31)	-
PTPN11	chr12:112856155-112947717	1259419.6 (0.0, 4107788603997230.5)	0.05 (0.02, 0.16)*	-

Druggable gene	Genomic coordinates	LDL-C (OR, 95% CI)	HDL-C (OR, 95% CI)	Triglycerides (OR, 95% CI)
PTPN13	chr4:87515468-87736324	-	1.03 (0.64, 1.65)	0.99 (0.7, 1.39)
PTPRJ	chr11:48002113-48189670	-	0.5 (0.35, 0.72)*	-
PVR	chr19:45147098-45166850	1.31 (1.12, 1.54)*†	0.32 (0.11, 0.91)*	-
PVRL2	chr19:45349432-45392485	1.43 (1.27, 1.63)*†	0.35 (0.24, 0.51)*	0.51 (0.19, 1.37)
RAF1	chr3:12625100-12705725	2.06 (1.48, 2.86)*	-	2.63 (0.79, 8.83)
RAF1	chr3:12625100-12705725	2.06 (1.48, 2.86)*	-	2.63 (0.79, 8.83)
RELA	chr11:65421067-65430565	-	0.09 (0.03, 0.3)*	16.79 (4.99, 56.51)*
RGS12	chr4:3294755-3441640	1.99 (1.31, 3.03)*	-	1.98 (1.35, 2.89)*
RHCE	chr1:25688740-25756683	0.71 (0.4, 1.26)	-	-
RHD	chr1:25598884-25656936	0.27 (0.07, 1.07)	-	-
RPL13A	chr19:49990811-49995565	-	-	1.95 (0.75, 5.07)
RPL17	chr18:47014851-47018906	0.61 (0.21, 1.81)	0.75 (0.52, 1.08)	-
RPL19	chr17:37356536-37360980	-	0.45 (0.23, 0.9)*	-
RPL5	chr1:93297582-93307481	0.43 (0.27, 0.68)*	-	-
RPL6	chr12:112842994-112856642	0.13 (0.06, 0.28)*	0.05 (0.02, 0.16)*	-
RPL7A	chr9:136215069-136218281	2.29 (1.57, 3.36)*†	-	-
RPS11	chr19:49999622-50002946	-	-	1.95 (0.75, 5.07)
RPS28	chr19:8386042-8388224	-	0.57 (0.3, 1.07)	-
RPS6KA1	chr1:26856252-26901521	-	1.7 (0.62, 4.68)	-
RPS9	chr19:54704610-54752862	-	0.95 (0.64, 1.41)	-
RSPO3	chr6:127439749-127518910	-	0.69 (0.45, 1.07)	1.1 (0.74, 1.65)
SAE1	chr19:47616531-47713886	-	0.23 (0.08, 0.65)*	3.19 (1.82, 5.58)*
SCARB1	chr12:125261402-125367214	5.16 (1.85, 14.39)*	0.42 (0.14, 1.26)	8.0 (2.41, 26.56)*
SCN1B	chr19:35521588-35531352	-	-	0.61 (0.24, 1.51)
SCUBE3	chr6:35182190-35220856	0.2 (0.06, 0.66)*	0.0 (0.0, 0.13)*	-
SEMA3F	chr3:50192478-50226508	-	0.4 (0.18, 0.89)*	-
SEMA3G	chr3:52467069-52479101	-	1.07 (0.45, 2.55)	1.16 (0.31, 4.34)
SERPINA1	chr14:94843084-94857030	0.65 (0.22, 1.9)	-	-
SERPINA10	chr14:94749650-94759608	0.73 (0.26, 2.1)	-	-
SERPINA6	chr14:94770585-94789731	0.65 (0.22, 1.9)	-	-
SFN	chr1:27189633-27190947	1.51 (0.67, 3.41)	0.49 (0.25, 0.97)*	1.91 (0.71, 5.17)
SFTA2	chr6:30899130-30899952	0.85 (0.4, 1.78)	1.97 (0.61, 6.41)	0.69 (0.25, 1.9)
SIDT2	chr11:117049449-117068160	1.62 (0.85, 3.11)	0.79 (0.51, 1.21)	1.03 (0.19, 5.6)
SIK3	chr11:116714118-116969153	1.15 (0.57, 2.31)	0.46 (0.29, 0.73)*†	1.08 (0.98, 1.18)
SLC12A3	chr16:56899119-56949762	1.94 (1.43, 2.63)*	0.89 (0.86, 0.93)*†	0.75 (0.24, 2.33)
SLC12A4	chr16:67977377-68003504	-	1.11 (0.92, 1.33)	-
SLC12A5	chr20:44650356-44688784	-	2.06 (0.18, 22.96)	0.4 (0.21, 0.76)*
SLC18A1	chr8:20002366-20040717	-	0.22 (0.08, 0.61)*	2.41 (1.66, 3.5)*
SLC22A1	chr6:160542821-160579750	4.39 (2.62, 7.36)*	-	224064.11 (5.16, 9728155751.75)*
SLC22A2	chr6:160592093-160698670	2.48 (1.85, 3.32)*	2.55 (1.16, 5.63)*	6.47 (2.56, 16.37)*
SLC22A3	chr6:160769300-160876014	5.13 (3.44, 7.66)*	2.4 (1.27, 4.53)*	4.42 (2.62, 7.45)*
SLC44A4	chr6:31830969-31846823	3.32 (1.25, 8.83)*	2.66 (0.73, 9.69)	0.23 (0.03, 1.59)
SLC5A6	chr2:27422455-27435826	-	-	0.86 (0.71, 1.05)
SLC9A1	chr1:27425306-27493472	1.03 (0.43, 2.46)	0.96 (0.22, 4.13)	-
SLCO1B1	chr12:21284136-21392180	-	-	0.22 (0.07, 0.65)*
SMARCA4	chr19:11071598-11176071	2.22 (1.98, 2.49)*†	0.01 (0.0, 0.02)*	-
SOST	chr17:41831099-41836156	-	0.93 (0.25, 3.55)	1.03 (0.39, 2.69)
ST3GAL4	chr11:126225535-126310239	2.25 (1.16, 4.39)*	0.03 (0.0, 0.17)*	-
STX1A	chr7:73113536-73134002	-	-	0.89 (0.62, 1.28)
TECTB	chr10:114043493-114064793	0.88 (0.3, 2.64)	0.63 (0.27, 1.5)	1.49 (0.9, 2.49)
TMED1	chr19:10943114-10946994	2.06 (1.5, 2.83)*†	-	-
TNF	chr6:31543344-31546113	2.03 (1.05, 3.93)*	-	1.21 (0.78, 1.9)
TNKS	chr8:9413424-9639856	-	-	0.79 (0.54, 1.16)
TNNC1	chr3:52485118-52488086	-	0.57 (0.36, 0.93)*	1.73 (0.56, 5.36)
TNNC2	chr20:44451853-44462384	-	1.21 (0.69, 2.12)	0.81 (0.43, 1.52)
TNXB	chr6:32008931-32083111	2.15 (1.55, 2.97)*	0.98 (0.21, 4.67)	1.64 (0.95, 2.82)
TOP1	chr20:39657458-39753127	2.3 (0.15, 35.62)	-	16.72 (4.19, 66.8)*
TSSK6	chr19:19623227-19626838	1.64 (1.14, 2.37)*	-	1.17 (0.99, 1.39)
TUBB	chr6:30687978-30693203	-	7.56 (1.18, 48.38)*	4.46 (2.13, 9.36)*
TYRO3	chr15:41849873-41871536	-	0.8 (0.49, 1.29)	-
UCN	chr2:27530268-27531313	-	-	1.35 (0.53, 3.44)
UGT1A1	chr2:234668894-234681945	1.33 (0.74, 2.39)	-	-
UGT1A10	chr2:234545100-234681951	1.95 (1.25, 3.05)*	-	-
UGT1A3	chr2:234637754-234681945	2.04 (1.27, 3.25)*	-	-
UGT1A4	chr2:234627424-234681945	2.03 (1.27, 3.23)*	-	-
UGT1A5	chr2:234621638-234681945	2.03 (1.27, 3.25)*	-	-
UGT1A6	chr2:234600253-234681946	1.91 (1.22, 3.0)*	-	-
UGT1A7	chr2:234590584-234681945	1.94 (1.22, 3.07)*	-	-

Druggable gene	Genomic coordinates	LDL-C (OR, 95% CI)	HDL-C (OR, 95% CI)	Triglycerides (OR, 95% CI)
UGT1A8	chr2:234526291-234681956	1.95 (1.23, 3.08)*	-	-
UGT1A9	chr2:234580499-234681946	1.94 (1.24, 3.05)*	-	-
VEGFA	chr6:43737921-43754224	-	0.22 (0.15, 0.3)*	4.16 (2.45, 7.08)*†
VIM	chr10:17270258-17279592	1.02 (0.34, 3.06)	0.77 (0.21, 2.78)	-
VKORC1	chr16:31102163-31107301	1.62 (0.45, 5.86)	-	1.32 (0.52, 3.35)
WNT9B	chr17:44910567-44964096	-	6.95 (2.1, 23.05)*	

****** - significant in the discovery analysis; †- significant in both original and validation study and concordant direction of effect. OR = odds ratio per 1-SD increase in LDL-C/HDL-C or triglycerides; CI = confidence interval.*

Appendix 7.C. Univariable MR estimates of drug targets with lipid records in clinicaltrials.gov and/or the British National Formulary (BNF).

