THE UNIVERSITY *of* EDINBURGH

Integrating Functional Genomics and Semi-Parametric Estimation to Identify Binding Variants Likely Causal for Altering Human Traits

Olivier Labayle

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2025

Abstract

Understanding the genetic architecture of complex human traits is a central challenge in modern genetics with applications in drug development and precision medicine. This thesis presents methodological advancements for the discovery of causal variants affecting human traits. These advancements are grounded in mathematical statistics and functional genomics and supported by extensive simulations and real-world data studies using the UK Biobank.

In the first part of this body of work we introduce a comprehensive mathematical framework for the analysis of genetic effects on traits or disease, including single variant effects, non-linear allelic effects, and higher-order interactions. Genetic effects are formally defined as causal estimands, yet remain difficult to identify, reasons for which are discussed. We then construct semi-parametric estimators for asymptotically unbiased and efficient estimation of associated statistical estimands. Finally, we propose a network approach, based on genetic relatedness to account for non-independent individuals. This statistical advancement is delivered within state-of-the-art software called TarGene. TarGene is designed to provide performant and reproducible semi-parametric estimation routines, scaling to biobank-scale datasets, and compatible with modern high-performance computing platforms.

In the second part, we investigate the empirical performance of these semi-parametric estimators in the context of population genetics, using UK Biobank data. Firstly, this is done via an extensive simulation study, leveraging flexible generative models that can adequately represent the data generating process. Practical violations of theoretical assumptions are illustrated as well as strategies for their mitigation. Secondly, we contrast semi-parametric estimates to published data produced by conventional parametric models. To this end, we perform a phenome-wide association study (768 traits) for a well-established variant with large effect size on the body-mass index (BMI). We observe that p-values obtained via parametric models are substantially smaller than those originating from semi-parametric methods. The absence of overlap between some semi-parametric confidence intervals and those originating from parametric models highlight inflated false discovery rates due to model misspecification. In addition, for 39 traits our method reveals non-linear allelic effects which are commonly overlooked by current practices in linear modelling.

Finally, we propose a paradigm based on functional genetics for the discovery of probable causal variants and the mechanism through which they act on human traits. These variants are likely to be causal for two main reasons: (i) they are experimentally shown to disrupt the binding of a specific transcription factor and are thus biologically active; and, (ii) their effect on traits is modulated via trans-acting variants that were associated with the same mechanism. As a pilot study, we use TarGene to discover putative causal variants acting through the vitamin D receptor. For these variants, a post-analysis is performed to gain more insight into the mechanism of action.

Overall, this thesis advances the field of population genetics in three ways. First, it provides a robust mathematical framework within which the main challenges in the field are formally defined. Second, it addresses the statistical estimation challenge by removing the need for parametric assumptions and delivers an open-source state-of-the-art software. Third, it proposes a paradigm based on functional genomics for the discovery of putative causal variants as well as the mechanism through which they act on human traits.

Lay Summary

Understanding how genetic variations affect human traits and disease is a key challenge in modern genetics. Unravelling this connection is crucial for advancing preventive and personalised medicine, as well as developing more effective drugs. In this thesis, we introduce new methodological advancements that move us closer to achieving these goals.

Over the past two decades, many genetic variants have been associated with human traits. However, association does not imply causation, and many of these variants may have no real impact on human health. New experimental technologies now allow us to identify genetic variants that influence how proteins interact with the DNA at a highly precise level. These interactions play a crucial role in controlling gene expression throughout the body, and the variants that affect them are likely important for health. In this thesis, we go beyond studying individual genetic variants and focus on understanding how combinations of two, three, or more variants work together to influence disease via these protein-DNA interactions.

Detecting these interactions is challenging, especially as the number of genetic variants increases. The complexity of these interactions also surpasses that of the single-variant effects studied in the past. To address these challenges, we employ a cutting-edge mathematical framework known as Targeted Learning, which combines machine learning with causal inference. Targeted Learning is both highly flexible and mathematically rigorous, and it has gained attention from regulatory bodies such as the U.S. Food and Drug Administration. In this thesis, we adapt and apply Targeted Learning to genetics extensively for the first time. We demonstrate its effectiveness using both simulated and real-world data, and we develop new software that can be widely used by the scientific community.

In conclusion, this work integrates advanced mathematical techniques with the biological knowledge accumulated over the years. The methods are rigorously

validated, and the practical outcome is the TarGene software, a state-of-the-art tool that is ready for use in genetic research.

Acknowledgements

Successfully completing a PhD requires more than just a few cups of coffee, it takes tenacity, a dash of passion, and, most importantly, the support of kind and generous people.

I have been fortunate to have an incredible supervisory team guiding me along the way. Thank you, Ava, for bringing me on board and convincing me that Targeted Learning truly brings us closer to uncovering the ground truth. Chris, your patience and encouragement have been invaluable, three years of debugging TarGene, and we finally got to do some biology! And Sjoerd, having your pure mathematical perspective was a vital addition to the journey. You all ensured I got the most out of this experience, and for that, I am deeply grateful.

A big thank you to my office and lab mates for making the journey more enjoyable with casual lunch breaks, after-work spike-ball, tennis, climbing, and pints. The list of names is long, but if you're reading this, I know you'll recognise yourselves.

Lastly, Emily, we boarded this journey as strangers, but soon found ourselves in the same carriage. Thank you for making the trip so comfortable. Your unwavering support rivals only that of my mom's.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Olivier Labayle*)

A ma mère, dont l'énergie inspire mon quotidien.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of genetic variants is crucial for understanding and improving human health outcomes. Some of these variants are in fact so important, that national health services such as the NHS in the United Kingdom, provide genetic testing to detect them. This is the case for some variants in the BRCA genes which greatly increase a woman's chance of developing breast cancer and ovarian cancer [119]. The benefit is that individuals can then make certain lifestyle changes to lower their risk, have regular screenings or opt for preventative treatments. In the vast majority of other cases however, the impact of genetic variations on human traits is much smaller and difficult to interpret. For instance, a recent study reported that at least 12 000 independent variations are significantly associated with height and explain $40-50\%$ of phenotypic variation [181].

Such predictions are traditionally obtained by mapping genetic variations to phenotypes using statistical procedures in large databases. For instance, the UK Biobank contains both genetic and epidemiological data for about 500 000 volunteers [20]. Similarly, the All of Us program in the United States is an effort to gather data from one million participants or more [156]. The development of such databases was largely driven by the realisation that they would enable more discoveries than pedigree-based studies due to increased sample sizes [127]. Together with the standardization of statistical methods, these analyses led to the discovery of hundreds of thousands of genetic variations associated with biomedical traits [28].

However, many reported associations are spurious or overestimated due to biases in data collection and statistical methodology. For instance, socio-economic traits can be confounded by familial relationships within cohorts [66], while ge-

netic ancestry remains a major source of bias in genetic analyses [183]. Moreover, traditional statistical methods rely on strong parametric assumptions, such as linearity and normality of measurement errors. When these assumptions are violated, association estimates may be biased, inflating false positive and false negative rates. This challenge is exacerbated by the increasing sample sizes of modern biobanks, where even subtle biases can become statistically significant. In real-world population genetics datasets, the relationship between genetic variation and traits is often complex and unlikely to be strictly linear. By defaulting to linear models, researchers risk oversimplifying genetic architectures and missing key non-linear interactions. While some recent methods attempt to mitigate these issues by incorporating additional covariates or adopting more flexible modelling strategies [85, 95, 96], they remain constrained by parametric assumptions. In contrast, semi-parametric methods offer a more data-driven approach, providing both flexibility and optimal inference in large samples [76]. This thesis explores the development, implementation, and evaluation of various semi-parametric estimators in the context of population genetics, leveraging their advantages to improve inference accuracy and robustness.

While statistical associations are valuable for screening and disease risk prediction, they are insufficient for understanding disease mechanisms and informing therapeutic development for two key reasons.

First, many of these associations are spurious and will not reveal causal variants, even if the previous biases are correctly mitigated. This is because genetic variations are not independent but occur in approximately independent blocks. This non-random inheritance is known as linkage disequilibrium [143], and complicates the identification of true causal variants from linked ones. Since these blocks can be large, the causal variants and potential mechanisms of action remain elusive. The mathematical framework of causation, pioneered by Judea Pearl [117] and Jamie Robins [128], offers a way to formalise the challenges faced in statistical genetics. It also holds promise for developing new methodologies to address these limitations. In this thesis, we conceptualise genetic effects as causal quantities defined within a causal model of inheritance. Using this model, we demonstrate why current datasets and estimation methods fall short in identifying causal variants. Additionally, we position the traditional adjustment strategy for population stratification, based on Principal Components Analysis, within this causal framework.

Second, even when a variant is causal, the mechanisms by which it influences the trait often remain unclear. This is largely because most association signals occur in non-coding regions of the genome, believed to affect traits through gene regulation [154]. Identifying these regulatory mechanisms is crucial, as they offer direct targets for therapeutic interventions. Though other mechanisms can play a role, gene regulation is often mediated by the binding of transcription factors proteins that interact with nearby DNA elements. Rather than a single protein, transcription is facilitated by complexes of interacting proteins and small molecules called ligands. In this thesis, we propose a general strategy focused on variants known to disrupt transcription factor binding. To overcome the challenges of linkage disequilibrium, we estimate the interactions between these variants and other modulating variants, which are specifically selected based on their impact on the transcription complex. We illustrate this approach with the vitamin D receptor, a specific transcription factor. In this case, since the ligand is vitamin D, a straightforward therapeutic strategy could involve vitamin D supplementation. Furthermore, since the binding disruption is likely to affect nearby genes, knockdowns or knockouts of these genes in model systems can further help validate the mechanism.

Finally, we note that the approach proposed in this thesis aligns with the Targeted Learning framework. Targeted Learning is an integrated approach that aims to generate real-world evidence from real-world data by combining methods from causal inference, machine learning, and statistical theory. This framework has gained considerable attention in the past decade with various applications ranging from HIV testing, COVID-19 vaccine efficacy analysis or child growth faltering in low-resource settings [55, 61, 98]. Furthermore, the Food and Drug Administration is currently evaluating the performance of Targeted Learning in clinical trials through a two-year demonstration project [57].

1.1 Thesis Outline

In chapter 2, we provide foundational material on population genetics, as well as causal and statistical inference. We discuss the challenge of linkage disequilibrium and present state-of-the-art statistical methods. The limitations of these methods are illustrated with a simple simulation, setting the stage for the next chapter.

In chapter 3, we adapt the Causal Roadmap [81] to population genetics, defin-

ing genetic effects as causal quantities and formalising the identification challenges posed by imperfect Mendelian inheritance. We also derive assumption-lean, semi-parametric estimators for estimating these genetic effects.

In Chapter 4, we evaluate the performance of these semi-parametric estimators using simulated data. Two simulation studies are conducted: the first assesses estimation strategies under the null hypothesis of no effect, while the second uses realistic generative models to mimic the original data.

In chapter 5, we introduce the TarGene software, a cutting-edge Nextflow pipeline designed for estimating genetic effects in population cohorts. The software supports traditional analyses like GWAS and PheWAS, as well as more targeted approaches.

In chapter 6, we compare the semi-parametric estimates to those obtained from linear models reported in population genetics databases. We replicate known single-variant effects and interactions, showing consistency with theoretical expectations.

In chapter 7, we apply the TarGene software to investigate the impact of variants that alter the binding of the vitamin D receptor on human traits. We explain why these variants are likely causal and propose potential gene targets through wich they may act.

Finally, in chapter 8, we conclude the thesis by reflecting on its contributions and outlining future research directions inspired by this work.

1.2 Outputs

Three main outputs have been publicly made available.

- A preprint of the method: Dispensing with unnecessary assumptions in population genetics analysis. Olivier Labayle, Kelsey Tetley-Campbell, Mark J. van der Laan, Chris P. Ponting, Sjoerd Viktor Beentjes, Ava Khamseh. bioRxiv 2022.09.12.507656; doi: https://doi.org/10.1101/2022.09.12.507656

- A Julia package for general purpose targeted minimum-loss estimation on tabular datasets, TMLE.jl, and an associated linux executable command-line interface TMLECLI.jl.

- A scalable Nextflow pipeline for the estimation of genetic effects in population cohorts via Targeted Learning, TarGene. Labayle Olivier, Tetley-

Campbell, Kelsey, Slaughter Joshua, Roskams-Hieter Breeshey, Beentjes Sjoerd., Khamseh Ava. and Ponting Chris.

1.3 Statement of Contributions

In this thesis, I use the pronoun "we" as a general recognition that the work presented in this thesis was inspired and mentored by the variety of people acknowledged further above. Biostatistics is a multidisciplinary discipline and this journey took many discussions. In this section I make precise the various significant contributions others have made to some of the work presented here.

- In chapter 3, the initial idea to apply targeted minimum loss-based estimation to the field of population genetics originates from Sjoerd Beentjes and Ava Khamseh. The work presents for the first time, an integrated road-map for the semi-parametric estimation of genetic effects, rooted in modern causal inference.

- The ideas and development of the simulations presented in chapter 4 are my own. They showcase the practical importance of the positivity condition in population genetics and lead to practical recommendations to obtain nominal coverage of the ground truth when using these estimators in genetics.

- In chapter 5, I present the TarGene software and its foundational components, reflecting on my three-year journey and highlighting key strategic insights gained along the way. It is the first software enabling the semi-parametric estimation of genetic effects at scale. It has now been used by at least 6 PhD students and postdoctoral researchers across the lab. Joshua Slaughter significantly contributed to the GWAS study design, that I supervised. Breeshey Roskams-Hieter enabled the use of traditional CSV traits' datasets.

- In chapter 6, the PCA plots were made by Kelsey Tetley-Campbell (PhD student). This chapter presents the first application of semi-parametric estimators to real-world population genetics data.

- Finally, the variant interaction-based framework, and the idea of integrating population genetics with functional genomics, presented in chapter 7 was

originated by Chris Ponting. The presentation however, corresponds to my own interpretation.

Unless otherwise stated, the rest of the elements in this thesis, in particular all implementations and UK Biobank analyses, are my own.

1.4 Notations

The following notations will be recurrent.

Mathematical Notations

- Y: any binary, ordinal or continuous trait or disease.
- $\mathbf{V}$: a set of genetic variants.
- $\mathbf{W}$: the set of variables confounding the effect of $\mathbf{V}$, e.g., population stratification.
- $\mathbf{C}$: a set of variables predictive of the outcome Y but not confounding the effect of $\mathbf{V}$.
- $\mathcal{M}$: A statistical model, that is, a set of probability distributions.
- P: A probability distribution in $\mathcal{M}$.
- P_0: The true probability distribution that generated the observed data, ideally in $\mathcal{M}$.
- $\Psi : \mathcal{M} \to \mathbb{R}^d$: a statistical functional, that is, a function that maps a distribution to the real numbers. This is how we formally define genetic effects.
- $\Psi_0 = \Psi(P_0)$: the true target parameter of interest, here, the value of the genetic effect we seek to estimate.
- $Q_Y : (\mathbf{V}, \mathbf{W}, \mathbf{C}) \mapsto \mathbb{E}[Y|\mathbf{V}, \mathbf{W}, \mathbf{C}]$: the outcome regression model.
- $P_{W,C} : (\mathbf{W}, \mathbf{C}) \mapsto P(\mathbf{W}, \mathbf{C})$: the joint marginal probability distribution of confounders and extra covariates.
- $G : (\mathbf{V}, \mathbf{W}) \mapsto P(\mathbf{V}|\mathbf{W})$: the conditional distribution of the variants given the confounders. It is known as the propensity score or treatment mechanism.

- Q: an arbitrary set of functions necessary to estimate Ψ. Very often, the estimation of Ψ does not require the whole P but only components of Ψ, called nuisance functions. In this thesis $Q = (Q_Y, P_{W,C}, G)$.
- The following notations for expected values are equivalent and will be used interchangeably: $\mathbb{E}_P[f] = \int f dP = Pf$.
- $L_0^2(P)$: the space of square integrable and measurable functions with respect to the measure P whose mean is 0.

Acronyms

- SNP: single nucleotide polymorphism, a genetic variant at a single base position in the DNA.
- GWAS: genome-wide association study, a single trait is tested for association across the genome.
- PheWAS: phenome-wide association study, a single variant is tested for association across many traits.
- PheWIS: phenome-wide interaction study, a single set of variants' interaction (e.g., a pair) is investigated across many traits.
- LD: linkage disequilibrium, describes the non-random association between alleles of genetic loci in proximity to each other.
- MAF: minor allele frequency (for a genetic variant).
- mRNA: messenger RNA, is a type of single-stranded RNA involved in protein synthesis.
- QTL: quantitative trait locus, a genetic region influencing phenotypic variation. QTLs are often prefixed by a letter indicating the type of phenotype they are associated with. For example, an eQTL is a QTL associated with differential gene expression (mRNA).
- ATE: average treatment effect.
- AIE: average interaction effect.
- GLM: generalised linear model.

- LMM: linear-mixed model.
- PCA: principal components analysis.
- TMLE: targeted minimum loss-based estimation (sometimes targeted maximum-likelihood estimation though this is not as general).
- OSE: one-step estimation.
- FWER: family-wise error rate.
- FDR: false discovery rate.
- GRM: genetic relationship matrix.
- VDR: vitamin D receptor.

Chapter 2

Background

2.1 From Mendelian Inheritance to Modern Population Genetics

2.1.1 The Experimental Evidence of Genetic Inheritance

The first evidence of genetic influence on traits can be traced back to Gregor Mendel's pioneering work with pea plants in 1865. Through his meticulous experiments, Mendel proposed the existence of a heritable substance, which he referred to as "elementen". Today, we know this substance as Deoxyribonucleic Acid (DNA), which is composed of functional units called genes. Each gene can exist in different forms, known as alleles, that determine the specific manifestation of a trait.

Mendel's model of heredity is encapsulated in three fundamental laws of inheritance:

1. **Law of Segregation** Each individual has two alleles for each trait, one from each parent. During the formation of gametes (sperm and egg cells), these two alleles separate (segregate) from each other so that each gamete carries only one allele for each gene. During fertilisation, the gametes from each parent fuse to form a new organism, restoring the paired condition.

2. **Law of Independence** The alleles of different genes assort independently of each other during gamete formation. This means the inheritance of one trait generally does not affect the inheritance of another.

3. **Law of Dominance** Some alleles are dominant, while others are recessive. An organism will exhibit the effect of the dominant allele if it possesses at least one copy of it.

Mendel's conclusions find strong support in our modern understanding of genetics, particularly in the process of meiosis, which is responsible for the formation of gametes. As illustrated in figure 2.1, meiosis results in the almost independent segregation of genetic material into daughter cells, a key mechanism that aligns with Mendel's laws of inheritance.

One of the primary critiques of Mendel's model was its implication that traits should be discrete, a notion seemingly at odds with the continuous variation observed in human traits. This apparent contradiction was later reconciled by Ronald Fisher, who demonstrated that the involvement of multiple genes in the expression of a single trait could account for the observed diversity [46]. Fisher's work laid the groundwork for quantitative genetics, showing that complex traits could indeed arise from the additive effects of many genes, each contributing a small amount to the overall phenotype.

Additionally, William Bateson's work on epistasis (genetic interactions), particularly his study on the colour of pea flowers, further refined Mendel's model. Bateson showed that the expression of a trait could be influenced by interactions between genes, where one gene could mask or modify the effect of another [9]. These insights paved the way for a more nuanced understanding of genetic complexity, including higher-order interactions between genes and the influence of environmental factors.

Together, these developments have enriched Mendel's original model, allowing for the incorporation of arbitrary complexity in genetic inheritance and providing a more accurate reflection of the diversity observed in nature. Recently, it has been proposed that gene regulatory networks are so highly interconnected that even genes outside of core pathways may contribute to disease, a hypothesis known as the omnigenic model [16].

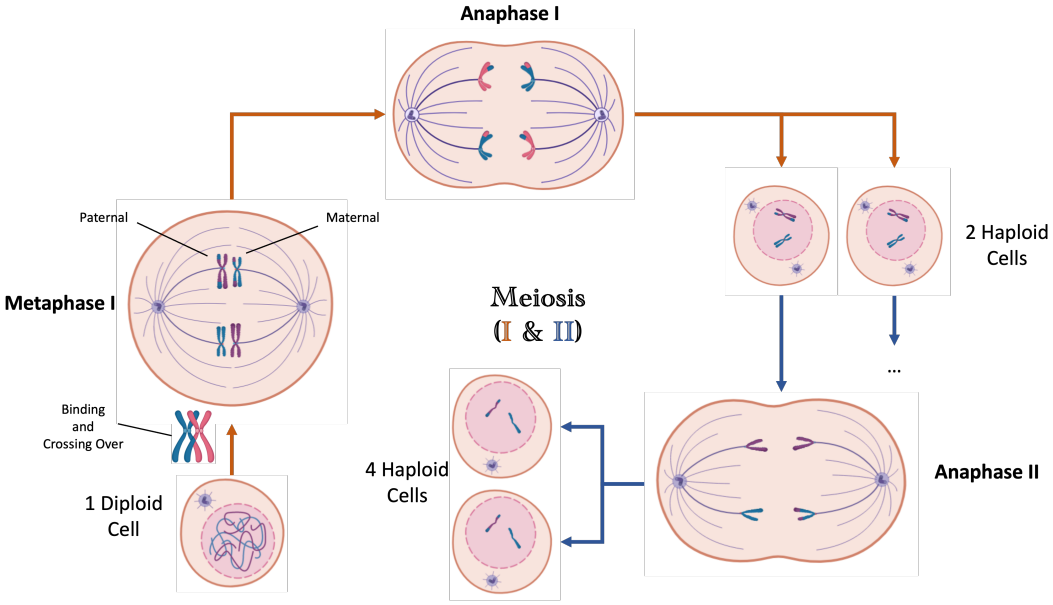

Figure 2.1: **Simplified Representation of Meiosis**. Meiosis, is the process by which two successive divisions of a diploid cell give rise to four haploid cells. It provides a molecular understanding of the laws of segregation and independence. During Meiosis I, after DNA replication, homologous (paternal and maternal) chromosomes pair up and bind together. Segments of homologous chromosomes are randomly exchanged, a process called crossing over. Crossing over is one of the main sources of genetic diversity in offspring, it explains the almost independent inheritance of genes on a given chromosome. During Metaphase I and Anaphase I, homologous chromosomes line up at the equator of the cell and are segregated to opposite sides of the cell. The end of the first Meiosis cycle results in two haploid cells containing random combinations of crossed-over parental chromosomes. At this point, each chromosome consists of two sister chromatids. These will be similarly separated in the second Meiosis cycle, which does not comprise a DNA replication stage. The entire process results in 4 haploid cells or gametes.

2.1.2 Measuring the Genome

From the discovery of the structure of DNA in 1953 [169], to the Human Genome Project initiated in 1990 [111], we finally obtained a complete sequence of a human genome in 2022 [109]. These scientific and technological advancements have enabled the transition from a pedigree-based understanding of genetics to a fine-grained molecular perspective. For instance, we now know the human genome has a total length of approximately 3.2 billion base pairs separated across 46 chromo-

somes, and contains between 20 000 to 25 000 genes. As of 2017, there were 324 million known variants from sequenced human genomes [140], but a typical human only differs from the reference at around 20 million base pairs [30]. These genetic variations are usually classified into two categories, single-nucleotide polymorphisms (SNPs) and structural variations. Single-nucleotide polymorphisms, by far the most frequent type of variations ($> 99\%$), are defined by the substitution of a single nucleotide at a specific position in the genome, e.g. the substitution of an adenine base with a thymine base $A \rightarrow T$. The substitution can occur within coding regions, in which case it can have dramatic impacts like Progeria. But in most cases, substitutions are found outside coding regions and are of unknown effect. Changes to the chromosomal structure like deletions, duplications, or insertions are known as structural variants. While less frequent, they can account for more variability in a genome due to the size of the impacted regions.

Central to these findings are the methodologies employed to measure genetic variation across populations and individuals: genotyping and sequencing. Each of these techniques offers unique strengths and applications, contributing to the broader field of genomics and personalised medicine.

Sequencing involves determining the precise order and type of nucleotides within a DNA molecule. Sequencing provides comprehensive information about an individual's entire genome or specific regions of interest. Next-Generation Sequencing (NGS) advances, such as Illumina's sequencing by synthesis and Oxford Nanopore's long-read sequencing, have dramatically increased sequencing throughput while reducing costs [99]. While whole-genome sequencing data will become increasingly available in the future, as illustrated by the recent release for all 500 000 UK Biobank participants, large scale population studies are still dominated by genotyping datasets.

Genotyping refers to the process of determining differences in the genetic makeup (genotype) of an individual by examining their DNA sequence at specific loci. This method primarily focuses on identifying single nucleotide polymorphisms (SNPs), which are the most common type of genetic variation. One of the most prevalent genotyping methods utilises micro-array technology, where DNA samples are hybridised to a micro-array chip containing probes for specific SNPs [58]. This technology allows the cost-effective simultaneous examination of hundreds of thousands of SNPs across the genome. Genotyping inevitably incurs a loss of information and some genetic variations will be unobserved. At

the same time, it was observed that nearby genetic variations were highly correlated, a phenomenon known as linkage disequilibrium [143]. By carefully choosing genotyped SNPs across the genome, and using comprehensive reference panels, imputation methods can enhance the resolution of genotyping assays [34]. For the UK Biobank, the Haplotype Reference Consortium, the UK10K and 1000 Genomes datasets were used as reference panels [27, 30, 86].

In this thesis, we do not consider whole-genome sequencing data which was only released for all participants in the UK Biobank in 2024. Instead, the genetic data comprises both genotyped and imputed genetic variations using the GRCh37 genome assembly. As will be explained later, stringent quality thresholds for imputation will be enforced.

2.1.3 From Causality to Association Studies

Since genetic experimentation in humans is both unpractical and unethical, the effect of genetic variations on human traits is investigated via statistical methods. Perhaps the most successful study design to date is the genome-wide association study (GWAS), whereby millions of variants are statistically tested for association with a given trait. This success has been notably driven by the ever growing sample size of modern studies, resulting in greater statistical power [161], that is greater probability to correctly reject the null hypothesis. This is first because the effect of genetic variations is typically small, hence requiring high precision to be detected. But foremost, due to the sheer number of hypotheses tested. Controlling the false discovery rate of hundreds of thousands of statistical tests requires extreme statistical confidence. In genome-wide association studies, it is typical to control the family-wise error rate, the probability of making one or more false discoveries across all tests. This strict criterion requires p-value $< 10^{-8}$ for genome-wide significance using the Bonferroni method [4]. The emergence of large-scale bio-banks, such as the UK Biobank [20], combining both genetic and phenotypic data for 500 000 individuals, have rendered the GWAS study-design scalable to many human traits. For example, the geneATLAS study published genetic associations for 118 non-binary and 660 binary traits of 452 264 participants of European ancestry [21].

One unequivocal conclusion from GWASs is that almost any complex trait that has been studied is associated with many different genetic variations, far

from the Mendelian perspective [168]. For instance, Type 2 Diabetes has been associated with more than 100 genetic variants [51]. Furthermore, these variants are often associated with multiple seemingly independent traits, a phenomenon known as pleiotropy. This highlights the complexity of human genetics and the need for a holistic approach since interventions could result in unintended effects.

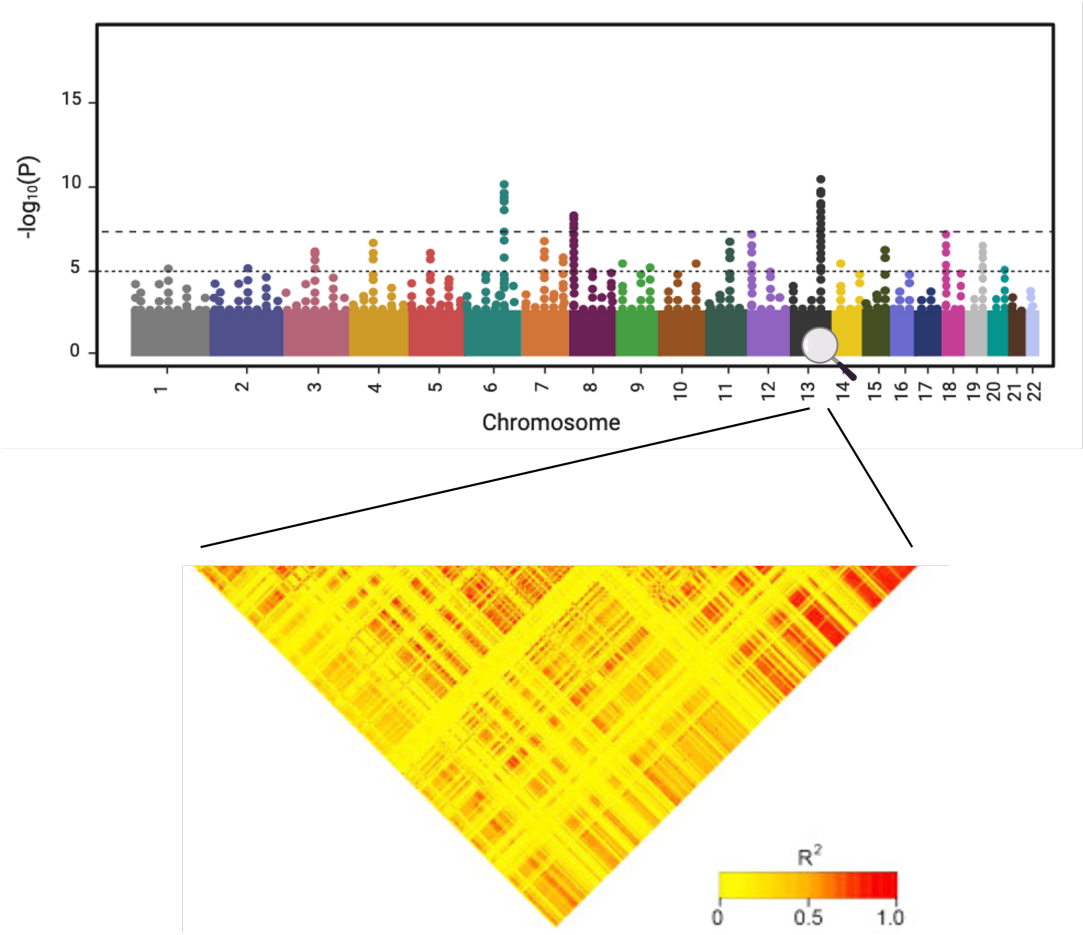

Figure 2.2: **A Manhattan Plot and Linkage-Disequilibrium Heatmap**. The output of GWAS, a list of p-values for each tested locus in the genome, is traditionally presented as a Manhattan plot (Top). P-values are organised by chromosome and position on the horizontal axis to ease visualisation. Due to linkage disequilibrium patterns, blocks of variants are identified and the causal variants are concealed. Strong association peaks can be observed on chromosome 6, 8 and 13. This dependence structure among variants can be further investigated via heatmaps (e.g. for chromosome 13 at the bottom). The dependence is reported using the square of the correlation coefficient between two variants (R^2).

The main limitation of the statistical paradigm is that it does neither reveal

causal variants, nor the biological mechanisms through which they act on human traits. Therapeutic intervention strategies are thus difficult to derive from GWASs alone. Notably, detected variants need not be causal due the crossing over of homologous chromosomes during Meiosis (see figure 2.1). Instead of being inherited completely independently, genetic variations are inherited in approximately independent blocks. The high degree of dependence within blocks, known as linkage disequilibrium (LD), renders causal variants indistinguishable from linked variants [143]. Furthermore, most detected variants lie in non-coding regions, and their mechanism of action is difficult to unravel. Since they do not alter protein sequences, it is believed that these variants impact traits through gene regulation (figure 2.3). Variant interaction studies, could help pinpoint the mechanism of actions but have been largely unsuccessful [161]. This is surprising since prior evidence has been shown in other organisms over the past 100 years, and interactions are ubiquitous at the biomolecular level. It was furthermore suggested, and validated in yeast experiments [166], that epistasis evolved as a compensation to point mutations, enhancing and stabilising health outcomes. The reason for this lack of success is mainly thought to be due to statistical limitations. First, the defects of commonly employed parametric methods have been pointed out [102]. Second, for a set of candidate variants, the multiple-hypotheses correction burden grows exponentially with the order of the interaction under consideration.

The success of statistical genetics has been tempered by the challenge of contextualising detected associations and their impact on human health outcomes. To bridge this gap, post-GWAS methodologies leveraging additional biomedical knowledge and advanced statistical methods have been developed [154].

2.1.4 The Search for Causal Variants and Mechanisms

We have seen that, due to LD patterns, GWASs identify regions of association rather than causal variants (figure 2.2). Statistical fine mapping techniques attempt to identify sets of likely causal variants by combining GWAS results and LD patterns. At first sight, the lead variant, with the most significant association, may be expected to be causal. Unfortunately, this is not necessarily the case, for instance if there are multiple causal variants in the associated region [134]. It has also been shown that fine-mapped variants across ethnicities do not always overlap, and that trans-ethnic fine-mapping can improve the resolution. Modern

methodologies, rooted in Bayesian statistics, attempt to combine various sources of information into informative priors and produce credible sets of causal variants.

The majority of variants in credible sets are located outside coding regions, and have unknown regulatory functions. The biological mechanisms driving changes in cellular and physiological functions thus remain unknown. Additionally, the specific tissue, cell type, and cell state relevant to these effects are also unidentified, further complicating the understanding of these mechanisms and development of novel therapeutics. Perhaps the most important part of the functional characterisation of a credible variant is the identification of the impacted genes. This investigation relies on combined information from biomolecular experiments, sometimes tissue specific and further association testing. If a variant is associated with a molecular trait, it is said to be a molecular quantitative trait locus (molQTL). Depending on the experiment, the variant can sometimes be causally implicated in the biological pathway, hence strengthening its potential implication in the trait investigated by the GWAS. We describe here some of these molQTLs, and how they can be used to contextualise fine-mapped variants.

An expression Quantitative Trait Locus (eQTL) is a genetic locus associated with the variation in gene expression levels, and more precisely messenger RNA (mRNA). eQTLs are identified via RNA sequencing (RNA-Seq) technologies [59] and association testing, and will also suffer from linkage disequilibrium. By linking GWAS hits to eQTLs, researchers can understand how these variants might contribute to disease through gene expression changes. For instance, in figure 2.3, a variation in a potentially distant enhancer can affect the gene's transcription. Fine grained resolution can further be obtained using specific tissues, cell types or even single cells. The GTEx project (Genotype-Tissue Expression [87]) is a notable initiative that maps eQTLs across a wide range of human tissues.

A splicing Quantitative Trait Locus (sQTL) is a genomic locus that explains variations in mRNA splicing patterns. Alternative splicing is the process by which the same gene can produce different mRNA transcripts and contribute to protein diversity and function. This diversity can then modulate cellular function and disease susceptibility. Methodologies and resources for the identification of sQTLs are similar to those of eQTLs.

A protein Quantitative Trait Locus (pQTL) is similar to an eQTL, but is associated with variation of protein levels. Since proteins are proximal to phenotypic traits, they might be of more direct relevance to human health. High-throughput techniques like mass spectrometry are used to quantify protein levels in different individuals, enabling the identification of pQTLs. However, mass spectrometry involves more complex sample preparation, data acquisition, and analysis workflows compared to RNA-Seq, which is why pQTLs are not as widespread as eQTLs.

A binding Quantitative Trait Locus (bQTL) is a specific region of the genome that is associated with variations in the binding affinity or binding patterns of regulatory proteins, such as transcription factors, to DNA. Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is often used to identify bQTLs [116]. A more recent method, ChIP-exo, offers a theoretical near base pair resolution [126]. Binding QTLs identified via ChIP-exo are thus likely causal of binding disruption. The ChIP-Atlas is a data mining and visualisation platform integrating thousands of ChIP-seq experiments, hence providing further insights into transcription regulation [110]. In chapter 7, we present an analysis framework based on bQTLs.

A methylation Quantitative Trait Locus (meQTL) is a specific genetic region of the genome that is associated with DNA methylation levels. DNA methylation is an epigenetic modification where methyl groups are added to cytosine bases in the DNA sequence, also contributing to the regulation of gene expression [103].

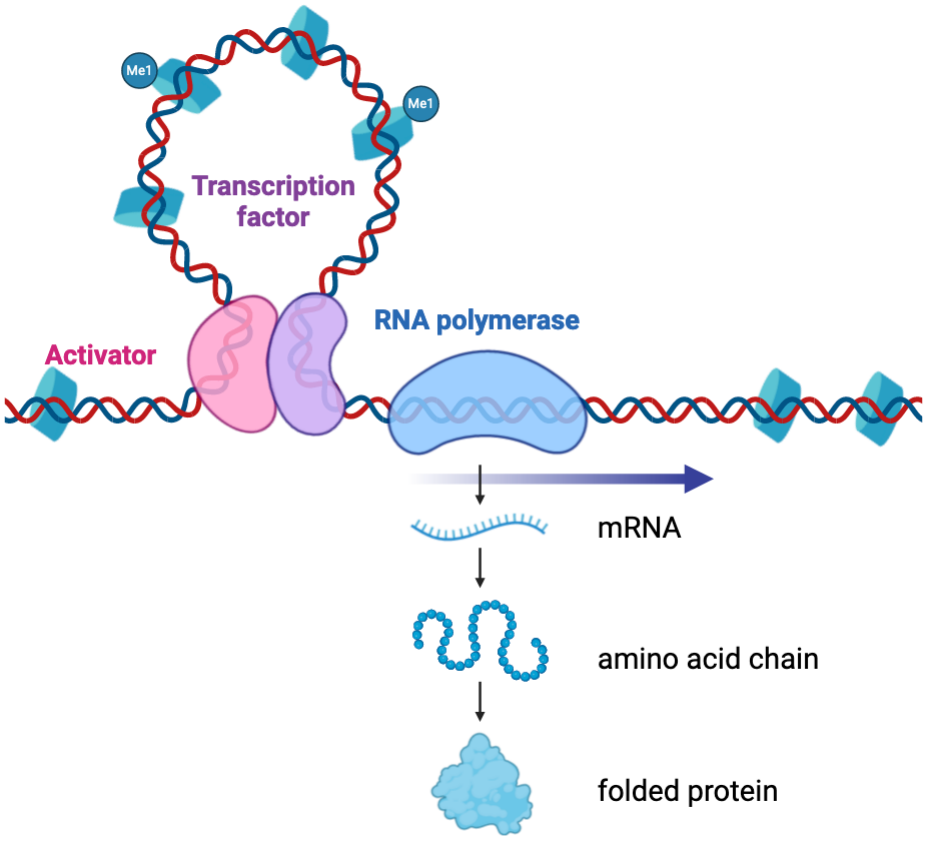

Figure 2.3: **An overview of Gene Regulation**. DNA is transcribed into mRNA by the RNA polymerase enzyme. In humans, the initial mRNA transcript (pre-mRNA) undergoes modifications such as splicing (not shown). The processed mRNA is then translated into a protein by ribosomes in the cytoplasm and undergoes folding and other modifications (e.g., phosphorylation) to become fully functional. Many genetic and non-genetic factors can influence this process. A genetic variation within a regulatory region can impact the transcription of the downstream gene. Promoters and enhancers are of particular importance in this context. A promoter is a DNA sequence located near the start of a gene that serves as a binding site for RNA polymerase and other transcription factors. Enhancers are DNA regions, often located far from the gene, that can facilitate gene transcription. In this example, an enhancer and the promoter are bound together by two transcription factors. This mechanism is made possible by DNA looping, which is controlled by epigenetic changes (e.g., histone methylation) that affect chromatin structure and DNA accessibility.

Given the diverse and intertwined nature of gene regulation, not only a single modality, but combined information will help unravel causal variants, their mechanisms and context of actions. These multiple modalities will also be leveraged

in more scalable and principled ways. For instance, probabilistic colocalisation techniques such as eCAVIAR, combine eQTLs and GWAS summary statistics to identify target genes [64].

2.1.5 Mendelian Randomisation

Mendelian randomisation (MR) is one set of methods that can help uncover causal mechanisms. While distinct from the focus of this work, MR provides a complementary approach by using genetic variation to explore whether specific exposures influence health, developmental, or social outcomes [41]. In MR, exposures are any factors reliably linked to genetic variation, such as measurable traits like body mass index (BMI) or gene expression levels in particular tissues. One of the key advantages of MR is that it can point to modifiable exposures, potentially highlighting targets for therapeutic intervention. However, MR does not directly identify causal variants. Instead, genetic variants serve as instruments to assess the effect of the exposure. Although MR is grounded in the instrumental variables method, a well-established theoretical framework, it often faces practical challenges in meeting its underlying assumptions. While the principles supporting MR are covered by the mathematical framework presented next, they will not be discussed further since they fall outside the primary scope of this thesis.

2.2 Causality

What does it mean for a genetic variant to be causal of a disease, and why is it important for human health? In the context of genetics, causality refers to a relationship where a genetic variant directly influences the likelihood of developing a disease. This concept is often grounded in counterfactual reasoning, a philosophical idea with origins tracing back to David Hume [68]. Counterfactual theories assert that causal relationships can be understood by considering what would happen in alternate scenarios. Specifically, would a person have developed a disease if they had a different genetic makeup?

This viewpoint is particularly useful since it implies the potential for a targeted intervention. If a genetic variant is truly causal, then altering that variant or its downstream effects could prevent or treat the disease. However, since we cannot observe these alternate realities directly, causal effects cannot be di-

rectly measured, a challenge known as the fundamental problem of causal inference [151].

Nevertheless, causal inference is not always beyond reach. With the right techniques and assumptions, one can make informed inferences about causality using observational data and in the absence of perturbation experiments. Modern mathematical frameworks, such as the Potential Outcomes framework [130], or Structural Causal Models [117], provide the foundation for these techniques. These models formalise the assumptions and conditions necessary to infer causality from both experimental and observational data.

Perhaps the most robust method for establishing causality is through experimentation. In health care and more generally in science, randomised control trials (RCTs) are considered the gold standard for establishing causality. This is because RCTs randomly assign participants to different treatments, effectively simulating counterfactual scenarios [36]. However, randomised controlled trials (RCTs) are not always feasible or sufficient for answering certain research questions. In human genetics, RCTs are often impractical and raise ethical concerns. Additionally, the population studied in an RCT may not be representative of the broader population, limiting the generalisability of the results [108]. Fortunately, the same causal frameworks that justify RCTs also allow us to draw causal inferences from observational data, which is becoming increasingly abundant in genetics research.

In this section, we focus on the Structural Causal Model framework developed by Judea Pearl. We explore the assumptions and methodologies required to apply the framework to the genetic data analysed in this thesis. The specific challenges and nuances related to genetic inheritance will be discussed in detail in Section 3.2.

2.2.1 Structural Causal Models

A structural causal model (SCM), is an idealised representation of the world, whereby the model explicitly encodes the mechanisms by which the observed state of word is produced. The original definition of a Structural Causal Model is as follow

Definition 2.2.1 (Causal Model)**.** *A causal model is a triple* $\mathbf{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$ *where:*

1. $\mathbf{U}$ *is a set of background variables, (also called exogenous), that are determined by factors outside the model.*

2. $\mathbf{V}$ *is a set* $\{V_1, V_2, ..., V_n\}$ *of variables, called endogenous, that are determined by variables in the model – that is, variables in* $\mathbf{U} \cup \mathbf{V}$ *and denoted by* $\mathbf{PA_i}$.

3. $\mathbf{F}$ *is a set of functions* $\{f_1, f_2, ..., f_n\}$ *such that each* f_i *is a mapping from the respective domains of* $U_i \cup \mathbf{PA_i}$ *to* V_i, *where* $U_i \subset \mathbf{U}$, $\mathbf{PA_i} \subset \mathbf{V} \setminus V_i$, *and the entire set* $\mathbf{F}$ *forms a mapping from* $\mathbf{U}$ *to* $\mathbf{V}$.

A causal model defines a directed acyclic graph over the sets $\{\mathbf{U}, \mathbf{V}\}$, where $\mathbf{F}$ represents the mechanisms that determine each variable in the system based on its parent variables. As an example, consider the structural causal model represented in figure 2.4 (left), and defined by $\mathbf{V} = \{W, T, C, Y\}$, $\mathbf{U} = \{U_W, U_T, U_C, U_Y\}$ and $\mathbf{F} = \{f_T, f_W, f_C, f_Y\}$, such that

$$
\begin{aligned}
W &= f_W(U_W) \\
C &= f_C(U_C) \\
T &= f_T(W, U_T) \\
Y &= f_Y(T, W, C, U_Y)
\end{aligned}
$$

In this model, one can view the variable T as a treatment variable whose effect on Y is the quantity of interest. In the context of genetics, T is understood as a genetic variant (and will later be denoted by V) and Y the disease status of an individual. The confounding variable W, such as population stratification, influences both the individual's genetics and their propensity to develop the disease. In contrast, the variable C only affects the disease status. Precise definitions for W and C are given in section 3.2.

The SCM approach is appealing because it is intuitive, similarly to the experimental approach, it defines causal effects in terms of interventions. More precisely, an intervention is represented by a causal submodel where some of the mechanisms have been changed. In this Thesis, the only type of submodel we will need, consists in ideal interventions where some mechanisms are replaced with constant functions.

Definition 2.2.2 (Causal Submodel)**.** *Let* $\mathbf{M}$ *be a causal model,* $\mathbf{T}$ *be a set of variables in* $\mathbf{V}$, *and* $\mathbf{t}$ *a particular realisation of* $\boldsymbol{T}$. *A submodel* $\mathbf{M_t}$ *of* $\mathbf{M}$ *is a causal model* $\mathbf{M_t} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F_t} \rangle$, *where*

$$\mathbf{F_t} = \{f_i : V_i \notin \mathbf{T}\} \cup \{\mathbf{T} = \mathbf{t}\}$$

Continuing with our previous example, the following $\mathbf{F}_t$ defines a causal submodel, also illustrated in figure 2.4 (right).

$$\begin{aligned} W &= f_W(U_W) \\ C &= f_C(U_C) \\ T &= t \\ Y &= f_Y(T, W, C, U_Y) \end{aligned}$$

From the genetic point of view, such an intervention can be thought as the hypothetical introduction of a germline mutation. Causal effects are thus defined by a given intervention and associated causal submodel. For two subsets $\mathbf{T}$ and $\mathbf{Y}$ in $\mathbf{V}$, the causal effect of $\mathbf{T}$ on $\mathbf{Y}$ is the solution for $\mathbf{Y}$ of the set of equation $\mathbf{F_t}$, that is $\mathbf{Y_t}(\mathbf{u}) = \mathbf{Y_{M_t}}(\mathbf{u})$. $\mathbf{Y_t}(\mathbf{u})$, is often referred to as the potential response to action $do(\mathbf{T} = \mathbf{t})$. This definition of causality is entirely deterministic, if the functions in $\mathbf{F_t}$ are known, causal effects can be computed for any $\mathbf{u} \in \mathbf{U}$.

