Year : 2021

Mendelian randomization and its application to genome-wide association studiesSarcopenia and major complications in patients undergoing oncologic colon surgery

Mounier Ninon

Mounier Ninon, 2021, Mendelian randomization and its application to genome-wide association studiesSarcopenia and major complications in patients undergoing oncologic colon surgery

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_03589C92E22E0

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Unisanté, Centre universitaire de médecine générale et santé publique
Département formation, recherche et innovation

Mendelian randomization and its application to genome-wide association studies

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

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

par

Ninon MOUNIER

Ingénieure diplômée de l'Institut National des Sciences Appliquées de Lyon

Jury

Prof. Francesco Panese, Président
Prof. Zoltán Kutalik, Directeur de thèse
Prof. Jacques Fellay, Expert
Prof. Matthew Robinson, Expert

Lausanne
(2021)

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Unisanté, Centre universitaire de médecine générale et santé publique
Département formation, recherche et innovation

Mendelian randomization and its application to genome-wide association studies

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

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

par

Ninon MOUNIER

Ingénieure diplômée de l'Institut National des Sciences Appliquées de Lyon

Jury

Prof. Francesco Panese, Président
Prof. Zoltán Kutalik, Directeur de thèse
Prof. Jacques Fellay, Expert
Prof. Matthew Robinson, Expert

Lausanne
(2021)

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

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

Président·e	Monsieur	Prof.	Francesco	**Panese**
Directeur·trice de thèse	Monsieur	Prof.	Zoltán	**Kutalik**
Expert·e·s	Monsieur	Prof.	Matthew	**Robinson**
	Monsieur	Prof.	Jacques	**Fellay**

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

Madame Ninon Sarah Mounier

Master, Université Claude Bernard / Lyon I, France

intitulée

Mendelian randomization and its application to genome-wide association studies

Date de l'examen : 10 août 2021

Date d'émission de l'imprimatur : Lausanne, le 09.09.2021

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Niko GELDNER
Directeur de l'Ecole Doctorale

Acknowledgements

You would think that after four and a half years of PhD, after all the ups and downs, after spending weeks working on this manuscript, writing and re-writing the different parts, writing the acknowledgments should be easy, right? Well, this is not. Because there are many people to thank and many things to say.

A PhD is a journey. It is sometimes difficult, always challenging but it can also be very rewarding. This journey is coming to an end and when looking back at it, I can not say how glad and how proud I am to have been able to take it. Here, I would like to thank all the people who contributed to make this journey possible and memorable. First and foremost, I am incredibly grateful to my supervisor, **Zoltán**. Thank you for giving me a chance to take this journey. Thank you for seeing that behind my shyness there was someone who also was motivated and really keen on coming aboard. You are a brilliant scientist and a remarkable human being. I have been very lucky and privileged to have you as a supervisor. In the past years, you have been understanding, kind and supportive. You helped me overcome my fears, you gave me intellectual freedom in my work and invaluable advice. I have learnt so much from working with you, thank you!

This has not been a solitary journey and I had the chance to have been surrounded by amazing people. I can not mention everyone here, but **past and current members of the Statistical Genetics Group**, thank you! I am particularly grateful to **Sina** and **Eleonora** who warmly welcomed me in the group when I joined and helped my find my place. Thank you for being such inspirational role models and for reminding me that anything is possible when you really want it. **Jonathan** and I started this journey together. We don't know where our next journeys will take us, but it has been a pleasure going through every step of this one with you. And finally, this journey took a different direction when **Liza** joined the group. You are an incredibly kind person, bringing joy wherever you go. I am happy to say that I found in you someone I could rely on and share my thoughts with, someone who has always been there to encourage me when I needed it, a friend.

Et puis, il y a toutes ces personnes qui n'ont pas directement fait partie de l'aventure, mais qui ont été présentes pour moi tout au long de celle-ci:

- Mes amis, qui ont eu la présence d'esprit de ne pas me demander comment se passait ma thèse *à chaque fois* que l'on se voyait et qui ont eu la gentillesse de toujours m'écouter quand j'avais besoin de parler. Merci d'être là, même si parfois nos vies prennent des directions différentes.
- Ma famille, en particuler **mes parents et ma soeur, Fantine**. Je n'ai jamais réussi à vous expliquer clairement ce que je faisais, mais vous avez toujours été là pour moi, bien avant que je commence ma thèse. Vous avez soutenu chacune de mes décisions et savoir que vous êtes fiers de moi vaut bien plus que tous les diplômes du monde.
- **Florent**. Tu es certainement la personne qui me connait le mieux et tu as été présent à chaque étape de cette aventure. Tu étais déjà passé par là, et tu savais à quel point cela pourrait être difficile pour moi par moments. Tu m'as écoutée parler de mes projets en boucle, tu as compris l'importance de lancer mes scripts sur le serveur le vendredi soir, tu m'as laissé me plaindre à chaque fois que quelque chose ne marchait pas. Tu m'as soutenue quand j'en avais besoin et je n'aurais pas pu y arriver sans toi. Merci d'avoir été là, merci d'être cette personne et merci d'avoir fait de **Baloo** et moi ta famille.

Abstract

Genetics aims is to study heredity: how traits are passed from one generation to the next and how genetic variations can lead to changes in phenotypes. Some phenotypes, called complex or quantitative traits, are under the control of both genetic and environmental factors. Examples of complex traits include quantitative phenotypes, such as height or cholesterol levels, as well as certain diseases, like diabetes or cardiovascular diseases. Genome-wide association studies (GWASs) are used to statistically test for the association between each genetic variant and a given phenotype. These studies confirmed that most complex traits are influenced by a large number of genetic variants, often exhibiting small effect sizes that can only be detected using large numbers of individuals. They also permitted the estimation of narrow-sense heritability, which is the proportion of phenotypic variation that can be attributed to these genetic variations. The results of such GWASs (association results for every genetic variant) are often made publicly available and they can be used to perform follow-up analyses, for example Mendelian randomization. Mendelian randomization aims at investigating causal relationships between complex traits and estimating the causal effect of one exposure on an outcome. This method mimicks randomized controlled trials and takes advantage of the fact that genetic variations are randomly distributed across the population. By using association results for genetic variants strongly associated with a given risk factor and measuring the effect of these variants on another trait or disease, Mendelian randomization can infer the existence and the strength of the causal relationship between them. Analyses helping to understand the genetics underlying complex traits and the relationships between them are key to precision medicine. Precision medicine is an approach that takes into account the genome sequence and the environmental exposures of each patient, to provide personalized prevention and treatment to each individual.

During my thesis, I have been involved in several projects aiming at developing statistical methods that rely on Mendelian randomization. In the first part, I worked on a Bayesian GWAS approach (*bGWAS*). The goal of this approach is to increase statistical power to discover variants associated with a trait by leveraging data from correlated risk factors. The idea is to combine (1) the causal effects of the risk factors on the trait of interest (estimated using Mendelian randomization) with (2) the association results of genetic variants with these risk factors, in order to estimate the prior effect of each variant on the trait of interest. This approach has been used to study the genetics underlying lifespan, taking into account various potential risk factors, such as body mass index, cholesterol levels, and several diseases for example. In the second part, I worked on developing Mendelian randomization extensions (*MRlap* and *LHC-MR*) that aim at tackling some of the most common sources of bias. These extensions allow for more robust causal effect estimates, when some of the Mendelian randomization assumptions are violated, as well as for an extension of the scope of application of Mendelian randomization.

Résumé

La génétique est l'étude de la transmission de traits héréditaires au sein d'une population. Un défi majeur de la génétique moderne est cependant d'expliquer le mécanisme exact par lequel les variations génétiques peuvent, ou non, se traduire par des variations phénotypiques. Ce défi est d'autant plus important dans le cas des traits dits «complexes», qui sont affectés à la fois par des facteurs génétiques et par des facteurs environnementaux. C'est le cas par exemple de la taille adulte, du taux de cholestérol ou encore de certaines maladies, comme le diabète. Les études d'association pangénomique, en anglais *genome-wide association studies* (GWASs), permettent de tester si des variants génétiques sont statistiquement associés à un phénotype donné. Ces études ont confirmé que la plupart des traits complexes sont influencés par un très large nombre de variants génétiques, dont chacun a souvent un faible effet qui n'aurait pas été détecté sans l'accès à de larges jeux de données. Elles ont également permis d'estimer la part de la variation phénotypique expliquée par l'ensemble des variants (héritabilité au sens étroit). Les résultats de ces GWASs sont souvent publiés sous forme de statistiques synthétiques (pour chaque variant génétique) qui peuvent être utilisées pour réaliser des analyses additionnelles, notamment des analyses de randomisation mendélienne. Celles-ci permettent d'étudier les relations de cause à effet entre différents traits complexes et d'estimer l'effet de causalité d'un trait sur un autre. Les variations génétiques étant théoriquement réparties de façon aléatoire dans une population, la randomisation mendélienne est une alternative aux essais cliniques randomisés. En utilisant les résultats d'association de variants génétiques associés spécifiquement avec un facteur de risque et en mesurant leurs effets sur un autre trait, la randomisation mendélienne permet d'établir une relation de cause à effet entre deux traits. Ces études, permettant la compréhension des causes génétiques à l'origine des traits complexes ainsi que des relations de cause à effet pouvant exister entre ceux-ci, ouvrent la voie au développement de la médecine de précision, une approche prenant en compte toutes les informations concernant un individu (génétiques et environnementales) pour proposer à chacun un diagnostic et un traitement personnalisés.

Durant mon doctorat, j'ai été impliquée dans différents projets visant à développer des approches techniques basées sur la randomisation mendélienne. Dans un premier temps, j'ai travaillé sur une méthode appelée GWAS bayésien (*bGWAS*). Cette méthode utilise des informations provenant de potentiel facteurs de risques identifiés a priori de façon à augmenter la puissance statistique de l'identification de variants génétiques associés à un trait d'intérêt. L'idée est de combiner (1) les effets de causalité des risques facteurs sur le trait d'intérêt (estimés en utilisant la randomisation mendélienne) et (2) les résultats d'association des variants génétiques avec ces facteurs de risque, pour estimer leur effet a priori sur le trait d'intérêt. Cette méthode a notamment été utilisée pour étudier les causes génétiques influençant l'espérance de vie, en prenant en compte plusieurs facteurs de risques tel que certaines maladies ou encore l'indice de masse corporel. Dans un second temps, j'ai travaillé sur des projets visant à proposer des extensions aux méthodes classiques de randomisation mendélienne (*MRlap* et *LHC-MR*) pour les rendre plus robustes à certaines sources de biais communément observées, avec pour but d'élargir les possibilités d'application de ces méthodes.

List of Abbreviations

General

CD/CV hypothesis : Common Disease / Common Variant hypothesis

CNV : Copy Number Variant

DAG : Directed Acyclic Graph

eQTL : expression Quantitative Trait Loci

GnomAD : Genome Aggregation Database

GRM : Genetic Relationship Matrix

GTEx : Genotype-Tissue Expression

GWAS : Genome Wide Association Study

GxE : Gene-Environment interactions

GxG : Gene-Gene interactions

HRC : Haplotype Reference Consortium

IV : Invstrumental Variable

IVW : Inverse-Variance Weighted

LD : Linkage Disequilibrium

LDSC : LD-score regression

MAF : Minor Allele Frequency

MR : Mendelian Randomization

MVMR : Multivariable Mendelian Randomization

OMIM : Online Mendelian Inheritance in Man

PGS : Polygenic Score

PheWAS : Phenome-Wide Association Study

RCT : Randomized Controlled Trial

SNV : Single-Nucleotide Variant

TopMED : Trans-omics for precision Medicine

Complex Traits

BMI : Body Mass Index

CRP : C-Reactive Protein

FEV1 : Forced Expiratory Volume in 1 second

FVC : Forced Vital Capacity

HDL : High-Density Lipoprotein

LDL : Low-Density Lipoprotein

SBP : Systolic Blood Pressure

Table of Contents

List of Figures

List of Tables

Introduction

Most phenotypic traits - such as height, blood pressure or lipid levels for example - and common diseases have a genetic basis. This means that they depend on genetic information a person inherits from his or her parents. This idea is not recent, as it goes back to the 19th century and Gregor Mendel's pioneering research on heredity. However, our understanding of human genetics has tremendously progressed in the last decades, with the improvement of biological approaches and technologies easing the access to genetic data, combined with the development of statistical methods to analyze them. These advancements have been instrumental in better understanding and quantifying the genetic contribution to phenotypic traits and are paving the way towards precision medicine.

The aim of this introduction is to provide an overview of the current knowledge and methods in statistical genetics, with a particular focus on approaches that use genetics to explore relationships between complex traits and diseases and methods that take advantage of these relationships.

Human genetics of complex traits

Genetic information is encoded in DNA, a molecule composed of several billions of nucleotides called base pairs. The sequence of these nucleotides can be determined using sequencing methods. The genotype of an individual is its complete set of genetic material. For humans, it corresponds to two alleles at each genetic position, or locus, with one allele inherited from each parent. DNA integrity during transmission in ensured through replication. DNA can be decomposed in functional units called genes (coding regions), which are scattered across the genome and separated by intergenic (non-coding) regions. Each gene corresponds to a single set of instructions, usually coding for a particular protein. DNA is identical in each cell of an individual, but the way the information is interpreted might differ depending on the cell type and the environmental conditions, through epigenetics modifications for example. This can happen during the transcription of a gene into messenger RNA, or when the latter is translated into protein. These proteins are key components of the cellular machinery and can influence the human phenotypes (Figure 1).

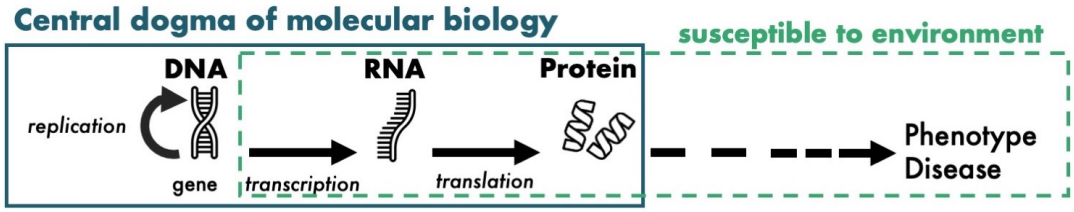


Figure 1: From genotype to phenotype.
Genetic information encoded in DNA influences phenotypes through intermediate processes (central dogma of molecular biology). Environmental factors might act on these processes, or directly affect the phenotype.

The sequencing of the first genome (International Human Genome Sequencing Consortium, 2001, 2004) has been a decisive step in human genetics and this initial draft has been frequently updated, with the very recent publication of the first truly complete sequence of a human genome (Nurk et al., 2021) (Figure 2). It also opened the way for other large-scale, worldwide sequencing projects. For example, the International HapMap Consortium (2003) aimed at identifying common patterns of genetic variation across humans (Figure 2). Genetic variations can be more or less common in the population and this is usually measured using minor allele frequency (MAF). They obtained, analyzed and shared the sequences of 269 individuals, from four different populations (The International HapMap Consortium, 2005). The 1000 Genomes Project Consortium (2010) also aimed at characterizing human genome sequence variations (Figure 2). The project was designed to develop and compare high-throughput sequencing approaches. They used low-coverage whole-genome sequencing, high-coverage whole-genome sequencing and exome (coding regions only) sequencing to study genetic variations across 1,000 individuals from seven populations. More recently, initiatives to aggregate and harmonize sequencing data such as the Haplotype Reference Consortium (2016) (HRC) or the Genome Aggregation Database (Karczewski et al., 2020) (GnomAD) have been proposed (Figure 2).

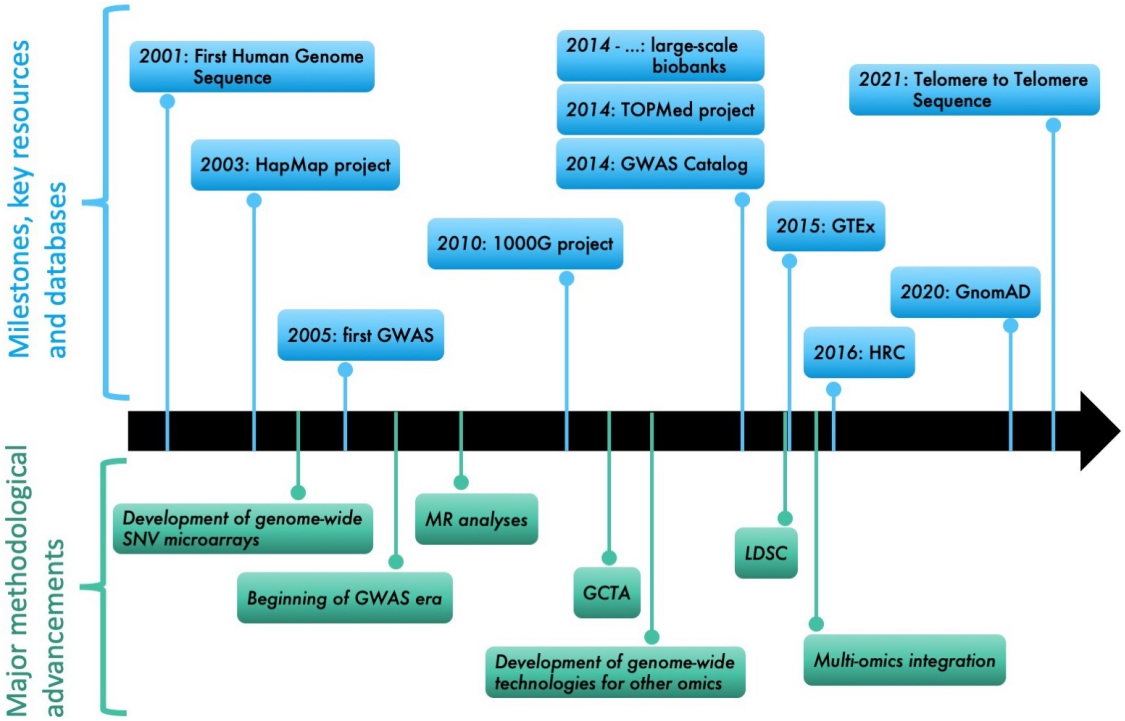


Figure 2: Timeline of the after genome sequencing era.
Milestones achieved, key resources and databases, and major technological and methodological advancements in human genetics from the last 20 years.
Note that these events will be introduced in different sections of the introduction.

These sequencing projects led to a better understanding of human genetic variation. It has been shown that from the 3 billion nucleotide pairs of the human genome, less than 3% (~84.7 million) are

single-nucleotide variants (SNVs, substitution of a single nucleotide) and only 0.3% (~8 million) of these variations are commonly observed (MAF $>$ 5%) (The 1000 Genomes Project Consortium, 2015). In a typical genome, the number of SNVs varies between 3.5 and 4.5 million (most of them being rare and observed only in a very small subset of individuals) and differs greatly between populations.
Large-scale sequencing studies also shed light on the linkage disequilibrium (LD) structure of the human genome. Genetic variations occurring at different loci are not always independent and LD is a term used to describe the correlation structure that exists between them (Pritchard & Przeworski, 2001). LD can sometimes be used to infer information about past events and population history (Tenesa et al., 2007). Since recombination frequency is positively correlated to physical distance, SNVs nearby tend to be inherited together and this means that knowing the genotype at one locus also provides information about the genotype at other loci nearby. This is of high interest when designing micro-arrays for genotyping (determination of the genotype at specific positions only) (Figure 2). It implies that much fewer SNVs actually need to be assayed and that these variants can capture the majority of common genetic variation and be used as tagging SNVs for other unassayed variants.

Complex traits are often defined in contrast to monogenic diseases. The genetics underlying monogenic diseases, also called Mendelian diseases, has been studied using linkage analyses since the 1970's. Monogenic diseases are often rare and genetically determined by a single gene, in which a mutation will lead to clinical symptoms in a high proportion of individuals carrying it (high penetrance). More than 6,000 monogenic diseases are currently referenced in the Online Mendelian Inheritance in Man (OMIM) database (Mckusick, 1998). The relationship between the transmission of a locus and the disease within families can be investigated using pedigrees. Pedigrees, which are genetic representations of a family tree, are used to analyze the pattern of inheritance of alleles. The identification of the genetic variation underlying the disease relies on co-segregation of markers (Pulst, 1999). This type of analyses has led to the identification of genes involved in familial hypercholesterolaemia (Berg & Heiberg, 1978), Huntington's disease (Gusella et al., 1983) and cystic fibrosis (Tsui et al., 1985), for example.
Monogenic diseases can be seen as rare familial forms of complex traits (Peltonen et al., 2006). Complex traits, also sometimes referred to as quantitative traits, are traits that exhibit a high variability and that can be influenced by both environmental and genetic factors. The extent of the genetic contribution to a trait variation is often quantified using heritability. The heritability of a trait corresponds to the proportion of variation of this trait that can be attributed to genetic variation. The broad-sense heritability includes additive and non-additive effects, such as dominant and epistatic (interaction between two or more loci) effects, and can be measured from family or twin studies, taking advantage of observed and expected resemblance between relatives (Visscher, 2008). Such studies have shown that most complex traits have a strong genetic component (Table 1).

Table 1: Broad-sense heritability estimates of four complex traits.

Trait	Heritability Estimate	Method	Reference
Height	0.65	sib-regression	Young et al. (2018)
Body mass index	0.39	sib-regression	Young et al. (2018)
Coronary heart disease	0.38 (females) - 0.57 (males)	twin study	Zdravkovic et al. (2002)
Schizophrenia	$\sim$ 0.80	twin study	Cardno & Gottesman (2000)

Therefore, any phenotypic trait that can be measured at the population level, or any predisposition to a disease that is common, is a complex trait as long as it is heritable. The genetics underlying complex traits involves several genetic variants, in multiple genes, and the effect sizes of these variants are expected to be smaller (Rowe & Tenesa, 2012). As a consequence, identification of the genetic basis of complex traits is usually performed using association studies and large sample sizes are needed to achieve statistical power because of the relatively small effect sizes.

Large-scale sequencing studies have helped to design efficient micro-arrays, making access to genetic data easier and cheaper. This allowed researchers to study genome-wide common genetic variations affecting complex traits, in an hypothesis-free manner, using association studies and large cohorts of unrelated individuals.

The genome-wide association studies era

The genetics underlying complex traits is commonly studied using genome-wide association studies (GWASs) (Hirschhorn & Daly, 2005). The association between each SNV and the phenotype is usually estimated using fixed effects linear regression where the phenotype is the outcome and the genotype (dosage, probability of carrying zero, one, or two copies of the effect allele) is the predictor (Figure 3 **A**-**B**). The effect size estimated by the model corresponds to the effect of carrying one additional copy of the SNV (additive effect). Covariates, such as age, sex or genetic principal components (reflecting population structure, i.e. a systematic difference in allele frequencies between subpopulations) (Price et al., 2006), can be used to increase estimation accuracy and correct for confounders. When the phenotype is not continuous, e.g. diseases, where individuals can be either sick ("case") or healthy ("controls"), case-control designs are common. In such situations, the cohort is enriched in cases to maximize statistical power and fixed effects logistic regression is used.

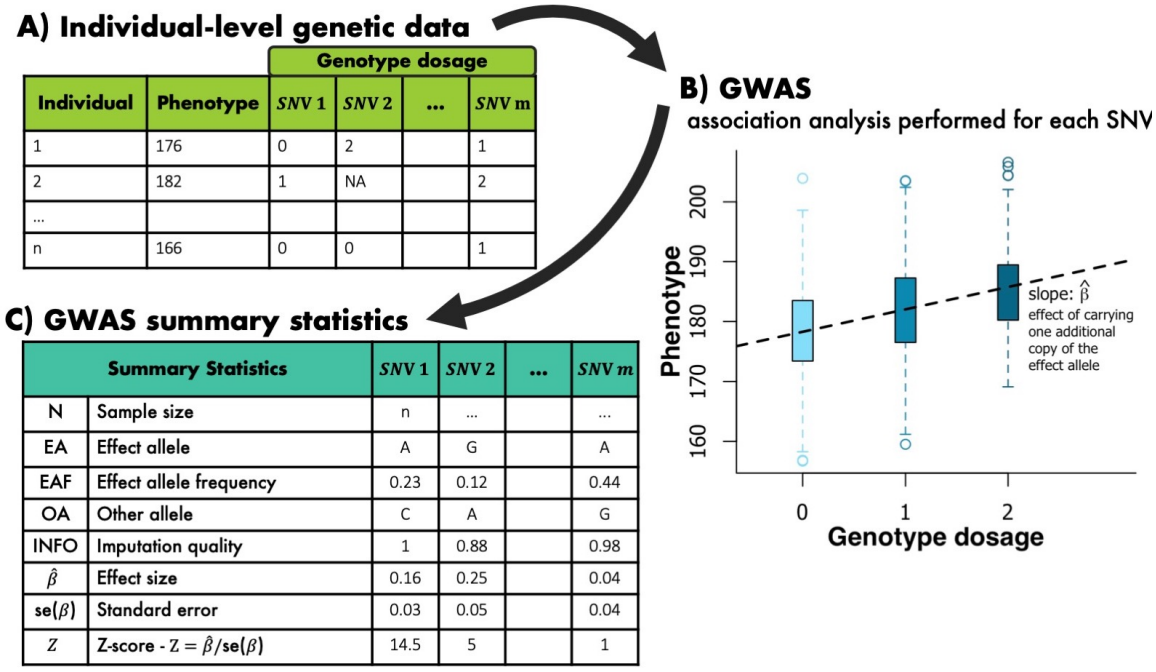


Figure 3: Individual-level data and GWAS summary statistics.
GWASs are perfomed using individual-level data (**A**), regressing phenotypic value on genotypic information (**B**). GWASs results, such as effect sizes, are often shared in the form of summary statistics (**C**).

The first example of GWAS in human genetics is the age-related macular degeneration analysis performed by Klein (2005) (Figure 2). They included 96 cases and 50 controls and screened 116,204 SNVs throughout the genome. Since then, a very large number of analyses has been performed and researchers have aimed at increasing the sample size in order to increase statistical power. Large-scale consortia have been created to combine data from various cohorts across the world, such as the GIANT consortium (2009) that focuses

on anthropometric traits or the Psychiatric Genomics Consortium (2008) that studies the genetic of psychiatric diseases. The last decade has seen the emergence of large population-based biobanks (Figure 2), providing genotypic and extensive phenotypic data, in different countries: the UK Biobank (Sudlow et al., 2015), the BioBank Japan (Nagai et al., 2017), the Estonian Biobank (Leitsalu et al., 2014) or FinnGen, a Finnish biobank (https://www.finngen.fi/en). This allows GWASs to be performed using hundreds of thousands individuals. The number of variants screened has also increased with the development of imputation methods that use LD to infer the genotype at loci that were not assayed (Marchini & Howie, 2010). The GWAS catalog (Buniello et al., 2019; Welter et al., 2014), that is a curated repository for GWAS results, currently contains 255,015 SNVs-phenotype association results from 5,002 publications (*April 20th 2021*) (Figure 2). More recently, UK Biobank association results between hundreds of traits and millions of SNVs have been made available through the GeneATLAS database (Canela-Xandri et al., 2018).

So far, GWAS have mostly permitted the identification of common variants having low to moderate effect sizes and penetrance, which is in line with the Common Disease / Common Variant (CD/CV) hypothesis (Figure 4) (McCarthy et al., 2008). It is hard to associate variants that are rare, because extremely large sample sizes are needed to actually observe these SNVs often enough in the population and perform statistical tests. Similarly, large sample sizes are required to achieve sufficient statistical power to identify variants having very small effect size. There are very few common variants exhibiting large effect sizes, as a consequence of negative selection. Those findings are unusual with the remarkable exception of variants associated with eye color (Sulem et al., 2007), for example.

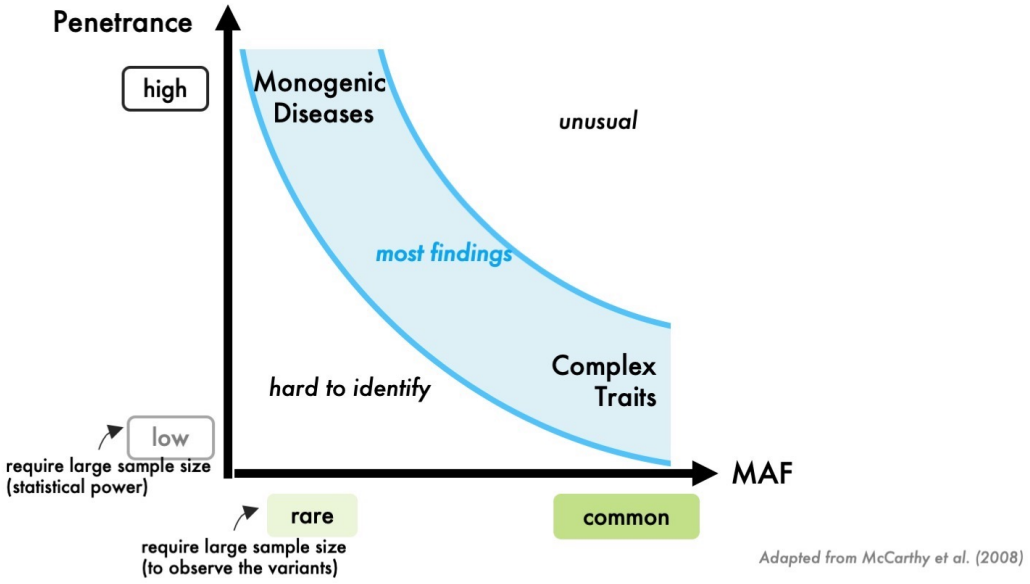


Figure 4: Illustration of the CD/CV hypothesis.
While most associations with monogenic diseases are rare variants of large effect size and penetrance, most GWAS findings for complex traits and common diseases correspond to common variants of relatively small effect sizes.

While fixed effects models (standard linear regression or logistic regression, in case of binary traits) is the most commonly used approach, it is also possible to use mixed effects models. In such frameworks, the association test between each genetic variant and the phenotype is performed by conditioning on a genetic relationship matrix (GRM). The GRM reflects the sample structure and is estimated from genome-wide SNV data (Yang et al., 2014). Because population stratification can be a major confounder in GWAS (Cardon & Palmer, 2003), standard linear regression analyses are restricted to unrelated individuals. Mixed effects models explicitly account for it and related individuals can be included, therefore allowing an increase in sample size. For the UK Biobank for example, analyses restricted to unrelated individuals used 344,397 individuals (Bycroft et al., 2018) whereas 456,422 individuals could be included using a mixed effects model (Jiang et al., 2019), leading to a 25% increase in sample size. However, such increase in sample size does not directly translate into a similar increase in statistical power because some of these individuals are related. Whole-genome regression models that consider all variants simultaneously in a mixed model framework have been very widely used for animal and plant breeding (Campos et al., 2013) and are now also applied to study complex traits genetics in humans (Mbatchou et al., 2020).
To be able to deal with biobank-scale datasets, specific software have been proposed (Table 2).

Table 2: GWAS software for large-scale datasets.

Software	Approach	Outcome	Reference
PLINK 1.9	fixed effects model	continuous & binary	Chang et al. (2015)
BGENIE	fixed effects model	continuous	Bycroft et al. (2018)
GEMMA	mixed effects model	continuous	Zhou & Stephens (2012)
BOLT-LMM	mixed effects model	continuous	Loh et al. (2015)
SAIGE	mixed effects model	binary	Zhou et al. (2018)
fastGWA	mixed effects model	continuous	Jiang et al. (2019)
BayesR and extensions	whole-genome regression	continuous	Moser et al. (2015) Banos et al. (2020) Patxot et al. (2020)
REGENIE	stacked block ridge regression	continuous & binary	Mbatchou et al. (2020)

GWASs have led to numerous discoveries in the past 15 years (Visscher et al., 2012, 2017) and it is likely that, with the amount of genetic and phenotypic data still increasing, many more discoveries are still to come (Loos, 2020). Furthermore, while sharing individual-level genetic data is often complicated because of privacy concerns, GWAS summary statistics (Figure 3 **C**) can be easily shared to perform meta-analyses to increase power, for example. Indeed, summary statistics are often made publicly available, directly by the consortia or through dedicated databases (Buniello et al., 2019; Leslie et al., 2014).
However, several limitations to GWASs have also been highlighted. First, it is important to state that

most GWASs to date have been focusing on populations of European ancestry (Mills & Rahal, 2020). As a consequence, findings are not always directly transferable to other populations because of differences in allele frequency or LD structure, for example (Sirugo et al., 2019). GWASs are often performed using a two-stage design: a discovery study followed up by some replication study in an independent sample. While most GWAS results have been highly replicable (Marigorta et al., 2018), some did fail to replicate in independent cohorts. This might partially be due to winner's curse, a phenomenon leading to a systematic overestimation of effects when thresholding is used (Palmer & Pe'er, 2017). It is also important to note that GWAS results are not immune to bias. As already mentioned, population stratification (if not correctly accounted for), but also assortative mating (a mating pattern in which individuals with similar phenotypes tend to mate with one another more frequently) (Robinson et al., 2017) or dynastic effects (when parental genotypes, even untransmitted parts, directly affect offspring phenotypes) (Kong et al., 2018) can lead to an inflation of the estimated effect sizes. It is also noteworthy to mention that individuals that are included in GWASs might not be representative of the general population. Individuals are often invited to participate and only the ones who are healthy enough or coming from higher socio-economic strata might do so. In addition, people who have serious health conditions might have passed away at younger ages and therefore could not be included in the recruitment process. As a consequence, GWAS results might suffer from what is called selection bias (Munafò et al., 2017). Evidence of sex-differential participation bias, for example, has been shown by Pirastu et al. (2021). One very recent study looked at how such biases could have affected recent findings about COVID-19 risk and severity and why appropriate sampling strategies are needed (Griffith et al., 2020). Finally, the lack of interpretability of the results is often criticized. Indeed, because of LD, GWASs identify genetic regions that are associated with a trait and it is hard to determine the causal variant, which might not even be assayed in the analysis (Altshuler et al., 2008). Moreover, the majority of GWAS hits lie in non-coding regions and do not directly affect the coding sequence of a gene (Edwards et al., 2013), hindering the understanding of the biology behind these association signals. Methods to help with the interpretation of GWAS results will be discussed in the next section.

A long road towards precision medicine

While this has not been the main focus of my work, it is important to mention downstream analyses that can be used to help understand the effects of genetic associations identified by GWASs. A better understanding of the biological mechanisms affecting complex traits and diseases is key to transfer GWAS findings into medicine. This would help in developing precision medicine, an approach that takes into account the genetics and the environment of each patient, to provide personalized prevention and treatment to each individual. GWAS findings can help improve four main areas of healthcare: prevention, diagnosis, therapy and prognosis (McCarthy et al., 2008). The different ways GWAS findings can benefit healthcare are summarized in Figure 5 and will be discussed more in details in this section.

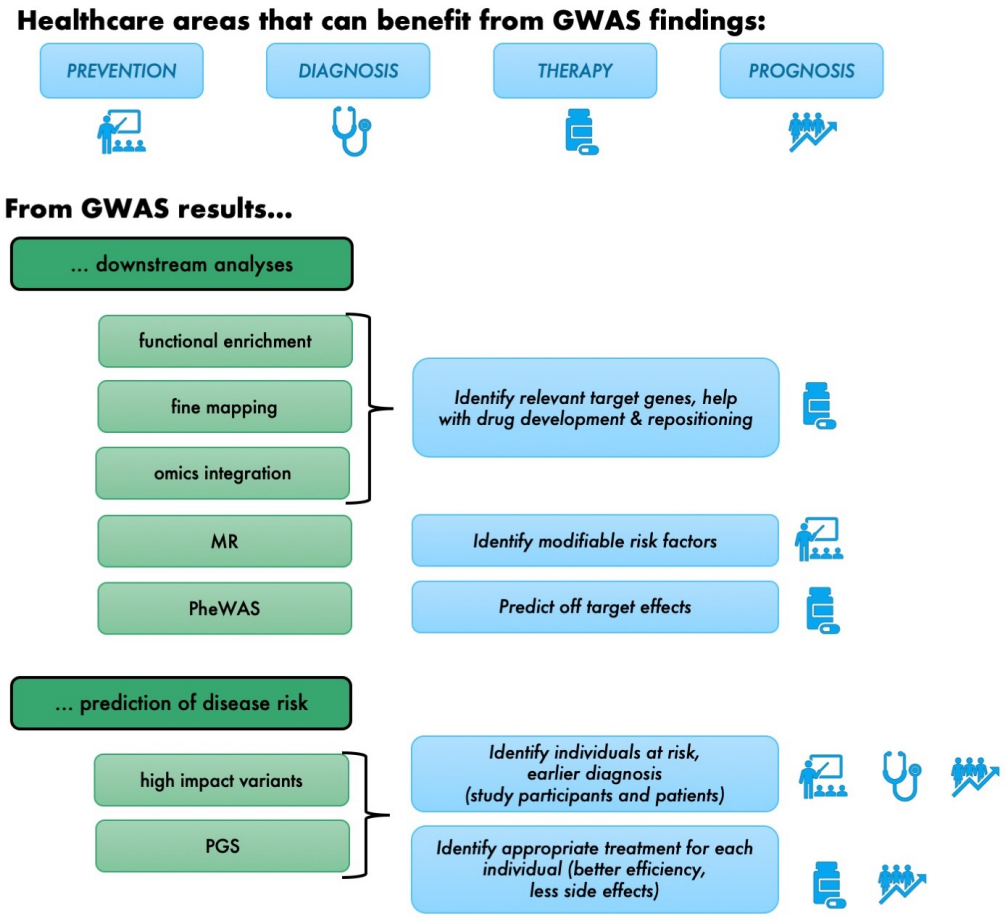


Figure 5: How healthcare can benefit from GWASs.
Four main healthcare areas can directly or undirectly benefit from GWASs findings: prevention, diagnosis, therapy and prognosis. Most of the benefits are coming from downstream analyses, helping with prevention and therapy, and prediction of disease risk, which might be beneficial to all four areas.

Note that two approaches, Mendelian Randomization (MR) and Phenome-wide association studies (PheWASs), will not be discussed in this section, as they are presented more in details in one of the following sections of this introduction. MR can help identify modifiable risk factors by shedding light on causal relationships between complex traits. Ultimately, acting on these risk factors could help reduce the risk of some diseases and improve prevention. PheWAS analyses, however, can provide information about potential off-target effects of drugs (Diogo et al., 2018; Richardson et al., 2020). If genetic variants near one gene affect several traits and diseases, and a drug for one of the disease is acting on this gene, it is likely that it would also affect the other traits, leading to potential side effects. Keeping this in mind can therefore help with drug design and therapy.

Risk prediction and polygenic scores

The most direct way of taking advantage of GWAS findings is to use the association effect sizes for disease risk prediction. Using an individual's genotype, it is possible to predict disease risk very early, even at birth.

For monogenic diseases, where a single gene is involved, prenatal screening can be performed to assess if the embryo is carrying mutations in this specific gene (Riet et al., 1997). For complex traits, it is often more difficult to implement screening strategies. Even when relevant genes can be identified, the number of genes involved is relatively large, each of them having small effect sizes. In such case, it makes more sense to look at the cumulative risk by combining several variant effect sizes into what is called a polygenic score (PGS). However, there are some complex traits for which rare variants have been found to have effects large enough to be of clinical relevance. The most common example is breast cancer. It has been shown that mutations in two genes, BRCA1 and BRCA2, strongly increase the risk of developing breast cancer (Chen & Parmigiani, 2007) and screening strategies have been implemented in several countries (Toland et al., 2018). This remains an unusual finding and PGS-based approaches are better suited for risk assessment in most complex traits. PGSs carry additional information and can come in addition to well-known high impact mutations, as it has been shown for breast cancer (Mars et al., 2020).

The potential utility of PGSs does not rely on an understanding of the mechanisms through which variants affect the phenotype. PGSs only rely on prediction accuracy from GWASs and several softwares have been proposed to build PGSs from GWAS results (Euesden et al., 2015; Vilhjálmsson et al., 2015). Whole-genome regression approaches that provide a genome-wide prediction by considering all variants simultaneously can also be used.
It is important to note that there are several challenges in using PGSs. First, there is no common metric to assess and interpret the quality of PGSs. Recently, a database gathering about 800 PGSs for 200 traits has been created to increase consistency and reproducibility in reporting PGSs (Lambert et al., 2021).

Since PGSs are built based on available GWAS data, often from European ancestry individuals, their predictive power has been shown to be significantly lower when used in other populations (Duncan et al., 2019). More diversity in GWASs is needed to build ancestry-specific PGSs or multi-ethnic PGSs and ensure that the use of PGSs would not increase health disparities (Martin et al., 2021).
For many complex traits, conventional risk factors (such as age, sex, smoking status, etc) are already used by clinicians to assess the disease risk of an individual. Lifestyle is a good predictor for many diseases but it is often noisily estimated and subject to misreporting. Genetic risk can be more accurately measured (even at birth) and PGSs should be used in combination with the more conventional risk factors. Studies looking at coronary artery disease or type two diabetes, for example, showed that prevalence of the disease was much higher among individuals having higher PGSs (Khera et al., 2018; Mahajan et al., 2018). A very recent study, looking at cardiovascular risk prediction showed that a targeted strategy using PGSs could help prevent 7% more cardiovascular disease events than conventional risk prediction alone (Sun et al., 2021), highlighting the high potential of using PGSs in a clinical setting.

PGSs could benefit healthcare in all four areas mentioned above. First, by identifying individuals at risk earlier, it would allow the set up of targeted prevention strategies. Early changes in behavior in these individuals could help improve their prognosis. PGSs can also shed light on different subtypes of some diseases, improving diagnosis and allowing individuals to have access to more appropriate treatments, depending on their genotype. Similarly, genotype data can be used to identify individuals more at risk of having side effects when using a particular drug, offering better therapy options.
It is also important to reflect upon how genomic results should be returned to research participants and patients. Some biobanks started to return results to participants. In FinnGen for example, a subset of participants have received information about their individual risk of atherosclerotic cardiovascular disease (Widén et al., 2020). These participants have been followed up a year and a half after getting back this information to assess how having access to it may have affected their behaviour. This study showed that communicating information about PGSs can motivate positive changes in health behavior in high-risk individuals. While the way disease risk prediction results are communicated to participants in this type of research studies is regulated and supervised, the fact that some direct-to-consumer services also give back health related information might be concerning (Majumder et al., 2021). Indeed, a set of good practices and a legislative framework are needed. Making sure that people who are getting disease risk prediction from genetic data have the knowledge to interpret these results is key to ensure that there is a real benefit for the individuals.

Identifying causal variants, relevant genes and pathways

A lot of efforts have been put into identifying the causal variants underlying GWAS signals, as these variants should be prioritized for further study (Schaid et al., 2018). This process is called fine-mapping

and it requires two types of data: GWAS summary statistics and an LD reference panel to model dependency between SNVs and infer credible sets of causal variants. Fine-mapping analyses have been shown to benefit from trans-ethnic analyses by taking advantage of differences in LD in different populations (Li & Keating, 2014). The two most commonly used softwares are CAVIAR (Hormozdiari et al., 2014) and FINEMAP (Benner et al., 2016).

Another approach to understand how GWAS hits are actually acting on the traits is to identify the genes that might be affected by the genetic variants. This is quite straightforward when a variant is located within a gene but most GWAS hits are lying in non-coding regions, making this task more challenging. Historically, the relevant gene for a specific SNV was assumed to be the closest one. More recently, identification of relevant genes from GWAS association signals has been done by integrating -omics data. These -omics data, such as gene expression data ("transcriptomics") or protein levels data ("proteomics") are measures of intermediate molecular phenotypes. As discussed earlier in the introduction, these phenotypes are very likely to have a key role in how genetic variations are affecting the phenotypes of interest. Recent developments in technologies allowing genome-wide measures of these -omics data have allowed researchers to gather large scale -omics data (Figure 2). For example, the Genotype-Tissue Expression (GTEx) project was set up to study tissue-specific gene expression and regulation, gathering data from 54 different tissues in nearly 1,000 individuals (Lonsdale et al., 2013) (Figure 2). The Trans-Omics for Precision Medicine (TOPMed) program, focusing on improving the understanding of heart, lung, blood, and sleep disorders and advancing precision medicine, collects whole-genome sequencing and various -omics data (TopMed program) (Figure 2). It is possible to perform association studies, similar to GWASs, to identify genetic variants influencing gene expression. Such variants are called eQTLs (expression Quantitative Trait Loci). Several methods have tried to use gene expression data and other -omics data to fill in the gap between genotype and phenotype. Some studies used colocalization approaches to match the eQTL and GWAS signals and estimate the probability that the association signals are due to the same genetic variants (Hormozdiari et al., 2016; Wallace, 2013). Transcriptome-wide association studies (TWAS) have also emerged to integrate GWAS results and gene expression datasets in order to identify gene–trait associations. These methods, like PrediXscan, for example, use eQTL data to predict gene expression levels and then test for the association between genetically predicted expression and disease risk in GWAS (Gamazon et al., 2015). More recently, a method was proposed to extend the TWAS approach to multiple genes (Mancuso et al., 2019). An alternative approach for transcriptome-wide analyses actually tries to go further and investigate the causal relationship between gene expression and phenotypes, instead of only looking at the similarity between the two (Porcu et al., 2019).

Functional enrichment analyses methods can help identify relevant pathways underlying the bi-

ology of complex traits. By using databases that reference biological pathways, genes involved in each pathway, or functional categories (Ashburner et al., 2000; Kanehisa, 2000), it is possible to test whether GWASs results are enriched for some of those. If a large proportion of SNVs associated with a phenotype are located near or within genes that are involved in the same pathway or in a certain type of functional category, it is likely that this pathway or this functional category plays an important role for this phenotype. Several softwares such as DEPICT (Pers et al., 2015), Pascal (Lamparter et al., 2016), stratified LDSC (Finucane et al., 2015) or web-platforms like FUMA (Watanabe et al., 2017) allow the direct use GWAS summary statistics to perform enrichment analyses.

These approaches aiming at identifying the causal variants, relevant genes and pathways can help find therapeutic targets and guide drug design. In some cases, drugs already acting on these genes or pathways might already exist and could be repurposed, overall improving therapy.

Genetic architecture and heritability models

Genetic architecture is a term used to refer to the landscape of genetic contributions to a given phenotype, influencing broad-sense heritability, such as the number of genetic variants affecting a trait (polygenicity), their allele frequency, their effect sizes and their interactions with each other and with the environment (Timpson et al., 2017). GWAS results confirmed the fact that complex traits are highly polygenic and that most genetic variants associated with these phenotypes are common variants of small effect sizes (CD/CV hypothesis) (McCarthy et al., 2008). More recently, it has been suggested that all complex traits and common diseases share a single omnigenic architecture (Boyle et al., 2017). As in the infinitesimal model proposed by Fisher more than a century ago (1918), that assumes that all variants have a (small) non-zero effect on every phenotype, the omnigenic model suggests that a large majority of the genome would be associated with every trait. However, it assumes that these associations are due to the interconnection of gene regulatory networks and intermediate phenotypes. This means that there would only be a few core genes for each trait and many non-core / peripheral genes, with smaller effect sizes, as they affect the trait only indirectly.

Gene-gene and gene-environment interactions

Interactions between variants (gene-gene interactions, GxG) and interactions with the environment (gene-environment interactions, GxE) are reasons why the same genetic variation might have different effects in different individuals. In case of GxG, the effect of the genetic variation might be influenced by the genetic background and the presence of other genetic variations in the same region or elsewhere in the genome. The effect of the genetic variation might be influenced by environmental factors, such as diet, exposure to pathogens, etc. There is little evidence for GxG in human complex traits (Wei et al., 2014). As additive effects identified by GWAS, GxG effects are expected to be small and very large sample size would be required for their identification. Moreover, the number of tests needed to explore genome-wide interactions is extremely large: assuming 5 millions observed SNVs, it would require more than 10,000 billions tests to investigate all pairwise interactions and significant associations could only be identified by further increasing the sample size. While large sample sizes are now becoming available, such analyses are still computationally challenging. On the other hand, lots of advancements have been made to GxE analyses in the past years (McAllister et al., 2017), leading to some meaningful discoveries. For example, the variants near FTO and MCAR genes that are associated with body mass index (BMI) have been shown to interact with lifestyle factors (Bjørnland et al., 2017; Corella et al., 2012). GxE analyses still suffer technical limitations and are often focusing on a restricted number of variants that have already been associated with the phenotype of interest: looking at them individually or combining them into PGSs. Moreover, environmental factors can be difficult to accurately measure and are often very noisy. Most analyses so far looked at the effect of specific environmental factors but it is also possible to look at

a global environmental load, even if the environmental factors are not measured. For example, such an approach showed significant contribution of interactions between PGSs and environmental load on several obesity-related traits (Sulc et al., 2020). Finally, some approaches are not focussing on specific variants or PGS and looked at the contribution of GxE across the genome (Robinson et al., 2017; Ni2019: Kerin & Marchini, 2020).

Heritiability estimation from whole-genome data

Genetic data can also be used to estimate narrow-sense heritability, the proportion of phenotypic variance due to additive genetic effects only. When looking at GWAS top hits, it is possible to use the effect sizes to estimate the proportion of variance explained and the corresponding narrow-sense heritability. Methods to estimate narrow-sense heritability from whole-genome data (also called SNV heritability), using all SNVs and not only top hits, can also be used.

There are several methods to estimate to estimate SNV heritability from individual-level data, with different modeling assumptions for the effect sizes (Zhu & Zhou, 2020).In this section, I will focus on the two more commonly used approaches: GCTA (Yang et al., 2011) (Figure 2) and LDAK (Speed et al., 2017). Both approaches are looking at all SNVs simultaneously to estimate the total contribution of genetics using linear mixed models, treating all SNVs as random effects and using the genetic relationship matrix. Their assumptions about the genetic architecture differ: GCTA assumes that all SNVs contribute equally to heritability (i.e. rare SNVs must have stronger effects to explain the same proportion of variance as common variants) whereas LDAK allows for different relationships between MAF and effect sizes. LDAK is more flexible, as it also allows heritability to vary according to LD level or genotyping quality.

As often in statistical genetics, similar approaches have been developed to work with summary statistics, as access to individual-level data can be complicated. The most popular method to estimate heritability from summary statistics is LD-score regression (LDSC) (Bulik-Sullivan et al., 2015) (Figure 2). The LDSC model is similar to the one assumed by GCTA, and several extensions have been proposed to account for functional annotations (Finucane et al., 2015) and LD level (Gazal et al., 2017), for example. The LDAK model has also been extended to summary statistics analysis (SumHer: Speed & Balding, 2018). The development of such approaches, combined with the availability of summary statistics from large cohorts like the UK Biobank have led to the creation of databases compiling SNV heritability for thousands of complex traits (UK Biobank heritability browser).

Missing heritability

When first comparing heritability estimates from GWAS top hits to the ones from twin or family studies, very large discrepancies were observed, leading to the missing heritability concept (Eichler et al., 2010). For example, the broad-sense heritability of height is estimated to be around 80%. Visscher (2008) found 40 loci associated with height, explaining only 5% of phenotypic variance, corresponding to only 6% of

the broad-sense heritability. Missing heritability can be attributed to different causes:

- inaccuracy from family/twin studies: shared environmental effects can lead to an overestimation of broad-sense estimates;
- SNVs having small effects and that are not significant in GWASs;
- rare variants and other types of variations - insertions and deletions, copy number variations (CNVs), structural variations (SVs) - that are not investigated in GWASs;
- non-additive effects: GxE and GxG effects are not captured by GWAS results.