Gene	LDL-C (OR, 95% CI)	HDL -C (OR, 95% CI)	Triglycerides (OR, 95% CI)	Clinical trial (Record type)	BNF (Record type)	Phase	Mechanism of action	Indication
ADRB1	-	1.67 (0.58, 4.8)	-	Outcome	Side effect	4	AGONIST	Pain, Asthma, Nasal Obstruction, Glaucoma, Obstructive Lung Diseases, Hemorrhage, Cardiovascular Diseases, Serum Sickness, Bronchial Spasm, Rhinitis, Seasonal Allergic, Urticaria, Heart Arrest, Angioedema, Sinusitis, Sepsis, Hypotension, Orthostatic
ADRB1	-	1.67 (0.58, 4.8)	-	Outcome	Side effect	4	ANTAGONIST	Angina Pectoris, Hypertension, Myocardial Infarction, Cardiovascular Diseases, Arrhythmias, Cardiac, Migraine Disorders, Open-Angle Glaucoma, Ocular Hypertension, Glaucoma, Heart Failure, Left Ventricular Dysfunction
ADRB1	-	1.67 (0.58, 4.8)	-	Outcome	Side effect	4	PARTIAL AGONIST	Cardiovascular Diseases
ESR1	-	2.11 (1.13, 3.93)*	-	Outcome	Side effect	4	AGONIST	Neoplasms, Hypogonadism, Menorrhagia, Primary Ovarian Insufficiency, Acne Vulgaris, Osteoporosis, Postmenopausal
ESR1	-	2.11 (1.13, 3.93)*	-	Outcome	Side effect	4	ANTAGONIST	Breast Neoplasms, Neoplasms
ESR1	-	2.11 (1.13, 3.93)*	-	Outcome	Side effect	4	MODULATOR	Infertility, Dyspareunia, Breast Neoplasms, Osteoporosis, Postmenopausal
TNF	2.03 (1.05, 3.93)*	-	1.21 (0.78, 1.9)	Outcome	Side effect	4	INHIBITOR	Ankylosing Spondylitis, Crohn Disease, Psoriasis, Rheumatoid Arthritis, Colitis, Ulcerative, Psoriatic Arthritis, Immune System Diseases, Juvenile Arthritis
FRK	0.76 (0.48, 1.21)	0.57 (0.17, 1.94)	-	Outcome	Side effect	4	INHIBITOR	Neoplasms, Precursor Cell Lymphoblastic Leukemia-Lymphoma
BLK	-	-	0.46 (0.31, 0.7)*	Outcome	Side effect	4	INHIBITOR	Precursor Cell Lymphoblastic Leukemia-Lymphoma, Neoplasms
DHODH	0.66 (0.44, 1.0)	-	7.42 (2.32, 23.71)*	Adverse event	Side effect	4	INHIBITOR	Rheumatoid Arthritis, Immune System Diseases, Multiple Sclerosis

Gene	LDL-C (OR, 95% CI)	HDL -C (OR, 95% CI)	Triglycerides (OR, 95% CI)	Clinical trial (Record type)	BNF (Record type)	Phase	Mechanism of action	Indication
HMGCR	1.22 (1.03, 1.45)*	-	-	Outcome	Indication	4	INHIBITOR	Cardiovascular Diseases, Hypercholesterolemia, Dyslipidemias, Hyperlipidemias, Coronary Artery Disease, Hyperlipoproteinemia Type II, Myocardial Infarction, Heart Failure, Hypertension, Stroke, Stable Angina, Angina Pectoris, Type 2 Diabetes Mellitus
NPC1L1	2.01 (1.48, 2.73)*†	-	2.56 (0.75, 8.68)	Outcome	Indication	4	INHIBITOR	Hypercholesterolemia, Hyperlipidemias, Cardiovascular Diseases
PPARG	1.67 (1.04, 2.68)*	0.71 (0.35, 1.48)	2.18 (1.14, 4.15)*	Outcome	Indication	4	AGONIST	Type 2 Diabetes Mellitus, Diabetes Mellitus, Colitis, Ulcerative, Cardiovascular Diseases
PPARA	3.77 (1.44, 9.85)*	-	-	Outcome	Indication	4	AGONIST	Cardiovascular Diseases, Hypercholesterolemia, Dyslipidemias
PCSK9	1.6 (1.45, 1.77)*†	-	-	Outcome	Indication	4	INHIBITOR	Hyperlipoproteinemia Type II, Coronary Artery Disease, Cardiovascular Diseases
INSR	-	0.69 (0.47, 1.04)	1.15 (0.83, 1.62)	Outcome	-	4	AGONIST	Diabetes Mellitus, Type 2 Diabetes Mellitus, Type 1 Diabetes Mellitus
NDUFS3	-	1.18 (0.83, 1.69)	-	Outcome	-	4	INHIBITOR	Diabetes Mellitus, Type 2 Diabetes Mellitus
NDUFA7	-	0.57 (0.3, 1.07)	-	Outcome	-	4	INHIBITOR	Diabetes Mellitus, Type 2 Diabetes Mellitus
NDUFA13	1.63 (1.13, 2.35)*	-	1.18 (1.0, 1.39)*†	Outcome	-	4	INHIBITOR	Diabetes Mellitus, Type 2 Diabetes Mellitus
ALDH2	0.14 (0.07, 0.29)*	-	-	Outcome	-	4	INHIBITOR	Ectoparasitic Infestations, Alcoholism
NISCH	-	0.57 (0.35, 0.93)*	1.16 (0.31, 4.34)	Outcome	-	4	AGONIST	Hypertension
ABCA1	2.05 (1.34, 3.15)*	1.41 (0.66, 3.0)	2.4 (1.29, 4.49)*	Outcome	-	4	INHIBITOR	Cardiovascular Diseases
PDE3A	-	0.79 (0.4, 1.58)	-	Outcome	-	4	INHIBITOR	Thrombosis, Obstructive Lung Diseases, Essential Thrombocythemia, Asthma, Cardiovascular Diseases, Coronary Artery Disease, Stroke
F2	0.17 (0.05, 0.59)*	0.57 (0.13, 2.43)	0.35 (0.13, 0.94)*	Outcome	-	4	INHIBITOR	Venous Thrombosis, Thrombosis, Unstable Angina, Thrombocytopenia, Atrial Fibrillation, Embolism, Stroke
TUBB	-	7.56 (1.18, 48.38)*	4.46 (2.13, 9.36)*	Adverse event	-	4	INHIBITOR	Breast Neoplasms, Neoplasms, Hodgkin Disease, Large-Cell Anaplastic Lymphoma, Non-Small-Cell Lung Carcinoma, Gout, Familial Mediterranean Fever
VEGFA	-	0.22 (0.15, 0.3)*	4.16 (2.45, 7.08)*†	Adverse event	-	4	ANTAGONIST	Retinal Neovascularization

Gene	LDL-C (OR, 95% CI)	HDL -C (OR, 95% CI)	Triglycerides (OR, 95% CI)	Clinical trial (Record type)	BNF (Record type)	Phase	Mechanism of action	Indication
VEGFA	-	0.22 (0.15, 0.3)*	4.16 (2.45, 7.08)*†	Adverse event	-	4	INHIBITOR	Diabetic Retinopathy, Retinal Neovascularization, Wet Macular Degeneration, Macular Edema, Colorectal Neoplasms, Neoplasms, Glioblastoma, Renal Cell Carcinoma, Non-Small-Cell Lung Carcinoma, Uterine Cervical Neoplasms
ERBB2	-	2.82 (0.25, 31.53)	-	Adverse event	-	4	INHIBITOR	Breast Neoplasms, Neoplasms, Non-Small-Cell Lung Carcinoma, Thyroid Neoplasms
RAF1	2.06 (1.48, 2.86)*	-	2.63 (0.79, 8.83)	Adverse event	-	4	INHIBITOR	Neoplasms
PSMC3	-	1.12 (0.9, 1.4)	0.68 (0.18, 2.53)	Adverse event	-	4	INHIBITOR	Multiple Myeloma, Neoplasms, Mantle-Cell Lymphoma
PSMA1	-	1.43 (0.5, 4.11)	-	Adverse event	-	4	INHIBITOR	Multiple Myeloma, Neoplasms, Mantle-Cell Lymphoma
PSMB8	2.23 (0.99, 4.98)	1.06 (0.31, 3.59)	1.09 (0.6, 1.97)	Adverse event	-	4	INHIBITOR	Multiple Myeloma, Neoplasms, Mantle-Cell Lymphoma
PSMB9	2.23 (0.99, 4.98)	1.06 (0.31, 3.59)	1.1 (0.6, 1.99)	Adverse event	-	4	INHIBITOR	Multiple Myeloma, Neoplasms, Mantle-Cell Lymphoma
PSMA5	2.47 (1.8, 3.39)*†	0.08 (0.02, 0.29)*	-	Adverse event	-	4	INHIBITOR	Multiple Myeloma, Neoplasms, Mantle-Cell Lymphoma
PSMB10	-	1.11 (0.91, 1.35)	-	Adverse event	-	4	INHIBITOR	Multiple Myeloma, Neoplasms, Mantle-Cell Lymphoma
ALOX5	-	1.74 (1.18, 2.58)*	-	Adverse event	-	4	INHIBITOR	Asthma, Ulcerative Colitis, Rheumatoid Arthritis, Juvenile Arthritis
CACNB1	-	0.38 (0.2, 0.72)*	-	Adverse event	-	4	BLOCKER	Cardiovascular Diseases
CACNB1	-	0.38 (0.2, 0.72)*	-	Adverse event	-	4	MODULATOR	Fibromyalgia, Seizures, Epilepsy, Neuralgia, Restless Legs Syndrome, Postherpetic Neuralgia
PLG	18.35 (5.47, 61.6)*	5.48 (0.07, 456.86)	0.75 (0.18, 3.14)	Adverse event	-	4	ACTIVATOR	Thrombosis, Pulmonary Embolism, Stroke, Myocardial Infarction, Heart Failure, Hepatic Veno-Occlusive Disease
PLG	18.35 (5.47, 61.6)*	5.48 (0.07, 456.86)	0.75 (0.18, 3.14)	Adverse event	-	4	INHIBITOR	Hemorrhage, Menorrhagia
ITGB3	1.64 (1.06, 2.52)*	2.79 (0.81, 9.62)	-	Adverse event	-	4	INHIBITOR	Thrombosis, Unstable Angina
MET	-	0.66 (0.43, 1.01)	1.09 (0.42, 2.87)	Adverse event	-	4	INHIBITOR	Thyroid Neoplasms, Non-Small-Cell Lung Carcinoma, Neoplasms
GSK3B	-	0.45 (0.16, 1.25)	-	Adverse event	-	4	INHIBITOR	Bipolar Disorder, Psychotic Disorders