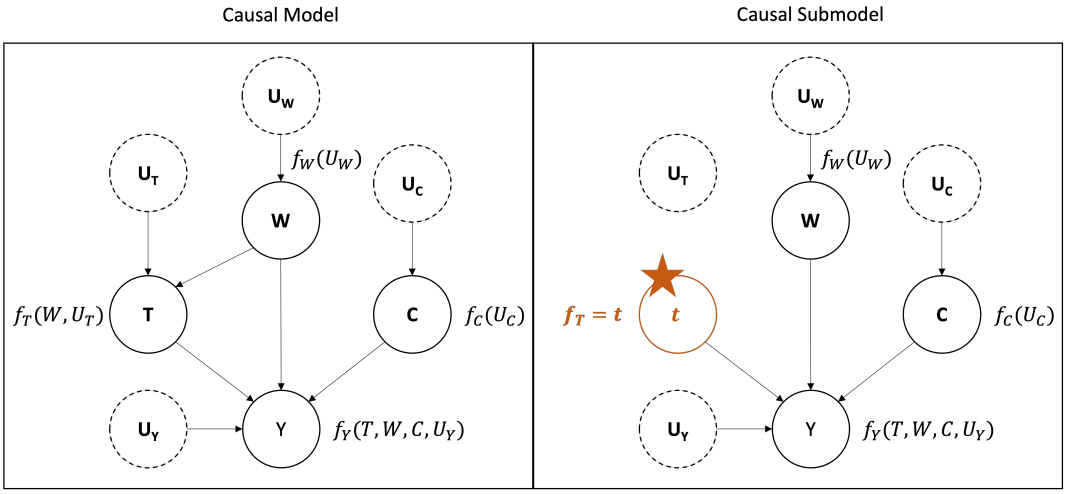

Figure 2.4: **A Causal Model and Submodel side by side.** In this model $\mathbf{V} = \{W, T, C, Y\}$ are endogenous variables and $\mathbf{U} = \{U_W, U_T, U_C, U_Y\}$ are exogenous variables. The causal model (left) represents our understanding of the natural world. The causal submodel (right) represents the effect of a hypothetical intervention on T. Importantly, since the intervention is perfect, all other independent mechanisms $\{f_W, f_C, f_Y\}$ remain invariant.

To reflect our imperfect knowledge of the physical world, we will equip the causal model with a probability distribution. A probabilistic causal model is a pair $\langle \mathbf{M}, P(\mathbf{u}) \rangle$ where $\mathbf{M}$ is a causal model and P is a probability distribution over the exogenous variables $\mathbf{U}$. Importantly, while the functions $\{f_1, f_2, ..., f_n\}$ are deterministic mechanisms, $P(\mathbf{U})$ induces a probability distribution over the endogenous variables $P(\mathbf{V})$. Then, causal effects can also defined from a probabilistic perspective as follows

$$P(\mathbf{Y_t} = \mathbf{y}) = P(\mathbf{Y}|do(\mathbf{t})) = P(\{\mathbf{u} : \mathbf{Y_t}(\mathbf{u}) = \mathbf{y}\}) \tag{2.1}$$

The event $\{\mathbf{u} : \mathbf{Y_t}(\mathbf{u}) = \mathbf{y}\}$ is a counterfactual event, it is understood as "The set of situations $\mathbf{u} \in \mathbf{U}$ for which $\mathbf{Y}$ would be $\mathbf{y}$, had $\mathbf{T}$ been $\mathbf{t}$". In other words, equation 2.1 represents the probability of the outcome $\mathbf{Y}$ had the intervention $\mathbf{T} = \mathbf{t}$ been performed.

This definition extends naturally to more complex effects, such as counterfactuals that are conditional on actual observations. For instance, $P(\mathbf{Y_t} = \mathbf{y}|\mathbf{t}') = P(\mathbf{Y}|do(\mathbf{t}), \mathbf{t}') = P(\{\mathbf{u} : \mathbf{Y_t}(\mathbf{u}) = \mathbf{y}, \mathbf{T}(\mathbf{u}) = \mathbf{t}'\})$ represents the probability that $\mathbf{Y} = \mathbf{y}$ after intervention $\mathbf{T} = \mathbf{t}$ but given the observation $\mathbf{T} = \mathbf{t}'$. At first sight, if $\mathbf{t} \neq \mathbf{t}'$, this statement may seem meaningless. This is not the case since $\mathbf{T}(\mathbf{u})$ and $\mathbf{Y_t}(\mathbf{u})$ are evaluated in two distinct submodels and correspond to an ordinary event in $\mathbf{U}$-space. The analysis of this type of counterfactuals is beyond the scope of this thesis and we will focus exclusively on causal effects derived from equation 2.1.

2.2.2 Identifiability

Apart from randomised controlled trials, causal submodels are typically hypothetical and do not match the observed data. For instance, in population genetics, the genome of embryos is not edited to understand the effect of such modification on disease risk. Fortunately, the mathematical approach to causation provides an analytical framework to understand when interventional questions can be answered from observational data alone. The ability to uniquely determine the causal effect of an intervention from the joint distribution of observed variables is called identifiability.

Definition 2.2.3 (Identifiability)**.** *The causal effect $P(\mathbf{Y_t})$ is identifiable if there exists a function h such that $P(\mathbf{Y_t}) = h(P(\mathbf{V}, \mathbf{U}))$.*

Do-calculus, developed by Judea Pearl, provides a set of rules and procedures to systematically identify causal effects from observational data when certain conditions are met. One of the most important results obtained from these rules, and the one we will use in this work, is the Backdoor Adjustment Theorem. The Backdoor Adjustment Theorem is a graphical criterion for identifying a set of variables that, when conditioned upon, can control for confounding.

Theorem 2.2.1 (Backdoor Adjustment Theorem)**.** *A set of variables* $\mathbf{Z}$ *satisfies the backdoor criterion relative to* $(\mathbf{T},\mathbf{Y})$ *if:*

1. *No node in* $\mathbf{Z}$ *is a descendant of* $\mathbf{T}$
2. $\mathbf{Z}$ *blocks every path between* $\mathbf{T}$ *and* $\mathbf{Y}$ *that contains an arrow into* $\mathbf{T}$

If a set of variables $\mathbf{Z}$ *satisfies the backdoor criterion relative to* $(\mathbf{T},\mathbf{Y})$*, then the causal effect of* $\mathbf{T}$ *on* $\mathbf{Y}$ *is identifiable and is given by the formula*

$$p(\mathbf{Y}|do(\mathbf{T})) = \int p(\mathbf{Y}|\mathbf{T},\mathbf{z}) \cdot p(\mathbf{z}) \cdot \mathbf{dz} \tag{2.2}$$

For instance, in the example of figure 2.4, and relative to (T,Y)

- $\{Y\}$ does not satisfy the backdoor criterion since it is a descendent of T.
- $\{C\}$ does not satisfy the backdoor criterion since it does not block the path $T \leftarrow W \rightarrow Y$ containing an arrow into T.
- The sets $\{W\}$ or $\{W,C\}$ satisfy the backdoor criterion since they blocks the previous path.

And hence the causal effect $P(Y_t)$ is given by any of the following

$$\begin{aligned} P(Y_t = y) &= \int p(Y|T,w) \cdot p(w) \cdot dw \\ &= \int p(Y|T,w,c) \cdot p(w,c) \cdot dw \cdot dc \end{aligned} \tag{2.3}$$

While these two formulations are equivalent, the inclusion of more relevant predictive variables can improve statistical significance [53], and the second equation is preferred.

Identification has thus effectively transformed a causal quantity corresponding to an unobserved intervention into a quantity involving only observed data from the real world. However, there are two implicit assumptions that need to be

satisfied for backdoor adjustment. First, since the adjustment set Z must block every backdoor path, there should be no unobserved confounders, an assumption known as unconfoundedness. This requirement is essentially unverifiable, engagement with domain experts is thus of particular importance to ascertain the validity of causal inferences. The implications of unconfoundedness in population genetics are discussed further in section 3.2.2. Additionally, some methods, broadly categorized under sensitivity analysis [84], aim to quantify the potential impact of unmeasured confounding. Second, the conditional distribution involved in the identified causal effect is by definition $p(\mathbf{y}|\mathbf{t},\mathbf{Z}) = \frac{P(\mathbf{y},\mathbf{t},\mathbf{Z})}{P(\mathbf{t}|\mathbf{Z})\cdot P(\mathbf{Z})}$. This means that, for each value of the confounding variables $\mathbf{Z}$, the treatment variable $\mathbf{T}=\mathbf{t}$ should have non-zero probability, otherwise the adjustment formula is not defined. This condition is known as Positivity and is formally defined by

Definition 2.2.4 (Positivity)**.** *Let* $\mathbf{T}=\mathbf{t}$ *be a treatment value of interest, and* $\mathbf{Z}$ *be a valid adjustment set. The positivity assumption states that*

$$\forall \mathbf{z} \in \mathbf{Z}, 0 < P(\mathbf{T}=\mathbf{t}|\mathbf{Z}=\mathbf{z}) < 1 \tag{2.4}$$

Note that the condition of definition 2.2.4 is only required for the treatment value of interest. For instance, if the research question focuses solely on the genotype AA of a specific variant, it is irrelevant whether positivity is satisfied for genotypes AC or CC. Positivity is particularly important in statistical genetics since some genetic variations can occur in less than 1% of individuals. Since the conditional distributions $p(\mathbf{T}=\mathbf{t}|\mathbf{Z}=\mathbf{z}), \mathbf{z} \in \mathcal{Z}$ are unknown, positivity is impossible to measure exactly. Instead, an approximation of the $\min_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{t}|\mathbf{z})$ could be obtained by fitting a flexible machine learning model. However, the quality of this measure depends heavily on the extrapolation capabilities of the fitted model to regions not supported by the data. In population genetics, lowly represented subpopulations or ethnicities can suffer from poor extrapolation and lead to biased results. Most UK Biobank based studies thereby only include white individuals in their analyses, and associated inferences are not applicable to the broad UK population. We follow this convention in this thesis, but show with simulations in chapter 4, that this may not be necessary with double robust semi-parametric estimators.

Furthermore, there is no consensus method for responding to positivity violations [118]. In genetics, since the cost of false positives is higher than that of false negatives, rare variants are typically discarded. The minor allele frequency is

commonly used as a proxy for positivity and typical thresholds are 0.01 and 0.05 for populations of size 100 000 and 10 000 respectively [94]. The use of the minor allele frequency instead of the minor genotype frequency has no theoretical basis, and is likely due to the use of linear models encoding the number of minor alleles in the model. With simulations, we will propose a heuristic marginal positivity constraint for semi-parametric estimators in chapter 4.

2.2.3 The Average Treatment Effect

While theoretically well-defined, the estimation of the full interventional conditional density of equation 2.3 is difficult if not assumed to be part of a restricted parametric family [77]. In practice, causal effects are often defined in terms of expected values, which give an incomplete picture, but for which robust estimation methods exist [76]. The most important of such effects is the Average Treatment Effect

Definition 2.2.5 (Average Treatment Effect (ATE))**.** *Let T and Y be two random variables with T taking values in $\{0,1\}$. The Average Treatment Effect of T on Y is given by*

$$ATE = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \int y \cdot p(y|do(T=1)) \cdot dy - \int y \cdot p(y|do(T=0)) \cdot dy \quad (2.5)$$

The ATE measures the average difference in outcomes between a treatment group and a control group across a population. Since the ATE also involves $P(Y_t)$, if the later is identifiable, so is the ATE. Continuing with the example of figure 2.4, the ATE is identifiable and given by

$$\begin{aligned} ATE &= \int y \cdot \left(\int p(y|T=1,w) \cdot p(w) \cdot dw \right) \cdot dy - \int y \cdot \left(\int p(y|T=0,w) \cdot p(w) \cdot dw \right) \cdot dy \\ &= \int \int y \cdot (p(y|T=1,w) - p(y|T=1,w)) \cdot dy \cdot dw \\ &= \mathbb{E}_W \left[\mathbb{E}[Y|T=1,W] - \mathbb{E}[Y|T=0,W] \right] \end{aligned} \quad (2.6)$$

The ATE has an intuitive interpretation in terms of linear models, assume the following mechanism for Y

$$Y = \beta \cdot T + \gamma \cdot W + U_y \quad (2.7)$$

such that $\mathbb{E}[U_y] = 0$. Then the ATE is simply the β coefficient associated with T in the linear model

$$\begin{aligned} ATE &= \mathbb{E}[\beta \cdot 1 + \gamma \cdot W] - \mathbb{E}[\beta \cdot 0 + \gamma \cdot W] \\ &= \beta + \gamma \cdot \mathbb{E}[W] - \gamma \cdot \mathbb{E}[W] \\ &= \beta \end{aligned} \tag{2.8}$$

The ATE, or β coefficient, is the main quantity of interest in statistical genetics as will become clear shortly. However, if the true model is not as in equation 2.7, the ATE may not be trivially read of as a single coefficient. Which is why it is useful to define the quantity model-independently.

2.3 Current Approaches in Statistical Genetics

We now return to the statistical (not causal) identification of genetic variations associated with human traits and diseases. Following from Fisher's work, the field of population genetics has been largely dominated by parametric models based on Linear-Gaussian models. We discuss the statistical foundations of these models, the rise of the Linear Mixed Model (LMM) and emergent modern ideas.

2.3.1 The Foundational Linear-Gaussian Model

Generalised-Linear-Models (GLMs), and in particular Linear-Models play such a crucial role in modern population genetics that we review them first. Most of the content in this section is adapted from classical textbooks in statistics [52, 101]. In its classical version, the Linear-Model assumes that the relationship between a continuous health outcome Y, and a set of covariates $\mathbf{W}$, is linear. A genetic variation V (we will from now on drop the more general T notation in favour of V), is just a special covariate of interest for which we would like to estimate the effect on Y. Finally, an error term ε, assumed normally distributed $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, accounts for the uncertainty in the generating process. The model can be written as

$$Y = \beta \cdot V + \gamma \cdot \mathbf{W} + \varepsilon \tag{2.9}$$

In this equation, the quantity of interest, also called estimand, is the parameter β which captures the effect of V on Y. In the previous section, we saw that under causal assumptions, β has a causal interpretation, the average treatment effect. This parameter is unknown to the scientist and must

be estimated from data. It turns out that the maximum-likelihood estimator $\hat{\beta}_{ml}, \hat{\gamma}_{ml}, \hat{\sigma}_{ml} = \arg\max_{\beta,\gamma,\sigma} \prod_{i=1}^{n} p(Y_i|V_i, \mathbf{W}_i; \beta, \gamma, \sigma)^2$, benefits from remarkable statistical properties; it yields a minimum-variance unbiased estimator (MVUE) for β. Its sampling distribution is given by

$$\begin{cases} \begin{bmatrix} \hat{\beta}_{ml} \\ \hat{\gamma}_{ml} \end{bmatrix} = (\mathbf{W}^T \cdot \mathbf{W})^{-1} \mathbf{W}^T y \\ \begin{bmatrix} \hat{\beta}_{ml} \\ \hat{\gamma}_{ml} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \beta \\ \gamma \end{bmatrix}, \sigma^2 (\mathbf{W}^T \cdot \mathbf{W})^{-1} \right) \end{cases} \tag{2.10}$$

Since σ^2 is also unknown, its maximum-likelihood estimate is used instead and the standardised $\hat{\beta}$ follows a student distribution t_{n-p} where p is the total number of covariates (including V).

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 [(\mathbf{W}^T \cdot \mathbf{W})^{-1}]_{11}}} \sim t_{n-p} \tag{2.11}$$

This essential statistic, known as a pivot, allows the construction of confidence intervals and to conduct hypothesis testing.

So far, we have only considered the case where Y is a continuous variable but most traits we are interested in are binary (e.g. disease). In this case, the conditional distribution of Y is conventionally modelled using a logistic regression

$$p(Y|V, \mathbf{W}) = \frac{1}{1 + e^{-(\beta \cdot V + \gamma \cdot \mathbf{W})}} \tag{2.12}$$

Note that in this case, β is not the average treatment effect anymore. It is interpreted as the change in the log-odds of the outcome $Y = 1$ for a unit increase in V, holding all other predictors constant. Unlike the continuous outcome case, maximum-likelihood estimation does not yield closed form solutions for the logistic regression model. Furthermore, the sampling distribution of the estimator is not exactly, but only asymptotically Gaussian. These two limitations illustrate how the departure from the linear assumption complicates statistical inference, even in parametric models.

So far, for both continuous and binary outcomes, we have yet to define the set of covariates $\mathbf{W}$. Ideally, this set must be a valid backdoor adjustment set as per equation 2.3. However, as will be discussed in section 3.2, this is impossible in practice. Instead, approximate heuristics were developed. For instance, it was

demonstrated that population stratification can lead to spurious associations, and Principal Component Analysis (PCA) has been identified as a method to capture this substructure [122]. In practice, principal components are first estimated from the entire genotyping array as a pre-processing step, and then included as covariates of the statistical model. While concerns have been raised about findings based on PCA, which we discuss further in section 3.2.3.1, it remains a widely accepted practice [40].

Finally, the central hypothesis implicitly assumed in this section is that of independent and identically distributed individuals. While this assumption may be innocuous in many fields, it is unlikely to hold in population genetics. Study participants can be related to each other through descent (e.g., individuals within the same family) or via cryptic genetic similarities. Addressing this limitation was the primary motivation behind developing the now widely accepted Linear-Mixed-Model, which we now discuss.

2.3.2 The Linear Mixed Model

Linear-Mixed-Models (LMMs) were first proposed as a unified method crossing the boundary between family-based and structured association samples [182]. Interestingly, LMMs propose to address two statistical issues at once: population stratification and cryptic relatedness between individuals. This is done by incorporating a genetic relationship matrix (GRM) into traditional generalised linear models. More precisely, and in its most traditional version, the LMM is a joint population level model with so-called fixed and random effects given by the following definition.

Definition 2.3.1 (Linear-Mixed-Model)**.** *A linear-mixed-model is defined by the following equation:*

$$\mathbf{Y} = \beta \cdot \mathbf{W} + \mathbf{Z} \cdot \mathbf{u} + \boldsymbol{\varepsilon} \tag{2.13}$$

with

- $\mathbf{Y}$ *is the outcome variable across all individuals in the population.*
- $\mathbf{W}$ *is a design matrix whose effect on* $\mathbf{Y}$ *is measured by* β
- $\mathbf{Z}$ *is the genetic relationship matrix whose effect on* $\mathbf{Y}$ *is measured by the random variable* $\mathbf{u}$

- $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$, *where* $\mathbf{G}$ *is a covariance matrix.*
- $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, *where* $\mathbf{R}$ *is another covariance matrix.*
- $Cov(\mathbf{u}, \boldsymbol{\varepsilon}) = \mathbf{0}$

Note that in this case, the variant V is incorporated into $\mathbf{W}$ to keep the notation compact.

The GRM measures the genetic similarity between individuals in the population using the sample correlation coefficient between their (centred and scaled) variants. The GRM, denoted $\mathbf{Z}$, is thus of size $N \times N$ where N is the number of individuals in the population. Given M variants, it is defined as:

$$Z_{ij} = \frac{1}{M-1} \sum_{k=1}^{M} \frac{(s_{ik} - 2p_k)(s_{jk} - 2p_k)}{2p_k(1-p_k)}. \tag{2.14}$$

Here $s_{ik} \in \{0, 1, 2\}$ denotes the number of copies of the reference allele for individual i at variant k, and $p_k \in (0, 1)$ denotes the frequency of the reference allele at variant k over the population of N individuals. In particular, the population average of s_{ik} equals twice the reference allele frequency at variant k, i.e., $2p_k$ (one for each strand copy), so

$$\frac{1}{N} \sum_{i=1}^{N} s_{ik} = 2p_k. \tag{2.15}$$

Thus $\tilde{s}_{ik} = s_{ik} - 2p_k$ is the zero-centred count of the number of copies of the reference allele of individual i at variant k. Considered as a random variable, $\tilde{s}_{ik}$ takes on three values. Assuming reference alleles are sampled binomially with mean frequency p_k, the standard deviation of $\tilde{s}_{ik}$ equals $\sqrt{2p_k(1-p_k)}$. This explains the additional factor in Equation 2.14 that scales the variables $\tilde{s}_{ik}$ and $\tilde{s}_{jk}$ so as to have unit variance. The GRM thus represents a measure of genetic dependence among individuals, and, it could also capture genetic confounding since it contains whole genome information.

Under definition 2.3.1, the outcome $\mathbf{Y}$ follows a multivariate Gaussian distribution with mean $E[\mathbf{Y}] = \boldsymbol{\beta} \cdot \mathbf{W}$ and variance $Var[\mathbf{Y}] = \boldsymbol{\Sigma} = \mathbf{Z} \cdot \mathbf{G} \cdot \mathbf{Z}^T + \mathbf{R}$. It is thus natural to estimate the parameters of this model via maximum-likelihood estimation. Since we are interested in the estimation of the fixed effect $\boldsymbol{\beta}$, we denote the nuisance variance parameters by $\boldsymbol{\eta}$ such that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\eta})$. It turns out that the value of the fixed effect that maximises the likelihood can be written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{W}^T \cdot \boldsymbol{\Sigma}(\boldsymbol{\eta})^{-1} \cdot \mathbf{W})^{-1} \cdot \mathbf{W}^T \cdot \boldsymbol{\Sigma}(\boldsymbol{\eta})^{-1} \mathbf{Y} \tag{2.16}$$

Since η is unknown, it needs to be estimated first. However, maximum-likelihood estimation leads to biased estimators for η. Unbiased estimates can be obtained with a slight adjustment by optimising the residual maximum-likelihood (REML), which optimises the likelihood of the residuals instead.

It is instructive to investigate the computational complexity of LMMs in their most general form as presented in definition 2.3.1. Optimisation of the REML uses iterative procedures that require inversion and computation of the determinant of the covariance matrix $\mathbf{\Sigma}$ at every step. Both operations have an approximate computational complexity of $O(N^3)$ resulting in a total computational complexity of $O(I \cdot N^3)$ where I is the number of iterations. As per equation 2.16, the estimation of the fixed effects also requires the inversion of $N \times N$ matrices also resulting in $O(N^3)$ complexity. For testing K variants the total complexity of a GWAS using LMMs is thus of order $O(K \cdot I \cdot N^3)$, scaling with the cube of the sample size. Given the increasing size of cohorts individuals and genotyping arrays, the UK Biobank contains approximately 500 000 individuals genotyped at around 800 000 loci, a GWAS would take years of CPU time to run [185].

Most of the past developments have thus focused on numerical approximations and optimisations to scale LMMs to larger cohorts. For example, the BOLT-LMM algorithm, reduced the complexity to $O(K \cdot N)$ using a variational approximation and employing a Bayesian mixture prior on variants' effect sizes [85]. Another method, fastGWA leverages a sparse GRM and scales to more than 2000 traits on array-genotyped and imputed samples from 450 000 individuals [71].

LMMs are thus attractive because they provide an attempt to address the two long standing problems of dependent individuals and confounding. One limitation of LMMs is that they still rely on a strong parametric assumptions. As we discuss shortly in section 2.4, this can limit their validity and the correctness of the effect sizes obtained from them. These limitations have also been pointed out by others, and several methods proposed to address them [32, 85, 95].

2.3.3 Recent Advances in Statistical Genetics

In this section we discuss three recent advances in statistical genetics that use more flexible and realistic methodologies. They either attempt to control for whole-genome confounding, and thus take us a step closer to causal inference, use more flexible modelling strategies, or both.

One interesting approach, called REGENIE, was presented in 2021 by Mbatchou et al. [95]. REGENIE is a two steps procedure that fits a whole-genome model for each phenotype of interest. In the first step, the genome is split in blocks and a ridge linear regression model of the outcome is fitted for each block using cross-validation. In a second step, the ridge regression models' predictions are used as covariates for association testing of each variant of interest. To avoid correlated covariates due to linkage disequilibrium, REGENIE only uses predictors from other chromosomes than the tested variant in the second step. This means that while the method uses the whole genome, only population structure confounding can effectively be controlled. REGENIE offers several computational benefits compared to LMMs; first it does not rely on the GRM and thus keeps memory requirements low, second it is easily parallelised. However, REGENIE still relies on potentially mis-specified linear/logistic models, the effect of which may be exacerbated by the two steps procedure.

DeepNull is a semi-parametric method which models non-linear covariate effects using neural networks [96]. Like REGENIE, DeepNull proceeds in two steps. In the first step, the effect of covariates on trait is fitted using a flexible neural network. In the second step, association testing is performed using a standard linear model and the neural network's prediction as extra covariate. This means that the effect of the genetic variant is still assumed to be linear and does not allow complex interactions between covariates and genetic variants. DeepNull is entirely subsumed by the approaches presented in section 2.4.

Finally, a entirely different framework, KnockoffGWAS, aims at controlling the false discovery rate in genome-wide association studies [137]. KnockoffGWAS works by generating knockoffs [7], indistinguishable negative control variables, that correct for linkage disequilibrium and account for population structure. As such, KnockoffGWAS is a significant step forward to the identification of causal variants. Furthermore, KnockoffGWAS does not rely on unrealistic assumptions regarding the conditional distribution of the outcome given the genome. Instead, it models the randomness in the genome using a Hidden-Markov model, similarly to imputation methods [93]. The caveat of knockoffs is that they do not yield effect sizes, and thus need to be used in conjunction with other methods to quantify the effect of interventions.

2.4 Semi-Parametric Estimation

We have seen that parametric methods are particularly appealing because they enable the construction of optimal estimators of model parameters, at least asymptotically. They are also interpretable, the scientific questions of interest can be directly answered from the value of these parameters. The theory of Ordinary Least Square (OLS) is in fact remarkably robust. In the absence of confounding, the OLS estimator is asymptotically unbiased for the ATE, and its variance is never worse than that of the difference in mean estimator [165]. This, even if the actual data generating process is not linear, justifying the use of linear models in randomised control trials where the unconfoundedness assumption is enforced by construction. In observational studies, in particular in population genetics, confounding is known to exist. Correct inference using parametric models thus relies on the assumption that the model is correctly specified [124]. To illustrate, we perform a simple simulation study, both in the absence or presence of confounding using a slightly mis-specified linear model. We then compare the sampling distributions of both the linear and the semi-parametric targeted maximum-likelihood estimator (see 2.4.3). The dataset is generated according to the following structural causal model

$$W \sim \mathcal{N}(1,1) \tag{2.17}$$

$$U_T \sim \mathcal{N}(0,1) \tag{2.18}$$

$$U_Y \sim \mathcal{N}(0,1) \tag{2.19}$$

$$T_U \sim \mathcal{B}(0.2) \quad \textbf{OR} \quad T_C = \mathbb{1}(W + U_t < 0) \tag{2.20}$$

$$Y = \alpha \cdot W + \beta \cdot T_{U|C} + \gamma \cdot W \cdot T_{U|C} + U_Y \tag{2.21}$$

while the parametric model used to fit the data is

$$Y = \alpha \cdot W + \beta \cdot T_{U|C} + U_Y \tag{2.22}$$

Thus, the only difference between the true generating process and the fitted model lies in the interaction term $\gamma \cdot W \cdot T_{U|C}$, where T_U is unconfounded and T_C is confounded by W. In the model of equation 2.21, the ATE can be computed exactly regardless of confounding, and is equal to $ATE_0 = \beta + \gamma \cdot \mathbb{E}[W] = \beta + \gamma$. We sample 1000 bootstraps of 1000 samples using $\alpha = 0, \beta = 1, \gamma = -1$ and present the resulting distributions in figure 2.5. As discussed, in the unconfounded scenario,

the OLS estimate is centred on the true ATE (but not β as one could expect). Note that the targeted maximum-likelihood estimation is also unbiased and has smaller variance, it is thus more powerful. In contrast, in the confounded scenario, the OLS estimates lie in between the ATE and the model's β coefficient, leading to incorrect inference and possibly increased error rates in large scale association studies. In contrast, the targeted maximum-likelihood estimator, while resorting to the same misspecified model, has smaller bias and covers the true effect. In this simple situation, inserting the missing interaction term into the fitted linear model would solve the problem. However, in practice, we do not know anything about the real nature of the relation $Y = f_Y(W, T)$ which might be arbitrarily complex. Instead of developing new modelling paradigms tailored to specific beliefs, semi-parametric estimators provide a one size fits all type of methodology. The lack of knowledge is bypassed by using flexible and data-adaptive machine-learning methods together with theoretical non-parametric statistics [133]. In population genetics, this could pave the way to the investigation of arbitrarily complex determinants of human health, including higher-order gene-gene and gene-environment interactions.

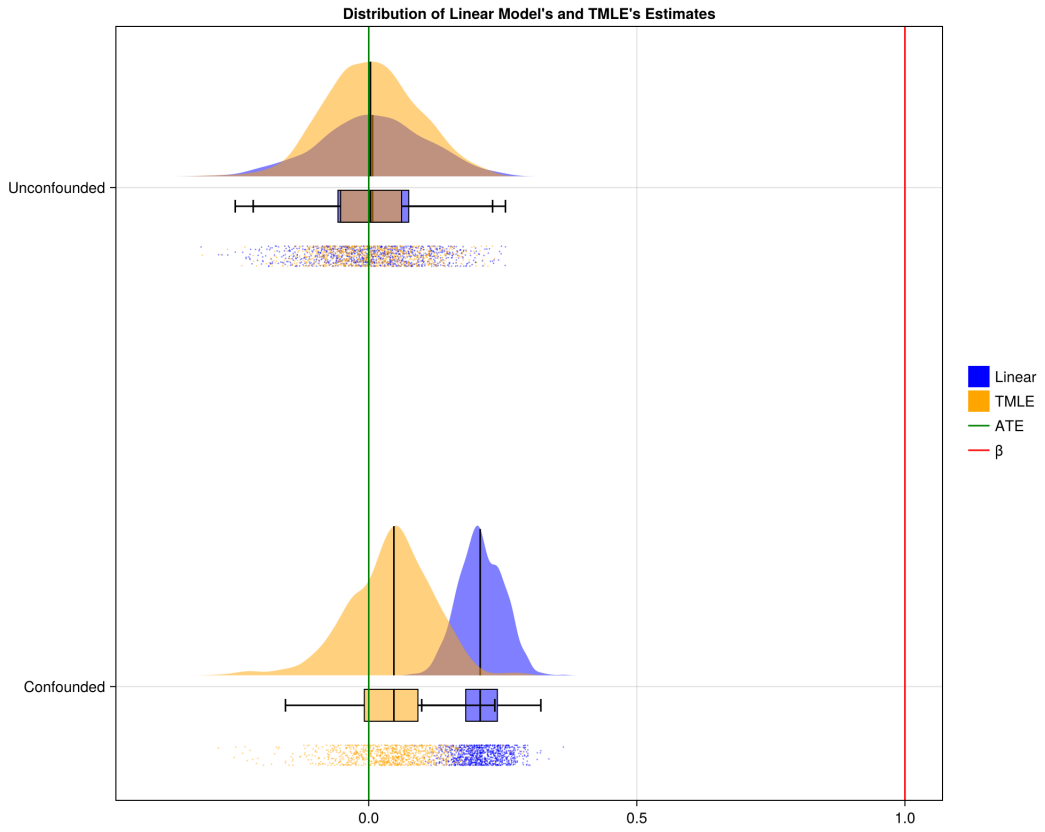

Figure 2.5: **Illustration of model mis-specification**. A bootstrap analysis comparing the linear model and the targeted maximum-likelihood estimator. In the unconfounded case, both estimators perform well. The targeted maximum-likelihood estimator has smaller variance (0.005) than the linear estimator (0.008). In the confounded case, the linear model is biased (0.210), it does not cover β either. The targeted maximum-likelihood estimator however, has smaller bias (0.041) and covers the ATE.

Non-parametric and semi-parametric statistical methods aim to provide robust and reliable inferences by considering larger classes of distributions. Ideally, we would like to work with the fully non-parametric model that encompasses all possible distributions. However, while some machine-learning algorithms are fully non-parametric [12], most methods rely on some form of parametrization (e.g., neural networks). We will thus use the semi-parametric terminology in this thesis, even though the estimation methods we are about to discuss are theoretically fully non-parametric.

In this chapter we present an overview of the two main semi-parametric estimation strategies used in this thesis. They are called the one-step estimator

(OSE) and targeted minimum loss-based estimator (TMLE), and an illustration of their behaviour is presented in figure 2.6. We emphasise on intuition and graphical exposition and defer mathematical details to section 3.4. In the semi-parametric framework, the estimands or objectives of the analysis, cannot be captured by model parameters. Instead they are more generally represented by statistical functionals. A functional is a function $\Psi : \mathcal{M} \mapsto \mathbb{R}$ that maps a data generating process, or distribution P, to the real numbers. It can be understood as a feature of this distribution. The natural world, is the realisation of a ground truth generating process we denote by P_0. We are thus interested in the value of the estimand at P_0, i.e., $\Psi_0 = \Psi(P_0) \in \mathbb{R}$. The ATE, presented in section 2.2.3, is an example of functional where the dependence on the generating process can be made explicit $ATE(P_0) = \mathbb{E}_{P_O}[Y_1] - \mathbb{E}_{P_O}[Y_0]$.

2.4.1 Plugin Estimation

Given an estimand of interest, an intuitive estimation strategy consists in estimating P_0 using machine-learning methods, obtain an estimate $\hat{P}$, and then evaluate $\Psi(\hat{P})$ as an estimate for Ψ_0. Despite being semi-parametric, this method often referred to as G-computation [128], or plugin estimation, has a few drawbacks. The first caveat is that, if the employed machine-learning model is mis-specified, the plugin estimate will remain biased. Secondly, the plugin estimator does not have a clear sampling distribution, thereby complicating uncertainty quantification and hypothesis testing. In principle, the non-parametric bootstrap [39] could be employed to obtain a sampling distribution, but it is practically unfeasible since computationally intensive machine-learning algorithms need to be fitted for each bootstrap sample (e.g., 1000 times).

To address these limitations, modern methods take advantage of the fact that most estimands of interest, like the average treatment effect, are "smooth" [76]. As may be expected, the main ingredient used by all methods is the gradient of Ψ, sometimes also called the efficient influence function. This gradient is a function that needs to be mathematically derived for each new estimand of interest. Fortunately, all genetic effects of interest in this thesis are linear combinations of a well-studied estimand: the mean under intervention (section 3.3.1). Since the gradient is a linear map, it is easily derived for linear combinations of this estimand. We present all necessary gradients in section 3.3.5.

Finally, while they differ in philosophy and may have different finite sample properties, the OSE and the TMLE are both asymptotically unbiased and have minimum variance (efficient). They are also asymptotically normal with variance given by the variance of the gradient of Ψ. Therefore, confidence intervals and hypothesis testing can be performed using traditional methods such as t-tests.

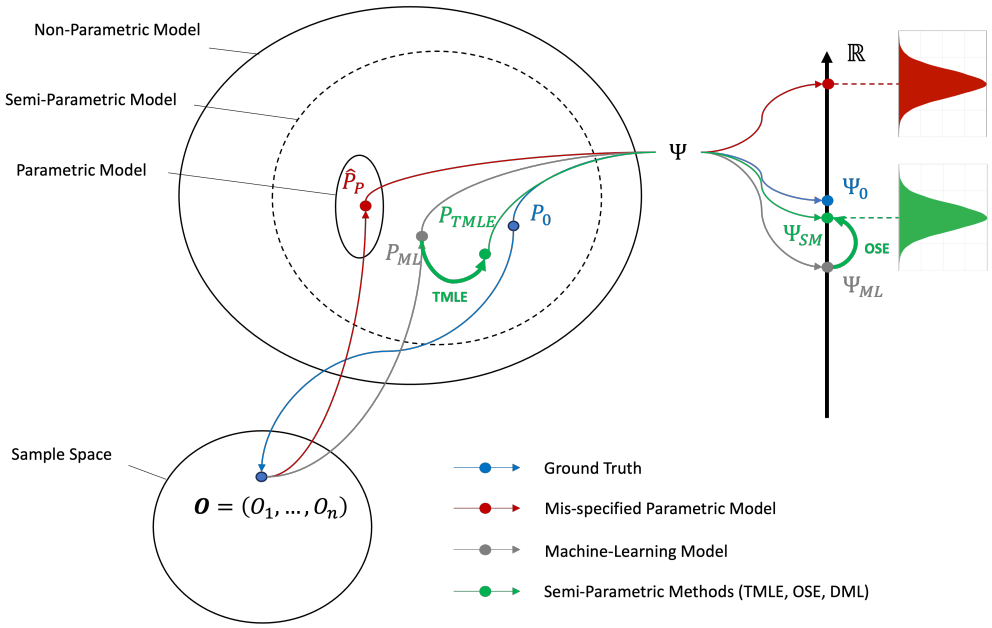

Figure 2.6: **Semi-Parametric Estimation Methods** A single realisation of the data $\mathbf{O} = (\mathbf{O}_1, \ldots, \mathbf{O}_n)$ is generated according to P_0 (blue). A misspecified parametric model (red) has no guarantee to provide correct inference. Semi-parametric methods (green) correct the bias resulting from an initial flexible machine-learning method (grey). The OSE corrects the plugin bias in the estimand's space whereas the TMLE corrects the bias in distribution space. For readability, a single semi-parametric estimate $\hat{\Psi}_{SM}$ is presented. In reality, while close, $\hat{\Psi}_{OSE}$ and $\hat{\Psi}_{TMLE}$ will exhibit finite sample differences. Both estimators are asymptotically normal and confidence intervals can be built using conventional methods.

2.4.2 One-Step Estimation

One-step estimation is a simple, yet effective way to obtain an asymptotically unbiased and efficient semi-parametric estimator for an estimand Ψ given an initial estimator $\hat{P}$ for P_0 [14, 76]. Since the initial plugin estimate $\hat{\Psi} = \Psi(\hat{P})$ is biased, the OSE simply estimates this residual bias and removes it from the initial estimate. That is, $\hat{\Psi}_{OSE} = \hat{\Psi} - \hat{Bias}_{\hat{\Psi}}$. The intuition is that the OSE takes

a Newton–Raphson step in the estimand's space (e.g. $\mathbb{R}$) using the gradient of Ψ.

The limitation of the OSE is that it is not a plugin estimator anymore and may fail to respect the natural bounds of the estimand. For example, if the estimand is a probability, it should lie between 0 and 1. In population genetics, effect sizes are typically very small, this theoretical limitation has thus little practical implications.

2.4.3 Targeted Minimum Loss-based Estimation

Targeted minimum loss-based estimation bypasses the limitation of the one-step estimator by performing the Newton–Raphson step in distribution space. It corrects the initial estimator $\hat{P}_{ML}$ to a targeted $\hat{P}_{TMLE}$ which is optimal for the estimand of interest. The correction step fluctuates the nuisance estimate $\hat{P}_{TMLE}$ to solve a set of estimating equations that are derived from the mathematical study of the estimand of interest Ψ. As such, the TMLE requires more preliminary mathematical work than the OSE. However, by linearity of the gradient, the fluctuations required in this thesis are easily obtained.

Then, TMLE computes Ψ at $\hat{P}_{TMLE}$, that is $\hat{\Psi}_{TMLE} = \Psi(\hat{P}_{TMLE})$, and is thus a plugin estimator. Heuristically, $\hat{P}_{TMLE}$ is obtained by optimally transporting the initial $\hat{P}_{ML}$ along the gradient of Ψ to minimise the residual estimation bias.

2.5 Discussion

In this chapter, we have emphasised the critical role of identifying genetic variants that causally influence human traits and diseases. Advances in modern genotyping and sequencing technologies, combined with large-scale epidemiological data, have enabled the discovery of genetic regions associated with complex traits. However, due to linkage disequilibrium and population-specific genetic ancestry, pinpointing the actual causal variants remains a challenge. Furthermore, understanding the biological mechanisms through which these variants exert their effects is essential for therapeutic development, yet requires further investigation.

Traditionally, these genetic associations have been identified using variations of linear models. However, we have demonstrated that when confounding variables are present, these associations can be biased if the true data-generating

process is not inherently linear. To overcome these limitations, modern semi-parametric methods offer a promising solution. By integrating flexible machine-learning techniques with semi-parametric statistical theory, these methods can mitigate bias while maintaining the asymptotic properties of the estimators.

In the following chapter, we will introduce a general framework designed to define and estimate genetic effects, providing a pathway to more accurate and actionable insights into the genetic architecture of complex traits.

Chapter 3

The Causal Roadmap of Population Genetics

Randomised controlled trials (RCTs) stand out as the favoured study design for estimating the causal effects of interventions. This preference arises from their near ideal experimental protocol, a randomised treatment assignment, tailored to minimise bias [60]. However, RCTs also present certain drawbacks. They often incur significant time and financial costs, potentially lack generalisability when the participants are not representative of the wider population and sometimes prove to be simply unfeasible. This is the case in population genetics, where a RCT would be highly unethical.

At the other end of the spectrum, modern technological advances have contributed to the expansion of real-world data (RWD), that is, data collected outside the context of highly controlled clinical trials. These massive and heterogeneous datasets have the potential to overcome some limitations of RCTs. The UK Biobank for instance, is a prospective study of 500 000 individuals that allows the investigation of the genetic determinants of diseases of middle and old age [147].

However, these datasets commonly exhibit structural biases such as confounding, missingness and noise; raising questions about their suitability for generating what is termed as real-world evidence (RWE). In particular, a major concern associated with RWD is that of internal and external validity, which are two forms of causal gaps (equation 3.1). Internal validity relates to whether the outcome of a study actually measures what it aims to measure, while external validity refers to how well the findings will apply to the wider population of interest.

Well aware of these challenges, regulatory agencies such as the United States

Food and Drug Administration or the National Institute for Health and Care Excellence in the UK, have formulated frameworks and guidelines for the generation of RWE [78, 108]. In 2023, Dang et al. proposed a 7 steps unified framework, termed the Causal Roadmap [81], to generate high-quality RWE. This framework is structured as follows. In step 1, the causal question of interest and associated causal estimand are defined. A causal model, incorporating background knowledge and information about the data generation process is built. In step 2, the actual data is considered. The goal is to understand if the proposed causal model is representative of the real data or needs to be refined, for example due to missingness. In step 3, identifiability is assessed (definition 2.2.3). Can the causal estimand be expressed as a statistical quantity in the proposed causal model? The difference between the causal estimand and the statistical estimand is often called the causal gap. While ideally null, the causal gap cannot be measured in observational studies since identifiability depends on unverifiable assumptions. It is thus of particular importance to build multi-disciplinary collaborations to assess the credibility of a causal model and assumptions leading to identifiability. In step 4, the study can decide to proceed and commit to a statistical estimand. This estimand should be as close as possible to the causal estimand of interest in order to minimise the causal gap. In step 5, a statistical model and estimator that respect available knowledge are chosen. For instance, known bounds on specific variables can be included in the statistical model. However, unrealistic assumptions such as linearity can lead to estimation error and should be avoided. This potential estimation error is often called the statistical gap. The total estimation error can thus be decomposed as

$$\text{Total Estimation Error} = \text{Causal Gap} + \text{Statistical Gap} \tag{3.1}$$

In step 6, sensitivity analyses attempt to quantify how the estimated results would change if the causal identification assumptions were violated. For example, assume a study reported a 95% confidence interval $0.5 - 0.7$ for a genetic effect on disease. Subsequently, sensitivity analysis shows that the causal gap can be no more than 0.3. It can be concluded that the genetic effect on disease is likely positive and can be further investigated. Finally, in step 7, alternative study designs can be considered. For instance if the causal or statistical gap are deemed too high and the risk induced by a false positive too consequential.

In this chapter, we illustrate how the Causal Roadmap can be applied to

analyse the effect of genetic variations on human health. Instead of following the order presented above, we start with step 2, and discuss the UK Biobank data first. This is because the UK Biobank is a prospective study for which all the data has already been collected. We next present a causal model of inheritance and the causal estimands of interest. We discuss the non-identifiability of these estimands, the associated implications, as well as ideas for future research. We propose and commit to a working causal model under which causal effects are identified. Finally, we derive general semi-parametric estimation strategies in order to minimise the statistical gap. Sensitivity analysis methodologies and alternative designs are not tackled and left for future work.

3.1 The UK Biobank

The UK Biobank is a large-scale biomedical database, aimed at improving the prevention, diagnosis, and treatment of a wide range of serious and life-threatening illnesses [147]. Established in 2006, it collects and stores biological samples and health-related information from about 500 000 volunteer participants aged, upon recruitment, between 40 and 69 years old across the United Kingdom. The data includes genetic, lifestyle, and health information, along with blood, urine, and saliva samples. The UK Biobank is unique due to its scale, scope, and the depth of data available, making it one of the most comprehensive health research resources globally. It is thus an exceptional resource to study how genetic and environmental factors contribute to the development of diseases.

However, like any prospective study, the UK Biobank suffers from some limitations. First, as a population study, individuals in the UK Biobank may be related to one another. For instance because participants can influence each other's decision to enrol. This means that the statistical assumption of independent units may not be true (see section 3.6). Second, the UK Biobank suffers from selection bias. This is particularly important to consider since this determines how research findings can be generalised to the broader population (external validity). For instance, it was shown that UK Biobank participants were more likely to be older, female, and to live in less socioeconomically deprived areas than non-participants [50]. This is a serious problem for equal access to healthcare since findings from the UK Biobank are less likely to impact poorly represented individuals. These issues relate to the study design (step 7 of the Roadmap) and future

studies will likely benefit from the experience gained from the UK Biobank. For instance, the new UK's health research programme, Our Future Health, aims to recruit a broader and more diverse part of the population [139]. Another criticism pointed out that the UK Biobank omits early and late life stages critical for understanding ageing [17]. It advocated funders and researchers not to neglect complementary small scale targeted study designs.

In this thesis, we acknowledge these limitations and propose a non-parametric method to account for the possible dependence between individuals (section 3.6). We now present the two main datasets used throughout, the genotyping dataset and the phenotyping dataset.

3.1.1 Genotyping Dataset

The genotyping data (see section 2.1.2), covers 489 212 participants and includes information on approximately 800 000 genetic variants (single nucleotide polymorphisms and short insertions and deletions) per participant. The data was primarily generated using two customised genotyping arrays: the Affymetrix UK Biobank Axiom Array and the Affymetrix UK BiLEVE Axiom Array. These arrays were designed to capture a broad range of genetic diversity, including common and rare variants, as well as specific variants relevant to the UK population.

Unlike sequencing, genotyping only provides a sparse representation of the whole genome. To enhance the comprehensiveness of the genetic data, the UK Biobank performs genotype imputation. It predicts the genotypes at ungenotyped variants using a reference panel, which in the case of the UK Biobank includes data from the Haplotype Reference Consortium and the 1000 Genomes Project [86, 142]. This process increases the number of genetic variants available for analysis to approximately 90 million. However, imputation quality at a specific locus, depends heavily on the variant's frequency. Figure 3.1 from the UK Biobank's initial release, shows that imputation quality drops significantly for variants with minor allele frequency (MAF) $< 5\%$. Given this uncertainty, it is current practice in genome-wide association studies to only consider frequent variants with MAF $> 5\%$. A more principled approach involves calculating an imputation quality score for each genetic variant and retaining only those variants that meet a specified threshold [93]. Various metrics have been devised,

each specific to the associated imputation method. Since the UK Biobank uses IMPUTE2 [67], the metric used is the INFO measure. It is defined as the ratio of the observed and complete information [93]. The score ranges from 0 to 1, with higher values indicating more reliable imputation. In this work, we only consider imputed variants with INFO > 0.9.