For height, increasing sample size, and therefore statistical power, as well as including SNVs that are not reaching genome-wide significance level, strongly reduces the gap between twin-based estimates and GWAS-based estimates (Wood et al., 2014). More recently, LDSC-based estimates suggested that SNV heritability for height is close to 50%, confirming that a huge proportion of missing heritability can actually be captured by SNVs having small effect sizes (Yengo et al., 2018).

Regarding other possible explanations for missing heritability, the contribution of rare variants (SNVs only) has been explored recently using whole-genome sequencing data (assessment of all SNVs and not only the ones tagged by the micro-arrays) (Wainschtein et al., 2019). In this paper, broad-sense heritability of height was recovered, but such estimates are likely to be biased by population stratification, and further analyses are needed to confirm these results. It is important to note that even if rare SNVs, or other types of variations, with large effects still need to be discovered, such variants might not explain a very large proportion of variance in the general population because of their low allele frequency. GxE has been shown to explain around 5% of the variance for traits such as BMI and blood pressure (Kerin & Marchini, 2020; Robinson et al., 2017). There is no evidence of GxG contributing to heritability so far, as much larger sample sizes are needed to detect it (Hivert et al., 2021).

Relationships between complex traits

From correlation...

Correlation between complex traits is pervasive and it is important to distinguish observational correlation, at the phenotypic level, from genetic correlation. Observational correlation, the correlation between two traits, is the sum of two things: environmental factors affecting both traits and genetic correlation. While heritability corresponds to the part of the variance of a trait that is due to genetics, genetic correlation corresponds to the part of the covariance between two traits that can be explained by genetics. Genetic correlation implies correlated effects of genetic variants on both traits. Such phenomenon is called pleiotropy, and it corresponds to one gene having an effect on multiple phenotypes (Sivakumaran et al., 2011). Here we will consider three types of pleiotropy:

- uncorrelated pleiotropy (horizontal pleiotropy), when the same variant (or two variants in LD that are inherited together) influences both traits through different mechanisms;
- correlated pleiotropy (horizontal pleiotropy), when the same variant is affecting both traits through the same mechanism, often assumed to be another trait called a genetic confounder;
- causality or mediated pleiotropy (vertical pleiotropy), when one trait is affecting the other trait.

PheWASs can be used to look at the association of a SNV with multiple phenotypes, enabling the identification of highly pleiotropic variants (Pendergrass et al., 2011). The role of horizontal pleiotropy in the genetic architecture of human traits has been studied by Jordan et al. (2019), showing that horizontal pleiotropy is driven by the extremely high polygenicity of complex traits. Note that only correlated pleiotropy (genetic confounding) and mediated pleiotropy (causality) can lead to genetic correlation between phenotypes (Solovieff et al., 2013), as uncorrelated pleiotropy is only occurring at particular loci and not throughout the genome.

LDSC can be used to study the shared genetic architecture and estimate the genetic correlation between complex traits (Bulik-Sullivan et al., 2015). Cross-trait LDSC has been applied to analyze the relationships between 49 complex traits, highlighting the existence of clusters of highly genetically correlated traits. When restricting the analysis to 24 traits (one per cluster), a lot of significant correlations were still observed.

It is possible to take advantage of these correlations to boost power in GWAS analyses. This can be done using multi-trait association testing models. There are three main approaches to do so: multivariate regression models, associations tests combining univariate effects and dimension reduction (Aschard et al., 2014; Stephens, 2013). Pleiotropic effects provide accumulative evidence for association between one variant and all the phenotypes. For example, the multi-trait analysis of GWAS (MTAG) approach (Turley et al., 2018), a multivariate regression model, has been applied to educational attainment and three related cognitive phenotypes (Lee et al., 2018). Results showed that not only did the multi-trait

approach help to discover new loci, but the multivariate effects also led to an increase in the variance explained by GWAS hits. It is also important to note that studying multiple correlated phenotypes together can help identify associations that are specific to one of the traits, with an increased power to detect SNVs with opposite effects on positively correlated traits (Aschard et al., 2014). A recent analysis, using different association tests to look at 36 traits found 322 new associations that were not previously reported by univariate analyses (Julienne et al., 2020). Dimension reduction, using principal components analyses for example, can also be used. By combining phenotypes, it is possible to identify major axes of phenotypic variation, that might correspond to more relevant phenotypes which can not not be directly observed (Ried et al., 2016). In a paper investigating obesity-related traits using 14 phenotypes, 4 major axes of variations were identified, corresponding to i) an overall increase in body size, ii) a decrease in height but an increase in fat mass, iii) a decrease in hip circumference and an in increase in waist circumference, iv) a decrease in height and an increase in BMI with an increase in lean mass (Sulc et al., 2020). Distinct sets of SNVs were found to be associated with each of these axes. Other approaches aim at disentangling the genetics underlying correlated traits by modeling unmeasured factor(s), to identify variants that are associated with the common factor(s). For example, GenomicSEM proposes a framework to model a common liability underlying the traits by using structural equation models and factor analyses (Grotzinger et al., 2019). It allows the identifcation of SNVs acting on the common liability, but also variants that have divergent effects on the different traits. When investigating five psychiatric traits, they found 27 SNVs significantly associated with the common liability, that were not previously identified in the univariate GWASs. Multi-traits approaches can also been used to increase accuracy prediction, as shown by Maier et al. (2018) for a wide range of traits.

In all these approaches, what matters is the existence of genetic correlation and the reasons that gave rise to these correlations are often ignored. However, understanding the underlying phenomenon leading to this correlation and distinguishing correlated pleiotropy effects from causation is important as it can help better understand the relationships between complex traits.

... to causation

Historically, the identification of causal relationships between complex traits has been done using randomized controlled trials (RCTs), aiming at identifying modifiable risk factors that can affect an outcome (Hill, 1952). RCTs are an epidemiological tool that rely on the observation and comparison of two groups: one in which an epidemiological intervention (change of behavior, drug administration, etc) is performed and one control group. People are assigned to one of the two groups randomly, minimizing the effects of potential confounders. However, RCTs are expensive, they can be difficult to set up, suffer compliance issues and some hypotheses are not testable because of practical or ethical reasons (Sanson-Fisher et al., 2007). MR is an alternative method that uses genetic data to explore causation.

It relies on the assumption that genetic information is transmitted from parents to offspring randomly, in a process that can not be influenced by environmental confounders and therefore provides a natural framework to explore causality (Lawlor et al., 2008) (Figure 2). MR uses genetic variants as instrumental variables (IVs), or instruments, to assess the existence and the strength of the causal effect of an exposure on a outcome. Causality between two traits can be represented using a directed acyclic graph (DAG), where G represents the IVs, X represents the exposure, Y represents the outcome, and U represents a genetic confounder (Figure 6).

MR relies on three main assumptions (Figure 6): (1) Relevance – IVs must be robustly associated with the exposure. (2) Exchangeability or Independence – IVs must not be associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – IVs must be independent of the outcome conditional on the exposure and all confounders of the exposure-outcome relationship (i.e. the only path between the IVs and the outcome is via the exposure).

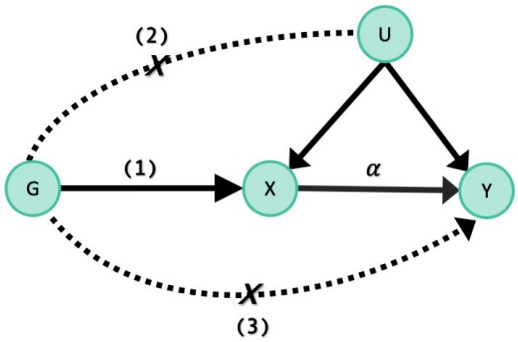


Figure 6: Main assumptions of Mendelian randomization.
(1) Relevance – IVs, denoted by G, are strongly associated with the exposure. (2) Exchangeability – G is not associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – G is independent of the outcome conditional on the exposure and all confounders of the exposure-outcome relationship (i.e. the only path between the IVs and the outcome is via the exposure).

A large number of MR analyses have been performed in the last decade. For example, it has been used to study the causal effect of lipid traits on coronary heart disease (Holmes et al., 2015), confirming epidemiological evidence for a causal relationship between LDL cholesterol and cardiovascular disease risk (Cholesterol Treatment Trialists' Collaborators et al., 2012). MR analyses refuted the causal role of smoking on depression, which could not have been studied using RCTs because of ethical concerns (Bjørngaard et al., 2012). MR has also been used to investigate the causal effect of liability to depression on brain white matter microstructure (Shen et al., 2020), which also would not have been possible using standard epidemiological approaches.

It is important to distinguish two types of MR methods: one-sample MR and two-sample MR,

also referred to as summary-data MR. One sample MR requires genetic data and phenotypic data (both for the exposure and the outcome) from a single cohort. It is usually performed using a two-stage least square regression. In the first stage, IVs are selected and used to obtain a prediction of the exposure and in the second stage, the outcome is regressed on the predicted exposure to obtain the causal effect estimate (Lawlor et al., 2008).
Two-sample MR only requires GWAS summary statistics and therefore has become very popular in the last decade, as those became widely available. The most common two-sample MR approach is the inverse-variance weighted (IVW) method (Burgess et al., 2013). Each IV provides a ratio estimate (combining effect sizes from GWAS on the exposure and on the outcome) and these ratio estimates can be combined using an IVW approach. When performing two-sample MR, GWAS summary statistics for both the exposure and the outcome are needed. These two sets of summary statistics should come from similar populations (i.e. same ancestry, age/sex distribution, etc) to ensure that the underlying causal effect between the exposure and the outcome is the same in both datasets (Lawlor, 2016). Additionally, it is expected to use two independent cohorts, with no sample overlap between the exposure and the outcome cohorts, to ensure that the direction of some of the potential biases, discussed afterwards, is towards the null (Burgess et al., 2016).

MR is a very active field: major biases that may affect MR estimates have been quite intensively studied and several approaches to tackle some of them have been proposed. Two main categories of biases can be defined, depending on whether they are inherent to MR or if they are coming from the estimation of the IVs effects on the exposure.

Here, I will describe four main sources of bias coming from the MR methodology and assumptions. The first one is called weak instrument bias. Most MR methods assume that SNV-exposure effects are measured without noise and this simplification leads to regression-dilution bias, which is particularly strong for weak instruments. For this reason, the bias introduced by this assumption is referred to as weak instrument bias and it becomes more and more severe as the average variance of the exposure explained by the IVs decreases (Burgess et al., 2011; Burgess & Thompson, 2011). To reduce the impact of weak instrument bias, the strength of the association between the IVs and the exposure can be assessed using the F-statistic. Rule of thumb is to have an F-statistic larger than 10. In one-sample MR, the direction of weak-instruments bias is towards the observational correlation, that includes correlation that may be induced by environmental confounders, whereas in two-sample MR, when non-overlapping samples are used, the direction of the bias is towards the null. This bias can be particularly strong when the sample size of the exposure GWAS is relatively small.
Secondly, pleiotropy can also be one major source of bias in MR. Under the INSIDE assumption (INstrument Strength Independent of Direct Effect), the pleiotropic effects (effects of the IVs on the

exposure and direct effects of the IVs on the outcome) are assumed to be independent and direct effects are assumed to be on average zero (Bowden et al., 2015), therefore IVW-based methods will produce consistent estimates. MR-Egger (Bowden et al., 2016) allows direct effects to have a non-zero offset. Correlated pleiotropy however, can lead to a violation of the exchangeability assumption if some of the IVs are in fact associated with a genetic confounder and not directly with the exposure. As the violation of this assumption is not directly testable, a lot of methods aiming at reducing or correcting for the impact of this source of bias have been described. Extensions of IVW-MR have been proposed, to make this assumption less stringent: median-based estimators for example, assume that more that 50% of the IVs are valid (Bowden et al., 2016), while the mode-based estimators assume that zero-pleiotropy instruments are the most frequent (Hartwig et al., 2017). More sophisticated methods, such as MR-RAPS, that models pleiotropy using a random effects model (Zhao et al., 2020), or CAUSE, that explicitly takes into account the existence of a confounding factor (Morrison et al., 2020), have been developed to tackle the biases induced by pleiotropy in MR.

The third potential source of bias is reverse causality. While identifying instrumental variables that are strongly associated with the exposure (relevance assumption) is relatively easy - IVs are selected based on the observed association between SNVs and the exposure - making sure that they are not acting on the outcome through other mechanisms (exclusion restriction assumption) can sometimes be more difficult, in particular in the existence of reverse causality (Lawlor et al., 2008). If the outcome is also causally affecting the exposure, variants that are primarily affecting the outcome might appear strongly associated with the exposure too, and could end up being used as IVs while being invalid. One way to detect such invalid IVs is to compare their associations with the exposure and with the outcome and to only include the ones that are more strongly associated with the exposure.

Finally, MR assumes that the causal effect is the same for everyone in the population (Sheehan & Didelez, 2019) and this might not be the case for different reasons: causal effect heterogeneity, non-linear effects or age-specific effects. Causal effect heterogeneity can occur when there are several distinct mechanisms by which the exposure acts on the outcome. This can be explicitly taken into account using the MR-clust approach (Foley et al., 2020), for example. This method has been used to investigate the causal effect of blood pressure on coronary artery disease risk, detecting four different clusters, corresponding to different mechanisms. In the case of non linear effects, the causal effect of the exposure on the outcome might vary non-linearly depending on the exposure level. Some one-sample MR methods, requiring individual-level data, have been proposed and showed the non-linear causal effect of BMI on all cause mortality risk, for example (Sun et al., 2019). Standard MR approaches suggest a positive effect, with higher BMI increasing mortality risk, while non-linear MR results show that in reality, extremes values of BMI, very low or very high, both increase mortality risk. Regarding time-specific effects, a recent study tried to dissect the effects of childhood and adulthood BMI on several diseases (Richardson et al., 2020). They showed for example that larger BMI in childhood was associated with lower breast cancer risk, while there was no evidence

of such protective effect later in life. In all these scenarios, it is not possible to define a direction for the bias, since what MR estimates is an overall effect, combining the different underlying true causal effects.

The other sources of bias are tightly linked to the effect sizes estimation obtained from the association testing between genetic variants and the exposure, in the GWAS or in the first stage of the one-sample MR approach.
IVs are often selected using a threshold, to ensure that only SNVs that are strongly associated with the exposure are used, to satisfy the relevance assumption. In such case, the effect size estimates for the IVs are biased because of winner's curse (Palmer & Pe'er, 2017; Zhong & Prentice, 2008). Winner's curse occurs when the same sample is used to select the IVs and estimate their effect on the exposure. In such case, the observed effect of an IV on the exposure is not an unbiased estimator for its true effect and is likely to be overestimated (in absolute value). In one-sample MR, winner's curse biases the causal effect estimates towards the observational correlation, whereas in two-sample MR, using non-overlapping samples, the direction of the bias is towards the null. One possibility to avoid winner's curse is to use two samples for the exposure: one to select the IVs and one to estimate their effect sizes, but this might not always be possible and this is not often done in practice.
As mentioned in a previous section, GWAS effects size estimates can be affected by selection bias, but also by population stratification, assortative mating or dynastic effects. As a consequence, these biases will be propagated to the causal effect estimates. Interestingly, it has recently been shown that in one-sample MR analyses, it can be beneficial to artificially induce and then correct for selection bias, as this allows the use of two-sample MR methods to one-sample individual level data (Barry et al., 2021). This allows the use of more advanced two-sample MR methods to deal with pleiotropy for example, and provides a framework in which the direction of weak instruments bias is towards the null. The effects of population stratification, assortative mating and dynastic effects on MR causal effect estimates have also been investigated and one-sample MR methods relying on family data (mother–father–offspring trios or sibling pairs), that provide robust causal effect estimates, have been gaining in popularity recently (Brumpton et al., 2020; Davies et al., 2019; Hartwig et al., 2018).

Finally, it is important to keep in mind that the way the different phenotypes are defined and measured can strongly change the way MR results should be interpreted. This is for example highlighted by the coarsening bias, occurring when the exposure is a categorical trait representing a potentially continuous liability. SNVs might affect the continuous latent trait without changing the status when looking at the categorical approximation and IVs might affect the outcome through their effects on the liability and not directly through an effect on the categorical trait. In such cases, one might be interested in identifying the causal effect of the latent exposure as it might be more relevant and easier to interpret. Tudball et al. (2021) presented an approach describing how to do it and clarifying under which conditions

this might be feasible. Another example is coming from what is sometimes referred to as collider bias (Aschard et al., 2015). It comes from the fact that GWASs often use heritable confounders as covariates. Adjusting for a heritable confounder can bias the effect size estimates of the SNVs, potentially leading to spurious associations signals for variants associated with the confounder (Day et al., 2016). These effect sizes would later on be used in MR for the IVs, while their biological relevance might be questioned.

One particularly interesting extension of MR analyses is multivariable MR (MVMR). It allows the simultaneous estimation of the causal effects of several exposures on a single outcome (Burgess et al., 2014; Burgess & Thompson, 2015). MVMR can actually be seen as a way to take advantage of pleiotropic genetic variants to estimate causal effects. Its assumptions are similar to the ones of standard MR analyses. IVs included in the analysis should be strongly associated with at least one of the exposures. MVMR also assumes that all the exposures are acting independently and that there is no causal relationship between one exposure and another. Violation of this assumption would not totally invalidate the approach, but it would change the interpretation of the causal effects of the two exposures on the outcome. However, the existence of a causal relationship between two of the exposures could correspond to a scenario where one of the exposures, exposure 2, acts both on the other exposure, exposure 1, and on the outcome. In such a case, univariate estimates of the causal effect of exposure 1 on the outcome would be biased, while MVMR provides a framework where genetic confounders can be accounted for (Sanderson et al., 2019). MVMR has been heavily applied to try to disentangle the effects of different lipid traits on coronary disease risks for example, but it can also be used for mediation analyses to estimate the part of the effect of the exposure on the outcome that is actually mediated by another trait (Carter et al., 2021; Sanderson, 2021). For example, Burgess et al. (2017) used such approach to show that the causal effect of age at menarche on breast cancer risk is partially mediated through BMI. The effect BMI on breast cancer risk (childhood vs adulthood BMI) has also been investigated using MVMR (Richardson et al., 2020).

MR has been proven to be a robust and powerful approach to assess causality, when all assumptions are met, and the MR-base platform (Hemani et al., 2018) aims at providing a unified framework to perfom MR analyses. MR can be used to try to partition genetic correlation and distinguish correlated pleiotropy from causality. However, to date, the extent to which causality is driving genetic correlation between complex traits has not been really investigated. MR allows not only the exploration of the causal relationships between complex traits but it can also be used to help fill the gap between genotype and phenotype when applied to intermediate molecular traits, as already mentioned in one of the previous sections.

Statistical genetics is a rapidly evolving field and the amount of data available to study the genetics of human complex traits has grown exponentially in the last two decades. Efficient statistical models and methods are needed to analyze and better understand this data as well as the biological mechanisms underlying the genetics of these phenotypes and diseases. In particular, approaches that use GWAS summary statistics, as those are more easily shareable, are likely to play a decisive role. During my thesis, I focused on method development, proposing MR-based approaches that use such summary statistics. In the first chapter of this dissertation, I will present a Bayesian GWAS approach, that relies on multivariable MR to boost GWAS power. In the second chapter, I will introduce two methods, MRlap and LHC-MR, that aim at tackling some of the most common biases in MR analyses.

Chapter 1
Bayesian GWAS

This method has first been developed by a previous member of the group (McDaid et al., 2017). I took over this project when I started and the first part of my PhD was dedicated to this Bayesian GWAS approach. This chapter lists manuscripts that were part of my efforts in developing and applying this method. For each manuscript I give a short summary and indicate my contribution.

The main aim of this Bayesian GWAS approach is to increase power by leveraging information from risk factors and by comparing the observed association statistics from the focal trait to prior effects. These prior effects are estimated using multivariable MR. Only risk factors having a significant causal effect on the focal trait, identified using a stepwise selection approach, are used to calculate the prior effects. Causal effects are estimated masking the focal chromosome to ensure independence and the prior effects are estimated as described in Figure 7.

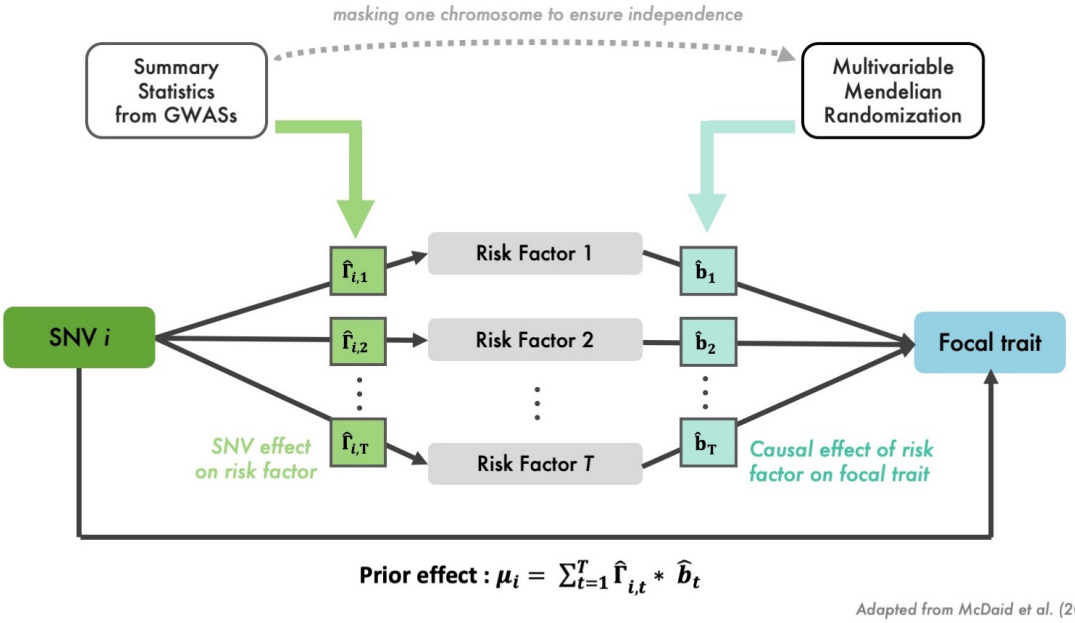


Figure 7: Prior estimation design.
For each SNV i, its prior effect on the focal trait is calculated as the product of the effect of SNV i on the risk factor ($\hat{\Gamma}_{i,t}$) and the causal effect of risk factor t on the focal trait ($\hat{b}_t$, estimated using multivariable MR) summed over all T risk factors identified using a step-wise selection approach.

Observed and prior effects are compared using Bayes factors. Significance is assessed by calculating the probability of observing a value larger than the observed Bayes factors (p-value) given the prior distribution. This approach allows increasing statistical power to identify SNVs that have consistent prior and observed effects.

Analysis of 1 million parental lifespans

The article *Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances* (Timmers et al. (2019) - see Appendix A) has been published together with collaborators from the University of Edinburgh. The aim was to study the biology underlying human lifespan, using parental lifespans. It directly builds up on the work published by the same team a few years before (Joshi et al., 2017) and on the first Bayesian GWAS publication (McDaid et al., 2017). This new analysis takes advantage of the second release of the UK Biobank to increase the sample size, including a total of 1 million parental lifespans, and therefore statistical power. This GWAS of parental survival identified new loci associated with lifespan. The Bayesian GWAS analysis identified 16 risk factors having a causal effect on lifespan and allowed the identification of 6 additional loci. Several downstream analyses have also been performed to help understand the underlying mechanisms, such as pathways and cell-types enrichment studies. Finally, the association results have been used to derive whole-genome PGS for survival, showing a mean lifespan difference of about five years between the top and bottom deciles. For this manuscript, I carried out the Bayesian GWAS analysis and the interpretation of its results. I have also performed an eQTL enrichment analysis. I wrote the corresponding sections of the manuscript.

The `bGWAS` R-package

After performing the lifespan analysis mentioned above, we realized that some parts of the approach could be improved and that new features could be added. We also thought that implementing this approach in a user-friendly R-package would allow people to use it more easily and benefit the community. The improvements that have been presented in the manuscript *bGWAS: an R package to perform Bayesian genome wide association studies* (Mounier & Kutalik (2020) - see Appendix B) are:

- analytic derivation for p-values, leading to a decrease in runtime and a gain in accuracy for very low p-values;
- derivation of posterior effects, combining prior and observed effects, that can be used for downstream analyses;
- derivation of direct effects that correspond to the part of the SNV effects that are not mediated through the risk factors.

The `bGWAS` R-package is now available on GitHub and we tried to make its usage as flexible as possible. Different types of input data for the focal trait are allowed, most of the parameters used for the analysis can be changed and there is a possibility to use a different set of risk factors.
I am the main author of this article. I worked on the statistical derivations needed for the improvements and the new features, I implemented the R-package and I wrote the manuscript.

Chapter 2
Two-Sample MR extensions

Two-sample MR is a powerful approach to assess causality but it relies on several assumptions and it can be subject to different types of bias. In this section, I would like to introduce two manuscripts I have been working on, both proposing extensions to the standard two-sample MR framework and aiming at tackling some sources of bias. The first manuscript presents an approach that allows the use of potentially overlapping samples, while accounting for weak instruments bias and winner's curse. The second manuscript tackles possible violations of the second MR assumption (exchangeability assumption) by accounting for the existence of a potential genetic confounder(s).

Sample overlap, winner's curse and weak instrument bias

The first manuscript, *Correction for sample overlap, winner's curse and weak instrument bias in two-sample Mendelian Randomization* (Mounier & Kutalik (2021) - see Appendix C), presents a direct extension of the standard IVW-MR approach, allowing for sample overlap between the exposure and the outcome GWASs. We wanted to propose a two-sample MR approach that is not restricted to non-overlapping samples, as it can sometimes be difficult to find compatible GWAS summary statistics. Moreover, the necessity of using non-overlapping samples prevents researchers from fully taking advantage of large biobanks since the same data can not be used for both the exposure and the outcome.

The two main sources of bias that are dependent on sample overlap are winner's curse and weak instrument bias. When using overlapping samples, the direction of the bias induced is towards the observational correlation, potentially leading to false positive effect estimates. When using non-overlapping samples, the direction of the bias is towards the null, which is conservative. This explains why non-overlapping samples should be preferred when these biases are not accounted for.

In this article, we analytically derived the expectation of the IVW-based effect estimate. This expectation depends on the following parameters:

- the true underlying causal effect;
- the sample size of the exposure GWAS;
- the threshold used to select the instruments;
- the genetic architecture of the exposure (polygenicity and per-variant heritability);
- the cross-trait LDSC intercept (which depends on the sample size of the exposure / outcome GWASs, the degree of sample overlap between the two and the observational correlation between the traits).

All the parameters, except the true underlying causal effect, can be estimated using GWAS summary statistics. The exact degree of sample overlap between the two sets of GWAS summary statistics does not need to be known, since it is estimated using the cross-trait LDSC intercept. From this expectation, we were able to derive a corrected causal effect estimate that corresponds to a debiased version of the IVW-based estimator. Using a wide range of simulation scenarios, we showed that IVW-based estimates can be biased both when using non-overlapping and fully overlapping samples. The corrected effects that we derived were significantly less biased and provided consistent causal effect estimates for all the degrees of overlap that we used. To assess how this approach would perform when applied to real data, we used the UK Biobank. The large sample size of the UK Biobank allowed us to sample subsets of individuals while varying the overlap degree between the exposure and the outcome datasets. We applied this sampling strategy to several trait pairs and compared IVW-based and corrected effects for different degrees of overlap. We found that when looking at the effect of BMI on systolic blood pressure, for example, there was a significant difference between IVW-based and corrected effect estimates when using non-overlapping samples (underestimation with IVW) but no significant difference when using fully overlapping samples. The correction we proposed has been implemented in an R-package (`MRlap`). We believe that this approach can potentially increase statistical power, by permitting the use of GWAS summary statistics of larger sample sizes while accounting for major biases, and by correcting downward bias in the estimation when non-overlapping samples are used.
I am the main author of this article. I worked on the statistical derivations, performed all simulations and analyses. I implemented the R-package and I wrote the manuscript.

Latent Heritable Confounder MR

The second article, *Simultaneous estimation of heritable confounding and bi-directional causal effects from GWAS summary statistics* (Darrous et al. (2020) - see Appendix D for an updated version of the manuscript), is a joint work with another PhD student from the group. The approach presented in this manuscript, called Latent Heritable Confounder MR (LHC-MR), aims at simultaneously estimating bi-directional causal effects and genetic confounding. This approach uses genome-wide GWAS summary statistics and structural equation modelling to describe the relationship between two observed traits (X and Y) and an unobserved genetic confounding factor (U) (Figure 8).
In addition to estimating bi-directional causal effects and confounding effects, the approach also estimates the genetic architecture of each trait (polygenicity and heritability). Potential sample overlap between the two sets of summary statistics is accounted for. Using genome-wide SNVs rather than strong instruments allows LHC-MR to use more information while avoiding winner's curse. Correlation between SNVs is taken into account by incorporating LD information into the model.
We compared LHC-MR to existing MR methods in a wide range of simulation scenarios. We showed

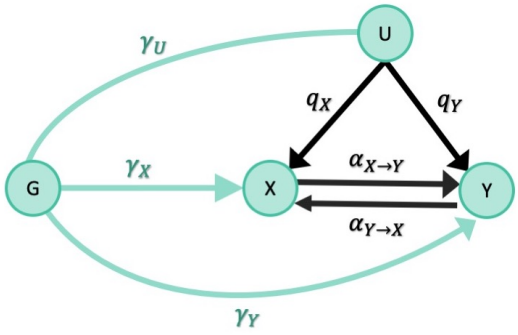


Figure 8: The Latent Heritable Confounder model.
X and Y are two observed traits. U is a genetic confounder of the X-Y relationship. G represents genome-wide genetic variants and γ_X, γ_Y, γ_U correspond to their effects on X, Y and U repectively. q_X and q_Y correspond to the effects of U on X and Y. Bi-directional causal effects, $\alpha_{X\to Y}$ and $\alpha_{Y\to X}$, are included in the model.

that LHC-MR outperforms several of them in most of the scenarios that we tested, including scenarios that violated its assumptions. We applied LHC-MR to 13 complex traits, using summary statistics from the UK Biobank and other large consortia. When comparing LHC-MR results to the ones from other MR methods, we observed that the majority of the causal effect estimates were consistent and LHC-MR identified more significant causal relationships. This can be explained by the fact that LHC-MR takes advantage of genome-wide information. We also identified a significant confounding effect for several trait pairs, one example being systolic blood pressure and HDL cholesterol level. In this case, when accounting for the confounder, we found a significant negative causal effect of HDL cholesterol level on systolic blood pressure ($\alpha = -0.13, p = 5.4 \cdot 10^{-5}$). This effect was missed by all standard MR approaches but seems to be in line with previous observational studies (Laaksonen et al., 2008; Sesso, 2005). In addition, LHC-MR provides a framework that naturally decomposes genetic correlation, allowing it to estimate causal effect-driven and confounder-driven contributions. The total genetic correlation estimates derived from LHC-MR were highly consistent with the ones obtained using LDSC and most of them seem to be driven by bi-directional causal effects.

I have been involved in this project since the beginning, I participated in the development of the approach, in the interpretation of the results and in writing the manuscript.

Discussion

In the introduction, I gave a broad overview of how human genetics progressed in the last two decades, focusing on statistical methods that have been proposed. Since the beginning of the GWAS era, a large amount of summary statistics have been made available. This a valuable resource and methods exploiting it are key to help understand the mechanisms underlying the biology of complex traits as well as the relationships between them.

Most of my thesis work consisted in developing MR-based methods that rely on summary statistics. The 4 articles that I presented here correspond to 3 different methods: a Bayesian GWAS approach, an IVW-MR extension accounting for sample overlap and the LHC-MR model. In the discussion, I would like to present how these 3 methods could be further expanded and combined with each other.

MRlap: current considerations and future work

The MRlap approach is still under active development and we would like to further investigate and improve several aspects of the method. Currently, performing an analysis using GWAS summary statistics for about 1 million SNPs and using the default parameters takes less than 5 minutes. In our simulations, the runtime does not significantly increase when increasing the number of instruments, using a less stringent threshold, for example. The most time-consuming steps are the LDSC fit, that depends on the number of SNPs in common between the GWAS summary statistics and the LD-score reference panel, and the standard error estimation of the corrected effect. Runtime for the standard error estimation mostly depends on the number of simulations (s) used in the sampling strategy. This parameter currently needs to be chosen before running the analysis. In order to improve estimation accuracy while optimising runtime, we are currently trying to automatically assess the optimal value for s in the pipeline.
It is important to mention that while our approach is very effective at reducing bias, the corrected effects obtained in our simulations are not unbiased. There are several hypotheses that could explain it and that we would like to explore. First, this could come from the fact that for each locus, top SNVs are used (lowest p-value) as instruments. However, those top SNVs might not necessarily be the causal ones and lower p-values could also correspond to SNVs that are highly correlated with other variants (*i.e.* large LD scores). This hypothesis could be tested by using another strategy to select instruments. For example, we could try to randomly select instruments for each locus instead of selecting the ones with the lowest p-values. The second reason that could explain why we can not completely remove the bias from the IVW estimates is related to the estimation of $\widehat{\pi}_x$. We noticed in our simulations that there can be some discrepancies between the $\widehat{\pi}_x$ estimates and the true value of π_x that we used to simulate effect sizes.

Our approach assumes that all causal variants are independent, which might not be true as there might exist several independent signals in a single region. This could explain why we are often underestimating the polygenicity and we would like to test if this also explains the bias that we observe for the corrected effects.

Sample overlap in Bayesian GWAS

The interest of using a Bayesian approach to increase power in GWASs has been illustrated with its applications to lifespan analyses (McDaid et al., 2017; Timmers et al., 2019). This could be further investigated by using subsets of different sample sizes from the UK Biobank. We noticed that Bayesian GWAS hits are often moderately associated with the focal trait in the conventional GWAS and we believe that the Bayesian approach allows a prioritization of SNVs that would be reaching significance if larger sample sizes were available. It would be interesting to assess if the Bayesian GWAS hits identified using only a subset of the UK Biobank are more likely to be validated when using the full sample than other variants moderately associated with the focal trait (enrichment analysis). To do this, we would like to focus on traits that are potentially affected by several risk factors, such as BMI or type 2 diabetes.

When implementing our approach into an R-package (Mounier & Kutalik, 2020), we shared the set of potential risk factors that we used for previous analyses. It contains summary statistics from 38 GWASs. This set of potential risk factors has been selected considering that in most cases the focal trait might be coming from the UK Biobank. Since our method uses a two-sample MR approach, we chose risk factors' GWASs coming from independent cohorts, avoiding sample overlap between the focal trait, the outcome of the MR analysis, and the risk factors used as exposures. However, most GWASs that are published are either results coming from UK Biobank analyses or meta-analyses that include UK Biobank data. As a consequence, it is almost impossible to find recent GWASs to add to our current set of risk factors and our approach can not take advantage of the most recent large-scale GWASs. This is what first motivated us to investigate the effect of sample overlap in two-sample MR.
While our work (Mounier & Kutalik, 2021) allowed us to better characterize the effect of increasing sample overlap on winner's curse and weak instrument bias, the correction that we were able to derive only works for univariable MR and can not be easily extended to the multivariable MR method that we use in our Bayesian GWAS approach. We are currently considering two different strategies that could be used to include more GWASs to our set of risk factors.

- **Correcting multivariable effect estimates**

The first strategy is to use the corrected causal effects obtained using univariable MR to correct multivariable effect estimates. However, part of the univariable observed effects can be mediated through other risk factors which causes univariable and multivariable effect estimates to sometimes differ. Therefore the corrected causal effect estimates from the univariable analyses can not directly be used. To overcome this problem, we are considering an approach similar to the approximate conditional analysis presented by Yang et al. (2012). By combining univariable causal effect estimates and the similarity matrix between risk factors, it is possible to approximate the multivariable causal effect estimates (Equation 1). In our case, we will first identify relevant risk factors using our current stepwise approach. Then, by replacing

the IVW-based effects from the univariable analysis by their corresponding corrected effects, it would allow us to obtain corrected multivariable effect estimates (Equation 2).

$$\hat{\boldsymbol{b}} = \left(\frac{\boldsymbol{X}\boldsymbol{X}'}{m}\right)^{-1} \cdot \hat{\boldsymbol{\alpha}} \quad (1)$$

$$\hat{\boldsymbol{b}}_{corrected} = \left(\frac{\boldsymbol{X}\boldsymbol{X}'}{m}\right)^{-1} \cdot \hat{\boldsymbol{\alpha}}_{corrected} \quad (2)$$

$\hat{\boldsymbol{b}}$: multivariable causal effect estimates,

$\hat{\boldsymbol{\alpha}}$: univariable causal effect estimates (IVW-based),

$\boldsymbol{X}$: t x m matrix containing the effects of each IV on each risk factor,

t being the number of risk factors, m being the number of IVs,

$\left(\frac{\boldsymbol{X}\boldsymbol{X}'}{m}\right)^{-1}$ is the similarity matrix between risk factors,

$\hat{\boldsymbol{b}}_{corrected}$: *corrected* multivariable causal effect estimates,

$\hat{\boldsymbol{\alpha}}_{corrected}$: *corrected* univariable causal effect estimates.

Further simulation analyses are needed to explore how this strategy would perform. It would also be interesting to use the UK Biobank data and a sampling strategy similar to the one used in Mounier & Kutalik (2021) to create samples of varying degrees of overlap for the different risk factors and the outcome, and test this strategy on real data.

- **Ignoring sample overlap**

Alternatively, some of our more recent results suggest that sample overlap could actually be ignored in our Bayesian GWAS approach. Indeed, the extent of the bias depends on the sample size of the exposure GWAS and using biobank-scale data would strongly reduce it (Equation 3).

$$\begin{aligned} E[\hat{\alpha}_{IVW}] &\approx \alpha \cdot \frac{\left(\pi_x \cdot \sigma_x^2\right) \cdot \left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\right)}{d} \\ &+ \lambda' \cdot \frac{\pi_x \cdot \left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\right) + \left(1 - \pi_x\right) \cdot 2 \cdot c}{d} \end{aligned} \quad (3)$$

with α being the true causal effect,

n_A being the exposure GWASs sample sizes,

π_x being the polygenicity of the exposure,

σ_x^2 being the per-variant heritability of the exposure,

λ' being a quantity closely related to LDSC cross-trait intercept,

a, b, c and d being constants that do not depend on α nor λ'.

We used UK Biobank summary statistics made available by the Neale lab (GWAS Results round 2) for 17 traits to investigate the causal relationship for 272 trait pairs. The exact degree of sample overlap

between the GWAS summary statistics is unknown (it depends on the amount of missing data for each trait) but is expected to be relatively high. We applied IVW-MR to each pair and also estimated the corrected effects. Overall, the agreement between IVW-based and corrected effect estimates is good and we found a significant difference between the two for about 16% (44/272) of the pairs (Figure 9). This difference was often relatively small and we observed that in the majority of the cases (25/44), the IVW-based estimates were smaller than the corrected effect estimates.

These results suggest than even when the degree of sample overlap is high, the bias from winner's curse and weak instrument bias can still go towards the null. This can be explained by looking at the formula for the expectation of the IVW estimator (Equation 3). The first term of this formula only depends on the true causal effect (α) and it corresponds to a bias towards the null. This bias is more severe when the exposure sample size (n_A) is small or when the trait is highly polygenic or less heritable (larger π_x and/or smaller σ_x^2). This first term will be balanced by the second term that will get larger as sample overlap increases. This second term is largely influenced by a quantity closely related to the LDSC cross-trait intercept ($\lambda' = (\alpha + \rho) \cdot \frac{n_{A \cap B}}{n_A \cdot n_B}$) and that directly depends on sample overlap. However, this value will reach a maximum when the degree of sample overlap is 1, i.e. $n_A = n_B = n_{A \cap B}$, which is unlikely to happen in most cases. It explains why we still often observe a bias towards the null even when using the full UK Biobank data.

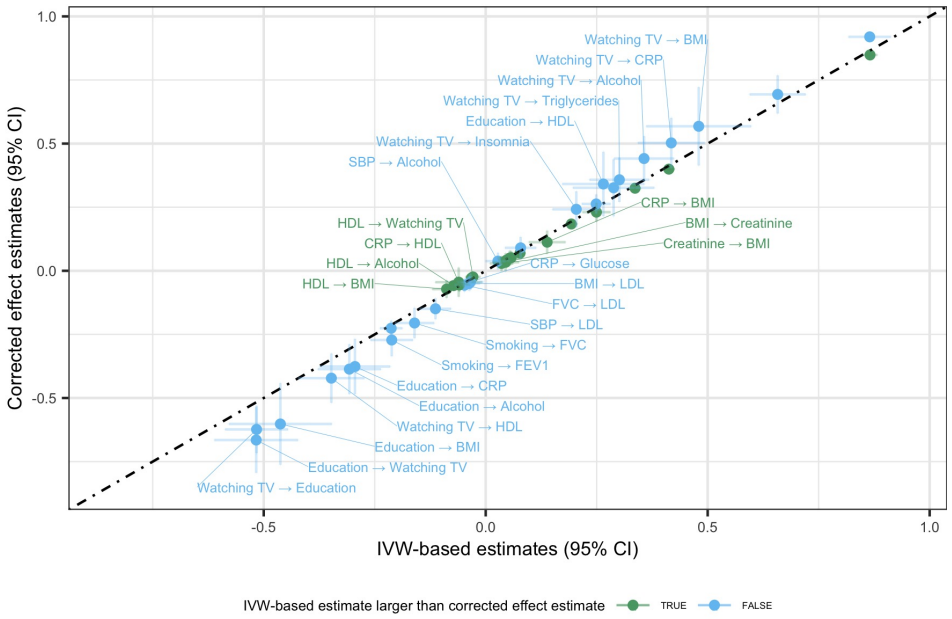


Figure 9: Comparison of IVW-based and corrected effect estimates.
Only the 44 trait pairs for which a significant difference is observed are represented (point estimates and 95% confidence intervals). The color indicates if the IVW-based estimate is larger (in absolute value) than the corrected effect estimate. Only trait pairs where the relative difference between the IVW-based and the corrected effects is larger than 15% are labelled. The dashed line represents the identity line.

Based on these observations, we believe that it should be possible to ignore sample overlap between the risk factors and the focal trait, and that the increase in power resulting from the increase in sample size for the risk factor GWASs would be more important than the relatively small bias that could arise from using overlapping samples in this case. It is also important to note that if the causal effect estimates were biased, this would not result in an increased false positive rate when using Bayes factors to identify associated SNVs. It would only reduce statistical power by making our prior effects less informative.

Interpretation of direct effects under the omnigenic model

One of the new Bayesian GWAS features that have been implemented in the `bGWAS` R-package is the estimation of direct effects (Mounier & Kutalik, 2020). Direct effects are defined as the part of the observed effect that is not mediated through the risk factors and hence cannot be explained by the prior. These direct effects have already been used in a project in collaboration with researchers from the University of Edinburgh (Pirastu et al., 2019). The aim of this project was to study the impact of food consumption on several health related phenotypes using MR. However, food consumption itself can be affected by several traits (risk factors or social determinants) and some of the instruments identified from the food consumption GWASs might have been invalid for MR analyses, if they were primarily acting on other traits. Using the Bayesian GWAS approach allowed us to identify genetic variants with only a direct effect on the food consumption phenotypes, increasing the reliability of subsequent MR analyses.

We believe that our Bayesian GWAS approach and the direct effect estimates that can be obtained offer other applications. They can provide a framework to investigate the omnigenic architecture model, for example (Boyle et al., 2017). Under the omnigenic model, a vast majority of the genome is expected to be associated with every trait. Most of these associations are indirect, through what are called peripheral genes, because of the interconnection of gene regulatory networks and intermediate phenotypes. Some significant proportion of the heritability of complex traits is mediated through gene expression (Yao et al., 2020) and the effects of SNVs acting on gene expression in trans (not in the direct vicinity of a gene) could support the omnigenic hypothesis (Liu et al., 2019). Integrating other -omics data would help better characterize gene regulatory networks.
As an alternative, our Bayesian GWAS approach could be used to take into account the effects of the peripheral genes, through the risk factors, into the prior. Since a part of the peripheral gene effects is supposed to be driven by intermediate phenotypes, these effects are likely to be captured by the prior when these intermediate phenotypes, or traits that are highly correlated with these intermediate phenotypes, are used as risk factors. By defining an adequate set of risk factors, including enough intermediate phenotypes, the direct effects should point towards the core genes. Interestingly, it would also be possible to estimate the heritability of the direct effects, to see what fraction of the total heritability of the trait is mediated through the risk factors and what fraction is due to the core genes identified using the direct effects. It is important to note that if a very large number of risk factors are used to create the prior, there is a risk of underestimating the causal effects, and therefore the part of the effects mediated through the risk factor because of regression dilution bias. We would like to start testing this approach using molecular traits, where signals are often more easy to interpret (Sinnott-Armstrong et al., 2021), before applying it to other complex traits.

The network-LHC model

The LHC-MR model investigates the relationships between two observed traits and a latent heritable confounder. It can be further extended to build a network, including multiple observed traits and latent heritable confounders (network-LHC).

In this network-LHC model, we will consider t continuous random variables: $X = (X_1, X_2, ..., X_t)$. These t random variables can be split into two groups: t_{obs} observed traits $(X_1, X_2, ..., X_{t_{obs}})$, and t_{latent} latent confounders $X_{t_{obs}+1}, ..., X_{t_{obs}+t_{latent}}$. The network-LHC model relies on the same assumptions as the LHC-MR model, including the fact that observed variables are not affecting the latent ones. In addition, network-LHC assumes that latent variables are not affecting each other (to ensure identifiability), and that each trait has a no causal effect on itself. This allows us to define the causal effect matrix, $\boldsymbol{D}$:

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{A} & \mathbf{0} \\ \boldsymbol{B} & \mathbf{0} \end{bmatrix} \quad \text{with} \quad \boldsymbol{A} \in \mathcal{R}^{t_{obs} \times t_{obs}}, \quad \text{and} \quad \boldsymbol{B} \in \mathcal{R}^{t_{latent} \times t_{obs}} \tag{4}$$

$\boldsymbol{A}$ contains the multivariable bi-directional causal effects between observed traits with $diag(\boldsymbol{A}) = \mathbf{0}$

$\boldsymbol{B}$ contains the causal effects of each latent variable on each observed trait with at least two elements of each row of $\boldsymbol{B}$ being non-zero.

We assume that the joint distribution of the traits can be modelled as follows:

$$X = X \cdot \boldsymbol{D} + G \cdot \boldsymbol{\Gamma} + E \qquad \text{with} \qquad E \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{5}$$

where G represents the genome-wide sequence data for genetic markers,

$\boldsymbol{\Gamma}$ contains the direct effect of each genetic marker on each trait,

E is the error term for each trait, with Σ being the residual error variance.

Similarly to what is done in LHC-MR, the likelihood function can be derived, allowing the estimation of the causal effects of observed variables on each other and the causal effects of the latent variables on the observed ones (all elements of the $\boldsymbol{D}$ matrix). In parallel, parameters describing the genetic architecture of each observed trait can also be estimated. This approach needs to be further tested using simulations. For real data, the idea would be to start using small subsets of highly correlated traits to create small networks. The main challenge will be to identify the optimal number of latent confounders. It is also important to note that since the latent variables need to be independent, they would correspond to an orthogonal decomposition of multiple confounders and it would make their interpretation difficult. The approach would still allow a more accurate estimation of bi-directional causal effects, accounting for unobserved confounding factors, and a better understanding of the relationships between complex traits.

Conclusion

Statistical genetics is a field that has been evolving rapidly but major challenges still need to be tackled. Researchers can now have access to large-scale data. While a large amount of GWAS summary statistics has been made publicly available, access to individual-level data can still be sometimes challenging because of ethical concerns.

One aspect that will be crucial in the next years is to gather high-quality data. Most of the data available at the moment is coming from biobanks or population-based cohorts. It would be important to consider other recruitment strategies to avoid selection bias, for example, and to incorporate a familial component into the recruitment strategies, to create large-scale family-based cohorts, with trio (parents/offspring) and/or siblings. Statistical genetics would also benefit from increasing diversity, by including individuals of different ancestries. Remarkably, there have been several initiatives going in this direction in the last years.

High-quality and comprehensive phenotypic data is also needed. One way of doing so is to connect different databases: cohorts with genotypic data and electronic health records, for example. Insurance data or medication purchases could also provide reliable phenotypic data. However, accessing this information comes with additional challenges regarding privacy protection and data security. Resources like the UK Biobank do provide deep phenotyping, but some traits are only measured in a small fraction of individuals. In addition, well-defined and accurately measured phenotypes can be particularly informative. Phenotyping should also focus on environmental factors, such as geolocalisation, exposure to noise pollution, access to sport facilities, etc. Some environmental factors, like physical activity for instance, are not always reliably captured when self-reported. In this case, access to accelerometer-based phenotypic measures for a larger number of individuals would improve the accuracy and the quality of the data. Additional large-scale -omics data, measured in different tissues, would strongly benefit the field. Moreover, longitudinal data, long time follow-up and repeated measurements are needed to better understand the dynamics of some traits, how diseases progress, or study the cumulative effect of environment on some phenotypes over time, for example. These data would allow the use of Granger-causality to test if two events are temporally related and triangulate causal evidence.

It is also important to note that with the recent advances in sequencing technologies, it is likely that whole genome sequencing data will become more widely available and more reliable in the next years. This will allow researchers to more easily study the contribution of rare variants but also other types of genetic variations such as CNVs and SVs.

Getting the full picture of a specific trait requires taking into account other traits that are affecting it, but also considering how changes in this trait might have further consequence on others.

Many biological mechanisms are entangled, making these points a prerequisite to fully understand the relationships between complex traits, and consequently to get a more comprehensive picture of human genetics and pave the way for genome-based precision medicine.

I would like to conclude by reflecting on what I have learned during my thesis. While the projects I have been involved in mostly revolved around method development, I believe that I have also acquired a broad understanding of the current challenges in human genetics. Indeed, I personally believe that new methods can not be designed without keeping in mind future applications. My genuine interest in this field and my general curiosity helped me keep in mind the larger picture while working on very specific topics.
I have been able to develop my autonomy while working on two methodological projects on my own (Mounier & Kutalik, 2020, 2021). In addition, I had the chance to work on two other statistical approaches in collaboration with students of the group (Darrous et al., 2020; Sulc et al., 2020), helping me improve my ability to work with others. Science is a collaborative effort and during my thesis I have also been involved in projects more focused on applying statistical genetics approach, both locally in Lausanne (Tomasoni et al., 2020) and internationally (Pirastu et al., 2019; Timmers et al., 2019).