Gene	LDL-C (OR, 95% CI)	HDL -C (OR, 95% CI)	Triglycerides (OR, 95% CI)	Clinical trial (Record type)	BNF (Record type)	Phase	Mechanism of action	Indication
FDFT1	-	-	0.88 (0.44, 1.73)	Outcome	-	3	INHIBITOR	Hypercholesterolemia, Lipid Metabolism Disorders, Type 2 Diabetes Mellitus
CETP	1.49 (1.29, 1.72)*	0.91 (0.87, 0.95)*†	1.98 (1.63, 2.4)*†	Outcome	-	3	INHIBITOR	Hypercholesterolemia, Lipid Metabolism Disorders, Hyperlipoproteinemia Type II, Coronary Disease, Cardiovascular Diseases, Acute Coronary Syndrome, Hyperlipidemias
ANGPTL3	1.21 (1.11, 1.33)*	1.61 (0.52, 5.01)	1.16 (1.08, 1.25)*	Outcome	-	3	INHIBITOR	Hyperlipoproteinemia Type II
AKT1	-	0.49 (0.18, 1.36)	-	Adverse event	-	3	INHIBITOR	Prostatic Neoplasms
SOST	-	0.93 (0.25, 3.55)	1.03 (0.39, 2.69)	Adverse event	-	3	INHIBITOR	Osteoporosis, Postmenopausal, Osteoporosis, Bone Diseases
CYP26A1	7.25 (4.25, 12.37)*	0.22 (0.09, 0.51)*	4.35 (2.79, 6.79)*	Adverse event	-	2	INHIBITOR	Psoriasis, Acne Vulgaris
LTA	2.03 (1.04, 3.97)*	-	1.22 (0.78, 1.9)	Adverse event	-	2	INHIBITOR	Rheumatoid Arthritis, Sjogren's Syndrome
LTB	2.01 (1.03, 3.93)*	1.2 (0.44, 3.25)	1.11 (0.76, 1.62)	Adverse event	-	2	INHIBITOR	Sjogren's Syndrome, Rheumatoid Arthritis
NR1H3	-	1.07 (0.97, 1.18)	0.86 (0.58, 1.27)	Outcome	-	1	AGONIST	Hypercholesterolemia
NR1H3	-	1.07 (0.97, 1.18)	0.86 (0.58, 1.27)	Outcome	-	1	MODULATOR	Hypercholesterolemia
TOP1	2.3 (0.15, 35.62)	-	16.72 (4.19, 66.8)*	Adverse event	-	4	INHIBITOR	Neoplasms

** indicates significance in the discovery analysis; † indicates significance in both original and validation study and concordant direction of effect. OR = CHD odds ratio per 1-SD increase in LDL-C/HDL-C or triglycerides; CI = confidence interval.*

Appendix 7.D. Multivariable drug target MR estimates. OR = CHD odds ratio per 1-SD increase in LDL-C/HDL-C or triglycerides; CI = confidence interval. An asterisk (*) indicates significant estimates.

Drug target gene	No. variants	Heterogeneity *p* value	LDL-C (OR, 95% CI)	HDL-C (OR, 95% CI)	Triglycerides (OR, 95% CI)
MARK4	6	7.47e-01	1.08 (0.82, 1.42)	1.01 (0.33, 3.12)	0.82 (0.18, 3.65)
GIPR	4	3.61e-01	2.49 (1.2, 5.19)*	0.45 (0.1, 2.07)	1.91 (0.32, 11.34)
NPC1L1	6	3.40e-01	1.5 (0.76, 2.96)	0.76 (0.11, 5.15)	1.11 (0.21, 5.92)
NR1H3	8	8.01e-01	1.4 (0.45, 4.34)	0.88 (0.73, 1.05)	0.51 (0.37, 0.7)*
SLC18A1	7	2.32e-01	5.31 (0.45, 62.01)	0.21 (0.01, 3.79)	0.37 (0.02, 9.01)
LPAR2	5	2.93e-01	0.07 (0.0, 25.72)	4.89 (0.02, 957.6)	16.68 (0.07, 4147.03)
CTSA	5	7.66e-01	0.09 (0.02, 0.48)*	77.68 (6.48, 931.29)*	117.3 (7.26, 1896.01)*
SLC12A3	19	4.16-01	4.02 (2.34, 6.93)*	1.16 (0.96, 1.4)	0.88 (0.34, 2.25)
PVR	10	2.15e-01	1.41 (0.93, 2.14)	1.71 (0.54, 5.41)	1.18 (0.3, 4.68)
SCARB1	10	2.48e-03	25.58 (1.06, 614.7)*	0.79 (0.47, 1.33)	0.06 (0.0, 1.2)
FDFT1	5	9.72e-01	1.87 (0.3, 11.58)	0.94 (0.1, 8.91)	0.93 (0.57, 1.5)
APOB	16	5.48e-04	1.54 (1.02, 2.33)*	2.48 (0.66, 9.29)	2.37 (0.78, 7.19)
GCKR	10	3.45e-01	3.11 (0.8, 12.13)	1.79 (0.24, 13.49)	0.84 (0.58, 1.24)
CAD	11	7.56e-01	4.64 (0.9, 24.08)	1.26 (0.26, 6.07)	0.8 (0.66, 0.97)*
CETP	36	3.00e-02	1.89 (0.83, 4.3)	1.0 (0.84, 1.19)	0.96 (0.46, 2.0)
PLTP	5	6.02e-01	0.18 (0.04, 0.8)*	24.18 (4.51, 129.56)*	28.65 (4.7, 174.74)*
MMP9	5	1.01e-01	0.01 (0.0, 0.06)*	5.65 (0.72, 44.19)	17.38 (1.99, 152.02)*
LIPG	20	8.64e-02	0.24 (0.13, 0.47)*	0.9 (0.74, 1.09)	5.09 (3.36, 7.7)*
DHODH	11	8.39e-01	2.18 (1.61, 2.95)*	2.08 (0.94, 4.59)	0.44 (0.16, 1.22)
LACTB	6	7.157e-01	0.21 (0.0, 14.0)	1.35 (0.11, 16.87)	1.61 (0.06, 43.74)
LILRB5	6	8.18e-01	1.63 (0.13, 20.4)	0.82 (0.5, 1.36)	1.1 (0.12, 10.51)
STX1A	5	7.75e-01	0.51 (0.13, 2.04)	0.34 (0.03, 3.29)	0.7 (0.26, 1.87)
HGFAC	7	9.98e-01	1.19 (0.13, 10.74)	0.15 (0.01, 2.31)	0.9 (0.07, 11.59)
DCPS	9	2.46e-01	0.72 (0.41, 1.23)	0.33 (0.16, 0.69)*	6.7 (1.54, 29.28)*
ST3GAL4	10	5.14e-03	1.46 (0.32, 6.8)	1.03 (0.15, 6.97)	6.72 (0.29, 156.11)
APOA5	14	6.87e-01	10.23 (6.21, 16.84)*	1.23 (0.93, 1.62)	0.73 (0.61, 0.88)*
APOA4	15	5.17e-01	1.53 (0.74, 3.14)	1.0 (0.74, 1.37)	1.11 (0.82, 1.5)
APOC3	18	1.86e-03	2.24 (1.18, 4.27)*	0.75 (0.62, 0.91)*	0.81 (0.69, 0.95)*
SLC22A2	16	2.07e-04	2.73 (1.66, 4.47)*	1.26 (0.4, 3.93)	1.38 (0.34, 5.63)
VEGFA	4	4.39e-01	0.27 (0.04, 1.67)	0.39 (0.0, 346.13)	2.39 (0.01, 1013.57)
HMGCR	7	4.30e-01	1.79 (1.28, 2.5)*	0.13 (0.01, 1.38)	0.31 (0.02, 4.26)
NRBP1	10	8.43e-01	0.68 (0.06, 7.53)	1.39 (0.09, 21.94)	0.91 (0.47, 1.76)
AMPD2	4	4.22e-01	2.43 (1.01, 5.81)*	0.05 (0.0, 15.18)	0.01 (0.0, 70.8)
APOA1	17	1.76e-02	2.21 (1.26, 3.87)*	0.84 (0.67, 1.04)	0.97 (0.74, 1.29)
PLG	5	5.49e-01	21.26 (12.83, 35.23)*	0.17 (0.04, 0.7)*	0.15 (0.03, 0.86)*
SLC12A5	4	9.71e-01	0.0 (0.0, 0.16)*	15.94 (0.81, 312.7)	73.81 (1.12, 4881.0)*
PEPD	6	8.91e-01	2.24 (0.66, 7.57)	0.48 (0.21, 1.09)	1.74 (0.72, 4.22)
ATG4C	8	6.51e-01	0.42 (0.12, 1.47)	4.32 (1.24, 15.04)*	0.96 (0.41, 2.25)
SMARCA4	15	8.61e-02	1.95 (1.8, 2.11)*	0.97 (0.49, 1.92)	0.13 (0.04, 0.44)*
ALDH1A2	42	4.49e-01	1.29 (0.75, 2.22)	1.1 (0.95, 1.28)	1.63 (1.03, 2.57)*
LDLR	18	6.01e-04	1.37 (1.15, 1.62)*	0.23 (0.05, 1.11)	0.12 (0.01, 1.03)
PVRL2	15	3.29e-02	1.29 (1.08, 1.54)*	1.12 (0.49, 2.53)	1.03 (0.72, 1.46)
APOE	14	6.96e-03	1.28 (1.16, 1.42)*	0.9 (0.56, 1.46)	0.87 (0.65, 1.17)
APOC1	14	2.86e-03	1.3 (1.17, 1.46)*	0.9 (0.52, 1.58)	0.81 (0.6, 1.08)
NCAN	6	4.55e-01	1.32 (0.2, 8.87)	0.42 (0.06, 2.82)	0.9 (0.14, 6.02)
LILRB2	8	7.81e-01	0.87 (0.18, 4.19)	1.02 (0.7, 1.49)	0.91 (0.13, 6.51)
PPARG	14	1.20e-02	2.77 (0.99, 7.76)	0.38 (0.15, 0.96)*	0.54 (0.15, 1.87)
ANGPTL3	5	6.85e-01	7.52 (0.07, 826.16)	0.43 (0.09, 1.97)	0.35 (0.01, 10.44)
AMPD3	5	7.32e-01	0.02 (0.0, 0.21)*	1.52 (0.55, 4.22)	5.21 (0.4, 67.63)
ACP2	9	6.75e-01	0.77 (0.38, 1.54)	0.84 (0.71, 0.98)*	0.56 (0.43, 0.73)*
DAGLA	9	4.84e-01	0.95 (0.67, 1.35)	1.48 (0.51, 4.3)	0.92 (0.43, 1.98)
BLK	4	2.53e-01	0.04 (0.0, 0.61)*	0.8 (0.01, 91.76)	0.35 (0.15, 0.84)*
CGREF1	5	5.99e-01	0.59 (0.1, 3.58)	1.68 (0.11, 26.04)	1.1 (0.87, 1.38)
SLC5A6	11	7.23e-01	2.7 (0.54, 13.63)	1.77 (0.36, 8.66)	0.83 (0.69, 1.0)
ATRAID	11	4.64e-01	2.49 (0.39, 15.91)	1.62 (0.27, 9.53)	0.86 (0.62, 1.18)
CBLN3	4	9.61e-02	0.71 (0.27, 1.91)	4.05 (0.08, 211.75)	51.45 (0.28, 9606.89)
PSMA5	4	7.79e-01	1.46 (0.66, 3.21)	0.12 (0.0, 5.13)	0.13 (0.0, 56.08)
CELSR2	23	2.91e-02	1.88 (1.5, 2.34)*	0.86 (0.25, 2.88)	1.84 (0.52, 6.56)
GALNT2	17	6.60e-03	0.98 (0.16, 6.02)	0.61 (0.12, 3.13)	0.94 (0.14, 6.14)
GDF7	4	1.89e-01	0.87 (0.39, 1.98)	1.16 (0.04, 37.71)	2.44 (0.05, 127.94)
KLHL8	8	9.95e-01	0.5 (0.07, 3.56)	1.51 (0.28, 8.02)	1.95 (1.08, 3.54)*
RSPO3	5	5.57e-01	0.02 (0.0, 0.7)*	11.04 (0.44, 279.24)	100.3 (0.84, 11945.38)
SLC22A3	11	9.02e-02	4.7 (3.44, 6.43)*	2.64 (1.45, 4.81)*	3.5 (1.96, 6.24)*