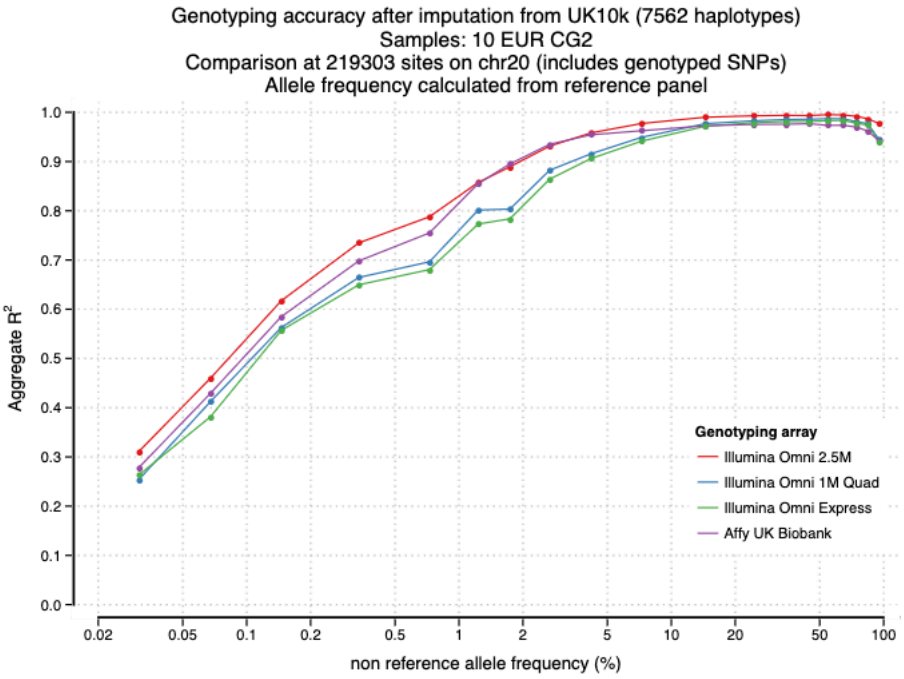

Figure 3.1: **Precision of Genotyping Imputation.** From UK Biobank release v1. Imputation quality drops quickly for rare variants. To mimic a typical imputation analysis, a pseudo-GWAS dataset was constructed by extracting the CG SNP genotypes at all the sites included on a given array. All sites not on the array were then imputed using the UK10K reference panel. Variants were stratified into allele frequency bins and the squared correlation (R^2) was calculated between the allele dosages at variants in each bin with the masked CG genotypes.

Despite this improvement, imputation still provides an incomplete representation of genetic variations. As a comparison, in 2024, the gnomAD project identified 644 267 978 high-confidence short nuclear variants from 76 215 diverse human genomes [22]. Out of these, 390 393 900 are rare (minor allele frequency $< 0.1\%$). This incomplete information will complicate the identification of causal effects as will be discussed in section 3.2.2. The recent release of whole genome sequences in November 2023, will present an unprecedented opportunity to explore rare genetic variations in greater depth than ever before.

3.1.2 Deep Phenotyping Dataset

The phenotyping dataset in the UK Biobank is a comprehensive collection of health outcomes and lifestyle data gathered from participants. It contains all non-genetic information that we use throughout this work. More precisely, we investigate around 110 non-binary and 660 binary traits as previously defined by [21]. Phenotypes in the UK Biobank are organised in Data-Fields which can contain information about multiple phenotypes. We first describe the fields we consider throughout this thesis and then explain how they are processed to define individuals' phenotypes.

3.1.2.1 UK Biobank Data-Fields

Diagnoses made during hospital inpatient admission (467 traits, Data-Fields 41204 and 41202) This category contains information relating to main and secondary diagnoses made during hospital inpatient admissions. Each trait in this category corresponds to a node or set of nodes in the International Classification of Diseases 10-th Revision (ICD10) ontology [113]. For example, "K41 Femoral hernia" is defined by any of the following diagnoses ("K410", "K411", "K412", "K413", "K414", "K419").

Self reported conditions (161 traits, Data-Field 20002) This category contains data obtained through a verbal interview by a trained nurse on past and current medical conditions, including type of cancer and other illnesses, the number of medical conditions, and date of diagnosis. This field can contain inaccuracies due to recall bias or intentional misreporting.

Self reported conditions (29 traits, Data-Field 40006) This category contains coded data on cancer incidence, obtained through linkage to national cancer registries. Because data is continually accruing, the number of cases may vary depending on the time the data is sent out to researchers.

Further Miscellaneous fields (113 traits) These traits do not correspond to diseases but to various lifestyle behaviours or biological measurements. For instance Data-Field 30180 corresponds to Lymphocyte percentage in a blood assay while Data-Field 1389 corresponds to Pork intake.

Note that the traits under investigation are not all independent. For instance a self-reported condition can also be diagnosed or be present in a cancer registry.

3.1.2.2 Data-Fields to Phenotypes

The information contained in the UK Biobank dataset is not directly usable for statistical analyses and needs to be processed. This is both because the data format is non standard and because some fields contain information about multiple phenotypes. Furthermore, multiple fields can also be combined to define custom phenotypes. This guide provides a practical description of the downloaded raw dataset. In this work, we process fields according to rules. These can be defined either at the field level, or based on its "value type" metadata, which indicates the data type the field contains.

Continuous and Integer Fields (value type $= 31, 11$ respectively) These contain one or more values per individual corresponding to the multiple visit assessments. The value of the field at the first visit assessment is extracted. For instance, the "Alcohol intake frequency." is simply the value from the "1558-0.0" column.

Ordinal Fields (1408, 1727, 1548, 728, 1717, 1389, 1478, 1518, 1558, 1349, 1359, 1369, 1379, 1329, 1339, 1239, 1687, 1697, 1319, 1498) Some variables are encoded as categorical by the UK Biobank but an ordinal interpretation seems more appropriate. For instance, "Lamb intake" ranges from 0 (Never) to 5 (Once or more daily). Similarly to continuous fields, the value of the field at the first visit assessment is extracted. Negative values (Do not know, Prefer not to answer) are treated as missing.

Categorical Fields (value type $= 21$) These variables cannot be turned into continuous variables and are usually represented via a coding. For instance, the field 21000, describes self-reported ethnic background. It uses coding 1001 for British individuals, 4001 for Caribbean individuals, 5 for Chinese individuals, etc. These variables can currently be processed in two different ways by TarGene (see chapter 5). Either the coding of each individual is extracted, or, if multiple codings are specified, an indicator of any of these codings is produced (1 or 0). For instance, "White" could be defined as any of the following codings 1001, 1002, 1003, that is British, Irish or Any other white background. Similarly to ordinal variables, the

value at the first visit assessment is extracted and negative values are considered missing.

Categorical Arrayed Fields (40006, 20002, 41202, 41204) Some fields in the UK Biobank comprise a list of codings for each individual at each visit. Each coding represents, for example, a disease or condition that was diagnosed or self reported. A phenotype is then defined by a set of codings. As an example, we define "oesophageal disorder" as the set of the following codings: $\{1134, 1139, 1140, 1141, 1138, 1474\}$ from Data-Field 20002. An individual annotated with any of these codings, at any assessment visit, is considered a case for "oesophageal disorder". Moreover, some fields share the same codings, this is the case for 41202 and 41204. In that situation it may be useful to aggregate these sources of information. For instance, "G20 Parkinson's disease" is defined by the single element set $\{G20\}$, in any of the 41202 and 41204 Data-Fields.

Together, these illustrate that the UK Biobank contains a rich set of epidemiological data of various data types. In the analysis methods discussed later, we slightly simplify this diversity by categorising traits as either continuous or binary. Consequently, ordinal phenotypes are not analysed with specific ordinal regression models but are instead treated as continuous variables. Although this approach facilitates streamlined analysis, more nuanced methods that account for the ordinal nature of certain traits could enhance the precision and reliability of the resulting inferences.

3.2 A Causal Model of Inheritance

In this section, we show how the formalism of section 2.2 can explicitly represent our knowledge of genetic inheritance. Using the model, we will demonstrate why the identification of causal variants is difficult. Finally, we will further show how current methods approach these difficulties and conclude with the working causal model used in the rest of the thesis.

3.2.1 The Causal Model of Genetic Inheritance

Now that we have discussed the available data, the UK Biobank, we can consider its integration within a causal modelling framework. The proposed causal model is inspired by related work in Mendelian Randomisation [152] and is presented in

figure 3.2. For readability, exogenous variables are omitted from the graph. We first provide a general description of the model before turning to the question of identifiability in the next section (3.2.2).

According to the model of figure 3.2, an offspring's trait (Y) is determined by their genome $G = (G_1, ..., G_p)$ as well as external environmental factors outside the model (not represented). The offspring's trait is also affected by the parents' phenotypes (P^M, P^F), via dynastic or familial effects. For example, parental alcohol consumption patterns can influence the child's alcohol consumption as well as other traits [175]. Note that while (P^M, P^F) can comprise the outcome of interest Y, they are not restricted to it and consist in all traits that influence both Y, and the reproduction event between the two parents.

An individual's genome is the combination of maternal and paternal haplotypes $G_i = (H_i^M, H_i^F)$ through assortative mating (R). A haplotype, is the result of meiosis (M^M, M^F), a specialised type of cell division that reduces the chromosome number in the original cell ((G^M, G^F)) by half (figure 2.1). Importantly, meiosis introduces genetic variation through crossing over and independent assortment of chromosomes. The immediate consequence is that genetic loci are not inherited independently from each other, a phenomenon known as Linkage Disequilibrium (LD), which was discussed in section 2.1.3.

Finally, the genetic ancestry variable (A) accounts for the fact that our Causal Model only represents a single generation. Genetic ancestry typically captures both dynastic effects as well as the evolutionary pressure that shaped the parental genotypes across generations. One element that is not captured by the model is the direct dynastic effect of grandparents on offsprings' phenotype (direct arrow between A and Y). This effect could be represented if the model explicitly represented this third generation. Here we assume this direct effect to be negligible. We also note that it has little justification beyond the third generation since great-grandparents are unlikely to contribute directly to the offspring's education.

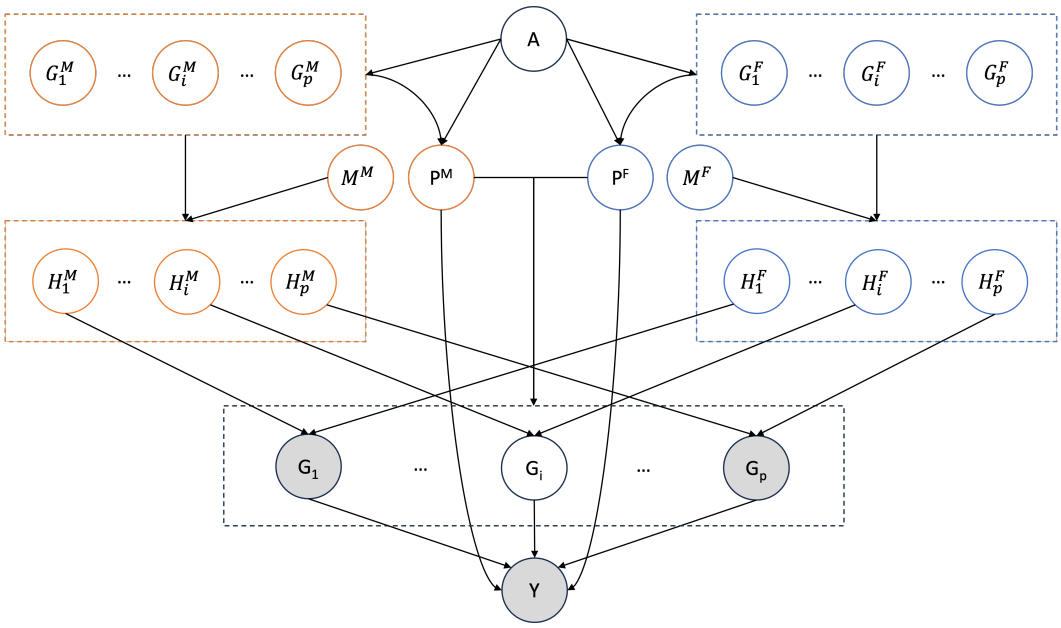

Figure 3.2: The proposed Causal Model of Genetic Inheritance. Filled nodes represent observed variables in the UK Biobank while transparent nodes are unobserved. Because we are relying on genotyping data, not all genetic variations are observed. Unlike trio studies, the parental genotypes are also unobserved. The model captures how genetic variations and traits are inherited from a generation to the next. Linkage disequilibrium, dynastic effects and genetic ancestry confound genetic association studies.

3.2.2 Non-Identifiability

Unfortunately, the genetic effects we seek to estimate are not identifiable via backdoor-adjustment within the model shown in figure 3.2 and the available UK Biobank data. A closer inspection of the graph reveals two main backdoor paths that cannot be adjusted for because they contain unobserved variables (transparent nodes).

The first backdoor path corresponds to dynastic effects, whereby an individual's trait is directly influenced by their parent's phenotypes. While family-based studies can mitigate bias from these confounding factors [18], the UK Biobank lacks extensive parental data, limiting our ability to adjust for these effects.

The second backdoor path arises from meiosis and the resulting genetic dependence across the genome. Instead of being independently inherited, genetic variations are inherited in approximately independent blocks, and variations within

a block are highly correlated [13]. These blocks can span up to several mega bases, thereby obfuscating the causal variants from their linked counterparts. In theory, variants linked to the variants under investigation make up a valid adjustment set for backdoor adjustment. However, in practice, this is complicated by two problems. First, since we only have access to genotyping data, many variations are unobserved and the adjustment set is incomplete. The advent of whole genome sequencing data can alleviate this complication but another problem remains. If the variants are perfectly, or almost perfectly correlated, backdoor adjustment will lead to violation of the positivity hypothesis (see section 2.2.2). Precisely, let $V_c = v_c$ be a candidate variant and genotype of interest, and $\mathbf{V}_b$ be a set of linked variants in the block. For positivity to hold, it must be true that $p(V_c = v_c|\mathbf{V}_b = \mathbf{v}_b) > 0$, for all combinations $\mathbf{v}_b$. This constraint is almost impossible to achieve given the size of linkage disequilibrium blocks and the high degree of correlation within these blocks.

In general, the identification of causal variants through formal identification strategies remains a challenging and open research direction. Instead, the field of population genetics has relied on heuristics that we now turn to.

3.2.3 Approximate Identifiability Via Heuristics

3.2.3.1 Confounders as Latent Variables

We have established that under the causal model depicted in Figure 3.2, the UK Biobank data (and similar bio banks) does not allow for the direct identification of causal genetic effects. This creates an inevitable gap between the causal effect we aim to estimate and the statistical quantity we can actually measure (equation 3.1). However, we have also argued that adjusting for genetic variants linked to those of interest can effectively block most backdoor paths, with the exception of dynastic effects. This concept is central to the adjustments proposed thus far to address confounding.

Rather than considering a high-dimensional set of linked variants—or potentially the entire genome—it may be sufficient to focus on a low-dimensional representation thereof. In population genetics, this is traditionally achieved using Principal Component Analysis (PCA)[1, 122]. Figure 3.3 illustrates that the first six principal components do indeed convey information about the ethnicity of white individuals in the UK Biobank.

Recent research suggests that this approach to representation learning is theoretically sound [167]. The authors argue that the high dimensionality of multiple causes (in this case, genetic variants) is not a curse but a blessing, as it provides additional information that can help disentangle the effects of individual causes. However, this method relies on the latent factor model accurately approximating $p(\mathbf{V})$, and thus providing an accurate causal abstraction. While Hidden Markov Models have proven effective for this purpose [67], it remains unclear whether PCA would offer the same advantages.

As discussed in Section 2.3, KnockoffGWAS also employs Hidden Markov Models to model $p(\mathbf{V})$ and effectively control the false discovery rate [136]. In this thesis, we do not seek to improve upon the representation learning method and instead rely on PCA throughout. Nonetheless, it is worth noting that such improvements would naturally fit within the proposed framework and represent a promising direction for future research.

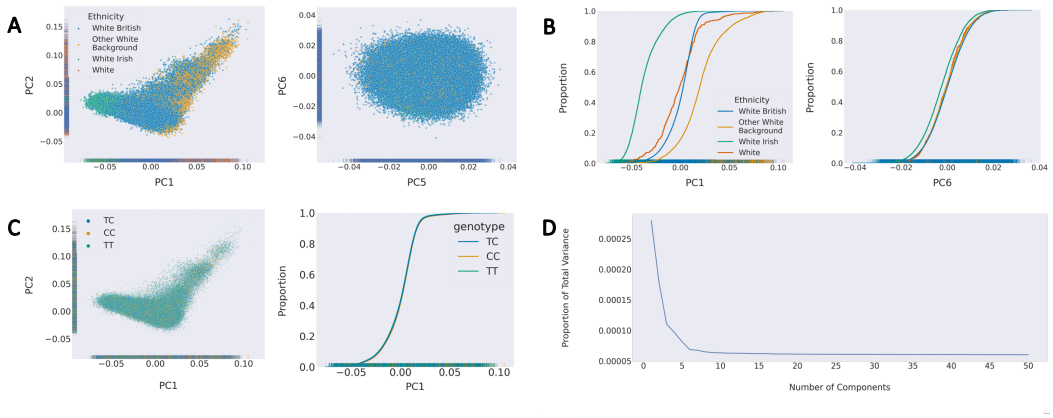

Figure 3.3: **Principal Component Analysis of the UK Biobank's white population** $(\mathbf{A})$ Principal component analysis labelled by ethnicity. Left: PC1 vs PC2 shows high level of population structure dependent on self-reported ethnicity. Right: PC5 vs PC6 shows a more symmetric shape suggesting that there is no ethnicity structure for PCs > 6. This is more clearly visible in $(\mathbf{B})$ via the cumulative distribution analysis of ethnicity for PC1 and PC6. Left: The cumulative distributions of PC1 conditioned on self-reported ethnicity differ, indicating that variation in ethnicity and variation in PC1 are dependent. Right: In PC6 this separation has disappeared. $(\mathbf{C})$ A variant specific analysis showing that this variant is not stratified in the population (the variant is rs1421085, see chapter 6). When this is the case, principal components are not confounding the genotype-phenotype relationship. $(\mathbf{D})$ This scree plot shows that the proportion of variance explained by each additional PC plateaus after 6 PCs, when subset on 'self-reported White' UK Biobank population, indicating that 6 PCs is sufficient to explain the population structure of this cohort.

3.2.3.2 Marginal Positivity Constraint

The final issue that needs to be addressed for the identification of causal estimands is that of positivity. As discussed in section 2.2.2, we will here as well rely on a heuristic, which we term the marginal positivity constraint.

Definition 3.2.1 (Marginal Positivity Constraint)**.** *Let Ψ be a causal quantity of interest involving a set of treatment variables $\boldsymbol{T}$ and associated treatment levels $\boldsymbol{t}_1, ...\boldsymbol{t}_l$. Ψ satisfies the marginal positivity constraint at the $\boldsymbol{\varepsilon}$ level if*

$$\forall i \in \{1, ..., l\}\, p(\boldsymbol{t}_i) > \boldsymbol{\varepsilon} \tag{3.2}$$

That is, instead of imposing a constraint on the original $p(\mathbf{T} = \mathbf{t}|\mathbf{w})$ for all $\mathbf{w} \in \mathcal{W}$, we enforce a constraint on the marginal $p(\mathbf{T} = \mathbf{t})$ for all genotypes defining

the estimand. Note that, if $\mathbf{T}$ is independent of $\mathbf{W}$, these quantities are the same. The benefit of the approximation is that unlike $p(\mathbf{T}=\mathbf{t}|\mathbf{w})$, the marginal positivity constraint can be estimated using the empirical distribution only (by counting). In contrast, $p(\mathbf{T}=\mathbf{t}|\mathbf{w})$ requires modelling assumptions due to the fact that principal components are continuous variables. The marginal positivity constraint is thus easier to obtain and independent of modelling assumptions and limitations. To illustrate, consider the following frequency table for two genetic variants rs10132320 and rs974766, and the marginal positivity level $\varepsilon = 0.005$.

		rs974766		
		AA	AC	CC
rs10132320	AA	0.84	0.15	0.006
	AG	0.006	0.003	0.0004
	GG	0.0007	0.0004	0.0001

Figure 3.4: An example of frequency table for two genetic variants for which 5 genotypes strata have a frequency lower than 0.005 (orange). All estimands involving these strata would fail to satisfy the marginal positivity constraint at the $\varepsilon = 0.005$ level.

Then, as an example, the causal effect

$$Y_{(AA,AA)} = p\big(Y|\text{do(rs10132320=AA), do(rs974766=AA)}\big) \tag{3.3}$$

satisfies the marginal positivity constraint but the causal effect $Y_{(AG,AC)}$ does not. Similarly the causal effect $Y_{(AA,AC)} - Y_{(AA,CC)}$ does satisfy the marginal positivity constraint but $Y_{(AA,AC)} - Y_{(GG,AA)}$ does not. As will be discussed shortly, most causal effects of interest in this thesis will be joint effects involving multiple components. For instance, $\boldsymbol{\Psi} = \big(Y_{(AA,AA)}, Y_{(AG,AA)}, Y_{(GG,AA)}\big)$, has 3 components. Rather than dropping $\boldsymbol{\Psi}$ entirely, it is enough to filter the components of $\boldsymbol{\Psi}$ that do not satisfy the positivity constraint, here $Y_{(GG,AA)}$.

Definition 3.2.2 (ε-constrained Joint Estimand)**.** *Let $\boldsymbol{\Psi} = (\Psi_1, ..., \Psi_k)$ be a joint estimand such that each Ψ_i involves a set of treatment variables $\boldsymbol{T}_i$ and associated treatment levels $\boldsymbol{t}_{i,1}, ..., \boldsymbol{t}_{i,l}$. The ε-constrained joint estimand of $\boldsymbol{\Psi}$ is*

$$\boldsymbol{\Psi}_\varepsilon = (\Psi_i : i \in \{1, ..., k\}, \Psi_i \textit{ satisfies the marginal positivity constraint at the } \varepsilon \textit{ level}) \tag{3.4}$$

In the previous example this yields $\Psi_{\boldsymbol{\varepsilon}} = \left(Y_{(AA,AA)}, Y_{(AG,AA)}\right)$. In this thesis, only $\boldsymbol{\varepsilon}$-constrained joint estimands are considered and a practical value for $\boldsymbol{\varepsilon}$ is inferred from simulations in chapter 4.

3.2.3.3 The Working Causal Structural Model

We have now discussed the main assumptions for identifiability, that is the absence of unobserved confounders and positivity. In this section we describe the causal model that is used in the rest of this thesis and under which causal genetic effects are identified. Since this model is only a practical approximation of model 3.2, we call it the working causal model.

Definition 3.2.3 (The Working Causal Structural Model)**.** *The Working Structural Model over the observed data unit* $\boldsymbol{O} = (\boldsymbol{W}, \boldsymbol{C}, \boldsymbol{V}, Y)$ *is given by the following set of structural equations:*

$$\begin{aligned} \boldsymbol{W} &= f_{\boldsymbol{W}}(U_{\boldsymbol{W}}) \\ \boldsymbol{C} &= f_{\boldsymbol{C}}(U_{\boldsymbol{C}}) \\ \boldsymbol{V} &= f_{\boldsymbol{V}}(\boldsymbol{W}, U_{\boldsymbol{V}}) \\ Y &= f_Y(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{C}, U_Y) \end{aligned} \tag{3.5}$$

Where:

- $\boldsymbol{W}$ *are the principal components obtained via PCA as per section 3.2.3.1.*
- $\boldsymbol{C} = \{$*Age when attended assessment centre, Genetic sex*$\}$ *corresponding to the UK Biobank Data-Field 21003 and Data-Field 22001 respectively.*
- $\boldsymbol{V}$ *is a set of genetic variants, directly genotyped or imputed, as explained in section 3.1.1.*
- Y *is any phenotype defined in section 3.1.2.*

A graphical representation of this model is presented in figure 3.5. Note that the set $\mathbf{V}$ of causal variants is unknown in practice and could be comprised of the whole genome if the omnigenic model is correct [16]. The goal of this work is to rigorously define and quantify various effects for subsets of $\mathbf{V}$. These subsets are typically small and comprise 1, 2 or 3 variants. Larger subsets would likely fail to satisfy the marginal positivity constraint. In this chapter we do not make precise

how these subsets are selected and focus on the formal definition and estimation of genetic effects. A general strategy to determine high-quality putative causal variants is presented in Chapter 7.

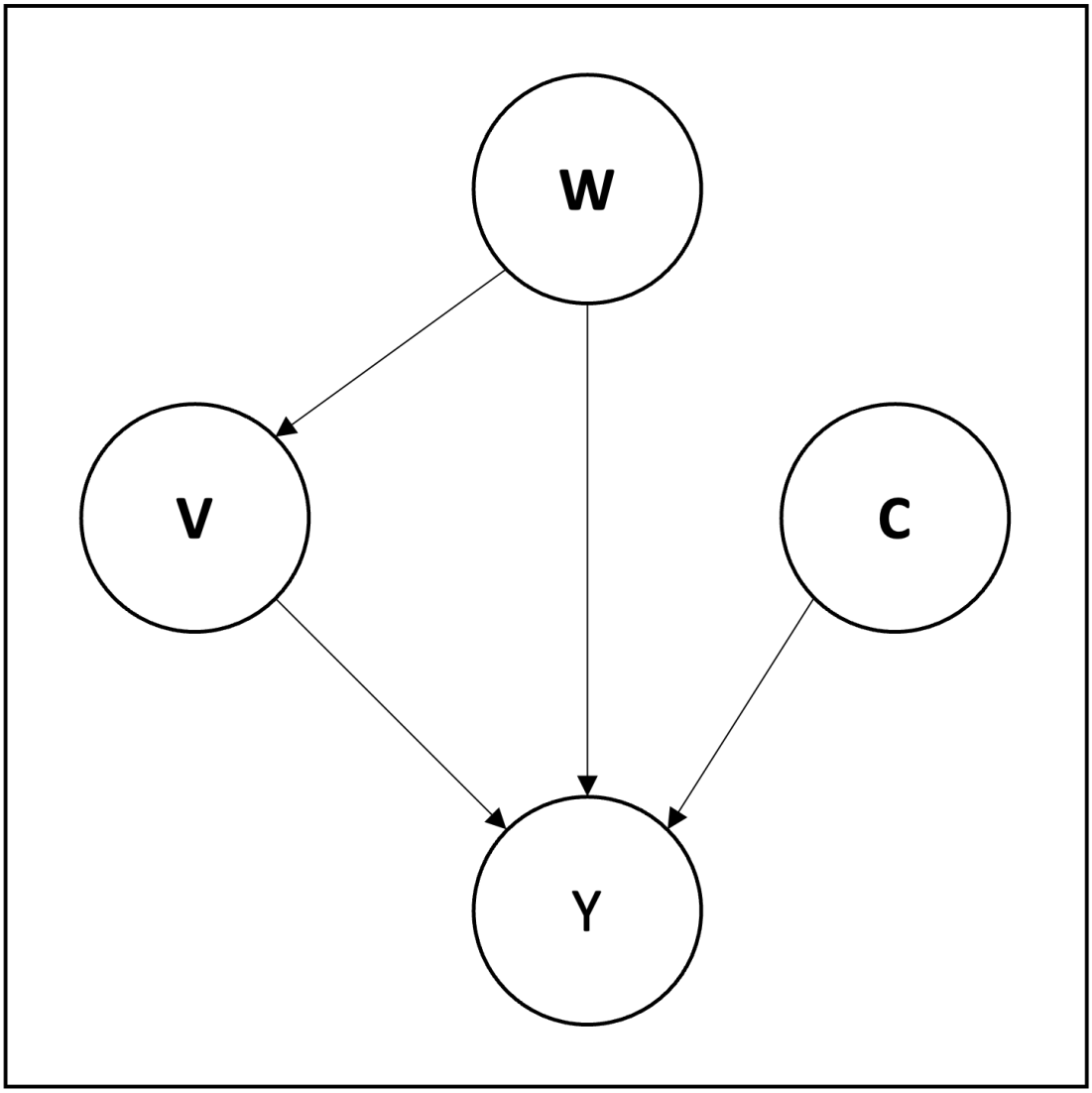

Figure 3.5: The Working Causal Model.

As per section 2.2.1, we further equip our causal model with an unknown ground truth probability distribution, or data generating process, which will be denoted by $P_0 \in \mathcal{M}$. Our lack of knowledge about P_0 is reflected in $\mathcal{M}$ which is simply the fully non-parametric model containing all distributions. For simplicity of exposure, we will assume that P_0 is absolutely continuous with respect to the Lebesgue measure μ, and denote its density by p_0. Similarly, for an arbitrary distribution P, its density will be denoted by p. Unlike the causal model 3.2, the working causal model does not contain any unobserved confounding variable. In

this model, the interventional conditional distribution $p\big(Y|do(\mathbf{V})\big)$ is identifiable via the backdoor adjustment theorem 2.2.1. This identifiability result is key since all genetic effects will be defined as functions of this interventional conditional distribution and will consequently be identifiable as well.

Finally, we assume that we observe not one, but n independent and identically distributed units such that $(\mathbf{O}_1, ..., \mathbf{O}_n) \sim P_0^n$. We also remind the reader that the realization of such a random variable is denoted by $\mathbf{o}_i$, or more simply by $\mathbf{o}$. This assumption will be relaxed in section 3.6.

3.3 Genetic Effects

We now turn to the definition of the causal effects that we study in this thesis. As discussed in section 2.4, these estimands are defined as functionals, that is functions of a probability distribution, $P \mapsto \Psi(P) \in \mathbb{R}^d$. As previously discussed, all these causal estimands are identifiable via backdoor adjustment and their statistical counterparts are presented as well. A variant, or set of variants is denoted by $\mathbf{V}$ while a specific realisation (or genotype) is denoted by $\mathbf{v}$. However, contrary to section 3.2.3.3, the set $\mathbf{V}$ is typically very small here, e.g. $\mathbf{card}(\mathbf{V}) = 1, 2, 3, ...$

3.3.1 The Mean Under intervention

The mean under intervention for an outcome Y and genotype $\mathbf{v}$, is not of special interest in itself, but is the building block for all estimands presented below. It is the mean value Y would have had, had the genotype of all individuals been set to $\mathbf{v}$ at birth. For binary outcomes like diseases, this is the probability of the outcome under this hypothetical intervention. It is defined by

$$\begin{aligned} \mathrm{MI}_{\mathbf{v}} &= \mathbb{E}[Y(\mathbf{v})] \\ &= \mathbb{E}[\mathbb{E}[Y|\mathbf{V} = \mathbf{v}, \mathbf{W} = \mathbf{w}]] \end{aligned} \tag{3.6}$$

where the second equality arises from the identification result. For a set of variants $\mathbf{V}$, there are multiple possible genotypes $\mathbf{v}$ an individual could have had. For G potential genotypes, the genotype-independent mean under intervention is

defined as the vector of means under interventions for each genotype:

$$\mathbf{MI} = \begin{bmatrix} \mathrm{MI}_{v_1} \\ \dots \\ \mathrm{MI}_{v_G} \end{bmatrix}. \tag{3.7}$$

Note that by definition of the genotype-independent mean under intervention, the estimand naturally captures dominance, which is of interest in genetic studies [74].

3.3.2 The Average Treatment Effect

The mean under intervention does not reveal whether a genotype is detrimental or beneficial for a specific outcome. To answer this question, we need to contrast two genotypes. This is the purpose of the Average Treatment Effect (ATE). For a set of variants $\mathbf{V}$, two genotypes $\mathbf{v}$ and $\mathbf{v'}$, and an outcome Y, it is defined as the difference in means under interventions

$$\begin{aligned} \mathrm{ATE}_{v \to v'} &= \mathrm{MI}_{\mathbf{v'}} - \mathrm{MI}_{\mathbf{v}} \\ &= \mathbb{E}\big[\mathbb{E}[Y|\mathbf{V} = \mathbf{v'}, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y|\mathbf{V} = \mathbf{v}, \mathbf{W} = \mathbf{w}]\big]. \end{aligned} \tag{3.8}$$

Similarly to the mean under intervention, many Average Treatment Effects can be defined from all potential genotypes. However these effects are not all linearly-independent, for example, it is easy to see that $\mathrm{ATE}_{v \to v'} = -\mathrm{ATE}_{v' \to v}$. Naively estimating all possible Average Treatment Effects has two drawbacks. First it is a waste of computational resources, there is no need to run an expensive estimation procedure to obtain $\mathrm{ATE}_{v' \to v}$ when it can be simply obtained by a sign flip. Second, because of this linear-dependence structure, the covariance matrix of estimates will not be full rank. This causes numerical instabilities when performing joint hypothesis testing or applying the delta-method on these multi-dimensional estimates (section 3.5.2).

To avoid these complications, we define the (genotype-independent) Average Treatment Effect as the Cartesian Product of single treatment ordered transitions. For illustration, let $\mathbf{V} = (V_1, V_2)$, with both variants having potential values $\{0, 1, 2\}$. The single treatment ordered transitions are $\{0 \to 1, 1 \to 2\}$ for both variants. Then, the (genotype-independent) Average Treatment Effect is given

by

$$\mathbf{ATE} = \begin{bmatrix} \text{ATE}_{(0,0)\to(1,1)} \\ \text{ATE}_{(0,1)\to(1,2)} \\ \text{ATE}_{(1,0)\to(2,1)} \\ \text{ATE}_{(1,1)\to(2,2)} \end{bmatrix} \tag{3.9}$$

We say that the **ATE** is non-zero if any of its component is non-zero.

3.3.3 Allelic Effect Difference

Bi-allelic variants represent a large proportion of known genetic variants, they are loci for which only two alleles have been observed in the population, say M and m for the major (most frequent) and minor alleles respectively. Because humans are diploid, this results in 3 different genotypes at each locus: $\{\text{MM}, \text{Mm}, \text{mm}\}$. Linear models typically assume that the effect of a variant on trait is a linear function of the number of minor alleles. In other words, they assume that $\text{ATE}_{\text{MM}\to\text{Mm}}$ and $\text{ATE}_{\text{Mm}\to\text{mm}}$ are equal. However there is no reason why this should be the case. To answer this question, we define the Allelic Effect Difference (AED) as a difference in Average Treatment Effects.

$$\text{AED} = \text{ATE}_{\text{Mm}\to\text{mm}} - \text{ATE}_{\text{MM}\to\text{Mm}} \tag{3.10}$$

There is a difference in allelic effects if the AED is different from 0.

3.3.4 Interactions: Pairwise and Higher-Order

For two genetic variants V_1 and V_2, an additive interaction effect measures the extent to which the joint effect of V_1 and V_2 exceeds the effect of each considered individually. Without loss of generality, let $(0, 1)$ be two possible genotypes for both V_1 and V_2. Implicitly, 0 represents the reference genotype or "control" value, and 1 the alternate genotype or "treatment" value. The Average Interaction Effect (AIE) is equivalently defined defined by any of the following:

$$\begin{aligned} \text{AIE}_{(0,0)\to(1,1)} &= \text{ATE}_{(0,0)\to(1,1)} - (\text{ATE}_{(0,0)\to(1,0)} + \text{ATE}_{(0,0)\to(0,1)}) \\ &= \text{MI}_{(1,1)} - \text{MI}_{(0,1)} - \text{MI}_{(1,0)} + \text{MI}_{(0,0)} \end{aligned} \tag{3.11}$$

Where the first equation makes it explicit that it is to be interpreted as a difference in effects and the second equation is used in practice.

This definition can also be extended to higher-orders [10]. Let $\mathbf{V} = (V_1, ..., V_n)$ be a set of n variants and two fully disjoint genotypes represented by $\mathbf{0}_n = (0, ..., 0)$ and $\mathbf{1}_n = (1, ..., 1)$. An arbitrary genotype is denoted by $\mathbf{v}$, for example $\mathbf{v} = (0, 1, 0, 1)$ is a valid 4-points genotype. Finally, let $\mathcal{K}_j = \{\mathbf{v} : \sum_{i=1}^{n} \mathbb{1}(\mathbf{v}_i = 1) = j\}$, the set of genotypes with j alternate values. The n-points additive interaction effect is then defined by:

$$AIE^{(n)} = \sum_{j=0}^{n} (-1)^{n-j} \sum_{v \in \mathcal{K}_j} MI_v \tag{3.12}$$

Let's illustrate with a 3-points interaction, grouping terms exactly as they appear in the sum:

$$\begin{aligned} AIE^{(3)} = & MI_{1,1,1} \\ & - (MI_{0,1,1} + MI_{1,0,1} + MI_{1,1,0}) \\ & + (MI_{0,0,1} + MI_{1,0,0} + MI_{0,1,0}) \\ & - MI_{0,0,0} \end{aligned} \tag{3.13}$$

Finally, just as for the **ATE**, because there are more than two fully disjoint genotypes, we can define a (genotype-independent) **AIE**. It is defined using the exact same strategy. Here is the counterpart example with two genetic variants

$$\mathbf{AIE} = \begin{bmatrix} \text{AIE}_{(0,0)\to(1,1)} \\ \text{AIE}_{(0,1)\to(1,2)} \\ \text{AIE}_{(1,0)\to(2,1)} \\ \text{AIE}_{(1,1)\to(2,2)} \end{bmatrix} \tag{3.14}$$

We say that the **AIE** is non-zero if any of its component is non-zero.

3.3.5 Smooth Functionals And Their Gradients

As will be become clear in the next section, in order to build semi-parametric asymptotically unbiased and efficient estimators, the estimands Ψ need to satisfy some form of smoothness. In other words, moving away from $P \in \mathcal{M}$ in any given direction should not affect the value of the estimand $\Psi(P)$ too much. Before we can discuss any notion of smoothness for Ψ we need to define what movements are acceptable in $\mathcal{M}$. These acceptable movements are defined as differentiable paths.

Definition 3.3.1 (Differentiable Path through P)**.** *A differentiable path, or simply a path, through P is a subset $\{P_\varepsilon\}_{\varepsilon\in[0,+\infty[} \subset \mathcal{M}$ such that $P_{\varepsilon=0} = P$, and there exists a measurable map h such that:*

$$\lim_{\varepsilon\to 0+} \int \Big[\frac{\sqrt{p_\varepsilon} - \sqrt{p}}{\varepsilon} - \frac{1}{2}h\sqrt{p}\Big]^2 \mathrm{d}\mu = 0 \tag{3.15}$$

The latter technical condition is known as differentiability in quadratic mean and the function h is called the score of the path. It is a fact that $h \in \mathcal{L}_0^2(P)$ and, if $\varepsilon \mapsto p_\varepsilon$ is differentiable, we also have for a unit data point o

$$h : o \mapsto \frac{\mathrm{d}}{\mathrm{d}\varepsilon}_{|\varepsilon=0} \log p_\varepsilon(o) \tag{3.16}$$

For example, for any $h \in \mathcal{L}_0^2(P)$, the following equation defines a valid (differentiable) path through P.

$$p_\varepsilon = (1 + h\varepsilon)p \tag{3.17}$$

This is because, since $h \in \mathcal{L}_0^2(P)$, $\int hp\mathrm{d}\mu = \int h dP = 0$. In general, p is orthogonal to all scores of paths through it, which motivates the following definition. The tangent space of the model $\mathcal{M}$ at P, denoted by $\mathrm{T}_{\mathcal{M},P}$, is the closure of the linear span of all scores of paths through P. Since we are working in the full non-parametric model, it can be shown that the tangent space is in fact equal to the entire $\mathcal{L}_0^2(P)$ [159].

We are now ready to define the required version of smoothness for our estimands, called pathwise differentiability.

Definition 3.3.2 (Pathwise Differentiability)**.** $\Psi : \mathcal{M} \to \mathbb{R}^d$ *is pathwise differentiable at P if there exists a continuous linear map $\mathrm{d}\Psi_P : L_0^2(P) \to \mathbb{R}^d$, such that for every $h \in \mathcal{L}_0^2(P)$ and submodel $\{P_\varepsilon\}_{\varepsilon\in[0,+\infty[}$, with score function h*

$$\mathrm{d}\Psi_P(h) = \frac{\mathrm{d}\Psi(P_\varepsilon)}{\mathrm{d}\varepsilon}|_{\varepsilon=0} = \lim_{\varepsilon\to 0+} \frac{\Psi(P_\varepsilon) - \Psi(P)}{\varepsilon} \tag{3.18}$$

Since $L_0^2(P)$ is a Hilbert space, by the Riesz representation theorem, there exist a unique function $D_{\Psi,P} \in \mathcal{L}_0^2(P)$, called the gradient of Ψ such that

$$\mathrm{d}\Psi(h) = \langle D_{\Psi,P}, h\rangle_P = \mathbb{E}_P[D_{\Psi,P}h] = \int D_{\Psi,P}h\mathrm{d}P \tag{3.19}$$

The gradient is a key ingredient to the construction of the semi-parametric estimators we are about to consider. To see why, consider a one-dimensional path

$\{P_\varepsilon\}_{\varepsilon\in[0,+\infty[} \subset \mathcal{M}$. In this submodel, the Cramér-Rao lower bound on the variance of any estimator is given by

$$\frac{(\frac{\mathrm{d}\Psi(p_\varepsilon)}{\mathrm{d}\varepsilon})^2|_{\varepsilon=0}}{\mathbb{E}_P[h^2]} = \frac{\langle D_{\Psi,P}, h\rangle_P}{\langle h,h\rangle_P} \tag{3.20}$$

The Cramér-Rao lower bound in the non-parametric model is necessarily higher than that of any submodel $\{P_\varepsilon\}_{\varepsilon\in[0,+\infty[} \subset \mathcal{M}$. It is formally defined as the supremum over all such submodels, or equivalently scores. Using the Cauchy-Schwartz inequality, we have

$$\sup_{h\in \mathrm{T}_{\mathcal{M},P}} \frac{\langle D_{\Psi,P}, h\rangle_P}{\langle h,h\rangle_P} = \mathbb{E}_P[D^2_{\Psi,P}] = Var[D_{\Psi,P}] \tag{3.21}$$

In other words, the variance of the gradient $D_{\Psi,P}$ represents the non-parametric optimal asymptotic variance for any estimator of Ψ. Estimators that achieve this variance are referred to as efficient. Note that $D_{\Psi,P}$ depends on both the estimand Ψ and the distribution P where it is evaluated. However, we will sometimes write D_Ψ when it does not matter where the gradient is evaluated.

Returning to genetic effects, since $\mathrm{d}\Psi_P$ is a linear operator, and all genetic effects are linear combinations of $MI_{\mathbf{v}}$, it is enough to derive the gradient for $MI_{\mathbf{v}}$. Fortunately, this gradient, $D_{MI_{\mathbf{v}},P_0}$, is a well studied quantity [77]. It is given by

$$D_{MI_{\mathbf{v}},\mathbb{P}}(\mathbf{V},\mathbf{W},\mathbf{C},Y) = \frac{\mathbb{1}(\mathbf{V}=\mathbf{v})}{G(\mathbf{V},\mathbf{W})}(Y - Q_Y(\mathbf{V},\mathbf{W},\mathbf{C})) + Q_Y(\mathbf{v},\mathbf{W},\mathbf{C}) - MI_{\mathbf{v}} \tag{3.22}$$

where

$$\begin{aligned} Q_Y(\mathbf{V},\mathbf{W},\mathbf{C}) &= \mathbb{E}_P[Y|\mathbf{V},\mathbf{W},\mathbf{C}] \\ G(\mathbf{V},\mathbf{W}) &= p(\mathbf{V}|\mathbf{W}) \end{aligned} \tag{3.23}$$

To illustrate how other gradients are obtained we consider the Average treatment Effect. Let $\mathbf{v}$ and $\mathbf{v}'$ be two distinct genotypes, the gradient of the Average treatment Effect is

$$\begin{aligned} D_{\mathrm{ATE}_{v\to v'},\mathbb{P}}(\mathbf{V},\mathbf{W},\mathbf{C},Y) &= (D_{MI_{\mathbf{v}'},\mathbb{P}} - D_{MI_{\mathbf{v}},\mathbb{P}})(\mathbf{V},\mathbf{W},\mathbf{C},Y) \\ &= \frac{\mathbb{1}(\mathbf{V}=\mathbf{v}') - \mathbb{1}(\mathbf{V}=\mathbf{v})}{G(\mathbf{V},\mathbf{W})}(Y - Q_Y(\mathbf{V},\mathbf{W},\mathbf{C})) \\ &\quad + (Q_Y(\mathbf{v}',\mathbf{W},\mathbf{C}) - Q_Y(\mathbf{v},\mathbf{W},\mathbf{C})) - \mathrm{ATE}_{v\to v'} \end{aligned} \tag{3.24}$$

The gradient of interactions can be obtained similarly but is omitted for readability.

3.4 Semi-Parametric Estimation

We are now nearly ready to define the estimation procedures that will be used throughout this thesis. First, we will outline the criteria for a good estimation strategy and demonstrate that it is sufficient to focus on regular and asymptotically linear estimators. Following this, we will define several such estimation strategies and discuss their respective properties in greater detail. The observed data has the structure presented in the working causal model, $\mathbf{O} = (\mathbf{O}_1, ..., \mathbf{O}_n) = (\mathbf{V}, \mathbf{W}, \mathbf{C}, Y)_{i=1...n} \in \mathbb{R}^{n \times q}$. It is generated according to a true but unknown distribution $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the non-parametric model.

3.4.1 Regular Asymptotically Linear Estimators

The goal of this section is to formalise what is meant by a good estimation strategy, and for that, we first need to define what an estimator is.

Definition 3.4.1 (Estimator)**.** *Let $\Psi_0 = \Psi(P_0) \in \mathbb{R}^d$, a statistical estimand of interest. An estimator $\hat{\Psi}$ of Ψ_0 is a measurable function $\hat{\Psi} : \mathbb{R}^{n \times q} \mapsto \mathbb{R}^d$.*

Loosely speaking, $\hat{\Psi}$ is a procedure that maps a dataset to an estimate of our estimand of interest. Since the dataset is a random variable, so is the estimator. For example, one can think of the UK Biobank as a random draw from the British population. A different draw would result in a different estimate. It will be convenient to see estimators as functions of the empirical distribution rather than the data itself. The empirical distribution, or empirical measure, is the distribution that assigns equal probability to each observation in a given sample. It is formally defined as

$$\mathbb{P}_n : A \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(\mathbf{O}_i) \tag{3.25}$$

where A is any measurable set, and $\mathbb{1}_A(\mathbf{O}_i)$ is an indicator function that equals 1 if $\mathbf{O}_i \in A$ and 0 otherwise. This is not restrictive since by the i.i.d assumption, there are no duplicates in the dataset and the empirical distribution entirely captures the dataset. We will thus use the notation $\hat{\Psi}_n = \hat{\Psi}(\mathbb{P}_n) = \hat{\Psi}(\mathbf{O}_1, ..., \mathbf{O}_n)$.