References

Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic mapping in human disease. *Science*, *322*(5903), 881–888. http://doi.org/10.1126/science.1156409

Aschard, H., Vilhjálmsson, B. J., Greliche, N., Morange, P.-E., Trégouët, D.-A., & Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics*, *94*(5), 662–676. http://doi.org/10.1016/j.ajhg.2014.03.016

Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L., & Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics*, *96*(2), 329–339. http://doi.org/10.1016/j.ajhg.2014.12.021

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. http://doi.org/10.1038/75556

Banos, D. T., McCartney, D. L., Patxot, M., Anchieri, L., Battram, T., Christiansen, C., ... Robinson, M. R. (2020). Bayesian reassessment of the epigenetic architecture of complex traits. *Nature Communications*, *11*(1). http://doi.org/10.1038/s41467-020-16520-1

Barry, C., Liu, J., Richmond, R., Rutter, M. K., Lawlor, D. A., Dudbridge, F., & Bowden, J. (2021). Exploiting collider bias to apply two-sample summary data mendelian randomization methods to one-sample individual level data. *medRxiv*. http://doi.org/10.1101/2020.10.20.20216358

Benner, C., Spencer, C. C. A., Havulinna, A. S., Salomaa, V., Ripatti, S., & Pirinen, M. (2016). FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, *32*(10), 1493–1501. http://doi.org/10.1093/bioinformatics/btw018

Berg, K., & Heiberg, A. (1978). Linkage between familial hypercholesterolemia with xanthomatosis and the C3 polymorphism confirmed. *Cytogenetic and Genome Research*, *22*(1-6), 621–623. http://doi.org/10.1159/000131037

Bjørngaard, J. H., Gunnell, D., Elvestad, M. B., Smith, G. D., Skorpen, F., Krokan, H., ... Romundstad, P. (2012). The causal role of smoking in anxiety and depression: A mendelian randomization analysis of the HUNT study. *Psychological Medicine*, *43*(4), 711–719. http://doi.org/10.1017/s0033291712001274

Bjørnland, T., Langaas, M., Grill, V., & Mostad, I. L. (2017). Assessing gene-environment interaction effects of FTO, MC4R and lifestyle factors on obesity using an extreme phenotype sampling design: Results from the HUNT study. *PLOS ONE*, *12*(4), e0175071. http://doi.org/10.1371/journal.pone.0175071

Bowden, J., Del Greco M., F., Minelli, C., Smith, G. D., Sheehan, N. A., & Thompson, J. R. (2016). Assessing the suitability of summary data for two-sample mendelian randomization analyses using MR-egger regression: The role of the I2 statistic. *International Journal of Epidemiology*, dyw220. http://doi.org/10.1093/ije/dyw220

Bowden, J., Smith, G. D., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, *44*(2), 512–525. http://doi.org/10.1093/ije/dyv080

Bowden, J., Smith, G. D., Haycock, P. C., & Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, *40*(4), 304–314. http://doi.org/10.1002/gepi.21965

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*(7), 1177–1186. http://doi.org/10.1016/j.cell.2017.05.038

Brumpton, B., Eleanor Sanderson, and, Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. Åberge, ... and, N. M. D. (2020). Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses. *Nature Communications*, *11*(1). http://doi.org/10.1038/s41467-020-17117-4

Bulik-Sullivan, B., Hilary K Finucane, and, Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., ... and. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, *47*(11), 1236–1241. http://doi.org/10.1038/ng.3406

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., ... Parkinson, H. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. http://doi.org/10.1093/nar/gky1120

Burgess, S., Butterworth, A., & Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, *37*(7), 658–665. http://doi.org/10.1002/gepi.21758

Burgess, S., Davies, N. M., & Thompson, S. G. (2016). Bias due to participant overlap in two-sample mendelian randomization. *Genetic Epidemiology*, *40*(7), 597–608. http://doi.org/10.1002/gepi.21998

Burgess, S., Freitag, D. F., Khan, H., Gorman, D. N., & Thompson, S. G. (2014). Using multivariable mendelian randomization to disentangle the causal effects of lipid fractions. *PLoS ONE*, *9*(10), e108891. http://doi.org/10.1371/journal.pone.0108891

Burgess, S., Thompson, D. J., Rees, J. M. B., Day, F. R., Perry, J. R., & Ong, K. K. (2017). Dissecting causal pathways using mendelian randomization with summarized genetic data: Application to age at menarche and risk of breast cancer. *Genetics*, *207*(2), 481–487. http://doi.org/10.1534/genetics.117.300191

Burgess, S., & Thompson, S. G. (2011). Bias in causal estimates from mendelian randomization studies with weak instruments. *Statistics in Medicine*, *30*(11), 1312–1323. http://doi.org/10.1002/sim.4197

Burgess, S., & Thompson, S. G. (2015). Multivariable mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology*, *181*(4), 251–260. http://doi.org/10.1093/aje/kwu283

Burgess, S., Thompson, S. G., & CRP CHD Genetics Collaboration. (2011). Avoiding bias from weak instruments in mendelian randomization studies. *International Journal of Epidemiology*, *40*(3), 755–764. http://doi.org/10.1093/ije/dyr036

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... Marchini, J. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. http://doi.org/10.1038/s41586-018-0579-z

Campos, G. de los, Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, *193*(2), 327–345. http://doi.org/10.1534/genetics.112.143313

Canela-Xandri, O., Rawlik, K., & Tenesa, A. (2018). An atlas of genetic associations in UK biobank. *Nature Genetics*, *50*(11), 1593–1599. http://doi.org/10.1038/s41588-018-0248-z

Cardno, A. G., & Gottesman, I. I. (2000). Twin studies of schizophrenia: From bow-and-arrow concordances to star wars mx and functional genomics. *American Journal of Medical Genetics*, *97*(1), 12–17. http://doi.org/10.1002/(sici)1096-8628(200021)97:1<12::aid-ajmg3>3.0.co;2-u

Cardon, L. R., & Palmer, L. J. (2003). Population stratification and spurious allelic association. *The Lancet*, *361*(9357), 598–604. http://doi.org/10.1016/s0140-6736(03)12520-2

Carter, A. R., Sanderson, E., Hammerton, G., Richmond, R. C., Smith, G. D., Heron, J., ... Howe, L. D. (2021). Mendelian randomisation for mediation analysis: Current methods and challenges for implementation. *European Journal of Epidemiology*, *36*(5), 465–478. http://doi.org/10.1007/s10654-021-00757-1

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). http://doi.org/10.1186/s13742-015-0047-8

Chen, S., & Parmigiani, G. (2007). Meta-analysis ofBRCA1andBRCA2Penetrance. *Journal of Clinical Oncology*, *25*(11), 1329–1333. http://doi.org/10.1200/jco.2006.09.1066

Cholesterol Treatment Trialists' Collaborators, Mihaylova, B., Emberson, J., Blackwell, L., Keech, A., Simes, J., ... Baigent, C. (2012). The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: Meta-analysis of individual data from 27 randomised trials. *The Lancet*, *380*(9841), 581–590. http://doi.org/10.1016/s0140-6736(12)60367-5

Corella, D., Ortega-Azorín, C., Sorlí, J. V., Covas, M. I., Carrasco, P., Salas-Salvadó, J., ... Estruch, R. (2012). Statistical and biological gene-lifestyle interactions of MC4R and FTO with diet and physical activity on obesity: New effects on alcohol consumption. *PLoS ONE*, *7*(12), e52344. http://doi.org/10.1371/journal.pone.0052344

Darrous, L., Mounier, N., & Kutalik, Z. (2020). Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. *medRxiv*. http://doi.org/10.1101/2020.01.27.20018929

Davies, N. M., Howe, L. J., Brumpton, B., Havdahl, A., Evans, D. M., & Smith, G. D. (2019). Within family mendelian randomization studies. *Human Molecular Genetics*, *28*(R2), R170–R179. http://doi.org/10.1093/hmg/ddz204

Day, F. R., Loh, P.-R., Scott, R. A., Ong, K. K., & Perry, J. R. B. (2016). A robust example of collider bias in a genetic association study. *The American Journal of Human Genetics*, *98*(2), 392–393. http://doi.org/10.1016/j.ajhg.2015.12.019

Diogo, D., Tian, C., Franklin, C. S., Alanne-Kinnunen, M., March, M., Spencer, C. C. A., ... Runz, H. (2018). Phenome-wide association studies across large population cohorts support drug target validation. *Nature Communications*, *9*(1). http://doi.org/10.1038/s41467-018-06540-3

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., ... Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, *10*(1). http://doi.org/10.1038/s41467-019-11112-0

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: Illuminating the dark road from association to function. *The American Journal of Human Genetics*, *93*(5), 779–797. http://doi.org/10.1016/j.ajhg.2013.10.012

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, *11*(6), 446–450.

http://doi.org/10.1038/nrg2809

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic risk score software. *Bioinformatics*, *31*(9), 1466–1468. http://doi.org/10.1093/bioinformatics/btu848

Finucane, H. K., Brendan Bulik-Sullivan, and, Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., ... and. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, *47*(11), 1228–1235. http://doi.org/10.1038/ng.3404

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*, 99--433. http://doi.org/10.1016/j.cell.2017.05.038

Foley, C. N., Mason, A. M., Kirk, P. D. W., & Burgess, S. (2020). MR-clust: Clustering of genetic variants in mendelian randomization with similar causal estimates. *Bioinformatics*, *37*(4), 531–541. http://doi.org/10.1093/bioinformatics/btaa778

Gamazon, E. R., Heather E Wheeler, and, Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., ... Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098. http://doi.org/10.1038/ng.3367

Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., ... Price, A. L. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics*, *49*(10), 1421–1427. http://doi.org/10.1038/ng.3954

Griffith, G. J., Morris, T. T., Tudball, M. J., Herbert, A., Mancano, G., Pike, L., ... Hemani, G. (2020). Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications*, *11*(1). http://doi.org/10.1038/s41467-020-19478-2

Grotzinger, A. D., Rhemtulla, M., Vlaming, R. de, Ritchie, S. J., Mallard, T. T., Hill, W. D., ... Tucker-Drob, E. M. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour*, *3*(5), 513–525. http://doi.org/10.1038/s41562-019-0566-x

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., ... Martin, J. B. (1983). A polymorphic DNA marker genetically linked to huntington's disease. *Nature*, *306*(5940), 234–238. http://doi.org/10.1038/306234a0

Hartwig, F. P., Davies, N. M., & Smith, G. D. (2018). Bias in mendelian randomization due to assortative mating. *Genetic Epidemiology*, *42*(7), 608–620. http://doi.org/10.1002/gepi.22138

Hartwig, F. P., Smith, G. D., & Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, *46*(6), 1985–1998. http://doi.org/10.1093/ije/dyx102

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., ... Haycock, P. C. (2018). The MR-base platform supports systematic causal inference across the human phenome. *eLife*, *7*. http://doi.org/10.7554/elife.34408

Hill, A. B. (1952). The clinical trial. *New England Journal of Medicine*, *247*(4), 113–119. http://doi.org/10.1056/nejm195207242470401

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*(2), 95–108. http://doi.org/10.1038/nrg1521

Hivert, V., Sidorenko, J., Rohart, F., Goddard, M. E., Yang, J., Wray, N. R., ... Visscher, P. M. (2021). Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *The American Journal of Human Genetics*, *108*(5), 786–798. http://doi.org/10.1016/j.ajhg.2021.02.014

Holmes, M. V., Asselbergs, F. W., Palmer, T. M., Drenos, F., Lanktree, M. B., Nelson, C. P., ... and, J. P. C. (2015). Mendelian randomization of blood lipids for coronary heart disease. *European Heart Journal*, *36*(9), 539–550. http://doi.org/10.1093/eurheartj/eht571

Hormozdiari, F., Bunt, M. van de, Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., ... Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics*, *99*(6), 1245–1260. http://doi.org/10.1016/j.ajhg.2016.10.003

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., & Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, *198*(2), 497–508. http://doi.org/10.1534/genetics.114.167908

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. http://doi.org/10.1038/35057062

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. http://doi.org/10.1038/nature03001

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, *51*(12), 1749–1755. http://doi.org/10.1038/s41588-019-0530-8

Jordan, D. M., Verbanck, M., & Do, R. (2019). HOPS: A quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. *Genome Biology*, *20*(1). http://doi.org/10.1186/s13059-019-1844-7

Joshi, P. K., Pirastu, N., Kentistou, K. A., Fischer, K., Hofer, E., Schraut, K. E., ... Wilson, J. F. (2017). Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nature Communications*, *8*(1). http://doi.org/10.1038/s41467-017-00934-5

Julienne, H., Laville, V., McCaw, Z. R., He, Z., Guillemot, V., Lasry, C., ... Aschard, H. (2020). Multitrait genetic-phenotype associations to connect disease variants and biological mechanisms. *bioRxiv*. http://doi.org/10.1101/2020.06.26.172999

Kanehisa, M. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. http://doi.org/10.1093/nar/28.1.27

Karczewski, K. J., Laurent C. Francioli, and, Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141, 456 humans. *Nature*, *581*(7809), 434–443. http://doi.org/10.1038/s41586-020-2308-7

Kerin, M., & Marchini, J. (2020). A non-linear regression method for estimation of gene–environment heritability. *Bioinformatics*, *36*(24), 5632–5639. http://doi.org/10.1093/bioinformatics/btaa1079

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., ... Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), 1219–1224. http://doi.org/10.1038/s41588-018-0183-z

Klein, R. J. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385–389. http://doi.org/10.1126/science.1109557

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsson, B. J., Young, A. I., Thorgeirsson, T. E., ... Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. *Science*, *359*(6374), 424–428. http://doi.org/10.1126/science.aan6877

Laaksonen, D. E., Niskanen, L., Nyyssonen, K., Lakka, T. A., Laukkanen, J. A., & Salonen, J. T. (2008). Dyslipidaemia as a predictor of hypertension in middle-aged men. *European Heart Journal*, *29*(20), 2561–2568. http://doi.org/10.1093/eurheartj/ehn061

Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., ... Inouye, M. (2021). The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*, *53*(4), 420–425. http://doi.org/10.1038/s41588-021-00783-5

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., & Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLOS Computational Biology*, *12*(1), e1004714. http://doi.org/10.1371/journal.pcbi.1004714

Lawlor, D. A. (2016). Commentary: Two-sample mendelian randomization: Opportunities and challenges. *International Journal of Epidemiology*, *45*(3), 908–915. http://doi.org/10.1093/ije/dyw127

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*. http://doi.org/10.1002/sim.3034

Lee, J. J., Robbee Wedow, and, Okbay, A., Kong, E., Maghzian, O., Zacher, M., ... and. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), 1112–1121. http://doi.org/10.1038/s41588-018-0147-3

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., ... Metspalu, A. (2014). Cohort profile: Estonian biobank of the estonian genome center, university of tartu. *International Journal of Epidemiology*, *44*(4), 1137–1147. http://doi.org/10.1093/ije/dyt268

Leslie, R., ODonnell, C. J., & Johnson, A. D. (2014). GRASP: Analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics*, *30*(12), i185–i194. http://doi.org/10.1093/bioinformatics/btu273

Li, Y. R., & Keating, B. J. (2014). Trans-ethnic genome-wide association studies: Advantages and challenges of mapping in diverse populations. *Genome Medicine*, *6*(10). http://doi.org/10.1186/s13073-014-0091-5

Liu, X., Li, Y. I., & Pritchard, J. K. (2019). Trans effects on gene expression can drive omnigenic inheritance. *Cell*, *177*(4), 1022–1034.e6. http://doi.org/10.1016/j.cell.2019.04.014

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., ... Price, A. L. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284–290. http://doi.org/10.1038/ng.3190

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., ... Moore, H. F. (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. http://doi.org/10.1038/ng.2653

Loos, R. J. F. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, *11*(1). http://doi.org/10.1038/s41467-020-19653-5

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., ... McCarthy, M. I. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, *50*(11), 1505–1513. http://doi.org/10.1038/s41588-018-0241-6

Maier, R. M., Zhu, Z., Lee, S. H., Trzaskowski, M., Ruderfer, D. M., Stahl, E. A., ... Robinson, M. R. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature Communications*, *9*(1). http://doi.org/10.1038/s41467-017-02769-6

Majumder, M. A., Guerrini, C. J., & McGuire, A. L. (2021). Direct-to-consumer genetic testing: Value and risk. *Annual Review of Medicine*, *72*(1), 151–166. http://doi.org/10.1146/annurev-med-070119-114727

Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., & Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics*, *51*(4), 675–682. http://doi.org/10.1038/s41588-019-0367-1

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. http://doi.org/10.1038/nrg2796

Marigorta, U. M., Rodríguez, J. A., Gibson, G., & Navarro, A. (2018). Replicability and prediction: Lessons and challenges from GWAS. *Trends in Genetics*, *34*(7), 504–517. http://doi.org/10.1016/j.tig.2018.03.005

Mars, N., Elisabeth Widén, and, Kerminen, S., Meretoja, T., Pirinen, M., Briotta Parolo, P. della, ... Ripatti, S. (2020). The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nature Communications*, *11*(1). http://doi.org/10.1038/s41467-020-19966-5

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2021). Publisher correction: Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, *53*(5), 763–763. http://doi.org/10.1038/s41588-021-00797-z

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., ... Marchini, J. (2020). Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv*. http://doi.org/10.1101/2020.06.19.162354

McAllister, K., Mechanic, L. E., Amos, C., Aschard, H., Blair, I. A., Chatterjee, N., ... and, J. S. W. (2017). Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *American Journal of Epidemiology*, *186*(7), 753–761. http://doi.org/10.1093/aje/kwx227

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. http://doi.org/10.1038/nrg2344

McDaid, A. F., Joshi, P. K., Porcu, E., Komljenovic, A., Li, H., Sorrentino, V., ... Kutalik, Z. (2017). Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nature Communications*, *8*, 15842. http://doi.org/10.1038/ncomms15842

Mckusick, V. A. (1998). Mendelian inheritance in man: A catalog of human genes and genetic disorders.

Mills, M. C., & Rahal, C. (2020). The GWAS diversity monitor tracks diversity by disease in real time. *Nature Genetics*, *52*(3), 242–243. http://doi.org/10.1038/s41588-020-0580-y

Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., & He, X. (2020). Publisher correction: Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, *52*(7), 750–750. http://doi.org/10.1038/s41588-020-0655-9

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLOS Genetics*, *11*(4), e1004969. http://doi.org/10.1371/journal.pgen.1004969

Mounier, N., & Kutalik, Z. (2020). bGWAS: An r package to perform bayesian genome wide association studies. *Bioinformatics*, *36*(15), 4374–4376. http://doi.org/10.1093/bioinformatics/btaa549

Mounier, N., & Kutalik, Z. (2021). Correction for sample overlap, winner's curse and weak instrument bias in two-sample mendelian randomization. *bioRxiv*. http://doi.org/10.1101/2021.03.26.437168

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Smith, G. D. (2017). Collider scope: When selection bias can substantially influence observed associations. *International Journal of Epidemiology*, *47*(1), 226–235. http://doi.org/10.1093/ije/dyx206

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., ... Yanai, H. (2017). Overview of the BioBank japan project: Study design and profile. *Journal of Epidemiology*, *27*(3), S2–S8. http://doi.org/10.1016/j.je.2016.12.005

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., ... Phillippy, A. M. (2021). The complete sequence of a human genome. *bioRxiv*. http://doi.org/10.1101/2021.05.26.445798

Palmer, C., & Pe'er, I. (2017). Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLOS Genetics*, *13*(7), e1006916. http://doi.org/10.1371/journal.pgen.1006916

Patxot, M., Banos, D. T., Kousathanas, A., Orliac, E. J., Ojavee, S. E., Moser, G., ... Robinson, M. R. (2020). Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits. *bioRxiv*. http://doi.org/10.1101/2020.09.04.20188433

Peltonen, L., Perola, M., Naukkarinen, J., & Palotie, A. (2006). Lessons from studying monogenic disease for common disease. *Human Molecular Genetics*, *15*(suppl_1), R67–R74. http://doi.org/10.1093/hmg/ddl060

Pendergrass, S. A., Brown-Gentry, K., Dudek, S. M., Torstenson, E. S., Ambite, J. L., Avery, C. L., ... Ritchie, M. D. (2011). The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genetic Epidemiology*, *35*(5), 410–422. http://doi.org/10.1002/gepi.20589

Pers, T. H., Juha M. Karjalainen, and, Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., ... Franke, L. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, *6*(1). http://doi.org/10.1038/ncomms6890

Pirastu, N., Mattia Cordioli, and, Nandakumar, P., Mignogna, G., Abdellaoui, A., Hollis, B., ... and. (2021). Genetic analyses identify widespread sex-differential participation bias. *Nature Genetics*, *53*(5), 663–671. http://doi.org/10.1038/s41588-021-00846-7

Pirastu, N., McDonnell, C., Grzeszkowiak, E. J., Mounier, N., Imamura, F., Merino, J., ... Wilson, J. F. (2019). Using genetic variation to disentangle the complex relationship between food intake and health outcomes. *bioRxiv*. http://doi.org/10.1101/829952

Porcu, E., Sina Rüeger, and, Lepik, K., Santoni, F. A., Reymond, A., & and, Z. K. (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications*, *10*(1). http://doi.org/10.1038/s41467-019-10936-0

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. http://doi.org/10.1038/ng1847

Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *The American Journal of Human Genetics*, *69*(1), 1–14. http://doi.org/10.1086/321275

Pulst, S. M. (1999). Genetic linkage analysis. *Archives of Neurology*, *56*(6), 667. http://doi.org/10.1001/archneur.56.6.667

Richardson, T. G., Hemani, G., Gaunt, T. R., Relton, C. L., & Smith, G. D. (2020). A transcriptome-wide mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nature Communications*, *11*(1). http://doi.org/10.1038/s41467-019-13921-9

Richardson, T. G., Sanderson, E., Elsworth, B., Tilling, K., & Smith, G. D. (2020). Use of genetic variation to separate the effects of early and later life adiposity on disease risk: Mendelian randomisation study. *BMJ*, m1203. http://doi.org/10.1136/bmj.m1203

Ried, J. S., M., J. J., Chu, A. Y., Bragg-Gresham, J. L., Dongen, J. van, Huffman, J. E., ... Loos, R. J. F. (2016). A principal component meta-analysis on multiple anthropometric traits identifies novel loci for body shape. *Nature Communications*, *7*(1). http://doi.org/10.1038/ncomms13357

Riet, A. A. van der, Hout, B. A. van, & Rutten, F. F. (1997). Cost effectiveness of DNA diagnosis for four monogenic diseases. *Journal of Medical Genetics*, *34*(9), 741–745. http://doi.org/10.1136/jmg.34.9.741

Robinson, M. R., Aaron Kleinman, and, Graff, M., Vinkhuyzen, A. A. E., Couper, D., Miller, M. B., ... and, P. M. V. (2017). Genetic evidence of assortative mating in humans. *Nature Human Behaviour*, *1*(1). http://doi.org/10.1038/s41562-016-0016

Robinson, M. R., Geoffrey English, and, Moser, G., Lloyd-Jones, L. R., Triplett, M. A., Zhu, Z., ... Visscher, P. M. (2017). Genotype–covariate interaction effects and the heritability of adult body mass index. *Nature Genetics*, *49*(8), 1174–1181. http://doi.org/10.1038/ng.3912

Rowe, S. J., & Tenesa, A. (2012). Human complex trait genetics: Lifting the lid of the genomics toolbox - from pathways to prediction. *Current Genomics*, *13*(3), 213–224. http://doi.org/10.2174/138920212800543101

Sanderson, E. (2021). Multivariable mendelian randomization and mediation. *Cold Spring Harbor Perspectives in Medicine*, *11*(2), a038984. http://doi.org/10.1101/cshperspect.a038984

Sanderson, E., Smith, G. D., Windmeijer, F., & Bowden, J. (2019). An examination of multivariable mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, *48*(3), 713–727. http://doi.org/10.1093/ije/dyy262

Sanson-Fisher, R. W., Bonevski, B., Green, L. W., & D'Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health interventions. *American Journal of Preventive Medicine*, *33*(2), 155–161. http://doi.org/10.1016/j.amepre.2007.04.007

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491–504. http://doi.org/10.1038/s41576-018-0016-z

Sesso, H. D. (2005). A prospective study of plasma lipid levels and hypertension in women. *Archives of Internal Medicine*, *165*(20), 2420. http://doi.org/10.1001/archinte.165.20.2420

Sheehan, N. A., & Didelez, V. (2019). Epidemiology, genetic epidemiology and mendelian randomisation: More need than ever to attend to detail. *Human Genetics*, *139*(1), 121–136. http://doi.org/10.1007/s00439-019-02027-3

Shen, X., David M. Howard, and, Adams, M. J., Hill, W. D., Clarke, T.-K., Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, ... McIntosh, A. M. (2020). A phenome-wide association and mendelian randomisation study of polygenic risk for depression in UK biobank. *Nature Communications*, *11*(1). http://doi.org/10.1038/s41467-020-16022-0

Sinnott-Armstrong, N., Naqvi, S., Rivas, M., & Pritchard, J. K. (2021). GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife*, *10*. http://doi.org/10.7554/elife.58615

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, *177*(1), 26–31. http://doi.org/10.1016/j.cell.2019.02.048

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., ... Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, *89*(5), 607–618. http://doi.org/10.1016/j.ajhg.2011.10.004

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics*, *14*(7), 483–495. http://doi.org/10.1038/nrg3461

Speed, D., & Balding, D. J. (2018). SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics*, *51*(2), 277–284. http://doi.org/10.1038/s41588-018-0279-5

Speed, D., Na Cai, and, Johnson, M. R., Nejentsev, S., & Balding, D. J. (2017). Reevaluation of SNP heritability in complex human traits. *Nature Genetics*, *49*(7), 986–992. http://doi.org/10.1038/ng.3865

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE*, *8*(7), e65245. http://doi.org/10.1371/journal.pone.0065245

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., ... Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, *12*(3), e1001779. http://doi.org/10.1371/journal.pmed.1001779

Sulc, J., Mounier, N., Günther, F., Winkler, T., Wood, A. R., Frayling, T. M., ... Kutalik, Z. (2020). Quantification of the overall contribution of gene-environment interaction for obesity-related traits. *Nature Communications*, *11*(1). http://doi.org/10.1038/s41467-020-15107-0

Sulc, J., Sonrel, A., Mounier, N., Auwerx, C., Marouli, E., Darrous, L., ... Kutalik, Z. (2020). Composite trait mendelian randomization reveals distinct metabolic and lifestyle consequences of differences in body shape. *medRxiv*. http://doi.org/10.1101/2020.09.03.20187567

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., ... Stefansson, K. (2007). Genetic determinants of hair, eye and skin pigmentation in europeans. *Nature Genetics*, *39*(12), 1443–1452. http://doi.org/10.1038/ng.2007.13

Sun, L., Pennells, L., Kaptoge, S., Nelson, C. P., Ritchie, S. C., Abraham, G., ... Angelantonio, E. D. (2021). Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLOS Medicine*, *18*(1), e1003498. http://doi.org/10.1371/journal.pmed.1003498

Sun, Y.-Q., Burgess, S., Staley, J. R., Wood, A. M., Bell, S., Kaptoge, S. K., ... Mai, X.-M. (2019). Body mass index and all cause mortality in HUNT and UK biobank studies: Linear and non-linear mendelian randomisation analyses. *BMJ*, l1042. http://doi.org/10.1136/bmj.l1042

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., & Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, *17*(4), 520–526. http://doi.org/10.1101/gr.6023607

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. http://doi.org/10.1038/nature09534

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. http://doi.org/10.1038/nature15393

The GIANT Consortium. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics*, *41*(1), 25–34. http://doi.org/10.1038/ng.287

The Haplotype Reference Consortium. (2016). A reference panel of 64, 976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. http://doi.org/10.1038/ng.3643

The International HapMap Consortium. (2003). The international HapMap project. *Nature*, *426*(6968), 789–796. http://doi.org/10.1038/nature02168

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. http://doi.org/10.1038/nature04226

The Psychiatric GWAS Consortium Steering Committee. (2008). A framework for interpreting genome-wide association studies of psychiatric disorders. *Molecular Psychiatry*, *14*(1), 10–17. http://doi.org/10.1038/mp.2008.126

Timmers, P. R., Mounier, N., Lall, K., Fischer, K., Ning, Z., Feng, X., ... and, P. K. J. (2019). Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife*, *8*. http://doi.org/10.7554/elife.39856

Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., & Richards, J. B. (2017). Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, *19*(2), 110–124. http://doi.org/10.1038/nrg.2017.101

Toland, A. E., Andrea Forman, and, Couch, F. J., Culver, J. O., Eccles, D. M., Foulkes, W. D., ... Brody, L. C. (2018). Clinical testing of BRCA1 and BRCA2: A worldwide snapshot of technological practices. *Npj Genomic Medicine*, *3*(1). http://doi.org/10.1038/s41525-018-0046-7

Tomasoni, M., Beyeler, M. J., Mounier, N., Porcu, E., Vela, S. O., Button, A. L., ... Bergmann, S. (2020). Genome-wide association studies of retinal vessel tortuosity identify 173 novel loci, capturing genes and pathways associated with disease and vascular tissue pathomechanics. *medRxiv*. http://doi.org/10.1101/2020.06.25.20139725

Tsui, L., Buchwald, M., Barker, D., Braman, J., Knowlton, R., Schumm, J., ... et, al. (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science*, *230*(4729), 1054–1057. http://doi.org/10.1126/science.2997931

Tudball, M. J., Bowden, J., Hughes, R. A., Ly, A., Munafò, M. R., Tilling, K., ... Smith, G. D. (2021). Mendelian randomisation with coarsened exposures. *Genetic Epidemiology*, *45*(3), 338–350. http://doi.org/10.1002/gepi.22376

Turley, P., Raymond K. Walters, and, Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., ... and, D. J. B. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, *50*(2), 229–237. http://doi.org/10.1038/s41588-017-0009-4

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., ... Zheng, W. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, *97*(4), 576–592. http://doi.org/10.1016/j.ajhg.2015.09.001

Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics*, *40*(5), 489–490. http://doi.org/10.1038/ng0508-489

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, *90*(1), 7–24. http://doi.org/10.1016/j.ajhg.2011.11.029

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M., & Yang, Kj. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. http://doi.org/10.1016/j.ajhg.2017.06.005

Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., Cupples, L. A., Shadyab, A. H., ... Visscher, P. M. (2019). Recovery of trait heritability from whole genome sequence data. *bioRxiv*. http://doi.org/10.1101/588020

Wallace, C. (2013). Statistical testing of shared genetic control for potentially related traits. *Genetic Epidemiology*, *37*(8), 802–813. http://doi.org/10.1002/gepi.21765

Watanabe, K., Taskesen, E., Bochoven, A. van, & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, *8*(1). http://doi.org/10.1038/s41467-017-01261-5

Wei, W.-H., Hemani, G., & Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, *15*(11), 722–733. http://doi.org/10.1038/nrg3747

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., ... Parkinson, H. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(D1), D1001–D1006. http://doi.org/10.1093/nar/gkt1229

Widén, E., Junna, N., Ruotsalainen, S., Surakka, I., Mars, N., Ripatti, P., ... Ripatti, S. (2020). Communicating polygenic and non-genetic risk for atherosclerotic cardiovascular disease - an observational follow-up study. *medRxiv*. http://doi.org/10.1101/2020.09.18.20197137

Wood, A. R., Tonu Esko, and, Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., ... and, and. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, *46*(11), 1173–1186. http://doi.org/10.1038/ng.3097

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76–82. http://doi.org/10.1016/j.ajhg.2010.11.011

Yang, J., Teresa Ferreira, and, Morris, A. P., Medland, S. E., Madden, P. A. F., Heath, A. C., ... and, P. M. V. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, *44*(4), 369–375. http://doi.org/10.1038/ng.2213

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, *46*(2), 100–106. http://doi.org/10.1038/ng.2876

Yao, D. W., O'Connor, L. J., Price, A. L., & Gusev, A. (2020). Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics*, *52*(6), 626–633. http://doi.org/10.1038/s41588-020-0625-2

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., ... and, P. M. V. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of european ancestry. *Human Molecular Genetics*, *27*(20), 3641–3649. http://doi.org/10.1093/hmg/ddy271

Young, A. I., Frigge, M. L., Gudbjartsson, D. F., Thorleifsson, G., Bjornsdottir, G., Sulem, P., ... Kong, A. (2018). Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics*, *50*(9), 1304–1310. http://doi.org/10.1038/s41588-018-0178-9

Zdravkovic, S., Wienke, A., Pedersen, N. L., Marenberg, M. E., Yashin, A. I., & Faire, U. D. (2002). Heritability of death from coronary heart disease: A 36-year follow-up of 20 966 swedish twins. *Journal of Internal Medicine*, *252*(3), 247–254. http://doi.org/10.1046/j.1365-2796.2002.01029.x

Zhao, Q., Wang, J., Hemani, G., Bowden, J., & Small, D. S. (2020). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, *48*(3). http://doi.org/10.1214/19-aos1866

Zhong, H., & Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, *9*(4), 621–634. http://doi.org/10.1093/biostatistics/kxn001

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., ... Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50*(9), 1335–1341. http://doi.org/10.1038/s41588-018-0184-y

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. http://doi.org/10.1038/ng.2310

Zhu, H., & Zhou, X. (2020). Statistical methods for SNP heritability estimation and partition: A review. *Computational and Structural Biotechnology Journal*, *18*, 1557–1568. http://doi.org/10.1016/j.csbj.2020.06.011

Appendix A

Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances

This article (Timmers et al., 2019) is presented in Chapter 1.

eLIFE elifesciences.org

RESEARCH COMMUNICATION

Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances

Paul RHJ Timmers[1], Ninon Mounier[2,3], Kristi Lall[4,5], Krista Fischer[4,5], Zheng Ning[6], Xiao Feng[7], Andrew D Bretherick[8], David W Clark[1], eQTLGen Consortium, Xia Shen[1,6,7], Tõnu Esko[4,9], Zoltán Kutalik[2,3], James F Wilson[1,8], Peter K Joshi[1,2]*

[1]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, United Kingdom; [2]Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland; [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [4]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; [5]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; [6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [7]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China; [8]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; [9]Broad Institute of Harvard and MIT, Cambridge, United States

Abstract We use a genome-wide association of 1 million parental lifespans of genotyped subjects and data on mortality risk factors to validate previously unreplicated findings near *CDKN2B-AS1*, *ATXN2/BRAP*, *FURIN/FES*, *ZW10*, *PSORS1C3*, and 13q21.31, and identify and replicate novel findings near *ABO*, *ZC3HC1*, and *IGF2R*. We also validate previous findings near 5q33.3/*EBF1* and *FOXO3*, whilst finding contradictory evidence at other loci. Gene set and cell-specific analyses show that expression in foetal brain cells and adult dorsolateral prefrontal cortex is enriched for lifespan variation, as are gene pathways involving lipid proteins and homeostasis, vesicle-mediated transport, and synaptic function. Individual genetic variants that increase dementia, cardiovascular disease, and lung cancer – but not other cancers – explain the most variance. Resulting polygenic scores show a mean lifespan difference of around five years of life across the deciles.

Editorial note: This article has been through an editorial process in which the authors decide how to respond to the issues raised during peer review. The Reviewing Editor's assessment is that all the issues have been addressed (see decision letter).

DOI: https://doi.org/10.7554/eLife.39856.001

***For correspondence:** peter.joshi@ed.ac.uk

Group author details: eQTLGen Consortium See page 31

Competing interest: See page 33

Funding: See page 33

Received: 05 July 2018
Accepted: 20 November 2018
Published: 15 January 2019

Introduction

Human lifespan is a highly complex trait, the product of myriad factors involving health, lifestyle, genetics, environment, and chance. The extent of the role of genetic variation in human lifespan has been widely debated (***van den Berg et al., 2017***), with estimates of broad sense heritability ranging from around 25% based on twin studies (***Ljungquist et al., 1998***; ***Herskind et al., 1996***; ***McGue et al., 1993***) (perhaps over-estimated [***Young et al., 2018***]) to around 16.1%, (narrow sense

eLife digest Ageing happens to us all, and as the cabaret singer Maurice Chevalier pointed out, "old age is not that bad when you consider the alternative". Yet, the growing ageing population of most developed countries presents challenges to healthcare systems and government finances. For many older people, long periods of ill health are part of the end of life, and so a better understanding of ageing could offer the opportunity to prolong healthy living into old age.

Ageing is complex and takes a long time to study – a lifetime in fact. This makes it difficult to discern its causes, among the countless possibilities based on an individual's genes, behaviour or environment. While thousands of regions in an individual's genetic makeup are known to influence their risk of different diseases, those that affect how long they will live have proved harder to disentangle. Timmers et al. sought to pinpoint such regions, and then use this information to predict, based on their DNA, whether someone had a better or worse chance of living longer than average.

The DNA of over 500,000 people was read to reveal the specific 'genetic fingerprints' of each participant. Then, after asking each of the participants how long both of their parents had lived, Timmers et al. pinpointed 12 DNA regions that affect lifespan. Five of these regions were new and had not been linked to lifespan before. Across the twelve as a whole several were known to be involved in Alzheimer's disease, smoking-related cancer or heart disease. Looking at the entire genome, Timmers et al. could then predict a lifespan score for each individual, and when they sorted participants into ten groups based on these scores they found that top group lived five years longer than the bottom, on average.

Many factors beside genetics influence how long a person will live and our lifespan cannot be read from our DNA alone. Nevertheless, Timmers et al. had hoped to narrow down their search and discover specific genes that directly influence how quickly people age, beyond diseases. If such genes exist, their effects were too small to be detected in this study. The next step will be to expand the study to include more participants, which will hopefully pinpoint further genomic regions and help disentangle the biology of ageing and disease.

DOI: https://doi.org/10.7554/eLife.39856.002

12.2%) based on large-scale population data (***Kaplanis et al., 2018***). One very recent study suggests it is much lower still (<7%) (***Ruby et al., 2018***), pointing to assortative mating as the source of resemblance amongst kin.

Despite this modest heritability, extensive research has gone into genome-wide association studies (GWAS) finding genetic variants influencing human survival, using a variety of trait definitions and study designs (***Deelen et al., 2011***; ***Sebastiani et al., 2012***; ***Beekman et al., 2013***; ***Broer et al., 2015***; ***Joshi et al., 2016***; ***Pilling et al., 2016***; ***Zeng et al., 2016***; ***Pilling et al., 2017***). GWAS have primarily focused on extreme cases of long-livedness (longevity) – individuals surviving past a certain age threshold – and scanning for differences in genetic variation from controls. While this case-control design has the advantage of focusing on highly statistically-informative individuals, who also often exhibit extreme healthspan and have potentially unique genetic attributes (***Sebastiani et al., 2013***; ***Sebastiani et al., 2016***), the exceptional nature of the phenotype precludes collection of large samples, and differences in definitions of longevity complicate meta-analysis. As a result, only two robustly replicated, genome-wide significant associations (near *APOE* and *FOXO3*) have been made to date (***Broer et al., 2015***; ***Deelen et al., 2014***).

An alternative approach is to study lifespan as a quantitative trait in the general population and use survival models (such as Cox proportional hazards [***Cox, 1972***]) to allow long-lived survivors to inform analysis. However, given the incidence of mortality in middle-aged subjects is low, studies have shifted to the use of parental lifespans with subject genotypes (an instance of Wacholder's kin-cohort method [***Wacholder et al., 1998***]), circumventing the long wait associated with studying age at death in a prospective study (***Joshi et al., 2016***; ***Pilling et al., 2016***). In addition, the recent increase in genotyped population cohorts around the world, and in particular the creation of UK Biobank (***Bycroft et al., 2017***), has raised GWAS sample sizes to hundreds of thousands of individuals, providing the statistical power necessary to detect genetic effects on mortality.

A third approach is to gather previously published GWAS on risk factors thought to possibly affect lifespan, such as smoking behaviour and cardiovascular disease (CVD), and estimate their actual independent, causal effects on mortality using Mendelian Randomisation. These causal estimates can then be used in a Bayesian framework to inform previously observed SNP associations with lifespan (**McDaid et al., 2017**).

Here, we blend these three approaches to studying lifespan and perform the largest GWAS on human lifespan to date. First, we leverage data from UK Biobank and 26 independent European-heritage population cohorts (**Joshi et al., 2017**) to carry out a GWAS of parental survival, quantified using Cox models. We then supplement this with data from 58 GWAS on mortality risk factors to conduct a Bayesian prior-informed GWAS (iGWAS). Finally, we use publicly available case-control longevity GWAS statistics to compare the genetics of lifespan and longevity and provide collective replication of our lifespan GWAS results.

We also examine the diseases associated with lifespan-altering variants and the effect of known disease variants on lifespan, to provide insight into the interplay between lifespan and disease. Finally, we use our GWAS results to implicate specific genes, biological pathways, and cell types, and use our findings to create and test whole-genome polygenic scores for survival.

Results

Genome-wide association analysis

We carried out GWAS of survival in a sample of 1,012,240 parents (60% deceased) of European ancestry from UK Biobank and a previously published meta-analysis of 26 additional population cohorts (LifeGen [**Joshi et al., 2017**]; **Table 1—source data 1**). We performed a sex-stratified analysis and then combined the allelic effects in fathers and mothers into a single parental survival association in two ways. First, we assumed genetic variants with common effect sizes (CES) for both parents, maximising power if the effect is indeed the same. Second, we allowed for sex-specific effect sizes (SSE), maximising power to detect sexually dimorphic variants, including those only affecting one sex. The latter encompasses a conventional sex-stratified analysis, but uses only one statistical test for the much more general alternative hypothesis that there is an effect in at least one sex.

We find 12 genomic regions with SNPs passing genome-wide significance for one or both analyses ($p < 2.5 \times 10^{-8}$, accounting for the two tests CES/SSE) (**Figure 1**; **Table 1**). Among these are five loci discovered here for the first time, at or near *MAGI3*, *KCNK3*, *HTT*, *HP*, and *LDLR*. Carrying one copy of a life-extending allele is associated with an increase in lifespan between 0.23 and 1.07 years (around 3 to 13 months). Despite our sample size exceeding 1 million phenotypes, a variant had to have a minor allele frequency exceeding 5% and an effect size of 0.35 years of life or more per allele for our study to detect it with 80% power.

We also attempted to validate novel lifespan SNPs discovered by **Pilling et al. (2017)** in UK Biobank at an individual level by using the LifeGen meta-analysis as independent replication sample. Testing 20 candidate SNPs for which we had data available, we find directionally consistent, nominally significant associations for six loci ($p < 0.05$, one-sided test), of which three have sex-specific effects. We also provide evidence against three putative loci but lack statistical power to assess the remaining 11 (**Figure 2**, **Figure 2—source data 1**).

We then used our full sample to test six candidate SNPs previously associated with longevity (**Zeng et al., 2016**; **Deelen et al., 2014**; **Flachsbart et al., 2009**; **Sebastiani et al., 2017**) for association with lifespan, and find directionally consistent evidence for SNPs near *FOXO3* and *EBF1*. The remaining SNPs did not associate with lifespan despite apparently adequate power to detect any effect similar to that originally reported (**Figure 2**, **Figure 2—source data 1**).

Finally, we tested a deletion, d3-GHR, reported to affect male lifespan by 10 years when homozygous (**Ben-Avraham et al., 2017**) by converting its effect size to one we expect to observe when fitting an additive model. We used a SNP tagging the deletion and estimated the expected effect size in a linear regression for the (postulated) recessive effect across the three genotypes, given their frequency (see Materials and methods). While this additive model reduces power relative to the correct model, our large sample size is more than able to offset the loss of power, and we find evidence d3-

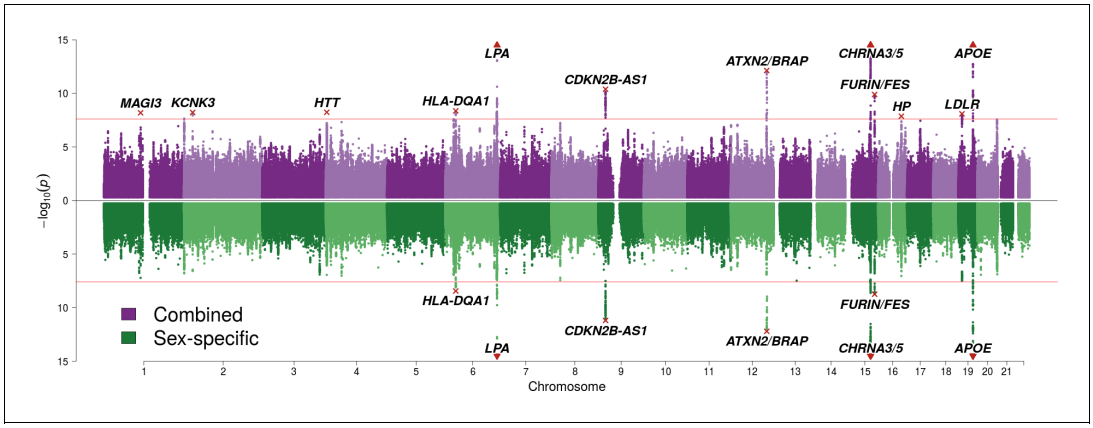


Figure 1. SNP associations with lifespan across both parents under the assumption of common and sex-specific effect sizes. Miami plot of genetic associations with joint parental survival. In purple are the associations under the assumption of common SNP effect sizes across sexes (CES); in green are the associations under the assumption of sex-specific effect sizes (SSE). P refers to the two-sided P values for association of allelic dosage on survival under the residualised Cox model. The red line represents our multiple testing-adjusted genome-wide significance threshold ($p = 2.5 \times 10^{-8}$). Annotated are the gene, set of genes, or cytogenetic band near the index SNP, marked in red. P values have been capped at $-\log_{10}(p) = 15$ to better visualise associations close to genome-wide significance. SNPs with P values beyond this cap (near *APOE*, *CHRNA3/5* and *LPA*) are represented by triangles.

DOI: https://doi.org/10.7554/eLife.39856.007

GHR does not associate with lifespan with any (recessive or additive) effect similar to that originally reported (***Figure 2***, ***Figure 2—source data 1***).

Mortality risk factor-informed GWAS (iGWAS)

We integrated 58 publicly available GWAS on mortality risk factors with our CES lifespan GWAS, creating Bayesian priors for each SNP effect based on causal effect estimates of 16 independent risk factors on lifespan. These included body mass index, blood biochemistry, CVD, type 2 diabetes, schizophrenia, multiple sclerosis, education levels, and smoking traits.

The integrated analysis reveals an additional seven genome-wide significant associations with lifespan (Bayes Factor permutation $p < 2.5 \times 10^{-8}$), of which SNPs near *TMEM18*, *GBX2/ASB18*, *IGF2R*, *POM12C*, *ZC3HC1*, and *ABO* are reported at genome-wide significance for the first time (***Figure 3***; ***Table 2***). A total of 82 independent SNPs associate with lifespan when allowing for a 1% false discovery rate (FDR) (***Table 2—source data 2***).

As has become increasingly common (***Pilling et al., 2017***), we attempted to replicate our genome-wide significant findings collectively, rather than individually. This is usually done by constructing polygenic risk scores from genotypic information in an independent cohort and testing for association with the trait of interest subject-by-subject. We used publicly available summary statistics on extreme longevity as an independent replication dataset (***Broer et al., 2015***; ***Deelen et al., 2014***), but lacking individual data from such studies, we calculated the collective effect of lifespan SNPs on longevity using the same method as inverse-variance meta-analysis two-sample Mendelian randomisation (MR) using summary statistics (***Hemani et al., 2018***), which gives equivalent results. Prior to doing this, all effects observed in the external longevity studies were converted to hazard ratios using the *APOE* variant effect size as an empirical conversion factor, to allow the longevity studies to be meta-analysed despite their different study designs (and to be adjusted for sample overlap; see Materials and methods).

Although the focus is on collective replication, our method has the advantage of transparency at an individual variant level, which is of particular importance for researchers seeking to follow-up individual loci. Remarkably, all lead lifespan variants show directional consistency with the independent longevity sample, and 4 SNPs or close proxies ($r^2 > 0.8$) reach nominal replication ($p < 0.05$, one-sided test) (***Figure 4—source data 1***). Of these, SNPs near *ABO*, *ZC3HC1*, and *IGF2R* are replicated for the first time, and thus appear to affect overall survival and survival to extreme age. The overall ratio of replication effect sizes to discovery effect sizes – excluding *APOE* – is 0.42 (95% CI 0.23–

Table 1. Twelve genome-wide significant associations with lifespan using UK Biobank and LifeGen.

Parental phenotypes from UK Biobank and LifeGen meta-analysis, described in ***Table 1—source data 1***, were tested for association with subject genotype. See ***Table 1—source data 2*** for LD Score regression intercept of each cohort separately and combined. Displayed here are loci associating with lifespan at genome-wide significance ($p < 2.5 \times 10^{-8}$). At or near – Gene, set of genes, or cytogenetic band nearest to the index SNP; rsID – The index SNP with the lowest P value in the standard or sex-specific effect (SSE) analysis. Chr – Chromosome; Position – Base-pair position on chromosome (GRCh37); A1 – the effect allele, increasing lifespan; Freq1 – Frequency of the A1 allele; Years1 – Years of life gained for carrying one copy of the A1 allele; SE – Standard Error; P – the P value for the Wald test of association between imputed dosage and cox model residual; Disease – Category of disease for known associations with SNP or close proxies ($r^2 > 0.6$), see ***Table 1—source data 3*** for details and references. Despite the well-known function of the *HTT* gene in Huntington's disease, SNPs within the identified locus near this gene have not been associated with the disease at genome-wide significance.

At or near	rsID	Chr	Position	A1	Freq1	Years1	SE	P	SSE P	Disease
MAGI3	rs1230666	1	114173410	G	0.85	0.3224	0.0555	6.4E-09	6.1E-08	Autoimmune
KCNK3	rs1275922	2	26932887	G	0.74	0.2579	0.0443	6.0E-09	2.7E-07	Cardiometabolic
HTT	rs61348208	4	3089564	T	0.39	0.2299	0.0395	5.8E-09	1.2E-07	-
HLA-DQA1	rs34967069	6	32591248	T	0.07	0.5613	0.0956	4.3E-09	3.6E-09	Autoimmune
LPA	rs10455872	6	161010118	A	0.92	0.7639	0.0743	8.5E-25	3.1E-24	Cardiometabolic
CDKN2B-AS1	rs1556516	9	22100176	G	0.50	0.2510	0.0386	7.5E-11	6.4E-12	Cardiometabolic
ATXN2/BRAP	rs11065979	12	112059557	C	0.56	0.2798	0.0393	1.0E-12	6.2E-13	Autoimmune/ Cardiometabolic
CHRNA3/5	rs8042849	15	78817929	T	0.65	0.4368	0.0410	1.6E-26	1.9E-30	Smoking-related
FURIN/FES	rs6224	15	91423543	G	0.52	0.2507	0.0390	1.3E-10	1.8E-09	Cardiometabolic
HP	rs12924886	16	72075593	A	0.80	0.2798	0.0493	1.4E-08	9.1E-08	Cardiometabolic
LDLR	rs142158911	19	11190534	A	0.12	0.3550	0.0616	8.1E-09	3.3E-08	Cardiometabolic
APOE	rs429358	19	45411941	T	0.85	1.0561	0.0546	3.1E-83	1.8E-85	Cardiometabolic/ Neuropsychiatric

DOI: https://doi.org/10.7554/eLife.39856.003

The following source data is available for Table 1:

Source data 1. Descriptive statistics of the cohorts and lives analysed.