Drug target gene	No. variants	Heterogeneity *p* value	LDL-C (OR, 95% CI)	HDL-C (OR, 95% CI)	Triglycerides (OR, 95% CI)
RPL7A	6	5.17e-01	2.39 (1.44, 3.99)*	5.41 (2.64, 11.1)*	2.09 (0.09, 46.73)
PTPRJ	4	8.41e-02	12.48 (1.22, 127.5)*	0.51 (0.28, 0.92)*	0.04 (0.0, 0.58)*
SIDT2	10	1.41e-05	5.26 (0.2, 139.08)	1.34 (0.52, 3.46)	1.08 (0.34, 3.41)
NAT2	9	6.67e-01	1.46 (0.2, 10.91)	3.77 (1.16, 12.22)*	1.39 (0.4, 4.86)
GPR61	8	9.57e-01	2.05 (1.62, 2.6)*	0.82 (0.32, 2.05)	0.47 (0.18, 1.19)
RGS12	8	9.93e-01	0.7 (0.2, 2.44)	0.3 (0.07, 1.26)	1.76 (0.56, 5.6)
CILP2	4	7.39e-01	55.61 (0.0, 15957345.62)	10.05 (0.02, 5027.9)	0.04 (0.0, 3522.79)
SIK3	17	1.57e-01	3.68 (2.02, 6.69)*	0.76 (0.65, 0.88)*	0.69 (0.59, 0.8)*
PCSK7	10	1.17e-03	22.24 (2.45, 201.73)*	0.96 (0.46, 2.02)	0.53 (0.25, 1.11)
PTPN13	5	1.26e-01	2.62 (0.34, 20.08)	3.67 (0.15, 91.55)	2.38 (0.19, 30.33)
UCN	8	7.82e-01	10.72 (1.76, 65.49)*	0.36 (0.03, 4.02)	0.57 (0.41, 0.78)*
CTSB	4	9.44e-01	1.68 (0.06, 47.93)	1.9 (0.01, 327.35)	0.9 (0.36, 2.26)
ABCA1	21	1.11e-02	2.09 (0.59, 7.36)	0.83 (0.58, 1.19)	3.33 (1.39, 7.96)*
LIPC	26	4.95e-01	1.45 (0.76, 2.76)	1.09 (0.93, 1.29)	1.72 (1.05, 2.81)*
C2	5	1.34e-01	0.05 (0.0, 22.82)	1.16 (0.4, 3.36)	0.41 (0.07, 2.33)
ANGPTL4	5	8.41e-01	2.8 (0.63, 12.39)	0.43 (0.05, 4.09)	0.94 (0.01, 122.83)
TNXB	5	7.22e-01	2.54 (1.45, 4.43)*	0.53 (0.07, 3.85)	1.05 (0.28, 3.92)
FEN1	11	6.08e-01	0.94 (0.69, 1.28)	1.81 (0.81, 4.06)	1.06 (0.62, 1.81)
GSTM4	4	6.94e-01	3.46 (2.01, 5.94)*	0.28 (0.07, 1.09)	0.14 (0.01, 3.47)
PCSK9	20	5.21e-03	2.39 (1.45, 3.96)*	1.01 (0.28, 3.6)	0.78 (0.24, 2.49)
LILRA3	9	5.22e-01	0.07 (0.01, 0.81)*	1.01 (0.65, 1.56)	0.87 (0.12, 6.31)
RPS9	6	8.20e-01	1.67 (0.17, 16.86)	0.83 (0.46, 1.47)	1.24 (0.1, 14.84)
FPR1	5	3.47e-01	0.65 (0.2, 2.12)	1.6 (0.31, 8.36)	0.66 (0.01, 32.0)
OBP2B	9	9.50e-01	1.25 (0.87, 1.8)	2.13 (0.75, 6.1)	0.07 (0.01, 0.62)*
INSR	6	7.00e-01	6.78 (0.6, 76.86)	19.37 (0.94, 400.01)	16.08 (1.15, 224.46)*
TNKS	4	2.69e-01	1.16 (0.05, 24.78)	0.25 (0.0, 14.36)	0.44 (0.12, 1.56)
SLC22A1	13	5.143e-05	2.73 (1.53, 4.9)*	1.16 (0.16, 8.61)	0.3 (0.04, 2.2)
LPL	27	6.78e-03	0.17 (0.03, 1.07)	0.28 (0.1, 0.78)*	0.48 (0.17, 1.35)
TSSK6	4	7.39e-01	55.37 (0.0, 16133127.76)	10.04 (0.02, 5084.4)	0.04 (0.0, 3586.37)
EMILIN3	7	7.29e-02	1.28 (0.69, 2.36)	0.4 (0.02, 7.21)	4.03 (0.18, 90.18)
NDUFA13	4	7.39e-01	55.95 (0.0, 16832999.76)	10.07 (0.02, 5139.88)	0.04 (0.0, 3654.32)
BACE1	6	4.52e-02	5.17 (0.34, 79.25)	1.7 (0.78, 3.73)	0.43 (0.17, 1.07)
LILRA5	9	7.70e-01	0.08 (0.01, 0.59)*	0.92 (0.59, 1.44)	0.58 (0.08, 4.38)
BCAM	12	1.28e-01	1.23 (0.99, 1.53)	1.42 (0.62, 3.28)	0.68 (0.49, 0.95)*
FPR3	5	3.56e-01	0.66 (0.2, 2.15)	1.58 (0.31, 7.93)	0.63 (0.01, 28.98)
HAPLN4	4	6.89e-02	4.0 (0.2, 80.12)	46.52 (0.79, 2724.55)	0.48 (0.03, 8.07)
HLA-DRB1	8	4.04e-02	1.01 (0.47, 2.17)	1.0 (0.57, 1.75)	1.3 (0.51, 3.3)
SFTA2	4	8.37e-02	0.62 (0.24, 1.63)	1.36 (0.35, 5.34)	1.92 (0.25, 14.86)
IGF2R	16	1.23e-04	3.75 (2.25, 6.25)*	0.35 (0.07, 1.73)	0.24 (0.07, 0.8)*
HSD17B11	4	4.23e-01	0.4 (0.02, 6.55)	0.61 (0.08, 4.8)	1.57 (0.33, 7.53)
LPA	9	1.47e-02	4.44 (2.43, 8.13)*	1.01 (0.47, 2.21)	1.03 (0.37, 2.91)
TOP1	7	6.86e-01	1.32 (0.81, 2.14)	0.47 (0.09, 2.45)	4.61 (0.49, 43.81)
PSMB8	7	8.47e-01	3.36 (1.76, 6.4)*	0.85 (0.21, 3.41)	0.73 (0.39, 1.38)
HLA-DRA	6	3.49e-01	1.06 (0.57, 1.95)	0.86 (0.39, 1.91)	7.48 (3.04, 18.39)*
NOTCH4	15	3.53e-02	1.1 (0.46, 2.67)	2.03 (1.5, 2.75)*	0.99 (0.54, 1.8)
AGER	11	6.88e-03	2.73 (0.8, 9.29)	3.13 (1.89, 5.19)*	2.27 (0.71, 7.25)
EHMT2	5	1.05=e-01	328.24 (7.85, 13716.69)*	0.53 (0.2, 1.39)	0.04 (0.0, 0.38)*
SLC44A4	5	1.07e-01	322.41 (7.63, 13629.22)*	0.53 (0.2, 1.4)	0.04 (0.0, 0.39)*
NEU1	4	NA	1.18 (0.0, 690.62)	0.91 (0.31, 2.68)	0.09 (0.01, 0.91)*
HSPA1B	4	NA	615016.56 (120.95, 3127274364.99)*	0.21 (0.01, 3.43)	0.11 (0.01, 1.23)
HSPA1A	4	NA	224722364.93 (797.77, 63301607264984.07)*	0.0 (0.0, 0.09)*	48.9 (1.62, 1472.63)*
C6orf25	4	NA	0.0 (0.0, 1.05)	194.1 (0.82, 45848.5)	0.3 (0.07, 1.29)
ABHD16A	4	NA	641.94 (13.43, 30693.48)*	0.04 (0.0, 0.39)*	2.54 (0.53, 12.2)
APOM	4	NA	292.41 (8.42, 10153.82)*	0.6 (0.19, 1.96)	0.26 (0.05, 1.23)
NCR3	4	4.22e-01	5.55 (0.71, 43.25)	0.21 (0.04, 0.97)*	0.63 (0.15, 2.66)
HLA-C	15	4.35e-04	1.31 (0.77, 2.22)	2.72 (1.09, 6.79)*	0.93 (0.7, 1.24)
C6orf15	11	8.77e-01	5.24 (2.57, 10.65)*	0.66 (0.45, 0.96)*	0.37 (0.21, 0.64)*
DDR1	4	1.52e-01	0.74 (0.25, 2.2)	7.09 (0.34, 146.43)	0.67 (0.06, 8.05)
GSTM2	4	7.11e-01	4.4 (0.99, 19.64)	0.28 (0.07, 1.13)	0.03 (0.0, 8312.8)
CEACAM16	13	3.85e-01	1.34 (1.01, 1.79)*	2.29 (0.72, 7.3)	1.21 (0.69, 2.11)
C4B	5	9.04e-01	2.29 (1.25, 4.18)*	0.34 (0.06, 1.97)	1.35 (0.77, 2.39)
APOC4-APOC2	11	1.35e-02	1.37 (1.21, 1.56)*	0.62 (0.39, 0.98)*	0.58 (0.44, 0.77)*
LTA	4	3.72e-01	7.6 (1.01, 57.24)*	0.26 (0.06, 1.12)	0.44 (0.11, 1.74)
LTB	4	3.76e-01	7.61 (1.04, 55.69)*	0.26 (0.06, 1.11)	0.44 (0.11, 1.69)
CYP21A2	4	6.71e-01	2.22 (1.09, 4.55)*	0.32 (0.05, 2.06)	1.47 (0.44, 4.93)
TNF	4	3.73e-01	7.69 (1.03, 57.64)*	0.26 (0.06, 1.11)	0.44 (0.11, 1.71)
HLA-B	10	2.98e-02	1.85 (1.2, 2.84)*	1.36 (0.65, 2.84)	1.1 (0.85, 1.42)
APOC2	12	4.74e-14	1.14 (0.71, 1.82)	0.48 (0.14, 1.67)	0.66 (0.32, 1.36)
HLA-DQA2	13	2.58e-01	1.06 (0.77, 1.44)	3.59 (2.12, 6.09)*	1.82 (1.22, 2.72)*