Definition 3.4.1 is not constructive, but it is clear that we would like to design procedures with good properties. An ideal, but unrealistic procedure would always point to the true estimand's value regardless of the input dataset. More pragmatically, we would like to know whether our estimator is accurate and precise. These concepts are captured by the notions of bias and variance respectively.

Definition 3.4.2 (Bias)**.** *The bias of an estimator $\hat{\Psi}_n$ with respect to an estimand Ψ_0, is the mean error of this estimator under P_0, it is defined as :* $B(\hat{\Psi}_n) = \mathbb{E}_{P_0}[\hat{\Psi}_n] - \Psi_0$.

Definition 3.4.3 (Variance)**.** *Since an estimator is a random variable, its variance is simply given by:* $\sigma^2 = Var[\hat{\Psi}_n]$.

Ideally, we are interested in unbiased estimators with minimum variance (efficient). However, in complex and realistic semi-parametric models, this is often an unattainable goal. Despite this challenge, it is possible to develop estimators that are asymptotically unbiased and efficient. To achieve this, we will focus on a specific class of estimators: those that are asymptotically normal

Definition 3.4.4. *Let $\Psi(P_0) \in \mathbb{R}$ be a parameter of the data generating process P_0. An estimator $\hat{\Psi}_n$ of $\Psi(P_0)$ is* asymptotically linear *if there exists a measurable random function $\phi_{P_0} \in \mathcal{L}_0^2(P_0)$ of the unit data structure such that:*

$$\sqrt{n}\big(\hat{\Psi}_n - \Psi(P_0)\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{P_0}(\mathbf{O}_i) + o_{P_0}(1). \tag{3.26}$$

Where $o_{P_0}(1)$ is a term that converges to zero in probability as $n \to \infty$.

In other words, the asymptotic difference between the estimator and the true estimand's value behaves as an average of i.i.d variables. The function ϕ_{P_0} is called the influence function of the estimator. It measures how much each observation contributes to the variance of the estimator. The benefit of restricting the scope to asymptotically linear estimators is that by the central limit theorem

$$\sqrt{n}(\hat{\Psi}_n - \Psi_0) \rightsquigarrow \mathcal{N}(0, Var[\phi_{P_0}(\mathbf{O})]) \tag{3.27}$$

This result will be useful to derive asymptotic Wald-type confidence intervals and perform hypothesis testing as will be discussed in section 3.5.

Are asymptotically linear optimal? Since $\phi_{P_0} \in \mathcal{L}_0^2(P_0)$, $\mathbb{E}[\phi_{P_0}(\mathbf{O})] = 0$, hence they are already asymptotically unbiased but are they efficient? The Hájek-Le Cam convolution theorem states that the most efficient regular estimator is guaranteed to be asymptotically linear [159]. Regularity is a desirable technical condition which ensures that an infinitesimal perturbation to P_0 does not change the asymptotic distribution of the estimator. Irregular estimators are pathological and may surpass a regular estimator at most on a set of Lebesgue measure zero [159]. For instance, a constant estimator ignoring the data has zero variance

and is thus super efficient, but is only correct for a single estimand's value. The Hodges' estimator is another famous irregular estimator [159]. For completeness, we provide the definition of a regular estimator below.

Definition 3.4.5 (Regular Estimator)**.** *An estimator $\hat{\Psi}$ is regular at P_0 for estimating $\Psi(P_0)$ if there exist a probability distribution $\mathcal{D}$ such that for all score functions $h \in \mathrm{T}_{\mathcal{M},P_0}$ and all paths $\{P_{\varepsilon_n=\frac{1}{\sqrt{n}}}\}_{n\in\mathbb{N}}$ with score h*

$$\sqrt{n}(\hat{\Psi}(\mathbb{P}_{\varepsilon_n=\frac{1}{\sqrt{n}}}) - \Psi(P_{\varepsilon_n=\frac{1}{\sqrt{n}}})) \rightsquigarrow \mathcal{D} \tag{3.28}$$

where $\mathbb{P}_{\varepsilon_n=\frac{1}{\sqrt{n}}}$ is the empirical distribution of $P_{\varepsilon_n=\frac{1}{\sqrt{n}}}$.

We will thus only consider regular asymptotically linear estimators whose limiting distribution is thus $\mathcal{D} = \mathcal{N}(0, Var[\phi_{P_0}(\mathbf{O})])$. It is thus clear that we are trying to build regular asymptotically linear estimators with influence curve having minimum variance. We have also seen in the previous section that the gradient of Ψ, D_{Ψ,P_0} attains the non-parametric efficiency bound. If we can construct estimators whose influence curves match this gradient, these estimators will be efficient. In this context, the influence curve is referred to as the efficient influence curve. The construction of such estimators is the focus of the next section.

3.4.2 Plugin Estimators

By defining an estimand as a functional map, we have also suggested a general estimation strategy. Let $\hat{P}$ be an estimator for P_0, then $\Psi(\hat{P})$ is called a plugin estimator for $\Psi_0 = \Psi(P_0)$. Plugin estimators are intuitive, but will usually be biased. Since the estimands of interest are pathwise differentiable, they admit a von-Mises expansion [45]:

$$\Psi(\hat{P}) - \Psi(P_0) = (\hat{P} - P_0) D_{\Psi,\hat{P}} + R_2(\hat{P}, P_0) \tag{3.29}$$

where D_Ψ is the gradient of Ψ, and $R_2(\hat{P}, P_0)$ is called the second-order remainder defined by this equation, and depends on products of differences of P_0 and $\hat{P}$. We can add and subtract $\mathbb{P}_n D_{\Psi,P_0}$ and $P_0 D_{\Psi,P_0}$ and reorganise to reach the following seemingly more complex expression:

$$\Psi(\hat{P}) - \Psi(P_0) = \mathbb{P}_n D_{\Psi,P_0} \underbrace{-\mathbb{P}_n D_{\Psi,\hat{P}}}_{\text{Plug-In Bias}} + \underbrace{(\mathbb{P}_n - P_0)(D_{\Psi,\hat{P}} - D_{\Psi,P_0})}_{\text{Empirical Process}} + \underbrace{R_2(\hat{P}, P_0)}_{\text{Second-Order}} \tag{3.30}$$

Careful inspection of equation 3.30, reveals a strategy to build regular asymptotically linear estimators. This is because the first term on the right hand side (not under-braced) is exactly what is needed to reach asymptotic linearity. In other words, if all terms under-braced are $o_{P_0}(1)$, $\Psi(\hat{P})$ is asymptotically linear with influence curve the gradient and is thus asymptotically unbiased and efficient. The following sections illustrate how this requirement can be achieved under mild conditions.

In practice, Ψ depends on the probability distribution P where it is evaluated only through certain components which are traditionally denoted by Q. In the case of $MI_{\mathbf{v}} = \mathbb{E}[\mathbb{E}[Y|\mathbf{V}=\mathbf{v},\mathbf{W},\mathbf{C}]]$, which we consider next, $Q = (Q_Y, Q_W, Q_C)$ where

$$\begin{aligned} Q_Y(\mathbf{V},\mathbf{W},\mathbf{C}) &= \mathbb{E}[Y|\mathbf{V},\mathbf{W},\mathbf{C}] \\ Q_W(\mathbf{W}) &= p(\mathbf{W}) \\ Q_C(\mathbf{C}) &= p(\mathbf{C}) \end{aligned} \tag{3.31}$$

The dependence on these factors can be made explicit

$$\begin{aligned} MI_{\mathbf{v}} &= \mathbb{E}[\mathbb{E}[Y|\mathbf{V}=\mathbf{v},\mathbf{W},\mathbf{C}]] \\ &= \int \mathbb{E}[Y|\mathbf{V}=\mathbf{v},\mathbf{w},\mathbf{c}]p(\mathbf{w})p(\mathbf{c})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{c} \\ &= \int Q_Y(\mathbf{v},\mathbf{w},\mathbf{c})Q_W(\mathbf{w})Q_C(\mathbf{c})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{c}. \end{aligned} \tag{3.32}$$

These functions are often called nuisance functions because they are not of direct interest, but need to be estimated to obtain $\hat{\Psi} = \Psi(\hat{Q}) = \Psi(\hat{P})$, where $\hat{P}$ is any distribution compatible with $\hat{Q}$. Furthermore, the gradient D_Ψ typically requires estimation of further nuisance functions. In the case of $MI_{\mathbf{v}}$, there is only one such function called the propensity score and defined by

$$G(\mathbf{V},\mathbf{W}) = P(\mathbf{V}|\mathbf{W}). \tag{3.33}$$

Thus, we will sometimes collectively refer to (Q,G) as the nuisance functions. Finally, if interest is restricted to a single genotype $\mathbf{v}$, the functions Q_Y and G can be simplified to $Q_{Y,\mathbf{v}}(\mathbf{W},\mathbf{C}) = Q_Y(\mathbf{v},\mathbf{W},\mathbf{C})$ and $G_{\mathbf{v}}(\mathbf{W}) = G(\mathbf{v},\mathbf{W})$ respectively. This scenario might arise in targeted studies designed to reduce the multiple testing burden, although it is not the primary focus of this thesis. Moreover, most software estimate the entire functions (Q_Y, G) based on the available data, rendering this distinction less relevant.

We now explain how each term in equation 3.30 can be controlled.

3.4.3 Plug-In Bias

The Plug-In bias $-\mathbb{P}_n D_{\Psi,\hat{P}}$ has received a particular interest because it is not possible to show that it is negligible for a general estimate $\hat{Q}$ of Q. Here we describe the two methods we use across the thesis to deal with this term.

3.4.3.1 One-Step Estimation

By definition, the Plug-In bias depends on two key elements: the empirical distribution $\mathbb{P}_n$ and the nuisance functions that the gradient D_Ψ relies on, specifically $\hat{Q}_Y$ and $\hat{G}$. The one-step estimator operates on a straightforward principle: it calculates the Plug-In bias and subtracts it from the initial estimate to correct for this bias.

Definition 3.4.6 (One-Step Estimator (OSE))**.** *Let Ψ be an estimand, and $\hat{P}$ an estimator for P_0, or equivalently $\hat{Q}$ an estimator for the relevant nuisance functions Q_0. The one-step estimator (OSE) is defined by*

$$\hat{\Psi}_{OSE} = \Psi(\hat{Q}) + \mathbb{P}_n D_{\Psi,\hat{Q}} \tag{3.34}$$

The one-step estimator (OSE) is an appealing approach due to its generality and simplicity. However, because it is not a plug-in estimator, it may sometimes produce estimates that fall outside the natural domain of the estimand Ψ. For example, if Y represents a binary trait, then $MI_{\mathbf{v}}$ should be interpreted as a probability and must be constrained between 0 and 1. The OSE, however, might yield values outside this natural range, leading to potentially invalid results. This limitation has led to the development of the estimator we consider next.

3.4.3.2 Targeted Minimum Loss-based Estimation

The general idea of the targeted minimum loss-based estimator (TMLE) is to preserve the plugin principle. Instead of correcting the Plug-In bias in the estimand's space, it aims to find a targeted $\hat{P}^*$ for which the bias is provably 0. That is, it corrects the bias term in distribution space. To that end, consider a path $\{\hat{P}_\varepsilon\}$ through $\hat{P}$, TMLE works by iteratively minimising an appropriate loss L for a cleverly chosen $\{\hat{P}_\varepsilon\}$. That is TMLE iteratively optimises

$$\underset{\varepsilon}{\operatorname{argmin}}\, \mathbb{P}_n L(\hat{P}_\varepsilon, O_i) \tag{3.35}$$

until convergence ($\varepsilon = 0$) [158].

First, because we are minimising L in a submodel that contains $\hat{P}$, each iteration is guaranteed not to deteriorate $\hat{P}$. Second, if the path is constructed to have score the gradient, convergence of the procedure indeed results in $\mathbb{P}_n D_{\Psi,\hat{P}^*} = 0$. This is because by definition, the score equals $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}|_{\varepsilon=0} \log p_\varepsilon$.

Instead of targeting the entire estimate $\hat{P}$, it is often enough to target some relevant nuisance functions. In our case, it can be shown that it is sufficient to correct Q_Y [158]. For $MI_{\mathbf{v}}$, it corresponds to optimising the squared loss or log-loss for continuous and binary outcomes Y respectively, in the following corresponding paths:

$$\begin{aligned} \hat{Q}_{Y,\varepsilon}(\mathbf{V},\mathbf{W},\mathbf{C}) &= \hat{Q}_Y(\mathbf{V},\mathbf{W},\mathbf{C}) + \varepsilon \hat{H}(\mathbf{V},\mathbf{W}) \\ \hat{Q}_{Y,\varepsilon}(\mathbf{V},\mathbf{W},\mathbf{C}) &= \frac{1}{1+e^{-(logit(\hat{Q}_Y(\mathbf{V},\mathbf{W},\mathbf{C}))+\varepsilon\hat{H}(\mathbf{V},\mathbf{W}))}} \end{aligned} \tag{3.36}$$

where $\hat{H}(\mathbf{V},\mathbf{W}) = \frac{\mathbb{1}(\mathbf{V}=\mathbf{v})}{\hat{G}(\mathbf{V},\mathbf{W})}$ is known as the clever covariate. Furthermore, in these paths it can be shown that convergence happens in only one step. In practice this can be done by fitting a generalised linear model using $\hat{H}$ as a covariate and $\hat{Q}_Y$ as an offset.

Definition 3.4.7 (Targeted Minimum Loss-based Estimator (TMLE))**.** *Let Ψ be an estimand, and $\hat{P}$ estimator for P_0, or equivalently $\hat{Q}$ an estimator for the relevant nuisance functions Q_0. Let ε^* be the solution to the optimisation problem of equation 3.35, with paths as defined by equation 3.36. The targeted minimum loss-based estimator (TMLE) is defined by*

$$\hat{\Psi}_{TMLE} = \Psi(\hat{Q}_{Y,\varepsilon^*}, \hat{Q}_W, \hat{Q}_C) \tag{3.37}$$

For the other estimands considered in this thesis and presented in section 3.3, it is enough to adjust the clever covariate $\hat{H}(\mathbf{V},\mathbf{W})$. For instance, consider the Average Treatment Effect $ATE_{\mathbf{v}\to\mathbf{v}'}$, the clever covariate is given by

$$\hat{H}(\mathbf{V},\mathbf{W}) = \frac{\mathbb{1}(\mathbf{V}=\mathbf{v}') - \mathbb{1}(\mathbf{V}=\mathbf{v})}{\hat{G}(\mathbf{V},\mathbf{W})} \tag{3.38}$$

Finally, it was shown that in practical cases of near positivity violations, an alternative, but equivalent procedure was more performant [144]. Instead of the natural loss associated with the paths of equations 3.36, a weighted loss is minimised. The weights are simply defined by the denominator of the clever covariate, that is $\frac{1}{\hat{G}(\mathbf{V},\mathbf{W})}$. In this case, the clever covariate reduces to $\hat{H}(\mathbf{V},\mathbf{W}) =$

$\mathbb{1}(\mathbf{V}=\mathbf{v})$ for $MI_{\mathbf{v}}$, or $\hat{H}(\mathbf{V},\mathbf{W})=\mathbb{1}(\mathbf{V}=\mathbf{v}')-\mathbb{1}(\mathbf{V}=\mathbf{v})$ for $ATE_{\mathbf{v}\to\mathbf{v}'}$. This TMLE variation, often called weighted-TMLE (wTMLE) is particularly relevant in fields such as genetics where some genetic variants are rare.

3.4.4 Empirical Process Term

We now move to the $(\mathbb{P}_n-P_0)(D_{\Psi,\hat{P}}-D_{\Psi,P_0})$ term, known as the empirical process term. In order to control this term, two conditions are required [77].

The first condition is that $D_{\Psi,\hat{P}}$ be $\mathcal{L}_2$-consistent for D_{Ψ,P_0}, that is

$$\int (D_{\Psi,\hat{P}}(o)-D_{\Psi,P_0}(o))^2 p_0(o)\mathrm{d}(o) \xrightarrow{P} 0 \tag{3.39}$$

In our case, since D_Ψ depends on Q_Y and G, it is enough to show that (i) $\hat{Q}_Y$ and $\hat{G}$ are $\mathcal{L}_2$-consistent and (ii) G is bounded away from 0 (positivity). Condition (i) is typically satisfied since most machine-learning algorithms are known to be $\mathcal{L}_2$-consistent [43, 89, 150]. The practical validity of condition (ii) is discussed in chapter 4.

The second condition is that the function class $\{D_{\Psi,P} : P\in\mathcal{M}\}$ and corresponding estimators are not too complex. The precise requirement is that the class of functions be Donsker [159]. Even though most machine-learning algorithms are not guaranteed to be Donsker, this is not necessarily an unachievable requirement. For example, functions of bounded variation are Donsker and Donsker classes are stable under many algebraic operations. Furthermore, it is also possible to bypass the Donsker class requirement altogether by leveraging sample splitting techniques [26]. The drawback is that such techniques are more computationally intensive, and may be unusable in large scale studies. They require splitting the data into K folds and estimating the nuisance functions independently on each fold. We present next the OSE and TMLE in their cross-validated versions. For a given sample i, we denote by $k(i)\in\{1,2,\ldots,K\}$ the fold it belongs to (called validation fold/set) and by $-k(i)$ the union of all remaining folds (called training set). Similarly, we denote by $\hat{Q}^k$ an estimator for Q obtained from samples in the validation fold k and $\hat{Q}^{-k}$ an estimator for Q obtained from samples in the (training) fold $\{1,...,K\}-\{k\}$.

3.4.4.1 Cross-Validated One-Step Estimator

Using the previous notation, the cross-validated one-step estimator (CV-OSE) can be compactly written as an average over the folds of sub one-step estimators

$$\begin{aligned}\hat{\Psi}_{\text{CV-OSE}} &= \sum_{k=1}^{K} \frac{N_k}{n}(\Psi(\hat{Q}_Y^{-k}, \hat{Q}_W^k) + \hat{\mathbb{P}}_n^k \hat{D}_\Psi^{-k}) \\ &= \frac{1}{n}\sum_{k=1}^{K} \sum_{\{i:k(i)=k\}} (\hat{Q}_Y^{-k}(\mathbf{V}_i = \mathbf{v}, \mathbf{W}_i, \mathbf{C}_i) + \hat{D}_\Psi^{-k}(\mathbf{V}_i, \mathbf{W}_i, \mathbf{C}_i, Y_i)),\end{aligned} \tag{3.40}$$

where the first equality holds generally for all estimands described in section 3.3 and the second equality is specific to $MI_{\mathbf{v}}$. The important thing to note is that for each sub one-step estimator, the sum runs over the validation samples while $\hat{Q}_Y^{-k}$ and $\hat{D}_\Psi^{-k}$ are estimated using the training samples.

3.4.4.2 CV-TMLE

Using the same notation as the CV-OSE, the targeted distribution is obtained by solving:

$$\boldsymbol{\varepsilon}^* = \arg\min_{\boldsymbol{\varepsilon}} \frac{1}{n}\sum_{k=1}^{K} \sum_{\{i:k(i)=k\}} L(Y_i, \hat{Q}_{Y,\boldsymbol{\varepsilon}}^{-k}(\mathbf{V}_i, \mathbf{W}_i, \mathbf{C}_i, Y_i)) \tag{3.41}$$

where $\hat{Q}_{Y,\boldsymbol{\varepsilon}}$ and L are the respective paths and losses for continuous and binary outcomes. This leads to a targeted $\hat{Q}_Y^*$ such that:

$$\forall i \in \{1, ..., n\}, \hat{Q}_Y^*(\mathbf{V}_i, \mathbf{W}_i, \mathbf{C}_i) = \hat{Q}_{Y,\boldsymbol{\varepsilon}^*}^{-k(i)}(\mathbf{V}_i, \mathbf{W}_i, \mathbf{C}_i) \tag{3.42}$$

That is, the predictions of $\hat{Q}_Y^*$ for sample i are based on the out of fold predictions of $\hat{Q}_Y^{-k(i)}$ and the "pooled" value of $\boldsymbol{\varepsilon}^*$.

Then, the cross-validated targeted minimum loss-based estimator (CV-TMLE) is

$$\begin{aligned}\hat{\Psi}_{\text{CV-TMLE}} &= \sum_{k=1}^{K} \frac{N_k}{n} \Psi(\hat{Q}_Y^*, \hat{Q}_W^k) \\ &= \frac{1}{n}\sum_{k=1}^{K} \sum_{\{i:k(i)=k\}} \hat{Q}_Y^*(\mathbf{V}_i = \mathbf{v}, \mathbf{W}_i, \mathbf{C}_i)\end{aligned} \tag{3.43}$$

Where again, the first equality holds generally for all estimands and the second equality is specific to $MI_{\mathbf{v}}$. Notice that, while $\hat{\Psi}_{\text{CV-TMLE}}$ is not a plugin estimator anymore, it still respects the natural range of the parameter because it is an average of plugin estimators.

3.4.5 Second-Order Remainder

The last term $R_2(\hat{P}, P_0)$ is the second-order remainder. Unlike the empirical process term, its behaviour is more specific to each estimand under consideration. In our case it can be shown that if $\hat{G}(\mathbf{v}, \mathbf{W}) \geq \varepsilon$ for some $\varepsilon > 0$, then for $MI_{\mathbf{v}}$

$$R_2(\hat{P}, P_0) \leq \frac{1}{\varepsilon} \cdot ||G(\mathbf{v}, \mathbf{W}) - \hat{G}(\mathbf{v}, \mathbf{W})||_2 \cdot ||Q_Y(\mathbf{v}, \mathbf{W}, \mathbf{C}) - \hat{Q}_Y(\mathbf{v}, \mathbf{W}, \mathbf{C})||_2 \quad (3.44)$$

Thus if $||G(\mathbf{v}, \mathbf{W}) - \hat{G}(\mathbf{v}, \mathbf{W})||_2 = o_p(n^{-\frac{1}{4}})$ and $||Q_Y(\mathbf{v}, \mathbf{W}, \mathbf{C}) - \hat{Q}_Y(\mathbf{v}, \mathbf{W}, \mathbf{C})||_2 = o_p(n^{-\frac{1}{4}})$, for example, then $R_2(\hat{P}, P_0) = o_p(n^{-\frac{1}{2}})$, as desired. But other combinations, with a nuisance function estimator converging faster than the other are also possible. Some algorithms, that unfortunately do not scale to large datasets, are known to converge at a rate faster than $o_p(n^{-\frac{1}{4}})$ [12]. In many other cases, some form of smoothness of the nuisance functions is required to obtain such rates. This property of the estimation problem is often called double robustness since, if we know one of the nuisance functions, it does not matter how well the other one is estimated.

We have seen that proper control of both the empirical process term and the second-order remainder depends on how well we can estimate the nuisance functions Q_Y and G. The following method, called Super Learning, provides a mean to achieve that goal.

3.4.6 Super Learning

Super-Learning, also known as Stack learning, is a machine learning approach that combines multiple models (referred to as base learners) to produce a single, more accurate predictive model [157, 177]. It operates on the principle that by leveraging the strengths of multiple algorithms, the overall performance can be improved beyond what could be achieved by any single model alone. The Super-Learner also benefits from theoretical properties, it performs asymptotically as well as the oracle, which is defined as the best estimator given the algorithms in the base learners. As such, it is a method of choice to make sure the empirical process term and second-order remainder are $o_p(\frac{1}{\sqrt{n}})$.

The method, presented in figure 3.6, begins by partitioning the dataset into K training and validation folds. A selection of base learners (e.g., linear regression, decision trees, neural networks) is then trained on each training fold and predictions generated for the associated validation fold. Since the validation folds

constitute a partition of the original dataset, the validation predictions are used as covariates to train a meta-learner. The final Super-Learner is obtained by retraining the base learners on the entire dataset and combining their predictions with the meta-learner.

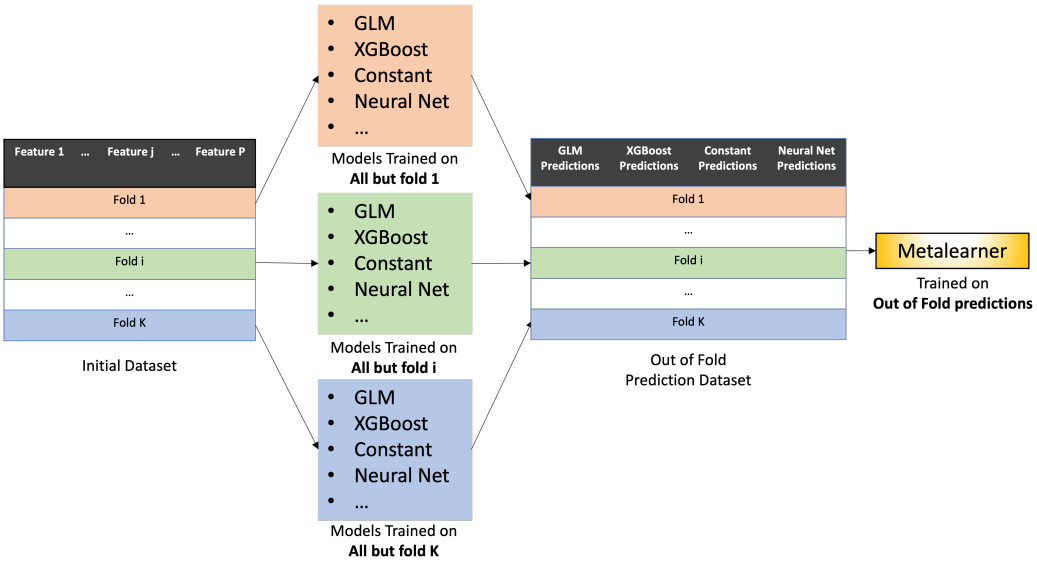

Figure 3.6: Super-Learning. The dataset is partitioned in K-folds and base learners trained on the resulting training sets. Predictions on the validation sets are concatenated to build the meta-learner's training set. All base learners are subsequently retrained on the entire dataset to build the final Super-Learner (not shown).

The Super-Learning algorithm was implemented in Julia and released as part of the open source MLJ ecosystem. We note that further levels of stacking are possible and could further improve performance [180], but were not investigated in this work.

3.5 Confidence Regions and Hypothesis Testing

Using the asymptotic normality of the estimators defined in the previous section we explain how confidence regions and hypothesis testing can be performed. We start with estimators of one-dimensional estimands, extend to multi-dimensional estimands and end with transformations of multi-dimensional estimands.

3.5.1 One-Dimensional Estimands

According to the previous section, for an estimand Ψ, the semi-parametric estimators we consider are asymptotically linear with influence curve the gradient D_{Ψ,P_0}. That is

$$\sqrt{n}(\hat{\Psi} - \Psi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_{\Psi,P_0}(\mathbf{O}_i) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \tag{3.45}$$

By the Central Limit Theorem [159]

$$\sqrt{n}(\hat{\Psi} - \Psi_0) \rightsquigarrow \mathcal{N}(0, Var[D_{\Psi,P_0}]). \tag{3.46}$$

Now, consider the sample variance of the gradient

$$\sigma_n^2 = \hat{Var}[D_{\Psi,\hat{P}}] = \frac{1}{n-1} \sum_{i=1}^{n} D_{\Psi,\hat{P}}(\mathbf{O}_i)^2 \tag{3.47}$$

σ_n^2 is a consistent estimator for $Var[D_{\Psi,P_0}]$, by Slutsky's Theorem and the Continuous Mapping Theorem [159]

$$\frac{\sqrt{n}(\hat{\Psi} - \Psi_0)}{\sigma_n} \rightsquigarrow \mathcal{N}(0,1) \tag{3.48}$$

This pivot can be used to build asymptotic confidence intervals of the form

$$\left[\hat{\Psi} - z_\alpha \frac{\sigma_n}{\sqrt{n}}, \hat{\Psi} + z_\alpha \frac{\sigma_n}{\sqrt{n}}\right], \tag{3.49}$$

where z_α denotes the α-quantile function of the standard normal distribution

3.5.2 Multi-Dimensional Estimands

Let $\mathbf{\Psi} = (\Psi_1, \ldots, \Psi_p)$ be a p-dimensional estimand and $\mathbf{D}_{\mathbf{\Psi},P_0}$ its gradient. $\mathbf{\Psi}$ and $\mathbf{D}_{\mathbf{\Psi},P_0}$ are p-dimensional vectors of one-dimensional estimands and gradients. If $\hat{\mathbf{\Psi}}$ is an asymptotically linear estimator with influence curve the gradient $\mathbf{D}_{\mathbf{\Psi}}$, then by the multi-dimensional Central Limit Theorem

$$\sqrt{n}(\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0) \rightsquigarrow \mathcal{N}(0, \mathbf{\Sigma}) \tag{3.50}$$

where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{D}_{\mathbf{\Psi},P_0}$. Similarly to the one-dimensional case, we would like to build a pivot for hypothesis testing and the definition of confidence regions. The Hotelling's T-squared statistic (T^2) is the generalisation

of the Student's T-statistic that is used in multivariate hypothesis testing. Let $\hat{\mathbf{\Sigma}}$ be the sample covariance matrix whose elements are given by

$$\hat{\mathbf{\Sigma}}_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} D_{\Psi_j,\hat{P}}(\mathbf{O}_i) D_{\Psi_k,\hat{P}}(\mathbf{O}_i) \tag{3.51}$$

The Hotelling's t-squared statistic is then defined as

$$t^2 = n(\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0)^T \hat{\mathbf{\Sigma}}^{-1} (\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0) \sim T^2_{p,n-1} \tag{3.52}$$

which can be equivalently represented by the well-tabulated F-statistic to compute p-values and confidence regions.

$$\frac{n-p}{p(n-1)} t^2 \sim F_{p,n-p} \tag{3.53}$$

Note that confidence regions are multi-dimensional and hence difficult to visualise for estimands of dimension more than 2.

3.5.3 Composition of Multi-Dimensional Estimands

Finally, suppose we have obtained an estimate for a p-dimensional estimand $\mathbf{\Psi}$ and wish to obtain an estimate for $f(\mathbf{\Psi})$ for some function f. Such an estimate can be obtained using the functional delta method [159]. Let $f : \mathbb{R}^p \to \mathbb{R}^m$ be a differentiable map at $\mathbf{\Psi}_0$. We have seen in the previous section that $\sqrt{n}(\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0) \rightsquigarrow \mathcal{N}(0, \mathbf{\Sigma})$. The delta method states that

$$\sqrt{n}(f(\hat{\mathbf{\Psi}}) - f(\mathbf{\Psi}_0)) \rightsquigarrow \mathcal{N}(\mathbf{0}_m, \nabla f(\mathbf{\Psi}_0) \mathbf{\Sigma} (\nabla f(\mathbf{\Psi}_0))^T) \tag{3.54}$$

where $\nabla f(\mathbf{\Psi}_0) : \mathbb{R}^p \to \mathbb{R}^m$ is a linear map such that by abusing notations and identifying the function with matrix multiplication $\nabla f(\mathbf{\Psi}_0)$ is the Jacobian of f at $\mathbf{\Psi}_0$.

In practice, gradient computation is efficiently handled using modern automatic differentiation techniques. This allows any differentiable function f to be evaluated without the need to explicitly derive its Jacobian matrix. Depending on the dimensionality of f, the methods discussed in the previous sections can then be applied to conduct hypothesis testing and to establish asymptotic confidence regions with precision.

3.5.4 False Discovery Rate Control

In statistical genetics, the number of tested hypotheses is large, in a traditional GWAS, this is in the order of millions. Even if each individual test is conducted with a 5% significance level, the overall chance of finding at least one false positive is much higher. Multiple hypotheses correction methods aim at controlling the number of false positives due to simultaneous testing. The Family-Wise Error Rate (FWER) and the False Discovery Rate (FDR) are two metrics used in multiple hypotheses testing to control the likelihood of making errors, they address different types of errors and have different goals.

The FWER is the probability of making at least one Type I error (false positive) across a family of tests. It is a stringent measure that aims to minimise the chance of any false positives occurring among the tests. The goal of controlling the FWER is thus to ensure that the likelihood of making one or more false positive errors remains below a specified threshold (e.g., 5%). It is particularly important in fields where false positives can have significant consequences, such as clinical trials. One of the simplest and most conservative methods to control the FWER is the Bonferroni correction [173]. It adjusts the significance level for each individual test by dividing the desired overall alpha level (e.g., 0.05) by the number of tests conducted. This ensures that the probability of making one or more Type I errors across all tests remains below the alpha level.

The FDR is the expected proportion of Type I errors (false positives) among the rejected hypotheses. It provides a less stringent control compared to FWER by focusing on the rate of false positives among the discoveries rather than ensuring no false positives at all. The goal of controlling the FDR is to allow some level of false positives while controlling their proportion among the rejected hypotheses. This approach is useful in large-scale testing scenarios, such as genomics where a small number of false positives is acceptable if it leads to a larger number of true discoveries. The most commonly used method to control the FDR is the Benjamini-Hochberg procedure [11], which ranks p-values and adjusts them to control the expected proportion of false discoveries.

In GWAS analyses, FWER control using the Bonferroni method is the standard multiple hypotheses correction strategy. It is believed to be one of the reasons for the replicability of genome-wide association results [82]. Indeed, a genome-wide ($\approx 10^6$ tests) significance p-value threshold of $5 \times 10^{-8} = 0.05/10^6$

is used to report significant variants. However, this approach is extremely stringent and will result in a loss of power to detect true positives in large studies [25]. In this thesis, we are ready to accept a small set of false discoveries and favour the FDR, using the Benjamini-Hochberg procedure throughout.

3.6 Sieve Variance Plateau Correction

Many statistical analyses rely on the assumption that the data units are independently sampled from the population; it is also often required that the data be identically distributed but this assumption is typically not essential. However, the independence assumption may no longer hold for participants in the UK Biobank since many of them are, to some extent, genetically related. Such genetic similarity can occur on a sub-population level due to ancestry (e.g., being white Irish), or on an individual level due to kinship (e.g., parents, children, cousins). In this section, we assume that the dependence among individuals is sufficiently weak so that the estimators of section 3.4 are still asymptotically linear. We propose however, to account for dependence by adjusting variance estimates via Sieve Variance Plateau (SVP) estimation [35]. Like Linear Mixed Models 2.3.2, SVP variance estimators are based on the genetic similarity between participants, however, they are fully non-parametric. In what follows, Z is the GRM computed as per equation 2.14, from a set of R genotyped and independent variants (`plink2 --indep-pairwise 1000 50 0.05`).

3.6.1 Sieve Plateau Variance Estimators

We first illustrate how dependence among individuals impacts the variance of semi-parametric estimators. The problem arises since for two random variables X_1 and X_2 the variance of their sum is not in general equal to the sum of their variances but includes a covariance term.

$$Var(X_1 \pm X_2) = Var(X_1) + Var(X_2) \pm 2 \cdot Cov(X_1, X_2). \tag{3.55}$$

The impact of the dependence between X_1 and X_2 is entirely captured by the covariance term $Cov(X_1, X_2)$. A more general version of equation 3.55, with n random variables, applied to the variance of asymptotically-linear estimators

shows that their variance is given by

$$\hat{\sigma}_n^2 = Var\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{\Psi,P_0}(\mathbf{O}_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov\left(D_{\Psi,P_0}(\mathbf{O}_i), D_{\Psi,P_0}(\mathbf{O}_j)\right). \quad (3.56)$$

Note that if $\mathbf{O}_i$ and $\mathbf{O}_j$ are truly independent, then so are $D_{\Psi,P_0}(\mathbf{O}_i)$ and $D_{\Psi,P_0}(\mathbf{O}_j)$ and $Cov\left(D_{\Psi,P_0}(\mathbf{O}_i), D_{\Psi,P_0}(\mathbf{O}_j)\right) = 0$ when $i \neq j$, hence not contributing to the estimator's variance. This idea is at the heart of SVP estimators [35]. The SVP estimator computes a variance estimate for a range of thresholds τ, by considering individuals to be genetically independent if their genetic distance exceeds τ. The genetic distance between a pair of individuals (i, j) is defined by $1 - Z_{i,j}$, where Z_{ij} is the sample correlation coefficient of equation 2.14. Since correlation is bounded, $|Z_{ij}| \leq 1$, the genetic distance is non-negative and never larger than two, i.e., $0 \leq d(i, j) \leq 2$. As the distance threshold τ increases, fewer individuals are assumed to be genetically independent. For instance, the estimate corresponding to a distance of $\tau = 0$ corresponds to the independence hypothesis, while a distance of $\tau = 1$ incorporates pairs of individuals who are not genetically correlated. Formally, given a threshold τ, a SVP estimator is given by

$$\hat{\sigma}_n^2(\tau) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\{d(i, j) \leq \tau\} \cdot D_{\Psi,\hat{Q}}(o_i) \cdot D_{\Psi,\hat{Q}}(o_j). \quad (3.57)$$

In practice, we construct the variance estimator $\hat{\sigma}_n^2(\tau)$ for a number of values of the cut-off τ, e.g., $\tau = 0, \frac{1}{k}, \frac{2}{k}, \ldots, \frac{k-1}{k}, 1$, thereby defining a curve. The maximum of this curve, attained at τ_0, is the most conservative estimate of the variance and constitutes the SVP variance estimator $\hat{\sigma}_n^2(\tau_0)$. While the curve can be non-monotonic and lead to points of lower variance, since the maximum of the curve is used as a final estimate, the SVP variance estimator is guaranteed to be at least as large as the initial variance estimator. An illustration of the method is presented in section 6.2.3. Confidence intervals can be built as before, using equation 3.49, and simply replacing $\hat{\sigma}_n^2$ with $\hat{\sigma}_n^2(\tau_0)$.

3.7 Discussion

In this chapter we looked at the field of population genetics from a causal inference perspective. The mathematical treatment of causation provided a clear analysis framework for the definition of genetic effects. Unfortunately, under the proposed inheritance model, these effects are not identifiable. The sparsity of genotyping

arrays results in unobserved, potentially causal, variations confounding the effect of genotyped variations. The recent release of whole genome sequencing data has the potential to address this issue, provided any two variations in the same linkage disequilibrium block are statistically distinguishable. We then proceeded with the traditional heuristic model, with population stratification control via Principal Component Analysis. In this model, genetic effects are identified and estimation can proceed. Rather than committing to a restrictive parametric model we defined more realistic semi-parametric estimators that leverage flexible machine-learning methods. Under specific conditions, these estimators are guaranteed to be asymptotically unbiased and efficient. These conditions notably include the fast convergence of machine-learning algorithms and the positivity constraint (section 2.2.2). This latter condition thus has both causal and statistical importance for identifiability and estimation respectively. In chapter 4, we devise simulations to understand the practical implications of such conditions. Since these estimators are asymptotically normal, the central limit theorem can be used to build asymptotic confidence regions and perform hypothesis testing. Finally, we proposed a network-based method to mitigate the hypothesis that individuals are independent. This method, called Sieve Variance Plateau, adjusts variance estimates using the genetic relationship matrix as a similarity measure between individuals.

Chapter 4

Real-World Data Based Simulation Study

We have seen that the asymptotic performance of semi-parametric estimators is theoretically optimal. In practice, asymptotic regimes are not necessarily achieved even for large sample sizes because some events are extremely rare. This is particularly prevalent in population genetics where some genetic variants and traits are found in less than 1% of individuals. In the absence of finite sample guarantees, simulation studies provide an effective way to validate statistical methods. An ideal simulation would both yield ground-truth values for the causal effects of interest and be representative of real data. Ground truth values are easily obtained via algebraic formulas when the generating process is parametric. In Population Genetics, this generating process is usually assumed to be linear-Gaussian [95]. However, others have recognised that these simulations are deprived of important features of the true generating process [115, 135]. As a result, conclusions obtained through parametric simulations may not generalise well when analysing real world data. A better trade-off can be obtained by modelling the generating process with more flexible semi-parametric generative models. While this approach does not yield an algebraic definition of the ground truth, an arbitrarily precise estimate can be obtained via Monte-Carlo sampling. Although this approach is less resource-efficient than having a closed form solution, it only needs to be performed once per simulation, making its computational cost negligible.

In this chapter, this is the approach we take. We analyse the performance of semi-parametric estimators in population genetics analyses through two types of simulations. For both simulations, the whole UK Biobank population (sec-

tion 3.1) is used and a variety of estimation tasks are examined. The first simulation, called the null simulation, aims at analysing the behaviour of the estimators when the null hypothesis is true. The rationale behind this simulation is that most genetic variants are believed to have no effect, and it is thus of particular importance that the type 1 error rate be controlled appropriately. This is especially true since experimental follow-up studies are typically expensive. In the second simulation, termed the realistic simulation, we fit a generating process according to the Working Causal Model using Neural Network based density estimators. We first describe the set of tasks we consider in these studies in section 4.1. Both simulations are then presented in more detail in section 4.2. Finally, we analyse the performance of our semi-parametric estimators in section 4.3. Note that these simulations do not aim at comparing semi-parametric estimators to the gold standard in population genetics. Rather, they aim at identifying a practical set of conditions under which semi-parametric estimators are believed to yield robust inferences. A large scale comparison of semi-parametric estimators and linear-based models in population genetics is left for future work.

4.1 Simulations' Estimands

For the simulation study to be most informative, it needs to be representative of a variety of estimation tasks that can be found in population genetics analyses. Two of the most prevalent tasks are single variant analyses, where the marginal effect of a variant across traits is estimated, and interaction analyses where two (or more) genetic variants are under investigation. These were defined by the Average Treatment Effect and Average Interaction Effect respectively in section 3.3. Genetic variations can be more or less frequent, sometimes less than 1% of the population. Since we already know from section 2.2.2 that positivity is an important criterion for causal inference, we expect that the effect of rare variations will be challenging to estimate. Finally, in the case of rare binary traits like diseases, it is also uncertain how the estimators will perform since the asymptotic regime may not have been reached.

To account for the aforementioned challenges, we selected 29 estimands, listed in Table 4.1, according to the following criteria. All included estimands were supported by previous evidence so that the realistic simulation may diverge from the null hypothesis of no effect. Note however, that this is not mandatory and

that the goal of the simulation is not the replication of these effect sizes.

For single variant effects, 5 different traits were chosen:

- Leukocyte count: Count.
- Body mass index (BMI): Continuous.
- Sarcoidosis (self-reported) and Sarcoidosis (D86): approximately 1000 cases each. We use both definitions as a sanity check since the disease is rare.
- Multiple sclerosis: 1900 cases.
- Other diseases of the digestive system (K90-K93): approximately 25000 cases.

For each trait, 2 variants were manually selected based on previous GWAS results (p-value $< 10^{-5}$) from the geneATLAS [21]. To increase the diversity of tasks, these two variants were not in linkage disequilibrium and had different minor-allele frequencies.

For interaction analyses we used previous results from the literature as follows:

- Skin colour was shown to be associated with (rs1805007, rs6088372), (rs1805005, rs6059655) and (rs1805008, rs1129038) [104].
- Parkinson's disease was associated with (rs1732170, rs456998, rs356219, rs8111699) and (rs11868112, rs6456121, rs356219) [44]
- Multiple sclerosis was associated with (rs10419224, rs59103106) [141]
- Psoriasis was associated with (rs974766, rs10132320) [141]

The set of all estimands we consider is summarised in Table 4.1 together with additional statistics. In summary, we consider 29 estimands corresponding to both Average Treatment Effects (ATE) and Average Interaction Effects (AIE). These estimands represent both rare and more frequent genetic variations as well as a variety of trait types including continuous, count and binary. We believe these to be sufficiently diverse to provide relevant information on the performance of semi-parametric estimators in many classical settings.

Type	Outcome	Outcome Type	Variants	Outcome Freq	Variants Min Freq	Joint Min Freq
AIE	Body mass index	Continuous	rs62107261;rs9940128		4.2e-04	
AIE	Leukocyte count	Continuous	rs3859191;rs9268219		3.3e-03	
AIE	Multiple sclerosis	Binary	rs10419224;rs59103106	4.0e-03	3.9e-04	2.1e-06
AIE	Multiple sclerosis	Binary	rs3129889;rs62295911	4.0e-03	8.4e-05	2.1e-06
AIE	Other diseases of the digestive system	Binary	rs3129716;rs72926466	5.3e-02	3.0e-04	1.8e-05
AIE	Parkinson's disease	Binary	rs11868112;rs356219;rs6456121	6.5e-03	2.1e-03	1.8e-05
AIE	Parkinson's disease	Binary	rs1732170;rs356219;rs456998;rs8111699	6.5e-03	1.4e-03	6.2e-06
AIE	Psoriasis	Binary	rs10132320;rs974766	1.0e-02	1.0e-04	4.1e-06
AIE	Sarcoidosis (D86)	Binary	rs148515035;rs502771	2.3e-03	2.5e-05	2.1e-06
AIE	Skin colour	Count	rs1129038;rs1805008		3.6e-04	
AIE	Skin colour	Count	rs1805005;rs6059655		1.6e-04	
AIE	Skin colour	Count	rs1805007;rs6088372		3.0e-04	
AIE	Type 2 diabetes	Binary	rs117737810;rs4506565	8.8e-03	7.8e-05	2.1e-06
AIE	psoriasis	Binary	rs10132320;rs974766	1.2e-02	1.0e-04	2.1e-06
AIE	sarcoidosis	Binary	rs148515035;rs502771	2.1e-03	2.5e-05	2.1e-06
ATE	Body mass index	Continuous	rs62107261		2.2e-03	
ATE	Body mass index	Continuous	rs9940128		1.8e-01	
ATE	Leukocyte count	Continuous	rs3859191		2.2e-01	
ATE	Leukocyte count	Continuous	rs9268219		1.5e-02	
ATE	Multiple sclerosis	Binary	rs3129889	4.0e-03	2.0e-02	2.5e-04
ATE	Multiple sclerosis	Binary	rs62295911	4.0e-03	3.1e-03	2.7e-05
ATE	Other diseases of the digestive system	Binary	rs3129716	5.3e-02	2.1e-02	1.9e-03
ATE	Other diseases of the digestive system	Binary	rs72926466	5.3e-02	1.6e-02	8.4e-04
ATE	Sarcoidosis (D86)	Binary	rs148515035	2.3e-03	3.7e-04	2.1e-06
ATE	Sarcoidosis (D86)	Binary	rs502771	2.3e-03	7.5e-02	3.1e-04
ATE	Type 2 diabetes	Binary	rs117737810	8.8e-03	9.2e-04	1.0e-05
ATE	Type 2 diabetes	Binary	rs4506565	8.8e-03	1.0e-01	1.4e-03
ATE	sarcoidosis	Binary	rs148515035	2.1e-03	3.7e-04	4.1e-06
ATE	sarcoidosis	Binary	rs502771	2.1e-03	7.5e-02	3.2e-04

Table 4.1: The 29 estimands used across the simulation study. The "Variants Min Freq" column represents the minor genotype frequency for the variants in the estimand. When the outcome is binary, the frequency is provided as well as the "Joint Min Freq". The later represents the minor frequency of joint (genotype, outcome).