Summary statistics for the 1,012,240 parental lifespans passing phenotypic QC (most notably, parent age > 40). In practice, fewer lives than these were analysed for some SNPs, as a SNP may not have passed QC in all cohorts (in particular LifeGen MAF > 1%). Ancestries in UK Biobank are self-declared, except in the case of Gen. British. Gen. British – Participants identified as genomically British by UK Biobank, based on their genomic profile. LifeGen – A consortium of 26 population cohorts of European Ancestry, with UK Biobank lives removed.

DOI: https://doi.org/10.7554/eLife.39856.004

Source data 2. LD-score regression intercepts for GWAS results.

Regression intercepts (standard error) of the GWAS summary statistics as calculated by LD-score regression, using LD scores from on average 457,407 SNPs from the UK Biobank array. CES – Results under the assumption of common effect sizes across sexes, SSE – Results allowing for sex-specific effects.

DOI: https://doi.org/10.7554/eLife.39856.005

Source data 3. Known associations with genome-wide significant lifespan loci.

Genome-wide significant associations from the GWAS catalog and PhenoScanner are reported for the lead SNP and proxies ($r^2 > 0.6$). Similar associations have been grouped, keeping the most significant association and the shortest trait name (Trait). At or near – Gene or cluster of genes in close proximity to lead SNP; A1 – the effect allele, increasing lifespan; A0 – the reference allele. Freq1- Frequency of the A1 allele in the original study, or if missing, averaged from all associations; Beta1 – the reported effect on the trait for carrying one copy of the A1 allele; SE – Standard Error; P – P value; Disease – the type of lifespan-shortening diseases linked to the trait, or 'other' if the link is unclear or multiple disease links exist.

DOI: https://doi.org/10.7554/eLife.39856.006

0.61; $p = 1.35 \times 10^{-5}$). The fact this ratio is significantly greater than zero indicates most lifespan SNPs are indeed longevity SNPs. However, the fact most SNPs have a ratio smaller than one indicates they may affect early mortality more than survival to extreme age, relative to *APOE* (which itself has a greater effect on late-life mortality than early mortality) (***Figure 4***).

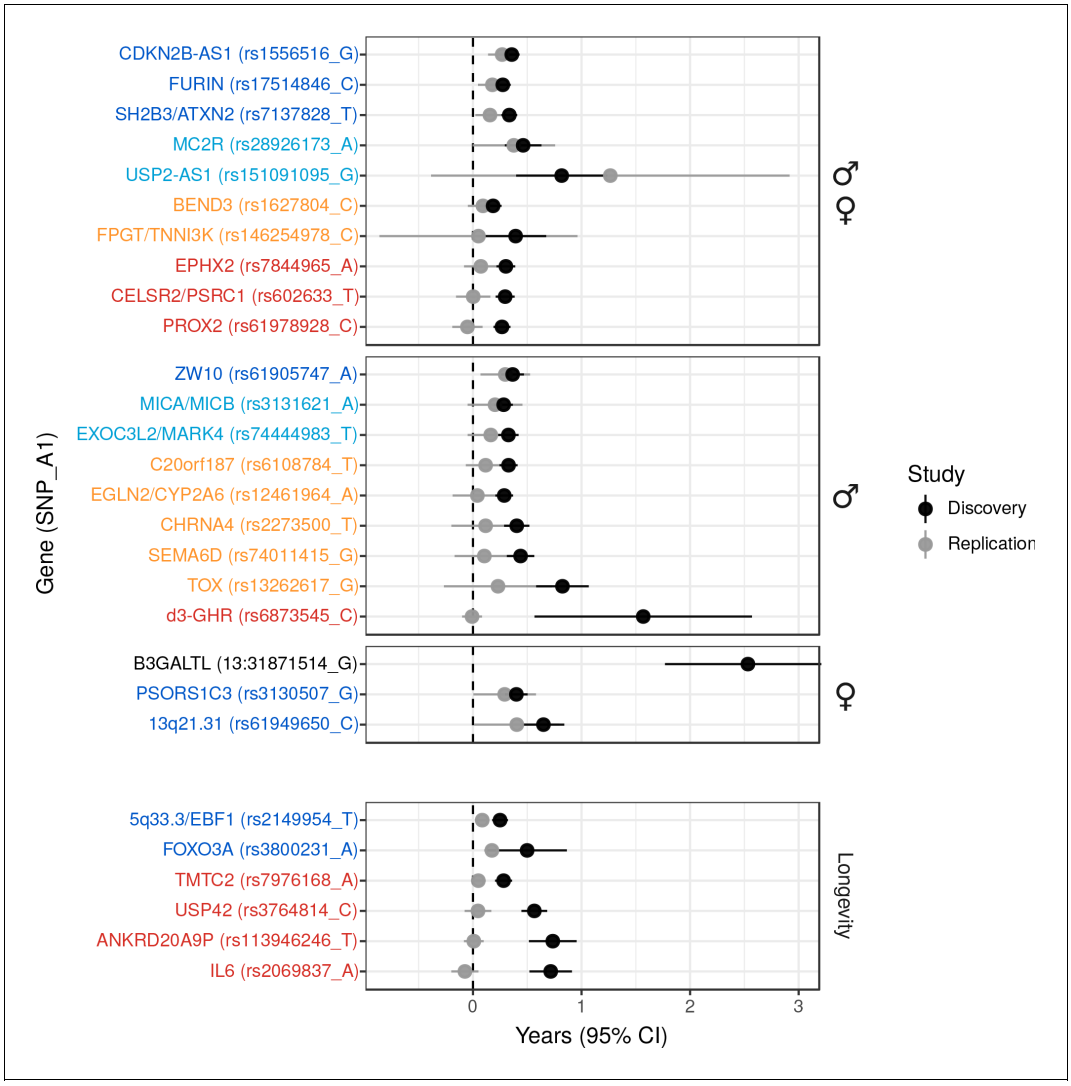


Figure 2. Validation of SNPs identified in other studies using independent samples of European descent. Discovery – Candidate SNPs or proxies ($r^2 > 0.95$) associated with lifespan (top panels, stratified by sex) and longevity (bottom panel) by previous studies (***Zeng et al., 2016***; ***Pilling et al., 2017***; ***Deelen et al., 2014***; ***Flachsbart et al., 2009***; ***Sebastiani et al., 2017***; ***Ben-Avraham et al., 2017***). Effect sizes have been rescaled to years of life to make direct comparisons between studies (see Materials and methods and ***Figure 2—figure supplement 1***). Replication – Independent samples, either the LifeGen meta-analysis to replicate ***Pilling et al. (2017)***, or the full dataset including UK Biobank. Gene names are as reported by discovery and have been coloured based on overlap between confidence intervals (CIs) of effect estimates. Dark blue – Nominal replication ($p < 0.05$, one-sided test). Light blue – CIs overlap ($P_{het} > 0.05$) and cover zero, but replication estimate is closer to discovery than zero. Yellow – CIs overlap ($P_{het} > 0.05$) and cover zero, and replication estimate is closer to zero than discovery. Red – CIs do not overlap ($P_{het} < 0.05$) and replication estimate covers zero. Black – no replication data.

DOI: https://doi.org/10.7554/eLife.39856.008

The following source data and figure supplement are available for figure 2:

Source data 1. Eight candidate lifespan regions replicate nominally ($p < 0.05$) in LifeGen or our full sample.

DOI: https://doi.org/10.7554/eLife.39856.010

Figure supplement 1. Concordance between inferred effect sizes from ***Pilling et al. (2017)*** and our estimated effect sizes in a largely overlapping UK Biobank sample.

DOI: https://doi.org/10.7554/eLife.39856.009

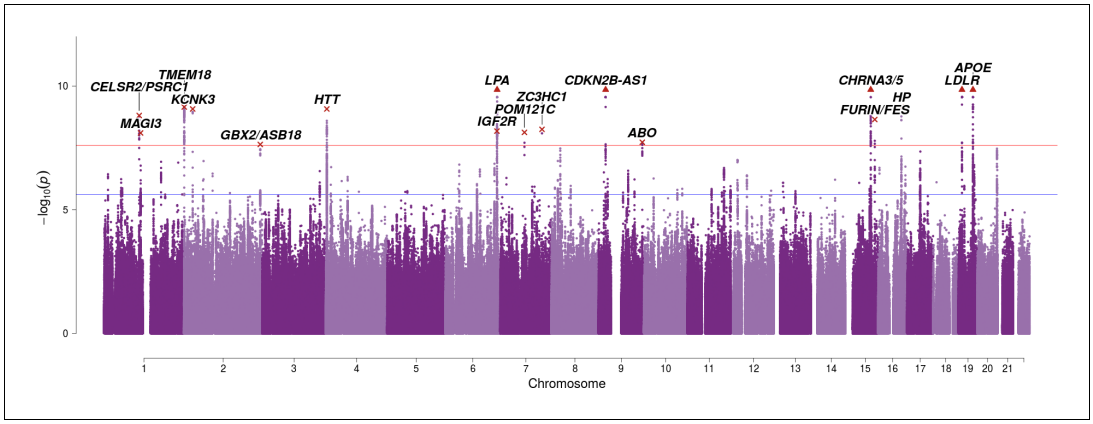


Figure 3. SNP associations with lifespan across both parents when taking into account prior information on mortality risk factors. Bayesian iGWAS was performed using observed associations from the lifespan GWAS and priors based on 16 traits selected by an AIC-based stepwise model. As the P values were assigned empirically using a permutation approach, the minimum P value is limited by the number of permutations; SNPs reaching this limit are represented by triangles. Annotated are the gene, cluster of genes, or cytogenetic band in close proximity to the top SNP. The red line represents the genome-wide significance threshold ($p = 2.5 \times 10^{-8}$). The blue line represents the 1% FDR threshold. ***Figure 3—figure supplement 1*** shows the associations of each genome-wide significant SNP with the 16 risk factors.

DOI: https://doi.org/10.7554/eLife.39856.011

The following figure supplement is available for figure 3:

Figure supplement 1. Heat map of the effect of genome-wide significant iGWAS SNPs on the mortality risk factors.

DOI: https://doi.org/10.7554/eLife.39856.012

Sex- and age-specific effects

We stratified our UK Biobank sample (for which we had individual level data) by sex and age bands to identify sex- and age-specific effects for survival SNPs discovered and/or replicated in this study. Although power was limited, as we sought contrasts in small effect sizes, we find 5 SNPs with differential effects on lifespan when stratified (FDR 5% across the 24 variants considered).

The effect of the *APOE* variant increases with age: the ε4 log hazard ratio on individuals older than 70 years is around 3 times greater than those between ages 40–70. In contrast, the effect of lead variants near *CHRNA3/5*, *CDKN2B-AS1*, and *ABO* tends to decline after age 60, at least when expressed as hazard ratios (***Figure 5A***).

Independent of age, lead variants near *APOE* and *PSORS1C3* also show an effect (lnHR) of 0.036; 0.038 greater in women (95% CI 0.013–0.059; 0.019–0.056, respectively), compared to men (***Figure 5B***). Notably, the SNP near *ZW10*, which was identified by ***Pilling et al. (2017)*** in fathers, and which replicated in LifeGen fathers, may affect men and women equally (95% CI years gained per effect allele, men 0.17–0.42, women 0.04–0.31), as measured in our meta-analysis of UK Biobank and LifeGen.

Causal genes and methylation sites

We used SMR-HEIDI to look for causal effects of gene expression or changes in methylation on lifespan within the 24 loci discovered or replicated in our study. Using blood eQTL summary statistics from two studies (***Westra et al., 2013***; ***Lloyd-Jones et al., 2017***), we suggest causal roles for expression of *PSRC1*, *SESN1*, *SH2B3*, *PSMA4*, *FURIN*, *FES*, and *KANK2* at 5% FDR (***Supplementary file 1***). GTEx tissue-wide expression data suggests further roles for 16 genes across 24 tissues, especially *FES* (nine tissues), *PMS2P3* (six tissues) and *PSORS1C1* (four tissues). Methylation data reveals roles for 44 CpG sites near nine loci, especially near the *PSORS1C3* locus (21 sites), *APOE* locus (nine sites), and *HLA-DQA1* locus (four sites) (***Supplementary file 2***).

We next used SOJO to perform conditional analysis on the same loci to find additional independent variants associated with lifespan. We find substantial allelic heterogeneity in several association intervals and identify an additional 335 variants, which increase out-of-sample explained variance from 0.095% to 0.169% (78% increase). *CELSR2/PSRC1*, *KCNK3*, *HLA-DQA1*, *LPA*, *ZW10*, *FURIN/*

Table 2. Bayesian GWAS using mortality risk factors reveals seven additional genome-wide significant variants.
At or near – Gene or set of genes nearest to the index SNP; rsID – The index SNP with the lowest P value in the risk factor-informed analysis. Chr – Chromosome; Position – Base-pair position on chromosome (GRCh37); A1 – the effect allele, increasing lifespan; Freq1 – Frequency of the A1 allele; Years1 – Years of life gained for carrying one copy of the A1 allele; SE – Standard Error; CES P – the P value for the Wald test of association between imputed dosage and cox model residual, under the assumption of common effects between sexes. Risk – mortality risk factors associated with the variant ($p < 3.81 \times 10^{-5}$, accounting for 82 independent SNPs and 16 independent factors). BF P – Empirical P value derived from permutating Bayes Factors. See ***Table 2—source data 1*** for the causal estimate of each risk factor. See ***Table 2—source data 2*** for all SNPs significant at FDR < 1%.

At or near	rsID	Chr	Position	A1	Freq1	Years1	SE	CES P	Risk	BF P
CELSR2/PSRC1	rs4970836	1	109821797	G	0.23	0.2234	0.0463	1.4E-06	LDL HDL CAD	1.6E-09
TMEM18	rs6744653	2	628524	A	0.17	0.2772	0.0511	5.8E-08	BMI	7.0E-10
GBX2/ASB18	rs10211471	2	237081854	C	0.80	0.2401	0.0493	1.1E-06	Education	2.3E-08
IGF2R	rs111333005	6	160487196	G	0.98	0.8665	0.1577	3.9E-08	LDL CAD	6.6E-09
POM121C	rs113160991	7	75094329	G	0.78	0.2541	0.0495	2.8E-07	BMI Insulin	7.5E-09
ZC3HC1	rs56179563	7	129685597	A	0.39	0.2107	0.0406	2.1E-07	CAD	5.6E-09
ABO	rs2519093	9	136141870	C	0.81	0.2244	0.0497	6.3E-06	LDL CAD	1.9E-08

DOI: https://doi.org/10.7554/eLife.39856.013

The following source data is available for Table 2:

Source data 1. Bayesian GWAS - Multivariate effect estimates for the 16 traits chosen by the AIC based stepwise model selection.
The multivariate MR identified 16 traits (58 tested, see ***McDaid et al., 2017*** for an exhaustive list) with significant causal effect on lifespan and used the effect estimates to create the prior assumption of the expected effect size of each variant on lifespan, in the (Bayesian) iGWAS. Effect Estimate – the estimated effect of standardized trait on standardized lifespan, in multivariate model. SE – the standard error of the estimated effect, in multivariate model. P – the P value (two sided) from MR, for testing association between standardized trait and standardized lifespan, in multivariate model.
DOI: https://doi.org/10.7554/eLife.39856.014

Source data 2. 82 SNPs significantly associated with lifespan at 1% FDR and the SNP's associations with risk factors.
Bayesian iGWAS was performed using observed association results from CES GWAS and priors from 16 risk factors selected by AIC based stepwise model selection. Bayes Factors were calculated to compare effect estimates observed in the conventional GWAS to the prior effect computed. Empirical P values were assigned using a permutation approach and further corrected for multiple testing using Benjamini-Hochberg correction. Chr – Chromosome, Position – Base-pair position on chromosome (GRCh37), A1 – Effect Allele, Freq1 – Frequency of the A1 allele (from conventional GWAS), Beta1 (from conventional GWAS), SE – Standard Error of Beta1, Years – Years of lifespan gained for carrying one copy of the A1 allele (from conventional GWAS), P – P value (from conventional GWAS), PriorEffect – Prior effect estimate calculated from the summary statistics data for the 16 risk factors identified, PriorSE – Standard Error of the prior effect estimate, LogBF – Log of the observed Bayes Factor, P_BF – Empirical P value from a permutation approach for the log Bayes Factor. Final columns show the P value of each SNP in the studies used to calculate the prior, if the P value is significant after Bonferroni multiple testing correction ($p < 3.81 \times 10^{-5}$, 82*16 tests) the cell is shaded green. Counts of these significant associations by SNP/trait are shown in the final column/row.
DOI: https://doi.org/10.7554/eLife.39856.015

FES, and *APOE* are amongst the most heterogeneous loci with at least 25 variants per locus showing independent effects (***Supplementary file 3***).

Disease and lifespan

We next sought to understand the link between our lifespan variants and disease. We looked up known associations with our top hits and proxies ($r^2 > 0.6$) in the GWAS catalog (***MacArthur et al., 2017***) and PhenoScanner (***Staley et al., 2016***), excluding loci identified in iGWAS as these used disease associations to build the effect priors. We also excluded trait associations discovered solely in UK Biobank, as the overlap with our sample could result in spurious association due to correlations between morbidity and mortality. Under these restrictions, we find alleles which increase lifespan associate with a reduction in cardiometabolic, autoimmune, smoking-related, and neuropsychiatric disease and their disease risk factors (***Table 1***, ***Table 1—source data 3***). None of the loci show any association with cancer other than lung cancer.

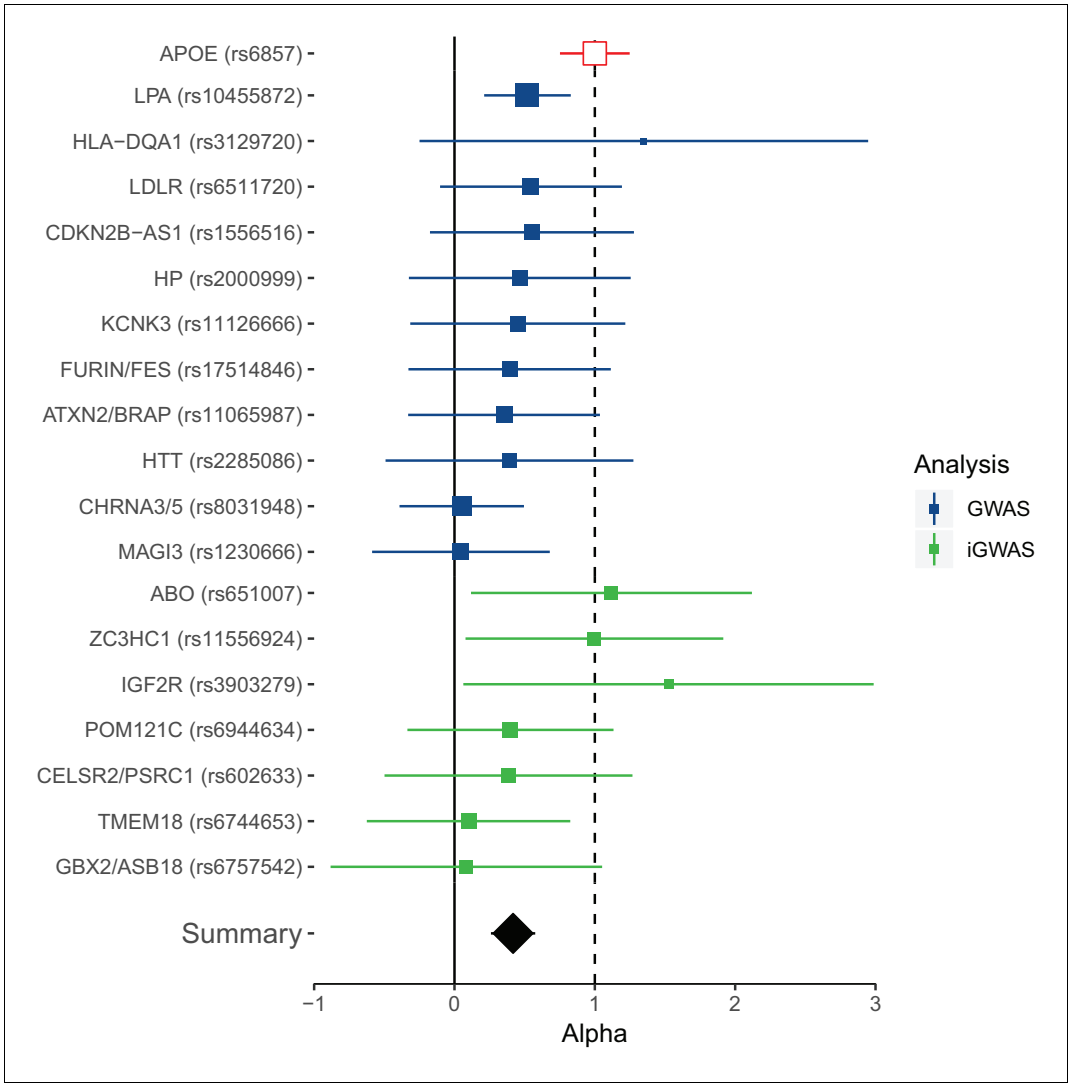


Figure 4. Collective replication of individual lifespan SNPs using GWAMAs for extreme long-livedness shows directional consistency in all cases. Forest plot of effect size ratios between genome-wide significant lifespan variants from our study and external longevity studies (***Broer et al., 2015***; ***Deelen et al., 2014***), having converted longevity effect sizes to our scale using APOE as benchmark (see Materials and methods and ***Figure 4—source data 1***). Alpha – ratio of replication to discovery effect sizes on the common scale and 95% CI (reflecting uncertainty in the numerator and denominator; P values are for one-sided test). A true (rather than estimated) ratio of 1 indicates the relationship between SNP effect on lifetime hazard and extreme longevity is the same as that of APOE, while a ratio of zero suggests no effect on longevity. A true ratio between 0 and 1 suggests a stronger effect on lifetime hazard than longevity relative to APOE. SNPs overlapping both 0 and 1 are individually underpowered. The inverse variance meta-analysis of alpha over all SNPs, excluding *APOE*, is 0.42 (95% 0.23 to 0.61; $p = 1.35 \times 10^{-5}$) for H_0 alpha = 0.

DOI: https://doi.org/10.7554/eLife.39856.016

The following source data is available for figure 4:

Source data 1. Replication of lead SNPs associating with lifespan using published longevity GWAS.

DOI: https://doi.org/10.7554/eLife.39856.017

We then looked up associations of the 81 iGWAS SNPs (1% FDR) with the risk factor GWAMAs used to inform the prior. While associations are *a priori* limited to the risk factors included in the iGWAS, the pattern of association is still of interest. We find loci show strong clustering in either blood lipids or CVD, show moderate clustering of metabolic and neurological traits, and show weak

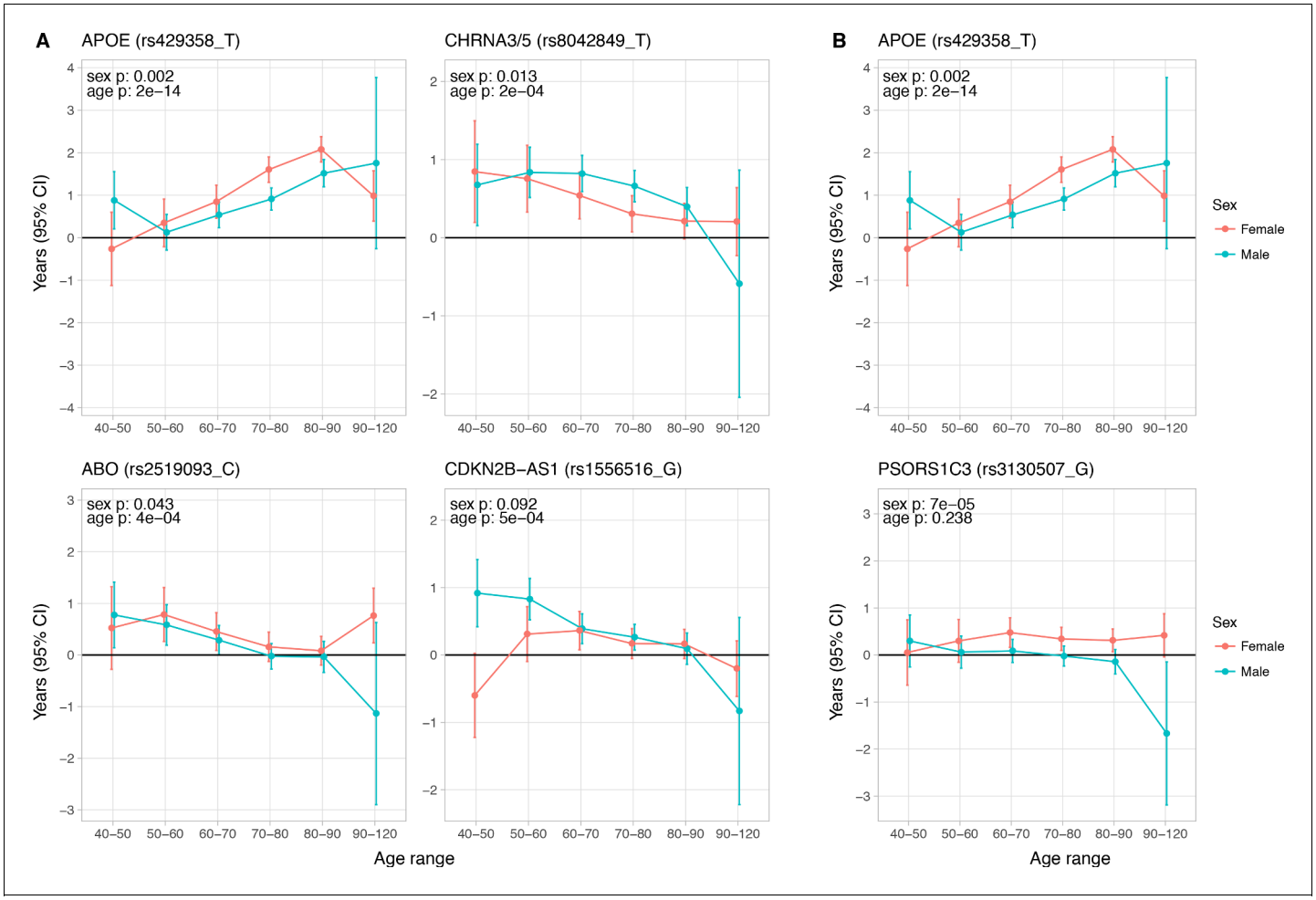


Figure 5. Age and sex specific effects on parent survival for 5 variants showing 5% FDR age- or sex-specificity of effect size from 23 lifespan-increasing variants. (**A**) Variants showing age-specific effects; (**B**) Variants showing sex-specific effects. Panel titles show the gene, cluster of genes, or cytogenetic band in close proximity to the index lifespan variant, with this variant and lifespan-increasing allele in parentheses. Beta – $\log_e$(protection ratio) for 1 copy of effect allele in self in the age band (i.e. 2 x observed due to 50% kinship). Note the varying scale of y-axis across panels. Age range: the range of ages over which beta was estimated. Sex p – nominal P value for association of effect size with sex. Age p – nominal P value for association of effect size with age.

DOI: https://doi.org/10.7554/eLife.39856.018

The following source data is available for figure 5:

Source data 1. Sex and age stratified effects on survival for 24 lifespan increasing variants.

DOI: https://doi.org/10.7554/eLife.39856.019

Source data 2. Effect sizes of sex and age moderators within fixed-effects with moderators' model of longevity alleles for 24 SNPs.

DOI: https://doi.org/10.7554/eLife.39856.020

but highly pleiotropic clustering amongst most of the remaining traits (see ***Figure 3—figure supplement 1*** for clustering of genome-wide significant SNPs).

In order to study the relative contribution of diseases to lifespan, we approached the question from the other end and looked up known associations for disease categories (CVD, type 2 diabetes, neurological disease, smoking-related traits, and cancers) in large numbers (>20 associations in each category) from the GWAS catalog (***MacArthur et al., 2017***) and used our GWAS to see if the disease loci associate with lifespan. Our measure was lifespan variance explained (LVE, years^2 [***Ljungquist et al., 1998***]) by the locus, which balances effect size against frequency, and is proportional to selection response and the GWAS test statistic and thus monotonic for risk of false positive lifespan associations. Taking each independent disease variant, we ordered them by LVE, excluding any secondary disease where the locus was pleiotropic.

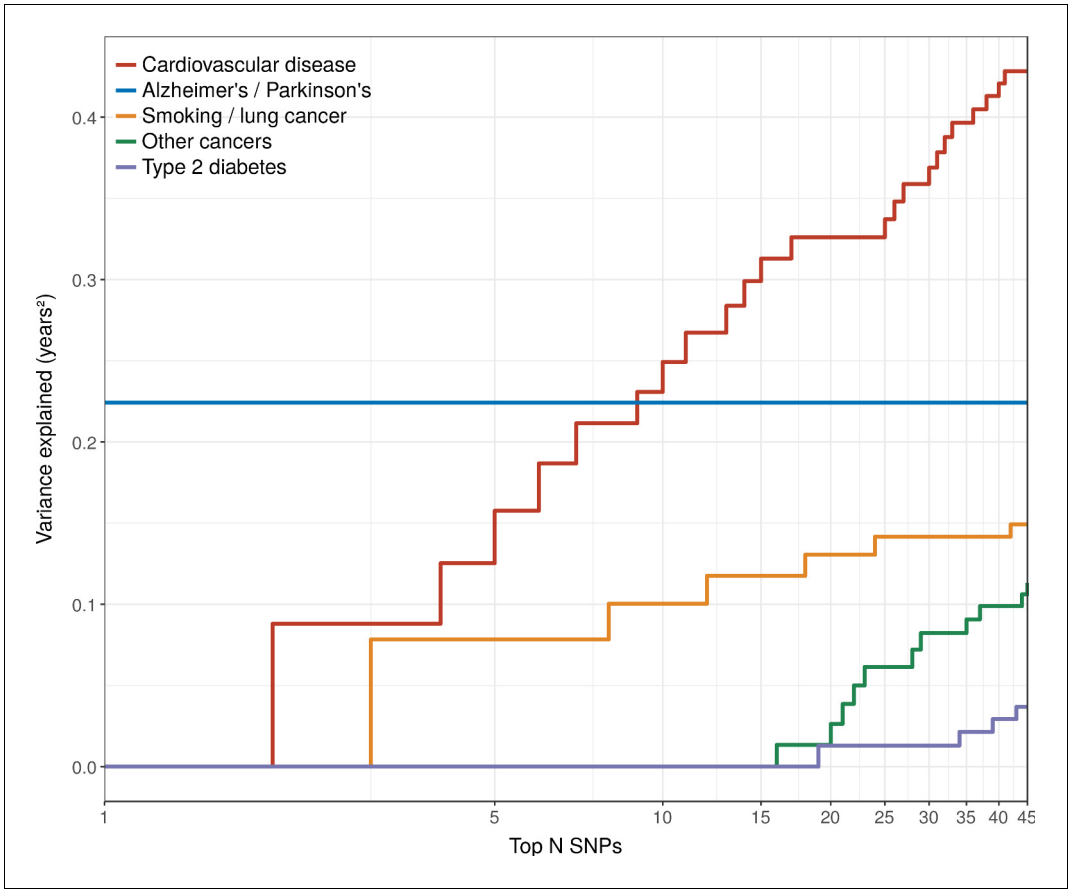


Figure 6. Disease loci explaining the most lifespan variance are protective for neurological disease, cardiovascular disease, and lung cancer. SNPs reported as genome-wide significant for disease in European population studies, ordered by their lifespan variance explained (LVE), show the cumulative effect of disease SNPs on variation in lifespan. An FDR cut-off of 1.55% is applied simultaneously across all diseases, allowing for one false positive association with lifespan among the 45 independent loci. Note the log scale on the X axis. Cardiovascular disease – SNPs associated with cardiovascular disease or myocardial infarction. Alzheimer's/Parkinson's – SNPs associated with Alzheimer's disease or Parkinson's disease. Smoking/lung cancer – SNPs associated with smoking behaviour, chronic obstructive pulmonary disease and lung adenocarcinomas. Other cancers – SNPs associated with cancers other than lung cancer (see ***Figure 7—source data 1*** for a full list). Type 2 diabetes – SNPs associated with type 2 diabetes.

DOI: https://doi.org/10.7554/eLife.39856.021

The Alzheimer's disease locus *APOE* shows the largest LVE (0.23 years^2), consistent with its most frequent discovery as a lifespan SNP in GWAS (***Joshi et al., 2016***; ***Pilling et al., 2017***; ***Deelen et al., 2014***; ***Deelen et al., 2013***). Of the 20 largest LVE SNPs, 12 and 4 associate with CVD and smoking/lung cancer, respectively, while only two associate with other cancers (near *ZW10* and *NRG1*; neither in the top 15 LVE SNPs). Cumulatively, the top 20/45 LVE SNPs explain 0.33/0.43 years^2 through CVD, 0.13/0.15 years^2 through smoking and lung cancer, and 0.03/0.11 years^2 through other cancers (***Figure 6***).

Strikingly, two of the three largest LVE loci for non-lung cancers (at or near *ATXN2/BRAP* and *CDKN2B-AS1*) show increased cancer protection associating with decreased lifespan (due to antagonistic pleiotropy with CVD), while the third (at or near *MAGI3*) also shows evidence of pleiotropy, having an association with CVD three times as strong as breast cancer, and in the same direction. In addition, 6 out of the 11 remaining cancer-protective loci which increase lifespan and pass FDR (near *ZW10*, *NRG1*, *C6orf106*, *HNF1A*, *C20orf187*, and *ABO*) also show significant associations with CVD but could not be tested for pleiotropy as we did not have data on the relative strength of association of every type of cancer against CVD, and thus (conservatively from the point of view of our

conclusion) remain counted as cancer SNPs (***Figure 7***, ***Figure 7—source data 1***). Visual inspection also reveals an interesting pattern in the SNPs that did not pass FDR correction for affecting lifespan: cardio-protective variants associate almost exclusively with increased lifespan, while cancer-protective variants appear to associate with lifespan in either direction (grey dots often appear below the x-axis for other cancers).

Together, the disease loci included in our study with significant effects on lifespan explain 0.95 years^2, or less than 1% of the phenotypic variance of lifespan of European parents in UK Biobank (123 years^2), and around 5% of the heritability.

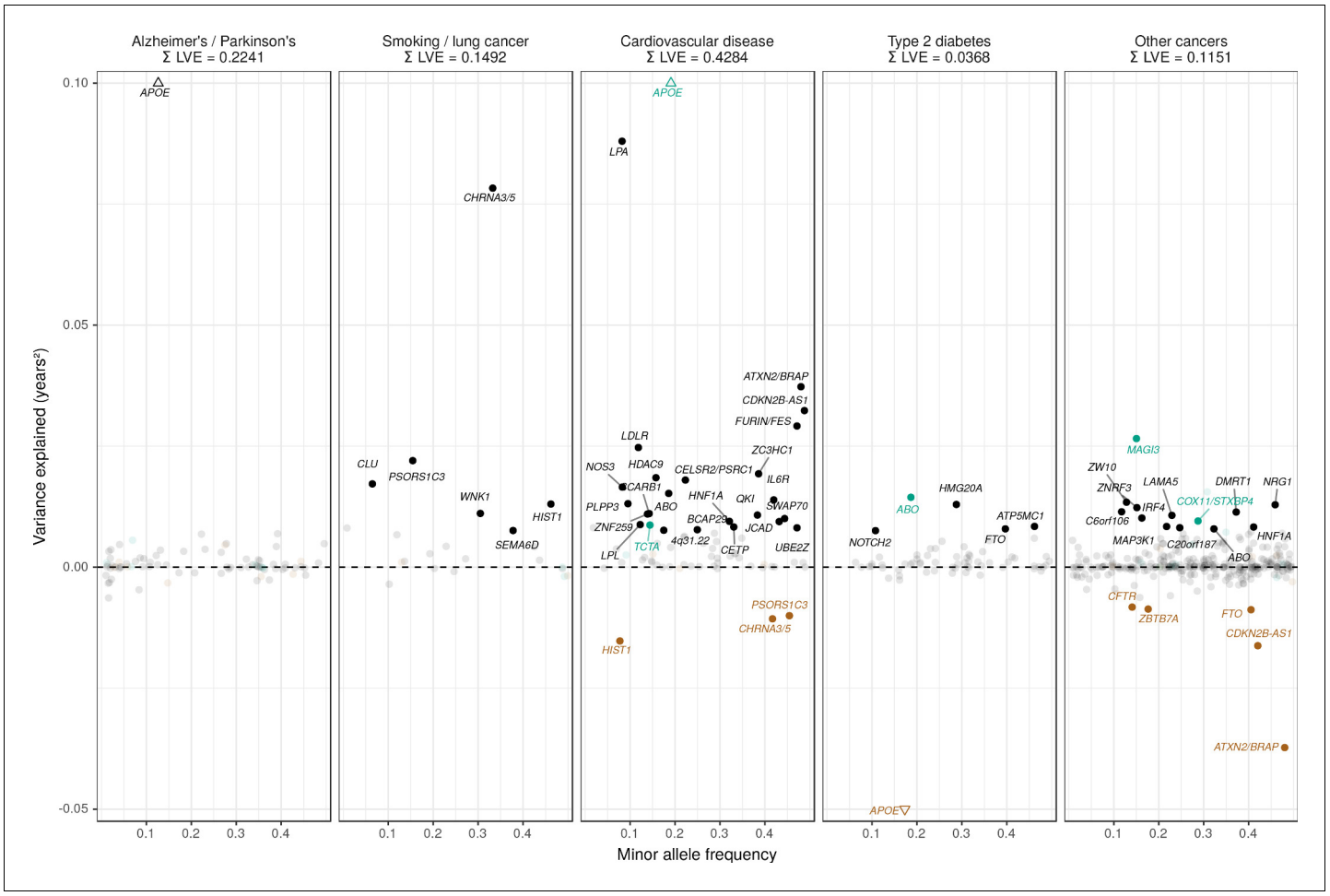


Figure 7. Lifespan variance explained by individual genome-wide significant disease SNPs within disease categories. Genome-wide significant disease SNPs from the GWAS catalog are plotted against the amount of lifespan variance explained (LVE), with disease-protective alleles signed positively when increasing lifespan and signed negatively when decreasing lifespan. SNPs with limited evidence of an effect on lifespan are greyed out: an FDR cut-off of 1.55% is applied simultaneously across all diseases, allowing for one false positive among all significant SNPs. Secondary pleiotropic SNPs (i.e. those associating more strongly with another one of the diseases, as assessed by PheWAS in UK Biobank) are coloured to indicate the main effect on increased lifespan seems to arise elsewhere. Of these, turquoise SNPs show one or more alternative disease associations in the same direction and at least twice as strong (double Z statistic – see Detailed Materials and methods) as the principal disease, while brown SNPs show one or more significant associations with alternative disease in the opposite direction that explains the negative association of the disease-protective SNP with lifespan. The variance explained by all SNPs in black is summed (∑LVE) by disease. Annotated are the gene, cluster of genes, or cytogenetic band near the lead SNPs. The Y axis has been capped to aid legibility of SNPs with smaller LVE: SNPs near APOE pass this cap and are represented by triangles. See ***Figure 7—source data 1*** for the full list of disease SNP associations.

DOI: https://doi.org/10.7554/eLife.39856.022

The following source data is available for figure 7:

Source data 1. List of genome-wide significant disease variants, their association with disease in UK Biobank and their lifespan variance explained.

DOI: https://doi.org/10.7554/eLife.39856.023

Cell type and pathway enrichment

We used stratified LD-score regression to assess whether cell type-specific regions of the genome are enriched for lifespan variants. As this method derives its power from SNP heritability, we limited the analysis to genomically British individuals in UK Biobank, which showed the lowest heterogeneity and the highest SNP heritability. At an FDR < 5%, we find enrichment in SNP heritability in five categories: two histone and two chromatin marks linked to male and female foetal brain cells, and one histone mark linked to the dorsolateral prefrontal cortex (DLPC) of the brain. Despite testing other cell types, such as heart, liver, and immune cells, no other categories are statistically significant after multiple testing correction (***Supplementary file 4***).

We also determined which biological pathways could explain the associations between our genetic variants and lifespan using three different methods, VEGAS, PASCAL, and DEPICT. VEGAS highlights 33 gene sets at an FDR < 5%, but neither PASCAL nor DEPICT (with SNP thresholds at $p < 5 \times 10^{-8}$ and $p < 1 \times 10^{-5}$) identify any gene sets passing multiple testing correction. The 33 gene sets highlighted by VEGAS are principally for blood lipid metabolism (21), with the majority involving lipoproteins (14) or homeostasis (4). Other noteworthy gene sets are neurological structure and function (5) and vesicle-mediated transport (3). Enrichment was also found for organic hydroxy compound transport, macromolecular complex remodelling, signalling events mediated by stem cell factor receptor (c-kit), and regulation of amyloid precursor protein catabolism (***Supplementary file 5***).

Finally, we performed an analysis to assess whether genes that have been shown to change their expression with age (***Peters et al., 2015***) are likely to have a causal effect on lifespan itself. Starting with a set of independent SNPs affecting gene expression (eQTLs), we created categories based on whether gene expression was age-dependent and whether the SNP was associated with lifespan in our study (at varying levels of significance). We find eQTLs associated with lifespan are 1.69 to 3.39 times more likely to have age-dependent gene expression, depending on the P value threshold used to define the set of lifespan SNPs (***Supplementary file 6***).

Out-of-sample lifespan PRS associations

We calculated polygenic risk scores (PRS) for lifespan for two subsamples of UK Biobank (Scottish individuals and a random selection of English/Welsh individuals), and one sample from the Estonian Biobank. The PRS were based on (recalculated) lifespan GWAS summary statistics that excluded these samples to ensure independence between training and testing datasets.

When including all independent markers, we find an increase of one standard deviation in PRS increases lifespan by 0.8 to 1.1 years, after doubling observed parent effect sizes to compensate for the imputation of their genotypes (see ***Table 3—source data 1*** for a comparison of performance of different PRS thresholds).

Correspondingly – again after doubling for parental imputation – we find a difference in median survival for the top and bottom deciles of PRS of 5.6/5.6 years for Scottish fathers/mothers, 6.4/4.8 for English and Welsh fathers/mothers and 3.0/2.8 for Estonian fathers/mothers. In the Estonian Biobank, where data is available for a wider range of subject ages (i.e. beyond median survival age) we find a contrast of 3.5/2.7 years in survival for male/female subjects, across the PRS tenth to first deciles (***Table 3***, ***Figure 8***).

Finally, as we did for individual variants, we looked at the age- and sex-specific nature of the PRS on parental lifespan and then tested for associations with (self-reported) age-related diseases in subjects and their kin. We find a high PRS has a larger protective effect on lifespan for mothers than fathers in UK Biobank subsamples ($p = 0.0071$), and has a larger protective effect on lifespan in younger age bands ($p = 0.0001$) (***Figure 9***), although in both cases, it should be borne in mind that women and younger people have a lower baseline hazard, so a greater improvement in hazard ratio does not necessarily mean a larger absolute protection.

We find that overall, higher PRS scores (i.e. genetically longer life) are associated with less heart disease, diabetes, hypertension, respiratory disease and lung cancer, but increased prevalence of Alzheimer's disease, Parkinson's disease, prostate cancer and breast cancer, the last three primarily in parents. We find no association between the score and prevalence of cancer in subjects. (***Figure 10***).

Table 3. Polygenic scores for lifespan associate with out-of-sample parent and subject lifespans.

A polygenic risk score (PRS) was made for each subject using GWAS results that did not include the subject sets under consideration. Subject or parent survival information (age entry, age exit, age of death, if applicable) was used to test the association between polygenic risk score and survival as (a) a continuous score and (b) by dichotomising the top and bottom decile scores. Population – Population sample of test dataset, where E and W is England and Wales; Kin – Individuals tested for association with polygenic score; N – Number of lives used for analysis; Deaths – Number of deaths; Beta – Effect size per PRS standard deviation, in $\log_e$(protection ratio), doubled in parents to reflect the expected effect in cohort subjects. SE – Standard error, doubled in parents to reflect the expected error in cohort subjects; Years – Estimated years of life gained per PRS standard deviation; P – P value of two-sided test of association; Contrast age at death – difference between the median lifespan of individuals in the top and bottom deciles of the score in year of life (observed parent contrast is again doubled to account for imputation of their genotypes).

Sample descriptives				Effect of polygenic score				Contrast age at death	
Population	Kin	N	Deaths	Beta	SE	Years	P	Men	Women
Scotland	Parents	46,936	33,196	0.107	0.011	1.07	4.2E-22	5.6	5.6
Scotland	Subjects	24,059	941	0.085	0.033	0.85	1.0E-02	-	-
E and W	Parents	58,070	39,347	0.133	0.010	1.33	7.3E-39	6.4	4.8
E and W	Subjects	29,815	760	0.098	0.037	0.98	7.1E-03	-	-
Estonia	Parents	61,728	29,660	0.099	0.012	0.99	2.5E-17	3.0	2.8
Estonia	Subjects	24,800	2894	0.087	0.019	0.87	2.6E-06	3.5	2.7
				Per standard deviation				Top vs. bottom 10%	

DOI: https://doi.org/10.7554/eLife.39856.024

The following source data is available for Table 3:

Source data 1. Polygenic survival scores in independent samples are most associated when including all markers.

A polygenic risk score was made for each subject using GWAS results that did not include the subject sets under consideration. Parent survival information (age and alive/dead status) was used to test the association between survival and several polygenic risk scores with different P value thresholds. Sample – Out-of-sample subsets of UK Biobank individuals used for PGRS association. N – Number of reported parental lifespans by sample individuals. Deaths – Number of reported parental deaths by sample individuals. Threshold – Criteria for SNPs to be included in the polygenic score. Beta – Log_e(protection ratio) per standard deviation of polygenic score, doubled to reflect the effect of the score on offspring survival. SE – standard error of the effect estimate. Mean Years – Mean years of life gained per standard deviation in PGRS. P – P value of the effect of the polygenic score on lifespan.

DOI: https://doi.org/10.7554/eLife.39856.025

Discussion

Applying the kin-cohort method in a GWAS and mortality risk factor iGWAS across UK Biobank and the LifeGen meta-analysis, we identified 11 novel genome-wide significant associations with lifespan and replicated six previously discovered loci. We also replicated long-standing longevity SNPs near *APOE*, *FOXO3*, and 5q33.3/*EBF1* – albeit with smaller effect sizes in the latter two cases – but found evidence of no association (at effect sizes originally published) with lifespan for more recently published longevity SNPs near *IL6*, *ANKRD20A9P*, *USP42*, and *TMTC2*. Conversely, all individual variants identified in our analyses showed directionally consistent effects in a meta-analysis of two European-ancestry studies of extreme longevity, and a test of association of a polygenic risk score of the variants was highly significant in the longevity dataset ($p < 1.5 \times 10^{-5}$).

Our findings validate the results of a previous Bayesian analysis performed on a subset (N = 116,279) of the present study's discovery sample (***McDaid et al., 2017***), which highlighted two loci which are now genome-wide significant in conventional GWAS in the present study's larger sample. iGWAS thus appears to be an effective method able to identify lifespan-associated variants in smaller samples than standard GWAS, albeit relying on known biology.

With the curious exception of a locus near *HTT* (the Huntington's disease gene), all lead SNPs are known to associate with autoimmune, cardiometabolic, neuropsychiatric, or smoking-related disease, and it is plausible these are the major pathways through which the variants affect lifespan. Whole-genome polygenic risk scores showed similar associations with disease, excluding late-onset disorders such as Alzheimer's and Parkinson's, where polygenic risk scores for extended lifespan increased risk (of survival to age at onset) of the disease.

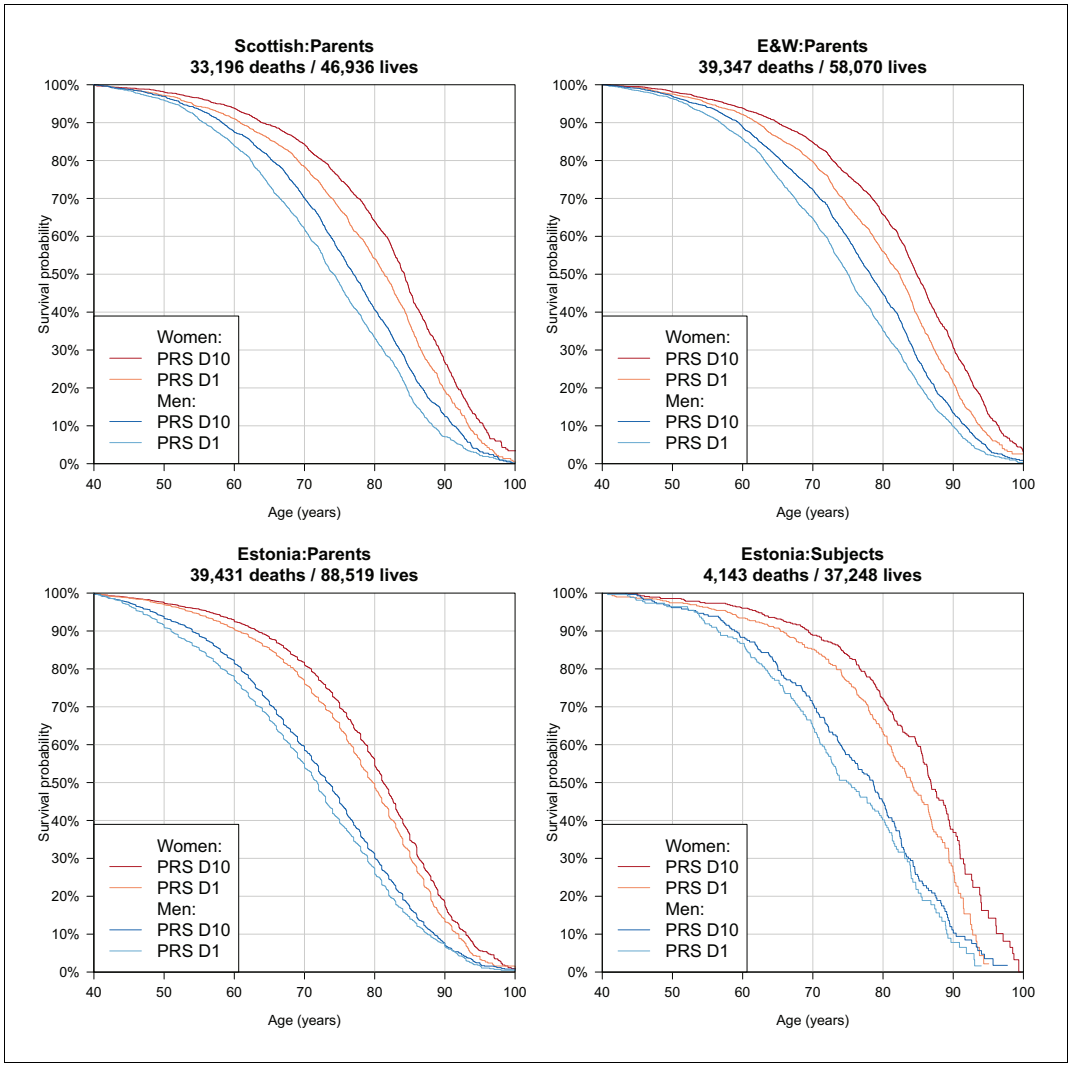


Figure 8. Survival curves for highest and lowest deciles of lifespan polygenic risk score. A polygenic risk score was made for each subject using GWAS results that did not include the subject sets under consideration. Subject or parent survival information (age entry, age exit, age of death (if applicable) was used to create Kaplan-Meier curves for the top and bottom deciles of score. In this figure (only) no adjustment has been made for the dilution of observed effects due to parent imputation from cohort subjects. Effect sizes in parent, if parent genotypes had been used, are expected to be twice that shown. E and W – England and Wales; PRS – polygenic risk score.