Drug target gene	No. variants	Heterogeneity *p* value	LDL-C (OR, 95% CI)	HDL-C (OR, 95% CI)	Triglycerides (OR, 95% CI)
LILRA4	6	8.07e-01	0.06 (0.0, 0.73)*	0.83 (0.49, 1.41)	0.53 (0.05, 5.16)
PSMB9	7	8.47e-01	3.35 (1.75, 6.41)*	0.85 (0.22, 3.38)	0.73 (0.38, 1.39)
HLA-DOB	8	9.52e-01	3.27 (1.97, 5.45)*	1.02 (0.41, 2.49)	0.75 (0.49, 1.15)
EGFL8	6	5.22e-04	1.49 (0.02, 110.15)	8.25 (0.57, 118.64)	0.72 (0.02, 24.37)
CFB	5	1.33e-01	0.05 (0.0, 20.51)	1.14 (0.39, 3.31)	0.4 (0.07, 2.29)
LILRA6	5	5.37e-01	1.63 (0.07, 39.37)	0.72 (0.19, 2.73)	0.46 (0.0, 112.3)
HP	11	8.77e-02	1.82 (1.15, 2.88)*	4.88 (1.89, 12.6)*	0.58 (0.13, 2.64)
ITGB3	4	7.49e-01	2.65 (0.34, 20.42)	0.64 (0.01, 61.4)	1.59 (0.17, 15.05)
RPL17	4	5.36e-01	0.86 (0.09, 8.06)	1.0 (0.47, 2.12)	8.34 (0.45, 155.7)

8 | Summary and Future research

I have used human genetic data linked to medical records, clinical biomarkers and molecular traits to investigate the added value of GWAS and drug target Mendelian randomisation to generate genetic evidence and inform genetically guided pharmaceutical research. This final chapter provides a summary of the findings from the work described in Chapters 4 to 7 and contextualises their contribution to genomic research in drug development.

8.1. Summary

Previous research has shown that human genomics could support drug development by generating evidence for target identification and validation[1–4]. In particular, genome-wide association studies (GWAS) have the potential to systematically and accurately identify disease-specific drug targets across the spectrum of human diseases which addresses one of the key productivity limiting steps in drug development.

In Chapter 4, I described the extent to which the causes of human disease have been addressed by genetic analyses, or by drug development, and the degree to which these efforts overlap. I found that only a small fraction of the 10,901 diseases curated in the human disease ontology has been investigated in drug development (13%; 1,370 out of 10,901) or GWAS (9%; 953 out of 10,901). For those diseases being pursued in clinical phase drug development, only 27% (369 out of 1,370) have been the subject of a GWAS. Furthermore, even for the 349 diseases that are the subject of ongoing clinical phase drug development and have been covered by GWAS, it remains uncertain how many specific target-indication pairings have genetic support. These findings showed poor alignment between the diseases studied by GWAS and those pursued in clinical phase drug development.

To help generate insights into how the GWAS and drug development efforts can be utilised in concert, a sample space of disease and targets was constructed in Chapter 4. The sample space included subsets of target-disease pairings that have been covered by clinical phase drug development and by GWAS which interrogate all possible targets by design. The aim of creating the sample space was to illustrate how some areas can be further exploited. For example, the intersection between targets of approved drugs and diseases studied by GWAS can help identify new indications for existing approved drugs. Similarly, the intersection between targets of drugs under clinical investigation and diseases studied by GWAS can lead

to potential repurposing opportunities of drugs that proved safe but lacked efficacy for their originally intended indication, or for indication expansion of approved drugs.

To increase the interest and investment in human genomics, robust evidence of the value of GWAS and GWAS-based approaches for drug target identification and validation is needed. In Chapter 5, I built on the previous findings described by Nelson *et al.*,[3] and King *et al.*,[2] and calculated an updated estimate of the probability of success in the drug development pipeline of drug target – indication pairings with genetic support. Using a 'truth' set of drug target-indication pairings, I provided further evidence that pairings with genetic support are twice more likely to get approved than those without genetic support (2.18; 95%CI: 1.86; 2.51).

Determining if an approved drug target-indication pair has been rediscovered by genetic associations with the intended indication is directly influenced by the definition of genetic evidence. In Chapter 5, I investigated a 'truth' set of drug target-indication pairings of approved drugs and found that using a stringent *p* value threshold to select significant associations may lead to an oversight of true genetic associations and relaxing the *p* value threshold to 5×10^{-6} increased the percent of rediscoveries by 32% on average. Moreover, in 21% of the genetic association - drug target gene - indication combinations explored the closest protein-coding gene was the target gene, and the target gene was in the top five closest genes in 43% of the cases.

Whilst the work described in Chapter 5 supports drug target identification by help map drug targets, the information that can be derived from a GWAS alone cannot be directly used to inform drug target validation as one cannot readily infer simply from the identification of the locus the mechanism of action of the drug (i.e., an inhibitor or activator for enzymes, or blocker or antagonist for receptor targets). To develop a drug targeting hypothesis for a new drug, the *cis*-Mendelian Randomization (MR) approach (also refer to as *drug target MR*)[5] has

been proposed. Several examples exist that describe the application of drug target MR using circulating protein levels to instrument the on target drug effect. However, the vast majority of successful drugs achieve their effect by binding to and modifying the activity or function of a protein[6]. Therefore, the drug target MR analyses that use as exposure circulating protein levels (pQTL) make the assumption that pQTL are a valid proxy of protein activity. In Chapter 6, I identified two proteins (BCHE and coagulation factor VII) for which genetic associations for protein levels and activity was available and showed that a strong correlation between genetic associations with activity and level for variants acting in *cis*- exist (Pearson's correlation coefficient for the BCHE was $\rho = 0.99$ and for coagulation factor VII $\rho = 0.96$).

Although only two proteins could be included in the comparison due to the lack of GWAS on protein activity, previous drug target MR studies that used pQTL data were able to recapitulate the mechanism of action of known drugs. Therefore, under the assumption that protein levels are a valid proxy for protein activity, I evaluated in Chapter 6 the performance of the drug target MR framework using a 'truth' set of drug target gene-indication pairings, where circulating levels of the target protein have been measured by a high-throughput proteomic platform and the indication has been studied by GWAS. After integrating information from GWAS on disease and clinical endpoints, and genetic associations on circulating protein levels measured by SOMAmers (i.e., short single-stranded oligonucleotides that bind with high affinity and specificity to a protein and enable the quantification of its levels), I identified a 'truth' set of 320 SOMAmer-drug target gene-traits pairs. The application of the drug target MR framework recapitulated the mechanism of action of several drug target gene – indication pairings under different models: PCSK9 and Coronary Artery Disease, Carotid intima media thickness, Carotid plaque, LDL cholesterol; ACE for Systolic and Diastolic blood pressure; AMY2A and type 2 diabetes Mellitus; ATP1B2 and atrial fibrillation; COMT and Parkinson's disease; F2 and prothrombin levels; IL1R1 and rheumatoid factor;

IMPA1 and bipolar disorder; PDE4A and forced expiratory volume in the first second (FEV1); PDE5A and prothrombin levels; PLG and activated partial thromboplastin time. In contrast, of the drug target gene-SOMAmer-trait combinations that returned significant MR results, 38-50% of the results were consistently in the unexpected direction of effect based on their reported mechanism of action (range 9-26 drug target gene - SOMAmer - trait). Several reasons that could explain the results were presented in Chapter 6, which included assay ambiguity or biological mechanism. However, the findings from the work in this thesis indicate that further research and validation are required before the pQTL-weighted drug target MR approach can be applied systematically for drug target validation.

While protein level exposures are potentially useful for drug target validation and defining a drug targeting hypothesis, if they are unavailable, genetic associations in *cis*- with well-established clinical biomarkers could be used in drug target MR analyses ('biomarker-weighted drug target MR') to inform drug target validation. Here the causal inference remains on the gene product at the target loci, however, the exposure trait may not necessarily be the mediator of the effect, so the drug targeting hypothesis cannot be directly established from the MR estimate direction. Whilst this may seem like a limitation, many GWAS of biomarkers are available allowing for the possibility of independent replication of MR associations. In Chapter 7, I combined publicly available GWAS datasets on blood lipids and coronary heart disease and to genetically validate and prioritise drug targets for CHD prevention. The aim was to illustrate the utility of biomarker-weighted drug target MR in high power settings with independent replication data to prioritise drug targets for CHD prevention. Out of the 341 drug targets identified through their association with blood lipids (HDL-C, LDL-C and triglycerides), 30 targets that might elicit beneficial effects in the prevention or treatment of CHD were robustly prioritized, including NPC1L1 and PCSK9, the targets of existing drugs used in CHD prevention.