4.2 Simulations

In this section, we motivate and describe the two types of generating processes that are used across the simulation study. A summary schematic illustration of both causal diagrams can be found in figure 4.1.

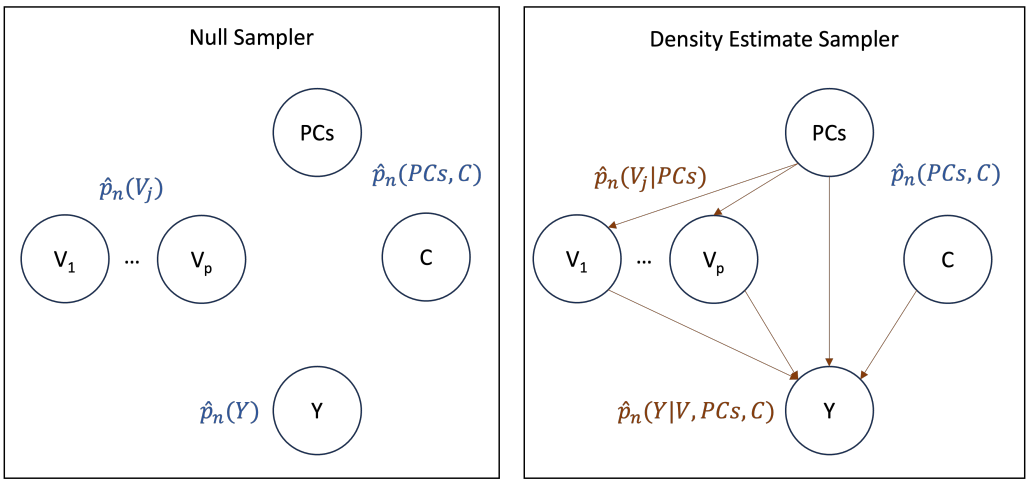

Figure 4.1: The two generating processes used for the study. Empirical marginal distributions are coloured in blue while learnt conditional densities are coloured in orange. In both cases (PCs, C) are sampled jointly using the empirical marginal distribution. The null sampler then independently samples from the empirical marginal distributions of each Y, V_j. This results in the theoretical null hypothesis of no effect. The density estimate sampler proceeds via ancestral sampling, first each V_j is sampled from $\hat{P}_n(V_j|PCs)$, then Y is sampled from $\hat{P}_n(Y|\mathbf{V}, PCs, C)$. The various causal effects can then be approximated via Monte-Carlo sampling using $\hat{P}_n\big(Y|do(\mathbf{V}), PCs, C\big)$.

4.2.1 The Null Simulation

The goal of the null generating process is to result in the theoretical null hypothesis of "no effect". The main rationale behind this simulation is that most of the genome is still believed to be irrelevant [125]. It is thus of particular importance that the proposed semi-parametric estimators are not too sensitive to noise in order to control the false discovery rate. Since we are interested in the effects of genetic variants on traits, the data generating process must satisfy: $\forall j, Y \perp\!\!\!\perp V_j$, where Y is a given trait. In practice, we enforce an even stronger condition, where all variables are pairwise independent. This is done by drawing independently n samples with replacement from the empirical marginal distribution of each variable. The only exception is that (PCs, C) are sampled jointly. This generating process, presented in figure 4.1 (left), hence preserves many characteristics of the original dataset, while resulting in the null hypothesis.

4.2.2 The Realistic Simulation

In order to understand how well semi-parametric estimators will perform in real-world scenarios, we aim to simulate new data which is as close as possible to the original dataset. For each estimand of interest, we assume a causal graph similar to that of figure 4.1 (right). Each estimand is thus associated with a natural generating process, or equivalently, set of densities. For example the interaction of (V_1, V_2) on Y requires three density estimates, namely: $\hat{P}_n(V_1|PCs)$, $\hat{P}_n(V_2|PCs)$ and $\hat{P}_n(Y|V_1, V_2, PCs, C)$. Similarly, the single variant effect of V_1 on Y requires: $\hat{P}_n(V_1|PCs)$ and $\hat{P}_n(Y|V_1, PCs, C)$. Once these conditional densities have been estimated, new data can be generated via ancestral sampling. Implicitly, the empirical distribution $\hat{P}_n(PCs, C)$ is always used to sample from the root nodes.

For the simulation to be realistic, the conditional density estimators must be able to capture the complexity of the data, which has two main implications. The first is that the Causal Model should include as many causal variables as possible to generate the corresponding children variables. The second is that the density estimators must be flexible and data adaptive to capture arbitrarily complex data generating processes. We now discuss how we address both challenges in turn.

4.2.2.1 Variable Selection

The Causal Model of figure 4.1 contains two hierarchical layers, or equivalently, structural equations. The first layer corresponds to the generation of genetic variants and the second layer to the generation of human traits. Ideally, for each layer, we would know and observe the exact set of causal variables. Unfortunately, this is not the case (it is the topic of this thesis), and we will rely on reasonable heuristics.

For genetic variants, as discussed in section 3.2, it is unclear what the causal variables are. However, in this simulation study, we are only interested in the statistical performance of estimators. Remembering that the total estimation error can be decomposed as

$$\text{Estimation Error} = \text{Causal Error} + \text{Statistical Error} \tag{4.1}$$

we are not evaluating the Causal Error in this simulation. Therefore, it is relatively safe to assume that we have the complete set of genetic causes. We will proceed under the assumption that this set consists of the first six principal components derived from the genotyping data, as discussed in section 3.2.3.1. The

primary risk of this assumption is that some genetic variants might not be influenced by these principal components. In such cases, although the simulation is unconfounded and may be less informative, the proposed estimators should still perform as intended. Conversely, if more than six principal components carry meaningful information, the simulated data may become less realistic. Specifically, for any given variant's $V = v$ and principal components' $\boldsymbol{PCs} = \boldsymbol{pcs}$, the minimum probability $\min_{v,\boldsymbol{pcs}} p(v|\boldsymbol{pcs})$ may decrease as the number of principal components increases, potentially exacerbating positivity violations. However, based on the forthcoming examples in Chapters 6 and 7, we anticipate that the impact of retaining more than six principal components will generally be minor.

Regarding trait variables, in essence, the whole genome could be causal as well as external environmental variables. While the construction of such a model would be extremely useful, it is beyond the scope of this thesis. Instead, we will need restrict the dimensionality of the problem and only consider a small subset of these putative causes. Environmental variables are simply kept to the traditional covariates, that is: age at assessment and genetic-sex. To include further potential causal variants from the whole genome, we use, once again, published GWAS results from the geneATLAS [21]. Precisely, we sub-sample a maximum of 50 variants from all variants associated (p-value $< 10^{-5}$) with the outcome of interest. Furthermore these variants must be at least 1 000 000 base pairs away from each other and have a minor allele frequency of at least 0.01. With respect to figure 4.1, these sub-sampled variants and environmental variables are both contained within the $\mathbf{C}$ variable. The fact that sub-sampled variants are not generated from principal components is another mild limitation (widely reducing computational burden). Since all variables in $\mathbf{C}$ are jointly sampled, the dependence structure within sub-sampled variants is preserved. However, sub-sampled variants and variants defining the causal estimands are indeed independent in this simulation. In a more realistic scenario they would only be independent once conditioned on principal components. This limitation is again related to the causal error component of the estimation error which is left for further work.

4.2.2.2 Model Selection

The second requirement for the simulation to be realistic is that the density estimators should capture complex patterns, which means the model class must be large. Neural networks, have been shown to be able to approximate a large

class of function and scale seamlessly to large datasets such as the UK Biobank [65]. We thus used two types of neural networks depending on the type of the density's outcome variable. For categorical variables, including binary outcomes, a one hidden-layer perceptron was employed, while for continuous variables we used a one layer mixture density network [15, 62]. In essence, the models were designed to be computationally efficient for fast training while remaining flexible enough to capture complex interactions between variables. In all cases, the size of the hidden layer was chosen via cross-validation, using a sieve [24]. That is, the size of the model was chosen data adaptively by sequentially increasing the hidden layer size (candidates: $[5, 10, 20, 40, 60, 80, 100, 120, 140]$) and early-stopping based on cross-validation performance [121]. To limit the computational burden, If the performance is not improved for a number of consecutive new hidden layer sizes ($maxSievePatience$), then the procedure stops and returns the current best model. For illustration, a simplified training procedure is described in Algorithm 1. The undefined "train" function, corresponds to each neural-network's training loop and also implicitly uses early-stopping to control the number of training epochs of each proposed architecture.

Algorithm 1 Sieve Neural Network Estimator

```
procedure SNNE(hiddenLayerSizes, dataset, maxSievePatience)
    trainingSet, validationSet ← split(dataset)
    bestModel ← build(hiddenLayerSizes[1])
    bestValidationLoss ← train(bestModel, trainingSet, validationSet)
    sievePatience ← 0
    for hiddenLayerSize ∈ hiddenLayerSizes[2 : end] do
        model ← build(hiddenLayerSize)
        validationLoss ← train(model, trainingSet, validationSet)
        if validationLoss ≤ bestValidationLoss then
            bestValidationLoss ← validationLoss
            bestModel ← model
            sievePatience ← 0
        else
            sievePatience ← sievePatience + 1
        end if
        if sievePatience == maxSievePatience then
            break
        end if
    end for
    return bestModel
end procedure
```

4.2.3 The Estimators

We have seen in section 3.4, that semi-parametric estimators exist in many different flavours. It is useful, and perhaps intuitive, to look at the semi-parametric estimators we study in this thesis, along three main axes (figure 4.2). On the first axis is the estimator type and we will consider three different estimators: the one-step estimator (OSE), the targeted maximum-likelihood estimator (TMLE) and the weighted targeted maximum-likelihood estimator (wTMLE). On the second axis is the resampling strategy, each estimator can be used in its canonical form (Canonical) or using cross-validation (CV). Finally, on the third axis are the models used to estimate the nuisance functions. In this study we only consider 4 types of modelling strategies.

1. GLMNet: Relies on the Generalised Linear Models [49] for both nuisance functions and incorporates all cross-terms involving genetic variations.

2. XGBoost: Relies on a scalable gradient tree boosting model [23] for both nuisance functions. An optimal model is selected based on cross-validation across a simple grid of hyper-parameters (max_depth, lambda) to prevent over-fitting.

3. SL: Super Learning (or Stack Learning) [177] is an ensemble method where models are combined in a cross-validation scheme to improve out of sample performance (section 3.4.6). Here, we combine the aforementioned XGBoost, GLMNet as well as a simple Generalised Linear Model for both nuisance functions.

4. G_SL_Q_GLMNet: Uses Super Learning for the estimation of the G function and GLMNet for the Q_Y function. The rationale behind this hybrid approach is the double robustness property of semi-parametric estimators together with the fact that G can be reused across multiple traits (section 5.3).

In all cases, the cross-validation scheme is a 3-folds stratified cross-validation, stratified across both variants of interest (not the sub-sampled variants) and outcome variables.

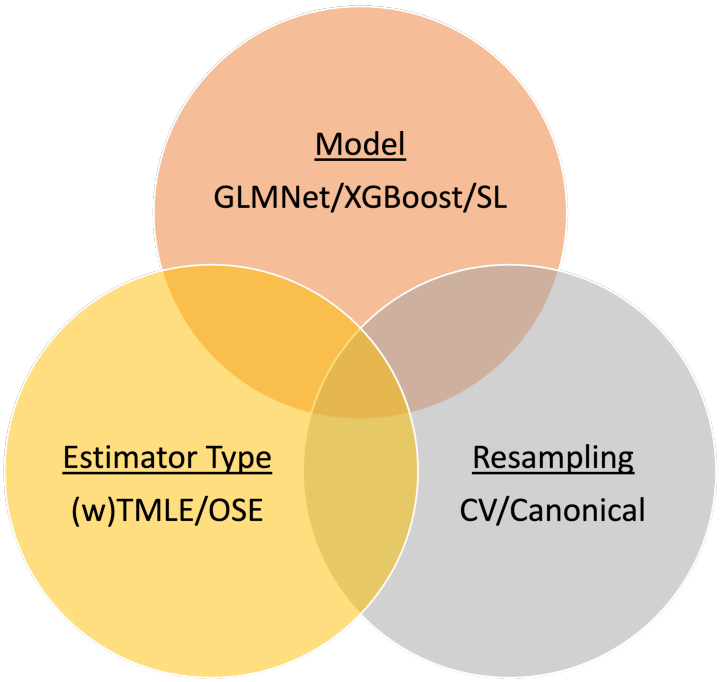

Figure 4.2: The three axes defining a semi-parametric estimators: Estimator Type, Resampling and Model used to learn the nuisance functions.

For the present simulation study, the set of estimators we investigate is simply the Cartesian product of all estimators along the three axes.

4.3 Results

To investigate the suitability of semi-parametric methods for various bio-banks, we perform the analysis across 3 different dataset sizes $(100\,000, 250\,000, 500\,000)$. A simulation task is thus a triple (Sample Size, Estimator, Estimand). For each task, a grid of 500 bootstrap samples was run. However, due to the high computational cost of this simulation study, not all tasks could be run to completion. This, using 1000 cores on the University of Edinburgh Eddie cluster for more than 3 weeks for each simulation. For both simulation studies, approximately $90-95\%$ of bootstrap samples were collected and tasks with less than 400 bootstrap samples discarded. The remaining 5% may seem trivial to obtain but this is not the case because all estimators and estimands are not computationally equal. Most discarded tasks correspond to cross-validated estimators using Super Learning, thereby inducing two nested loops of cross-validation with multiple model fits within them. Running only a few bootstrap samples for these tasks could

last for more than 2 days. For completeness, the list of discarded tasks for both simulations is presented in Supplementary Table 6.

4.3.1 Null Simulation

As discussed above, the goal of the null simulation is to ensure the false discovery rate is under control when resorting to semi-parametric estimators in population genetics analyses. Since there is effectively no relation between genetic variants and outcomes, model misspecification is practically impossible. All proposed estimators should thus provide perfect coverage probability at the nominal confidence level (e.g. 95%) across all tasks. Figure 4.3-A shows that this is not the case. For all estimators, there is at least one estimand for which the coverage is below 70%. This means that, either the asymptotic regime has not been reached, or, the positivity assumption is violated (see definition 2.2.4). In practice, machine-learning methods extrapolate for each confounding variables' values and positivity is guaranteed for every single genotype. However, this extrapolation is dangerous since it applies to regions that are not supported by the data and could lead to the bias observed here [118]. Furthermore, rare genetic variations in finite samples can result in numerical instabilities that can also affect inferences. These issues, can be understood broadly as a practical violation of the positivity constraint. Exact control of positivity is challenging since it depends on the true, but usually unknown value of $p(T = t|W = w)$ for all $t, w \in \mathcal{T}, \mathcal{W}$. In the null simulation however, $p(T|W) = p(T)$ since T and W are sampled independently. It is thus equivalent to control for the marginal positivity constraint defined in definition 3.2.1. Since all estimands in the simulation are joint estimands, we do so by considering only ε-constrained estimands (definition 3.2.2). Figure 4.3-B presents coverage results for this restricted set with $\varepsilon = 0.01$, the traditional GWAS minor allele frequency threshold. The estimation results now align perfectly with the nominal 95% coverage probability, confirming that the issue was indeed related to positivity violation.

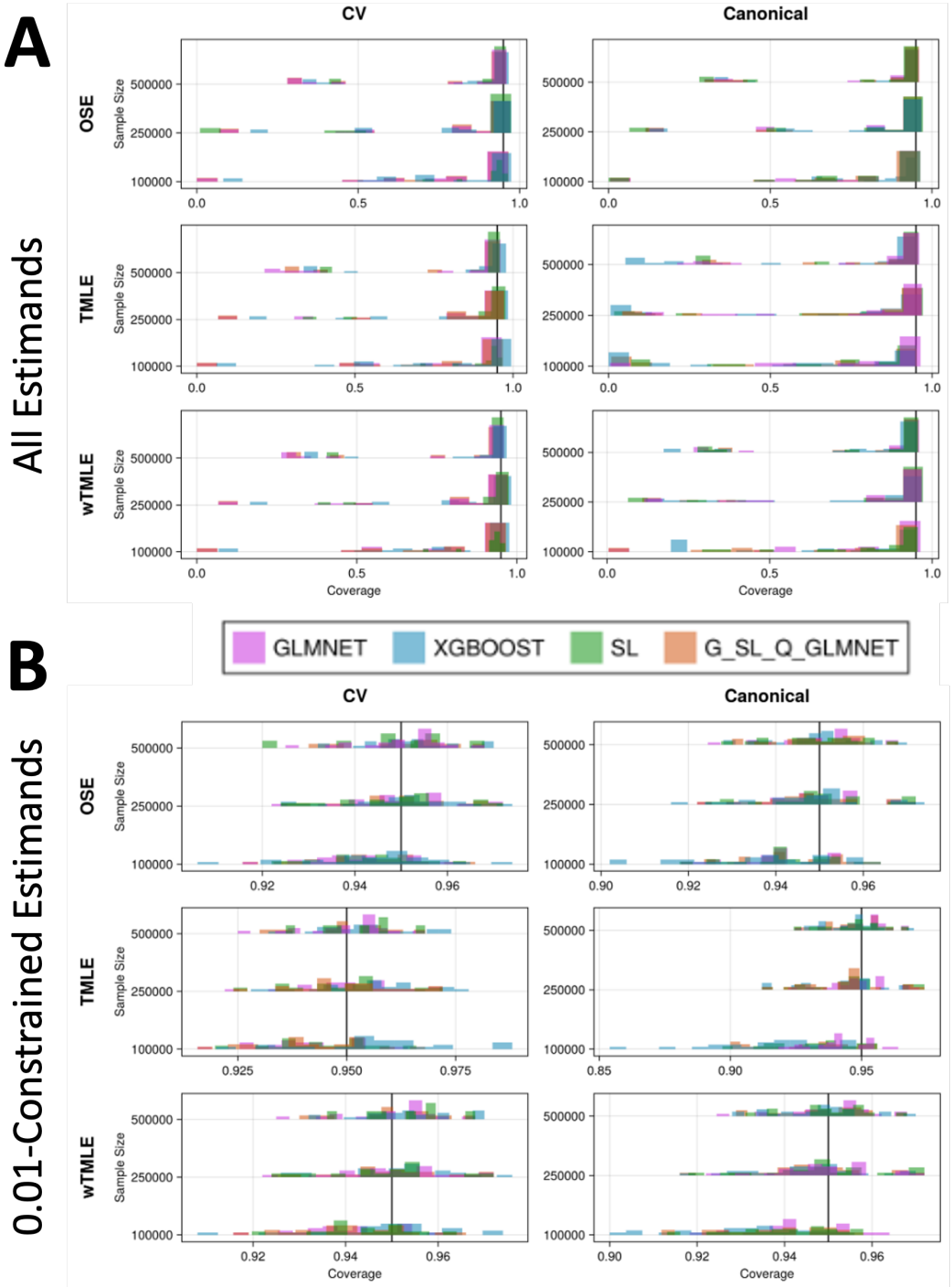

Figure 4.3: **(A) Estimation results across all tasks (Estimands, Estimators, Sample Sizes).** The two columns correspond to either cross-validated (CV) or canonical estimators. Rows correspond to all three variations of semi-parametric estimators (OSE, TMLE, wTMLE). Each histogram represents the mean coverage distribution across all estimands for the given model specification. The figure shows that while most of the mass is concentrated on the expected 95% level, some estimands suffer from low coverage. **(B) Estimation results with 0.01-constrained estimands.** The figure is organised exactly as figure (A) but the estimands are filtered to contain only components that satisfy a 0.01 marginal positivity threshold. The mass is now fully centred on the nominal 95% level with little variations around it (Note the difference in x-axis limits).

While traditionally used for most GWAS studies with over 100 000 samples, the 0.01 threshold is somewhat arbitrary and can be refined. It must also be emphasised that the GWAS threshold does not bound the same quantity. The GWAS threshold is usually set on the minor allele frequency while here it is constraining the joint variants' frequency. For instance, consider the rarest variant in the simulation, rs148515035 with minor allele frequency 0.019 (A). The genotypes frequencies for this variant are however: (TT, 0.96), (TA, 0.037) and (AA, 0.0003). While this variant would pass the GWAS threshold, only the TA $\rightarrow$ TT genotype change would be considered according to our results. Even though minor allele frequencies and genotypes frequencies are related, only the latter defines a treatment level and is supported by theory. The proposed $\boldsymbol{\varepsilon}$-constraint thus provides a refined criterion for the selection of candidate variants whose effects are estimable. Figure 4.4-A shows that the value $\boldsymbol{\varepsilon} = 0.005$ is in fact sufficient to guarantee nominal coverage. Furthermore, there is no noticeable difference across estimators or models used to learn the nuisance functions (figure 4.4-B). This is expected in this simulation where model mis-specification is impossible.

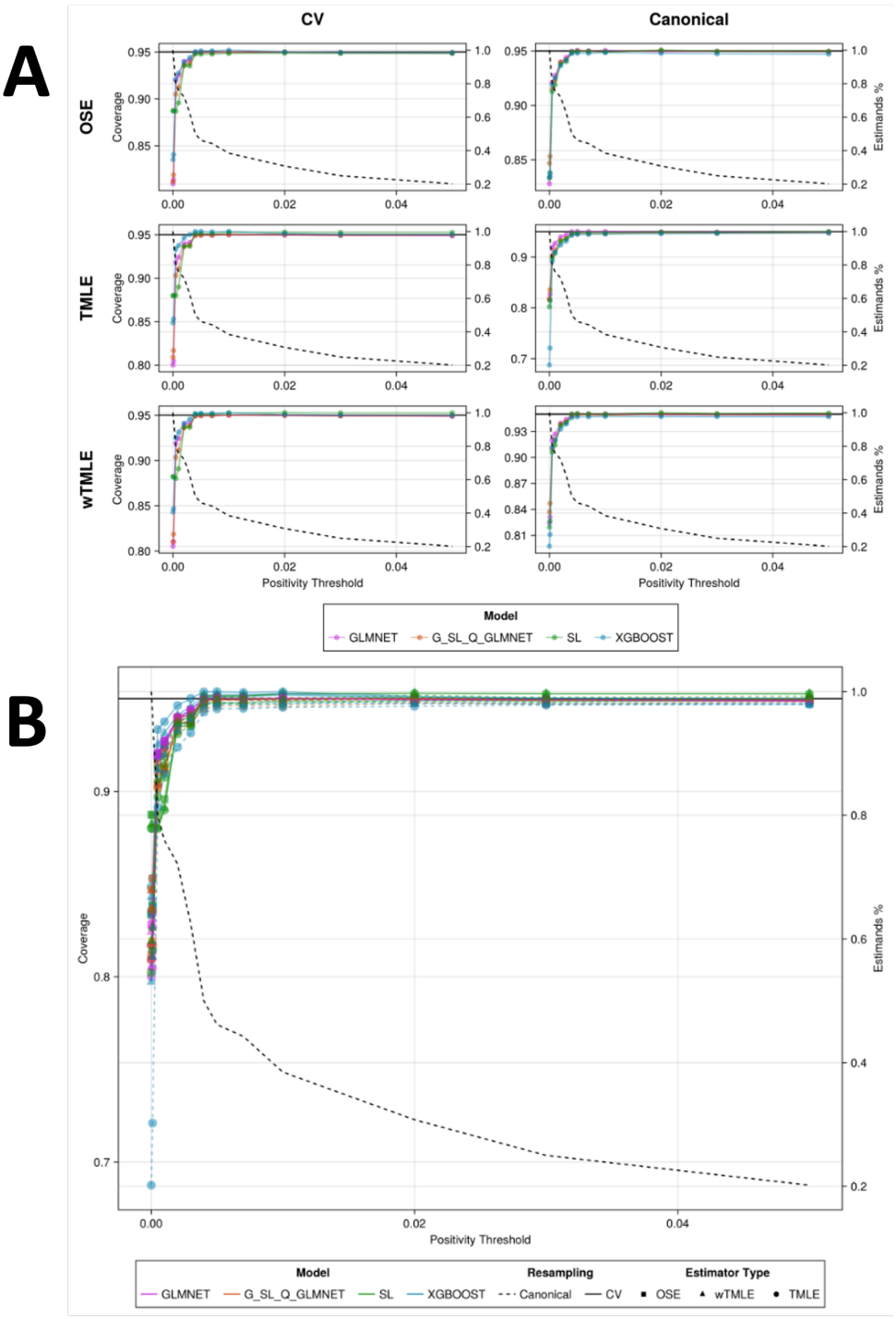

Figure 4.4: **(A) Mean coverage across all estimands and estimators for various positivity thresholds.** Each coloured line corresponds to a different sample size and shows that a positivity threshold of 0.005 is sufficient to reach the nominal confidence level. The black, dotted, and decreasing line, indicates which fraction of the estimands are preserved after application of each threshold. With a threshold of 0.005, around 50% of the estimands are preserved. **(B) Mean coverage across all estimands for various positivity thresholds (sample size $= 500\,000$).** Dashed and solid lines correspond to Canonical and Cross-Validated estimators respectively. Similarly, square, triangle and circle markers correspond to TMLE, wTMLE and OSE. Cross-validated estimators slightly outperform their canonical counterpart.

The cost of the positivity constraint is the reduction of estimable estimands. This cost is represented by the sharply decreasing dashed line in figure 4.4. At the level $\varepsilon = 0.005$, 3 out of 29 joint estimands are fully dropped (10%), and 46% of the total number of one dimensional estimands remain (Estimands Ratio, right y-axis). Fully dropping a joint estimand is more likely to happen when multiple variants are involved. In this case, this is the interaction of (rs10419224, rs59103106) on Multiple sclerosis and the interactions of (rs10132320, rs974766) on psoriasis and L40 psoriasis. For single variant effects, the estimation of changes corresponding to minor-minor $\rightarrow$ minor-major genotypes may sometimes be compromised. However, in most cases, single variant effects and interactions can still be partially investigated (90% of multi-dimensional estimands remain).

4.3.2 Realistic Simulation

We now shift our focus to the results of the realistic simulation, where we assume that the genetic variants influence traits, potentially in complex ways. This complexity is modelled using the SNNE algorithm 1, which we validate first before evaluating the performance of the semi-parametric estimators.

4.3.2.1 Validation of Density Estimates

In section 4.2.2.2, we introduced a general density estimation method leveraging Neural Networks, for both the outcomes and the variants' conditional densities. Despite the known challenges of training Neural Networks and their susceptibility to over-fit [56], we demonstrate that the proposed Sieve Neural Network Estimator (SNNE) consistently outperforms a baseline Generalised Linear Model (GLM) on both training and validation sets. The relative improvement of SNNE over the GLM is illustrated in figure 4.5. For all densities, the training error is always lower for the SNNE as would be expected from a complex model. Furthermore, in most cases, the SNNE also reduces validation error, indicating that early stopping effectively mitigates over-fitting. Exceptions to this improvement occur only with binary outcomes, where the baseline is a logistic regression, a model that already incorporates non-linear terms and benefits from more efficient optimisation algorithms.

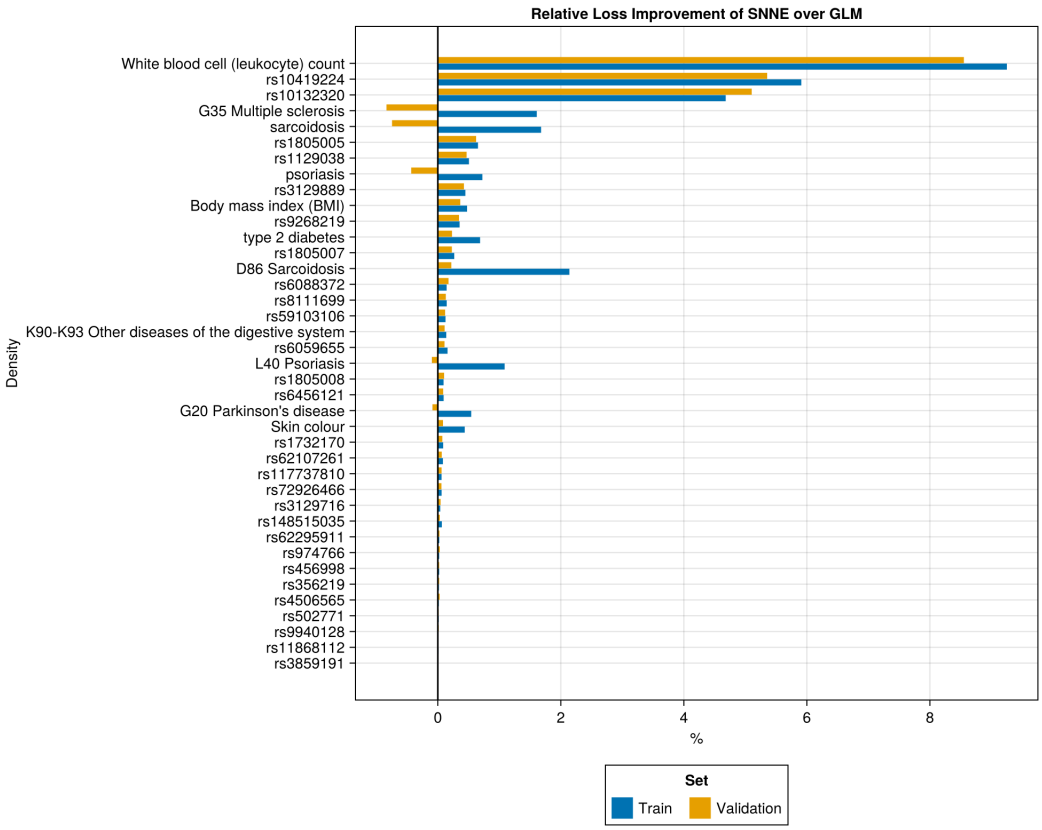

Figure 4.5: Comparison of the empirical loss between the proposed Sieve Neural Network Estimator and a Generalised Linear Model baseline. For each density (y-axis), results are presented as a relative improvement of the SNNE over the GLM (x-axis). Bars facing to the right of the thick 0-line indicate an improvement while bars facing to the left indicate a deterioration of the loss. Both Train (Blue) and Validation (Yellow) set improvements are presented. These results validate the proposed density estimation strategy as an effective flexible and data-adaptive method.

In conclusion, the proposed SNNE approach offers a straightforward yet effective method for conditional density estimation. It provides a flexible, data-adaptive solution for modelling complex data-generating processes while mitigating the risk of over-fitting. It is important to note that we did not provide any statistical performance measures in this analysis; only point estimates were presented. This is because the objective of this analysis was validation rather than statistical model comparison.

4.3.2.2 Coverage Results

Before looking at coverage results, it is important to keep in mind the conditions under which the estimators are guaranteed to provide valid inference at the nominal rate. From section 3.4, these conditions were:

1. Positivity.

2. $\mathcal{L}_2$ convergence in probability of both Q_Y and G at a combined rate of $o_P(\frac{1}{\sqrt{n}})$.

3. Neither Q_Y nor G is too complex, or cross-validation must be used at the estimator level (Resampling=CV).

The case of positivity was discussed extensively in the previous section, especially in the light of finite samples. As expected, figure 4.6 shows that the issue remains in the realistic simulation case. These coverage results are in fact much more contrasted, even after application of a 0.01 positivity threshold. This could indicate that in this case, the required positivity threshold should be higher, or, that conditions (2) and (3) are not always satisfied. Similarly to the null simulation, the coverage results across multiple positivity thresholds, presented in figure 4.7-A, reveal that the same 0.005 threshold is enough to reach the asymptotic regime and is unlikely to cause the issue.

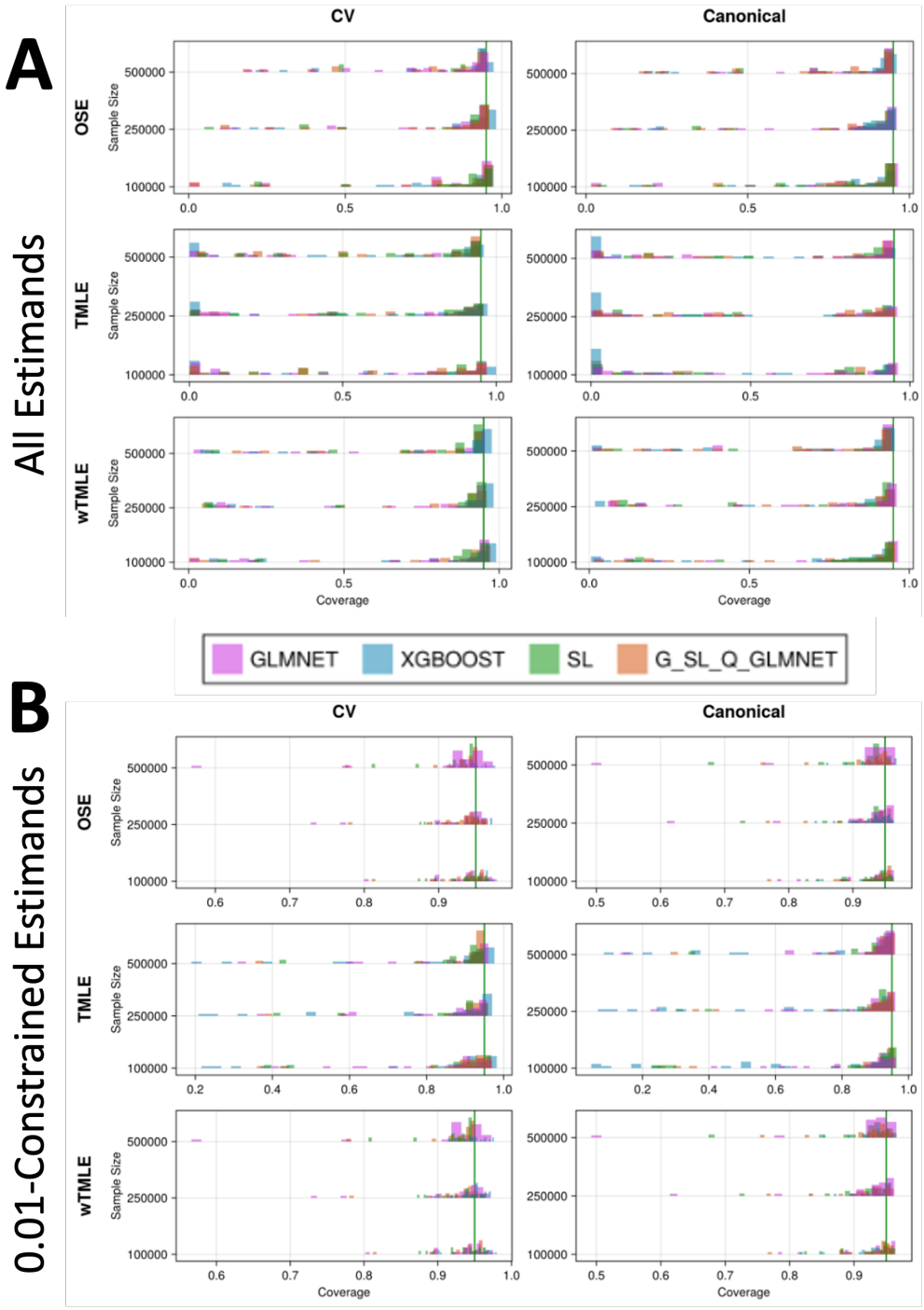

Figure 4.6: **(A) Estimation results across all tasks (Estimands, Estimators, Sample Sizes).** The two columns correspond to either cross-validated (CV) or canonical estimators. Rows correspond to all three variations of semi-parametric estimators (OSE, TMLE, wTMLE). Each histogram represents the mean coverage distribution across all estimands for the given model specification. The figure shows that while most of the mass is concentrated on the expected 95% level, some estimands suffer from low coverage. **(B) Estimation results with 0.01-constrained estimands.** The figure is organised exactly as figure (A) but the estimands are filtered to contain only components that satisfy a 0.01 marginal positivity threshold. The mass is now fully centred on the nominal 95% level with little variations around it (Note the difference in x-axis limits).

One unambiguous conclusion from this simulation is that the TMLE, either in its canonical or cross-validated form, is largely outperformed by the wTMLE and the OSE. Its coverage probability is never better than 0.90%. This phenomenon was also reported in [144], which initially motivated the wTMLE as less sensitive to practical positivity violations. One could wonder why this issue did not manifest in the null simulation. The wTMLE only differs from TMLE in that it fits a weighted loss function with weights $b(T,W) = \frac{1}{p(T|W)}$ instead of including it in the "clever covariate". In the null simulation case, we had by construction, $b(T,W) = \frac{1}{p(T|W)} = \frac{1}{p(T)}$, and thus the values of the weights were much less likely to be extreme and lead to numerical instabilities. Given this comparatively poor performance, we disregard the TMLE in favour of the wTMLE and OSE in the remaining of this chapter.

Regarding conditions (2) and (3), since the data generating process uses neural networks that are never used to learn either Q_Y or G, $\mathcal{L}_2$ convergence is never guaranteed. In other words, the estimators suffer from model mis-specification. This is voluntarily the case since in population genetics, the true generating process is also unknown. Nevertheless, figure 4.7-B, shows that the estimators' coverage probability is still close to the 95% nominal rate, with a slight advantage for XGBoost-based estimators. According to figure 4.5, neural network-based densities largely outperformed their GLM counterparts indicating that the learnt densities are more likely to be captured by XGBoost. The coverage results when a Super Learner is used for G and a GLMNet for Q_Y are also coherent with this analysis. This is since according to condition (2), only one of the algorithm needs to converge fast enough, here supposedly the Super Learner. However, the final case when Super Learning is used for both Q_Y and G surprisingly under-performs in this simulation. Further understanding of this behaviour could be obtained by careful examination of the theoretical properties of the Super Learner, which is beyond the scope of this thesis. Finally, regarding the resampling strategy, cross-validated estimators always outperform their canonical counterpart. This can also be understood by contrasting the complexity of the data generating process (Donsker class) and algorithms used in the estimators. In practice, this is because cross-validated estimators limit over-fitting and tend to have larger variance.

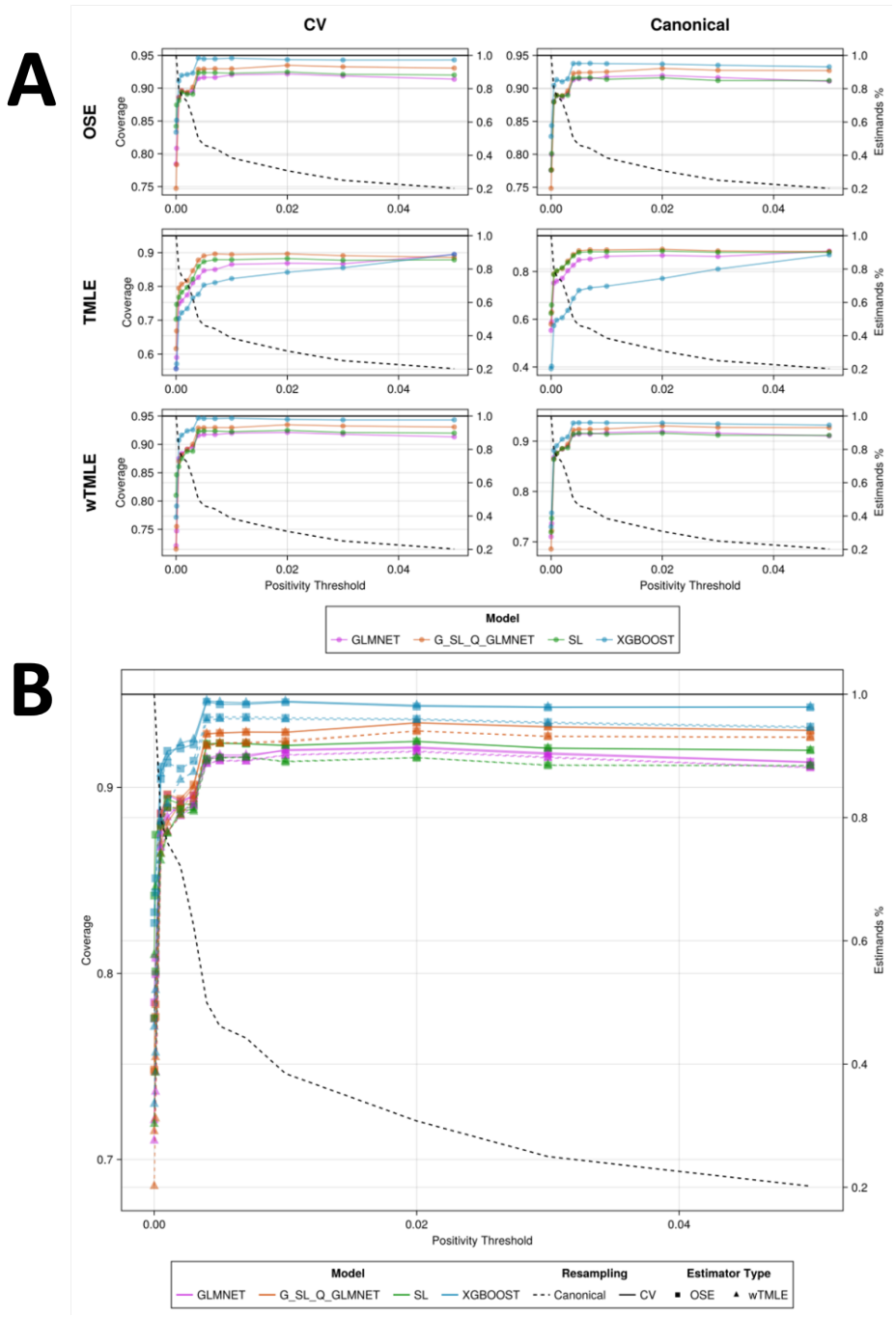

Figure 4.7: **(A) Mean coverage across all estimands and estimators for various positivity thresholds.** The TMLE is largely outperformed by its wTMLE counterpart and the OSE, never reaching more than 90% mean coverage. **(B) Focus on OSE (rectangles) and wTMLE (triangles) (sample size $= 500\,000$).** No noticeable difference can be seen between these two estimators. CV-XGBoost estimators are the top performing methods.

In this section we have seen that the TMLE is a poor estimation strategy in

population genetics studies. The alternative wTMLE and OSE have better coverage probability and are almost indistinguishable. We have also seen that modelling both Q_Y and G using XGBoost leads to almost nominal coverage. However, it is important to note that while not presented on the figures (for readability), the standard deviation on the coverage point estimates is approximately 1.5%. A larger bootstrap sample for some of the top performing estimators would help make the above statements more precise. In the next section we investigate the statistical power of this estimation strategy.

4.3.2.3 Power Analysis

Coverage analyses are particularly informative but not sufficient to evaluate the quality of an estimator. Indeed, an estimator with infinite variance would have perfect coverage but would be poorly informative. Since the variance of an estimator depends on the estimand, it is easier to look at statistical power. That is, the ability to reject the null hypothesis when the alternative hypothesis is true. In population genetics, as discussed in the null simulation section, the null hypothesis is usually that of no effect. In figure 4.8, we present the power of wTMLE using XGBoost in both cross-validated and canonical versions across 0.01-constrained estimands for a sample size of 500 000. Estimands to the left of the dashed line represent Average Treatment Effects (ATEs) and estimands to the right Average Interaction Effects (AIEs). As mentioned in the previous section, because the variance of cross-validated estimators tends to be higher, their power is slightly decreased. Most importantly this analysis shows that the power to detect ATEs is largely superior to that of detecting AIEs. The latter is almost everywhere 0 apart from the interaction between rs1129038 and rs1805008 on skin colour for which the AIE is large (≈ 0.136).

This loss of power is not unique to semi-parametric estimators, genetic interactions are notoriously difficult to detect [5]. To understand why, remember from section 3.4 that the variance of our estimators is given by the variance of the gradient $D_{\Psi,\hat{P}}$. This in turn is inversely proportional to the propensity score $G(\mathbf{V},\mathbf{W}) = P(\mathbf{V}|\mathbf{W})$. For interactions, $\mathbf{V}$ contains at least two variants, say V_1 and V_2. If these variants are conditionally independent given W (unconfoundedness assumption), the variance is inversely proportional to the product $P(V_1|\mathbf{W}) \cdot P(V_2|\mathbf{W})$. For equal effect sizes, achieving the same power for pairwise interactions as for single variant effects thus requires n^2 samples. More generally

for interactions of order p, this grows as n^p. This result is particularly detrimental to the functional genomics paradigm we propose in chapter 7, which is based on interactions.

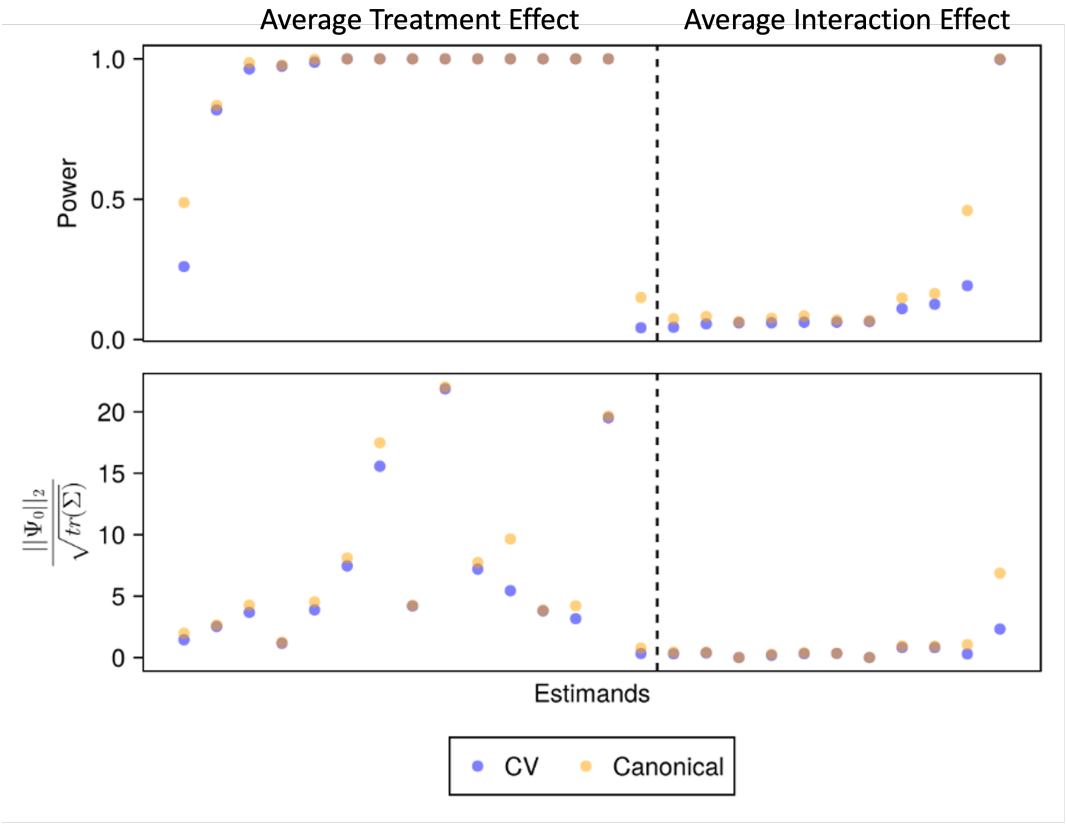

Figure 4.8: **Power Analysis.** Using wTMLE and XGBoost in both cross-validated and canonical versions across 0.01-constrained estimands for a sample size of $500\,000$. Estimands are organised on the x-axis with ATEs to the left of the dashed line and AIEs to the right. The top plot shows that the power to detect ATEs is high while the power to detect AIEs is almost everywhere 0. The bottom plot represents an estimate of the signal to noise ratio for each estimand. Because estimands are multi-dimensional, the signal is captured by the square norm and the noise using the trace of the covariance matrix. As expected, this signal to noise ratio is larger for ATEs.