DOI: https://doi.org/10.7554/eLife.39856.026

The following figure supplement is available for figure 8:

Figure supplement 1. Survival Curves for highest and lowest deciles of lifespan polygenic risk score in UK Biobank subjects.

DOI: https://doi.org/10.7554/eLife.39856.027

Genetic variants affecting lifespan were enriched for pathways involving the transport, homeostasis and metabolism of lipoprotein particles, validating previous reports (***McDaid et al., 2017***). We also identified new pathways including vesicle transport, metabolism of acylglycerol and sterols, and synaptic and dendritic function. We discovered genomic regions with epigenetic marks determining cell differentiation into foetal brain and DLPC cells were enriched for genetic variants affecting lifespan. Finally, we showed that we can use our GWAS results to construct a polygenic risk score, which makes 3 to 5 year distinctions in life expectancy at birth between individuals from the score's top and bottom deciles.

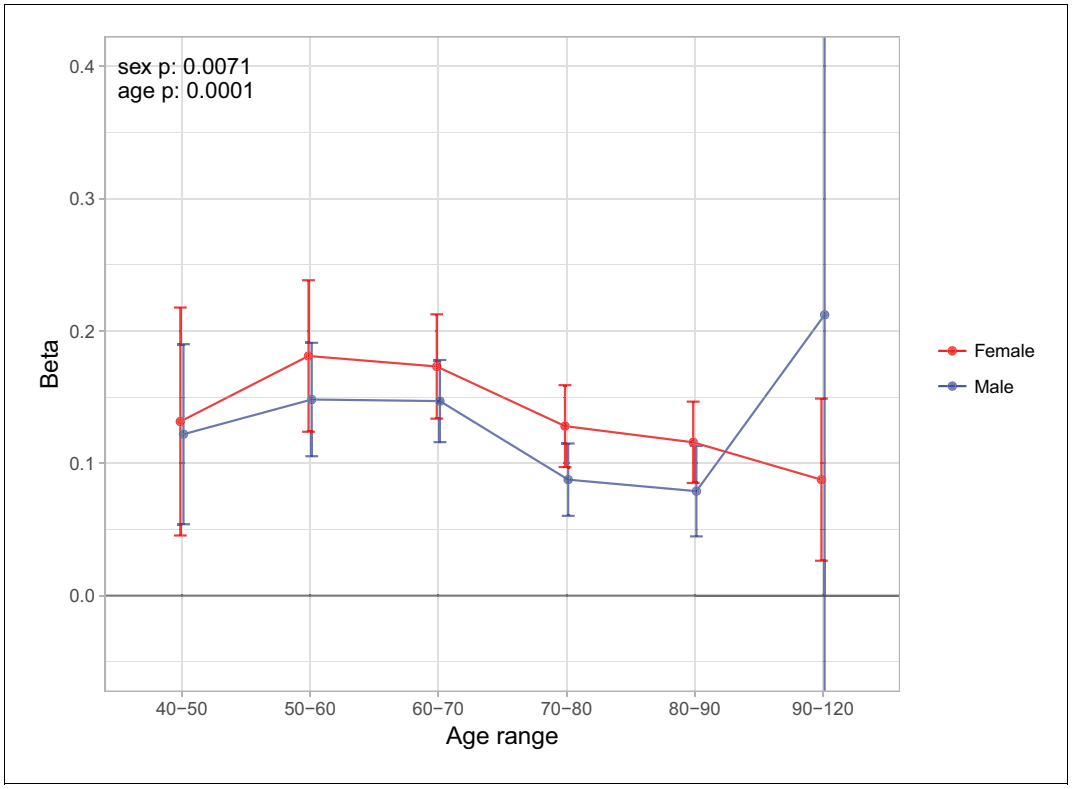


Figure 9. Sex and age specific effects of polygenic survival score (PRS) on parental lifespan in UK Biobank. The effect of out-of-sample PRS on parental lifespan stratified by sex and age was estimated for Scottish and English/Welsh subsamples individually (see ***Figure 9—figure supplement 1***) and subsequently meta-analysed. The estimate for the PRS on father lifespan in the highest age range has very wide confidence intervals (CI) due to the limited number of fathers surviving past 90 years of age. The beta 95% CI for this estimate is –0.15 to 0.57. Beta – $\log_e$(protection ratio) for one standard deviation of PRS for increased lifespan in self in the age band (i.e. 2 x observed due to 50% kinship), bounds shown are 95% CI; Age range – the range of ages over which beta was estimated; sex p – P value for association of effect size with sex; age p – P value for association of effect size with age.

DOI: https://doi.org/10.7554/eLife.39856.028

The following source data and figure supplement are available for figure 9:

Source data 1. Sex and age-stratified association of polygenic score on lifespan.

DOI: https://doi.org/10.7554/eLife.39856.029

Figure supplement 1. Sex and age specific effects of polygenic survival score (PRS) on parental lifespan of Scottish and English/Welsh subsamples of UK Biobank.

DOI: https://doi.org/10.7554/eLife.39856.030

Despite studying over 1 million lives, our standard GWAS only identified 12 variants influencing lifespan at genome-wide significance. This contrasts with height (another highly polygenic trait) where a study of around 250,000 individuals by ***Wood et al., 2014***. found 423 loci. This difference can partly be explained by the much lower heritability of lifespan (0.12; ***Kaplanis et al., 2018***) (cf. 0.8 for height [***Wood et al., 2014***]), consistent with evolution having a stronger influence on the total heritability of traits more closely related to fitness and limiting effect sizes. In addition, the use of indirect genotypes (the kin-cohort method) reduces the effective sample size to 1/4 for the parent-offspring design.

When considering these limitations, we calculate our study was equal in power to a height study of only around 23,224 individuals, were lifespan to have a similar genetic architecture to height (see Materials and methods). Under this assumption, we would require a sample size of around 10 million parents (or equivalently 445,000 nonagenarian cases, with even more controls) to detect a similar number of loci as Wood *et al*. At the same time, our inability to replicate several previous borderline

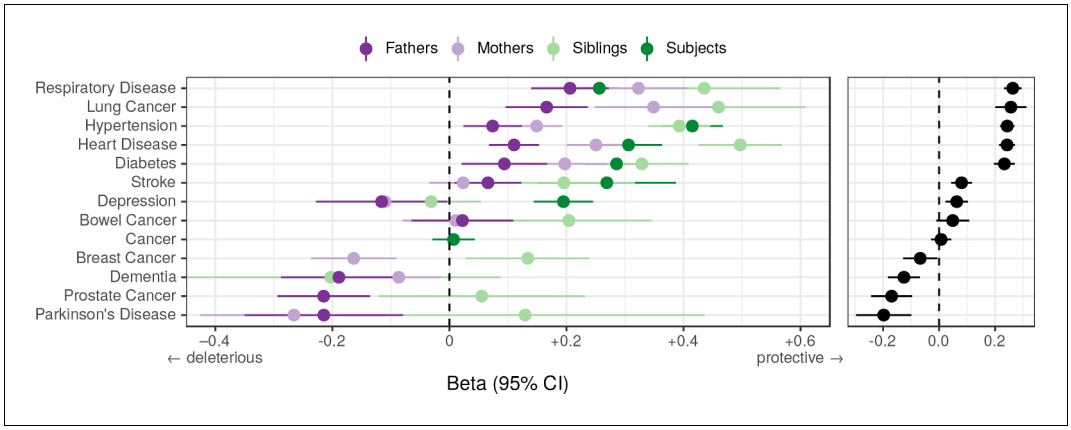


Figure 10. Associations between polygenic lifespan score and diseases of UK Biobank subjects and their kin. Logistic regression was performed on standardised polygenic survival score (all variants) and 21 disease traits reported by 24,059 Scottish and 29,815 English/Welsh out-of-sample individuals about themselves and their kin. For grouping of UK Biobank disease codes, see ***Figure 10—source data 1***. Displayed here are inverse-variance meta-analysed estimates of the diseases for which multiple sources of data were available (i.e. parents and/or siblings; see ***Figure 10—figure supplement 1*** for all associations). 'Cancer' is only in subjects, whilst the specific subtypes are analysed for kin. The left panel shows disease estimates for each kin separately; the right panel shows the combined estimate, with standard errors adjusted for correlation between family members. Diseases have been ordered by magnitude of effect size (combined estimate). Beta – log odds reduction ratio of disease per standard deviation of polygenic survival score, where a negative beta indicates a deleterious effect of score on disease prevalence (lifetime so far), and positive beta indicates a protective effect on disease. Effect sizes for first degree relatives have been doubled. Cancer – Binary cancer phenotype (any cancer, yes/no).

DOI: https://doi.org/10.7554/eLife.39856.031

The following source data and figure supplement are available for figure 10:

Source data 1. Grouping of UK Biobank disease codes into diseases and major disease categories.

DOI: https://doi.org/10.7554/eLife.39856.032

Source data 2. Associations of polygenic score with diseases in UK Biobank.

DOI: https://doi.org/10.7554/eLife.39856.033

Figure supplement 1. Associations between polygenic survival score and diseases of individuals and their kin from Scottish and English/Welsh subsamples of UK Biobank.

DOI: https://doi.org/10.7554/eLife.39856.034

significant longevity and lifespan findings suggests research into survival in general requires substantial increases in power to robustly identify loci.

Meta-analysis of mothers and fathers, permitting common or sex-specific effect sizes, of course, doubled effective sample size, with slight attenuation to reflect the observed correlation (~10%) between father and mother traits (consistent with previous studies [***Kaplanis et al., 2018***]). This correlation indicates the presence of assortative mating on traits which correlate with lifespan (as lifespan itself is of course not observed until later), or post-pairing environmental convergence. We note that in principle, assortative mating could lead to allelic correlations at causal loci for the contributing traits, causing departures from Hardy-Weinberg equilibrium, and increasing the genotypic variance and thus power to detect association. However, in practice, at least for lifespan, the effects are too small for the effect to be material.

The association of lifespan variants with well-known, life-shortening diseases (cardiovascular, autoimmune, smoking-related diseases and lung cancer; ***Mathers et al., 2018***) is not surprising, but the paucity of associations with other forms of cancer – without pleiotropic effects on CVD – is. This paucity suggests cancer deaths may often be due to (perhaps many) rarer variants or environmental exposures, although effect sizes might simply be slightly below our cut-off threshold to detect. Disappointingly, the variants and pathways we identified do not appear to underpin a generalised form of ageing independent of disease.

Our finding that lifespan genetics are enriched for lipid metabolism genes is in line with expectations, given lipid metabolites – especially cholesterol metabolites – have well-established effects on

atherosclerosis, type 2 diabetes, Alzheimer's disease, osteoporosis, and age-related cancers (**Zarrouk et al., 2014**). **Pilling et al. (2017)** implicated nicotinic acetylcholine receptor pathways in human lifespan, which we detected at nominal significance ($p = 2 \times 10^{-4}$) but not at 5% FDR correction (q = 0.0556). Instead we highlighted more general synapse and dendrite pathways and identified foetal brain and DLPC cells as important in ageing. The DLPC is involved in smoking addiction (**Hayashi et al., 2013**), dietary self-control (**Lowe et al., 2014**), and is susceptible to neurodegeneration (**Morrison and Baxter, 2012**), which could explain why genetic variation for lifespan is specifically enriched in these cells, mediated through smoking-related, cardiometabolic, and neuropsychiatric disease.

Much work has been done implicating *FOXO3* as an ageing gene in model organisms (**Kenyon et al., 1993**; **Hwangbo et al., 2004**), however we found the association in humans at that locus may be driven by expression of *SESN1* (admittedly a finding restricted to peripheral blood tissue). *SESN1* is a gene connected to the *FOXO3* promoter via chromatin interactions and is involved in the response to reactive oxygen species and mTORC1 inhibition (**Donlon et al., 2017**). While fine-mapping studies have specifically found genetic variation within the locus causes differential expression of *FOXO3* itself (**Flachsbart et al., 2017**; **Grossi et al., 2018**), this does not rule out the effect of co-expression of *SESN1*. More powered tissue-specific expression data and experimental work on *SESN1* vs. *FOXO3* could elucidate the causal mechanism. For now, results from model organisms seem to leave the preponderance of evidence for *FOXO3*.

Our results suggest disease-associated lifespan variants reduce the chances of extreme long-livedness, but remain agnostic as to the more interesting two-part question: are there longevity variants that have little effect on lifespan in the normal range (**Sebastiani et al., 2016**), and if so, do they control underlying ageing processes? We note, the genetic overlap between lifespan and extreme long-livedness is high (0.73), but not complete (**McDaid et al., 2017**). Regardless of this, only a small part of the heritability of both lifespan and longevity has thus far been explained by GWAS. It thus remains plausible that an enlarged long-livedness or lifespan study will find variants controlling the rate of ageing and associated pathways. Curiously, we find little evidence of SNPs of large deleterious effect on lifespan acting with antagonistic pleiotropy on other fitness and developmental component traits, despite long-standing theoretical suggestions to the contrary (**Williams, 1957**). However, we did not examine mortality before the age of 40, or mortality of individuals without offspring (by definition as we were examining parental lifespans), which may exhibit this phenomenon. For the time being, our findings that the improved polygenic risk score for lifespan was associated with an increased prevalence of Alzheimer's disease, Parkinson's disease, and prostate and breast cancer, means we appear to be predominantly measuring a propensity for longer life through avoidance of early disease-induced mortality, rather than healthy ageing or fertility costs.

Whilst it has previously been shown that transcriptomic age calculated based on age-related genes is meaningful in the sense that its deviation from the chronological age is associated with biological features linked to ageing (**Peters et al., 2015**), the role of these genes in ageing was unclear. A gene might change expression with age because (i) it is a biological clock (higher expression tracking biological ageing, but not influencing ageing or disease); (ii) it is a response to the consequences of ageing (e.g. a protective response to CVD); (iii) it is an indicator of selection bias: if low expression is life-shortening, older people with low expression tend to be eliminated from the study, hence the average expression level of older age groups is higher. However, our results now show that the differential expression of many of the age-related genes discovered by **Peters et al., 2015** are not only biomarkers of ageing, but are also enriched for direct effects on lifespan.

There is increasing interest in polygenic risk scores, and their potential clinical utility for some diseases appears to be similar to some Mendelian mutations (albeit such monogenic tests are usually only applied in the context of family history; **Khera et al., 2018**). At first sight, the magnitude of the distinctions in our genetic lifespan score (5 years of life between top and bottom deciles, for both the parent and subject generations) are quite small compared with variability in individual lifespans. However, these distinctions are potentially material at a group level, for example, actuarially. The implied distinction in price (14%; Methods) is greater than some recently reported annuity profit margins (8.9%) (**Legal General Group PLC, 2017**). In our view, the legal and ethical frameworks (at least in the UK [**Association of British Insurers and UK Government, 2014**]) are presently underdeveloped for genome-wide scores, whether for disease or lifespan and this needs to be urgently

addressed. At the same time, although material in isolation, our lifespan associations may only have practical utility in many applications if they provide additional information than that provided by conventional clinical risk measures (e.g. the Framingham score [*D'Agostino et al., 2008*]). Such an assessment has been beyond the scope of this work, in part as such risk measures are not readily available for the parents (rather than subjects) studied.

One limitation of our study was the power reduction caused by the exclusion of relatives in our study, rather than linear mixed modelling (LMM) with a term for kinship as measured by the genomic relationship matrix (GRM) (*Pilling et al., 2017*; *Loh et al., 2015*). However, as the correct adjustment is not derivable under the kin-cohort method, we felt this was the best approach. To see that the normal adjustment is not correct, consider two siblings. The phenotypes under study are of course identical (as the parents are the same), but the expected correlation under the mixed model would only be 50% of the heritability. Simply excluding siblings, however, is not sufficient. For example, consider two offspring subjects who are first cousins descended from two full brothers. The GRM entry in this situation is 12.5% whilst the appropriate relatedness factor for the father trait is 50% and the mother 0%. Exclusion of relatives thus appears the most straightforward solution, although if a pedigree were available, not just a GRM, accurate LMM might have been feasible.

The analysis of parent lifespans has enabled us to probe mortality for a generation whose lives are mostly complete and attain increased power in a survival GWAS. However, changes in the environment (and thus the relative importance of each genetic susceptibility, for example following the smoking ban) inevitably mean we have less certainty about associations with prospective lifespans for the present generation of middle-aged people, or a different population (with perhaps different relative importance of disease or even overall heritability of lifespan). The 21% reduction in the effect size of the association between our PRS for the UK offspring generation supports this idea, although the estimated contrast in hazard ratios across the deciles was not reduced, which may be a statistical artefact or due to the different periods of life probed. The lower explanatory power of the PRS in Estonia may reflect the differing alleles and LD patterns between the UK training data and the Estonian test data, but also the different environments, in particular the sources of mortality in that country in the Soviet, and early post-Soviet era.

In conclusion, recent genomic susceptibility to death in the normal age range seems rooted in modern diseases: Alzheimer's, CVD and lung cancer; in turn arising from our modern – long-lived, obesogenic and tobacco-laden – environment, however the keys to the distinct traits of ageing and extreme longevity remain elusive. At the same time, genomic information alone can now make material distinctions at a group level in variations in expected length of life, although the limited individual accuracy of these distinctions is far from reaching genetic determinism of that most (self-) interesting of traits – your lifespan.

Materials and methods

Summary

GWAS

For genetically British ancestry (as identified by UK Biobank using genomic PCA) and each self-reported European ethnicity in UK Biobank (including self-declared British but not genetically British ancestry), independent association analyses were performed between unrelated subjects' genotypes (MAF > 0.005; HRC imputed SNPs only; ~9 million markers) and parent survival using age and alive/dead status in residualised Cox models, as described in *Joshi et al. (2017)*. To account for parental genotype imputation, effect sizes were doubled, yielding log hazard ratios for the allele in carriers themselves. These values were negated to obtain a measure of log protection ratio, where higher values indicate longer life. While methods exist to account for related individuals using linear mixed models, such as BOLT-LMM (*Loh et al., 2015*), these are not accurate when trying to account for relatedness between parents (See Detailed Materials and methods).

Mother and father survival information was combined in two separate ways, essentially assuming the effects were the same in men and women, or allowing for sex-specific effect sizes (SSE), with appropriate allowance for the covariance amongst the traits. For the first analysis we summed parental survival residuals; for the second analysis we used MANOVA, implemented in MultiABEL (*Shen et al., 2015*).

For LifeGen, where individual-level data was not available, parent survival summary statistics were combined using conventional fixed-effects meta-analysis, adjusted to account for the correlation between survival traits (estimated from summary-level data). For SSE, the same procedure was followed as for the UK Biobank samples, with correlation between traits again estimated from summary-level data. The GWAS statistics showed acceptable inflation, as measured by their LD-score regression intercept (<1.06, ***Table 1—source data 2***).

Candidate SNP replication

Effect sizes from longevity studies were converted to our scale using an empirical conversion factor, based on the observed relationships between longevity and hazard ratio at the most significant variant at or near *APOE*, observed in the candidate SNPs study and our data (***Joshi et al., 2017***). These studies were then meta-analysed using inverse variance weighting and standard errors were inflated to account for sample overlap (see Detailed Methods)

Estimates reported in ***Pilling et al. (2017)*** were based on rank-normalized Martingale residuals, unadjusted for the proportion dead, which – for individual parents – could be converted to our scale by multiplying by sqrt(c)/c, where c is the proportion dead in the original study (see Detailed Methods for derivation). Combined parent estimates were converted using the same method as the one used for longevity studies.

The deletion reported by ***Ben-Avraham et al. (2017)*** is perfectly tagged by a SNP that we used to assess replication. Assuming a recessive effect and parental imputation, we derived the expected additive effect to be $\hat{\beta}_C = \hat{\beta}_{CC}\frac{q^2}{q^2+2pq}$, where $\hat{\beta}_C$ is the effect we expect to observe under our additive model, $\hat{\beta}_{CC}$ is the homozygous effect reported in the original study, q is the C allele frequency, and p is $1-q$ (see Detailed Materials and methods for derivation).

iGWAS

58 GWAS on mortality risk factors were used to create Bayesian priors for the SNP effects observed in the CES study, as described in ***McDaid et al. (2017)***. Mendelian randomisation was used to estimate causal effects of independent risk factors on lifespan, and these estimates were combined with the risk factor GWAS to calculate priors for each SNP. Priors were multiplied with observed Z statistics and used to generate Bayes factors. Observed Z statistics were then permuted, leading to 7.2 billion null Bayes factors (using the same priors), which were used to assess significance.

Sex and age stratified analysis

Cox survival models, adjusting for the same covariates as the standard GWAS, were used to test SNP dosage against survival of UK Biobank genomically British fathers and mothers, separately. The analysis was split into age bands, where any parent who died at an age younger than the age band was excluded and any parent who died beyond the age band was treated as alive. Using the R package 'metafor', moderator effects of sex and age on hazard ratio could be estimated while taking into account the estimate uncertainty (see Detailed Materials and methods for formula).

Causal genes and methylation sites

SMR-HEIDI (***Zhu et al., 2016***) tests were performed on CES statistics to implicate causal genes and methylation sites. Summary-level data from two studies on gene expression in blood (***Westra et al., 2013***; ***Lloyd-Jones et al., 2017***) and data on gene expression in 48 tissues from the GTEx consortium (***Battle et al., 2017***) were tested to find causal links between gene expression and lifespan. Similarly, data from a genome-wide methylation study (***McRae et al., 2017***) was used to find causal links between CpG sites and lifespan. All results from the SMR test passing a 5% FDR threshold where the HEIDI test p>0.05 were reported.

Conditional analysis

SOJO (***Ning et al., 2017***) was used to fine-map the genetic signals in 1 Mb regions around lead SNPs reaching genome-wide significance and candidate SNPs reaching nominal significance in our study. The analysis was based on CES statistics from UK Biobank genomically British individuals, using the LifeGen meta-analysis results to optimise the LASSO regression tuning parameters. For

each parameter, a polygenic score was built and the proportion of predictable variance from the regional polygenic score in the validation sample was calculated.

Disease association analysis

The GWAS catalog (**MacArthur et al., 2017**) and PhenoScanner (**Staley et al., 2016**) were checked for known genome-wide significant associations with our GWAS hits and proxies (r2 >0.6) in European samples. Associations discovered in UK Biobank by **Churchhouse and Neale, 2018** were omitted from the PhenoScanner database as the findings have not been replicated, and the large sample overlap with our own study could result in false positive associations, due to phenotypic correlations between morbidity and mortality. Triallelic SNPs and associations without effect sizes were excluded before near-identical traits were grouped together, discarding all but the strongest association and keeping the shortest trait name. For example, 'Lung cancer', 'Familial lung cancer', and 'Small cell lung cancer' were grouped and renamed to 'Lung cancer'. The remaining associations were classified into disease categories based on keywords and subsequent manual curation.

Lifespan variance explained by disease SNPs

The GWAS catalog (**MacArthur et al., 2017**) was checked for disease associations discovered in European ancestry studies, which were grouped into broad disease categories based on keywords and manual curation (see **Figure 7—source data 1** and Detailed Methods). Associations were pruned by distance (500 kb) and LD (r^2 <0.1), keeping the SNP most strongly associated with lifespan in the CES GWAS. Where possible this SNP was tested against diseases in UK Biobank subjects and their family to test for pleiotropy (see Detailed Matrials and methods). Significance of associations with lifespan was determined by setting an FDR threshold that allowed for one false positive among all independent SNPs tested ($q \leq 0.022$). Lifespan variance explained (LVE) was calculated as $2pqa^2$, where p and q are the frequencies of the effect and reference alleles in our lifespan GWAS, and a is the SNP effect size in years of life (**Falconer et al., 1996**).

Cell type enrichment

Stratified LD-score regression (**Finucane et al., 2015**) was used to test for cell type-specific enrichment in lifespan heritability. As the power of this method depends on SNP heritability, standard LD-score regression (**Finucane et al., 2015**) was first used to check which of our samples (UK Biobank, LifeGen, or the combined cohort) had the highest SNP heritability. Lifespan summary statistics from UK Biobank genomically British individuals were then analysed using the procedure described in **Finucane et al., 2015**, and P values were adjusted for multiple testing using the Benjamin-Hochberg procedure.

Pathway enrichment

VEGAS2 v2.01.17 (**Mishra and Macgregor, 2015**) was used to calculate gene scores using SNPs genotyped in UK Biobank, based on summary statistics from the full CES cohort and the default software settings. VEGAS2Pathway was then used to check for pathway enrichment using those gene scores and the default list of gene sets (**Mishra and MacGregor, 2017**).

DEPICT (**Pers et al., 2015**) was also used to map genes to lifespan loci and check for pathway enrichment in the combined cohort CES GWAS. Default analysis was run for regions with genome-wide significant (p < 5e-8) variants in the first analysis, and genome-wide suggestive (p < 1e-5) variants in the second analysis, excluding the MHC in both cases.

PASCAL (**Lamparter et al., 2016**) was used with the same summary statistics and gene sets as DEPICT, except the gene probabilities within the sets were dichotomized (Z > 3) as described in **Marouli et al., 2017**.

For each software independently, pathway enrichment was adjusted for multiple testing using the Benjamin-Hochberg procedure.

Age-related eQTL enrichment

Combined cohort CES lifespan statistics were matched to eQTLs associated with the expression of at least one gene ($p < 10^{-3}$) in a dataset from the eQTLGen Consortium (31,684 individuals) (**Võsa et al., 2018**). Data on age-related expression (**Peters et al., 2015**) allowed eQTLs to be

divided into four categories based on association with age and/or lifespan. Fisher's exact test was used check if age-related eQTLs were enriched for associations with lifespan.

Polygenic score analysis

Polygenic risk scores (PRS) for lifespan were calculated for two subsamples of UK Biobank (24,059 Scottish individuals and a random 29,815 English/Welsh individuals), and 36,499 individuals from the Estonian Biobank, using combined cohort CES lifespan summary statistics that excluded these samples. PRSice 2.0.14.beta (**Euesden et al., 2015**) was used to construct the scores from genotyped SNPs in UK Biobank and imputed data from the Estonian Biobank, pruned by LD (r^2 = 0.1) and distance (250 kb). Polygenic scores were Z standardised.

Cox proportional hazard models were used to fit parental survival against polygenic score, adjusted for subject sex; assessment centre; genotyping batch and array; and 10 principal components. Parental hazard ratios were converted into subject years of life as described in the GWAS method section.

Logistic regression models were used to fit polygenic score against the same self-reported UK Biobank disease categories used for individual SNPs. Effect estimates of first-degree relatives were doubled to account for imputation of genotypes and then meta-analysed using inverse variance weighting, adjusting for trait correlations between family members.

Postulation of equivalent sample size in height GWAS

The use of parent imputation, low trait heritability, and incomplete proportion dead all reduce the power to detect effect sizes. The equivalent sample size in a hypothetical, completely heritable trait with otherwise identical genetic architecture would be the original sample size, diluted (i.e. multiplied) by the heritability (0.122) (**Kaplanis et al., 2018**), the r^2 of offspring genotype on parent genotype (0.250) and the proportion dead (0.602). This gives an equivalent sample size of 18,579 from the 1,012,240 parent lifespans. We then calculated sample size for height that would have the same properties, accounting for the heritability of height (0.8) (**Wood et al., 2014**): 23,224 (i.e. 18,579/0.8). We next calculated the P value that would have been reported by Wood *et al*'s 250,000 sample size height GWAMA (**Wood et al., 2014**) for a SNP that was just significant in a hypothetical 23,224 sample height GWAMA: $p < 1.8 \times 10^{-72}$. Six distinct loci passed this significance threshold in Wood *et al*'s results.

With 17,893 nonagenarians, **Deelen et al. (2014)** attained a P value of 2.33×10^{-26} at rs4420638. With 1.012 m parents we attained a P value of 1.75×10^{-64}. Other things being equal a nonagenarian sample size of 44,500 thus appears to be equally powered to one million parents.

Sensitivity of annuity prices to age

Market annuity rates for life annuities in January 2018 written to 55, 60, 65, and 70 year olds were obtained from the sharing pensions website http://www.sharingpensions.com/annuity_rates.htm (accessed 22 January 2018) and were £4158, £4680, £5476, £6075, £7105 respectively per year for a £100,000 purchase price. The arithmetic average increase from one quinquennial age to the next is 14 percent.

Data availability

Individual phenotypic and genetic data is available from UK Biobank upon application: https://www.ukbiobank.ac.uk/register-apply/. Phenotypes used in this work include subject age, sex, ethnicity, relatedness, genotyping batch, array, and principal components, as well as parental age and alive/dead status. Also included are self-reported diseases of subjects and their families. A full description of our application can be found at https://www.ukbiobank.ac.uk/2015/02/dr-james-wilson-university-of-edinburgh-centre-for-population-sciences/. The results that support our findings, in particular, the GWAS summary statistics we generated for >1 million parental lifespans in this study are available at http://dx.doi.org/10.7488/ds/2463. Gene expression data used in the age-related gene analysis is being made available by the eQTLGen Consortium, see http://www.eqtlgen.org/ and **Võsa et al., 2018**. Single tissue gene expression data used in the SMR-Heidi analysis can be found on the GTEx website https://gtexportal.org/home/datasets, under GTEx_Analysis_v7_eQTL.tar.gz.

Details

Data sources

Our UK Biobank dataset consisted of 409,700 British individuals (determined by genomic PCA) and 48,656 European individuals of self-reported (but not genetic) British, Irish, and other White European descent. Details on genotyping marker and sample QC are described in ***Bycroft et al., 2017***. Subjects completed a questionnaire which included questions on adoption status, parental age, and parental deaths. For our analysis, we excluded individuals who were adopted or otherwise unclear about their adoption status (N = 7,279), individuals who did not report their parental ages (N = 2,995), and individuals both of whose parents died before the age of 40 and which were therefore more likely due to accident or injury (N = 4,472). We further excluded one of each pair from related individuals (third degree or closer; N = 88,354) from every relative pair reported by UK Biobank, leaving 443,610 individuals for the final analysis. Although exclusion of relatives reduces sample size, we were concerned that linear mixed modelling to account for relatedness might not be fully appropriate under the kin-cohort model. Consider the parental phenotypic correlation for two full sibling subjects ($r^2 = 1$) or the maternal genetic covariance amongst two subjects who are the offspring of two brothers ($r^2 = 0$): the heritability/GRM implied covariance is incorrect for both cases (although in the sibling case, it may be correct on average). Individuals passing QC reported a total of 339,732 paternal and 351,889 maternal lifespans, ranging from 40 to 107 years of life, that is 691,621 lives in total (***Table 1—source data 1***).

Our second dataset was the publicly available summary statistics from LifeGen, a consortium of 26 population cohorts investigating genomic effects on parental lifespans (***Joshi et al., 2017***). LifeGen had included results from UK Biobank, but the UK Biobank GWAS data were removed here, giving GWAS summary statistics for 160,461 father and 160,158 mother lifespans in the form of log hazard ratios. Combined, our datasets had 1,012,240 lives.

UK Biobank Genome-Wide association study

In each separate UK Biobank ethnicity, we carried out association analysis between genotype (MAF > 0.005; HRC imputed SNPs only; ~9 million markers) and parent age and alive/dead status, effectively analysing the effect of genotype in offspring on parent survival, given survival to at least age 40, using Cox Proportional Hazards models. The following model was assumed to hold:

$$h(x) = h_0(x)e^{\beta X + \gamma 1 Z 1 + \ldots \gamma k Z k} \quad (1)$$

Where x is (parent) age, h_0 the baseline hazard and X the offspring genotype (coded 0,1,2), beta the $\log_e$(hazard ratio) associated with X and Z_1-Z_k the covariates, with corresponding effect sizes y_1-y_k. The covariates were genotyping batch and array, the first forty principal components of relatedness, as provided by UK Biobank, and subject sex (but not age, as we were analysing parent age).

To facilitate practical runtimes, the Martingale residuals of the Cox model were calculated for father and mothers separately and divided by the proportion dead to give estimates of the hazard ratio (***Therneau et al., 1990***) giving a residual trait suitable for GWAS (for more details of the residual method see ***Joshi et al., 2017***). Effect sizes observed under this model, for a SNP in offspring, are half that of the actual effect size in the parent carrying the variant (***Joshi et al., 2017***). Reported effect sizes (and their SE) have therefore been doubled to give the effect sizes in carriers themselves, giving an estimate of the log hazard ratios (or often, with sign reversed, log protective ratios). These estimates are suitable for meta-analysis and allow direct comparison with the log hazard ratios from LifeGen.

Analysis of association between genotype and survival across both parents was made under two contrasting assumptions and associated models, which had to adjust for the covariance amongst parent traits, preventing conventional unadjusted inverse-variance meta-analysis. Firstly, we assumed that the hazard ratio was the same for both sexes, that is a common effect size across sexes (CES). If there were no correlation amongst parents' traits, this could have been done by straightforward inverse variance meta-analysis of the single parent results. However, to account for the covariance amongst father and mother lifespans, we calculated a total parent residual, the sum of individual parent residuals, for each subject (i.e. offspring). Under the common effect assumption, the combined trait's effect size is twice that in the single parent, and the variance of the combined trait, automatically and appropriately reflects the parents' covariance, amongst the two parents, giving a

residual trait suitable for GWAS, with an effect size equal to that in a carrier of the variant, and correct standard error. Secondly, we assumed that, there might be sex-specific effect sizes (SSE) in fathers and mothers. Under the SSE assumption, individual parental GWAS were carried out, and the summary statistics results were meta-analysed using MANOVA, accounting for the correlation amongst the parent traits and the sample overlap (broadly complete), but agnostic as to whether the effect size was similar or different in each parent, giving a P value against the null hypothesis that both effect sizes are zero, but, naturally, no estimate of a single common beta. This procedure was carried out using the R package MultiABEL and summary-level data (***Shen et al., 2015***). The procedure requires an estimate of the correlation amongst the traits (in this case parent residuals), which was measured directly (r = 0.1). The procedure automatically estimates the variance of the traits from summary level data (Mother residuals σ^2 = 6.74; Father residuals σ^2 = 5.25)

For the LifeGen results, the SSE procedure to combine results was identical to UK Biobank (Mother residuals σ^2 = 14.12; Father residuals σ^2 = 18.75), except the trait correlation was derived from summary level data instead of measured directly (r = 0.1). This was done by taking the correlations in effect estimates from independent SNPs from the summary statistics of the individual parents, which equals the trait correlation, assuming full sample overlap (which is slightly conservative). Similarly, since we did not have access to individual level (residual) data, it was not possible to carry out a single total parent residual GWAS under the CES assumption. Instead we meta-analysed the single parent effect sizes using inverse variance meta-analysis, but adjusted the standard errors to reflect the correlation amongst the traits (r) as follows:

$$\mathrm{SE}(\hat{\beta}) = \mathrm{SE}_0(\hat{\beta})\sqrt{1+r}$$

Where $\mathrm{SE}_0(\hat{\beta})$ is the usual (uncorrected) inverse-variance weighted meta-analysis standard error, ignoring the correlation amongst the estimates and $\mathrm{SE}(\hat{\beta})$ is the corrected estimate used.

This is slightly conservative as

$$\mathrm{Variance}(\hat{\beta}) = \mathrm{Variance}_0(\hat{\beta})\left(1+\frac{2rs_1s_2}{s_1^2+s_2^2}\right) <= \mathrm{Variance}_0(\hat{\beta})(1+r) \quad (2)$$

which follows straight forwardly from $\hat{\beta} = \frac{P_1\hat{\beta}_1+P_2\hat{\beta}_2}{P_1+P_2}$.

Where s_1 and s_2 are the standard error of the individual estimates and P_1, P_2 their associated precisions (i.e. reciprocal of the variance). Equation (2) is always conservative, but exact if $s_1 = s_2$. In practice s_1 and s_2 were similar, as the sample sizes, allele frequencies and variance in the traits for the two parents were very similar.

As we were using unrelated populations and fitting forty principal components to control for population structure, material inflation of test statistics due to structure or relatedness was not to be expected. This was confirmed using the intercept of LD-score regression (***Bulik-Sullivan et al., 2015***) for the summary statistics as shown in ***Table 1—source data 2***. We have tried to use a consistent approach to the direction of lifespan altering effects: positive implies longer life, consistent with previous studies of long-livedness (***Deelen et al., 2014***). Our base measure was thus a protection ratio, directly mirroring the cox hazard ratio. Effect sizes (betas) are typically –$\log_e$(cox hazard ratio), which we denote the $\log_e$(protection ratio). Years of life gained were estimated as 10 * log protection ratio, in accordance with a long-standing actuarial rule of thumb and recently verified (***Joshi et al., 2017***).

Candidate SNP replication

We sought to reproduce and replicate genome-wide significant associations reported by ***Pilling et al. (2017)***, who recently published a GWAS on the same UK Biobank data, but using a slightly different method. Rather than excluding relatives, Pilling *et al.* used BOLT-LMM and the genomic relationship matrix in subjects, to approximately account for covariance amongst parental phenotype. Pilling *et al.* also analysed parents separately as well as jointly, using a last survivor phenotype. Despite these factors, reproduction (obtaining the same result from almost the same data) was straightforward and consistent, once effect sizes were placed on the same scale (see below and ***Figure 2—figure supplement 1***), confirming our re-scaling was correct. To try to independently replicate their results, we used the consortium, LifeGen, excluding individuals from UK Biobank.

Pilling et al. (2017) performed multiple parental survival GWAS in UK Biobank, identifying 14 loci using combined parent lifespan and 11 loci using individual parent lifespan. Their study design involved rank-normalising Martingale residuals before regressing against genotype, which does not give an estimate of the $\log_e$(hazard ratio) (lnHR), nor, we believe, another naturally interpretable scale of effects, as the scale is now dependent on the proportion dead. Simulations (not shown) suggested $sd \approx \sqrt{c}$ for some Martingale residual distributions, where sd is the standard deviation of the distribution and c is the proportion dead. As multiplying the untransformed Martingale residual distribution by 1/c gives an estimate of the hazard ratio (***Joshi et al., 2017***; ***Therneau et al., 1990***), for individual parents, we could convert Pilling *et al.*'s effect sizes by multiplying them by sqrt(c) to return them to the Martingale residual scale (which still depends on the study structure) and then by 1/c to place them on the lnHR scale, using the proportion dead from Pilling *et al.*'s study descriptives. Further multiplication by 2 allows conversion from a subject-parent effect to an effect in self. The cumulative scale conversion allowing for all three of these effects was to multiply Pilling *et al.*'s effect sizes by 2.5863/2.2869 in mothers/fathers, respectively, placing them on a lnHR scale for effects in male/female subjects. The joint life parent phenotype does not appear to have a straightforward conversion to lnHR in self. Instead, we used an empirical estimate derived from effect sizes comparison of the APOE allele between Pilling *et al.*'s discovery sample and our own UK Biobank Gen. British sample (both parents combined), which were largely overlapping: to get from Pilling *et al.*'s effect size to $\log_e$ HR, we had to multiply their effect sizes by 1.9699 for *APOE* and used this ratio for other alleles, which should be completely valid under the proportional hazard assumption. Whilst this scheme may appear a little *ad hoc* (the use of simulation and *APOE*), it was confirmed empirically: visual inspection indeed showed hazard ratios from our own UK Biobank Gen. British sample calculations and inferred hazard ratios from Pilling were highly concordant (noting the concordance for joint life at *APOE*, which was pre-defined to be perfectly concordant by our procedure, is not, of itself, evidence) (***Figure 2—figure supplement 1***).

Flachsbart et al. (2009), ***Deelen et al. (2014)***, and ***Broer et al. (2015)*** tested extreme longevity cases (95–110 years, $\geq$85 years, $\geq$90 years, respectively) against controls (60–75 years, 65 years, deceased at 55–80 years, respectively), identifying SNPs at or near *FOXO3* and 5q33.3/*EBF1*. As done previously (***Joshi et al., 2017***), we estimated the relationship between study-specific longevity log odds ratio and log hazard ratio empirically using the most powered *APOE* variant overlapping with our own study, assuming increased odds of surviving to extreme age is due to a reduction in lifetime mortality risk. For Flachsbart *et al.* and Deelen *et al.*, we used rs4420638_G (reported log OR –0.33 (***Deelen et al., 2014***), our lnHR 0.086). Inverting the sign to give $\log_e$(protection ratio) estimates, the conversion estimate used was 3.82. For Broer *et al.*, we used rs6857_T (reported log OR –0.20 (***Broer et al., 2015***), our lnHR 0.087), with a conversion estimate of 0.43, again yielding a $\log_e$(protection ratio).

Ben-Avraham et al. (2017) reported a deletion in Growth Hormone Receptor exon 3 (*d3-GHR*) associated with an increase of 10 years in male lifespan when homozygous. This deletion is tagged by rs6873545_C (***McKay et al., 2007***), which is present in the UK Biobank and LifeGen population sample at a frequency of 26.9% (q). Considering the association is recessive and we are imputing father genotypes, we converted the reported effect size into expected years of life per allele as follows:

If the subject genotype is CT, the parent contributing the C allele has 50% chance of being the father and $\frac{q^2}{q^2+2pq}$ chance of being homozygous. If the subject genotype is CC, the father has 100% chance of contributing the C allele and again has $\frac{q^2}{q^2+2pq}$ chance of being homozygous. We therefore expect the relationship to be $\hat{\beta}_C = \frac{1}{2}\hat{\beta}_{CC}\frac{q^2}{q^2+2pq}$, where $\hat{\beta}_C$ is the observed effect per subject allele on father lifespan and $\hat{\beta}_{CC}$ is the reported effect of the homozygous deletion in the father. As before, doubling the allele effect gives an estimate of the effect of the allele on subject lifespan, which finally yielded a converted estimate of 0.155. That is, under Ben-Avrahim *et al.*'s assumptions on inheritance patterns, if their estimate of effect size in minor homozygotes is correct, we should see under the additive model an effect size of 0.155 years, or a $\log_e$(hazard ratio) of –0.015, and correspondingly scaled standard errors (note we are assuming that the effect is actually recessive, and estimating how that effect should appear if an additive model were fitted).

Standard errors were calculated from inferred betas and reported P values, assuming a two-sided test with a normally-distributed estimator. Confidence interval overlap was then assessed using a two-sided test on the estimate difference (P_{diff}), using a Z statistic from the difference divided by the standard error of the difference.

iGWAS

We performed a Bayesian Genome-Wide Association Study using the CES GWAS results and summary statistics on 58 risk factor GWASs (imputed, leading to 7.2 million SNPs in common between all the studies), as described by ***McDaid et al. (2017)***. To calculate our prior for SNPs on a given chromosome, first we used a multivariate Mendelian Randomization (masking the focal chromosome) to identify the risk factors significantly influencing lifespan and estimate their causal effect. This identified 16 risk factors independently causally contributing to lifespan (see ***Table 2—source data 1*** for the causal effect estimates). Next, assuming that a SNP affects lifespan through its effects on the 16 risk factors, prior effects estimates were estimated as the sum of the products of the causal effect estimates of the 16 significant risk factors on lifespan and the effect of the SNP on each risk factor. We added one to the prior effect variance formula described in ***McDaid et al. (2017)*** to account for the fact that prior effects are estimated using observed Z-scores, and not true Z-scores, with $Z_{obs} \sim \mathcal{N}(Z_{true},\ 1)$.

We computed Bayes factors by combining the prior effects and the observed association Z statistics. The significance of the Bayes factors was assessed using a permutation approach to calculate P values, by comparing observed Bayes factors to 7.2 billion null Bayes factors. These null Bayes factors were estimated using 1000 null sets of Z statistics combined with the same priors. These empirical P values were then adjusted for multiple testing using the Benjamini-Hochberg procedure.

Replication in extreme long-livedness

To replicate our novel lifespan findings, we inverse-variance meta-analysed summary statistics from ***Deelen et al. (2014)*** and ***Broer et al. (2015)***, having converted their effect sizes to a common scale (see Candidate SNP replication). These effect sizes were given or could be estimated from P value, effect direction and N, as well as the SNPs MAF.

Both longevity studies and the LifeGen consortium contain individuals from the Rotterdam Study, but due to differences in trait definitions, we could not inflate standard errors directly based on sample overlap. Instead, we calculated the covariance in null SNPs (Z < 1) between each study (r ~ 0.01) and then adjusted the standard errors based on equation 5 from ***Lin and Sullivan (2009)***:

$$\mathrm{Variance}(\hat{\beta}) = \mathrm{Variance}_0(\hat{\beta}) + 2 * \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} w_n w_m \,\mathrm{Cov}(\hat{\beta}_\mathrm{n}, \hat{\beta}_\mathrm{m})$$

Where $\mathrm{Variance}_0$ is the unadjusted variance of the SNP effect $\hat{\beta}$ after meta-analysis, w is the inverse variance weight of the SNP, N is the number of studies, and Cov $(\hat{\beta}_\mathrm{n}, \hat{\beta}_\mathrm{m})$ is the null SNP covariance between each study.

Test of the hypothesis that the effect was zero, was one sided, with alternate hypothesis that the effect had the same sign as in discovery. Effect sizes in discovery and replication were then compared by calculating the ratio (alpha) of replication effect sizes to discovery effect sizes:

$$\alpha = \frac{\beta_\mathrm{rep}}{\beta_\mathrm{disc}}$$

and its standard error using the following formula, reflecting the Taylor series expansion of the denominator for SE:

$$\mathrm{SE}_\alpha = \sqrt{\frac{\mathrm{SE}_\mathrm{rep}^2}{\beta_\mathrm{disc}^2} + \frac{\beta_\mathrm{rep}^2\,\mathrm{SE}_\mathrm{disc}^2}{\left(\beta_\mathrm{disc}^2\right)^2}}$$

where *rep* and *disc* are replication and discovery, respectively. Alpha was then inverse-variance meta-analysed across all SNPs to test for collective evidence that the discovery SNPs influence longevity.

Age and sex-stratified effects

Calculation of age and sex stratified effect sizes was done using the full Cox model (***Equation 1***), imputed dosages and the package 'Survival' in R. We split the full analysis into age decades from 40 to 90 and a wider band, 90–120, beyond that, excluding any parent who died at an age younger than the age band and treating any parent who died beyond the age band as alive at the end of the age band. We thus had, across independent periods of life, estimates of the hazard ratio by decade of age and parent sex, along with standard errors. This gave estimates of the hazard ratio beta(age band, sex) in each age band and sex.

We tested the effect of age and sex, by fitting the linear model beta(age band, sex) = intercept + beta1 x ageband + beta2 x sex + e, where e is independent, but with known variance (the square of the SE in the age/sex stratified model fit) and using the rma function from the R package 'metafor' which uses known variances of dependent variables. The process is more easily understood by examining the age and sex related effect size graphs, and recognising we are fitting age and sex as explanatory variables, considering the standard error of each point shown.

Causal gene prediction

In order to more accurately implicate causal genes and methylation sites from the detected loci associated with human lifespan, Summary-level Mendelian Randomisation (SMR) and HEterogeneity In InDependent Instruments (HEIDI) tests (***Zhu et al., 2016***) were performed on our CES GWAS statistics. Three separate analyses were performed. First, cis-eQTL scan results from peripheral blood tissue from two previous studies, the Westra data (***Westra et al., 2013***) and CAGE data (***Lloyd-Jones et al., 2017***), were used to prioritize causal genes. Second, cis-eQTL signals (SNPs with FDR < 0.05) for 48 tissues from the GTEx consortium (***Battle et al., 2017***) were used to prioritize causal genes in multiple tissues. Third, genome-wide methylation QTL (mQTL) scan signals in blood tissue from the Brisbane Systems Genetics Study and Lothian Birth Cohort (***McRae et al., 2017***) were used to predict causal CpG sites associated with human lifespan loci. All results from SMR test passing a 5% FDR threshold where the HEIDI test p > 0.05 were reported.

Fine-mapping using LASSO regression

Selection Operator for JOint multi-SNP analysis (SOJO) (***Ning et al., 2017***) was used to perform conditional fine-mapping analysis of the lifespan loci. The SOJO procedure implements LASSO regression for each locus, which outperforms standard stepwise selection procedure (e.g. GCTA-COJO), based on summary association statistics and the European-ancestry 1000 Genomes samples for LD reference. We based the SOJO analysis on our CES summary association statistics from UK Biobank and used the LifeGen summary statistics as validation sample to optimise the LASSO tuning parameters for each locus. Loci were defined prior to analysis as 1 Mb windows centred at each top variant from the GWAS. For each locus, based on UK Biobank data, we recorded the first 30 variants entering the model and the tuning parameters for these entering points along the LASSO path, as well as the LASSO results at the tuning parameters. For each recorded tuning parameter, we then built a polygenic score and computed the proportion of predictable variance from the regional polygenic score in the validation sample. The best out-of-sample R squared is reported, together with the selected variants per locus.

Lifespan variance explained

We sought an independent set of disease-associated SNPs to assess which diseases had the greatest effect on lifespan. A large number of SNPs per disease category, especially other cancers, were used to ensure that diseases were not under-represented when testing for association with lifespan. The latest, genome-wide significant disease SNPs from European ancestry studies were retrieved from the GWAS catalog (14 March 2018), based on string matching within reported trait names. For Alzheimer's/Parkinson's disease, these were 'alzh' and 'parkin'; for CVD, these were 'myocard', 'cvd', 'cardiovascular', 'coronary', and 'artery disease'; for Type 2 diabetes this was 'type 2 diabetes'; for cancers, this was 'cancer', 'noma', 'ioma', 'tumo[u]r', and 'leukemia'. Cancers were then divided in Lung cancer and Other cancers based on the presence or absence of the keyword 'lung'. The Smoking/Lung cancer category was created by adding traits containing the keywords 'smoking' and 'chronic obstructive' to the lung cancers. Each category was manually checked to include only

associations with the diseases themselves or biomarkers of the diseases. Although some throat cancers are often caused by smoking and alcohol consumption, we did not treat these as smoking loci; in practice, this choice had no effect as the only significant throat cancer locus (oesophageal cancer near CFTR) was discounted as secondary pleiotropic – see below.