8.2. Research in context

The success of GWAS as a method is multi factorial from the completion of the Human Genome Project in 2003[7], the rapid declining cost of genotyping, and the increasing number of international consortia and joint-pharma partnerships. This coupled with the accessibility of public data repositories has given the opportunity to screen multiple drug targets against multiple diseases. However, despite the growing interest of the pharmaceutical industry in using human genomic data to help prioritise drug development programmes and reduce the risk of clinical-stage failure, genetic studies of human diseases and pharmaceutical research and development have largely proceeded independently.

Therefore, the analysis presented in Chapter 4, where the extent to which the causes of human disease have been addressed by genetic analyses, or by drug development, and the degree to which these efforts overlap was investigated, could have several applications. First, it could be used to inform future drug development programmes direction if they are seeking to exploit existing genetic evidence. Secondly, it identified diseases without effective treatments that could be prioritised in large-scale GWAS or sequencing studies to help stimulate drug development in the disease area. Third, the sample space of human targets and diseases could help identify opportunities to expand the indications for approved drugs or discover repurposing opportunities for the many safe drugs that failed in clinical trials because of lack of efficacy in the originally intended indication.

In addition, the sample space of human targets and diseases could also inform *de novo* drug development for druggable targets and disease indication pairings that have yet to be investigated. In particular, soluble or secreted protein targets could especially benefit from having genetic support for a particular disease since such proteins are readily targeted by monoclonal antibodies or peptides, which typically exhibit higher selectivity and reduced

development timelines compared to small molecules[8]. Information on the set of human secreted proteins (the human 'secretome'[9]) is available in the public domain, and researchers and the pharmaceutical industry could use these resources to identify high priority putative circulating protein targets.

Part of the analysis in Chapter 4 was restricted to the genes encoding druggable protein targets (the 'druggable genome'[4]), which is currently defined as the set of *proteins* with potential to be modulated by a drug-like small molecule or monoclonal antibody. However, novel therapeutic modalities, such as RNA silencing or gene editing, are likely to expand the range of potential druggable targets[10–12]. In addition, artificial intelligence and the application of data-driven approaches and computer modelling have revealed protein motifs unknown before, turning undruggable protein targets into druggable ones[13].

The increasing interest of the pharmaceutical industry in human genomics has been driven by several retrospective studies showing that selecting genetically supported drug targets could double the success rate in clinical development. In Chapter 5, it was shown that 898 drugs exist with a license for 371 therapeutic indications. Out of 371 therapeutic indications, 144 have been well-studied by GWAS, and thus, offered an unique opportunity to retrospectively investigate how many of the 1,969 drug target gene – indication pairings had been rediscovered by GWAS. Previous work by Nelson *et al*., 2015[3] and King *et al.,* 2019[2] that used a similar approach and study design showed that targets with genetic evidence from GWAS were more likely to be successful in clinical development as indicated by the ratios of the probability of progressing in the drug development pipeline given genetic support to the probability of progressing without genetic support of 1.8 (95% CI: 1.3; 2.3) and 1.4 (95% CI: 1.1; 1.7), respectively. The findings from Chapter 5, which were based on two approaches for genetic evidence and the larger dataset of target-indication pairings to date, confirmed that the

probability of a target-indication pair with genetic support progressing from phase I to approval to the probability of progressing without genetic evidence is greater than two-fold (2.18; 95%CI: 1.86; 2.51). In line with previous observations, I also found that the probability of progression given genetic support increases along the clinical phases and that the lack of genetic support had the greatest impact from phase II to phase III (P(S-|G-) / P(S-|G+) = 1.40, 95%CI: 1.28; 1.56), where drugs are typically tested for clinical efficacy. Notably, variability was found among the proportion of approved target-indications pairs by indication area. Such stratification was also performed by Nelson *et al*.,[3] however, the rank of disease areas by genetic support presented in this thesis differed from the previous publication. For example, while Nelson *et al.,*[3] showed that target – indication pairs in the musculoskeletal disease area had the highest degree of genetic support, the analysis in Chapter 5 identified target - cardiovascular indication pairings as the ones with the highest support. The differences observed could be explained by the larger dataset used in Chapter 5, which, for instance, included 115 approved target – musculoskeletal indication pairings in contrast to the 11 identified by Nelson *et al*.[3].

Of note, these studies rely on assigning genetic associations from GWAS data to a causal gene, which remains a challenge because association signals from variants in high linkage disequilibrium (LD) may span multiple genes. Several approaches have been proposed to assign GWAS signals to genes (e.g., co-localisation using eQTL data[14]), however, physical proximity remains the simplest and most widely used approach to map association signals to causal genes[15,16]. While the closest gene may not always be the putative causal gene[17,18], there are several studies of 'truth' sets of genes with well validated causal relationships to disease that have shown that the closest gene to a GWAS signal is the causal gene in about two-thirds of cases[15]. Furthermore, in Chapter 5, I used a 'truth' set of approved drug target-indication pairings where the indication has been studied by GWAS[4,19] and showed that relative distance

(i.e., distance rank of the druggable gene from the GWAS SNP) to the gene rediscovered more drug target-indication pairs than the use of the absolute distance. Similarly, a study by Mountjoy *et al*., 2021[19] and funded by OpenTargets which evaluated different genomic features to assign GWAS signals to causal genes reported that, the 'mean distance' feature was the most predictive (which combined a distance and Bayes factor approach[20]), where the distance relative to other genes was more important than the absolute distance. In addition, the use of absolute distance to map association signals to genes is challenged by the lack of consensus on how much the genomic region should extend around the potential causal gene. In Chapter 5, I found that in 27% of the drug target gene – indication pairs, genetic associations with the indication were within 1 Mega base pair (Mbp) from the drug target gene, and that increasing the genomic distance beyond 1 Mbp led to a change in the curve from exponential to logarithmic suggesting that further increasing the region would lead to rediscoveries, however, at the cost of increasing the median number of protein-coding genes between the target gene and the genetic association. In fact, in a recent publication Fauman *et al.*[21] estimated a distance cut-off of 944 kbp (95%CI 767-1,161) separating the *cis* (i.e., the QTL is acting through the cognate gene) and *trans* (i.e., assumes that the QTL is acting through an intermediate gene) regimes, which in line with the findings in Chapter 5, suggests that approaches for mapping genetic associations to genes based on distance should be restricted to a maximum of 1 Mbp.

While genetic associations obtained through GWAS can support target identification, they cannot be readily used to infer the mechanism of action of a drug targeting the protein encoded by the associated gene. One would have identified a potential target for a particular disease, but how to perturb the target to obtain the intended effect cannot be drawn from a GWAS association, even if the causal gene can be inferred with certainty. To inform the design of an inhibitor or activator (blocker or antagonist for receptor targets), the *cis*-Mendelian Randomization (MR) approach ('drug target MR')[5] has been proposed. Since the vast majority

of successful drugs achieve their effect by binding to and modifying the activity of a protein[6], an ideal drug target MR analysis would assess the effect of modulating protein activity or function with respect to disease risk using genetic instruments in the encoding gene. This would determine whether and by how much an increase or decrease in the protein function or activity impacts disease risk, suggesting a plausible mechanism of action for the drug. However, GWAS data on protein activity is scarce and there are not examples in the literature of MR analysis where the exposure has been instrumented using genetic associations with protein activity. Recently, genetic associations with circulating protein levels have been used instead as a proxy for protein function or activity.

To explore such assumption, in Chapter 6 I identified two proteins (BCHE and coagulation factor VII) for which genetic associations with protein levels and activity was available and found a strong correlation between genetic variants acting in *cis*- (Pearson's correlation coefficient for the BCHE was $\rho = 0.99$ and for coagulation factor VII $\rho = 0.96$). Several examples support this, such as the drug target Mendelian randomization of CETP or PCSK9 protein concentration which replicated on-target effects previously reported in clinical trials[5,22].

To test the generalisability of the drug target MR framework described by Schmidt *et al*., 2020[5] to multiple targets and diseases, I performed a systematic evaluation of the performance of pQTL-weighted drug target MR analyses using a 'truth' set of 160 licensed drug target – indication pairings for which pQTL associations were available for the target protein and the intended indication or a clinically-relevant disease biomarker had been studied in GWAS. Only 11-13% of the combinations explored across all possible parameters (i.e., 27/234 in the sensitivity analysis 1 and 16/121 in the sensitivity analysis 3) recapitulated the known mechanism of action of the approved drug. Nonetheless, this represents a 2-fold increase

compared to a previous study by Zheng *et al*.[23], in which MR was applied to 1,002 proteins and 225 phenotypes, and identified four drug target gene – approved indication pairs for which the MR recapitulated the mechanism of action out of 73 pairs with potential to be rediscovered. Two of the four drug target gene – approved indication pairs were also rediscovered by the analysis presented in Chapter 6: PCSK9 for hypercholesterolemia and hyperlipidaemia; and ACE for hypertension. The other two rediscoveries of Zheng *et al*.,[23] were TNFRSF11A and osteoporosis; and IL12B for psoriatic arthritis and psoriasis. In the work described in Chapter 6, TNFRSF11A showed a concordant direction of effect when using heel bone mineral density as the outcome, however the association did not reach the significance threshold. On the other hand, the association between IL12B and psoriasis was in the unanticipated direction of effect in some of the scenarios, although most of the combinations analysed did not yield significant results. Out of the two drug target gene – indication pairs found by Zheng *et al*., [23] in the unexpected direction of effect, IL6R was also in the unanticipated direction of effect in the analysis described in Chapter 6, while PROC was not analysed because it is not recorded as the target of an approved drug in ChEMBL. In their work, Zheng *et al*.,[23] in line with the discussion in Chapter 6, indicated that for IL6R the alleles associated with higher soluble protein levels have been shown to also lead to lower intracellular pathway activation[24], suggesting consistency of direction with the therapeutic approach.