4.3.2.4 Conclusion

In this chapter we have performed two simulations based on real-world data. Via the null simulation, we have seen that semi-parametric estimators adequately control the false discovery rate at the marginal positivity constraint of 0.005. This threshold was further confirmed in the realistic simulation based on density

estimators and is recommended for all future studies. Regarding estimator types, this simulation also led to the conclusion that the wTMLE or OSE should be preferred against the classic TMLE. Similarly, estimators using XGBoost as a nuisance functions learning algorithm provided better coverage than other evaluated models. A further coverage benefit was observed for cross-validated estimators which was only undermined by a slight loss of power.

In practice, cross-validated estimators induce a further computational burden and their use in large scale studies is unlikely. This is because models need to be fitted multiple times but mostly because it prevents reuse of nuisance functions across traits. The latter point is due to the fact that the cross-validation scheme is stratified across both treatment and outcome variables (section 5.3).

In conclusion, from these analyses, the following recommendation can be made regarding semi-parametric estimation strategies in population genetics:

- Small scale study: CV-wTMLE-XGBoost (or OSE).
- Large scale study: Canonical-wTMLE-XGBoost (or OSE).

using a positivity constraint of at least 0.005.

The simulations presented in this chapter are an integral part of the TarGene software presented in chapter 5. While the results provided are comprehensive, they do not encompass all possible practical scenarios. For example, in this simulation, the rarest trait examined was sarcoidosis, leading to a minimum of only 20 cases in the smallest genotype bin. The findings suggest that such rare events can be safely analysed, but they do not provide insights into even rarer occurrences. This was not explored further due to the lack of theoretical support at this time. The positivity constraint limits only the genotype frequency, not the joint frequency of genotype and outcome. However, some studies have reported a potential decline in performance when investigating rare outcomes [6]. Therefore, it is advisable to either exclude or closely scrutinise rarer events in future research as part of a post-analysis.

Finally, the fields of machine-learning and semi-parametric estimation are rapidly evolving. New machine-learning methods can be implemented and analysed seamlessly within this simulation framework. The highly adaptive lasso represents such an advancement because it is fully non-parametric and satisfies both conditions (2) and (3) [12]. The main problem of this algorithm is that it

does not scale to large datasets. However recent research seems to indicate that a scalable lassoed gradient boosting tree method would benefit from the same properties [2]. New targeted estimators, tailored to population genetics can also be investigated. One extremely interesting direction is the Positivity-Collaborative-TMLE, which proposes to adapt the propensity score model (G) to control for positivity violations [73]. This estimator could lead to an improved positivity constraint threshold and enable the correct estimation of more genetic effects.

Chapter 5

The TarGene Software

5.1 Why TarGene?

At the time of writing, there was no existing software for semi-parametric estimation of genetic effects. Fast software for genome-wide association study like REGENIE [95], GCTA [179], BOLT-LMM [85] exist. However they all rely on strong parametric assumptions like linearity and do not enable the large variety of genetic effects presented in this thesis. For targeted minimum loss-based estimation, tmle3 [31] was the most widely used package and inspired some design ideas present in TMLE.jl. However, it lacks support for higher-order interactions, is not tailored for population genetics, and is not implemented in a high-performance programming language. Given the scalability requirements of population genetics analyses, this gap was a true opportunity for me to contribute to the open-source community, as part of my PhD. The result is the TarGene software, a set of Nextflow [38] workflows enabling the analysis of genetic effects from a semi-parametric lens. Nextflow is a workflow management system that enables the scalable and reproducible execution of data analysis pipelines, with built-in support for parallelization, containerization, and cloud computing. In this chapter, we describe the software from multiple standpoints. We first describe TarGene from a user perspective including the user interface and provide some benchmarks as well. We then show how TarGene adheres to high-quality software standards, both from a development and design point of view. We end the discussion with opportunities for future developments.

5.2 User Interface

As part of each TarGene software release, an extensive versioned documentation is also available online at: TarGene Documentation. This documentation describes all workflows available in TarGene and their associated parameters as well as best practices and tutorials. Since TarGene is a set of Nextflow pipelines, the user interface is simply given by the Nextflow user interface. The most commonly used is the command line interface, which enables both the management and execution of pipelines. While we refer to the Nextflow Documentation for detailed information, we briefly present here some TarGene recommendations for users. A typical TarGene run is given by the following command-line

```
nextflow run https://github.com/TARGENE/targene-pipeline/ \
        -r VERSION \
        -entry WORKFLOW_NAME \
        -profile PROFILE \
        -c RUN_CONFIG
```

where

1. `VERSION` describes the required TarGene version (e.g. v0.10.0), if left unspecified, the latest development version is used.
2. `WORKFLOW_NAME` is the specific pipeline that will be executed (see section 5.2.2).
3. `PROFILE` is the platform specific configuration (see section 5.2.1). If no profile currently exists in TarGene for the platform, this configuration can be provided with an additional file via `-c PLATFORM_CONFIG`.
4. `RUN_CONFIG` is the run specific configuration for the scientific question of interest (see section 5.2.2).

This means that instead of installing TarGene, users install the latest Nextflow version instead. TarGene will then be available to them as well as other Nextflow pipelines. TarGene also depends on containerisation technologies and supports both Docker and Singularity. This choice depends on the specific platform where TarGene will be run and high-performance computing platforms usually rely on Singularity.

Note that the previous command does not provide any parameter, this is because these parameters are workflow specific, and we recommend to provide them within configuration files that can be version controlled. For clarity, we recommend to split these configuration files into two separate files.

5.2.1 Platform Specific Configuration

The first configuration file describes the platform onto which the workflows are intended to be executed, and can be shared across all runs. Nextflow supports an extensive list of High-Performance-Computing (HPC) platforms including custom platforms, cloud providers or local execution. For University of Edinburgh users, we provide a ready to use configuration (`-profile eddie`) for the Eddie Cluster and no file is needed. The All of US workbench is also supported natively and can be opted in by using the `-profile allofus` option.

5.2.2 Run Specific Configuration

The second file corresponds to the actual run, or scientific question of interest and is usually contained within a `.nextflow.config` file in the run directory. If this is the case, the `-c RUN_CONFIG` can be omitted since Nextflow will automatically use the `.nextflow.config` file. This file specifies the workflow's specific parameters (identified by `WORKFLOW_NAME`) and point to relevant additional data files. Since TarGene is likely to evolve in the future we point to the Workflow Documentation, for up-to-date information on available workflows. Here, we briefly describe the two main workflows available at the time of writing: the discovery workflow and the simulation workflows.

Discovery Workflow. Defined by `WORKFLOW_NAME=TARGENE` (or simply `-entry` omitted). This is the principal workflow and is used for the estimation of genetic effects. The configuration file provided by a user for a GWAS could look like the following:

```
params {
    // Definition of estimands
    ESTIMANDS_CONFIG = "gwas_config.yaml"

    // UK-Biobank data
    BED_FILES = "unphased_bed/ukb_chr{1,2,3}.{bed,bim,fam}"
    TRAITS_DATASET = "dataset.csv"
}
```

Hence, the main ingredient required by TarGene is the description of the estimands that need to be estimated. There are currently 4 main ways to define these estimands, all via the `ESTIMANDS_CONFIG` parameter pointing to a separate YAML file. The way estimands are generated is defined by the `type` value at the top level of the YAML file.

1. The `type=gwas` mode will estimate the effect of all variants in the separately provided genotype files on the traits of interest.

2. The `type=flat` mode will estimate the effect of all variants provided in the `ESTIMANDS_CONFIG` on the traits of interest.

3. The `type=group` mode will estimate the joint or interaction effect of all variants, organised in groups, and provided in the `ESTIMANDS_CONFIG` on the traits of interest. This mode is particularly useful to investigate the genetic effects mediated by transcription factors as illustrated in chapter 7.

4. Otherwise, the `ESTIMANDS_CONFIG` is assumed to provide a list of fully defined estimands.

For a GWAS the `ESTIMANDS_CONFIG` file could thus simply contain:

```
type: gwas
```

Simulation Workflows. Defined by `WORKFLOW_NAME=NULL_SIMULATION` or `WORKFLOW_NAME=REALISTIC_SIMULATION`. These workflows enable the evaluation of semi-parametric estimators in a controlled environment where the data generating process is known. This can be for the evaluation of a specific estimator, estimand, sample size or combination thereof. These two types of simulations are described further in chapter 4. In this case, the `ESTIMANDS_CONFIG` must contain a list of fully defined estimands as described above 4.

5.3 Benchmark

Semi-Parametric estimation relies on computationally intensive machine-learning algorithms. However, performance is critical for routinely conducted large scale genetic studies like PheWAS or GWAS that seek to perform hundreds of thousands of association tests. We show below that, access to modern computing resources, makes TarGene possible in those large scale settings. In our case, all runs were performed on the Edinburgh high-performance Eddie cluster. In this section we investigate TarGene's runtime for the two most common types of genetic studies: GWAS and PheWAS. In both cases we are thus computing the Average Treatment Effect for each individual variant on trait by comparing the major/minor to the major/major genotype. Covariates were set to include the first 6 principal components, age and sex. Finally, the benchmark is conducted on a dataset of 500,000 samples, similar in scale to the UK Biobank, using a dual-core compute node representative of a modern laptop. Remember from section 3.4, that for the estimation of an Average Treatment Effect, we need to estimate two nuisance functions: the outcome mean Q_Y and the propensity score g. Computational complexity is thus dictated by the choice of machine-learning algorithms employed for these nuisance functions. Here, we investigate the computational performance of the canonical targeted minimum loss-based estimator for 4 nuisance functions' estimation strategies from the most naive to the most comprehensive.

- GLM: Standard generalized linear model
- GLMNet: GLM with regularization hyperparameter tuning over 3-folds cross-validation.

- XGBoost: The gradient boosting trees method with hyperparameter tuning over 10 different settings in a 3-folds cross-validation scheme.
- SL: Super Learning including both the previous XGBoost and GLMNet combined with an outer 3-folds cross-validation.

We first focus on the PheWAS setting for which runtime estimates are provided in Table 5.1 for 1 variant and 768 traits. It is important to emphasise that in a typical PheWAS the estimation of the propensity score, g, only needs to be performed once and can be re-used across all traits. The computational complexity is thus driven by the estimation of each regression, Q, and associated targeted steps. We also note that the same remark holds for the targeting steps corresponding to the various genetic changes. Computing the effects of the additional major/minor $\rightarrow$ minor/minor and major/major $\rightarrow$ minor/minor would only cost two additional targeting steps while re-using the current $\hat{Q}$. In all cases, running a PheWAS using TarGene is possible even without access to a high-performance computing platform.

Learning Algorithm	Time (hours)
GLM	2.2
GLMNet	4.5
XGBoost	8.8
SL	30

Table 5.1: PheWAS runtime for various nuisance functions' estimation strategies.

We now turn to the GWAS setting for which runtime estimates are provided in Table 5.2. In this case, Q and g need to be estimated for each variant. In order to obtain a run time estimate for a GWAS it is thus sufficient to compute the run time for one variant and simply multiply by the number of genotyped variants in a typical GWAS; here we take 600 000. However, because the runtime of the propensity score fit varies depending on the variant, we instead run the TMLE process over 100 variants and report the mean run time as a more accurate estimate. We find that, while it would be impossible to run a GWAS on a personal laptop, access to a modern computing platform makes this kind of study feasible using TarGene. On the Eddie cluster, the user limit defaults to 1000 processes, the GWAS time is thus only $\frac{1}{5}$ of the projected time in table 5.2.

Learning Algorithm	Unit Time (seconds)	Projected GWAS Time 200 cores (hours)
GLM	13	10
GLMNet	57	48
XGBoost	95	72
SL	451	375

Table 5.2: GWAS runtime. The unit time corresponds to a single variant/trait pair. The projected GWAS time assumes 600 000 variants and 200 folds parallelization.

The one-step estimator offers a more efficient alternative, particularly in scenarios with limited computational resources, as it bypasses the need to fit a generalised linear model during the targeted step. This reduction in computational overhead makes it a preferable choice when resources are constrained.

However, when employing cross-validated estimators, challenges arise with reusing nuisance functions across different estimands. The primary complication stems from the joint outer cross-validation scheme, which is typically stratified to manage the imbalance among variants and outcomes. This stratification process makes it impractical to reuse nuisance functions across multiple outcomes, adding to the complexity and resource demands of the estimation process.

Compared to existing software capable of running a GWAS in just a few minutes, the approach taken in this thesis incurs a significantly higher computational cost. This increased burden also results in greater carbon emissions, an undesirable side effect. Future efforts should prioritize finding computational shortcuts, similar to the advancements made for linear mixed models in recent years (see chapter 2). Additionally, most of my development time was dedicated to ensuring algorithmic correctness rather than optimizing runtime, leaving ample opportunity for further efficiency improvements, some of which are discussed in Section 5.6.

5.4 Development Methodology

5.4.1 The Agile Philosophy

Scientific software development is highly exploratory and a risky endeavour by nature. Agile principles promote flexibility, collaboration, and continuous improvement, which are essential in rapidly changing environments such as scientific

research. The development of TarGene followed the Agile Manifesto [47] which reads

- Individuals and Interactions over processes and tools.
- Working software over comprehensive documentation.
- Customer collaboration over contract negotiation.
- Responding to change over following a plan.

These principles do not mean that the items on the right have no value, simply that the items on the left are prioritised by the method. More precisely, as I was the only developer, I followed a framework called Extreme Programming (XP) which is centred around simplicity, communication and consistent feedback. The project was developed in multiple incremental iterations, of no more than two weeks, using the general cycle of Figure 5.1.

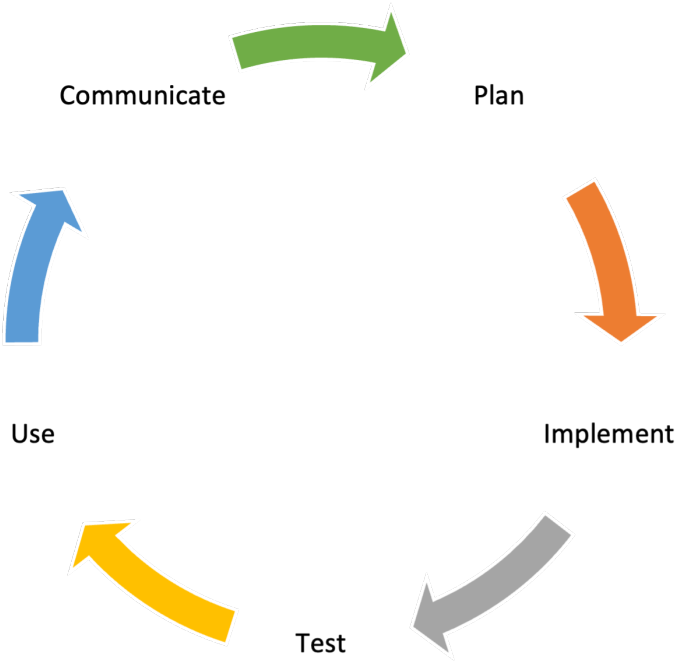

Figure 5.1: The Software Development Cycle. Each cycle lasts for around two weeks. Weekly team meetings were used for communication and planning. TarGene was used very early on in the development cycle, allowing for rapid feedback.

Part of TarGene's development success was due to the early adoption of the software across the lab. Its continuous use enabled rapid feedback and enhancements to be made within each new release.

5.4.2 Programming Practices

In order to improve code quality and increase efficiency, TarGene was built with modern coding standards in mind. We highlight here the three main standards that have guided the development of the software: testing, version control and continuous integration and deployment.

5.4.2.1 Testing

Software testing is the process used to evaluate the functionality, performance, and quality of a software application or system. It is an essential component of the software development life-cycle, that helps to identify and mitigate risks. It is particularly relevant in complex scientific systems, where the validity of results is difficult to verify. In TarGene, there are currently three main levels of tests that ensure the robustness of the system: unit tests, integration tests and end-to-end tests. These testing levels can be seen as ordered within a hierarchy (figure 5.2). Unit tests are typically found at the lowest level, they test individual functions within a module. Integration tests verify that the different parts of a module work well together. Finally, end-to-end tests replicate a user behaviour with the system. However, since TarGene is organised in multiple modules (section 5.5.1), the frontier between the various levels can become fuzzy. For instance an end-to-end test at the module level (e.g. in TMLE.jl) could be viewed as a unit test at the targene-pipeline level.

In order to clarify, we now take the example of TMLE.jl, the module providing routines for semi-parametric estimation, and refer back to section 3.4 for mathematical details.

5.4.2.1.1 Unit Test There are many low level functions in TMLE.jl, but perhaps a good example for a unit test is the so-called "clever covariate" which plays a pivotal role in targeted minimum loss-based estimation. The Julia code for it can be found in appendix A.1.1. The function takes as an input an estimand (Ψ), a propensity score (Gs), a dataset, two keyword arguments and outputs the cor-

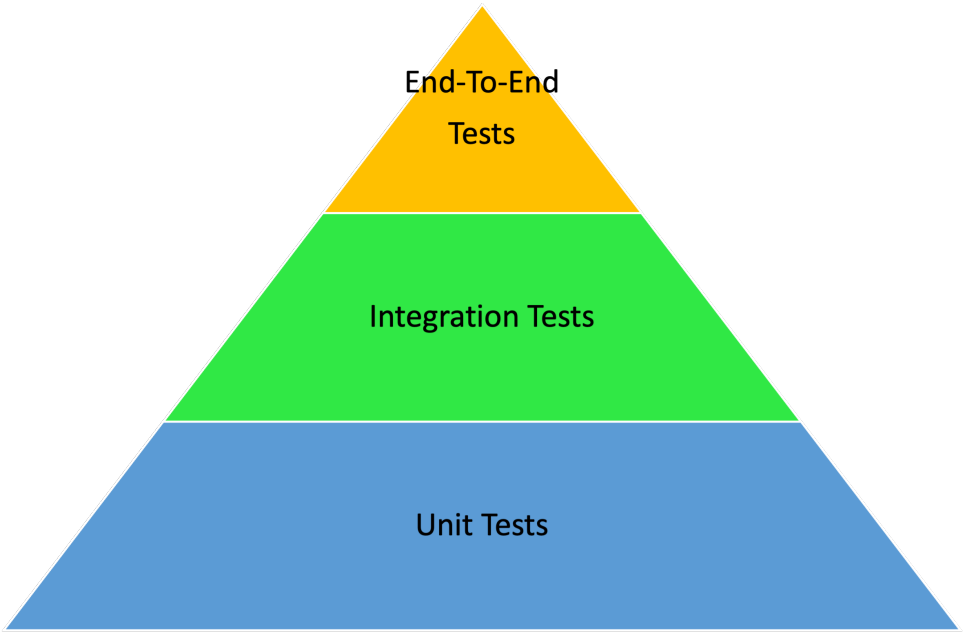

Figure 5.2: The Testing Hierarchy. Unit tests are found at the lowest level while end-to-end tests make sure the whole system works correctly. In between, are integration tests, ensuring correct function across processes.

responding "clever covariate" and associated "weights". In particular, depending on the type of Ψ and the weighted_fluctuation keyword argument, the outputs will differ. We illustrate with the Average Treatment Effect and one treatment variable for which the "clever covariate" and "weights" are given by:

$$\begin{cases} \text{if weighted\_fluctuation} = \text{false} & , \begin{cases} H(W,T) & = \frac{(-1)^{T==b}}{G(T|W)} \\ W(W,T) & = 1 \end{cases} \\ \text{if weighted\_fluctuation} = \text{true} & , \begin{cases} H(W,T) & = (-1)^{T==b} \\ W(W,T) & = \frac{1}{G(T|W)} \end{cases} \end{cases} \tag{5.1}$$

which is exactly the purpose of the set of tests in appendix A.1.2, which uses a constant mean model for G to keep the output simple and deterministic.

5.4.2.1.2 Integration Test Building on the unit test present above, the "clever covariate" is typically used to fluctuate an initial machine learning model (appendix A.2.1). By definition, the fluctuated model is obtained by minimising the loss corresponding to the outcome of interest (for continuous variables this is the mean-squared error).

Because the initial model, corresponding to the value $\varepsilon = 0$, is a possible

solution of this optimisation problem, the training loss of the fluctuated model can only be smaller than that of the initial model. This is an example of a test, among others, that is currently in place and is shown in appendix A.2.2.

5.4.2.1.3 End-to-End Test We finally turn to an example of an end-to-end test. As described in section 3.4, the semi-parametric estimators under consideration satisfy the double-robustness property. For any two machine-learning models (Q, G), only one needs to be correctly estimated to provide valid inference. This is considered an end-to-end test because a TMLE.jl user would create an estimator, run it on some data and expect coverage of the ground truth. This property can be tested in practice using simulation data. Consider the following data generating process:

$$\begin{aligned} W &\sim \mathcal{N}(0,1) \\ T &\sim \mathcal{B}(\frac{1}{1+e^{1+W}}) \\ Y &\sim \mathcal{N}(\alpha+\beta\cdot T+\gamma\cdot W, \sigma^2) \end{aligned} \tag{5.2}$$

If we let $G(W) = p(T|W)$ and $Q(W,T) = \mathbb{E}[Y|W,T]$, a constant mean model for G or Q would be misspecified but a logistic or linear regression, respectively, would be correctly specified. Testing each of these two scenarios where only one of G or Q is correctly specified is part of TMLE.jl's test suite. The code, too long to be included here is omitted, but available here.

5.4.2.2 Version Control

Version control, is a system that manages changes to documents, files, or code over time. It allows multiple contributors to collaborate on a project by tracking modifications, preserving the history of changes, and facilitating coordination among team members. Git has emerged as a leader in version control because of its decentralised architecture and branching mechanism. For instance, a contributor can make changes to any part of the TarGene code base without ever affecting the past releases. Finally, Git also integrates seamlessly with text editors and remote code storage platforms such as Github which makes it effortless to use in practice. The basic workflow is presented below in Figure 5.3.

In TarGene, we follow semantic versioning, which is a convention specifying that given a version number MAJOR.MINOR.PATCH, one should increment the:

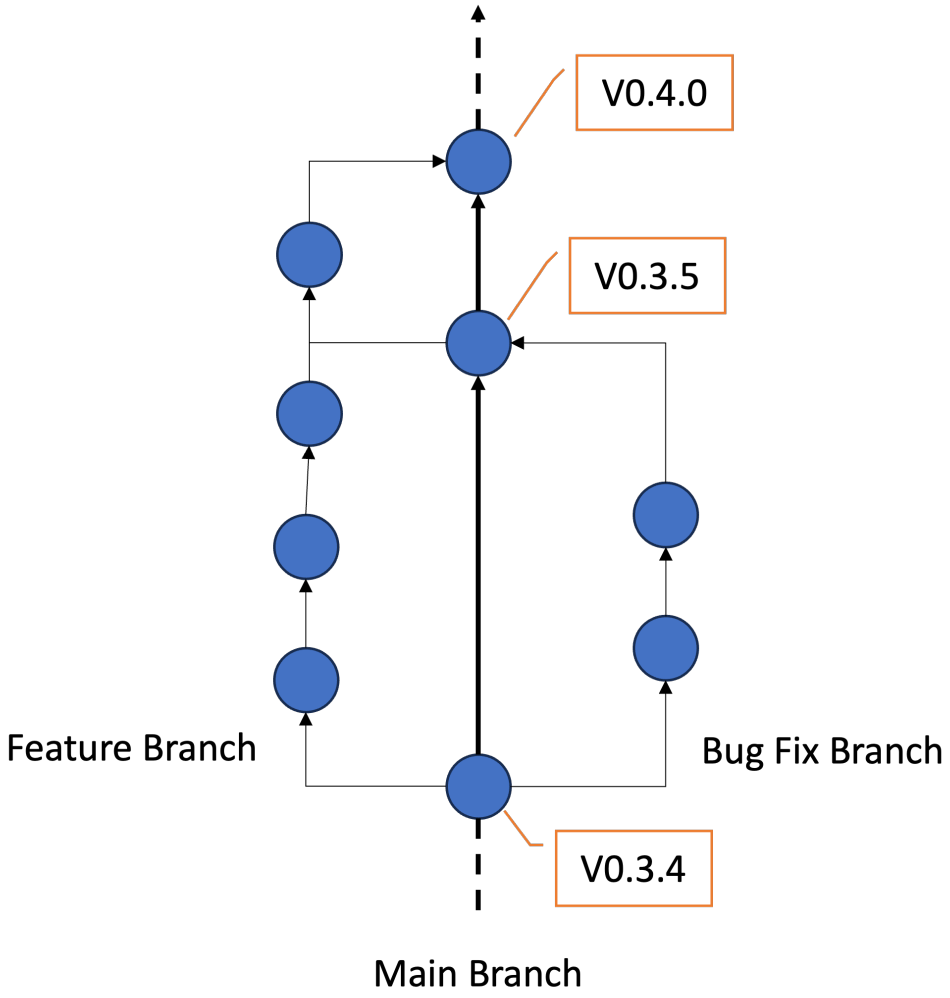

Figure 5.3: Git Workflow. Lines represent branches, circles code changes committed that branch. Specific commits on the main branch are annotated with a tag release version. These versions of the code are eventually downloaded and installed by users.

1. MAJOR version when making incompatible API (public functionalities available to users) changes
2. MINOR version when adding functionality in a backward compatible manner
3. PATCH version when making backward compatible bug fixes

This convention is particularly useful to indicate to downstream users and software that depend on your on a package, whether they can update their dependencies without compromising their own results. For example, at the time of writing TargetedEstimation.jl@v0.8.3 depends on TMLE.jl@v0.16. A bug fix or new feature in TMLE.jl would lead to a new minor or patch version, say TMLE.jl@v0.16.1, indicating that TargetedEstimation.jl@v0.8.3 can safely use this new version. However, if an API change was made in TMLE.jl, resulting in a TMLE.jl@v1.0.0, TargetedEstimation.jl@v0.8.3 would never use this version of TMLE.jl because it could result in unexpected behaviour.

Since TarGene is a Nextflow pipeline, any change to the pipeline parameters (e.g. name change or deletion) would be considered a MAJOR change. A new pipeline parameter providing a new functionality would be considered a MINOR change. Finally, note that this convention only applies from version $\geq$ 1.x.y, 0.x.y version are considered unstable and anything may change with any increment. However in TarGene, while version 1 has not yet been reached, we still implicitly follow the stable version semantics for added transparency.

5.4.2.3 Continuous Integration and Continuous Delivery

Testing and versioning are powerful tools that can be leveraged even further via automation. Continuous Integration (CI) and Continuous Deployment (CD) are practices used in software development to automate and streamline the process of building, testing, and deploying code changes. The process aims at improving the quality of the software as well as reducing the burden of manual release and deployment. Since TarGene is organised in modules, a TarGene release can be comprised of:

- Changes to the pipeline's code-base
- One or multiple modules releases

- Both

A module release corresponds to the build of a Docker image containing the functionalities of that module (typically a command-line interface). If the image build is successful, the image is published on Docker Hub with a tag respecting semantic versioning. This means that a module patch does not require an update at the pipeline level which will automatically use the most recent compatible version.

For a major module update or a new feature, end-to-end pipeline tests are executed using the Docker images specified by the pipeline version. If successful, a new TarGene release tag is created and the new version is immediately accessible to users. In practice, the entire CI/CD, presented in figure 5.4 is described by a set of YAML files and further managed by Github Actions.

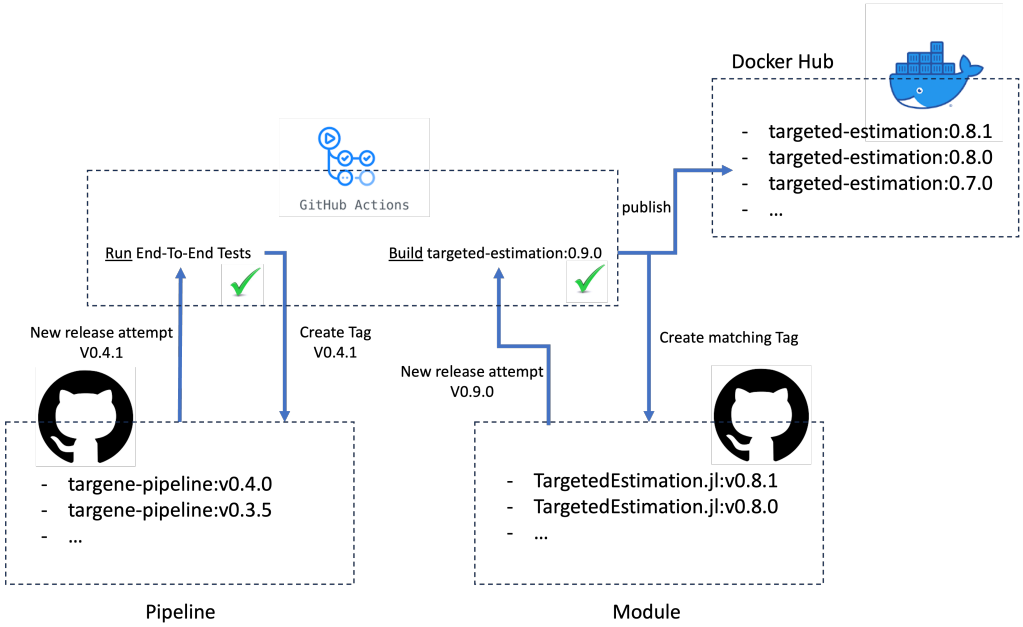

Figure 5.4: TarGene's Continuous Integration and Continuous Deployment process. Semantic versioning is used within the pipeline as well to minimise continuous testing burden. The entire process is handled through Github Actions.

5.5 TarGene's Design

When creating new software, several guiding principles can help ensure its success and effectiveness. The FAIR (Findable, Accessible, Interoperable and Reusable)

principles, initially proposed for data management [174] were later adapted to software [8] with the goal to maximise the value of research outputs. Before diving into technical specifics, we briefly outline how TarGene adheres to these principles.

- **Findable**: TarGene is entirely open source and hosted on Github within the TarGene organisation ensuring it is assigned a unique global identifier. The complete history of software versions along with associated documentation is readily accessible and easily searchable.

- **Accessible**: TarGene can be accessed via standard protocols and necessitates minimal dependencies (Nextflow and Docker or Singularity), which are ubiquitous on modern High-Performance Computing platforms where the software is intended to be utilised.

- **Interoperable**: TarGene operates seamlessly with raw data sources and produces outputs in standard formats. For instance, in the case of the UK Biobank, it directly employs the data downloaded from the website. Users have the flexibility to choose output formats; currently, JSON and/or HDF5 are available and can be further processed using external software.

- **Reusable**: TarGene is distributed under the MIT license and adheres to community standards, facilitating re-usability and reproducibility.

5.5.1 Hierarchical Design

TarGene was designed using a top-down, hierarchical approach. This methodology begins with a broad overview of the system and gradually decomposes it into smaller, more manageable components. At the highest level lies the targene-pipeline, which serves as the primary interface for user interaction. Rather than performing specific tasks on its own, this component specifies how various modules should interact and under what environment they should operate. A graphical presentation of the software's architecture is presented in figure 5.5.

Following the hierarchical approach it became apparent that a module for semi-parametric estimation was necessary. This led to the development of the TMLECLI.jl module from which the TMLE.jl package was later extracted and officially released within the Julia registry. This is because TMLE.jl is compatible

with any data table source and can be used outside of TarGene. One of the benefits of this hierarchical approach is its flexibility for iteration and refinement. For example, the initial scope of TMLECLI.jl was limited to targeted minimum loss-based estimation and was later extended to include one-step estimation.

A second module, UKBMain.jl, focuses on extracting data from the UK Biobank. As the UK Biobank data is relatively stable, this module has undergone only minor revisions throughout the development process. Additionally, the modular approach makes it straightforward to create new modules to support other biobanks or data sources. For instance, a PhD student in the lab later added support for custom cohorts, which required only minimal modification of the pipeline's code organisation.

TargeneCore.jl provides all the necessary quality control and data preprocessing necessary to transform raw datasets into a high quality tabular format that can be handled by TMLECLI.jl. It is also responsible for the generation of estimation inputs such as the validation of the estimands and batching, hence ensuring efficient parallelisation.

The latest module, Simulations.jl, builds on previous developments to enable the controlled evaluation of semi parametric estimators in various population genetics contexts.

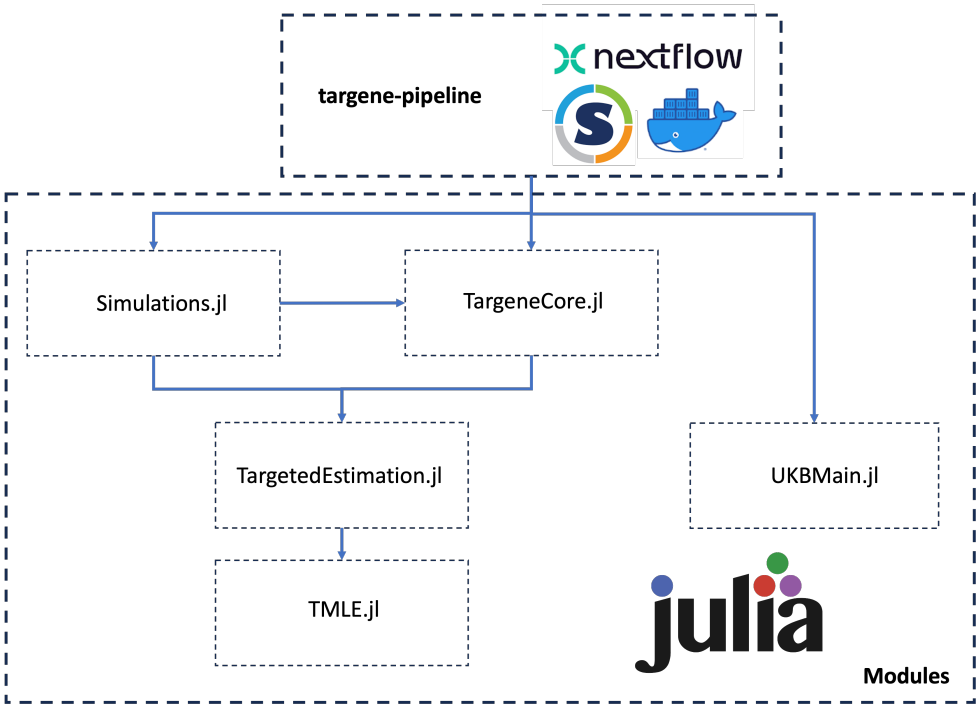

Figure 5.5: TarGene's Architectural Design. TarGene is a Nextflow pipeline using Docker or Singularity to containerise and execute the functionalities provided by the dependent modules. Each module is itself a plain Julia Package and an associated command-line interface.

5.5.2 The Technological Stack

When designing a software, technological choices must be made. These choices can be constrained by the knowledge and skills of the development team but are ideally motivated by how well they address the software's requirements.

The primary technological decision revolved around selecting the technology for the user interface. Workflow management systems offer many advantages that integrate perfectly with the FAIR principles and the hierarchical approach to software design. Nextflow [38] was a natural choice as it is a standard in the computational biology community and was already in use in the lab. Its main features are:

- **Portability**: It is platform-independent, allowing workflows to be executed consistently across different computing environments, including local machines, clusters or cloud platforms.
- **Scalability**: It enables seamless scaling of workflows to handle large-scale

data processing tasks. They automatically parallelize workflow tasks across available compute resources, leveraging multi-core CPUs, distributed computing clusters, or cloud computing instances to accelerate computation and improve throughput.

- **Fault Tolerance**: It handles failures, errors, and interruptions gracefully during workflow execution. For instance, they automatically retry failed tasks and resume execution from checkpoints.

- **Reproducibility**: It encapsulates the workflow logic, dependencies, and execution environments in a version-controlled manner. Workflows can be easily shared, and rerun with consistent results, facilitating collaboration, validation, and replication of research findings.

- **Integration with Containers**: It integrates with containerisation technologies such as Docker and Singularity, enabling workflows to be executed within isolated and reproducible software environments. Containers encapsulate workflow dependencies, software tools, and runtime environments, ensuring consistency and compatibility across different computing platforms.

Because we rely on containerisation, we have the flexibility to employ any technology or programming language for each autonomous module within TarGene. Given the computationally demanding nature of Causal Inference methods, an optimal choice would be a high-performance language like C or C++. However, rapid prototyping was also a major project constraint, essential for swift progress and risk mitigation. Hence, Julia emerged as a natural fit and was adopted uniformly across all modules.

5.6 Discussion

TarGene is a novel software designed for the estimation of genetic effects on human health outcomes. Some of the key features of TarGene are:

1. Minimal modelling assumptions. The software draws from a robust mathematical framework, leveraging methods from machine-learning and semi-parametric statistics. Under minimal assumptions, the reported estimates

are guaranteed to be asymptotically unbiased for the statistical estimands they target.

2. Support for complex interactions. The software is capable of estimating higher-order interactions between genetic variants, offering a more nuanced understanding of how multiple genetic factors contribute to complex traits.

3. Scalability. TarGene is designed to be highly scalable, with support for high-performance computing environments, which is essential for analysing large datasets.

The software's flexibility mirrors the adaptability of the underlying mathematical estimation framework it employs. This design ensures that the implementation of new estimands and their corresponding estimators is highly feasible in the near future. One estimand of particular interest to geneticists is heritability, which quantifies the proportion of phenotypic variation attributable to genetic factors relative to the total phenotypic variation. Traditionally, heritability is defined within the context of linear models, but integrating this concept into TarGene would require developing a non-parametric definition along with an appropriate estimator. This extension would not only enhance the software's utility for genetic research but also broaden the range of genetic effects that can be accurately estimated.

One limitation of the software is that it does not fully address the causal gap identified in equation 3.1. It currently relies on Principal Component Analysis (PCA) to model latent confounders, which may not capture the full complexity of the genetic data. To more effectively reduce this causal gap, more advanced methods such as Hidden Markov Models or Variational Auto-Encoders could be used. These methods offer a richer representation of the underlying genetic structure, leading to more accurate estimations of causal effects. Importantly, these models generate a probability distribution over genetic variants, and thus, propensity score models. This, could streamline the downstream estimation process, reducing computational costs.

Additionally, instead of estimating a marginal outcome model for each estimand separately, the use of large multivariate models that encompass multiple variants could be advantageous. These models could be estimated once and stored in a registry, allowing for rapid marginalisation to obtain the specific outcome models needed for each estimand.

Such innovations not only enhance computational efficiency but also move closer to achieving full backdoor adjustment, thereby enabling more robust causal inferences.

Chapter 6

Validation and Evaluation with the UK Biobank

To investigate further the implications of our proposed statistical method, we performed a series of analyses using the UK Biobank and restricting the population to white individuals to be consistent with other studies such as the geneATLAS [21].

First, we contrast our approach with the gold standard Linear Mixed Model's method by performing a phenome-wide association study for a well studied variant in the FTO gene region: rs1421085. Second, we reveal gene by environment interactions between rs1421085 and two deprivation indices. Third, we replicate pairwise genetic interactions previously reported for hair colour [105] and report additional evidence of interactions for both skin and hair colour. And fourth, we show how TarGene can investigate higher-order interactions using multiple loci related to the vitamin D receptor (VDR).

6.1 Estimators

For all analyses in this chapter, we used the 3 canonical estimators presented in section 3.4, that is, TMLE, wTMLE and OSE.

The nuisance functions Q and G were estimated using the Super-Learning strategy discussed in section 3.4.6. This is because this real-world study was performed prior to the recommendations resulting from the simulations of chapter 4. Note however, that while maybe not optimal, Super-Learning is still well supported by our simulation results and is the recommended strategy to mitigate model mis-specification [158]. We used the following Super-Learning speci-

fications: (i) k-fold cross-validation or stratified k-fold cross-validation based on the outcome type (continuous or binary, respectively), here $3 \leq k \leq 20$, selected adaptively based on the rarest class of each outcome [120], and, (ii) included the constant fit, a regularized logistic/linear regression (ridge, $\lambda = 1$), a gradient-boosted tree (`n_round` $= 100$, default parameters otherwise), and HAL with hyper-parameters `max_degree` $= 1$, `smoothness_orders` $= 1$, `lambda` $= 30$ [79], as base learners. However, we note that for the optimal performance of HAL in more bespoke analyses, the parameter λ, tuning the total variation norm of the fit, should be left unspecified so that it is chosen by the algorithm's internal cross-validation.

As per section 3.2.3.1, for confounding adjustment, we used the first 6 PCs computed from the genotypes. We also added the age and the sex of each individual as explanatory variables for the outcome model $\hat{Q}$.

To investigate the impact of Sieve Variance Plateau correction (section 3.6), variance estimates were corrected using 100 genetic similarity thresholds.

Finally, significant results are reported after appropriate multiple hypotheses adjustment using the Benjamini-Hochberg method.

Full run configurations details can be found online.

6.2 Phenome-Wide Association Study

In order to investigate how our method compares to current gold standards Linear Mixed Model's (LMM), we performed a phenome-wide association study (PheWAS) using UK Biobank data. We chose a well studied variant, rs1421085, located in the first intron of the *FTO* gene. For this variant, the T to C nucleotide substitution has been predicted to disrupt the repression of *IRX3* and/or *IRX5*, thereby leading to a developmental shift from browning to whitening programs and loss of mitochondrial thermogenesis [29]. This variant has also been associated with several related traits such as BMI and obesity [48].

Remember from section 3.3.2, that the effect of a variant on trait is defined as the Average Treatment Effect of a genotype change. Since rs1421085 has two different alleles, the three possible genotype changes we estimate are: TT $\rightarrow$ TC, TC $\rightarrow$ CC and TT $\rightarrow$ CC. Furthermore, this definition naturally suggests another question, is the effect of the first C substitution (TT $\rightarrow$ TC) is equal to the effect of the second substitution (TC $\rightarrow$ CC)? In section 3.3.3, we showed that such

a question can be answered via the Allelic Effect Difference, and prior evidence was reported for rs1421085 in previous research [178].

In order to evaluate our method, we compare it to LMM results published in the geneATLAS [21] for 768 traits. Unfortunately, large scale studies based on LMM usually neglect allelic effect differences, assuming the two genotype changes have the same effect, and hindering a direct comparison. As a workaround, we compare the geneATLAS results to our TT $\rightarrow$ TC estimate, noting that they would indeed be comparable if the Allelic Effect Difference is 0.

6.2.1 Comparing Effect Sizes

First, as can be seen from figure 6.1B (left), semi-parametric point estimates are largely aligned with those produced by LMMs. This result, is perfectly aligned with the information presented until now. We have established in section 2.4 that, in the absence of confounding, linear models are generally robust to model misspecification, offering reliable estimates even when the true relationship between variables is not perfectly linear. Additionally, in section 3.2.3.1, we demonstrated that rs1421085 is not stratified, indicating that principal components do not confound the effect of rs1421085. Therefore, for this variant, the main source of differences between semi-parametric and linear estimates will likely be due to allelic effect differences, which is unlikely to affect most traits.

While we discuss allelic effect differences in the next section, we illustrate the difference between our reported TT $\rightarrow$ TC effect for body mass index, and the effects present in GWAS catalogues [19] in figure 6.1A. The figure shows that all three double-robust estimators are concordant and report statistically lower effect sizes than the linear/linear-mixed models. Note however, that because Neale V2 (linear model) and GeneATLAS (linear-mixed model) do not report standard deviations for this variant and trait, only point estimates are displayed. The red confidence interval corresponds to a linear model which was fitted with the same covariates as ours and is $\approx 33\%$ smaller than that of the double-robust estimators, illustrating a potential underestimation of parametric estimators' variance.

This phenomenon is more easily seen in figure 6.1B (right), showing that the distribution of p-values is shifted towards less significant values as compared to the GeneATLAS. After FDR correction for the 768 traits, we find fewer significant results at the 0.05 level (63 traits) than the GeneATLAS (159 traits). This

indicates that the variance obtained from LMMs is likely to be underestimated and the false discovery rate inflated. A summary table of all significant estimation results after multiple-testing adjustment is provided in Supplementary Table 1.

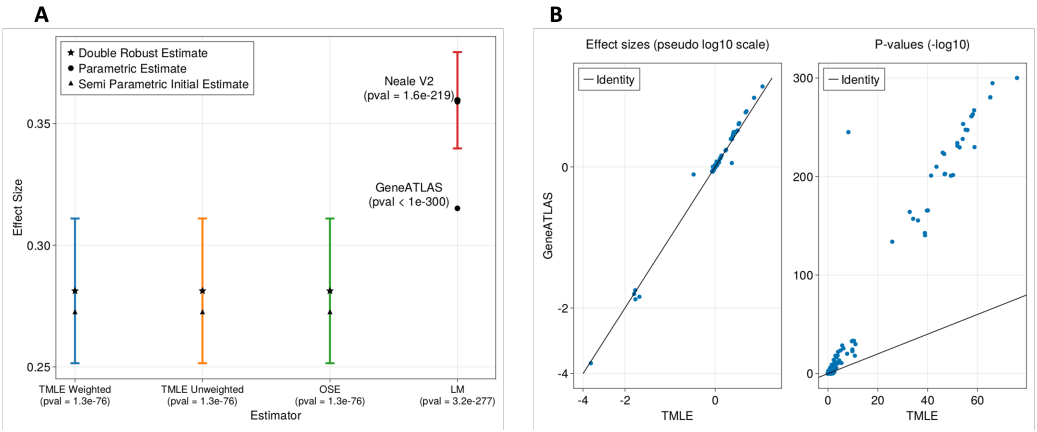

Figure 6.1: **Comparison of semi-parametric estimators and Linear (Mixed) Models on UK Biobank.** (A) **Inference results**. Comparison of methods to estimate the effect size of rs1421085 on body mass index (BMI; UK Biobank Data-Field 23104). All double robust estimators share the same initial fit and apply different targeting strategies: weighted TMLE (blue), unweighted TMLE (orange), OSE (green). The three estimates are all concordant and exhibit a statistically lower effect size than the linear model based inference (red). Neale V2 and GeneATLAS use a linear model and linear-mixed model respectively but do not report standard deviations. We refit the Neale V2 linear model on this data to obtain a confidence interval. We note that the central value of GeneATLAS does not lie within the 95% confidence interval of Neale V2. This could be because the GeneATLAS' model is slightly more flexible, thus yielding estimates closer to the non-parametric estimates we report. In contrast, all three double robust estimators are in complete agreement. (B) **Comparison with GeneATLAS**. Comparison of effect sizes (left) and p-values (right) reported by targeted minimum loss-based estimation and GeneATLAS (LMM). Effect sizes are concordant overall on this study but our p-values are more conservative. While it could be tempting to believe that more complex semi-parametric procedures yield higher p-values, this is not what we observed in the simulation of section 2.4. The most likely explanation for this behaviour is that model misspecification is leading to over-optimistic p-values and further false discoveries.