SNPs missing from the CES summary statistics were imputed from the closest proxy (min. $r^2 > 0.9$) or averaged from multiple proxies if equally close. SNPs without effect sizes, SNPs matching neither our reference nor effect alleles, and SNPs with reported frequencies differing by more than 0.3 from our own were excluded. The remaining SNPs were subdivided into independent ($r^2 < 0.1$) loci 500 kb apart, keeping the SNP most strongly associated with lifespan in the CES GWAS – thus proportional only to the lifespan GWAS test statistic and independent from the number of disease SNPs per category. Lastly, where possible, loci were tested for association with their disease category in UK Biobank, using self-reported diseases of 325,292 unrelated, genomically British subjects, their siblings, and each parent separately. Diseases of subject relatives were already coded into broad disease categories by UK Biobank. For subjects themselves, ICD codes had been recorded which we grouped into similar categories (hypertension, cerebral infarction, heart disease, diabetes, dementia, depression, stress, pulmonary disease, and cancer, in accordance with ***Figure 10—source data 1***, although cancer in subjects was more directly taken as the trait of reporting number of cancers > 0). The trait of reporting these diseases (separately for each relative and the subject themselves) was then tested for association with genotypic dosage for the GWAS catalog disease SNPs. The model fitted was a logistic regression of not reporting the disease, using the same covariates as the CES GWAS with the addition of subject age, and estimated the log odds ratio of protection from disease in UK Biobank for each copy of the disease-protective allele in the GWAS catalog. Effects reported in the GWAS catalog for which we found the pooled estimate from our association study was in the opposite direction were flipped (if $p < 0.05$) or discarded (if $p \geq 0.05$).

Our final dataset consisted of 555 disease SNPs (81 neurological, 72 cardiovascular, 65 diabetes, 22 smoking/lung cancer, and 315 other cancers). Lifespan variance explained (LVE) was calculated as $2pqa^2$, where p and q are the frequencies of the effect and reference alleles in our lifespan GWAS, and a is the SNP effect size in years of life (***Falconer et al., 1996***). To assess pleiotropy, SNPs were tested against other disease categories, and where possible, the relative strengths of standardised associations between disease categories were compared. SNPs associating more strongly with another disease, as defined by a Z statistic more than double that of the original disease, were marked as pleiotropic and secondary. Whilst strength of association would not normally be perceived as appropriately measured in this way (odds ratio being more conventional and independent of prevalence), here we are interested in the excess number of disease cases in the population due to the variant, so any locus with a moderate OR for a highly prevalent disease is judged more causative of that disease than a locus with a (somewhat) higher OR for a very rare disease, as the number of attributable cases will be lower. The Z statistic captures this – given that p and q are obviously the same (same SNP, same population). Correspondingly, for diseases only present in one sex, the other sex was treated as all controls. Whilst this halves the apparent effect size, the required measure is the amount of disease caused across the whole population. A SNP conferring similar attributable counts of CVD and breast cancer in women, but also CVD in men, is causing CVD more than cancer across the population. Correspondingly, selection pressure on the breast cancer effect is half that for a matching effect in both sexes. SNPs conferring both an increase in disease and an increase in lifespan were marked as antagonistically pleiotropic. Unsurprisingly, in practice, there were one or more other diseases reduced by the SNP and therefore the reported disease-increasing association was considered secondary. Total LVE per disease category was calculated by summing SNPs not marked as secondary and with significant effects on lifespan, where significance was determined by setting an FDR threshold that allowed for one false positive among all SNPs tested ($q \leq 0.016$, 60 SNPs). To compare the cumulative LVE of the top LVE loci, all non-secondary association SNPs from the disease categories were pooled and again subdivided into independent loci ($r^2 < 0.1$) 500 kb apart. Applying an FDR threshold with the same criteria ($q \leq 0.022$), a total of 45 (1 neurological, 23 cardiovascular, four diabetes, six smoking/lung cancer, and 11 other cancer) independent loci remained and their LVE was summed by disease category.

Cell type and pathway enrichment

Stratified LD-score regression (**Finucane et al., 2015**) partitions SNP heritability into regions linked to specific tissues and cell types, such as super-enhancers and histone marks, and then assesses whether the SNPs in these regions contribute disproportionately to the total SNP heritability. Standard LD-score regression (**Bulik-Sullivan et al., 2015**) indicated that between the different samples (UK Biobank and/or LifeGen) and analyses (CES or SSE), the CES results from UK Biobank genomically British ancestry individuals had the highest SNP heritability, plausibly due to its uniformity of population sample. These statistics were analysed using the procedure described by **Finucane et al., 2015**, which included limiting the regressions to HapMap3 SNPs with MAF > 0.05 to reduce statistical noise. Results from all cell types were merged and then adjusted for multiple testing using Benjamini–Hochberg (FDR 5%).

The full CES dataset was subjected to gene-based tests, which used up to 10^6 SNP permutations per gene to assign P values to 26,056 genes, as implemented by VEGAS2 v2.01.17 (**Mishra and Macgregor, 2015**). Only directly genotyped SNPs from the UK Biobank array were used to facilitate practical runtimes. Using the default settings, all SNPs located within genes (relative to the 5' and 3' UTR) were included. Scored genes were then tested for enrichment in 9741 pathways from the NCBI BioSystems Database with up to 10^8 gene permutations per pathway using VEGAS2Pathway (**Mishra and MacGregor, 2017**). Pathway enrichment P values were automatically adjusted for pathway size (empirical P) and further adjusted for multiple testing using Benjamini-Hochberg (FDR 5%).

DEPICT was also used to create a list of genes; however, this method uses independent SNPs passing a P value threshold to define lifespan loci and then attempts to map 18,922 genes to them. Gene prioritization and subsequent gene set enrichment is done for 14,461 probabilistically-defined reconstituted gene sets, which are tested for enrichment under the self-contained null hypothesis (**Pers et al., 2015**). Two separate analyses were performed on the CES summary statistics, using independent SNPs (>500 kb between top SNPs) which were present in the DEPICT database. The first analysis used a genome-wide significance threshold (GW DEPICT analysis) and mapped genes to 10 loci, automatically excluding the major histocompatibility complex (MHC) region. The second used a suggestive significance threshold ($p < 10^{-5}$), which yielded 93 loci and mapped genes to 91 of these, again excluding the MHC region. To test if pathways were significantly enriched at a 5% FDR threshold, we used the values calculated by DEPICT, already adjusted for the non-independence of the gene sets tested.

PASCAL was used with the same summary statistics and gene sets as DEPICT, except the gene probabilities within the sets were dichotomized (Z > 3) (**Marouli et al., 2017**), leading to the analysis of the same 14,461 pathways. PASCAL transformed SNP P values into gene-based P values (with default method '–genescoring=sum') for 21,516 genes (**Lamparter et al., 2016**). When testing the pathways for overrepresentation of high gene scores, the P values are estimated under the competitive null hypothesis (**Maciejewski, 2014**). These pathway empirical P values were further adjusted for multiple testing using Benjamini-Hochberg procedure.

Age-related eQTLs enrichment

We identified SNPs in our CES GWAS that were eQTLs that is associated with the expression of at least one gene with $p < 10^{-4}$ in a dataset by the eQTLGen Consortium (n = 31,684 individuals) (**Võsa et al., 2018**). A total of 3577 eQTLs after distance pruning (500 kb) were present, of which 755 were associated with genes differentially expressed with age (**Peters et al., 2015**). We used Fisher's exact test to determine, amongst the set of eQTLs, if SNPs which were associated with lifespan (at varying thresholds of statistical significance) were enriched for SNPs associated with genes whose expression is age-related.

Polygenic lifespan score associations

We used the CES GWAS, excluding (one at a time) all Scottish populations (whether from Scottish UK Biobank assessment centres or Scottish LifeGen cohorts), Estonian populations and a random 10% of UK Biobank English and Welsh subjects to create polygenic risk scores using PRSice (**Euesden et al., 2015**), where the test subjects had not been part of the training data. As we find polygenic risk scores developed using all ($p \leq 1$) independent ($r^2 < 0.1$) SNPs (PRSP1), rather than

those passing a tighter significance threshold are most associated (highest standardised effect size; see ***Table 3—source data 1*** for comparison between thresholds), these were used in the analysis.

To make cross-validated lifespan associations using polygenic scores, our unrelated, genomically British sample was partitioned into training and test sets. The first test set consisted of Scottish individuals from UK Biobank, as defined by assessment centre or northings and eastings falling within Scotland (N = 24,059). The second set consisted of a random subset of the remaining English and Welsh population, reproducibly sampled based on the last digit of their UK Biobank identification number (#7, N = 29,815). The training set was constructed by excluding these two populations, as well as excluding individuals from Generation Scotland, from our GWAS and recalculating estimates of beta on that subset.

A third independent validation set was constructed by excluding the EGCUT cohort from the LifeGen sample and using the remaining data to test lifespan in the newly genotyped EGCUT cohort (***Leitsalu et al., 2015***), using unrelated individuals only (N = 36,499).

Polygenic survival scores were constructed using PRSice 2.0.14.beta (***Euesden et al., 2015***) in a two-step process. First, lifespan SNPs were LD-clumped based on an r^2threshold of 0.1 and a window size of 250 kb. To facilitate practical run times of PRSice clumping, only directly genotyped SNPs were used in the Scottish and English/Welsh subsets. The Estonian sample was genotyped on four different arrays with limited overlap, so here imputed data (with imputation measure R2 > 0.9) was used and clumped with PLINK directly (r^2 = 0.1; window = 1000 kb). The clumped SNPs (85,539 in UK Biobank, 68,234 in Estonia) were then further pruned based on several different P value thresholds, to find the most informative subset. For all individuals, a polygenic score was calculated as the sum of SNP dosages (of SNPs passing the P value threshold) multiplied by their estimated allele effect. These scores were then standardised to allow for associations to be expressed in standard deviations in polygenic scores.

Polygenic scores of test cohorts were regressed against lifespan and alive/dead status using a cox proportional hazards model, adjusted for sex, assessment centre, batch, array, and 10 principal components. Where parental lifespan was used, hazard ratios were doubled to gain an estimate of the polygenic score on own mortality. Scores were also regressed against self-reported diseases in UK Biobank subjects, their siblings, and each parent separately, using a logistic regression adjusted for the same covariates as in the lifespan analysis plus subject age. As with previous disease associations, estimates were transformed so positive associations indicate a protective or life-extending effect, and effect estimates of first degree relatives were doubled. Meta-analysis of estimates between cohorts was done using inverse variance weighting. Where estimates between kin were meta-analysed, standard errors were adjusted for correlation between family members. This involved multiplying standard errors by $\sqrt{1+r}$ for each correlation (r) with the reference kin (***Equation 2***), which appears slightly conservative. As correlations between family member diseases were very low (range 0.0005 to 0.1048), in practice, this adjustment had no effect.

URLs

MultiABEL: https://github.com/xiashen/MultiABEL/
LDSC: https://github.com/bulik/ldsc
SMR/HEIDI: https://cnsgenomics.com/software/smr/
SOJO: https://github.com/zhenin/sojo/
DEPICT: https://www.broadinstitute.org/mpg/depict/
PASCAL: https://www2.unil.ch/cbg/index.php?title=Pascal
GTEx: https://gtexportal.org/home/datasets

Acknowledgments

We thank the UK Biobank Resource, approved under application 8304; we acknowledge funding from the UK Medical Research Council Human Genetics Unit, Wellcome Trust PhD Training Fellowship for Clinicians - the Edinburgh Clinical Academic Track (ECAT) programme (204979/Z/16/Z), the Medical Research Council Doctoral Training Programme in Precision Medicine (MR/N013166/1) and the AXA research fund. We thank Tom Haller of the University of Tartu, for tailoring RegScan so we could use it with compressed files (***Haller et al., 2015***). We would also like to thank the researchers, funders and participants of the LifeGen consortium (***Joshi et al., 2017***).

Additional information

Group author details

eQTLGen Consortium

M Agbessi: Ontario Institute for Cancer Research, Toronto, Canada; **H Ahsan**: Department of Public Health Sciences, University of Chicago, Chicago, United States; **I Alves**: Ontario Institute for Cancer Research, Toronto, Canada; **A Andiappan**: Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore; **P Awadalla**: Ontario Institute for Cancer Research, Toronto, Canada; **A Battle**: Department of Computer Science, Johns Hopkins University, Baltimore, United States; **MJ Bonder**: Department of Genetics University, Medical Centre Groningen, Groningen, The Netherlands; **D Boomsma**: Vrije Universiteit, Amsterdam, The Netherlands; **M Christiansen**: Cardiovascular Health Research Unit, University of Washington, Seattle, United States; **A Claringbould**: Department of Genetics University, Medical Centre Groningen, Groningen, The Netherlands; **P Deelen**: Department of Genetics University, Medical Centre Groningen, Groningen, The Netherlands; **J van Dongen**: Vrije Universiteit, Amsterdam, The Netherlands; **T Esko**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **M Favé**: Ontario Institute for Cancer Research, Toronto, Canada; **L Franke**: Department of Genetics University, Medical Centre Groningen, Groningen, The Netherlands; **T Frayling**: Exeter Medical School, University of Exeter, Exeter, United Kingdom; **SA Gharib**: Department of Medicine, University of Washington, Seattle, United States; **G Gibson**: School of Biological Sciences, Georgia Institute of Technology, Atlanta, United States; **G Hemani**: MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; **R Jansen**: Vrije Universiteit, Amsterdam, The Netherlands; **A Kalnapenkis**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **S Kasela**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **J Kettunen**: University of Helsinki, Helsinki, Finland; **Y Kim**: Department of Computer Science, Johns Hopkins University, Baltimore, United States; **H Kirsten**: Institut für Medizinische Informatik, Statistik und Epidemiologie, LIFE – Leipzig ResearchCenter for Civilization Diseases, Universität Leipzig, Leipzig, Germany; **P Kovacs**: IFB Adiposity Diseases, Universität Leipzig, Leipzig, Germany; **K Krohn**: Interdisciplinary Center for Clinical Research, Faculty of Medicine, Universität Leipzig, Leipzig, Germany; **J Kronberg-Guzman**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **V Kukushkina**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **Z Kutalik**: Lausanne University Hospital, Lausanne, Switzerland; **M Kähönen**: Department of Clinical Physiology and Faculty of Medicine and Life Sciences, Tampere University Hospital and University of Tampere, Tampere, Finland; **B Lee**: Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore; **T Lehtimäki**: Department of Clinical Chemistry, Fimlab Laboratories and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland; **M Loeffler**: Institut für Medizinische Informatik, Statistik und Epidemiologie, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany; **U Marigorta**: School of Biological Sciences, Georgia Institute of Technology, Atlanta, United States; **A Metspalu**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **J van Meurs**: Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands; **L Milani**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **M Müller-Nurasyid**: Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; **M Nauck**: Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany; **M Nivard**: Vrije Universiteit, Amsterdam, The Netherlands; **B Penninx**: Vrije Universiteit, Amsterdam, The Netherlands; **M Perola**: National Institute for Health and Welfare, University of Helsinki, Helsinki, Finland; **N Pervjakova**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **B Pierce**: Department of Public Health Sciences, University of Chicago, Chicago, United States; **J Powell**: Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; **H Prokisch**: Institute of Human Genetics, Helmholtz Zentrum München, München, Germany; **BM Psaty**: Departments of Epidemiology, Medicine, and Health Services, Cardiovascular Health Research Unit, University of Washington, Seattle, United States; **O Raitakari**: Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital and University of Turku, Turku, Finland; **S Ring**: School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom; **S Ripatti**: University of Helsinki, Helsinki, Finland; **O Rotzschke**: Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore; **S**

Ruëger: Lausanne University Hospital, Lausanne, Switzerland; **A Saha**: Department of Computer Science, Johns Hopkins University, Baltimore, United States; **M Scholz**: Institut für Medizinische InformatiK, Statistik und Epidemiologie, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany; **K Schramm**: Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; **I Seppälä**: Department of Clinical Chemistry, Fimlab Laboratories and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland; **M Stumvoll**: Department of Medicine, Universität Leipzig, Leipzig, Germany; **P Sullivan**: Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; **A Teumer**: Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany; **J Thiery**: Institute for Laboratory Medicine, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany; **L Tong**: Department of Public Health Sciences, University of Chicago, Chicago, United States; **A Tönjes**: Department of Medicine, Universität Leipzig, Leipzig, Germany; **J Verlouw**: Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands; **PM Visscher**: Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; **U Võsa**: Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands; **U Völker**: Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany; **H Yaghootkar**: Exeter Medical School, University of Exeter, Exeter, United Kingdom; **J Yang**: Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; **B Zeng**: School of Biological Sciences, Georgia Institute of Technology, Atlanta, United States; **F Zhang**: Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; **M Agbessi**: Ontario Institute for Cancer Research, Toronto, Canada; **H Ahsan**: Department of Public Health Sciences, University of Chicago, Chicago, United States; **I Alves**: Ontario Institute for Cancer Research, Toronto, Canada; **A Andiappan**: Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore; **P Awadalla**: Ontario Institute for Cancer Research, Toronto, Canada; **A Battle**: Department of Computer Science, Johns Hopkins University, Baltimore, United States; **MJ Bonder**: Department of Genetics University, Medical Centre Groningen, Groningen, The Netherlands; **D Boomsma**: Vrije Universiteit, Amsterdam, The Netherlands; **M Christiansen**: Cardiovascular Health Research Unit, University of Washington, Seattle, United States; **A Claringbould**: Department of Genetics University, Medical Centre Groningen, Groningen, The Netherlands; **P Deelen**: Department of Genetics University, Medical Centre Groningen, Groningen, The Netherlands; **J van Dongen**: Vrije Universiteit, Amsterdam, The Netherlands; **T Esko**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **M Favé**: Ontario Institute for Cancer Research, Toronto, Canada; **L Franke**: Department of Genetics University, Medical Centre Groningen, Groningen, The Netherlands; **T Frayling**: Exeter Medical School, University of Exeter, Exeter, United Kingdom; **SA Gharib**: Department of Medicine, University of Washington, Seattle, United States; **G Gibson**: School of Biological Sciences, Georgia Institute of Technology, Atlanta, United States; **G Hemani**: MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; **R Jansen**: Vrije Universiteit, Amsterdam, The Netherlands; **A Kalnapenkis**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **S Kasela**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **J Kettunen**: University of Helsinki, Helsinki, Finland; **Y Kim**: Department of Computer Science, Johns Hopkins University, Baltimore, United States; **H Kirsten**: Institut für Medizinische Informatik, Statistik und Epidemiologie, LIFE – Leipzig ResearchCenter for Civilization Diseases, Universität Leipzig, Leipzig, Germany; **P Kovacs**: IFB Adiposity Diseases, Universität Leipzig, Leipzig, Germany; **K Krohn**: Interdisciplinary Center for Clinical Research, Faculty of Medicine, Universität Leipzig, Leipzig, Germany; **J Kronberg-Guzman**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **V Kukushkina**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **Z Kutalik**: Lausanne University Hospital, Lausanne, Switzerland; **M Kähönen**: Department of Clinical Physiology and Faculty of Medicine and Life Sciences, Tampere University Hospital and University of Tampere, Tampere, Finland; **B Lee**: Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore; **T Lehtimäki**: Department of Clinical Chemistry, Fimlab Laboratories and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland; **M Loeffler**: Institut für Medizinische Informatik, Statistik und Epidemiologie, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany; **U Marigorta**: School of Biological Sciences, Georgia Institute of Technology, Atlanta, United States; **A Metspalu**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **J van Meurs**: Department of Internal

Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands; **L Milani**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **M Müller-Nurasyid**: Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; **M Nauck**: Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany; **M Nivard**: Vrije Universiteit, Amsterdam, The Netherlands; **B Penninx**: Vrije Universiteit, Amsterdam, The Netherlands; **M Perola**: National Institute for Health and Welfare, University of Helsinki, Helsinki, Finland; **N Pervjakova**: Estonian Genome Center, University of Tartu, Tartu, Estonia; **B Pierce**: Department of Public Health Sciences, University of Chicago, Chicago, United States; **J Powell**: Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; **H Prokisch**: Institute of Human Genetics, Helmholtz Zentrum München, München, Germany; **BM Psaty**: Departments of Epidemiology, Medicine, and Health Services, Cardiovascular Health Research Unit, University of Washington, Seattle, United States; **O Raitakari**: Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital and University of Turku, Turku, Finland; **S Ring**: School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom; **S Ripatti**: University of Helsinki, Helsinki, Finland; **O Rotzschke**: Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore; **S Ruëger**: Lausanne University Hospital, Lausanne, Switzerland; **A Saha**: Department of Computer Science, Johns Hopkins University, Baltimore, United States; **M Scholz**: Institut für Medizinische InformatiK, Statistik und Epidemiologie, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany; **K Schramm**: Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; **I Seppälä**: Department of Clinical Chemistry, Fimlab Laboratories and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland; **M Stumvoll**: Department of Medicine, Universität Leipzig, Leipzig, Germany; **P Sullivan**: Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; **A Teumer**: Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany; **J Thiery**: Institute for Laboratory Medicine, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany; **L Tong**: Department of Public Health Sciences, University of Chicago, Chicago, United States; **A Tönjes**: Department of Medicine, Universität Leipzig, Leipzig, Germany; **J Verlouw**: Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands; **PM Visscher**: Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; **U Võsa**: Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands; **U Völker**: Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany; **H Yaghootkar**: Exeter Medical School, University of Exeter, Exeter, United Kingdom; **J Yang**: Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; **B Zeng**: School of Biological Sciences, Georgia Institute of Technology, Atlanta, United States; **F Zhang**: Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia

Competing interests

Tõnu Esko: Reviewing Editor, eLife. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
Medical Research Council	DTP in Precision Medicine MR/N013166/1,HGU QTL in health and disease	Paul RHJ Timmers Andrew D Bretherick David W ClarkeQTLGen ConsortiumJames F Wilson
Estonian Research Competency Council	PUT 1665	Kristi Lall Krista Fischer
Wellcome Trust	PhD Training Fellowship for Clinicians	Andrew D Bretherick
Edinburgh Clinical Academic Track	204979/Z/16/Z	Andrew D Bretherick
Svenska Forskningsrådet Formas	2014-00371	Xia Shen

Svenska Forskningsrådet Formas	2017-02543	Xia Shen
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung	31003A_169929	Zoltán Kutalik
SystemsX.ch	51RTP0_151019	Zoltán Kutalik
AXA Research Fund		Peter K Joshi

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Paul RHJ Timmers, Ninon Mounier, Formal analysis, Investigation, Visualization, Writing—original draft; Kristi Lall, Formal analysis, Investigation, Writing—original draft; Krista Fischer, Xia Shen, Conceptualization, Formal analysis, Supervision, Investigation, Visualization, Methodology, Writing—original draft; Zheng Ning, Xiao Feng, Formal analysis, Visualization, Writing—review and editing; Andrew D Bretherick, David W Clark, Software, Formal analysis, Writing—review and editing; Tõnu Esko, Resources, Supervision, Funding acquisition, Writing—original draft; Zoltán Kutalik, Conceptualization, Resources, Software, Investigation, Methodology, Writing—original draft; James F Wilson, Conceptualization, Resources, Supervision, Funding acquisition, Writing—original draft; Peter K Joshi, Conceptualization, Supervision, Validation, Methodology, Writing—original draft, Writing—review and editing

Author ORCIDs

Paul RHJ Timmers http://orcid.org/0000-0002-5197-1267
Xia Shen http://orcid.org/0000-0003-4390-1979
Tõnu Esko http://orcid.org/0000-0003-1982-6569
James F Wilson http://orcid.org/0000-0001-5751-9178
Peter K Joshi http://orcid.org/0000-0002-6361-5059

Ethics

Human subjects: This work used existing datasets, for which ethical approval had been gathered for genetic investigation at the time of collection.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.39856.047
Author response https://doi.org/10.7554/eLife.39856.048

Additional files

Supplementary files

- Supplementary file 1. Loci with significantly predicted candidate genes using SMR-HEIDI test and two eQTL datasets (blood tissue). This PDF contains a table and plots of the lifespan GWAS and eQTL signals genes from Westra and CAGE eQTL studies that pass FDR < 5% threshold for the SMR test and p>0.05 threshold for HEIDI test.

DOI: https://doi.org/10.7554/eLife.39856.035

- Supplementary file 2. Predicted causal elements by SMR-HEIDI using expression and methylation QTL data. This Excel workbook contains two sheets. The first sheet lists the genes and tissues prioritised by SMR-HEIDI using GTEx expression data, while the second sheet lists the methylation sites prioritised using mQTL data.

DOI: https://doi.org/10.7554/eLife.39856.036

- Supplementary file 3. Evidence of allelic heterogeneity of the lifespan loci via identification of secondary associations using SOJO. This Excel workbook contains two sheets. The first sheet is a summary of the number of additional variants identified by SOJO per locus. The second sheet is a

detailed list of all variants identified by SOJO per locus, their allele frequencies, and their independent effects.

DOI: https://doi.org/10.7554/eLife.39856.037

• Supplementary file 4. Cell types enriched for lifespan heritability identified by stratified LD-score regression. This Excel workbook contains a list of all cell types tested for lifespan heritability enrichment by Stratified LD-score regression and includes a list of coefficients for each cel type.

DOI: https://doi.org/10.7554/eLife.39856.038

• Supplementary file 5. Gene sets highlighted by VEGAS2Pathway and corresponding results from DEPICT and PASCAL (FDR < 5%). This Excel workbook contains two sheets. The first sheet lists the gene sets highlighted by VEGAS2Pathway as enriched for lifespan genes at FDR 5%, with corresponding results for DEPICT and PASCAL enrichment analyses. The second sheet lists all VEGAS gene sets significant at FDR < 15%, with a list of genes included in each set.

DOI: https://doi.org/10.7554/eLife.39856.039

• Supplementary file 6. eQTL SNPs associated with lifespan for genes whose expression varies with age. This PDF contains a table and figure on the age-related gene expression analysis.

DOI: https://doi.org/10.7554/eLife.39856.040

• Transparent reporting form

DOI: https://doi.org/10.7554/eLife.39856.041

Data availability

Individual phenotypic and genetic data is available from UK Biobank upon application: https://www.ukbiobank.ac.uk/register-apply/. Phenotypes used in this work include subject age, sex, ethnicity, relatedness, genotyping batch, array, and principal components, as well as parental age and alive/dead status. Also included are self-reported diseases of subjects and their families. A full description of our application can be found at https://www.ukbiobank.ac.uk/2015/02/dr-james-wilson-university-of-edinburgh-centre-for-population-sciences/. The results that support our findings, in particular, the GWAS summary statistics we generated for >1 million parental lifespans in this study are available at http://dx.doi.org/10.7488/ds/2463. Gene expression data used in the age-related gene analysis is being made available by the eQTLGen Consortium, see http://www.eqtlgen.org/ and Võsa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. bioRxiv 447367 (2018). Single tissue gene expression data used in the SMR-Heidi analysis can be found on the GTEx website https://gtexportal.org/home/datasets, under GTEx_Analysis_v7_eQTL.tar.gz. Source data files for main figures and tables are included as supplementary material, other than GWAS summary statistics.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Timmers PRHJ, Läll K, Fischer K, Ning Z, Feng X, Bretherick A, Clark DW, Shen X, Xia Esko T, Kutalik Z, Wilson JF, Joshi PK	2018	Genomic underpinnings of lifespan allow prediction and reveal basis in modern risks	http://dx.doi.org/10.7488/ds/2463	Edinburgh DataShare, 10.7488/ds/2463

The following previously published dataset was used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
GTEx Consortium	2017	Single tissue gene expression data used in the SMR-Heidi analysis	https://gtexportal.org/home/datasets	Genotype-Tissue Expression, GTEx_Analysis_v7_eQTL.tar.gz

References

Association of British Insurers and UK Government. 2014. *Concordat and Moratorium on Genetics and Insurance*: ABI.

Battle A, Brown CD, Engelhardt BE, Montgomery SB, GTEx ConsortiumLaboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working GroupStatistical Methods groups—Analysis Working GroupEnhancing GTEx (eGTEx) groupsNIH Common FundNIH/NCINIH/NHGRINIH/NIMHNIH/NIDABiospecimen Collection Source Site—NDRIBiospecimen Collection Source Site—RPCIBiospecimen Core Resource—VARIBrain Bank Repository—University of Miami Brain Endowment BankLeidos Biomedical—Project ManagementELSI StudyGenome Browser Data Integration &Visualization—EBIGenome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa CruzLead analysts:Laboratory, Data Analysis &Coordinating Center (LDACC):NIH program management:Biospecimen collection:Pathology:eQTL manuscript working group:, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis &Coordinating Center (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:.2017. Genetic effects on gene expression across human tissues. *Nature* **550**:204. DOI: https://doi.org/10.1038/nature24277, PMID: 29022597

Beekman M, Blanché H, Perola M, Hervonen A, Bezrukov V, Sikora E, Flachsbart F, Christiansen L, De Craen AJ, Kirkwood TB, Rea IM, Poulain M, Robine JM, Valensin S, Stazi MA, Passarino G, Deiana L, Gonos ES, Paternoster L, Sørensen TI, et al. 2013. Genome-wide linkage analysis for human longevity: genetics of healthy aging study. *Aging Cell* **12**:184–193. DOI: https://doi.org/10.1111/acel.12039, PMID: 23286790

Ben-Avraham D, Govindaraju DR, Budagov T, Fradin D, Durda P, Liu B, Ott S, Gutman D, Sharvit L, Kaplan R, Bougnères P, Reiner A, Shuldiner AR, Cohen P, Barzilai N, Atzmon G. 2017. The GH receptor exon 3 deletion is a marker of male-specific exceptional longevity associated with increased GH sensitivity and taller stature. *Science Advances* **3**:e1602025. DOI: https://doi.org/10.1126/sciadv.1602025, PMID: 28630896

Broer L, Buchman AS, Deelen J, Evans DS, Faul JD, Lunetta KL, Sebastiani P, Smith JA, Smith AV, Tanaka T, Yu L, Arnold AM, Aspelund T, Benjamin EJ, De Jager PL, Eirkisdottir G, Evans DA, Garcia ME, Hofman A, Kaplan RC, et al. 2015. GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. *The Journals of Gerontology: Series A* **70**:110–118. DOI: https://doi.org/10.1093/gerona/glu166

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM, Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2015. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**:291–295. DOI: https://doi.org/10.1038/ng.3211, PMID: 25642630

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, McVean G, Leslie S, Donnelly P, Marchini J. 2017. Genome-wide genetic data on ~500,000 UK biobank participants. *bioRxiv*. DOI: https://doi.org/10.1101/166298

Churchhouse C, Neale B. 2018. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK biobank. http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank [Accessed November 2, 2018].

Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, Downes K, Barrett JC, Healy BC, Mychaleckyj JC, Warram JH, Todd JA. 2008. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics* **40**:1399–1401. DOI: https://doi.org/10.1038/ng.249, PMID: 18978792

Cox DR. 1972. Regression models and Life-Tables. *Journal of the Royal Statistical Society: Series B* **34**:187–202.

D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. 2008. General cardiovascular risk profile for use in primary care: the framingham heart study. *Circulation* **117**:743–753. DOI: https://doi.org/10.1161/CIRCULATIONAHA.107.699579, PMID: 18212285

Deelen J, Beekman M, Uh HW, Helmer Q, Kuningas M, Christiansen L, Kremer D, van der Breggen R, Suchiman HE, Lakenberg N, van den Akker EB, Passtoors WM, Tiemeier H, van Heemst D, de Craen AJ, Rivadeneira F, de Geus EJ, Perola M, van der Ouderaa FJ, Gunn DA, et al. 2011. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging Cell* **10**:686–698. DOI: https://doi.org/10.1111/j.1474-9726.2011.00705.x, PMID: 21418511

Deelen J, Beekman M, Capri M, Franceschi C, Slagboom PE. 2013. Identifying the genomic determinants of aging and longevity in human population studies: progress and challenges. *BioEssays* **35**:386–396. DOI: https://doi.org/10.1002/bies.201200148, PMID: 23423909

Deelen J, Beekman M, Uh HW, Broer L, Ayers KL, Tan Q, Kamatani Y, Bennet AM, Tamm R, Trompet S, Guðbjartsson DF, Flachsbart F, Rose G, Viktorin A, Fischer K, Nygaard M, Cordell HJ, Crocco P, van den Akker EB, Böhringer S, et al. 2014. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Human Molecular Genetics* **23**:4420–4432. DOI: https://doi.org/10.1093/hmg/ddu139, PMID: 24688116

Donlon TA, Morris BJ, Chen R, Masaki KH, Allsopp RC, Willcox DC, Elliott A, Willcox BJ. 2017. FOXO3 longevity interactome on chromosome 6. *Aging Cell* **16**:1016–1025. DOI: https://doi.org/10.1111/acel.12625, PMID: 28722347

Euesden J, Lewis CM, O'Reilly PF. 2015. PRSice: polygenic risk score software. *Bioinformatics* **31**:1466–1468. DOI: https://doi.org/10.1093/bioinformatics/btu848, PMID: 25550326

Falconer DS, Mackay TFC, Frankham R. 1996. *Trends in genetics : TIG*. In: *Introduction to Quantitative Genetics*. **12** Cell Press.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, Ripke S, Day FR, Purcell S, Stahl E, Lindstrom S, Perry JR, Okada Y, Raychaudhuri S, Daly MJ, Patterson N, et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**:1228–1235. DOI: https://doi.org/10.1038/ng.3404, PMID: 26414678

Flachsbart F, Caliebe A, Kleindorp R, Blanché H, von Eller-Eberstein H, Nikolaus S, Schreiber S, Nebel A. 2009. Association of FOXO3A variation with human longevity confirmed in german centenarians. *Proceedings of the National Academy of Sciences* **106**:2700–2705. DOI: https://doi.org/10.1073/pnas.0809594106, PMID: 19196970

Flachsbart F, Dose J, Gentschew L, Geismann C, Caliebe A, Knecht C, Nygaard M, Badarinarayan N, ElSharawy A, May S, Luzius A, Torres GG, Jentzsch M, Forster M, Häsler R, Pallauf K, Lieb W, Derbois C, Galan P, Drichel D, et al. 2017. Identification and characterization of two functional variants in the human longevity gene FOXO3. *Nature Communications* **8**:2063. DOI: https://doi.org/10.1038/s41467-017-02183-y, PMID: 29234056

Grossi V, Forte G, Sanese P, Peserico A, Tezil T, Lepore Signorile M, Fasano C, Lovaglio R, Bagnulo R, Loconte DC, Susca FC, Resta N, Simone C. 2018. The longevity SNP rs2802292 uncovered: hsf1 activates stress-dependent expression of FOXO3 through an intronic enhancer. *Nucleic Acids Research* **46**:5587–5600. DOI: https://doi.org/10.1093/nar/gky331, PMID: 29733381

Haller T, Kals M, Esko T, Mägi R, Fischer K. 2015. RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Briefings in Bioinformatics* **16**:39–44. DOI: https://doi.org/10.1093/bib/bbt066, PMID: 24008273

Hayashi T, Ko JH, Strafella AP, Dagher A. 2013. Dorsolateral prefrontal and orbitofrontal cortex interactions during self-control of cigarette craving. *Proceedings of the National Academy of Sciences* **110**:4422–4427. DOI: https://doi.org/10.1073/pnas.1212185110, PMID: 23359677

Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, Tan VY, Yarmolinsky J, Shihab HA, Timpson NJ, Evans DM, Relton C, Martin RM, Davey Smith G, Gaunt TR, Haycock PC. 2018. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**:e34408. DOI: https://doi.org/10.7554/eLife.34408, PMID: 29846171

Herskind AM, McGue M, Holm NV, Sørensen TI, Harvald B, Vaupel JW. 1996. The heritability of human longevity: a population-based study of 2872 danish twin pairs born 1870-1900. *Human Genetics* **97**:319–323. DOI: https://doi.org/10.1007/BF02185763, PMID: 8786073

Hwangbo DS, Gershman B, Gersham B, Tu MP, Palmer M, Tatar M. 2004. Drosophila dFOXO controls lifespan and regulates insulin signalling in brain and fat body. *Nature* **429**:562–566. DOI: https://doi.org/10.1038/nature02549, PMID: 15175753

Joshi PK, Fischer K, Schraut KE, Campbell H, Esko T, Wilson JF. 2016. Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. *Nature Communications* **7**:11174. DOI: https://doi.org/10.1038/ncomms11174, PMID: 27029810

Joshi PK, Pirastu N, Kentistou KA, Fischer K, Hofer E, Schraut KE, Clark DW, Nutile T, Barnes CLK, Timmers P, Shen X, Gandin I, McDaid AF, Hansen TF, Gordon SD, Giulianini F, Boutin TS, Abdellaoui A, Zhao W, Medina-Gomez C, et al. 2017. Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nature Communications* **8**:910. DOI: https://doi.org/10.1038/s41467-017-00934-5, PMID: 29030599

Kaplanis J, Gordon A, Wahl M, Gershovits M, Markus B, Sheikh M, Gymrek M, Bhatia G, MarArthur DG, Price A, Erlich Y. 2018. Quantitative analysis of population-scale family trees with millions of relatives. *bioRxiv*. DOI: https://doi.org/10.1101/106427

Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R. 1993. A C. elegans mutant that lives twice as long as wild type. *Nature* **366**:461–464. DOI: https://doi.org/10.1038/366461a0, PMID: 8247153

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50**:1219–1224. DOI: https://doi.org/10.1038/s41588-018-0183-z, PMID: 30104762

Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. 2016. Fast and rigorous computation of gene and pathway scores from SNP-Based summary statistics. *PLOS Computational Biology* **12**:e1004714. DOI: https://doi.org/10.1371/journal.pcbi.1004714, PMID: 26808494

Legal General Group PLC. 2017. *Interim Management Report*: Legal and General Group.

Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, Perola M, Ng PC, Mägi R, Milani L, Fischer K, Metspalu A. 2015. Cohort profile: estonian biobank of the estonian genome center, university of tartu. *International Journal of Epidemiology* **44**:1137–1147. DOI: https://doi.org/10.1093/ije/dyt268, PMID: 24518929

Lin DY, Sullivan PF. 2009. Meta-analysis of genome-wide association studies with overlapping subjects. *The American Journal of Human Genetics* **85**:862–872. DOI: https://doi.org/10.1016/j.ajhg.2009.11.001, PMID: 20004761

Livingstone SJ, Levin D, Looker HC, Lindsay RS, Wild SH, Joss N, Leese G, Leslie P, McCrimmon RJ, Metcalfe W, McKnight JA, Morris AD, Pearson DW, Petrie JR, Philip S, Sattar NA, Traynor JP, Colhoun HM, Scottish Diabetes Research Network epidemiology group, Scottish Renal Registry. 2015. Estimated life expectancy in a scottish cohort with type 1 diabetes, 2008-2010. *Jama* **313**:37–44. DOI: https://doi.org/10.1001/jama.2014.16425, PMID: 25562264

Ljungquist B, Berg S, Lanke J, McClearn GE, Pedersen NL. 1998. The effect of genetic factors for longevity: a comparison of identical and fraternal twins in the swedish twin registry. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **53A**:M441–M446. DOI: https://doi.org/10.1093/gerona/53A.6.M441

Lloyd-Jones LR, Holloway A, McRae A, Yang J, Small K, Zhao J, Zeng B, Bakshi A, Metspalu A, Dermitzakis M, Gibson G, Spector T, Montgomery G, Esko T, Visscher PM, Powell JE. 2017. The genetic architecture of gene expression in peripheral blood. *The American Journal of Human Genetics* **100**:371. DOI: https://doi.org/10.1016/j.ajhg.2017.01.026, PMID: 28157541

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N, Price AL. 2015. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**:284–290. DOI: https://doi.org/10.1038/ng.3190, PMID: 25642633

Lowe CJ, Hall PA, Staines WR. 2014. The effects of continuous theta burst stimulation to the left dorsolateral prefrontal cortex on executive function, food cravings, and snack food consumption. *Psychosomatic Medicine* **76**:503–511. DOI: https://doi.org/10.1097/PSY.0000000000000090, PMID: 25215552

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. 2017. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research* **45**:D896–D901. DOI: https://doi.org/10.1093/nar/gkw1133, PMID: 27899670

Maciejewski H. 2014. Gene set analysis methods: statistical models and methodological differences. *Briefings in Bioinformatics* **15**:504–518. DOI: https://doi.org/10.1093/bib/bbt002, PMID: 23413432

Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, Fine RS, Lu Y, Schurmann C, Highland HM, Rüeger S, Thorleifsson G, Justice AE, Lamparter D, Stirrups KE, Turcot V, Young KL, Winkler TW, Esko T, Karaderi T, et al. 2017. Rare and low-frequency coding variants alter human adult height. *Nature* **542**:186–190. DOI: https://doi.org/10.1038/nature21039, PMID: 28146470

Mathers C, Stevens GA, Mahanani WR, Fat DM, Hogan D. 2018. *Global Health Estimates 2016: Deaths by Cause Age, Sex, by Country and by Region*: WHO.

McDaid AF, Joshi PK, Porcu E, Komljenovic A, Li H, Sorrentino V, Litovchenko M, Bevers RPJ, Rüeger S, Reymond A, Bochud M, Deplancke B, Williams RW, Robinson-Rechavi M, Paccaud F, Rousson V, Auwerx J, Wilson JF, Kutalik Z. 2017. Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nature Communications* **8**:15842. DOI: https://doi.org/10.1038/ncomms15842, PMID: 28748955

McGue M, Vaupel JW, Holm N, Harvald B. 1993. Longevity is moderately heritable in a sample of danish twins born 1870-1880. *Journal of Gerontology* **48**:B237–B244. DOI: https://doi.org/10.1093/geronj/48.6.B237, PMID: 8227991

McKay JD, Kaaks R, Johansson M, Biessy C, Wiklund F, Balter K, Adami H-O, Boillot C, Gioia-Patricola L, Canzian F, Stattin P, Gronberg H. 2007. Haplotype-Based analysis of common variation in the growth hormone receptor gene and prostate cancer risk. *Cancer Epidemiology Biomarkers & Prevention* **16**:169–173. DOI: https://doi.org/10.1158/1055-9965.EPI-06-0320

McRae A, Marioni RE, Shah S, Yang J, Powell JE, Harris SE, Gibson J, Henders AK, Bowdler L, Painter JN, Murphy L, Martin NG, Starr JM, Wray NR, Deary IJ, Visscher PM, Montgomery GW. 2017. Identification of 55, 000 replicated DNA methylation QTL. *bioRxiv*. DOI: https://doi.org/10.1101/166710

Mishra A, Macgregor S. 2015. VEGAS2: software for more flexible Gene-Based testing. *Twin Research and Human Genetics* **18**:86–91. DOI: https://doi.org/10.1017/thg.2014.79, PMID: 25518859

Mishra A, MacGregor S. 2017. A novel approach for pathway analysis of GWAS data highlights role of BMP signaling and muscle cell differentiation in colorectal cancer susceptibility. *Twin Research and Human Genetics* **20**:1–9. DOI: https://doi.org/10.1017/thg.2016.100, PMID: 28105966

Morrison JH, Baxter MG. 2012. The ageing cortical synapse: hallmarks and implications for cognitive decline. *Nature Reviews Neuroscience* **13**:240–250. DOI: https://doi.org/10.1038/nrn3200, PMID: 22395804

Myllykangas-Luosujärvi R, Aho K, Kautiainen H, Isomäki H. 1995. Shortening of life span and causes of excess mortality in a population-based series of subjects with rheumatoid arthritis. *Clinical and Experimental Rheumatology* **13**:149–153. PMID: 7656460

Ning Z, Lee Y, Joshi PK, Wilson JF, Pawitan Y, Shen X. 2017. A selection operator for summary association statistics reveals allelic heterogeneity of complex traits. *The American Journal of Human Genetics* **101**:903–912. DOI: https://doi.org/10.1016/j.ajhg.2017.09.027, PMID: 29198721

Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, Lui JC, Vedantam S, Gustafsson S, Esko T, Frayling T, Speliotes EK, Boehnke M, Raychaudhuri S, Fehrmann RS, Hirschhorn JN, Franke L, Genetic Investigation of ANthropometric Traits (GIANT) Consortium. 2015. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications* **6**:5890. DOI: https://doi.org/10.1038/ncomms6890, PMID: 25597830

Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, Reinmaa E, Sutphin GL, Zhernakova A, Schramm K, Wilson YA, Kobes S, Tukiainen T, Ramos YF, Göring HH, Fornage M, Liu Y, Gharib SA, Stranger BE, De Jager PL, et al. 2015. The transcriptional landscape of age in human peripheral blood. *Nature Communications* **6**:8570. DOI: https://doi.org/10.1038/ncomms9570, PMID: 26490707

Pilling LC, Atkins JL, Bowman K, Jones SE, Tyrrell J, Beaumont RN, Ruth KS, Tuke MA, Yaghootkar H, Wood AR, Freathy RM, Murray A, Weedon MN, Xue L, Lunetta K, Murabito JM, Harries LW, Robine JM, Brayne C, Kuchel GA, et al. 2016. Human longevity is influenced by many genetic variants: evidence from 75,000 UK biobank participants. *Aging* **8**:547–560. DOI: https://doi.org/10.18632/aging.100930, PMID: 27015805

Pilling LC, Kuo CL, Sicinski K, Tamosauskaite J, Kuchel GA, Harries LW, Herd P, Wallace R, Ferrucci L, Melzer D. 2017. Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging* **9**:2504–2520. DOI: https://doi.org/10.18632/aging.101334, PMID: 29227965

Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, Gianniny L, Korman BD, Padyukov L, Kurreeman FA, Chang M, Catanese JJ, Ding B, Wong S, van der Helm-van Mil AH, Neale BM, Coblyn J, Cui J, Tak PP, Wolbink GJ, et al. 2008. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nature Genetics* **40**:1216–1223. DOI: https://doi.org/10.1038/ng.233, PMID: 18794853

Ruby JG, Wright KM, Rand KA, Kermany A, Noto K, Curtis D, Varner N, Garrigan D, Slinkov D, Dorfman I, Granka JM, Byrnes J, Myres N, Ball C. 2018. Estimates of the heritability of human longevity are substantially inflated due to assortative mating. *Genetics* **210**:1109–1124. DOI: https://doi.org/10.1534/genetics.118.301613, PMID: 30401766

Sebastiani P, Solovieff N, Dewan AT, Walsh KM, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wilk JB, Myers RH, Steinberg MH, Montano M, Baldwin CT, Hoh J, Perls TT. 2012. Genetic signatures of exceptional longevity in humans. *PLoS ONE* **7**:e29848. DOI: https://doi.org/10.1371/journal.pone.0029848, PMID: 22279548

Sebastiani P, Sun FX, Andersen SL, Lee JH, Wojczynski MK, Sanders JL, Yashin A, Newman AB, Perls TT. 2013. Families enriched for exceptional longevity also have increased Health-Span: findings from the long life family study. *Frontiers in Public Health* **1**:38. DOI: https://doi.org/10.3389/fpubh.2013.00038, PMID: 24350207

Sebastiani P, Nussbaum L, Andersen SL, Black MJ, Perls TT. 2016. Increasing sibling relative risk of survival to older and older ages and the importance of precise definitions of "Aging," "Life Span," and "Longevity". *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **71**:340–346. DOI: https://doi.org/10.1093/gerona/glv020

Sebastiani P, Gurinovich A, Bae H, Andersen S, Malovini A, Atzmon G, Villa F, Kraja AT, Ben-Avraham D, Barzilai N, Puca A, Perls TT. 2017. Four Genome-Wide association studies identify new extreme longevity variants. *The Journals of Gerontology: Series A* **72**:1453–1464. DOI: https://doi.org/10.1093/gerona/glx027

Shen X, Wang X, Ning Z, Tsepilov Y, Shirali M, Smith BH, Hocking LJ, Padmanabhan S, Hayward C, Porteous DJ, Pawitan Y, Haley CS, Aulchenko YS, Generation Scotland. 2015. Simple multi-trait analysis identifies novel loci associated with growth and obesity measures. *bioRXiv*. DOI: https://doi.org/10.1101/022269

Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, Paul DS, Freitag D, Burgess S, Danesh J, Young R, Butterworth AS. 2016. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**:3207–3209. DOI: https://doi.org/10.1093/bioinformatics/btw373, PMID: 27318201

Therneau TM, Grambsch PM, Fleming TR. 1990. Martingale-based residuals for survival models. *Biometrika* **77**:147–160. DOI: https://doi.org/10.1093/biomet/77.1.147

van den Berg N, Beekman M, Smith KR, Janssens A, Slagboom PE. 2017. Historical demography and longevity genetics: back to the future. *Ageing Research Reviews* **38**:28–39. DOI: https://doi.org/10.1016/j.arr.2017.06.005, PMID: 28689042

Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Kasela S, Pervjakova N, Alvaes I, Fave MJ, Agbessi M, Christiansen M, Jansen R, Seppälä I, Tong L, Teumer A, Schramm K, et al. 2018. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*. DOI: https://doi.org/10.1101/447367

Wacholder S, Hartge P, Struewing JP, Pee D, McAdams M, Brody L, Tucker M. 1998. The kin-cohort study for estimating penetrance. *American Journal of Epidemiology* **148**:623–630. DOI: https://doi.org/10.1093/aje/148.7.623, PMID: 9778168

Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JE, Zhernakova A, Zhernakova DV, Veldink JH, Van den Berg LH, Karjalainen J, Withoff S, Uitterlinden AG, Hofman A, Rivadeneira F, Hoen PAC', et al. 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics* **45**:1238–1243. DOI: https://doi.org/10.1038/ng.2756, PMID: 24013639

Williams GC. 1957. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* **11**:398–411.

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, Amin N, Buchkovich ML, Croteau-Chonka DC, Day FR, Duan Y, Fall T, Fehrmann R, Ferreira T, Jackson AU, Karjalainen J, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**:1173–1186. DOI: https://doi.org/10.1038/ng.3097, PMID: 25282103

Young AI, Frigge ML, Gudbjartsson DF, Thorleifsson G, Bjornsdottir G, Sulem P, Masson G, Thorsteinsdottir U, Stefansson K, Kong A. 2018. Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics* **50**:1304–1310. DOI: https://doi.org/10.1038/s41588-018-0178-9, PMID: 30104764

Zarrouk A, Vejux A, Mackrill J, O'Callaghan Y, Hammami M, O'Brien N, Lizard G. 2014. Involvement of oxysterols in age-related diseases and ageing processes. *Ageing Research Reviews* **18**:148–162. DOI: https://doi.org/10.1016/j.arr.2014.09.006, PMID: 25305550

Zeng Y, Nie C, Min J, Liu X, Li M, Chen H, Xu H, Wang M, Ni T, Li Y, Yan H, Zhang JP, Song C, Chi LQ, Wang HM, Dong J, Zheng GY, Lin L, Qian F, Qi Y, et al. 2016. Novel loci and pathways significantly associated with longevity. *Scientific Reports* **6**:21243. DOI: https://doi.org/10.1038/srep21243, PMID: 26912274

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, Yang J. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**:481–487. DOI: https://doi.org/10.1038/ng.3538, PMID: 27019110

Appendix B

bGWAS: an R package to perform Bayesian genome wide association studies

This article (Mounier & Kutalik, 2020) is presented in Chapter 1.

Bioinformatics, 36(15), 2020, 4374–4376
doi: 10.1093/bioinformatics/btaa549
Advance Access Publication Date: 29 May 2020
Applications Note

Genetics and population analysis

bGWAS: an R package to perform Bayesian genome wide association studies

Ninon Mounier [1,2] **and Zoltán Kutalik** [1,2,*]

[1]Department of Training, Research and Innovation, University Center for Primary Care and Public Health, Lausanne 1010, Switzerland and [2]Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on December 6, 2019; revised on May 18, 2020; editorial decision on May 19, 2020; accepted on May 25, 2020

Abstract

Summary: Increasing sample size is not the only strategy to improve discovery in Genome Wide Association Studies (GWASs) and we propose here an approach that leverages published studies of related traits to improve inference. Our Bayesian GWAS method derives informative prior effects by leveraging GWASs of related risk factors and their causal effect estimates on the focal trait using multivariable Mendelian randomization. These prior effects are combined with the observed effects to yield Bayes Factors, posterior and direct effects. The approach not only increases power, but also has the potential to dissect direct and indirect biological mechanisms.