In addition to PCKS9 and ACE, the analysis in Chapter 6 further identified nine target-indication pairs that consistently showed a concordant direction of effect under all the models. The rediscovery of IL1R1 and rheumatoid factor and PLG and activated partial thromboplastin was in line with the afore mentioned study by Schmidt *et al*.[5], which used a set of selected positive controls to illustrate the drug target MR framework. Another pair, ACE and hypertension, was rediscovered by Zheng *et al.*[23] using pQTL data and such genetic associations have also been used to instrument the effect of modifying ACE circulating levels

on different outcomes[25], including susceptibility to SARS-CoV-2 infection or COVID-19 severity[26]. Two of the drug targets rediscovered were phosphodiesterases (PDE4A and FEV1, PDE5A and prothrombin levels), and although MR studies have been performed on different outcomes[27], these have not included the intended indication. Similarly for coagulation factor II, as MR studies on coagulation factors and the intended indication (venous thrombosis) have been published using intermediate traits such as activated thromboplastin time as the exposure[28], however, drug target MR analyses using F2 pQTL data have not previously been reported. The remaining pairs (AMY2A and type 2 diabetes Mellitus; ATP1B2 and atrial fibrillation; COMT and Parkinson's disease) have not been previously studied in drug target MR analyses of the intended indication using pQTL data.

On the other hand, 38-50% of the pairs with significant MR estimates were consistently in the unexpected direction of effect based on their reported mechanism of action (range 9/24-26/53 drug target gene-SOMAmer-trait under the different scenarios). These findings highlight that additional evaluation and refinement of the pQTL-weighted drug target MR methodology is required. Other techniques such as co-localization[29] could be used to source additional evidence by investigating if the estimate obtained in the drug target MR analysis is not attributable to genetic confounding through a variant in linkage disequilibrium[30].

Another issue of the pQTL-weighted drug target MR analysis described in Chapter 6 was the lack of significant genetic associations that could be used to instrument the drug effect, an issue that affected 86 of the 320 SOMAmer-drug target gene-trait combinations in the dataset. The commercialisation of cost-effective high-throughput technologies for protein measurement and the linkage of national biobank to electronic health records would enable larger sample sizes, and thus increase the power to detect significant associations. That is the promise of the genetic associations identified by deCODE genetics using the SomaLogic 5K platform in

35,559 Icelanders[31], or the UK Biobank Pharma Proteomics Project[32] which aims to measure circulating concentrations of up to 1,500 plasma proteins in approximately 53,000 UK Biobank participants using the Olink technology. Furthermore, the number of proteins covered by the proteomics platform is increasing, with the latest SomaLogic and Olink assays measuring up to 7,000[33] and approximately 3,000 proteins[34], respectively, which will allow for an increase coverage of the sample space of target and human diseases in drug target MR analyses.

While opportunities for pQTL data in drug target MR analysis continue being evaluated, drug target MR analyses using genetic associations with 'biomarkers' downstream to the target protein could be used to prioritise drug targets. Biomarker-weighted drug target MR analyses do not provide evidence on whether the biomarker used for the weighting itself mediates disease, but they inform on the validity of the drug target for a disease, regardless of the mediating pathway. This approach was used in Chapter 7 to systematically prioritise drug targets for CHD prevention. Out of the 341 drug targets identified through their association with blood lipids (HDL-C, LDL-C and triglycerides), 30 targets that might elicit beneficial effects in the prevention or treatment of CHD were robustly prioritized, including NPC1L1 and PCSK9, the targets of licensed drugs used in CHD prevention. Of note, the mechanism of action of PCSK9 through LDL-cholesterol was also rediscovered by the pQTL-weighted drug target MR in Chapter 6 which supports the assumption that the effect on a clinically-validated biomarkers could be a valid proxy for protein concentration or activity. The same analysis prioritised other potential targets such as the lipoprotein lipase (LPL), a target that could potentially decrease CHD risk based on the univariable MR findings, with an effect potentially mediated by HDL-C, or another non-LDL-C pathway. Several pharmacological attempts have been pursued to target LPL[35,36], and the approval of gene therapy interventions and the known indirect activation of LPL by drugs targeting other proteins, such as fibrates[37] and metformin[38],

suggest that the previous failure of compounds targeting LPL in initial trials may have been idiosyncratic.

The potential of 'biomarker-weighted drug target MR' was illustrated in Chapter 7 using genetic association data on blood lipids and CHD data, however, the approach could also be extended to other areas where GWAS of diseases and biomarkers thought to be potentially affected by the drug target are available. For example, 'biomarker-weighted drug target MR' could leverage the increasing available data on cardiovascular biomarkers to evaluate the causal role of drug targets, such as carotid artery intima media thickness and carotid plaque, in atherosclerosis, following up on associations described in several studies[39,40], to identify potential new indications for anti-inflammatory agents established in the treatment of autoimmune conditions.

8.3. Thesis strengths and weaknesses

The work described in this thesis have a series of strengths and limitations that were discussed at length within each results chapters. Here I summarise those that were present throughout all the analyses.

One of the strengths of the analyses is that most of the data used were available in the public domain which facilitates the revisiting of the estimates if needed, reproducibility and look up of canonical examples. These datasets included repositories of genetic associations, databases of drug information and clinical trial data, and published lists of druggable genes. Information from these disparate sources was integrated in the thesis using different anchoring ontologies or coding systems. For example, human diseases, drug indications and phenotypes investigated by GWAS were connected using the UMLS system. By using the UMLS as an anchoring ontology to standardise the diseases across data sources and including a step of manual curation of the disease terms and areas, the error due to inaccurate mapping cross-databases was reduced. The effort of harmonising the disease nomenclature facilitated the stratification of the analyses in Chapter 4 and 5 by disease area. This represents an additional strength of the work presented as allowed for the identification of disease areas with unmet clinical need (Chapter 4), or disease areas where targets had the greater genetic support.

In addition, I created a dataset of 32,022 drug target-indication pairs using data from ChEMBL v25 and the druggable genome to estimate the value of genetic support in phase progression and to derive a 'truth' set of approved drug target-indication pairs. The dataset included 10,000 more pairings compared to the target-indication pairs reported by King *et al*., 2019[2] (21,934) and that used by Nelson *et al.*[3], 2015 (19,085). When filtering for those indications that had been studied by GWAS, the dataset included 18,065 drug target-indication

pairs in contrast to the 820 investigated by Nelson *et al*., (precise numbers from the King *et al*., were not provided in the publication). One of the potential reasons for the increased sample size may be that the analysis performed in this thesis included not only GWAS data from studies based on research-based case ascertainment, but also genetic associations from electronic health records (UK Biobank). Also in terms of sample size, the analysis presented in Chapter 5 utilised genetic associations with the levels of almost 5,000 circulating proteins measured in a large cohort (10,708 participants). The high number of targets of approved drugs with available pQTL allowed for a large scale evaluation of the drug target MR framework using 160 drug target-indication pairings. The only similar systematic analysis used a proteomic platform for 1,000 proteins and thus, could only evaluate 73 approved drug target - indication pairings.

Lastly, multiple testing in the MR analyses was addressed in a number of complementary ways throughout the thesis. In Chapter 6, several sensitivity analyses were performed using different conditions for the parameters. In Chapter 7, multiple sources of evidence were combined to prioritise drug targets. For example, to assess the potential for false positive results, the distribution of the exposure-specific *p* values was tested against the uniform distribution expected under the null hypothesis[41]. In addition, the findings were validated with independent data sources and a second drug target MR was conducted. Also, a multivariable extension of the inverse-variance weighted (IVW) and MR Egger methods was applied in Chapter 7 to further validate the findings, although in some cases imprecise estimates were obtained in line with previous studies which attributed this to the inclusion of highly correlated exposures in the model[42].

There are some general limitations to the analyses presented in the preceding chapters. First, information on drugs in preclinical or clinical development may be incomplete or not

available in the public domain, which may lead to an underestimation of the number of diseases studied in drug development, particularly for the preclinical candidates which did not progress to clinical trials. Second, there are several reasons for drug discontinuation besides lack of efficacy, including safety concerns, strategic decisions or the compound failing to show extra benefits compared to another treatment. This could affect the estimates derived in Chapter 5 as it was assumed that drug target-indication pairs not progressing in the development pipeline were primarily due to lack of efficacy. Another potential source of bias is that genetic evidence from GWAS may already be used to inform drug development. However, in line with the argument presented by Nelson *et al*., 2015[3] and due to the long timelines in drug development (on average 10 years), the impact of this bias would not inflate the estimate but rather underestimate the value of genetic support as it would increase the number of drugs with genetic support in the early phases of the development process.

From Chapter 4 to 6, diseases and indications studied by GWAS were identified using information in the GWAS Catalog. However, the set of diseases/indications may not include certain indications that may have been studied by GWAS but whose summary statistics had not been not deposited in the GWAS Catalog. Even for those GWAS traits included in the analysis, genetic associations may have been missed due to sample sizes not being large enough to detect all the responsible genes; or due to incomplete genomic coverage by the genotyping array. Furthermore, summary statistics deposited in the GWAS Catalog may be incomplete and lack essential information for the MR analyses, such as effect sizes or effect/reference alleles.

In the pQTL-weighted drug target MR analyses described in Chapter 6, it was assumed that protein expression levels (pQTL) can be used as a proxy of protein activity or function. While two examples are provided at the beginning of the chapter which supports such assumption, this has not been studied in detail due to the lack of GWAS data on protein activity.

Moreover, protein levels corresponded to circulating protein in plasma, although some proteins are not secreted or circulating in plasma, and therefore, their presence in the blood tissue rather than indicate physiological conditions. Since the function of these proteins should take place in a different tissue, it is unclear if the levels in plasma recapitulate those in the drug effector tissue, or, on the contrary, they are unrelated to their function and should not be used to infer the effect of modifying such protein by a drug.

The drug target MR approach used in Chapter 6 and 7, which utilised genetic variants in *cis*- to construct the genetic instrument. As described in Chapter 2, this approach is less prone to violation of the horizontal pleiotropy assumption than MR analyses with *trans* instruments[5]. However, *cis*-MR also requires some decisions to be made regarding instrument selection: defining the locus of interest, the significance threshold for the association with the exposure and the LD threshold to prune correlated instruments. The evaluation of the drug target MR framework in Chapter 6 suggested that the choice of parameters should be made on a *case-by-case* basis. Therefore a window of 50 kbp and LD threshold of 0.4 were used, which showed the most consistent estimates in a grid-search in the discovery data using the four positive control examples: PCSK9, NPC1L1, HMGCR and CETP. Based on previous studies showing that using less stringent *p* value thresholds often results in improved performance in *cis*-MR settings (i.e., effect in the anticipated direction), the threshold below genome-wide significance was relaxed to select the genetic associations to instrument the exposure; and accounted for LD correlation by pruning and LD modelling during the MR analysis[5,43].