6.2.2 Allelic Effect Differences

We find 39 traits for which rs1421085 displays a significant Allelic Effect Difference, 35 of which are highly correlated with BMI. For instance, we find that the departure from homozygous TT to heterozygous TC is associated with an increase of 0.78 kg (95% CI: $0.69 - 0.86$). In comparison, the departure from heterozygous TC to homozygous CC is associated with a significantly larger increase of 1.33 kg (95% CI: $1.20 - 1.45$). For illustration, a subset of significant non-linear traits is presented in Fig. 6.2; Supplementary Table 2 contains the complete list. As might be expected, most estimates reported by GeneATLAS, based on a linear LMM approach, fall in-between estimates from our two scenarios, representative of an averaging effect.

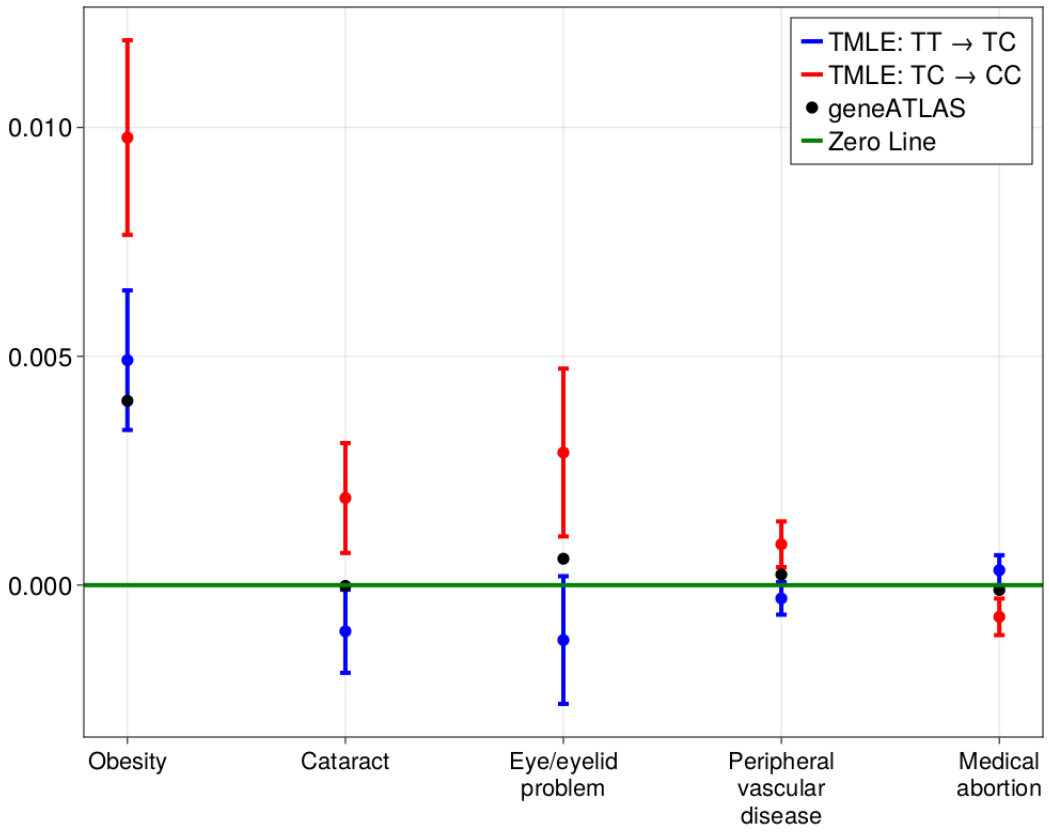

Figure 6.2: **Non-Linear effects**. A selection of traits for which rs1421085 TT → TC and TC → CC effect estimates are significantly different; Supplementary Table 2 contains the complete list. Effect sizes are reported with associated 95% confidence intervals together with estimates from GeneATLAS' LMM fits (black data points). The latter almost always fall in-between our TT → TC and TC → CC estimates, indicative of an averaging effect.

Notably, some traits seem to display opposite effect sizes (before multiple testing adjustment) for the two allelic changes TT $\rightarrow$ TC and TC $\rightarrow$ CC. Thus TarGene can capture variant-trait pairs displaying heterozygote advantage [63]. Such patterns cannot be detected by a linear model assuming equal allelic effect sizes.

6.2.3 Sieve Plateau Variance Estimation

In section 3.6, we proposed to account for the dependence among individuals in the variance estimates of semi-parametric estimators using Sieve Plateau (SP) variance estimation [35]. The SP method is computationally intensive, it requires, (i) computation of the GRM, a $450\,000 \times 450\,000$ matrix (ii) for each estimand and each threshold τ, matrices multiplications involving GRM components and influence curves. However, since SP only increases variance estimates, it is sufficient to revise estimates that may be significant at a given threshold (e.g. p-value ≤ 0.05). In Fig. 6.3, we show the effect of this correction for all effect sizes obtained for rs1421085 with initial p-value ≤ 0.07 (A). The p-values resulting from both the iid (red) and the sieve variance plateau estimators (blue) are reported and essentially show no remarkable difference. An example sieve plateau variance curve for body-mass index shows an increased variance of the influence curve of approximately 1.5 points. Note however, that the variance of the estimators themselves is obtained by division of $n \approx 450000$, further emphasising the marginal effect.

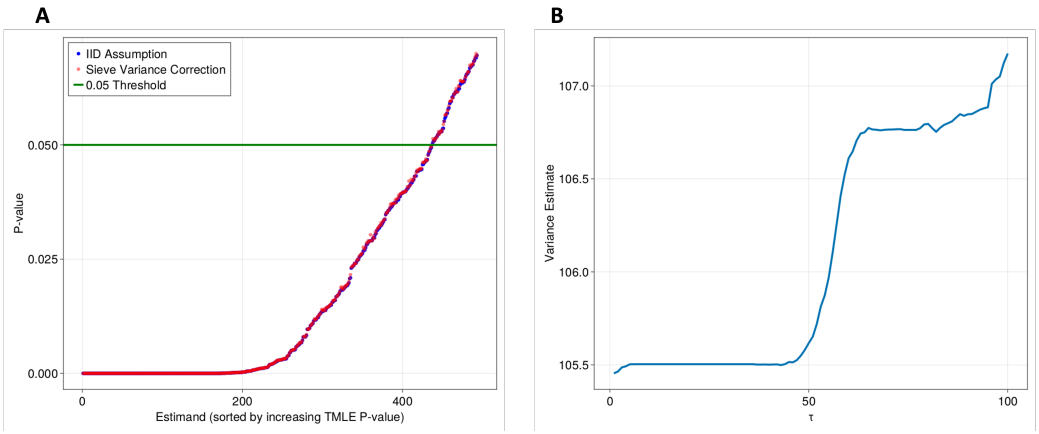

Figure 6.3: (A) **Impact of Sieve Plateau Correction** P-values obtained from two variance estimation methods for rs1421085. In red, the individuals in the UK Biobank are assumed to be independent and identically distributed (iid), while in blue, a sieve correction method is applied to account for the population dependence structure. Each p-value corresponds to a specific estimand of interest for which the initial iid estimate was under the 0.07 threshold. (B) **Sample Sieve Plateau Variance curve** for body mass index across 100 different thresholds.

This result is not necessarily surprising since only 2.8% of UK Biobank participants may be relatives [107]. The covariance terms between related individuals is thus likely to be negligible as compared to the individual variance terms in equation 3.57. The impact of the method in a more diverse population or in family-based studies is an interesting research direction which is not investigated in this Thesis.

6.3 Analysis of Interactions

6.3.1 Gene-Environment and Body-Mass-Index

We confirmed in section 6.2 that rs1421085 is significantly associated with various BMI related traits. Since BMI has also been associated with area-based deprivation [153], it is natural to investigate the potential interactions between rs1421085 and deprivation. There are currently two main measures of deprivation in the UK: the Townsend Deprivation Index (TDI) and the Index of Multiple Deprivation (IMD) used in [153]. A discussion of the advantages and limitations of those indices however, is beyond the scope of this thesis. We used these in-

dices in two separate phenome-wide interaction studies (PheWIS), one between rs1421085 and TDI and one between rs1421085 and IMD. Since deprivation indices are continuous quantities, we discretized them using quintiles and compared the most extreme quintiles. For rs1421085, there are still three genotype groups to compare: TT, TC, CC.

We found 21 significant BMI related traits after FDR correction that are captured by both TDI and IMD (Supplementary Table 3). For instance, whilst we have seen that an increase in the number of C alleles in an individual is associated with an increase in body weight and that most deprived individuals are more likely to be overweight, the interaction of these factors is super additive: an increase of 1.07 kg (p-value: 1.09×10^{6}, adjusted p-value: 1.4×10^{-3}) for TDI, and an increase of 0.91 kg (p-value: 3.89×10^{-5}, adjusted p-value: 8.0×10^{-3}) for IMD.

6.3.2 Epistasis and Hair-Colour

Detection of epistatis in complex traits is challenging [170]. For example, epistatic interactions are expected to be much smaller than main effect sizes, which can already be small for polygenic traits. In this section, we explore the potential for semi-parametric estimators to reveal such interactions. For that purpose, we rely on a study investigating hair colour, in which nine pairs of variants were reported to be statistically interacting with red-hair using a logistic regression model and a likelihood ratio test [105]. We note, however, that the likelihood ratio test statistic measures interactions on a multiplicative scale while we investigate interactions on an additive scale which is often of more direct public health relevance [160]. In particular, the existence of interactions on one scale does not imply the existence of interactions on the other scale. We found that five of the nine reported epistatic results (Table 6.1) are also revealed by semi-parametric estimation methods. Two were not reproduced and two were not computed because they did not pass the marginal positivity threshold $\varepsilon = 0.01$ (section 3.2.3.2) used throughout the study.

Variant 1	Variant 2	Effect-size	P-value
rs1805005 (GG →GT)	rs6059655 (GG →AG)	1.8×10^{-2}	1.0×10^{-19}
rs1805007 (CC →CT)	rs6088372 (CC →CT)	3.0×10^{-2}	1.4×10^{-40}
rs1805008 (CC →CT)	rs1129038 (TT →CT)	-1.9×10^{-2}	2.9×10^{-15}
rs2228479 (GG →GA)	rs6059655 (GG →AG)	-1.6×10^{-2}	1.5×10^{-24}
rs885479 (GG →GA)	rs6059655 (GG →AG)	-1.5×10^{-2}	1.6×10^{-16}

Table 6.1: Summary table of reproduced significant results for red hair color.

In total, we find 27 significant epistatic signals for traits corresponding to either skin or hair colour (Supplementary Table 4). This is expected because hair and skin colour are known to co-vary [148].

6.3.3 Higher-Order Interactions for Targeting Biological Mechanisms

Finally, as a foreshadowing of chapter 7, we investigate a complex biological mechanism using higher-order (3-points) interactions. Precisely, we focus on VDR, a nuclear hormone receptor that binds 25-hydroxyvitamin D (25OHD), the active form of vitamin D, and the retinoid-X receptor (RXRA). This complex can then enter the nucleus and regulate gene transcription programmes. Because this mechanism depends on three interacting molecules, it provides a natural field for investigating epistasis. We identified three genetic variants that were previously associated with differential expression of each molecule: A to C (rs7971418) is associated with increased levels of VDR mRNA; G to T (rs1045570) is associated with increased levels of RXRA mRNA [163]; and, C to T (rs3755967) has been associated with decreased 25-hydroxyvitamin D [72]. Although no interaction between these variants was detected after FDR adjustment (FDR < 0.05, Supplementary Table 5), 47, 42 and 39 pairwise interactions, and 21 3-point interactions were significant in single tests, hence demonstrating the potential of the method.

6.4 Conclusion

In this chapter, we compared semi-parametric and linear estimates using real-world data from the UK Biobank. Our analysis confirmed that, in the absence of confounding factors, linear models are robust with respect to model misspecification. However, these models can suffer from semantic limitations because they estimate model parameters rather than quantities of direct scientific interest. This issue was highlighted through examples of allelic effect differences, where the standard parameterization of linear models may not fully capture the underlying biological phenomena. We note however that in this case, a different parameterization (one-hot encoding) would likely resolve the issue.

We also demonstrated the strengths of our proposed semi-parametric methods in replicating known genetic effects and discovering new gene-environment interactions. Additionally, we briefly explored the potential of these methods to investigate higher-order interactions, which could offer deeper insights into biological mechanisms. This exploration will be expanded upon in chapter 7.

Chapter 7

Effects Mediated by the Vitamin-D Receptor

Over the past twenty years, genome-wide association studies (GWAS) have successfully identified many regions of the genome associated with human traits [90]. However, the ever growing number of reported associations has only been followed by limited mechanistic insights, essential for therapeutics development. As discussed in chapter 2, this is largely due to the fact that most genetic associations fall within non-coding regions of the genome. Rather than affecting the protein sequence of genes, the identified variants are believed to modulate their regulation. Transcription factors play a crucial role in gene regulation by controlling the transcription of specific genes. They are proteins that bind to specific DNA sequences near the genes, particularly in regions called promoters or enhancers, to regulate the process by which DNA is transcribed into messenger RNA (mRNA). In this chapter we propose a framework to investigate the impact of transcription factors' disrupted binding on human traits. The framework is exemplified with a specific transcription factor, the vitamin D receptor (VDR). We present the general framework in section 7.1, define the VDR study scope in section 7.2 and showcase our findings in section 7.3.

7.1 Transcription Factors' Differential Binding

To ease the hunt, instead of looking for causal variants and mechanisms among the numerous GWAS hits, we could restrict the attention to variants already implicated in transcription regulation from the beginning of the analysis. In sec-

tion 2.1.4, we described a variety of ways variants may impact molecular phenotypes, these were called molQTLs. However, molQTLs are not necessarily causal as their identification may exclusively rely on statistical association testing. For instance eQTLs, just like GWAS hits, may not impact the differential expression of genes but be in linkage disequilibrium with variants that do.

Binding quantitative trait loci (bQTLs), represent a particularly interesting type of molQTLs. They are specific regions of the genome that are associated with variations in the binding affinity of regulatory proteins from ChIP-exo experiments. However, unlike some other molQTLs, modern ChIP-exo technology capture proteins together with their binding site with a near base-pair resolution. This precise physical connection, allows us to causally implicate these loci in binding disruptions. As a result, bQTLs are natural candidates for prioritisation and further study to understand their impact on human traits.

It is important to note however, that binding disruption alone does not necessarily lead to physiological consequences, such as disease. This, even if a bQTL, e.g., bQTL_1, is associated with a disease. This connection could be coincidental, possibly due to another bQTL_2, disrupting the binding of a different transcription factor but in linkage disequilibrium with bQTL_1.

To mitigate this limitation, and strengthen our confidence in the proposed mechanism, we estimate the interacting effect on human traits between bQTLs and distant genetic variants called trans-acting variants. The benefit of the approach is that the entire interaction would need to be confounded for the bQTL to be non-causal of trait. While this is not strictly impossible, the chances drop rapidly with the interaction order. The reason for this is because the alternative explanatory variants would need to interact through a common alternative mechanism and from a distance, i.e., not physically. To clarify, the potential scenarios are presented in figure 7.1, and described below.

1. The (bQTL, trans-acting variant) or (bQTL, alt-QTL_2) interaction is causal. The first case is ideal because we identified two causal variants instead of one. In the second case, the linked alt-QTL_2 is causal. This is not a concern because the variant of interest is the bQTL. The trans-acting variant does not need to be causal itself.

2. The (alt-QTL_1, trans-acting variant) or (alt-QTL_1, alt-QTL_2) interaction is causal. These situations are problematic but unexpected. In particular,

alt-QTL$_1$ needs to act on the same trait through an alternative mechanism but also be interacting with a distant alt-QTL$_2$ in linkage disequilibrium with the trans-acting variant.

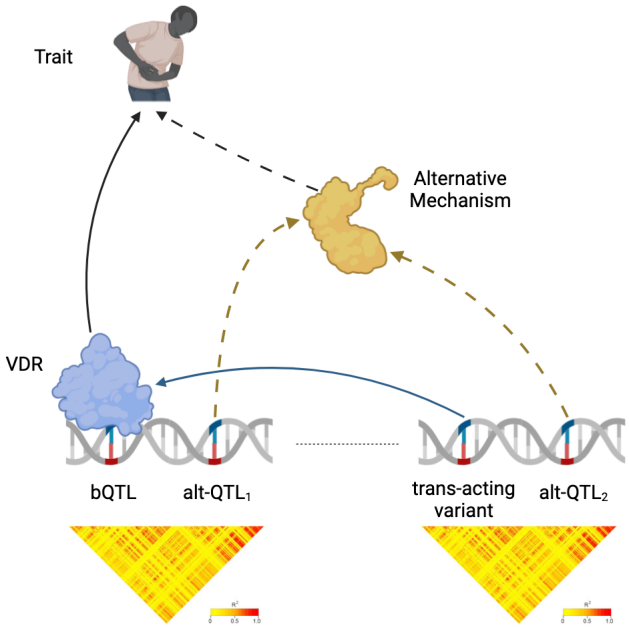

Figure 7.1: **Why trans-interactions likely reveal causal variants and mechanisms.**

Note that the absence of physical interaction between bQTLs and trans-acting variants does not pose a statistical detection problem. Statistical interactions can exist without physical interactions [160]. The caveat, discussed in section 4.3.2.3, is that power to detect interactions drops exponentially with the order of interactions. Heuristically, statistical power is traded-off for causal identification power at the rate of the interaction order.

If distant interactions are unexpected, how can we hope to detect them in the first place? In particular, naively scanning the entire genome for all possible trans-acting variants would lead to hundreds of thousands of pairwise interaction tests for a single bQTL. The associated multiple hypothesis correction burden would likely annihilate any significant signal. The solution is to carefully select trans-acting variants previously associated with the bQTL's mechanism. In particular, because the trans-acting variants need not be causal, only a single trans-acting variant per linkage disequilibrium block needs to be tested. Furthermore, since both bQTLs and trans-acting variants are associated with the same mechanism,

the existence of a putative confounding mechanism is even more unlikely.

In the following section we exemplify the above framework with a specific mechanism, the binding of the vitamin D receptor transcription factor.

7.2 The Vitamin D Receptor

The vitamin D receptor (VDR) is a protein found in cells throughout the body that binds to calcitriol, the active form of vitamin D. It belongs to the nuclear receptor superfamily, which includes receptors for various hormones and signaling molecules. When activated by vitamin D, the VDR forms a complex with another protein called the retinoid X receptor (RXR). This complex then binds to specific DNA sequences called vitamin D response elements in the promoter regions of target genes, influencing their transcription.

The primary function of the VDR is to regulate the expression of genes involved in various biological processes, including calcium homeostasis, bone metabolism, immune function, and cell growth and differentiation. Vitamin D, which can be obtained from dietary sources or synthesised in the skin through exposure to sunlight, serves as the ligand that activates the VDR.

The activation of VDR-mediated gene expression by vitamin D is essential for maintaining optimal health. Deficiencies or dysfunctions in the VDR or vitamin D metabolism can lead to various health problems, including skeletal disorders like rickets and osteomalacia, as well as an increased risk of autoimmune diseases, cancer, and other chronic conditions [33, 92, 164].

In this study, we thus hypothesise that VDR bQTLs affect human traits and diseases through binding disruption and gene regulation modification. As per the previous section, we also suppose that this effect is modulated by trans-acting variants. Since the VDR complex has three constituents, these trans-acting variants could either be associated with VDR, RXRA or vitamin D, which is illustrated in figure 7.2. In this case, we will also limit the study to pairwise interactions which are thus defined by a triple (bQTL, trans-acting variant, trait). In this section, we explain how we built a list for each element in this triple. Then, the total scope of the study is simply defined by the set of all pairwise (bQTL, trans-acting variant) combinations against all traits.

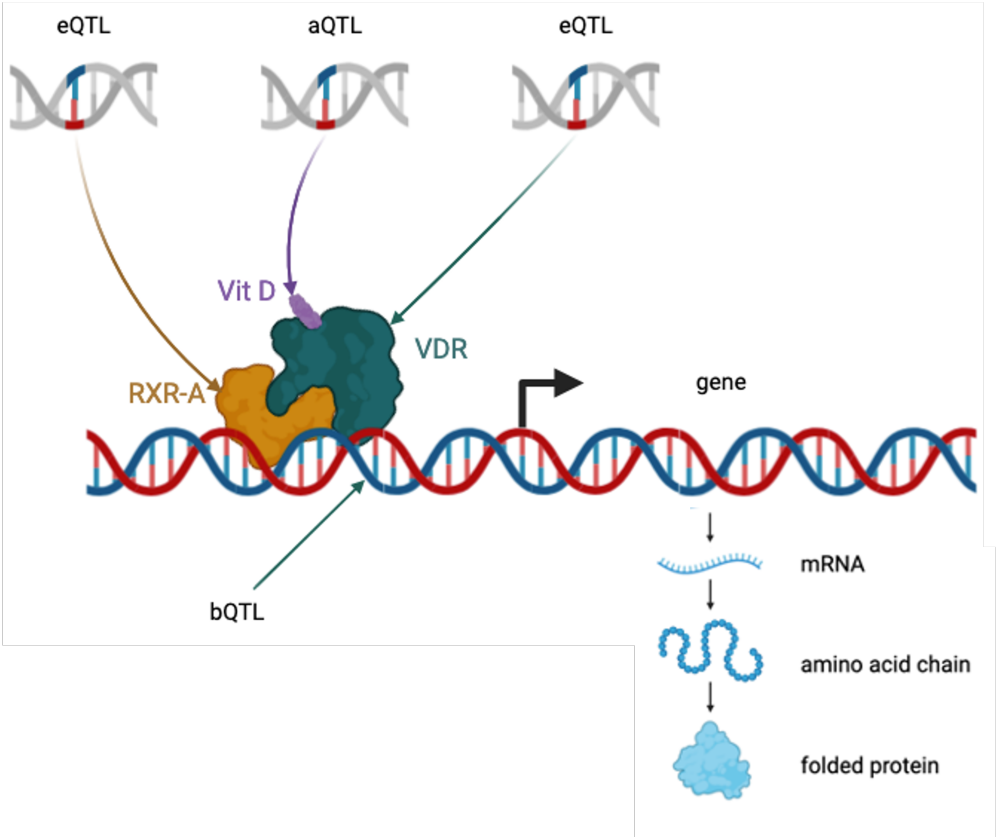

Figure 7.2: **Gene regulation by the VDR-RXRA complex.** The VDR-RXRA is a protein complex that binds the vitamin D ligand and regulates the transcription of many genes. Genetic variations can modulate this regulation in many ways. For instance, possibly remote eQTLs or abundance QTLs (aQTLs) can affect the availability of any of the molecule. A bQTL is a variant which alters the binding of VDR to the DNA strand.

7.2.1 Binding Quantitative Trait Loci

Binding Quantitative Trait Loci (bQTLs) are the most important elements in our framework because they are supported by direct physical evidence. These variants are further filtered according to a set of rules making them mechanistically interpretable. These rules were defined and encapsulated in a Nextflow pipeline by Breeshey Roskams-Hieter, Øyvind Almelid and Chris Ponting. The list we consider here thus corresponds to the output of this pipeline for the VDR transcription factor and is provided in supplementary table 7. We summarise these rules below.

1. bQTLs show evidence of allele specific binding. They are identified using chromatin immunoprecipitation followed by massively parallel DNA sequencing assays (ChIP-seq). For VDR, the ChIP-Exo data originates from lymphoblastoid cell lines from the International HapMap Project [54, 123]. Statistical evidence of differential binding was then measured using BaalChIP, a Bayesian approach accounting for DNA copy-number changes [132]. The quantity reported by BaalChIP is the corrected allelic ratio (CAR), it measures the preference for binding to either the reference (REF) or alternate (ALT) allele while correcting for biological and technical biases. A higher CAR (> 0.5) indicates higher affinity for the REF allele and lower CAR (< 0.5) higher affinity for the ALT allele.

2. bQTLs map to a high-quality motif. The set of bQTLs is then filtered to keep only those that land within a DNA motif. A motif is a short widespread DNA sequence assumed to have a biological function. The motifs we consider are either known motifs from the JASPAR database [131], or newly identified from the ChIP-seq data using NoPeak [97]. For each mapped motif, a mapping score is obtained for both the REF and ALT alleles of a given bQTL. A motif score difference (MSD) is computed by taking the difference between the REF motif score and the ALT motif score. A higher MSD thus corresponds to the REF allele being less disruptive of the motif than the ALT allele. A motif is then said to be high-quality if the Spearman's Correlation Coefficient between the CAR and MSD is statistically positive. A high-quality motif is thus statistically concordant with the differential binding observed across all bQTLs. Only bQTLs mapping to high-quality motifs are considered further.

3. The motif disruption is concordant with the binding disruption. Finally, a (bQTL, motif) pair is concordant if both $\text{CAR} > 0.5$ and $\text{MSD} > 0$ or $\text{CAR} \leq 0.5$ and $\text{MSD} \leq 0$. That is, the ALT or REF allele disrupts both the motif and the binding affinity of the transcription factor. Only concordant (bQTL, motif) pairs are considered for statistical testing.

Together, these steps make sure that the effect of the variant is interpretable: the variant alters the transcription factor's binding by disrupting a binding motif. For VDR, no new motif was discovered from the ChIP-seq data and all bQTLs map to a unique JASPAR motif, presented in figure 7.3.

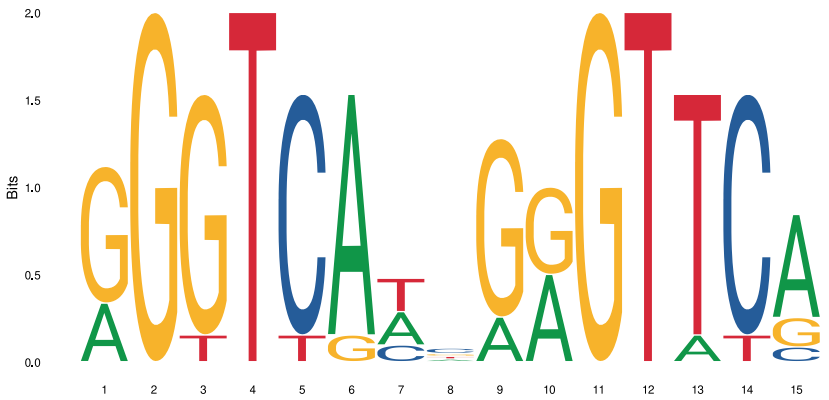

Figure 7.3: **VDR motif.** The unique mapped high-quality motif across all bQTLs under investigation. Only bQTLs whose ALT or REF allele disrupts both this motif and the binding affinity of VDR are considered.

7.2.2 Trans-Acting Variants

The trans-acting variants we consider here correspond to a curated list built from a literature review. They correspond to variants that have been associated with differential expression of a protein within the VDR-RXRA complex, or abundance of the vitamin D ligand. As such, they measure the recruitment availability of the complex to the transcription factor's binding site. A brief description of each trans-acting variant is provided in table 7.1; a complete description including the publication of origin can be found in supplementary table 8.

ID	CHROM	LABEL	LOCUS
rs3847987	chr12	vitamin D levels	VDR
rs117913124	chr11	vitamin D levels	CYP2R1
rs2228570	chr12	VDR Structure and activity (aQTL)	VDR
rs3755967	chr4	vitamin D level	GC
rs12785878	chr11	vitamin D level	NADSYN1
rs10741657	chr11	vitamin D level	CYP2R1
rs17216707	chr20	vitamin D level	CYP24A1
rs10745742	chr12	vitamin D level	AMDHD1
rs8018720	chr14	vitamin D level	SEC23A
rs11168319	chr12	VDR eQTL	VDR

Table 7.1: **VDR trans-acting variants** For each trans-acting variant, the label indicates which molecular phenotype it is associated with and the locus the closest gene.

7.2.3 Human Traits

In this study we focus on the biological mechanism and are trait agnostic. For each (bQTL, trans-acting variant) pair, we consider all 770 traits described in section 3.1.2. These comprised 110 non-binary and 660 binary traits as previously defined by the geneATLAS [21].

7.2.4 Estimation Strategy

Regarding estimation, We use the practical recommendations from the simulations of chapter 4. The reasonable scale of the study enables the use of the most expensive cross-validated estimators which benefited from better coverage. We thus use the cross-validated weighted targeted minimum loss-based estimator, where the cross-validation scheme is a 3-folds stratified cross-validation. To estimate the nuisance functions Q_Y and G, the XGBoost model is used. The XGBoost model's max-depth and lambda hyper-parameters are selected according to an identical internal cross-validation strategy. Finally, since the one-step estimator comes at almost zero additional computational cost, we report it as well as a control procedure.

Regarding positivity violations, we adopt a conservative strategy and use a

marginal positivity constraint of 0.01. As a reminder, a threshold of 0.005 was found to be sufficient in the simulation studies for UK Biobank's size datasets.

7.3 Results

To present a first overview of the results, it is instructive to display them in a Q-Q plot. That is we compare the distribution of obtained p-values to the p-values we would have obtained if the null hypothesis was true. Figure 7.4 shows that, if we consider all tests collectively, there is no departure from the null hypothesis. This is also the result we obtain if we perform multiple hypotheses correction using either the Bonferroni (FWER) or Benjamini-Hochberg (FDR) method across all tests (section 3.5.4).

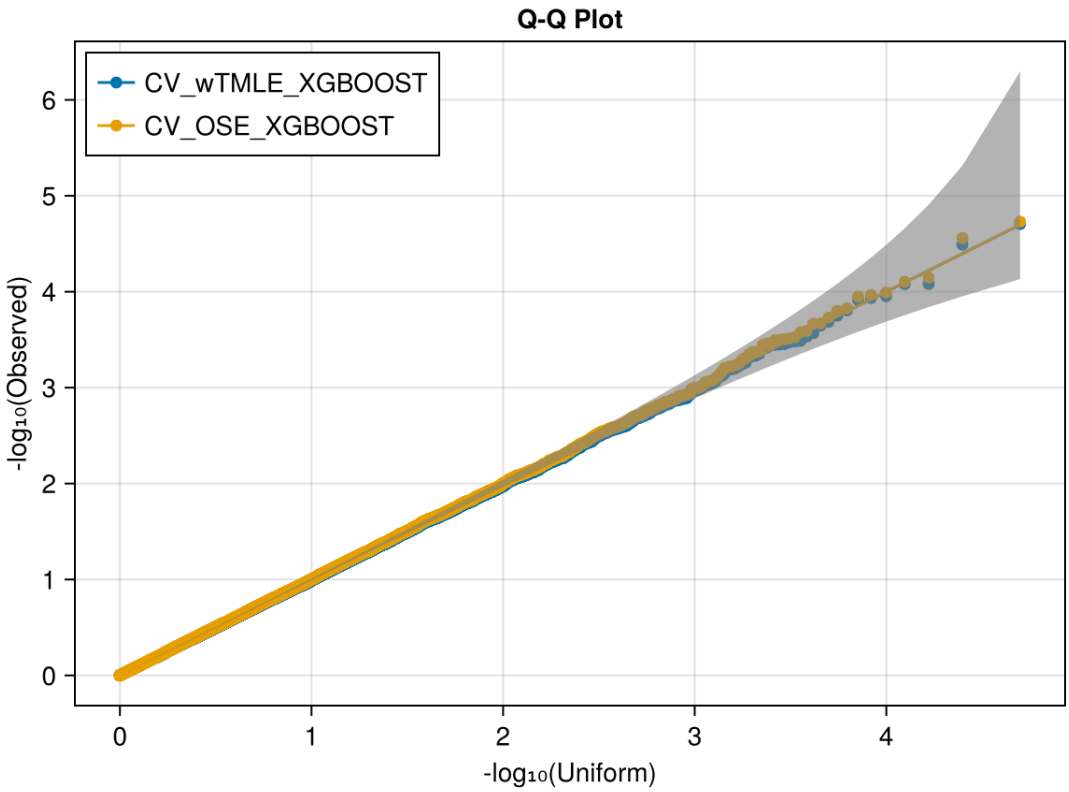

Figure 7.4: **Q-Q plot of all p-values**. The quantiles of the empirical p-values are plotted against the quantiles of the theoretical p-values if the null distribution was true. These two distributions are essentially indistinguishable, illustrating that, when considering all tests collectively, the null hypothesis of no effect is true.

If the tolerance for false discoveries is low, one could stop the analysis here.

However in this case, we are ready to accept a few false positives providing we can make some interesting discoveries. We thus adopt a risk tolerant strategy and control for the FDR per trait at the 0.05 level. That is, for each trait, the expected fraction of false positives should be no larger than 5%. This might reflect the perspective of a clinician focused on a single trait, who is not inclined to be penalised for unrelated traits in the current study (e.g., alcohol intake or other irrelevant factors). Furthermore, only interactions with sufficiently frequent traits are considered for post analysis. By sufficiently frequent we mean at least 20 cases in the smallest (bQTL, trans-acting variant) genotype group.

In total, 22 interactions are considered significant based on the aforementioned criteria. They correspond to 5 out of 18 bQTLs, and 21 out of 770 traits. In the remainder of this chapter we investigate the potential genes whose expression changes are associated with bQTLs, and how these genes are related to traits. To that end, we use eQTLGen, a database of eQTLs aiming to understand the genetic architecture of blood gene expression [163]. Precisely, we search the database for genes for which bQTLs are also cis-eQTLs. This is motivated by the fact that bQTLs are expected to be located in promoter or enhancer regions, and impact the expression or nearby genes. The complete list of interactions, enriched with additional information, is provided in supplementary table 9.

7.3.1 Effect of rs9846571 through eIF4E3

It is expected that causal bQTLs be associated with multiple related traits and it is the case for rs9846571. The interaction between rs9846571 and rs2228570 is significant for both F32 Depressive episode (p-value=0.00012) and F30-F39 Mood disorders (p-value=0.00033), where the former is a subset of the latter. Another related associated trait with this bQTL is mania/bipolar disorder/manic depression (p-value=0.00063). It shares 33% and 72% cases overlap with F32 Depressive episode and F30-F39 Mood disorders respectively. However the interaction is detected with another trans-acting variant: rs10745742. Finally, this bQTL is also associated with two other seemingly unrelated traits: infection of nervous system and K76 Other diseases of liver.

From eQTLGen, rs9846571 is a cis-eQTL for a single gene, eIF4E3, a member of the Eukaryotic Translation Initiation Factor 4E Family (eIF4E), which plays a crucial role in protein synthesis. As illustrated in figure 7.5, members of the

eIF4E family form a complex with eIF4G and eIF4A to bind to the 5' cap of messenger RNAs which will be presented to the ribosome for translation. As such, it may be expected that rs9846571, or a linked variant, be reported to be a protein QTL (pQTL) for other genes. Unfortunately, pQTL databases are pretty rare and low powered relative to eQTL databases. A search through the genomic atlas of the human plasma proteome [149] did not reveal any association.

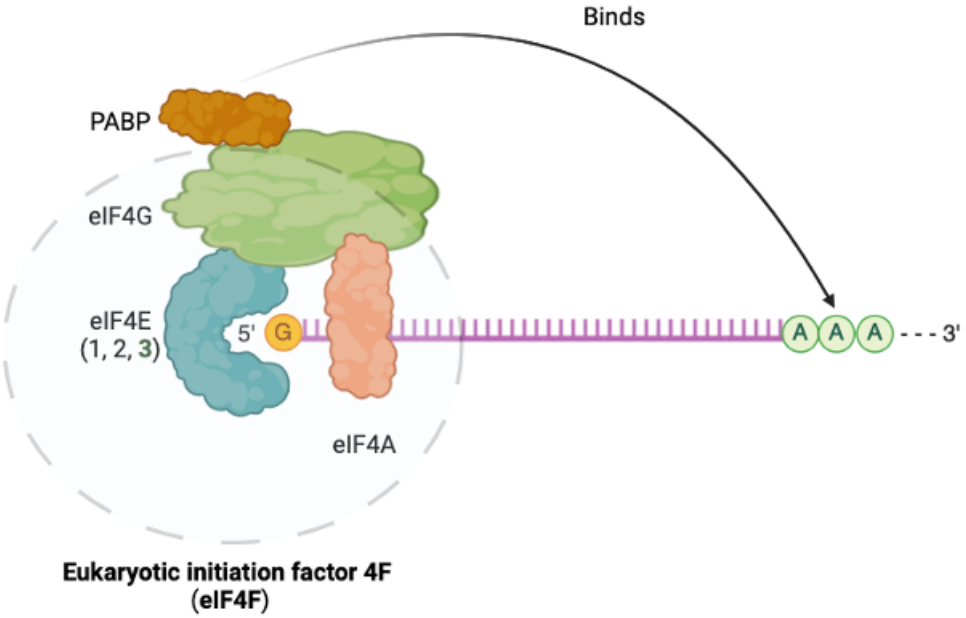

Figure 7.5: **The eIF4F complex.** This family of protein mediates protein synthesis by binding the 5' cap of messenger RNAs.

The most commonly studied eIF4E protein is eIF4E1. The eIF4E2 and eIF4E3 homologs have been shown to complement the translation process in a condition-specific way. Note that such condition-specific effects can only be revealed by interaction studies such as this one. For instance, eIF4E2 associates with other factors to re-program translation in response to hypoxia-induced stress [155]. Regarding eIF4E3, it was shown to form an active translation complex upon cellular stress treatment using Torin1 [172]. Torin1 is an inhibitor of the mTOR protein kinase which is believed to be responsible for the mechanism presented in figure 7.6. If this is the case, we might be able to detect an interaction between an eQTL for mTOR and rs9846571 for the same traits. Unfortunately, an interaction study between rs4845986, the lead cis-eQTL for mTOR in eQTLGen (p-value $<$ $1e-166$), and rs9846571, did not reveal such interactions (p-values > 0.6).

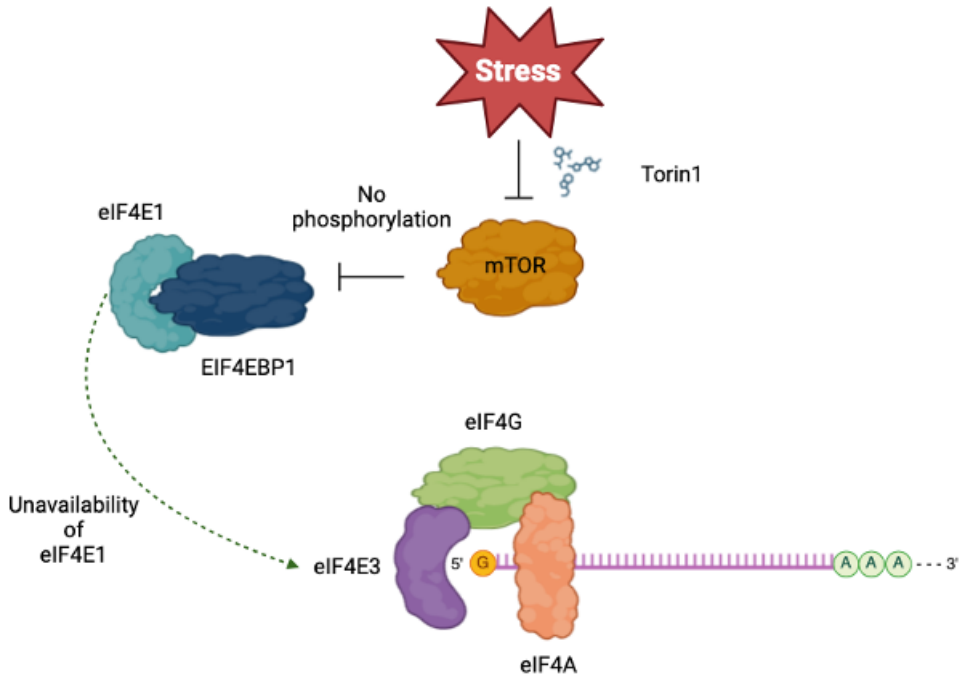

Figure 7.6: **eIF4E3 takes over under Torin1 stress.** Torin1 stimulation inhibits the action of mTOR, a protein kinase phosphorylating EIF4EBP1. When unphosphorylated, EIF4EBP1 competes for the binding of EIF4E1. The otherwise poorly active eIF4E3, can then mediate protein synthesis.

Other studies have suggested that eIF4E3, unlike the eIF4E1 growth factor, acts as a tumour suppressor [114, 162]. These studies suggest that eIF4E3 is indeed competing with eIF4E1 through an atypical cap-binding mode. The 5' cap of RNA is a special modification added to the 5' end of eukaryotic messenger RNA shortly after transcription begins. It plays several important roles in RNA stability, processing, and function. Expression of eIF4E3 is also reduced in these cancer cells indicating that this cap binding plasticity may underlie a clinically relevant inhibitory mechanism. While we do not detect association with cancer, this putative inhibitory mechanism might be shared and requires more research.

7.3.2 rs76057752 and Blood Cells

Significant interactions for rs76057752 are found with essentially two related traits each with a different trans-acting variant. The interaction between rs76057752 and rs11168319 on monocyte count (and percentage), and the interaction between rs76057752 and rs10741657 on high light scatter reticulocyte count (and percent-

age). Monocytes are white blood cells that can differentiate into macrophages while reticulocytes are immature red blood cells. They are both produced in the bone marrow, the primary site of new blood cell production.

rs76057752 is also a cis-eQTL for two genes: *SMG9* and *KCNN4*. The *SMG9* gene encodes a protein that is part of the nonsense-mediated mRNA decay (NMD) pathway, a cellular surveillance mechanism that degrades faulty mRNAs and regulates gene expression. In mice, hematopoietic-specific deletion of a core NMD factor, led to the rapid, complete, and lasting extinction of all hematopoietic stem and progenitor populations [171]. In contrast, more differentiated cells were only mildly affected, suggesting that NMD is mainly essential for proliferating cells.

The *KCNN4* gene encodes the calcium-activated potassium channel KCa3.1, which plays a key role in the regulation of potassium transport in various cells, including red blood cells and immune cells [83, 91]. It is implicated in cellular processes such as volume regulation, cell proliferation, and migration. Even though not directly supported by our results, the functions of KCa3.1 suggests its implication in related diseases of immune imbalance [83]. Similarly, *KCNN4* has been implicated in macrophage multinucleation, identifying it as a potential therapeutic target for inhibition of bone resorption and chronic inflammation [75].

7.3.3 rs17160772 Potential Driver of Ciliopathies

rs17160772 is found to be associated with three apparently distinct traits and is furthermore an eQTL for four different genes. The most significant eQTLGen association is found for TTC26 (p-value=$1.6e-106$), an intraflagellar transport protein required for transport of motility-related proteins into flagella. Similarly, it was shown that cilia in TTC26-mutated patient cells display variable length and impaired function [138]. Cilia are small organelles bound to the membrane of most eukaryotic cells that can either be motile or non-motile. In the respiratory epithelium, cilia help clear the mucus from the respiratory tract (figure 7.7). They remove inhaled particles including pathogens before they can reach the lungs. It is thus perhaps not surprising that one of the detected traits with rs17160772 is K20 Oesophagitis. The second trait found for this bQTL is B35-B49 Mycoses. In fact, the inflammation of the oesophagus can be of fungal origin, like oesophageal candidiasis. In the UK Biobank, there is a non negligible 15% of individuals with mycoses that are also diagnosed with oesophagitis.

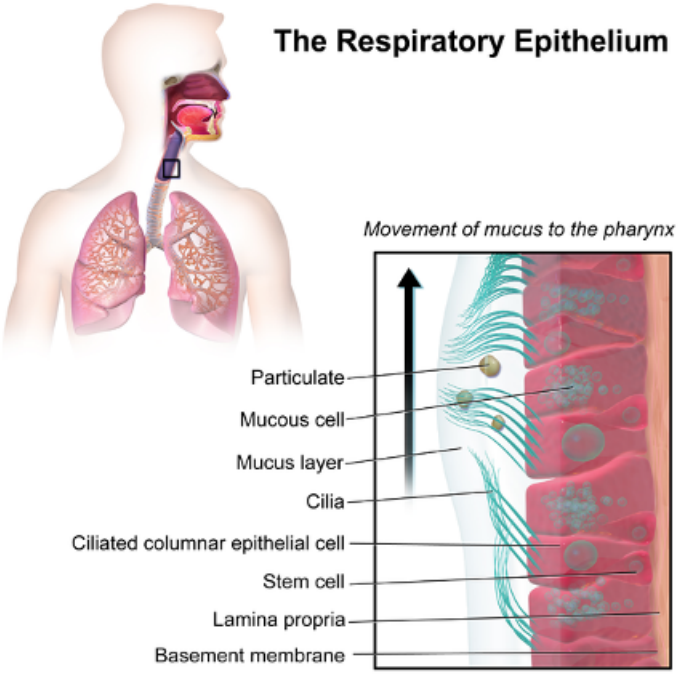

Figure 7.7: **The respiratory epithelium**. Cilia help to clear the mucus from the respiratory tract. (Image copied from wikipedia, free to use under the Creative Commons Attribution 3.0 Unported license)

Cilia have also been implicated in bone disorders, a group of conditions denoted by skeletal ciliopathies. Polydactyly, short limbs, short ribs, scoliosis, a narrow thorax, and numerous anomalies in bone and cartilage, have been discovered in ciliopathies [80]. Furthermore, a TTC26 splice variant was recently shown to lead to a novel ciliopathy syndrome with skeletal manifestations [3]. These skeletal conditions are also supported by the last detected trait we find for this bQTL: fracture of pelvis and lower limb.

7.3.4 The Impact of rs6580323 on Myelination

Surprisingly, rs6580323 is not an eQTL for any gene but is in an intron of the neuregulin gene *NRG2*. Neuregulins are cell-cell signaling proteins that are ligands for receptor tyrosine kinases of the ErbB family [42]. In particular, NRG2 is secreted by astrocytes and promotes survival and outgrowth of neurons via ErbB3 [106]. Members of the neuregulin family contribute to various biological processes including differentiation, migration, and myelination. Myelin is the material that surrounds nerve cell axons to insulate them and increase the rate at

which electrical impulses are passed. The loss of the myelin sheath, insulating the nerves, is the hallmark of some neurodegenerative autoimmune diseases like multiple sclerosis. Even though the exact effect of rs6580323 on *NRG2* remains elusive, we also find that this bQTL is associated with G35-G37 demyelinating diseases which includes multiple sclerosis (p-value=$8.3e-5$). Moreover, a Mendelian randomisation study showed that lowered vitamin D levels were strongly associated with increased susceptibility to multiple sclerosis [100].