Availability and implementation: bGWAS package is freely available under a GPL-2 License, and can be accessed, alongside with user guides and tutorials, from https://github.com/n-mounier/bGWAS.

Contact: zoltan.kutalik@unil.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the last decade, Genome Wide Association Studies (GWASs) have been widely used to identify genetic variants, usually single nucleotide polymorphisms (SNPs), associated with complex traits. These GWASs led to a large number of discoveries, helping to better understand the underlying biology of the studied traits (Visscher *et al.*, 2017). However, large sample sizes (typically > 1 million) are needed to achieve sufficient power to identify SNPs with small to moderate effects.

Besides ever-increasing sample sizes one can borrow strength from studies of related traits or risk factors (RFs). To leverage this information, several methods have already been published, such as MTAG (Turley *et al.*, 2018) or GenomicSEM (Grotzinger *et al.*, 2019) for example and we developed a Bayesian GWAS approach, first described by McDaid *et al.* (2017). The aim of our approach is to increase power by comparing the observed Z-statistics from the focal phenotype (representing association strength) to prior effects using Bayes factors (BFs) and computing the corresponding *P*-values. Prior effects are estimated from publicly available GWASs for RFs showing a significant multivariable causal effect (similar to Sanderson *et al.*, 2019) on the focal phenotype, established by Mendelian randomization (MR). Such approach has previously been used to identify new loci associated with lifespan (McDaid *et al.*, 2017; Timmers *et al.*, 2019).

Here, we present substantial improvements to the method and its implementation in an R package bGWAS. We optimized the causal effect estimation and improved the step-wise selection approach used to identify relevant RFs. We derived and implemented a fast analytical approach to accurately estimate BF *P*-values. Notably, the method now also provides posterior- and direct effect estimates (not acting through the RFs) that can be used for downstream analyses.

2 Materials and methods

The approach consists of five main steps: (i) Identification of relevant RFs, (ii) Out-of-sample estimation of prior effects, (iii) Computation of BFs and their respective *P*-values, (iv) Estimation of posterior and direct effects, (v) Extraction and visualization of the results (Supplementary Fig. S1).

2.1 Identification of relevant RFs

In the first step we identify relevant RFs to build the prior. The package currently includes 38 publicly available GWASs, which can easily be modified to include additional RFs. Using the package, all available RFs can be displayed (list_priorGWASs()) and an arbitrary subset can be selected (select_priorGWASs()). First, RFs with non-significant ($P > 0.05$) univariable causal effects are removed. Then, a step-wise selection approach applied to multivariable MR models identifies RFs that are jointly affecting the phenotype. Since Akaike information criterion (AIC)-based model comparison assumes equal

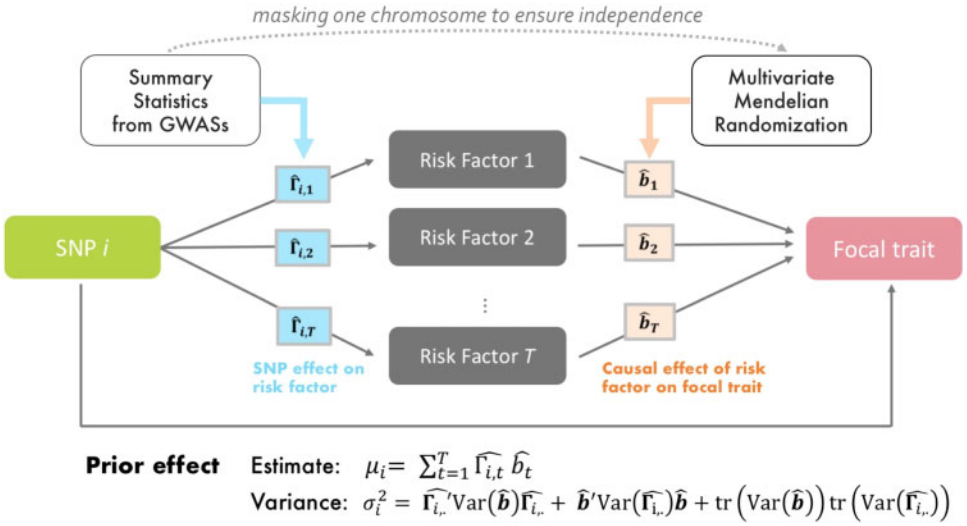


Fig. 1. Prior estimation design. For each SNP i, its prior effect on the focal trait is calculated as the product of the effect of SNP i on the RF t ($\hat{\Gamma}_{i,t}$) and the causal effect of RF t on the focal trait ($\hat{b}_t$, estimated using multivariable MR), summed over all T RFs identified in the step-wise selection approach. In our implementation, we use $\mathbf{I}_\mathbf{T}$, a T×T identity matrix, as an approximation of $\mathrm{Var}(\hat{\Gamma}_{i,\cdot})$ to estimate σ_i^2. Adapted from McDaid *et al.* (2017)

number of observations (instruments) in the compared models (which is not the case for two-sample MR), we rather implemented a *P*-value-based step-wise selection approach to identify all the RFs that have a significant multivariable causal effect on our focal phenotype (see Supplementary Section S1).

2.2 Out-of-sample estimation of prior effects

After identifying the RFs, prior effect estimates (μ) and standard errors σ are calculated for each SNP by multiplying SNP-RF effects with RF-trait causal effect estimates. To ensure that priors are independent of the observed association for SNP i, we estimate multivariable causal effects based on SNPs that do not lie on the same chromosome as SNP i (Fig. 1). Shrinking SNP-RF effects before estimating the prior leads to poorer priors (Supplementary Fig. S2).

2.3 Computation of BFs and their respective *P*-values

We use BFs to quantify the evidence in favor of the prior by comparing two competing hypotheses. Both our null and our alternative hypotheses are assuming that for a SNP i, the observed Z-statistic z_i is following a normal distribution. Under H_0, this distribution is centered on zero and has a variance of 1, whereas under H_1, the distribution is centered on μ_i and has a variance of σ_i^2 (prior parameters). The BFs can be derived in closed form (Equation 1) (Murphy, 2007).

$$\mathrm{BF}_i = \mathrm{BF}(z_i; \mu_i : \sigma_i) = \frac{L(z_i\,;\,\mu_i\,;\,1+\sigma_i^2)}{L(z_i\,;\,0\,;\,1)} \tag{1}$$

with $L(z\,;\,\mu\,;\,\sigma^2)$: the density of z under the corresponding Gaussian distribution.

Since BF alone does not readily control type I error rate, we also compute a corresponding *P*-value. The *P*-value $p_{\mathrm{BF}-i}$ represents the probability of observing any null BF (obtained for standard normal Z statistics and the same genome-wide priors) larger than the observed BF_i. We have now analytically derived the *P*-values and sample only certain percentiles of the null BF distribution (see Supplementary Section S2), yielding highly concordant *P*-value estimates with the ones from the (>8-times slower) gold-standard permutation approach (Supplementary Fig. S3).

2.4 Estimation of posterior and direct effects

The posterior effect μ'_{p-i} and posterior standard error σ_{p-i} can be easily derived for each SNP i (Equation 2) (Murphy, 2007).

$$\mu_{p-i} = \frac{\sigma_i^2}{\sigma_i^2+1}\left(\frac{\mu_i}{\sigma_i^2}+z_i\right) \quad \text{and} \quad \sigma_{p-i} = \sqrt{\frac{\sigma_i^2}{\sigma_i^2+1}} \tag{2}$$

We define the direct effect μ_{d-i} (and its standard error σ_{d-i}) as the part of the observed effect that is not mediated through the RFs and hence cannot be explained by the prior (Equation 3).

$$\mu_{d-i} = z_i - \mu_i \quad \text{and} \quad \sigma_{d-i} = \sqrt{\sigma_i^2+1} \tag{3}$$

Analogous formulae based on observed effect sizes and standard errors (instead of Z statistics) are implemented and provided in Supplementary Section S3.

2.5 Extraction and visualization of the results

We implemented dedicated functions in the bGWAS package to list, visualize and interpret the results (Supplementary Fig. S1). FDR threshold and SNP-pruning stringency can be set in the bGWAS() function to produce a final list of associated markers. Summary statistics (BFs, prior, posterior and direct effect) can be extracted from the returned *bGWAS* object using the extract_results_bGWAS() function. RFs causal effects can be obtained using extract_MRcoeffs_bGWAS() or visualized using the coefficients_plot_bGWAS() function. manhattan_plot_bGWAS() automatically creates a Manhattan plot and heatmap_bGWAS() illustrates through which RFs SNPs are exerting their (prior) effects on the focal phenotype.

3 Application to lifespan

In order to see how the improved method [implemented in the bGWAS R-package (v1.0.2)] compares to the original one (McDaid *et al.*, 2017), we applied both to the summary statistics from a GWAS on lifespan (Timmers *et al.*, 2019), which already included the latter application. A full description of the analysis and the results are available in Supplementary Section S4.

In the new analysis, we identified five RFs with a significant causal effect on lifespan (Supplementary Fig. S5): years of schooling, LDL cholesterol, diastolic blood pressure, coronary artery disease and body mass index. The priors obtained from the improved method are more informative: the squared correlation between prior and observed effects has improved to 0.377 from 0.082. In the new analysis, 28 SNPs reached genome-wide significance ($p_{\mathrm{BF}} < 5e^{-8}$). Among these variants, 15 are not identified by the conventional GWAS at the same threshold and 11 of them have never been

reported in any previous lifespan GWAS (Supplementary Table S1). Four of the seven SNPs identified based on the old BF *P*-value in Timmers *et al.* (2019) are confirmed in the new analysis, and three have low prior effects (Supplementary Table S2) due to the change in RFs. We identified nine additional loci with significant posterior effect ($p_p < 5e^{-8}$) (Supplementary Table S3) and four loci with significant direct effect ($p_d < 5e^{-8}$) (Supplementary Table S4), including the highly pleiotropic APOE locus, which might be acting on lifespan through RFs not included in the analysis (e.g. Alzheimer's disease) (Belloy *et al.*, 2019).

4 Conclusion

Leveraging information from related traits is an efficient approach to increase the power of GWAS of complex traits, which is now fully implemented in the bGWAS R package. Through an application to lifespan GWAS, we have demonstrated that this approach could lead to meaningful new discoveries in lifespan genetics and dissect direct from indirect mechanisms.

Acknowledgements

The authors thank Nicola Pirastu for the helpful discussions, as well as Jonathan Sulc and Sina Rüeger for the helpful comments on the documentation.

Funding

This work was supported by the Swiss National Science Foundation [31003A-143914] [310030-189147].

Conflict of Interest: none declared.

References

Belloy,M.E. *et al.* (2019) A quarter century of APOE and Alzheimer's disease: progress to date and the path forward. *Neuron*, **101**, 820–838.

Grotzinger,A.D. *et al.* (2019) Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.*, **3**, 513–525.

McDaid,A.F. *et al.* (2017) Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nat. Commun.*, **8**, 15842.

Murphy,K.P. (2007) Conjugate Bayesian analysis of the Gaussian distribution. *Technical report*.

Sanderson,E. *et al.* (2019) An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int. J. Epidemiol.*, **48**, 713–727.

Timmers,P.R. *et al.* (2019) Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife*, **8**, e39856.

Turley,P. *et al.* (2018) Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.*, **50**, 229–237.

Visscher,P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.

Supplementary Material

bGWAS: an R package to perform Bayesian Genome Wide Association Studies.

1 - Stepwise selection approach

In the initial approach described in McDaid et al. (2017), a stepwise selection using Akaike information criterion (AIC) was performed to identify the best model. By comparing the different models using AIC measures, risk factors (RFs) could be either added or removed. AIC assumes an equal number of observations in the models compared, but here, each model includes a specific subset of RFs and hence a different number of corresponding instruments (i.e. observations). The more RFs are included in a model, the more instruments are likely to be used and therefore AIC is not suited to compare the different models.

We implemented a forward-backward stepwise selection approach, based only on p-values. Instead of using AIC to compare models, we compare the p-values of each RF when added to the model. We start with a model containing only the RF having the strongest univariable causal effect on our focal phenotype and try to add the remaining RFs (the ones with a nominally significant univariable causal effect) one by one. At each step, we identify the RF showing the most significant effect when added. If the corresponding p-value is below a specific user-specified threshold (by default Bonferroni corrected p-value<0.05, accounting for the number of tested risk factors) the RF is added to the multivariable model. Then, the multivariable linear regression p-values of all currently included RFs are examined. If at least one p-value becomes larger that the threshold, the RF with the least significant p-value is removed from the model. This procedure is repeated until convergence, i.e. when none of the excluded RFs is significant when added to the multivariable model and all included RFs have a significant multivariable causal effect on our focal phenotype.

2 - P-value derivation and approximation

Let z_i be the observed z-statistic, μ_i and σ_i^2 be the prior effect estimate and variance for SNP i. BF_i can be defined as:

$$\text{BF}_i = \text{BF}(z_i\ ;\ \mu_i : \sigma_i\ ) = \frac{L(z_i\ ;\ \mu_i\ ;\ 1+\sigma_i^2)}{L(z_i\ ;\ 0\ ;\ 1)}$$

with $L(z\ ;\ \mu\ ;\ \sigma^2)$: the density of z under the corresponding Gaussian distribution.

The p-value p_{BF-i} can be analytically derived using a polynomial of z of degree 2 (Equations 1, 2 and 3).

$$p_{BF-i} = \mathbf{Pr}(\text{any BF} > \text{BF}_i) = \frac{\left(\sum_k \mathbf{Pr}(\text{BF}(z,\ \mu_k\ ,\sigma_k) > \text{BF}_i)\right)}{M_{SNPs}} \quad \text{for any k} \in [1 : M_{SNPs}] \tag{1}$$

with M_{SNPs} : the number of SNPs in the dataset.

$$\begin{aligned}
\mathbf{Pr}(\text{BF}(z,\ \mu_k\ ,\sigma_k) > \text{BF}_i) &= \mathbf{Pr}\left(\frac{1}{\sqrt{1+\sigma_k^2}} \cdot \exp\left(z^2 \cdot \left(\frac{-1}{2(1+\sigma_k^2)} + \frac{1}{2}\right) + z\frac{\mu_k}{1+\sigma_k^2} + \frac{-\mu_k^2}{2(1+\sigma_k^2)}\right) > \text{BF}_i\right) \\
&= \mathbf{Pr}\left(z^2 \cdot \left(\frac{-1}{2(1+\sigma_k^2)} + \frac{1}{2}\right) + z\frac{\mu_k}{1+\sigma_{\mu_k}^2} + \left(\frac{-\mu_k^2}{2(1+\sigma_k^2)} - log\big(\text{BF}_i \cdot \sqrt{1+\sigma_k^2}\big)\right) > 0\right) \\
&= \mathbf{Pr}\Big(\text{Pol}(z \mid \mu_k,\ \sigma_k^2) > 0\Big)
\end{aligned} \tag{2}$$

$$\begin{aligned}
&\text{with Pol}(z \mid \mu_k,\ \sigma_k^2) = a_k \cdot z^2 + b_k \cdot z + c_k \\
&\text{and } a_k = \frac{-1}{2(1+\sigma_k^2)} + \frac{1}{2},\ b_k = \frac{\mu_k}{1+\sigma_k^2},\ c_k = \frac{-\mu_k^2}{2(1+\sigma_{\mu_k}^2)} - log\big(\text{BF}_i \cdot \sqrt{1+\sigma_k^2}\big)
\end{aligned} \tag{3}$$

Let Δ_k be the discriminant and $x_{1k,2k}$ the roots of $Pol(z|\mu_k, \sigma_k^2)$ (Equation 4). It can be shown that $a_k > 0$ and therefore the inequality $Pol(z|\mu_k, \sigma_k^2) > 0$ is true for $z < x_{1k}$ and $z > x_{2k}$. In cases when $\Delta_k < 0$ (given that the coefficient of the quadratic term is positive) the equality holds for all values of z, hence the probability equals to 1. Therefore, we only need to estimate $\mathbf{Pr}\big(z < x_{1k} \cup z > x_{2k}\big)$ when $\Delta_k \geq 0$. We know that under the null hypothesis $z \sim \mathcal{N}(0,1)$. By using Φ the Gaussian cumulative distribution function, each term of p_{BF-i} can be analytically calculated (Equation 5).

$$\begin{aligned}
\Delta_k &= b_k^2 - 4a_kc_k \\
x_{1k,2k} &= \frac{-b_k \pm \sqrt{\Delta_k}}{2a_k}
\end{aligned} \tag{4}$$

$$\begin{aligned}\mathbf{Pr}(\mathrm{BF}(z,\,\mu_k\,,\sigma_k) > \mathrm{BF}_i) &= \mathbf{Pr}\Big(z < x_{1k} \cup z > x_{2k}\Big) \\ &= \mathbf{Pr}\Big(z < x_{1k}\Big) + \mathbf{Pr}\Big(z > x_{2k}\Big) \\ &= \Phi(x_{1k}) + 1 - \Phi(x_{2k})\end{aligned} \tag{5}$$

However, getting the p-value for a single SNP using this formula would require to solve the polynomial inequality for each SNP k. If $M_{SNPs} = 1000000$, getting the p-values for all SNPs would then require the estimation of 10^{12} probabilities, which would not be less time-consuming than a permutation approach. To overcome this limitation, we decided to use an approximation to estimate the p-values. Instead of solving the polynomial inequality for each SNP k, we solved it for a set of non-linear quantiles defined as $\{0\,,\ \frac{10^{[-10\text{ to }0\text{ by }0.1]}}{2}\ ,\ 1\text{-}\frac{10^{[0\text{ to }-9.9\text{ by }-0.1]}}{2}\ ,\ 1\}$ of μ and a single value of σ^2. Quantiles of μ are estimated from the full set of prior effects estimates, and the mean value of the prior effects standard errors is used for σ. We also ensured that the subset of SNPs identified as significant was not influenced by the approximation, by using the full formula for the SNPs that are close to the significance threshold. The true p-values calculated from the exact formula are compared to the approximated ones and used if any significance discrepancy is observed. To do so, SNPs are ordered based on their BF value. The p-value of the first SNP reaching significance threshold (lowest BF value among significant SNPs) is re-estimated using the exact formula. If the true p-value also reaches significance threshold, the approximated value is kept, and no further SNP above significance are examined. If the true p-value for this SNP does not reach significance threshold, the exact formula is used and the same procedure is repeated considering the next lowest BF value(s) until observing an agreement in significance (i.e. both being significant) between the approximated and the true p-values. A similar approach is applied to SNPs not reaching significance threshold, the exact formula is used to re-estimate the p-value of the first non-significant SNP (largest BF value among non-significant SNPs). If the true p-value reaches significance threshold, the exact formula is used, and the same procedure is repeated considering the next largest BF value(s) until observing an agreement in significance between the approximated and the true p-values This procedure relies on the fact that both the approximation and the exact formula are monotonous. Therefore, there is no need to re-estimate all p-values because if a SNP reaches significance level, a SNP with a larger BF will also be significant and vice versa.

As seen on Figure S3, the approximation is slightly liberal and tends to slightly overestimate the -log10 p-values compared to the permutation approach. Near the genome wide significance threshold, some p-values are re-estimated using the exact formula (points highlighted in red) creating two parallel lines. Note however that this overestimation is not more serious than the discrepancy between the permutation-based approximation and the exact P-value.

3 - Rescaling prior, posterior and direct effects

Let $\boldsymbol{\beta}$, $\boldsymbol{se}$ and $\boldsymbol{z}$ be the observed GWAS summary statitics, with:

$$\beta_i : \text{observed effect size for SNP } i$$

$$se_i : \text{observed standard error for SNP } i$$

$$z_i : \text{observed z-statistic for SNP } i \text{ with } z_i = \frac{\beta_i}{se_i}$$

The prior, posterior and direct effects described in our Bayesian approach are estimated using $\boldsymbol{z}$. Therefore, these effects can not directly be compared to the observed effect size $\boldsymbol{\beta}$ from the GWAS. Here, we derive $\boldsymbol{\beta'}$, $\boldsymbol{\beta_p}$, $\boldsymbol{\beta_d}$ and $\boldsymbol{se'}$, $\boldsymbol{se_p}$, $\boldsymbol{se_d}$, the rescaled effect sizes and standard errors for prior, posterior and direct effects respectively.

Prior effects

Let define:

$$\mu_i : \text{prior effect estimate, on the z-statistics scale, for SNP } i$$

$$\sigma_i^2 : \text{prior effect variance, on the z-statistics scale, for SNP } i$$

The rescaled effect sizes and standard errors, $\boldsymbol{\beta_{prior}}$ and $\boldsymbol{se_{prior}}$, can be estimated as described in Equation 6.

$$\begin{aligned} \beta'_i &= \frac{\beta_i}{z_i} \cdot \mu_i = se_i \cdot \mu_i \\ se'_i &= \frac{\beta_i}{z_i} \cdot \sigma_i = se_i \cdot \sigma_i \end{aligned} \tag{6}$$

Posterior effects

Let define:

$$\mu_{p-i} : \text{posterior effect estimate, on the z-statistics scale, for SNP } i$$

$$\sigma_{p-i}^2 : \text{posterior effect variance, on the z-statistics scale, for SNP } i$$

The rescaled effect sizes and standard errors, $\boldsymbol{\beta_p}$ and $\boldsymbol{se_p}$, can be estimated as described in Equation 7.

$$\begin{aligned} \beta_{p-i} &= \frac{\beta_i}{z_i} \cdot \mu_{p-i} = se_i \cdot \mu_{p-i} \\ se_{p-i} &= \frac{\beta_i}{z_i} \cdot \sigma_{p-i} = se_i \cdot \sigma_{p-i} \end{aligned} \tag{7}$$

Direct effects

Let define:

μ_{d-i} : direct effect estimate, on the z-statistics scale, for SNP i

σ^2_{d-i} : direct effect variance, on the z-statistics scale, for SNP i

The rescaled effect sizes and standard errors, $\boldsymbol{\beta_d}$ and $\boldsymbol{se_d}$, can be estimated as described in Equation 8.

$$\begin{aligned} \beta_{d-i} &= \frac{\beta_i}{z_i} \cdot \mu_{d-i} = se_i \cdot \mu_{d-i} \\ se_{d-i} &= \frac{\beta_i}{z_i} \cdot \sigma_{d-i} = se_i \cdot \sigma_{d-i} \end{aligned} \tag{8}$$

4 - Application to Lifespan

In order to see how the improvements made to the method would affect the results, we decided to apply our approach to lifespan, using data from Timmers et al. (2019). In Timmers et al. (2019), we already performed a Bayesian GWAS, using the method described in McDaid et al. (2017), leading to the identification of 7 new variants. Here, we used the same GWAS summary statistics and the improved method implemented in the bGWAS R-package (v1.0.2), with default parameters, to perform a new analysis. The source code and more detailed results are available on GitHub (https://github.com/n-mounier/bGWAS/blob/master/doc/Lifespan_Analysis.md).

Risk Factors used

First, we compared the risk factors (RFs) selected to build the prior. In Timmers et al. (2019), 16 RFs were used whereas only 5 were included in the new analysis (Supp Fig 5): years of schooling (multivariable effect: 0.167 - SE: 0.0246), LDL cholesterol (multivariable effect: -0.0905 - SE: 0.0155), diastolic blood pressure (multivariate effect: -0.243 - SE: 0.056), coronary artery disease (multivariable effect: -0.301 - SE: 0..0493) and body mass index (multivariable effect: -0.147 - SE: 0.0201). This is explained by the different stepwise selection approach, but also by the fact that, now, only RFs reaching nominal significance in the univariable MR model are considered to be added in the multivariate model (for example, "College Completion" - univariable p-value = 0.19 and "HDL Cholesterol" - univariable p-value = 0.06 were not tested in our new analysis). Additionally, using default parameters a minimal number of 3 instruments is required for each RF to be considered, explaining why smoking traits are not used to create the prior in our new analysis.

Results - Bayes Factors

In our new analysis, we identified 28 SNPs acting through the selected RFs ($p_{BF} < 5e-8$). 15 of the 28 genome-wide significant loci are missed by the conventional GWAS, using same p-value threshold of 5e-8 to assess significance and are described in Table S1. 4 of them (near CELSR2, TMEM18, ZC3HC1 and ABO) were already significant in Timmers et al. (2019) bGWAS results. The 11 other variants identified in this analysis (near IL6R, BCL11A, SLC4A7, TRAIP, MIR2113, PINX1, TNKS, BNC2, CUX2, PDE3A and PGPEP1) are reported to be associated with lifespan for the first time.

Among the 15 new variants, 8 are associated with at least one of the RFs (from the summary statistics used to create the prior) :

- the variant near CELSR2 is associated with LDL cholesterol,

- the variants near TMEM18 and PGPEP1 are associated with body mass index,

- the variants near BCL11A, TRAIP and MIR2113 are associated with years of schooling,

- the variant near ZC3HC1 is associated with coronary artery disease,

- the variant near ABO is associated with both LDL cholesterol and coronary artery disease.

Using the GWAS Catalog to further investigate the remaining variants and their neighbouring regions, we found that in more recent studies variants in LD with the variant identified near IL6R have been associated with coronary artery disease (Harst and Verweij 2018), variants in LD with the variants identified near SLC4A7, PINX1 and TNKS have been associated with Diastolic Blood Pressure (Wain et al. 2017; Feitosa et al. 2018). The other loci (near BNC2, CUX2 and PDE3A) have not been associated with any of the risk factors, and are likely acting on lifespan through moderate effects on several risk factors (pleiotropic effects).

We also further compared our results with the ones from Timmers et al. (2019) (see Table S2). We showed that prior effects estimate for SNPs identified by at least one of the two analyses are highly correlated (Pearson's correlation, 0.828), but overall the prior effects from the new analysis are more informative. Using the improved method, the squared correlation between prior and observed effects (among SNPs with some evidence for effect on lifespan ($p_{obs} < 0.001$)) has improved to 0.377 from 0.082 in the previous version. Small differences in terms of p-values are likely to be explained by the different BF distributions implied by the prior effects differences. Moreover, the 4 variants identified in Timmers et al. (2019) that did not replicate in our new analyses all show smaller prior effect in our analyses, probably because of pleiotropic effects on some RFs that are not included in our new analysis.

Results - Posterior Effects

28 SNPs were exhibiting significant ($p_p < 5e-8$) posterior effects. 9 of them (near USP4, GNPDA2, MAD1L1, SPR1, HNF1A, ATXN2L, FTO, UBE2Z, RMC1) were missed by the conventional GWAS and by the identification based on BFs, using same p-value threshold of 5e-8 to assess significance and are described in Table S3. All of these 9 variants are known to be associated with at least one of the RFs used to create the prior, explaining their large prior and hence posterior effects.

Results - Direct Effects

Finally, we identified 4 SNPs (near APOE, HYKK, LPA and RAD52) having significant direct effects ($p_d < 5e-8$) (see Table S4). Amongst these variants, 3 are likely to act through RFs that were not included in our subset of RFs and therefore not used to create the prior. The variants near APOE and HYKK both have very low prior effect estimates. APOE locus is likely affecting lifespan through its effects on Alzheimer's disease (Nazarian, Yashin, and Kulminski 2019) for example whereas the HYKK region is known to be

associated with smoking, pulmonary diseases and cancers (Furberg et al. 2010; Hobbs et al. 2017; McKay et al. 2017), that could also influence lifespan. The variant near LPA however does have a moderate prior effect, but not strong enough to entirely explain its effect on lifespan. Its effects on lipoprotein levels (Mack et al. 2017) for example could also influence lifespan. The variant near RAD52, however has a quite strong effect on lifespan in the conventional GWAS but a small prior effect in the other direction. It means that this variant is not acting on lifespan through the RFs used to create the prior. Interestingly, there is no strong association reported for this region, and the discrepancy between observed and prior effects could be due to some direct effect on lifespan.

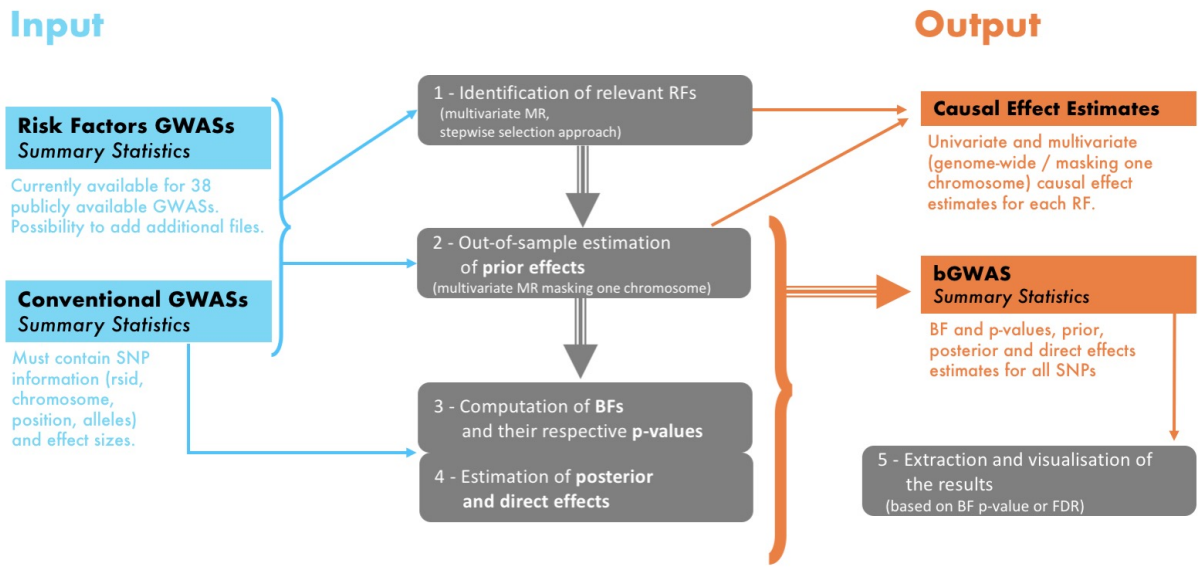


Figure S1: **bGWAS Workflow.**
To perform an analysis with `bGWAS`, two inputs are needed: GWAS Summary Statistics for potential risks factors (RFs - currently available for 38 publicly available studies, but additional studies can easily be added) and GWAS summary statistics for the conventional GWAS to be tested. These two sets of summary statistics are used in step 1) to identify the relevant RFs using a multivariable step-wise Mendelian Randomisation (MR) approach, and in step 2) to estimate the prior effects (multivariable MR model masking one chromosome to ensure independance). These two MR steps allow us to obtain the causal effect estimates of each risk factor on the trait of interest. By combining observed effects from the conventional GWAS and prior effects, Bayes Factors and their corresponding p-values are estimated in step 3), as well as posterior and direct effects in step 4). All these values are part of the bGWAS Summary Statistics that are returned by the main `bGWAS()` function. Finally, associated SNPs reaching the user-specified significance threshold (p-value or false discovery rate (FDR)) are identified in step 5).

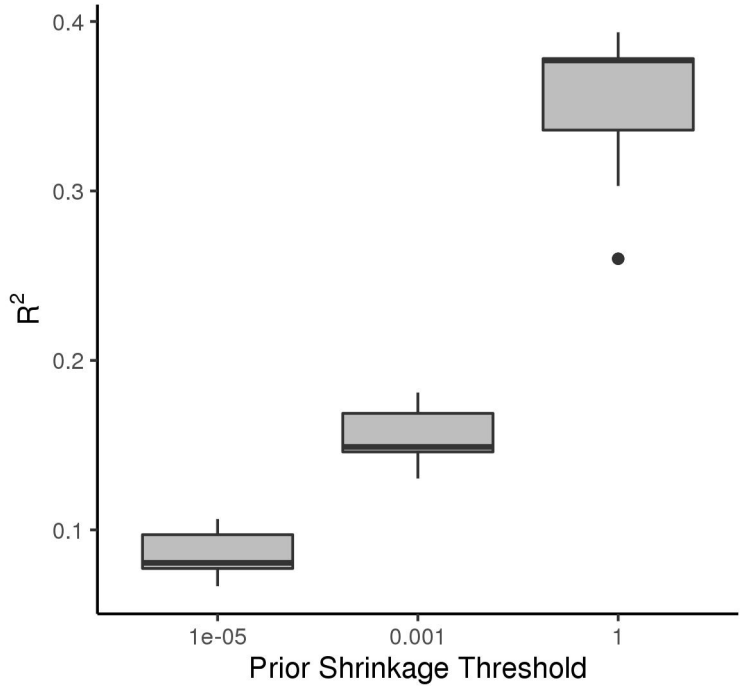


Figure S2: **Effect of shrinkage before prior estimation on prior quality (lifespan dataset).** To assess how the different parameters used in the analysis influenced the prior quality, we performed the Bayesian analysis using different values for several parameters (threshold to select MR instruments, minimum number of instruments needed for each RF, pruning distance/LD for instruments, shrinkage before MR step, p-value threshold for stepwise selection, shrinkage before prior estimation). Most of these parameters had a small effect on the squared correlation between observed and prior effects (R2, used as measure of the prior quality). The only parameter having a strong effect is the shrinkage threshold used to set to 0 all SNP risk factors effects before estimating the prior.

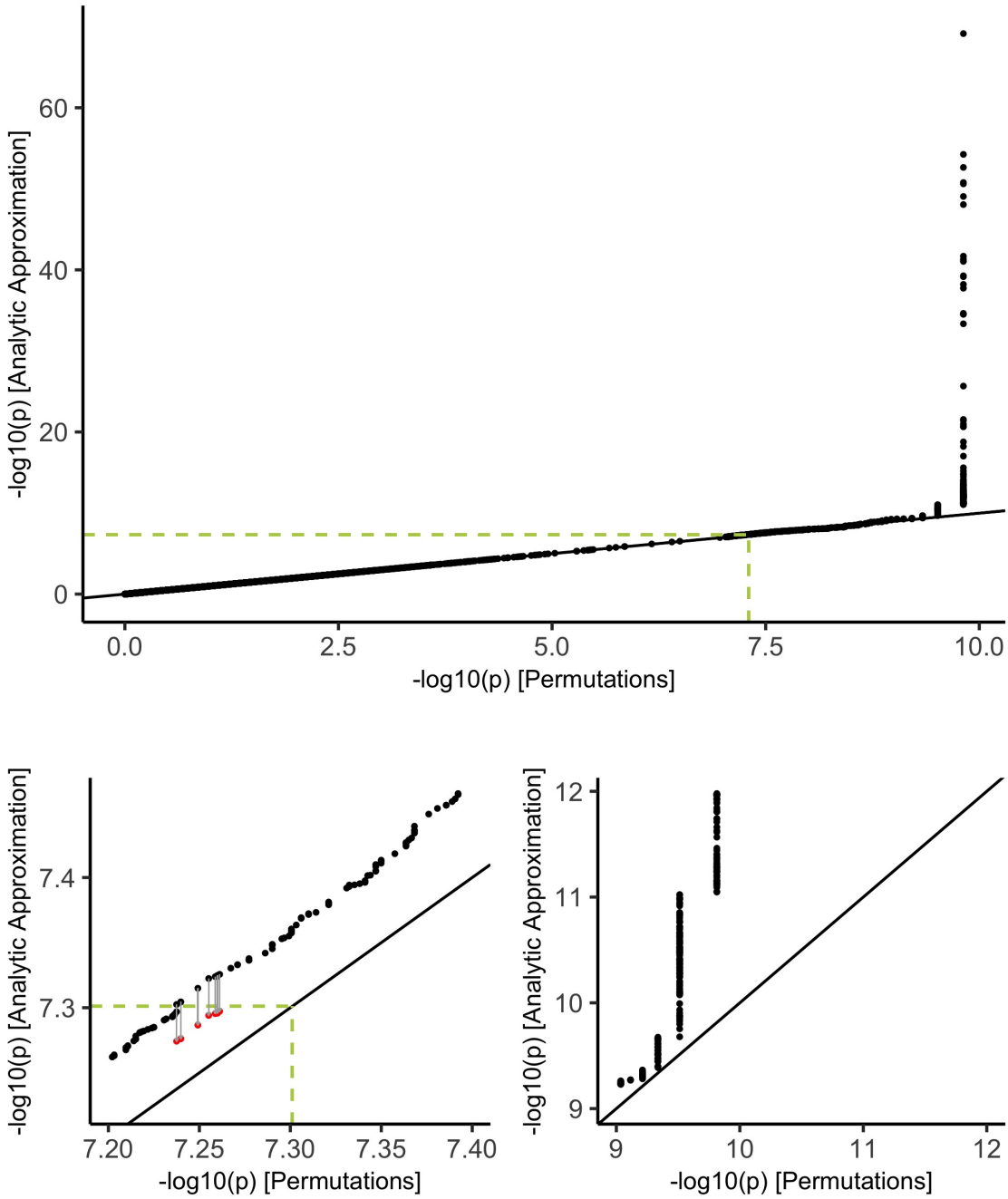


Figure S3: **Comparison of p-values estimation using different approaches (lifespan dataset).** Negative log10 p-values obtained using the analytical approach are plotted against the ones obtained from the permutation approach (correlation of 0.9999). The dotted grey line represents the maximal negative log10 p-value that can be estimated by the permutation approach (6,513,704 SNPs – 1,000 permutations, negative log10 p-value limit of 9.81). The dotted green lines represent the genome wide significance threshold (negative log10 p-value of 7.3). Results are shown for all SNPs (top panel) as well as for subsets of SNPs near genome-wide significance threshold (bottom left panel) and near permutation limit (bottom right panel). On the bottom left panel, the black points correspond to the p-values obtained using the approximation and p-values re-estimated using the exact formula are highlighted in red.

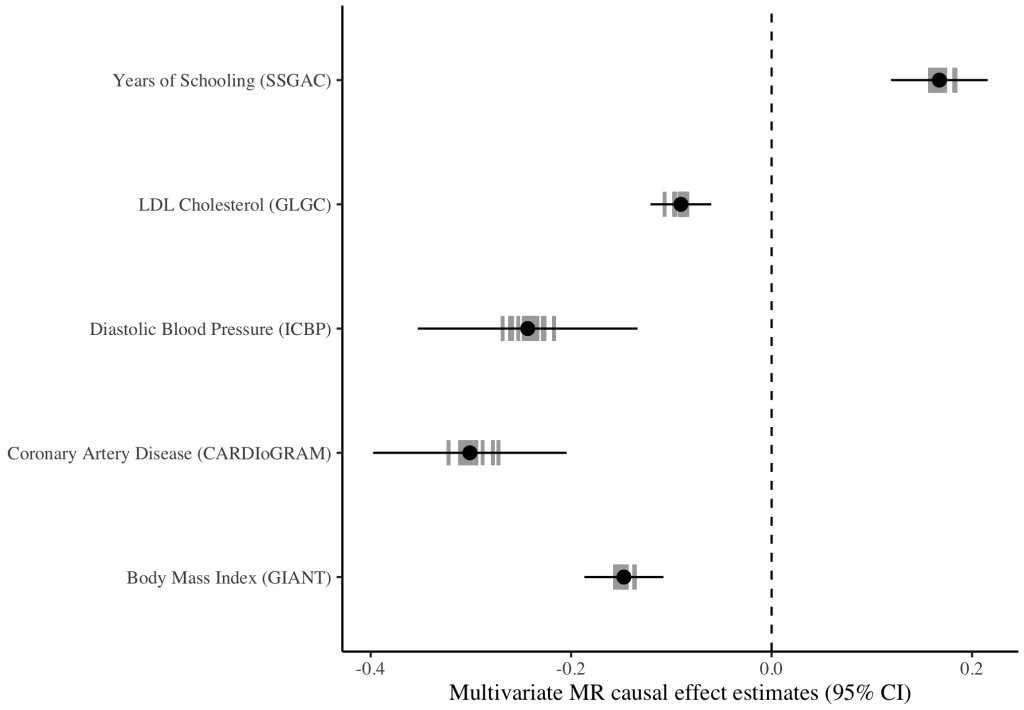


Figure S4: **Coefficients Plot for the five risk factors used to create the prior.**
For each risk factor, the multivariable causal effect estimate and the 95% interval from the multivariable MR model using all chromosomes (black dot and bars) as well as the 22 per-chromosome estimates (grey bars) are represented.

Table S1: **Fifteen new loci associated with lifespan based on Bayes Factors.**
For each loci, we reported information about the variant ("rsid") with the lowest p-value : nearest gene from this variant ("at or near"), chromosomic position ("chr", "pos", from GRCh37), effect and reference alleles ("alt/ref") as well as the observed association statitics ("z"), the prior effect estimate ("μ") and standard error ("σ"), the Bayes Factor ("BF") and the corresponding p-value ("p_{BF}").
Only loci not reaching genome-wide significance in the conventional GWAS (new loci) are reported.

rsid	at or near	chr	pos	alt/ref	z	μ	σ	BF	p_{BF}
rs646776	CELSR2	1	109818530	T/C	-4.908	-4.822	0.748	1.36e+05	1.21e-11
rs7536152	IL6R	1	154423909	A/G	-4.627	-1.636	0.544	1.24e+03	2.86e-08
rs6719980	TMEM18	2	651507	T/C	-5.407	-1.906	0.581	1.97e+04	2.45e-10
rs7599488	BCL11A	2	60718347	T/C	-4.663	-1.956	0.521	2.62e+03	7.32e-09
rs13082711	SLC4A7	3	27537909	T/C	4.428	1.954	0.616	1.67e+03	1.65e-08
rs2271961	TRAIP	3	49878113	T/C	3.824	3.085	0.572	1.06e+03	3.87e-08
rs4580876	MIR2113	6	98322872	A/G	4.112	2.915	0.526	2.37e+03	8.76e-09
rs11556924	ZC3HC1	7	129663496	T/C	5.062	3.276	0.648	1.00e+05	1.93e-11
rs10104032	TNKS	8	9616664	A/C	-4.232	-2.169	0.533	1.31e+03	2.61e-08
rs11986845	PINX1	8	10691318	T/C	-4.237	-2.428	0.544	1.96e+03	1.23e-08
rs59234174	BNC2	9	16730258	T/C	-5.127	-1.492	0.508	2.38e+03	8.70e-09
rs2519093	ABO	9	136141870	T/C	-4.517	-2.275	0.596	3.62e+03	4.13e-09
rs10841520	PDE3A	12	20586395	T/C	4.944	1.334	0.522	1.08e+03	3.75e-08
rs10849925	CUX2	12	111495518	A/G	5.193	1.881	0.605	1.11e+04	6.24e-10
rs12459965	PGPEP1	19	18452195	T/C	4.426	2.781	0.533	5.51e+03	2.00e-09

Table S2: **Comparison of the results for loci identified by at least one of the two analyses.**
For each loci, we reported information about the variant ("rsid") with the lowest p-value : nearest gene from this variant ("at or near"), chromosomic position ("chr", "pos", from GRCh37), as well as the observed association statitics (z), the prior effect estimate ("μ"), the Bayes Factor ("BF") and the corresponding p-value ("p_{BF}"). For each variant, there is one value of prior effect estimate, Bayes Factor and p-value from our analysis using bGWAS v.1.0.2 ("new") and one from Timmers *et al* ("old").

rsid	at or near	chr	pos	z	μ (new)	μ (old)	BF (new)	BF (old))	p_{BF} (new)	p_{BF} (old)	Remarks
rs646776	CELSR2	1	109818530	-4.908	-4.822	-6.902	1.36e+05	4.94e+04	1.36e+05	2.39e-09	both analyses
rs1230666	MAGI3	1	114173410	-5.805	-1.433	0.000	1.03e+04	1.84e+04	1.03e+04	7.88e-09	both analyses
rs6719980	TMEM18	2	651507	-5.407	-1.906	-2.107	1.97e+04	1.73e+05	1.97e+04	7.04e-10	both analyses
rs1275922	KCNK3	2	26932887	-5.817	-0.850	-0.816	1.15e+03	1.11e+05	1.15e+03	8.45e-10	both analyses
rs61348208	HTT	4	3089564	5.823	1.479	0.972	1.10e+04	1.42e+05	1.10e+04	8.45e-10	both analyses
rs10455872	LPA	6	161010118	10.282	2.150	2.534	2.42e+12	4.93e+17	2.42e+12	1.41e-10	both analyses
rs11556924	ZC3HC1	7	129663496	5.062	3.276	1.446	1.00e+05	1.76e+04	1.00e+05	8.16e-09	both analyses
rs1333045	CDKN2B-AS1	9	22119195	6.256	2.981	2.743	1.05e+07	2.00e+07	1.05e+07	1.41e-10	both analyses
rs2519093	ABO	9	136141870	-4.517	-2.275	-3.020	3.62e+03	1.09e+04	3.62e+03	1.87e-08	both analyses
rs8042849	CHRNA3/5	15	78817929	10.659	0.265	1.175	6.99e+05	1.48e+19	6.99e+05	1.41e-10	both analyses
rs8039305	FURIN/FES	15	91422543	6.414	1.473	-0.194	6.40e+04	5.06e+04	6.40e+04	2.25e-09	both analyses
rs12924886	HP	16	72075593	5.679	2.455	1.477	1.51e+05	2.05e+05	1.51e+05	7.04e-10	both analyses
rs6511720	LDLR	19	11202306	5.631	3.787	3.804	2.07e+06	2.51e+06	2.07e+06	1.41e-10	both analyses
rs429358	APOE	19	45411941	19.328	1.854	4.026	1.11e+37	2.92e+61	1.11e+37	1.41e-10	both analyses
rs7536152	IL6R	1	154423909	-4.627	-1.636	0.000	1.24e+03	4.32e+02	1.24e+03	3.86e-06	new signal
rs7599488	BCL11A	2	60718347	-4.663	-1.956	-1.290	2.62e+03	3.66e+03	2.62e+03	1.08e-07	new signal
rs2271961	TRAIP	3	49878113	3.824	3.085	1.757	1.06e+03	4.07e+02	1.06e+03	4.27e-06	new signal
rs4580876	MIR2113	6	98322872	4.112	2.915	1.792	2.37e+03	1.02e+03	2.37e+03	9.20e-07	new signal
rs10104032	TNKS	8	9616664	-4.232	-2.169	-0.124	1.31e+03	1.97e+02	1.31e+03	1.46e-05	new signal
rs11986845	PINX1	8	10691318	-4.237	-2.428	0.000	1.96e+03	1.63e+02	1.96e+03	2.00e-05	new signal
rs59234174	BNC2	9	16730258	-5.127	-1.492	0.000	2.38e+03	1.94e+03	2.38e+03	3.09e-07	new signal
rs10841520	PDE3A	12	20586395	4.944	1.334	0.000	1.08e+03	1.08e+03	1.08e+03	8.31e-07	new signal
rs12459965	PGPEP1	19	18452195	4.426	2.781	0.857	5.51e+03	9.20e+02	5.51e+03	1.08e-06	new signal
rs10211471	GBX2/ASB18	2	237081854	-4.871	-0.862	-1.433	2.15e+02	9.26e+03	2.15e+02	2.29e-08	Timmers et al only
rs111333005	LPA/IGF2R	6	160487196	-5.496	-0.183	-0.614	3.62e+01	2.02e+04	3.62e+01	6.62e-09	Timmers et al only
rs113160991	POM21C	7	75094329	-5.137	-0.963	-1.276	4.77e+02	1.94e+04	4.77e+02	7.46e-09	Timmers et al only

Table S3: **Nine new loci associated with lifespan based on posterior effects.**
For each loci, we reported information about the variant ("rsid") with the lowest p-value : nearest gene from this variant ("at or near"), chromosomic position ("chr", "pos", from GRCh37), effect and reference alleles ("alt/ref") as well as the observed association statitics (z), the posterior effect estimate (μ_p) and standard error (σ_p) and the corresponding p-value (p_p).
Only loci not reaching genome-wide significance in the conventional GWAS or based on Bayes Factors (new loci) are reported.

rsid	at or near	chr	pos	alt/ref	z	μ_p	σ_p	p_p
rs13086611	USP4	3	49385417	A/T	-3.019	-3.043	0.503	1.46e-09
rs13130484	GNPDA2	4	45175691	T/C	-3.161	-3.270	0.487	1.86e-11
rs34809719	MAD1L1	7	2028968	T/G	3.227	2.774	0.483	8.97e-09
rs964184	ZPR1	11	116648917	C/G	3.791	3.165	0.497	1.96e-10
rs1183910	HNF1A	12	121420807	A/G	-3.676	-2.915	0.481	1.33e-09
rs8049439	ATXN2L	16	28837515	T/C	2.809	2.949	0.480	8.05e-10
rs1421085	FTO	16	53800954	T/C	3.550	3.760	0.612	7.89e-10
rs999474	UBE2Z	17	46987665	A/G	3.502	2.674	0.469	1.22e-08
rs303757	RMC1	18	21078716	T/G	2.793	2.983	0.478	4.37e-10

Table S4: **Four loci associated with lifespan based on direct effects.**
For each loci, we reported information about the variant ("rsid") with the lowest p-value : nearest gene from this variant ("at or near"), chromosomic position ("chr", "pos", from GRCh37), effect and reference alleles ("alt/ref") as well as the observed association statitics (z), the posterior effect estimate (μ_d) and standard error (σ_d) and the corresponding p-value (p_d).

rsid	at or near	chr	pos	alt/ref	z	μ_d	σ_d	p_d
rs55730499	LPA	6	161005610	T/C	-10.258	-8.295	1.166	1.13e-12
rs7307680	RAD52	12	1052488	A/G	-5.286	-6.196	1.134	4.64e-08
rs8042849	HYKK	15	78817929	T/C	10.659	10.395	1.118	1.41e-20
rs429358	APOE	19	45411941	T/C	19.328	17.473	1.228	5.79e-46

References

Feitosa, Mary F., Aldi T. Kraja, Daniel I. Chasman, Yun J. Sung, Thomas W. Winkler, Ioanna Ntalla, Xiuqing Guo, et al. 2018. "Novel Genetic Associations for Blood Pressure Identified via Gene-Alcohol Interaction in up to 570K Individuals Across Multiple Ancestries." Edited by Helena Kuivaniemi. *PLOS ONE* 13 (6): e0198166. https://doi.org/10.1371/journal.pone.0198166.

Furberg, Helena, YunJung Kim, Jennifer Dackor, Eric Boerwinkle, Nora Franceschini, Diego Ardissino, Luisa Bernardinelli, et al. 2010. "Genome-Wide Meta-Analyses Identify Multiple Loci Associated with Smoking Behavior." *Nature Genetics* 42 (5): 441–47. https://doi.org/10.1038/ng.571.

Harst, Pim van der, and Niek Verweij. 2018. "Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease." *Circulation Research* 122 (3): 433–43. https://doi.org/10.1161/CIRCRESAHA.117.312086.

Hobbs, Brian D, and Kim de Jong, Maxime Lamontagne, Yohan Bossé, Nick Shrine, Marıa Soler Artigas, Louise V Wain, et al. 2017. "Genetic Loci Associated with Chronic Obstructive Pulmonary Disease Overlap with Loci for Lung Function and Pulmonary Fibrosis." *Nature Genetics* 49 (3): 426–32. https://doi.org/10.1038/ng.3752.