Lastly, some of the analyses presented in this thesis only included genes regarded as encoding druggable proteins which currently comprise approximately 25% of all protein coding genes[4]. As knowledge advances, additional proteins will become druggable, and

alternative therapeutic strategies such as antisense oligonucleotides and gene therapy may extend the range of mechanisms that can be targeted.

8.4. Concluding comments

The findings from this thesis have the potential to inform prioritisation strategies in drug development and future research so the investment and impact of human genetic studies can be maximised. It provides an overall picture of the drugs, targets and indications where genetic data exist and could be harnessed to genetically validate approved drugs or identify opportunities for indication expansion, repurposing or *de novo* drug development. It also demonstrates through retrospective analyses of drug target-indication pairings that those with genetic support are enriched among successful drug development programmes. Several molecular traits, including proteomics and other clinically relevant biomarkers, are now being measured and linked to medical records and genetic data in large cohort studies and national biobanks. Therefore, it is possible to optimise traditional approaches in genetic epidemiology, such as Mendelian randomisation, to harness genome-wide association studies and provide robust evidence of target efficacy in early stages of the drug development process. The drug target MR framework using genetic associations with protein levels holds the promise of genetically validating drug target – indication pairs by the systematic interrogation of every potential drug target with available pQTL data against all the potential indications studied by GWAS. Further work is still needed to fully understand and validate the approach before it can be applied systematically, but several case studies have been described in this thesis which illustrate its potential and support future research on the topic.

8.5. Future research

The findings from this thesis have the potential to generate future research in several directions.

The points of convergence or divergence between genomic research and drug development efforts identified in the sample space of all the human drug targets and diseases could have multiple applications: to inform future drug development programmes direction if they are seeking to exploit existing genetic evidence; to promote large-scale GWAS or sequencing studies to help stimulate drug development in diseases without an approved treatment; to identify opportunities to expand the indications for approved drugs or repurposing opportunities for the many safe drugs that failed in clinical trials due to lack of efficacy in the originally intended indication.

The declining cost of high-throughput technologies for protein quantification and the linkage of molecular measures to genetic data and electronic health records offer opportunities to conduct GWAS in a large number of patients and also on quantitative traits in healthy subjects to identify genetic associations that may explain differences, for example in protein levels. At the end of 2021, deCODE made available to the public genetic associations for almost 5,000 proteins measured in 35,559 Icelanders using the SomaLogic v4 platform[31]. These data could be meta-analysed with other pQTL GWAS to increase the sample size and increase the power to detect significant associations. It could also be leveraged in drug target MR studies to replicate the findings described in Chapter 6 or to identify opportunities for expansion of indications for those drug targets-approved indications that were consistently in the anticipated direction of effect. Similar work could be conducted using the genetic associations with 1,500 plasma proteins measured in approximately 53,000 UK Biobank participants using the Olink

technology. That resource, which has not been released at the time of this thesis, would also allow for cross-platform comparison.

Measurements of circulating levels will soon be available for a wider range of proteins based on the latest assays announced by SomaLogic v4.1 (7,000 proteins[33]) and Olink (3,000 proteins)[34]. Such data will increase the coverage of the sample space of target and human diseases in drug target MR analyses.

Future research should also focus on the methodological aspects of GWAS and MR approaches. Twenty years after the publication of the first GWAS[44], it remains unclear what is the most optimal method to map association signals to causal genes and several gold-standard datasets have been used to explore the different methodologies. These 'truth' sets include genes whose perturbation causes a Mendelian form of a common disease[45], the set of expression and protein QTLs[21], curated metabolite QTLs[15], manually curated examples from the literature[19], and approved drug target-indication pairings where the indication has been studied by GWAS[4,19]. As more data become available, these datasets are likely to expand and thus offer a larger sample size to test different methods. In addition, novel approaches may emerge that outperform the current mapping methods which are mostly based on the relative distance to the gene. Research on MR techniques will also benefit from the enhanced mapping between genetic association – causal gene, as it will ensure that valid genetic variants are selected to construct the genetic instrument. Similar 'truth' sets could be used in future work to evaluate the performance of the drug target MR framework and inform the design, parameter selection and interpretation of the findings.

The full integration of genome-wide association studies and related applications in the drug development pipeline is still very much a work-in-progress. This thesis anticipates that

the mining of data from genome-wide association studies will help address the efficiency and productivity problem in the pharmaceutical industry.

8.6. References

1. Hingorani, A. D. *et al.* Improving the odds of drug development success through human genomics: modelling study. *Sci Rep* **9**, 18911 (2019).
2. King, E. A., Davis, J. W. & Degner, J. F. 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**, e1008489 (2019).
3. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856–860 (2015).
4. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, (2017).
5. Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nature Communications* **11**, 3255 (2020).
6. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery* **1**, 727–730 (2002).
7. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
8. Shepard, H. M., Phillips, G. L., Thanos, C. D. & Feldmann, M. Developments in therapy with monoclonal antibodies and related proteins. *Clin Med (Lond)* **17**, 220–232 (2017).
9. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nature Biotechnology* **28**, 1248–1250 (2010).
10. Zaafar, D., Elemary, T., Hady, Y. A. & Essawy, A. RNA-targeting Therapy: A Promising Approach to Reach Non-Druggable Targets. *Biomedical and Pharmacology Journal* **14**, 1781–1790 (2021).
11. Fellmann, C., Gowen, B. G., Lin, P.-C., Doudna, J. A. & Corn, J. E. Cornerstones of CRISPR-Cas in drug discovery and therapy. *Nat Rev Drug Discov* **16**, 89–100 (2017).

12. Schneider, M. *et al.* The PROTACtable genome. *Nat Rev Drug Discov* **20**, 789–797 (2021).
13. He, H., Liu, B., Luo, H., Zhang, T. & Jiang, J. Big data and artificial intelligence discover novel drugs targeting proteins without 3D structure and overcome the undruggable targets. *Stroke Vasc Neurol* **5**, (2020).
14. Giambartolomei, C. *et al.* A Bayesian Framework for Multiple Trait Colo-calization from Summary Association Statistics. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty147.
15. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, e3–e3 (2019).
16. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* 1–6 (2021) doi:10.1038/s41586-021-03446-x.
17. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat Commun* **10**, 3300 (2019).
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. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* 1–7 (2021) doi:10.1038/s41588-021-00945-5.
20. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294–1301 (2012).
21. Fauman, E. B. & Hyde, C. An optimal variant to gene distance window derived from an empirical definition of cis and trans protein QTLs. 2022.03.07.483314 Preprint at https://doi.org/10.1101/2022.03.07.483314 (2022).

22. Schmidt, A. F. *et al.* Cholesteryl ester transfer protein (CETP) as a drug target for cardiovascular disease. *Nat Commun* **12**, 5640 (2021).
23. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature Genetics* **52**, 1122–1131 (2020).
24. Ferreira, R. C. *et al.* Functional IL6R 358Ala allele impairs classical IL-6 receptor signaling and influences risk of diverse inflammatory diseases. *PLoS Genet* **9**, e1003444 (2013).
25. Gill, D. *et al.* ACE inhibition and cardiometabolic risk factors, lung ACE2 and TMPRSS2 gene expression, and plasma ACE2 levels: a Mendelian randomization study. *Royal Society Open Science* **7**, 200958.
26. Butler-Laporte, G. *et al.* The effect of angiotensin-converting enzyme levels on COVID-19 susceptibility and severity: a Mendelian randomization study. *Int J Epidemiol* **50**, 75–86 (2020).
27. Gaziano, L. *et al.* Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. *Nat Med* **27**, 668–676 (2021).
28. Yuan, S. *et al.* Genetically Proxied Inhibition of Coagulation Factors and Risk of Cardiovascular Disease: A Mendelian Randomization Study. *J Am Heart Assoc* **10**, e019644 (2021).
29. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
30. Gill, D. & Burgess, S. The evolution of mendelian randomization for investigating drug effects. *PLoS Med* **19**, e1003898 (2022).
31. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* **53**, 1712–1721 (2021).

32. UK Biobank launches one of the largest scientific studies. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/uk-biobank-launches-one-of-the-largest-scientific-studies.
33. The SomaScan Platform - Our Science - Platform. *SomaLogic* https://somalogic.com/somascan-platform/.
34. Olink Explore 3072. *Olink* https://www.olink.com/products-services/explore/.
35. Tsutsumi, K. *et al.* The novel compound NO-1886 increases lipoprotein lipase activity with resulting elevation of high density lipoprotein cholesterol, and long-term administration inhibits atherogenesis in the coronary arteries of rats with experimental atherosclerosis. *The Journal of clinical investigation* (1993) doi:10.1172/JCI116582.
36. Yin, W. & Tsutsumi, K. Lipoprotein Lipase Activator NO-1886. *Cardiovascular Drug Reviews* **21**, 133–142 (2003).
37. Schoonjans, K. *et al.* PPARalpha and PPARgamma activators direct a distinct tissue-specific transcriptional response via a PPRE in the lipoprotein lipase gene. *EMBO J.* **15**, 5336–5348 (1996).
38. Ohira, M. *et al.* Effect of metformin on serum lipoprotein lipase mass levels and LDL particle size in type 2 diabetes mellitus patients. *Diabetes Research and Clinical Practice* **78**, 34–41 (2007).
39. Franceschini, N. *et al.* GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nat Commun* **9**, 5141 (2018).
40. Yeung, M. W. *et al.* Twenty-Five Novel Loci for Carotid Intima-Media Thickness: A Genome-Wide Association Study in >45 000 Individuals and Meta-Analysis of >100 000 Individuals. *Arterioscler Thromb Vasc Biol* **42**, 484–501 (2022).

41. Storey, J. D. A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 479–498 (2002).
42. Rees, J. M. B., Wood, A. M. & Burgess, S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Stat Med* **36**, 4705–4718 (2017).
43. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
44. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat Genet* **32**, 650–654 (2002).
45. Forgetta, V. *et al.* An effector index to predict target genes at GWAS loci. *Hum Genet* (2022) doi:10.1007/s00439-022-02434-z.