We also find that rs6580323 is associated with Monocyte percentage, the precursors of macrophages. In some forms of multiple sclerosis, macrophages are thought to be responsible for tissue damage and in particular demyelination [88, 145]. At the ligand level, vitamin D has also been shown to inhibit the production of pro-inflammatory cytokines in monocytes [176].

7.3.5 rs178399 and Degenerative Neurological Problems

Finally, the last bQTL for which we report interaction associations is associated with a total of seven different traits and is an eQTL for four different genes. Here, we only consider the interaction between rs178399 and rs12785878 on chronic/degenerative neurological problem (p-value=$3.2e-5$). This is possibly supported by two closely-linked genes for which rs178399 is also an eQTL: *SERPINB1* and *NQO2*. *SERPINB1* encodes a serine protease inhibitor which specifically inhibits neutrophil elastase. It was identified by a Mendelian randomisation study as a potential drug target for Alzheimer disease [146]. A GWAS also reported multiple associations between variants around *SERPINB1* and amyloid beta, a hallmark of Alzheimer disease [37]. Similarly, rs12785878 is an eQTL for *NQO2*, a gene encoding a quinone detoxifying enzyme which may be toxifying in some contexts [70]. This might be relevant to Parkinson's disease, since this toxifying function of *NQO2* was suggested to contribute to dopaminergic degeneration [69].

7.4 Discussion

In this chapter we have presented a powerful analysis framework for the discovery of binding quantitative trait loci which affect human traits. We also applied the methodology to an important transcription factor, the vitamin D receptor.

Unfortunately, the preliminary results in this case remain limited. In part, this could be simply due to the absence of genuine interactions in this specific study, we have already argued that trans-interactions are expected to be rare. However, this could also be related to the limited power associated with interaction detection as discussed in chapter 4.

The other conclusion of this analysis is that the results are difficult to contextualise in the current biomedical landscape. There are two mains reasons for this. The first is that the analyses were based on the assumption that the variant would impact the expression of a nearby gene. This was investigated using a database of eQTLs based on blood samples. We have seen that in many cases indeed, the variant is also a cis-eQTL for some genes. However, the implication of the gene on trait was not always straightforward. This could be due to the fact that genes may have tissue-specific behaviours. The GTEx database provides tissue specific QTLs but is not as well powered as eQTLGen. Similarly, these genes may only be expressed in response to specific stimuli. Some of the experiments we described above, move in this direction but more will be needed to deepen our understanding of human biology.

The second reason is that interactions are difficult interpret or contextualise because they would only be revealed by tailored experiments which have not been made or are impossible. As an example, consider the interaction between rs6580323 and rs12785878 on Monocyte percentage (p-value=0.00048). The interaction effect of the respective CC→CT and GT→TT changes, is -0.06 with a p-value of 0.0023. In contrast, the interaction effect of the CC→CT and GG→GT changes is 0.11 with a p-value of 0.013. Note that in this cases, both changes lead to more binding and more availability of vitamin-D but seem to have opposite synergistic effects. One may argue that the interaction may be spurious, but this phenomenon is not at all unexpected. This is because interactions measure the extent to which the effect of the two factors together exceeds the effect of each considered individually. These can be different in different genotypes' contexts which in this example are given by the transactor, i.e. GT→TT and GG→GT. That is, in this case, while interactions may be useful to detect causal variants they do not inform on the effect of the bQTL on traits. The natural candidate estimand for this question is the Average Treatment Effect (ATE). Unfortunately, estimating the ATE of each bQTL on traits for the 22 significant interactions did not reveal any effect, as illustrated in figure 7.8.

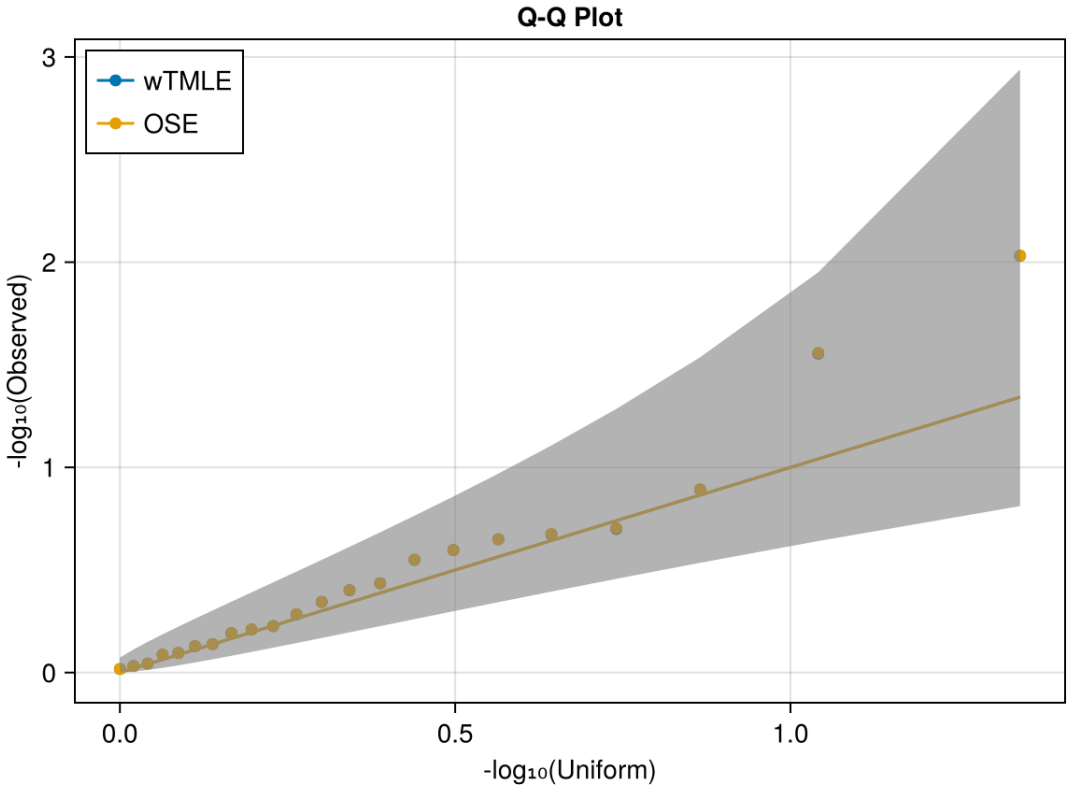

Figure 7.8: **Single variant Effects of bQTLs.** The interaction effect does not inform on the effect of the bQTL on trait but the Average Treatment Effect does. This Q-Q plot presents the p-values associated with the Average Treatment Effect of all significant 22 (bQTL, outcome) pairs.

However, this does not necessarily mean that the bQTLs are not functional. Their effect could only become apparent in specific genomic contexts, here embodied by the trans-acting variants' genotypes. This is actually what the interaction reveals. What is left to do is to estimate the bQTLs' effect in these contexts. To make things concrete, let us return to the previous example and consider the bQTL's CC→CT change and the trans-acting variant's GT→TT change. If this interaction is significant it means that the $\text{ATE}_{\text{CC}\rightarrow\text{CT,GT}}$ is different from the $\text{ATE}_{\text{CC}\rightarrow\text{CT,TT}}$. To obtain a comprehensive understanding of the mechanism, we thus also need to estimate these two effects. They represent the effects of the bQTL when the trans-acting variant is fixed to a specific genotype and are similar, but not exactly equal to conditional effects, because the intervention also bears on the trans-acting variant. This has not been done at this point but represents an interesting follow-up idea.

Chapter 8

Conclusion

This work presents a significant contribution to the field of population genetics. It integrates within the broader goal to develop a better understanding of the effect of genetic variations on human traits, and speed up the development of novel therapies. The achievement of such advancements requires three main ingredients; (i) the identification of causal variants, (ii) the identification of the biological mechanisms through which they act on traits, and, (iii) a quantification of these effects. While, we addressed all three axes, the main contribution of this thesis is on quantification.

8.1 Genetic Effect Quantification

Up to this date, the field of statistical genetics has been dominated by the use of parametric models. Under confounding, the inferences arising from these models can be biased and lead to inflated error rates. Furthermore, this inflation will be exacerbated by the ever growing sample sizes of modern biobanks. In this work, we showed that modern semi-parametric estimation methods are poised to solve this problem. This is because they enable highly flexible modelling strategies while guaranteeing asymptotically optimal inferences with minimal bias. The attention was restricted to two existing methods and their variations: the targeted minimum loss-based and the one-step estimators.

The empirical performance of these methods was evaluated in both simulated and real-world data. The simulations revealed that the estimators attained nominal coverage provided genetic variations were sufficiently frequent. In the precise setup of these simulations, this meant that 50% of genetic effects could be

confidently estimated. Larger sample sizes will inevitably lead to an increase in the proportion of estimable effects. Using real-world UK Biobank data, we contrasted semi-parametric estimates with currently reported linear estimates. We showed that under unconfoundedness, the effect sizes were largely concordant. However, we also saw that the approximations made by linear models can still be detrimental to inferences when differential allelic effects exist.

These estimators have further been delivered as open-source software to the community in two ways. First via the TMLE.jl package, released in the official Julia registry and its companion command-line executable TMLECLI.jl (linux only). These software are compatible with any tabular dataset, and can be used outside of the context of population genetics. The more focused TarGene software, is an end-to-end scalable and reproducible Nextflow pipeline for the estimation of genetic effects on human traits. It is already in use across the lab, enabling the research of three other PhD students and two postdoctoral researcher.

In this thesis, we saw that semi-parametric estimators exist in various flavours each with slightly different strengths and weaknesses. One particular variation which has not been studied here is the collaborative targeted minimum loss-based estimator (c-TMLE). Instead of only optimising the outcome model towards the effect of interest, the c-TMLE also selects an optimal propensity score. This could be extremely beneficial in order to mitigate the issues related to positivity violations. For instance, it was reported that adaptive propensity score truncation led to both improved point estimation and coverage results when positivity was practically violated [73].

Finally, the present work focused on the two quantities of interest in statistical geneticists, the average treatment effect (ATE) and average interaction effect (AIE). Together, these estimands can answer a large variety of causal questions aiming to improve human health. However, we saw that conditional effects would likely help the interpretation and contextualisation of detected interactions. Another related question, which was not tackled here, is that of heritability. It measures the proportion of phenotypic variation that can be explained by genetic variation. It is also known as the fraction of explained variance in the broader statistical literature. This question naturally falls within the Targeted Learning framework presented in this thesis and might require only little work. This is because the numerator of this quantity is the mean squared error which is already the optimisation objective of regression models. A machine-learning estimate

might thus already be targeted. In contrast to the ATE and AIE however, heritability relies on high dimensional $\mathbf{V}$. This means that large, potentially whole genome models need to be built, which we discuss next.

8.2 Identification of Causal Variants

We have seen in chapter 3 that, due to linkage disequilibrium and unobserved variations, causal variants cannot be identified via backdoor adjustment. The emergence of whole genome sequencing data, as recently released for all UK Biobank participants, has the potential to resolve both issues.

Indeed, a speculative whole genome regression model would effectively block all genetic backdoor paths in the causal model of figure 3.2. Such models could take advantage, for instance, of the recent successes of large language models [184]. The high dimension of the input data however, is likely to lead to technical difficulties. More realistically, the approach taken by REGENIE [95], consisting in modelling blocks of genetic variations, represents an easy to implement and efficient alternative. A whole genome models may also benefit from improved computational efficiency. This is because each outcome model would be fitted only once and cached for all other estimation queries. Marginal regression models, required for semi-parametric estimation, can be obtained by averaging whole-genome models across variants. The performance of such strategy however remains to be evaluated.

Similarly, a whole genome based propensity score model would likely improve PCA based findings. Hidden Markov models have been successfully used for imputation and are a natural first approach [93], but large language models could also represent a great opportunity. On this front too, a single model could be fitted once and cached across all estimation tasks, largely reducing the computational burden.

Large, whole genome models thus represent an interesting opportunity for the estimation of genetic effects. However, the increased dimensionality of the input space will likely lead to new challenges as well. Beside the technical engineering difficulties inherent to large dimensions, the increased number of variants will create a natural playground for practical positivity violations. Indeed, if two variants are almost perfectly linked, they are statistically (and causally) indistinguishable. An exact empirical evaluation of such phenomenon will be necessary,

and the precision with which causal variants could be identified may be limited. For this reason, priors, based on experimental evidence will remain necessary.

8.3 Identification of Causal Mechanisms

In chapter 7, we presented such a prior. It is based on the identification of variants that impact biological pathways, namely transcription factor's binding. We also showed that their implication in human trait could be strengthened via statistical interactions. However, the loss of statistical power, both inherent to higher-order interactions, and associated multiple testing burden, required further priors. In this thesis we used trans-acting variants curated from the literature to drastically reduce that space for the vitamin D receptor. Modern tools, like ChatGPT [112] could enable the scaling of the approach to many transcription factors. Alternatively, trans-acting variants could be identified in a more principled way using QTL databases. This is in fact an ongoing project, where all trans-actors are sourced from the eQTLGen database, a database of eQTLs from blood samples [163].

Note that the previous approach does not rely on the physical interaction between genetic variants. This is not a problem since statistical interactions can exist without physical interactions. However, physical interactions could be an interesting novel biological prior. In particular, promoter-enhancer interactions identified from high-throughput chromosome conformation capture (Hi-C) are natural candidates for transcription modulation [129].

With modern experimental and mathematical advancements, the field of population genetics is poised to evolve rapidly in the next decade. We believe the work presented in this thesis is perfectly aligned with these advancements and will shape this future.

Appendix A

Appendix 1 - Code Samples

A.1 Unit Testing

A.1.1 The clever covariate function

```
function clever_covariate_and_weights(
    Ψ::StatisticalCMCompositeEstimand,
    Gs::Tuple{Vararg{ConditionalDistributionEstimate}},
    dataset;
    ps_lowerbound=1e-8,
    weighted_fluctuation=false
    )
    T = selectcols(dataset, (p.estimand.outcome for p in Gs))
    indic_vals = indicator_values(indicator_fns(Ψ), T)
    weights = balancing_weights(Gs, dataset;
        ps_lowerbound=ps_lowerbound)
    if weighted_fluctuation
        return indic_vals, weights
    end
    indic_vals .*= weights
    return indic_vals, ones(size(weights, 1))
end
```

A.1.2 A clever covariate unit-test

```
@testset "Test clever_covariate_and_weights: ATE" begin
    Ψ = ATE(
        outcome=:Y,
        treatment_values=(T=(case="a", control="b"),),
        treatment_confounders=(T=[:W],),
    )
    dataset = (
        T = categorical(["a", "b", "c", "a", "a", "b", "a"]),
        Y = [1., 2., 3, 4, 5, 6, 7],
        W = rand(7),
    )
    distr_estimator = TMLE.MLConditionalDistributionEstimator(
        ConstantClassifier()
    )
    distr_estimate = distr_estimator(
        TMLE.ConditionalDistribution(:T, [:W]),
        dataset,
        verbosity=0
    )
    weighted_fluctuation = true
    cov, w = TMLE.clever_covariate_and_weights(
        Ψ,
        (distr_estimate,),
        dataset;
        weighted_fluctuation=weighted_fluctuation
    )
    @test cov == [1.0, -1.0, 0.0, 1.0, 1.0, -1.0, 1.0]
    @test w == [1.75, 3.5, 7.0, 1.75, 1.75, 3.5, 1.75]
    weighted_fluctuation = false
    cov, w = TMLE.clever_covariate_and_weights(
        Ψ,
        (distr_estimate,),
        dataset;
        weighted_fluctuation=weighted_fluctuation
    )
    @test cov == [1.75, -3.5, 0.0, 1.75, 1.75, -3.5, 1.75]
    @test w == ones(7)
end
```

A.2 Integration Testing

A.2.1 The fluctuation function

```
function MLJBase.fit(model::Fluctuation, verbosity, X, y)
    clever_covariate_and_offset, weights =
        clever_covariate_offset_and_weights(model, X)
    mach = machine(
        one_dimensional_path(scitype(y)),
        clever_covariate_and_offset,
        y,
        weights,
        cache=model.cache
        )
    fit!(mach, verbosity=verbosity)

    fitresult = (
        one_dimensional_path = mach,
        )
    cache = (
        weighted_covariate = clever_covariate_and_offset.covariate .* weights,
        training_expected_value = expected_value(predict(mach))
        )
    return fitresult, cache, nothing
end
```

A.2.2 A fluctuation test

```
@testset "Test Fluctuation reduces loss" begin
    Ψ = ATE(
        outcome=:Y,
        treatment_confounders=(T=[:W],),
        treatment_values=(T=(case="a", control="b"),),
    )
    dataset = DataFrame(
        T = categorical(["a", "b", "c", "a", "a", "b", "a"]),
        Y = [1., 2., 3, 4, 5, 6, 7],
        W = rand(7),
    )
    nuisance_factors = TMLE.CMRelevantFactors(
        TMLE.ConditionalDistribution(:Y, [:T, :W]),
        TMLE.ConditionalDistribution(:T, [:W])
    )
    nuisance_factors_estimator = TMLE.CMRelevantFactorsEstimator(
        nothing,
        (Y=with_encoder(ConstantRegressor()), T = ConstantClassifier())
    )
    nuisance_factors_estimate = nuisance_factors_estimator(
        nuisance_factors,
        dataset
    )
    X = dataset[!, collect(nuisance_factors.outcome_mean.parents)]
    y = dataset[!, nuisance_factors.outcome_mean.outcome]
    mse_initial = sum(
        (TMLE.expected_value(nuisance_factors_estimate.outcome_mean, X) .- y).^2
    )
    expected_weights = [1.75, 3.5, 7., 1.75, 1.75, 3.5, 1.75]
    expected_covariate = [1., -1., 0.0, 1., 1., -1., 1.]
    fluctuation = TMLE.Fluctuation(Ψ, nuisance_factors_estimate;
        weighted=true,
    )
    fitresult, cache, report = MLJBase.fit(fluctuation, 0, X, y)
    ypred = MLJBase.predict(fluctuation, fitresult, X)
    mse_fluct = sum((TMLE.expected_value(ypred) .- y).^2)
    @test mse_fluct < mse_initial
end
```

Bibliography

[1] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[2] Mark van der Laan Alejandro Schuler Yi Li. "Lassoed Tree Boosting". In: *arXiv:2205.10697* (2022).

[3] Majid Alfadhel et al. "Identification of the TTC26 splice variant in a novel complex ciliopathy syndrome with biliary, renal, neurological, and skeletal manifestations". In: *Molecular Syndromology* 12.3 (2021), pp. 133–140.

[4] Richard A Armstrong. "When to use the Bonferroni correction". In: *Ophthalmic and Physiological Optics* 34.5 (2014), pp. 502–508. DOI: 10.1111/opo.12131. URL: https://onlinelibrary.wiley.com/doi/10.1111/opo.12131.

[5] Hugues Aschard. "A perspective on interaction effects in genetic association studies". In: *Genetic epidemiology* 40.8 (2016), pp. 678–688.

[6] Laura Balzer et al. "Estimating effects with rare outcomes and high dimensional covariates: knowledge is power". In: *Epidemiologic methods* 5.1 (2016), pp. 1–18.

[7] Rina Foygel Barber and Emmanuel J Candès. "Controlling the false discovery rate via knockoffs". In: *The Annals of statistics* (2015), pp. 2055–2085.

[8] Michelle Barker et al. "Introducing the FAIR Principles for research software". In: *Scientific Data* 9.1 (2022), p. 622.

[9] William Bateson. "Heredity and variation in modern lights". In: *Darwin and modern science* (1909).

[10] Sjoerd Viktor Beentjes and Ava Khamseh. "Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium". In: *Physical Review E* 102.5 (2020), p. 053314.

[11] Yoav Benjamini and Yosef Hochberg. "On the adaptive control of the false discovery rate in multiple testing with independent statistics". In: *Journal of educational and Behavioral Statistics* 25.1 (2000), pp. 60–83.

[12] David Benkeser and Mark Van Der Laan. "The highly adaptive lasso estimator". In: *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE. 2016, pp. 689–696.

[13] Tomaz Berisa and Joseph K Pickrell. "Approximately independent linkage disequilibrium blocks in human populations". In: *Bioinformatics* 32.2 (2016), p. 283.

[14] Peter J Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Springer, 1993.

[15] Christopher M Bishop. "Mixture density networks". In: (1994).

[16] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. "An expanded view of complex traits: from polygenic to omnigenic". In: *Cell* 169.7 (2017), pp. 1177–1186.

[17] Carol Brayne and Terrie E Moffitt. "The limitations of large-scale volunteer databases to address inequalities and global challenges in health and aging". In: *Nature Aging* 2.9 (2022), pp. 775–783.

[18] Ben Brumpton et al. "Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses". In: *Nature communications* 11.1 (2020), pp. 1–13.

[19] Annalisa Buniello et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic acids research* 47.D1 (2019), pp. D1005–D1012.

[20] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (2018), pp. 203–209.

[21] Oriol Canela-Xandri, Konrad Rawlik, and Albert Tenesa. "An atlas of genetic associations in UK Biobank". In: *Nature genetics* 50.11 (2018), pp. 1593–1599.

[22] Siwei Chen et al. “A genomic mutational constraint map using variation in 76,156 human genomes”. In: *Nature* 625.7993 (2024), pp. 92–100.

[23] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[24] Xiaohong Chen. “Large sample sieve estimation of semi-nonparametric models”. In: *Handbook of econometrics* 6 (2007), pp. 5549–5632.

[25] Zhongsheng Chen et al. “Revisiting the genome-wide significance threshold for common variant GWAS”. In: *G3* 11.2 (2021), jkaa056.

[26] Victor Chernozhukov et al. *Double/debiased machine learning for treatment and structural parameters*. 2018.

[27] Statistics group Ciampi Antonio 8 Greenwood Celia MT (co-chair) 7 8 14 19 Hendricks Audrey E. 1 12 Li Rui 7 13 14 Metrustry Sarah 5 Oualkacha Karim 80 Tachmazidou Ioanna 1 Xu ChangJiang 7 8 Zeggini Eleftheria (co-chair) 1 et al. “The UK10K project identifies rare variants in health and disease”. In: *Nature* 526.7571 (2015), pp. 82–90.

[28] Melina Claussnitzer et al. “A brief history of human disease genetics”. In: *Nature* 577 (2020), pp. 179–189. DOI: 10.1038/s41586-019-1879-7. URL: https://www.nature.com/articles/s41586-019-1879-7.

[29] Melina Claussnitzer et al. “FTO Obesity Variant Circuitry and Adipocyte Browning in Humans”. In: *The New England journal of medicine* 373.10 (Sept. 2015), pp. 895–907. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1502214.

[30] 1000 Genomes Project Consortium et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), p. 68.

[31] Jeremy Coyle. *tmle3: The Extensible TMLE Framework*. R package version 0.2.0. 2021. DOI: 10.5281/zenodo.4603358. URL: https://github.com/tlverse/tmle3.

[32] Claire Dandine-Roulland and Hervé Perdry. “The use of the linear mixed model in human genetics”. In: *Human heredity* 80.4 (2016), pp. 196–206.

[33] Gholamreza Daryabor, Nasser Gholijani, and Fatemeh Rezaei Kahmini. “A review of the critical role of vitamin D axis on the immune system”. In: *Experimental and Molecular Pathology* 132 (2023), p. 104866.

[34] Sayantan Das et al. "Next-generation genotype imputation service and methods". In: *Nature genetics* 48.10 (2016), pp. 1284–1287.

[35] Molly M. Davies and Mark J. van der Laan. "Sieve Plateau Variance Estimators: A New Approach to Confidence Interval Estimation for Dependent Data". In: *U.C. Berkeley Division of Biostatistics Working Paper Series* Working Paper 322 (2014). URL: https://biostats.bepress.com/ucbbiostat/paper322.

[36] Angus Deaton and Nancy Cartwright. "Understanding and misunderstanding randomized controlled trials". In: *Social science & medicine* 210 (2018), pp. 2–21.

[37] Yuetiva Deming et al. "Genome-wide association study identifies four novel loci associated with Alzheimer's endophenotypes and disease modifiers". In: *Acta neuropathologica* 133.5 (2017), pp. 839–856.

[38] Paolo Di Tommaso et al. "Nextflow enables reproducible computational workflows". In: *Nature biotechnology* 35.4 (2017), pp. 316–319.

[39] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[40] Eran Elhaik. "Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated". In: *Scientific Reports* 12.1 (2022), p. 14683.

[41] Connor A Emdin, Amit V Khera, and Sekar Kathiresan. "Mendelian randomization". In: *Jama* 318.19 (2017), pp. 1925–1926.

[42] Douglas L Falls. "Neuregulins: functions, forms, and signaling strategies". In: *The EGF Receptor Family* (2003), pp. 15–31.

[43] Max H Farrell, Tengyuan Liang, and Sanjog Misra. "Deep neural networks for estimation and inference: application to causal effects and other semiparametric estimands". In: *arXiv preprint arXiv:1809.09953* 20 (2018).

[44] Rubén Fernández-Santiago et al. "SNCA and mTOR pathway single nucleotide polymorphisms interact to modulate the age at onset of Parkinson's disease". In: *Movement Disorders* 34.9 (2019), pp. 1333–1344.

[45] Luisa Turrin Fernholz. *Von Mises calculus for statistical functionals*. Vol. 19. Springer Science & Business Media, 2012.

[46] Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1999.

[47] Martin Fowler, Jim Highsmith, et al. "The agile manifesto". In: *Software development* 9.8 (2001), pp. 28–35.

[48] Timothy M. Frayling et al. "A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity". In: *Science* 316.5826 (May 2007), pp. 889–894. DOI: 10.1126/science.1141634.

[49] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1 (2010), p. 1.

[50] Anna Fry et al. "Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population". In: *American journal of epidemiology* 186.9 (2017), pp. 1026–1034.

[51] Christian Fuchsberger et al. "The genetic architecture of type 2 diabetes". In: *Nature* 536.7614 (2016), pp. 41–47.

[52] E. García-Portugués. *Notes for Predictive Modeling*. Version 5.10.1. ISBN 978-84-09-29679-8. 2024. URL: https://bookdown.org/egarpor/PM-UC3M/.

[53] Andrew Gelman. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

[54] Richard A Gibbs et al. "The international HapMap project". In: (2003).

[55] Peter B Gilbert et al. "Immune correlates analysis of the mRNA-1273 COVID-19 vaccine efficacy clinical trial". In: *Science* 375.6576 (2022), pp. 43–50.

[56] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[57] Susan Gruber et al. "Targeted learning: toward a future informed by real-world evidence". In: *Statistics in Biopharmaceutical Research* 16.1 (2024), pp. 11–25.

[58] Kevin L Gunderson et al. “A genome-wide scalable SNP genotyping assay using microarray technology”. In: *Nature genetics* 37.5 (2005), pp. 549–554.

[59] Ashraful Haque et al. “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications”. In: *Genome medicine* 9 (2017), pp. 1–12.

[60] Eduardo Hariton and Joseph J Locascio. “Randomised controlled trials—the gold standard for effectiveness research”. In: *BJOG: an international journal of obstetrics and gynaecology* 125.13 (2018), p. 1716.

[61] Diane V Havlir et al. “HIV testing and treatment with the use of a community health approach in rural Africa”. In: *New England Journal of Medicine* 381.3 (2019), pp. 219–229.

[62] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1998.

[63] Philip W. Hedrick. “What is the evidence for heterozygote advantage selection?” In: *Trends in Ecology & Evolution* 27.12 (2022/09/11 2012), pp. 698–704.

[64] Farhad Hormozdiari et al. “Colocalization of GWAS and eQTL signals detects target genes”. In: *The American Journal of Human Genetics* 99.6 (2016), pp. 1245–1260.

[65] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Multilayer feedforward networks are universal approximators”. In: *Neural networks* 2.5 (1989), pp. 359–366.

[66] Laurence J. Howe et al. “Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects”. In: *Nature Genetics* 54.5 (2022), pp. 581–592. DOI: 10.1038/s41588-022-01062-7. URL: https://www.nature.com/articles/s41588-022-01062-7.

[67] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. “A flexible and accurate genotype imputation method for the next generation of genome-wide association studies”. In: *PLoS genetics* 5.6 (2009), e1000529.

[68] David Hume. “An enquiry concerning human understanding”. In: *Seven masterpieces of philosophy*. Routledge, 2016, pp. 183–276.

[69] Elzbieta Janda et al. “Autophagy and neuroprotection in astrocytes exposed to 6-hydroxydopamine is negatively regulated by NQO2: Relevance to parkinson’s disease”. In: *Scientific Reports* 13.1 (2023), p. 21624.

[70] Elzbieta Janda et al. “Molecular pharmacology of NRH: quinone oxidoreductase 2: A detoxifying enzyme acting as an undercover toxifying enzyme”. In: *Molecular pharmacology* 98.5 (2020), pp. 620–633.

[71] Longda Jiang et al. “A resource-efficient tool for mixed model association analysis of large-scale data”. In: *Nature genetics* 51.12 (2019), pp. 1749–1755.

[72] Xia Jiang et al. “Genome-Wide Association Study in 79,366 European-ancestry Individuals Informs the Genetic Architecture of 25-Hydroxyvitamin D Levels”. In: *Nature Communications* 9.1 (Jan. 2018), p. 260. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02662-2.

[73] Cheng Ju, Joshua Schwab, and Mark J van der Laan. “On adaptive propensity score truncation in causal inference”. In: *Statistical methods in medical research* 28.6 (2019), pp. 1741–1760.

[74] Masahiro Kanai et al. “Analysis of genetic dominance in the UK Biobank”. In: *Science* 376.6598 (2022), eabn8455. DOI: 10.1126/science.abn8455. URL: https://www.science.org/doi/full/10.1126/science.abn8455.

[75] Heeseog Kang et al. “Kcnn4 is a regulator of macrophage multinucleation in bone homeostasis and inflammatory disease”. In: *Cell reports* 8.4 (2014), pp. 1210–1224.

[76] Edward H Kennedy. “Semiparametric doubly robust targeted double machine learning: a review”. In: *arXiv preprint arXiv:2203.06469* (2022).

[77] Edward H Kennedy, Sivaraman Balakrishnan, and LA Wasserman. “Semiparametric counterfactual density estimation”. In: *Biometrika* 110.4 (2023), pp. 875–896.

[78] David C Klonoff. “The new FDA real-world evidence program to support development of drugs and biologics”. In: *Journal of diabetes science and technology* 14.2 (2020), pp. 345–349.

[79] Mark J. van der Laan. “A generally efficient targeted minimum loss based estimator based on the highly adaptive Lasso”. In: *Int. J. Biostat.* 13.2 (2017), pp. 20150097, 35.

[80] Bowen Lai et al. “Skeletal ciliopathy: pathogenesis and related signaling pathways”. In: *Molecular and Cellular Biochemistry* 479.4 (2024), pp. 811–823.

[81] Dang LE et al. “A causal roadmap for generating high-quality real-world evidence”. In: *J Clin Transl Sci.* (2023). DOI: 10.1017/cts.2023.635.

[82] Xihong Lin. “Learning Lessons on Reproducibility and Replicability in Large Scale Genome-Wide Association Studies”. In: *Harvard Data Science Review* 2.4 (2020), 10.1162/99608f92.33703976. DOI: 10.1162/99608f92.33703976. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10869125/.

[83] Yi Lin et al. “Regulatory role of KCa3. 1 in immune cell function and its emerging association with rheumatoid arthritis”. In: *Frontiers in Immunology* 13 (2022), p. 997621.

[84] Weiwei Liu, S Janet Kuramoto, and Elizabeth A Stuart. “An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research”. In: *Prevention science* 14 (2013), pp. 570–580.

[85] Po-Ru Loh et al. “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. In: *Nature genetics* 47.3 (2015), pp. 284–290.

[86] Po-Ru Loh et al. “Reference-based phasing using the Haplotype Reference Consortium panel”. In: *Nature genetics* 48.11 (2016), pp. 1443–1448.

[87] John Lonsdale et al. “The genotype-tissue expression (GTEx) project”. In: *Nature genetics* 45.6 (2013), pp. 580–585.

[88] Claudia Lucchinetti et al. “Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination”. In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 47.6 (2000), pp. 707–717.

[89] Ye Luo, Martin Spindler, and Jannis Kück. “High-Dimensional L_2 Boosting: Rate of Convergence”. In: *arXiv preprint arXiv:1602.08927* (2016).

[90] Jacqueline MacArthur et al. “The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)”. In: *Nucleic acids research* 45.D1 (2017), pp. D896–D901.

[91] Anthony D Maher and Philip W Kuchel. "The Gardos channel: a review of the Ca2+-activated K+ channel in human erythrocytes". In: *The international journal of biochemistry & cell biology* 35.8 (2003), pp. 1182–1197.

[92] Peter J Malloy, J Wesley Pike, and David Feldman. "The vitamin D receptor and the syndrome of hereditary 1, 25-dihydroxyvitamin D-resistant rickets". In: *Endocrine reviews* 20.2 (1999), pp. 156–188.

[93] Jonathan Marchini and Bryan Howie. "Genotype imputation for genome-wide association studies". In: *Nature Reviews Genetics* 11.7 (2010), pp. 499–511.

[94] Andries T Marees et al. "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis". In: *International journal of methods in psychiatric research* 27.2 (2018), e1608.

[95] Joelle Mbatchou et al. "Computationally efficient whole-genome regression for quantitative and binary traits". In: *Nature genetics* 53.7 (2021), pp. 1097–1103.

[96] Zachary R McCaw et al. "DeepNull models non-linear covariate effects to improve phenotypic prediction and association power". In: *Nature communications* 13.1 (2022), p. 241.

[97] Michael Menzel et al. "NoPeak: k-mer-based motif discovery in ChIP-Seq data without peak calling". In: *Bioinformatics* 37.5 (2021), pp. 596–602.

[98] Andrew Mertens et al. "Causes and consequences of child growth faltering in low-resource settings". In: *Nature* 621.7979 (2023), pp. 568–576.

[99] Michael L Metzker. "Sequencing technologies—the next generation". In: *Nature reviews genetics* 11.1 (2010), pp. 31–46.

[100] Lauren E Mokry et al. "Vitamin D and risk of multiple sclerosis: a Mendelian randomization study". In: *PLoS medicine* 12.8 (2015), e1001866.

[101] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[102] Jason H Moore. "The ubiquitous nature of epistasis in determining susceptibility to common human diseases". In: *Human heredity* 56.1-3 (2003), pp. 73–82.

[103] Lisa D Moore, Thuc Le, and Guoping Fan. "DNA methylation and its basic function". In: *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.

[104] Michael D Morgan et al. "Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability". In: *Nature communications* 9.1 (2018), p. 5271.

[105] Michael D. Morgan et al. "Genome-Wide Study of Hair Colour in UK Biobank Explains Most of the SNP Heritability". In: *Nature Communications* 9 (Dec. 2018), p. 5271. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07691-z.

[106] Norihiko Nakano et al. "NTAK/neuregulin-2 secreted by astrocytes promotes survival and neurite outgrowth of neurons via ErbB3". In: *Neuroscience letters* 622 (2016), pp. 88–94.

[107] Ardalan Naseri et al. "Personalized genealogical history of UK individuals inferred from biobank-scale IBD segments". In: *BMC biology* 19 (2021), pp. 1–12.

[108] NICE. *NICE real-world evidence framework (ECD9)*. Tech. rep. NICE, 2024.

[109] Sergey Nurk et al. "The complete sequence of a human genome". In: *Science* 376.6588 (2022), pp. 44–53.

[110] Shinya Oki et al. "Ch IP-Atlas: a data-mining suite powered by full integration of public Ch IP-seq data". In: *EMBO reports* 19.12 (2018), e46255.

[111] Maynard V Olson. "The human genome project." In: *Proceedings of the National Academy of Sciences* 90.10 (1993), pp. 4338–4344.

[112] OpenAI. *ChatGPT: GPT-4*. Large language model. 2023. URL: https://chat.openai.com/.

[113] World Health Organization. *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. Vol. 1. World Health Organization, 1992.

[114] Michael J Osborne et al. "eIF4E3 acts as a tumor suppressor by utilizing an atypical mode of methyl-7-guanosine cap recognition". In: *Proceedings of the National Academy of Sciences* 110.10 (2013), pp. 3877–3882.

[115] Harsh Parikh et al. "Validating causal inference methods". In: *International conference on machine learning*. PMLR. 2022, pp. 17346–17358.

[116] Peter J Park. "ChIP–seq: advantages and challenges of a maturing technology". In: *Nature reviews genetics* 10.10 (2009), pp. 669–680.

[117] Judea Pearl. *Causality*. Cambridge university press, 2009.

[118] Maya L Petersen et al. "Diagnosing and responding to violations in the positivity assumption". In: *Statistical methods in medical research* 21.1 (2012), pp. 31–54.

[119] Nancie Petrucelli, Mary B Daly, and Tuya Pal. "BRCA1-and BRCA2-associated hereditary breast and ovarian cancer". In: (2022).

[120] Rachael V. Phillips et al. *Practical Considerations for Specifying a Super Learner*. Apr. 2022. DOI: 10.48550/arXiv.2204.06139. arXiv: 2204.06139 [stat].

[121] Lutz Prechelt. "Automatic early stopping using cross validation: quantifying the criteria". In: *Neural networks* 11.4 (1998), pp. 761–767.

[122] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature genetics* 38.8 (2006), pp. 904–909.

[123] Sreeram V Ramagopalan et al. "A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution". In: *Genome research* 20.10 (2010), pp. 1352–1360.

[124] James Bernard Ramsey. "Tests for specification errors in classical linear least-squares regression analysis". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 31.2 (1969), pp. 350–371.

[125] Chris M Rands et al. "8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage". In: *PLoS genetics* 10.7 (2014), e1004525.

[126] Ho Sung Rhee and B Franklin Pugh. "ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy". In: *Current protocols in molecular biology* 100.1 (2012), pp. 21–24.

[127] Neil Risch and Kathleen Merikangas. "The future of genetic studies of complex human diseases". In: *Science* 273.5281 (1996), pp. 1516–1517. DOI: 10.1126/science.273.5281.1516. URL: https://www.science.org/doi/10.1126/science.273.5281.1516.

[128] James Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical modelling* 7.9-12 (1986), pp. 1393–1512.

[129] Gil Ron et al. "Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains". In: *Nature communications* 8.1 (2017), p. 2237.

[130] Donald B Rubin. "Causal inference using potential outcomes: Design, modeling, decisions". In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331.

[131] Albin Sandelin et al. "JASPAR: an open-access database for eukaryotic transcription factor binding profiles". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D91–D94.

[132] Ines de Santiago et al. "BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes". In: *Genome biology* 18 (2017), pp. 1–17.

[133] Iqbal H Sarker. "Machine learning: Algorithms, real-world applications and research directions". In: *SN computer science* 2.3 (2021), p. 160.

[134] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. "From genome-wide associations to candidate causal variants by statistical fine-mapping". In: *Nature Reviews Genetics* 19.8 (2018), pp. 491–504.

[135] Alejandro Schuler et al. "Synth-validation: Selecting the best causal inference method for a given dataset". In: *arXiv preprint arXiv:1711.00083* (2017).

[136] Matteo Sesia, Chiara Sabatti, and Emmanuel J Candès. "Gene hunting with knockoffs for hidden markov models". In: *arXiv preprint arXiv:1706.04677* (2017).

[137] Matteo Sesia et al. “False discovery rate control in genome-wide association studies with population structure”. In: *Proceedings of the National Academy of Sciences* 118.40 (2021), e2105841118.

[138] Ranad Shaheen et al. “Biallelic mutations in tetratricopeptide repeat domain 26 (intraflagellar transport 56) cause severe biliary ciliopathy in humans”. In: *Hepatology* 71.6 (2020), pp. 2067–2079.

[139] Jacob S Sherkow. “Regulatory sandboxes and the public health”. In: *U. Ill. L. Rev.* (2022), p. 357.

[140] Stephen T Sherry et al. “dbSNP: the NCBI database of genetic variation”. In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[141] Pankhuri Singhal et al. “Evidence of epistasis in regions of long-range linkage disequilibrium across five complex diseases in the UK Biobank and eMERGE datasets”. In: *The American Journal of Human Genetics* 110.4 (2023), pp. 575–591.

[142] Nayanah Siva. “1000 Genomes project.” In: *Nature biotechnology* 26.3 (2008), pp. 256–257.

[143] Montgomery Slatkin. “Linkage disequilibrium—understanding the evolutionary past and mapping the medical future”. In: *Nature Reviews Genetics* 9.6 (2008), pp. 477–485.

[144] Oleg Sofrygin and Mark J van der Laan. “Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population”. In: *Journal of causal inference* 5.1 (2017), p. 20160003.

[145] Lawrence Steinman. “Multiple sclerosis: a coordinated immunological attack against myelin in the central nervous system”. In: *Cell* 85.3 (1996), pp. 299–302.

[146] Wei-Ming Su et al. “Systematic druggable genome-wide Mendelian randomisation identifies therapeutic targets for Alzheimer’s disease”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 94.11 (2023), pp. 954–961.

[147] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *PLoS medicine* 12.3 (2015), e1001779.

[148] Patrick Sulem et al. "Genetic Determinants of Hair, Eye and Skin Pigmentation in Europeans". In: *Nature Genetics* 39.12 (Dec. 2007), pp. 1443–1452. ISSN: 1546-1718. DOI: 10.1038/ng.2007.13.

[149] Benjamin B Sun et al. "Genomic atlas of the human plasma proteome". In: *Nature* 558.7708 (2018), pp. 73–79.

[150] Vasilis Syrgkanis and Manolis Zampetakis. "Estimation and inference with trees and forests in high dimensions". In: *Conference on learning theory*. PMLR. 2020, pp. 3453–3454.

[151] Rocío Titiunik. "Can big data solve the fundamental problem of causal inference?" In: *PS: Political Science & Politics* 48.1 (2015), pp. 75–79.

[152] Matthew J Tudball, George Davey Smith, and Qingyuan Zhao. "Almost exact Mendelian randomization". In: *arXiv preprint arXiv:2208.14035* (2022).

[153] Alwan N.A. Twaits A. "The association between area-based deprivation and change in body-mass index over time in primary school children: a population-based cohort study in Hampshire, UK." In: *International Journal of Obesity* 44 (2020), pp. 628–636.

[154] Emil Uffelmann et al. "Genome-wide association studies". In: *Nature Reviews Methods Primers* 1.1 (2021), p. 59.

[155] James Uniacke et al. "An oxygen-regulated switch in the protein synthesis machinery". In: *Nature* 486.7401 (2012), pp. 126–129.

[156] All of Us Research Program Investigators. "The "All of Us" research program". In: *New England Journal of Medicine* 381.7 (2019), pp. 668–676.

[157] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1 (2007).

[158] Mark J Van der Laan and Sherri Rose. *Targeted learning in data science*. Springer, 2018.

[159] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[160] Tyler J. VanderWeele and Mirjam J.. Knol. "A Tutorial on Interaction". In: *Epidemiologic Methods* 3.1 (2014), pp. 33–72. DOI: https://doi.org/10.1515/em-2013-0005.

[161] Peter M Visscher et al. “10 years of GWAS discovery: biology, function, and translation”. In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[162] Laurent Volpon et al. “eIF4E3, a new actor in mRNA metabolism and tumor suppression”. In: *Cell Cycle* 12.8 (2013), pp. 1159–1160.

[163] Urmo Võsa et al. “Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression”. In: *Nature genetics* 53.9 (2021), pp. 1300–1310.

[164] Matthias Wacker and Michael F Holick. “Vitamin D—effects on skeletal and extraskeletal health and the need for supplementation”. In: *Nutrients* 5.1 (2013), pp. 111–148.

[165] Stefan Wager. *Stats 361: Causal inference*. Tech. rep. Technical report, Technical report, Stanford University, 2020. URL: https …, 2020.

[166] Andreas Wagner. “Robustness against mutations in genetic networks of yeast”. In: *Nature genetics* 24.4 (2000), pp. 355–361.

[167] Yixin Wang and David M Blei. “The blessings of multiple causes”. In: *Journal of the American Statistical Association* 114.528 (2019), pp. 1574–1596.

[168] Kyoko Watanabe et al. “A global overview of pleiotropy and genetic architecture in complex traits”. In: *Nature genetics* 51.9 (2019), pp. 1339–1348.

[169] James D Watson and Francis HC Crick. “Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid”. In: *Nature* 171.4356 (1953), pp. 737–738.

[170] Wen-Hua Wei, Gibran Hemani, and Chris S. Haley. “Detecting epistasis in human complex traits”. In: *Nature Reviews Genetics* 15.11 (2014), pp. 722–733.

[171] Joachim Weischenfeldt et al. “NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements”. In: *Genes & development* 22.10 (2008), pp. 1381–1396.

[172] Benjamin Weiss et al. “eIF4E3 forms an active eIF4F complex during stresses (eIF4FS) targeting mTOR and re-programs the translatome”. In: *Nucleic Acids Research* 49.9 (2021), pp. 5159–5176.

[173] Eric W Weisstein. “Bonferroni correction”. In: *https://mathworld. wolfram. com/* (2004).

[174] Mark D Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.1 (2016), pp. 1–9.

[175] Michael Windle. “Effect of parental drinking on adolescents”. In: *Alcohol Health and Research World* 20.3 (1996), p. 181.

[176] Thea K Wöbke, Bernd L Sorg, and Dieter Steinhilber. “Vitamin D in inflammatory diseases”. In: *Frontiers in physiology* 5 (2014), p. 244.

[177] David H Wolpert. “Stacked generalization”. In: *Neural networks* 5.2 (1992), pp. 241–259.

[178] Andrew R. Wood et al. “Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively”. In: *Diabetologia* 59.6 (2016), pp. 1214–1221.

[179] Jian Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. In: *The American Journal of Human Genetics* 88.1 (2011), pp. 76–82.

[180] Yuling Yao et al. “Bayesian hierarchical stacking: Some models are (somewhere) useful”. In: *Bayesian Analysis* (2021). DOI: 10.1214/21-BA1287. URL: https://arxiv.org/abs/2101.08954.

[181] Loïc Yengo et al. “A saturated map of common genetic variants associated with human height”. In: *Nature* 610.7933 (2022), pp. 704–712.

[182] Jianming Yu et al. “A unified mixed-model method for association mapping that accounts for multiple levels of relatedness”. In: *Nature genetics* 38.2 (2006), pp. 203–208.

[183] Noah Zaitlen and Peter Kraft. “Improving genetic prediction by leveraging genetic correlations among human diseases and traits”. In: *Nature Communications* 11 (2020). DOI: 10.1038/s41467-019-13882-z. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6942007/.

[184] Wayne Xin Zhao et al. “A survey of large language models”. In: *arXiv preprint arXiv:2303.18223* (2023).

[185] Xiang Zhou and Matthew Stephens. “Genome-wide efficient mixed-model analysis for association studies”. In: *Nature genetics* 44.7 (2012), pp. 821–824.