Mack, Salome, Stefan Coassin, Rico Rueedi, Noha A. Yousri, Ilkka Seppälä, Christian Gieger, Sebastian Schönherr, et al. 2017. "A Genome-Wide Association Meta-Analysis on Lipoprotein (a) Concentrations Adjusted for Apolipoprotein (a) Isoforms." *Journal of Lipid Research* 58 (9): 1834–44. https://doi.org/10.1194/jlr.m076232.

McDaid, Aaron F., Peter K. Joshi, Eleonora Porcu, Andrea Komljenovic, Hao Li, Vincenzo Sorrentino, Maria Litovchenko, et al. 2017. "Bayesian Association Scan Reveals Loci Associated with Human Lifespan and Linked Biomarkers." *Nature Communications* 8 (July): 15842. https://doi.org/10.1038/ncomms15842.

McKay, James D, and Rayjean J Hung, Younghun Han, Xuchen Zong, Robert Carreras-Torres, David C Christiani, Neil E Caporaso, et al. 2017. "Large-Scale Association Analysis Identifies New Lung Cancer Susceptibility Loci and Heterogeneity in Genetic Susceptibility Across Histological Subtypes." *Nature Genetics* 49 (7): 1126–32. https://doi.org/10.1038/ng.3892.

Nazarian, Alireza, Anatoliy I. Yashin, and Alexander M. Kulminski. 2019. "Genome-Wide Analysis of Genetic Predisposition to Alzheimer's Disease and Related Sex Disparities." *Alzheimers Research & Therapy* 11 (1). https://doi.org/10.1186/s13195-018-0458-8.

Timmers, Paul RHJ, Ninon Mounier, Kristi Lall, Krista Fischer, Zheng Ning, Xiao Feng, Andrew D Bretherick, et al. 2019. "Genomics of 1 Million Parent Lifespans Implicates Novel Pathways and Common Diseases and Distinguishes Survival Chances." *eLife* 8 (January): e39856. https://doi.org/10.7554/elife.

39856.

Wain, Louise V., Ahmad Vaez, Rick Jansen, Roby Joehanes, Peter J. van der Most, A. Mesut Erzurumluoglu, Paul F. O'Reilly, et al. 2017. "Novel Blood Pressure Locus and Gene Discovery Using Genome-Wide Association Study and Expression Data Sets from Blood and the Kidney." *Hypertension* 70 (3): e4–e19. https://doi.org/10.1161/HYPERTENSIONAHA.117.09438.

Appendix C

Correction for sample overlap, winner's curse and weak instrument bias in two-sample Mendelian Randomization

This article (Mounier & Kutalik, 2021) is presented in Chapter 2.

Note that this is a preprint and additional work has been conducted since its publication. Some of the latest results are actually included in the Discussion.

Correction for sample overlap, winner's curse and weak instrument bias in two-sample Mendelian Randomization

Ninon Mounier[1,2] **Zoltán Kutalik**[1,2,3,*]

Abstract

Inverse-variance weighted two-sample Mendelian Randomization (IVW-MR) is the most widely used approach that uses genome-wide association studies summary statistics to infer the existence and strength of the causal effect between an exposure and an outcome. Estimates from this approach can be subject to different biases due to: (i) the overlap between the exposure and outcome samples; (ii) the use of weak instruments and winner's curse. We developed a method that aims at tackling all these biases together. Assuming spike-and-slab genomic architecture and leveraging LD-score regression and other techniques, we could analytically derive and reliably estimate the bias of IVW-MR using association summary statistics only. This allowed us to apply a bias correction to IVW-MR estimates, which we tested using simulated data for a wide range of realistic scenarios. In all the explored scenarios, our correction reduced the bias, in some situations by as much as 30 folds. When applied to real data on obesity-related exposures, we observed significant differences between IVW-based and corrected effects, both for non-overlapping and fully overlapping samples. While most studies are extremely careful to avoid any sample overlap when performing two-sample MR analysis, we have demonstrated that the incurred bias is much less substantial than the one due to weak instruments or winner's curse, which are often ignored.

[1]University Center for Primary Care and Public Health, University of Lausanne, 1010, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland

[3]Department of Computational Biology, University of Lausanne, Lausanne, 1015, Switzerland

[*]Correspondence should be addressed to zoltan.kutalik@unil.ch

1 Introduction

Mendelian Randomization (MR) is a method that uses genetic variants (typically single nucleotide polymorphisms, SNPs) as instrumental variables (IVs) to infer the existence and strength of the causal effect between an exposure and an outcome[1]. In particular, two-sample summary data MR[2], that requires solely genome-wide association study (GWAS) summary statistics has become increasingly popular. The reason for this is that in the last decade, GWASs have drastically increased in sample size[3]. The resulting summary statistics are often publicly available allowing not only the identification of genetic variants independently associated with a particular exposure, i.e. IVs, but also the look up of the variants' effects on the outcome trait in different samples. Each IV provide an independent estimate for the causal effect and these estimates can then be combined using an inverse variance-weighting (IVW) approach[2]. MR relies on three main assumptions (Figure 1): (1) Relevance – IVs must be robustly associated with the exposure. (2) Exchangeability – IVs must not be associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – IVs must be independent of the outcome conditional on the exposure and all confounders of the exposure-outcome relationship (i.e. the only path between the IVs and the outcome is via the exposure).

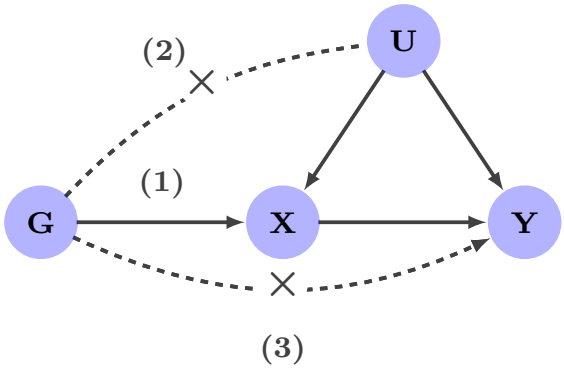


Figure 1: Main assumptions of Mendelian randomization. (1) Relevance – IVs, denoted by G, are strongly associated with the exposure. (2) Exchangeability – G is not associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – G is independent of the outcome conditional on the exposure and all confounders of the exposure-outcome relationship (i.e. the only path between the IVs and the outcome is via the exposure).

When the exposure and the outcome are coming from overlapping samples, environmental confounding factor(s) can bias the causal effect estimate toward the observational correlation, which includes the correlation induced by the environmental confounder(s)[4]. Hence, sample overlap will introduce an association between the IVs and the confounder, resulting in a violation of the second assumption of MR. Nowadays, large consortium summary statistics are often used for both the exposure and the outcome, and it is not always possible to ensure that the two samples used are not overlapping. The need for non-overlapping samples does not only substantially reduce the sample size used to generate the summary statistics, but in the era of large biobanks (such as UK-Biobank[5], UKBB) it is becoming increasingly difficult to find recent meta-analysis

datasets for an exposure and outcome without overlapping samples.

Sample overlap is not the only source of bias that needs to be considered. Two-sample MR estimates obtained from non-overlapping samples are known to be subject to weak instrument bias (toward the null). Most MR methods assume that SNP-exposure effects are measured without noise (NO Measurement Error, NOME assumption). This simplification leads to regression-dilution bias, which is particularly strong for weak instruments. For this reason, the bias introduced by the NOME assumption is referred to as weak instrument bias and it becomes more and more severe as the average variance of the exposure explained by the IVs decreases[6]. When combined with sample overlap, the effect of weak instrument bias will move toward the observational correlation[7,8]. Using IVs strongly associated with the exposure and/or increasing the sample size can mitigate weak instrument bias[7]. Although the exact multiplicative bias due to the NOME assumption can be expressed analytically (proportional to the inverse of the F-statistic), the estimator for the multiplicative constant has typically high variance and works poorly in practice[9]. The simulation-extrapolation based SIMEX method proved to yield more robust corrections both for IVW[10] and MR-Egger estimates[9].

In addition, MR estimates are subject to winner's curse bias, which occurs when the same sample is used to select IVs and estimate their effect on the exposure. In such case, the observed IV effect on the exposure is not an unbiased estimator for its true effect and is likely to be overestimated (in absolute value). This would affect the causal effect estimate (underestimation in non-overlapping sample and overestimation in fully overlapping samples)[8]. Using a third independent sample to select instruments, and therefore avoid winner's curse, is not always possible. Based on the expectation of truncated normal distribution[11], a correction can be applied for the SNP-exposure effect sizes. However, the additional estimator variance such correction entails can outweigh the benefit of the reduced bias, which can be mitigated by directly maximising the conditional likelihood[12]. Still, all these methods account for winner's curse for a single SNP, but do not model the winner's curse bias induced by the IV selection process from millions of potential markers, which is far more complex.

While previous approaches aimed at tackling one bias at a time, the intricate way these different sources of biases interplay with each other remains poorly understood, and there is currently no method that simultaneously tackles them.

In this paper, we will first introduce a two-sample MR framework that takes into account these three sources of bias simultaneously in order to obtain a corrected causal effect estimate. We will then test our approach and compare the proposed correction of the IVW-MR causal effect estimate against its uncorrected counterpart using a wide range of simulation settings. Finally, to demonstrate its utility, we will apply our approach to obesity-related traits using UKBB data.

2 Methods

2.1 Expectation of the causal effect estimate

Let X and Y denote two random variables representing two complex traits. We intend to use MR to estimate the causal effect of X on Y. We will use m Linkage Disequilibrium (LD)-independent genetic variants as IVs. Their corresponding genotype data is denoted by G and its jth column by g_j (rows representing individuals). To simplify notation we assume that $E(X) = E(Y) = E(G) = 0$ and $Var(X) = Var(Y) = Var(G) = 1$. Let us assume that X is observed in sample A of sample size n_A, Y is measured in sample B of sample size n_B with an overlap of $n_{A\cap B}$ individuals between the two samples. The vector of realisations of Z^C is denoted by $\boldsymbol{z}^C$ for all variables ($Z = X, Y, G, g$) and samples ($C = A, B, A \cap B$). Let us assume the following models:

$$\begin{aligned} \boldsymbol{x}^A &= G^A \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon_x^A} \\ \boldsymbol{x}^B &= G^B \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon_x^B} \\ \boldsymbol{y}^B &= \alpha \cdot \boldsymbol{x}^B + G^B \cdot \boldsymbol{\gamma_y} + \boldsymbol{\epsilon_y^B} \end{aligned} \tag{1}$$

where $\boldsymbol{\gamma_x}$ are the effect sizes of the instruments on X, $\boldsymbol{\gamma_y}$ are their pleiotropic effects on Y. Assuming that there is a single environmental confounder U acting linearly on both traits (as used for simulations) the error term can split into two parts : $\boldsymbol{\epsilon_x^C} = \kappa_x \cdot \boldsymbol{u}^C + \boldsymbol{\varepsilon}^C$ and $\boldsymbol{\epsilon_y^C} = \kappa_y \cdot \boldsymbol{u}^C + \boldsymbol{\varepsilon}^C$, where κ_x and κ_y refer to the effect of U on X and Y respectively, ε^C is independent of the confounder and C can take the values A, B or $A \cap B$ as above.

Under the INSIDE assumption[13] (INstrument Strength Independent of Direct Effect, i.e. horizontal pleiotropic effects are independent of the direct effect), $cov(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}) = 0$ and $E[\boldsymbol{\gamma_y}] = 0$. We denote $\rho := cov(\boldsymbol{\epsilon_x}, \boldsymbol{\epsilon_y}) = \kappa_x \cdot \kappa_y$. It corresponds to the part of the phenotypic correlation (r) due to a (non-genetic) confounder ($r = \rho + \alpha$). The univariate effect size estimates from GWASs for genetic variant j are as follows:

$$\begin{aligned} (\widehat{\beta}_x^A)_j &= \frac{1}{n_A} \cdot (\boldsymbol{g}_j^A)' \cdot \boldsymbol{x}^A = \frac{1}{n_A} \cdot (\boldsymbol{g}_j^A)' \cdot \left(G^A \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon_x^A}\right) \\ (\widehat{\beta}_y^B)_j &= \frac{1}{n_B} \cdot (\boldsymbol{g}_j^B)' \cdot \boldsymbol{y}^B = \frac{1}{n_B} \cdot (\boldsymbol{g}_j^B)' \cdot \left(\alpha \cdot \boldsymbol{x}^B + G^B \cdot \boldsymbol{\gamma_y} + \boldsymbol{\epsilon_y^B}\right) \end{aligned} \tag{2}$$

where the genotype data for genetic variant j for individuals in sample A is denoted by $\boldsymbol{g}_j^A$.

Let us now consider the fixed-effect inverse-variance weighting meta-analysis for the ratio estimates for the causal effect α. Each IV j provides a ratio estimate:

$$\widehat{\alpha}_j = \frac{(\widehat{\beta}_y^B)_j}{(\widehat{\beta}_x^A)_j} \tag{3}$$

$$Var(\widehat{\alpha}_j) = \frac{Var((\widehat{\beta}_y^B)_j)}{(\widehat{\beta}_x^A)_j^2} = \frac{(1 - \alpha^2 - \gamma_{y_j}^2)}{n_B \cdot (\widehat{\beta}_x^A)_j^2} \tag{4}$$

Hence the weights (w_j) of IV j for estimating the IVW causal effect are of the form:

$$w_j = \frac{1}{Var(\widehat{\alpha}_j)} = \frac{n_B \cdot (\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)} \tag{5}$$

Finally, the estimate can be written in the following form:

$$\begin{aligned} \widehat{\alpha}_{IVW} &= \frac{\sum_{j=1}^m \widehat{\alpha}_j \cdot w_j}{\sum_{j=1}^m w_j} = \frac{\sum_{j=1}^m \frac{(\widehat{\beta}_y^B)_j}{(\widehat{\beta}_x^A)_j} \times \left(\frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)}\right)}{\sum_{j=1}^m \frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)}} = \frac{\sum_{j=1}^m \frac{(\widehat{\beta}_y^B)_j \cdot (\widehat{\beta}_x^A)_j}{(1-\alpha^2-\gamma_{y_j}^2)}}{\sum_{j=1}^m \frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)}} \\ &\approx \frac{\sum_{j=1}^m \frac{(\widehat{\beta}_y^B)_j \cdot (\widehat{\beta}_x^A)_j}{(1-\alpha^2)}}{\sum_{j=1}^m \frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2)}} = \frac{\sum_{j=1}^m (\widehat{\beta}_y^B)_j \cdot (\widehat{\beta}_x^A)_j}{\sum_{j=1}^m (\widehat{\beta}_x^A)_j^2} \end{aligned} \tag{6}$$

Here, the last approximation is based on the realistic assumption that the individual pleiotropic effect of each SNP is very small. To account for winner's Curse, we need to consider the probability of being selected for each genetic variant. Let us consider a threshold T ($T = -\Phi^{-1}(p/2) \approx 5.45$ for $p = 5 \times 10^{-8}$, genome-wide significance threshold, for example) and use only IVs with $|(\widehat{\beta}_x^A)_j| \cdot \sqrt{n_A} > T$. By denoting $S_j := \left\{|(\widehat{\beta}_x^A)_j| \cdot \sqrt{n_A} > T\right\}$, the causal effect estimate (Equation 6) changes to:

$$\widehat{\alpha}_{IVW} \approx \frac{\sum_{j=1}^M ((\widehat{\beta}_y^B)_j \,|S_j) \cdot ((\widehat{\beta}_x^A)_j \,|S_j) \cdot Pr(S_j)}{\sum_{j=1}^M ((\widehat{\beta}_x^A)_j \,|S_j)^2 \cdot Pr(S_j)} \tag{7}$$

Note that while m denoted the number of IVs, M represents the number of genome-wide variants from which IVs are selected. By approximating the expectation of a ratio by the ratio of expectations, the expectation of the causal effect estimate (Equation 7) can be written as:

$$\begin{aligned} E[\widehat{\alpha}_{IVW}] &\approx \frac{\sum_{j=1}^M E\left[((\widehat{\beta}_y^B)_j \,|S_j) \cdot ((\widehat{\beta}_x^A)_j \,|S_j)\right] \cdot Pr(S_j)}{\sum_{j=1}^M E\left[((\widehat{\beta}_x^A)_j \,|S_j)^2\right] \cdot Pr(S_j)} \\ &= \frac{\sum_{j=1}^M s_j \cdot Pr(S_j)}{\sum_{j=1}^M t_j \cdot Pr(S_j)} \end{aligned} \tag{8}$$

The values of s_j, t_j and $Pr(S_j)$ can be analytically derived (presented in Supplementary Section A), and we show that Equation 8 translates to:

$$\begin{aligned} E[\hat{\alpha}_{IVW}] \approx \quad & \alpha \cdot \frac{\left(\pi_x \cdot \sigma_x^2\right) \cdot \left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\right)}{d(n_A, T, \pi_x, \sigma_x^2)} \\ + \quad & \lambda' \cdot \frac{\pi_x \cdot \left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\right) + \left(1 - \pi_x\right) \cdot 2 \cdot c}{d(n_A, T, \pi_x, \sigma_x^2)} \end{aligned} \tag{9}$$

where π_x and σ_x^2 are characteristics of the genetic architecture of trait X (respectively, a measure of the polygenicity and the per-variant heritability, see Equation S35) and λ' is a quantity closely related to the cross-trait LD score regression (LDSC) intercept[14] (λ) : $\lambda' = \frac{\lambda}{\sqrt{n_A \cdot n_B}}$.

The constants a, b and c do not depend on the causal effect α nor the sample overlap since $n_{A\cap B}$ is only affecting λ' (see Equations S33 and S51). The same is true for the denominator $(d(n_A, T, \pi_x, \sigma_x^2))$ (see Equation S52).

We can see that in the absence of sample overlap, the second term is equal to 0, so the first term corresponds to the causal effect estimated when only weak instruments and winner's curse are biasing the estimate. In this case, $E[\hat{\alpha}_{IVW}]$ is lower than α and the observed effect will be biased toward the null (Figure S1). The only parameters affecting the bias are π_x and h_x^2, n_A and T. When π_x is smaller, or h_x^2 is larger, then IVs have stronger effects, leading to a smaller bias when all other parameters are kept constant. As expected, since these are commonly used approaches to limit weak instrument bias, using a more stringent threshold and increasing the exposure sample size both reduce the bias.

When there is sample overlap (with % overlap defined as $\frac{n_{A\cap B}}{\sqrt{n_A \cdot n_B}}$), the expression of $E[\hat{\alpha}_{IVW}]$ is more complex. The magnitude of the bias will not only depend on the parameters described for non-overlapping samples (Figures S2, S3 and S4), but also on the confounder's effect (ρ) that can affect both the magnitude and the direction of the bias. For example, when the percentage of overlap is relatively low (20%) and ρ has the same sign as α, then the estimate will be further biased toward the null, whereas a confounder acting in the opposite direction (as the causal effect) will reduce the bias (Figure S2). When the percentage of overlap increases (Figures S3 and S4), the effect of sample overlap gets stronger, and the bias direction directly depends on ρ. We can see that there exists a value of ρ (that depends on all the other parameters) for which the two opposing forces (weak instruments, sample overlap) contributing to the biases cancel out.

All parameters except α are known or can be estimated from the data. We assume that sample sizes for X and Y (respectively n_A and n_B) are known, as well as the threshold used to select IVs (T). Parameter λ' can be estimated from cross-trait LDSC[14]. π_x and σ_x^2 are estimated by matching the denominator of this formula $d(n_A, T, \pi_x, \sigma_x^2)$ to the denominator of equation 8 (see Supplementary Section B for details).

From Equation 10, we can derive a corrected effect for the causal effect :

$$\hat{\alpha}_c = \frac{\hat{\alpha}_{IVW} \cdot d(n_A, T, \pi_x, \sigma_x^2) - \lambda' \cdot (\pi_x \cdot \big(2 \cdot a + b \cdot (1 + \sigma_x^2 \cdot n_A)\big) + \big(1 - \pi_x\big) \cdot 2 \cdot c)}{\big(\pi_x \cdot \sigma_x^2\big) \cdot \big(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\big)} \tag{10}$$

We also derived the standard error of the corrected effect as well as the covariance between observed and corrected effects in Supplementary Section C. Note that the formula proposed by Burgess *et al.*[4] (Supplementary Section D) under the null is a special case of ours when all instruments are selected based on external data (i.e. there is no winner's curse).

This approach has been implemented in an R-package (`MRlap`).

2.2 Simulations

We used UKBB[5] genotypic data and restricted our analyses to unrelated individuals of British ancestry (identified using genomic principal components) and HapMap3 genetic variants[15]

($M \approx 1,150,000$) to simulate phenotypic data. From this set of 379,530 individuals, we first sampled the exposure dataset (n_A individuals) and 5 different outcome datasets of sample size n_B, with an overlap with the exposure dataset varying (from no overlap to full overlap, increasing in increments of 25%). Next, causal SNPs for the exposure were randomly drawn from the set of 1,150,000 genetic variants, based on the polygenicity of X (π_x) and their effects were simulated using the heritability of X (h^2_x) as follows:

$$\boldsymbol{\gamma_x} \sim \begin{cases} 0 & \text{with probability } 1-\pi_x, \text{ for non-causal variants} \\ \mathcal{N}\left(0, \frac{h^2_x}{M \cdot \pi_x}\right) & \text{with probability } \pi_x, \text{ for causal variants} \end{cases} \tag{11}$$

For simplicity, we assumed that there were no direct genetic effects on the outcome. Then, phenotypic data for X and Y were simulated for all individuals included in the exposure or in any of the outcome samples, taking into account the effect of the confounder U on X and Y (respectively κ_x and κ_y) and the causal effect of X on Y (α), using the following design:

$$\boldsymbol{u} \sim \mathcal{N}(0,1) \tag{12}$$

$$\boldsymbol{x} = G \cdot \boldsymbol{\gamma_x} + \boldsymbol{u} \cdot \kappa_x + \epsilon_x \tag{13}$$
$$\text{with } \epsilon_x \sim \mathcal{N}(0, 1-(h^2_x + \kappa^2_x))$$

$$\boldsymbol{y} = \alpha \cdot \boldsymbol{x} + \boldsymbol{u} \cdot \kappa_y + \epsilon_y \tag{14}$$
$$\text{with } \epsilon_y \sim \mathcal{N}(0, 1-(\alpha^2 + \kappa^2_y + 2 \cdot \alpha \cdot \kappa_y \cdot \kappa_x))$$

Note that these settings ensure that both X and Y have a zero mean and a variance of 1. A GWAS was performed for each sample (the exposure one and the five outcome ones) using BGENIE[16]. The GWAS summary statistics obtained were used to perform downstream analyses. MR analyses were performed using the Two-Sample MR R-package[17]: IVs were selected based on their observed effect (with a varying T threshold) and distance-pruned (500kb) to obtain a set of independent IVs. LDSC analyses were perfomed using the LDSC implementation from the GenomicSEM R-package[18] and the 1000G LD-scores[19]. Corrected effects were estimated using Equation 10.

For each parameter setting we tested, 100 datasets were simulated. Our standard parameter settings consisted of simulating data for $n_A = 20,000$ and $n_B = 20,000$ individuals. X was simulated with moderate polygenicity and large heritability ($\pi_x = 0.001$ and $h^2_x = 0.4$). U had a moderate effect on both X and Y ($\kappa_x = 0.3$, $\kappa_y = 0.5$), leading to a correlation (ρ) of 0.15 induced by the confounder. X had a moderate direct causal effect on Y ($\alpha = 0.2$).
In addition to these standard settings, we explored various other parameter values. We investigated the effect of a confounder acting in an opposite direction ($\kappa_x = -0.3$ and $\kappa_y = 0.5$) and tested different strengths for the confounding factor (weaker : $\kappa_x = 0.15$ and $\kappa_y = 0.3$, and stronger : $\kappa_x = 0.5$ and $\kappa_y = 0.8$). We also simulated data for a scenario without any causal relationship between X and Y ($\alpha = 0$). Finally, we explored a scenario with more realistic parameters: larger sample sizes ($n_A = 100,000$, $n_B = 100,000$), increased polygenicity

($\pi_x = 0.005$), lower heritability ($h_x^2 = 0.2$) and a smaller causal effect ($\alpha = 0.1$). For each scenario, observed and corrected causal effects were compared for different degrees of sample overlap and different instrument selection thresholds. Results quality was assessed using root-mean square error (RMSE).

For a given instrument selection threshold, we obtained 500 causal effect estimates: one for each of the 100 simulated data sets for each of the five sample overlap proportion scenarios. Causal effect estimates should ideally not depend on the extent of overlap between the exposure and outcome samples. To quantify the extend to which this holds, we grouped estimates according to which sample overlap proportion they came from and compared the between group variance relative to the within group variance of the estimates. A method that is robust to overlap between the exposure and outcome samples will have small between group variance relative to the variance of the estimator (characterised by the within group variance).

Finally, we tested for differences between observed and corrected effects using the following test statistics:

$$t_{\text{diff}} = \frac{\hat{\alpha}_{IVW} - \hat{\alpha}_c}{\sqrt{Var(\hat{\alpha}_{IVW}) + Var(\hat{\alpha}_c) - 2 \cdot Cov(\hat{\alpha}_{IVW}, \hat{\alpha}_c)}} \tag{15}$$

2.3 Application to UKBB

To assess the effect of sample overlap on real data, we used a design very similar to the one used for simulations. We used both genotypic and phenotypic data from UKBB[5] and restricted our analyses to the same subsets of individuals and genetic variants. From this set of individuals, we first sampled the exposure dataset (100,000 individuals) and 5 different outcome datasets (100,000 individuals), where the overlap with the exposure dataset varied (from no overlap to full overlap, increasing in increments of 25% - in the case of unequal sample sizes, the percentage of overlap for the samples will proportionally increase from 0 to the maximum value attainable given the difference in sample sizes). Note that we always used the full set of individuals to create the 100,000-individuals samples, so the percentage of missing data in each sample will be the same as in the UKBB. Therefore, the total number of individuals with phenotypic data (effective sample size) will vary depending on the traits. We used four obesity-related traits: body mass index (BMI), systolic blood pressure (SBP), number of cigarettes previously smoked daily (smoking) and alcohol intake frequency (alcohol). For smoking, answers "prefer not answer" and "less than one a day" were considered as missing. For alcohol, answers were recoded to correspond to an increased intake frequency, and answers "do not know" were considered missing. Details about the pairs of traits analysed are available in Table 1.

Phenotypic data was normalised (inverse-normal quantile transformed) and subsequently adjusted for the following covariates: sex, age, age×age and the first 40 principal components. Similarly to what we did for simulations, a GWAS was performed for each sample (the exposure dataset and the five outcome datasets) using BGENIE[16]. The GWAS summary statistics obtained were then used to perform the downstream MR analyses using the Two-Sample

Exposure	Outcome	Field ID Exposure	Field ID Outcome	N Exposure	N Outcome
BMI	SBP	21001	4080	93,219	99,490
SBP	BMI	4080	21001	99,491	93,231
BMI	smoking	21001	2887	99,490	23,507
BMI	alcohol	21001	1558	99,493	99,808

Table 1: Description of the pairs of traits analysed
For each pair of trait analysed, we reported the field IDs corresponding to the exposure and the outcome, as well as the mean sample size (across the 100 repetitions, and the 5 overlaps for the outcome) of the phenotypic data used.

MR R-package[17]: IVs were selected based on their observed effect (exploring various selection thresholds T) and distance-pruned (500kb) to obtain a set of independent IVs. In addition, since in this case a reverse causal effect (from the outcome on the exposure) could exist, we filtered out variants that were significantly more strongly associated with the outcome than with the exposure in order to remove potentially invalid IVs. LDSC analyses were performed using the LDSC implementation from the GenomicSEM R-package[18] and the 1000G LD-scores[14]. Corrected effects were estimated using Equation (10).

We repeated this sampling approach a 100 times. For each repetition, observed and corrected causal effects were compared using the within groups and between groups variances and we tested for differences between observed and corrected effects using Equation 15.

3 Results

3.1 Overview of the method

We propose a two-sample MR framework that takes into account three sources of bias simultaneously : weak instrument bias, sample overlap and winner's curse. We analytically derived the expectation of the observed effect for IVW-based MR estimate:

$$E[\widehat{\alpha}_{IVW}] = f(\alpha, n_A, n_B, T, \pi_X, \sigma^2_X, \lambda)$$

which depends on the true causal effect size (α), the sample sizes of the exposure and outcome GWASs (n_A, n_B), the threshold used to select IVs (T), the cross-trait LDSC intercept (λ, which depends on the degree of the sample overlap, the true causal effect and the strength of the confounder) and the genetic architecture of the exposure (polygenicity π_X and per-variant heritability $\sigma^2_X = \frac{h^2_X}{M \cdot \pi_X}$, M being the number of variants). All parameters (except α) are either known or can be estimated from the data. This allows us to adjust the IVW causal effect estimate with the aim of making it unbiased.

3.2 Simulations

Simulation results under our standard settings show a large discrepancy of the observed causal effects (IVW-based estimate) estimated using different degrees of sample overlap, while the corrected effects are more closely aligned with the true causal effects (Figure 2 - A, Table S1). We observe a 10% overestimation of the causal effect for fully overlapping samples, and a 15% underestimation of the causal effect for non-overlapping samples, due to a combination of winner's curse and weak instrument bias. As expected from Equation 10, the bias is larger when using less stringent thresholds T. For all the thresholds tested, the ratio of the between groups and the within groups variances is larger (up to 26 times) for observed effects than for corrected effects (Table 2), highlighting the differences in observed effects when estimated using different degrees of overlaps. The fact that the within group variance is 1.3 times higher for the corrected effects is due to the slightly increased variance of the bias-corrected estimator. The RMSE of the observed effects is very dependent on the degree of overlap (being larger for non-overlapping and fully overlapping samples) while the RMSE of the corrected effects is consistent across varying degrees of overlap and up to 1.75 times lower for non-overlapping samples (Figure 2 - B). The corrected effects are significantly different from the observed effects for all overlaps values except 50% and all thresholds (Table S1). The absolute bias of the observed effects goes up to 0.03 while for the corrected effect it is smaller than 0.012 for all overlaps and thresholds.

The bias of observed effects depends on the strength of the confounder. If the confounder is weak, observed effects are mostly biased toward the null for low overlaps because of winner's curse and weak instrument bias (Figure S5, Table S2, Table S3). When we simulate a stronger confounder, the bias of the observed effects for fully overlapping samples increases (Figure S6, Table S4, Table S5). In both cases, the corrected and the observed effects are significantly

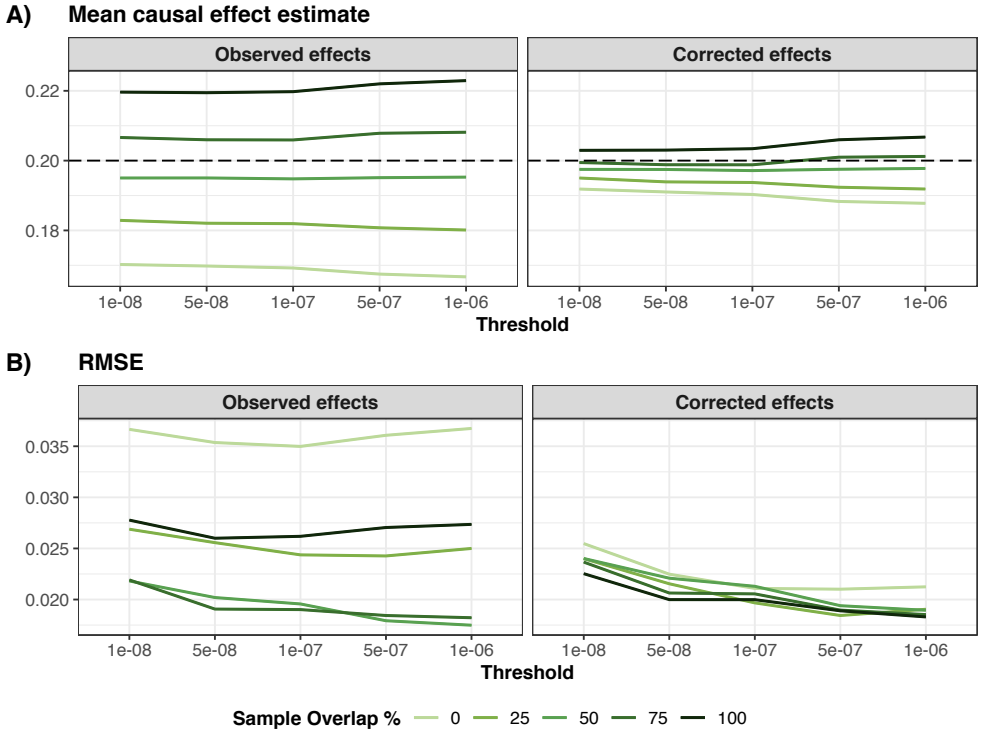


Figure 2: Simulation results for standard settings
$n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0.2$
Panel A) shows the mean observed and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represent the true causal effect). Panel B) shows the mean RMSE obtained for observed and corrected effect for each overlap and threshold.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	51.5	0.000437	0.0375	86	0.000558	0.00178	3.20
5e-08	65.3	0.000339	0.0380	112	0.000434	0.00212	4.88
1e-07	72.3	0.000308	0.0391	127	0.000394	0.00248	6.30
5e-07	92.5	0.000260	0.0463	178	0.000331	0.00484	14.64
1e-06	102.9	0.000251	0.0493	196	0.000320	0.00562	17.56

Table 2: Analysis of variance for standard settings
$n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0.2$
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.

different for almost all overlaps and thresholds, and the corrected effects are substantially less biased than the observed effects (Table S2, Table S4). When the confounder effect (ρ) and the causal effect (α) are acting in opposite directions, the results are particularly interesting because winner's curse, weak instrument bias and sample overlap are all biasing the results toward the null (Figure 3 - A, Table S6). In this case, observed effects are more similar across the different

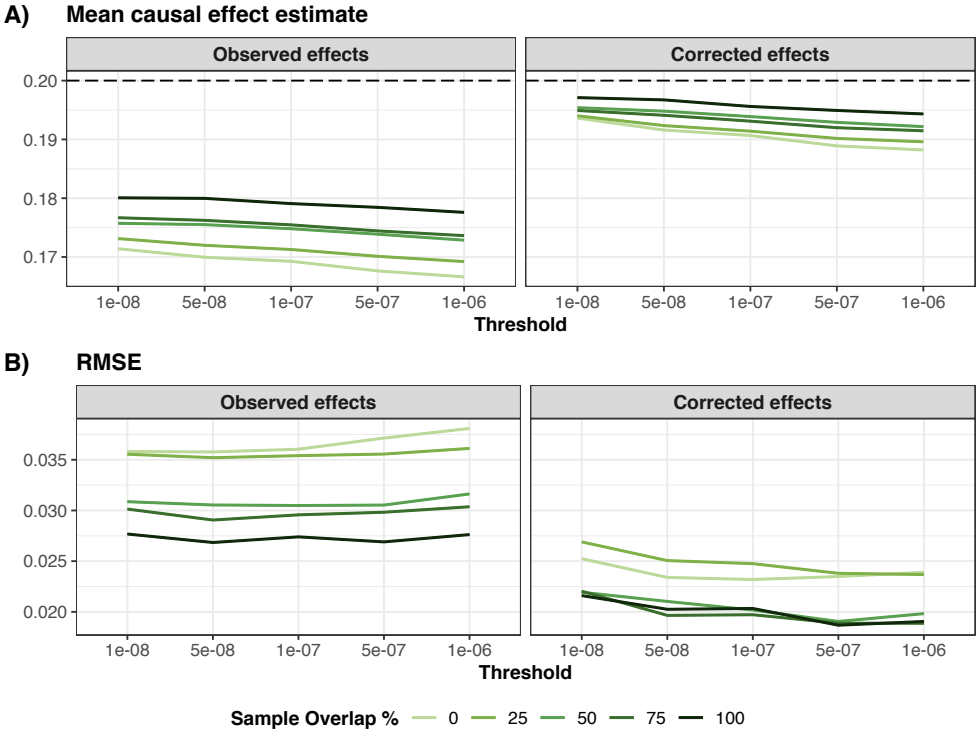


Figure 3: Simulation results for a scenario with a negative confounder $n_A = n_B = 20,000, \pi_x = 0.001, h^2_x = 0.4, \kappa_x = -0.3, \kappa_y = 0.5, \alpha = 0.2$
Panel A) shows the mean observed and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represent the true causal effect). Panel B) shows the mean RMSE obtained for observed and corrected effect for each overlap and threshold.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	51.2	0.000425	0.00112	2.63	0.000538	0.000187	0.348
5e-08	65.3	0.000356	0.00152	4.26	0.000447	0.000413	0.925
1e-07	72.7	0.000336	0.00146	4.35	0.000423	0.000389	0.920
5e-07	92.7	0.000293	0.00174	5.94	0.000369	0.000552	1.496
1e-06	102.7	0.000292	0.00178	6.10	0.000370	0.000559	1.512

Table 3: Analysis of variance for a scenario with a negative confounder $n_A = n_B = 20,000, \pi_x = 0.001, h^2_x = 0.4, \kappa_x = -0.3, \kappa_y = 0.5, \alpha = 0.2$ For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.

degrees of sample overlap tested, but all are underestimating the true causal effect. For this reason, we do not observe such a large decrease in the heterogeneity of the estimates across different sample overlaps compared to the corrected effects (ratio between the between groups and the within groups variance is 5 times larger for observed effects), but the latter ones show lower RMSE and bias for all overlaps and thresholds (Figure 3 - B, Table 3). In this scenario, observed and corrected effects significantly differ for all overlaps and thresholds, with an average

underestimation of 13% for observed effects (Table S6).

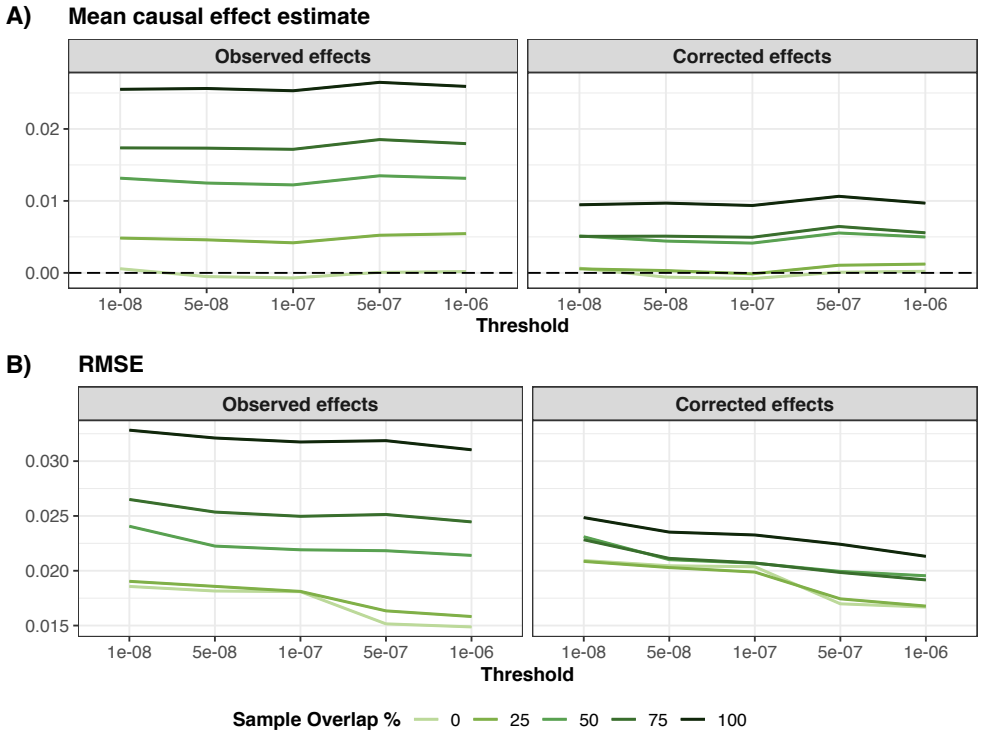


Figure 4: Simulation results in the absence of a causal effect
$n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0$
Panel A) shows the mean observed and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represent the true causal effect). Panel B) shows the mean RMSE obtained for observed and corrected effect for each overlap and threshold.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	52.2	0.000387	0.00985	25.4	0.000485	0.00138	2.84
5e-08	65.3	0.000345	0.01064	30.8	0.000431	0.00170	3.96
1e-07	72.0	0.000336	0.01064	31.6	0.000420	0.00171	4.08
5e-07	92.1	0.000276	0.01099	39.8	0.000343	0.00184	5.35
1e-06	102.2	0.000262	0.01029	39.3	0.000326	0.00143	4.40

Table 4: Analysis of variance in the absence of a causal effect
$n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0$
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.

In the absence of a true causal relationship between the exposure and the outcome, observed effects from non-overlapping samples are unbiased. However, for fully overlapping samples, the

observed effects are biased toward the confounder's effect (Figure 4, Table 4). We showed that for large overlap percentages ($\geq 50\%$), the corrected effects are significantly different from the observed effects (60% smaller), and they are less biased for all overlaps and thresholds (Table S7).

Results obtained using more realistic parameters in terms of sample sizes, genetic architecture and causal effect strength show a similar pattern. We observed an important bias of observed causal effects, mostly when estimated from non-overlapping (22% underestimation) or fully overlapping samples (30% overestimation), that is strongly reduced when using our correction (Figure S7, Table S8, Table S9). Corrected effects significantly differ from observed effects for the most extreme overlaps values (0%, 75%, 100%) for which corrected effects are on average 5 times less biased than observed effects.

3.3 Application to UKBB

We tested our method on UKBB obesity-related traits using a similar approach and splitting the full dataset into samples of varying degrees of overlap. When looking at the observed effect of BMI on SBP (Figure 5), we observed that the estimates obtained using different p-value thresholds vary considerably, independently of sample overlap. Even though we expect an increase in bias when reducing the threshold used, as shown in simulations, here we see that for non-overlapping sample the observed effect is larger for less stringent thresholds. This is inconsistent with winner's curse and weak instrument bias that lean the estimate toward the null and we would expect the observed effects for less stringent thresholds to be smaller (in absolute value). Hence, we believe that this phenomenon is not related to any of the biases discussed here and is due to other reasons such as the existence of multiple causal effects depending on exposure sub-type or the presence of a heritable confounder (see Discussion). Here we will focus on the results obtained using a p-value threshold of 5e-8.

When using IVs reaching genome-wide significance, the observed effects range between 0.095 for non-overlapping samples, and 0.13 for fully overlapping samples (Table S10). After correction, the range of the estimated effect is about 2 times smaller (0.108 - 0.125). The better agreement of corrected effects across overlaps can be seen by looking at the ratio between the between groups and the within groups variance which is 4 times larger for observed effects (Table S11). For this relationship, winner's curse and weak instrument bias seem to be the strongest sources of bias, arising when using non-overlapping samples. The difference between observed and corrected effect is significant ($p_{\text{diff}} = 0.0048$), and standard two-sample MR underestimates the causal effect by about 25% (the observed effect is 0.095 while the corrected effect is 0.126). For fully overlapping samples, there is no significant difference between observed and corrected effects. We observed a similar pattern when investigating the effect of BMI on smoking (Figure S8, Table S12 , Table S13) where the largest bias occurs for non-overlapping samples. The observed causal effect at p=5e-8 is 0.131 for non-overlapping samples while results after correction point toward a causal effect of 0.172 (underestimation of 24% - significant difference between observed and corrected effects, $p_{\text{diff}} = 0.0115$). We do not see a significant difference between observed and corrected effects for larger overlap values, but it is important to note that in this case, because

of the impact of missing data on our design, the largest possible overlap was only 48.5%.

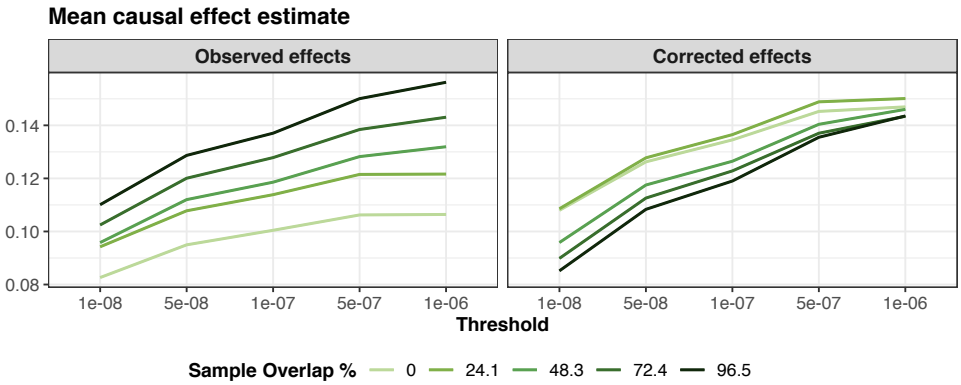


Figure 5: Effect of BMI on SBP
This figure shows the mean observed and corrected effect for each overlap and threshold obtained from 100 different sampled datasets.

When investigating the causal effect of SBP on BMI, the observed effects using a p-value threshold of 5e-8 were ranging from -0.0454 (for non overlapping samples) to 0.0031 (for fully overlapping samples) (Figure 6, Table S14). However, the corrected effects for different overlaps show a strong agreement, with a ratio between the between groups and the within groups variance that is is 360 times larger than that of observed effects (Table S15). The corrected effects across all degrees of sample overlap point toward a causal effect of -0.065 with the difference between observed and corrected effects being significant for larger overlaps ($p_{\text{diff}} = 0.028$ for 72.4% overlap, and $p_{\text{diff}} = 0.012$ for 96.5% overlap). It is known that SBP and BMI share risk factors, such as sedentary lifestyle[20], that would increase both SBP and BMI. Such a confounder would induce a positive environmental correlation between the traits, leading to a bias toward the null for all overlaps (all corrected effects are stronger than observed effects). We have shown that the extent of the bias induced by the existence of a confounder when using overlapping samples depends on several parameters, including the genetic architecture of the exposure. This could explain why we see a significant effect of the correction for fully overlapping samples when looking at the effect of SBP on BMI and not for the reverse direction (BMI on SBP). With the heritability of BMI being larger than that of SBP, we expect a smaller bias when this trait is used as an exposure. Our results also implicate the existence the existence of an environmental confounder biasing the causal effect estimate of BMI on alcohol consumption. At p=5e-8, observed effects range between -0.187 for non overlapping samples and -0.218 for fully overlapping samples, whereas the corrected effects are larger (-0.253 to -0.329) (Figure S9, Table S16). In this case however, the between groups variance is larger for corrected effects (Table S17). The difference between observed and corrected effects is significant for all overlaps, and the corrected effects being stronger than the observed ones hints at the existence of an environmental confounder having a concordant effect on BMI and alcohol frequency intake, biasing all estimates toward the null (as shown in simulations, Figure 3). While we could not identify any plausible confounder of

this relationship, its existence is supported by the fact that the phenotypic correlation between BMI and alcohol frequency (-0.13 among the 379,530 genetically British individuals in the UK Biobank) is weaker than the standardised causal effect.

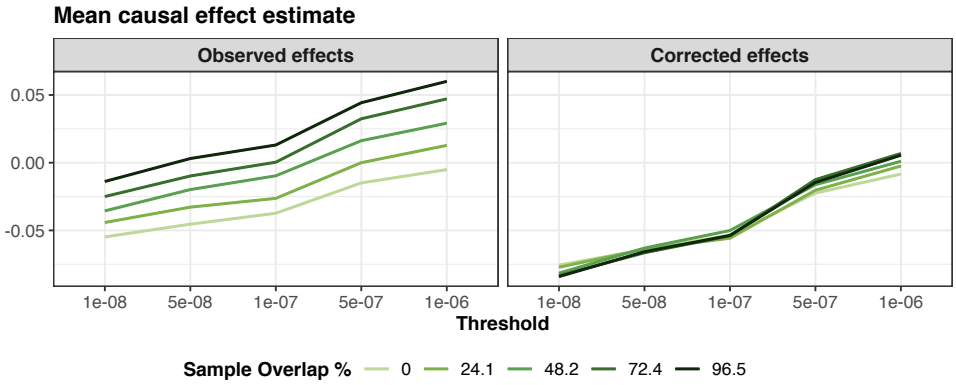


Figure 6: Effect of SBP on BMI
This figure shows the mean observed and corrected effect for each overlap and threshold obtained from 100 different sampled datasets.

4 Discussion

We developed a method that simultaneously accounts for biases arising from winner's curse, weak instruments and sample overlap, even when the degree of overlap is unknown. To the best of our knowledge, no other method can handle all these biases simultaneously while using summary statistics from exposure and outcome GWASs with arbitrary sample overlap. We tested our approach using a wide range of simulations scenarios: varying the strength of the causal effect, the strength of the confounder effect, sample sizes for the exposure and the outcome, as well as the genetic architecture of the exposure and demonstrated that both estimates for non-overlapping and fully overlapping samples can be biased. The direction and the magnitude of the bias depends on sample overlap and is strongly influenced by the effect of the confounder. When the confounder and the causal effect are acting in the same direction, observed effects are overestimated for fully-overlapping samples and underestimated for non-overlapping samples. However, when they are acting in opposite directions, these are underestimated for all overlaps because the direction of the three sources of biases is toward the null. We also showed that in the absence of a causal effect, results from overlapping samples would be biased, potentially leading to elevated type I error.

The correction we proposed worked remarkably well under all scenarios, and allows to drastically reduce the bias. For standard settings for example, we observed a 15% overestimation for fully overlapping samples and 10% underestimation for non-overlapping samples, that were respectively reduced to a 5% overestimation and a 2% underestimation after correction. We also found significant differences between observed and corrected effects for fully-overlapping samples under all scenarios. For non-overlapping samples, observed and corrected effects were significantly different under all scenarios expect in the case of the absence of causal effect. The decreased bias comes with increased variance (due to the correction) however in all of our simulation scenarios it reduced estimation error (RMSE) for at least one, if not all, sample overlap degrees. Moreover, while the RMSE of observed effects strongly depends on the degree of overlap (because the bias is overlap-dependent), the RMSE of corrected effects is very similar for all overlaps.

For real data, strong discrepancies were observed for low degrees overlap (i.e. winner's Curse and weak instrument bias) in three out of the four relationships we looked at, with significant differences between observed and corrected effects. This means that standard two-sample MR settings are often leading to an underestimation of the true causal effect, that can be corrected using our approach. We also demonstrated that while most studies are extremely keen on avoiding any sample overlap while performing two-sample MR analysis fearing potential bias, it is often much less substantial than weak instrument bias or winner's curse. In addition, we identified two trait pairs (SBP→BMI and BMI→alcohol use frequency) for which estimates for fully-overlapping samples are biased by an environmental confounder. In both cases, the confounder and the causal effect were acting in opposite direction, leading to an underestimation of the IVW-based effect for all overlaps. We have also highlighted that there is an important heterogeneity in

causal effect estimates that vary with the IV selection threshold, due to heterogeneity in the estimates between the groups of genetic variants used for different thresholds. This can happen if there is a strong phenotypic heterogeneity in the exposure, in which case different groups of IVs could be affecting the exposure through different pathways[21]. Alternatively, in the presence of a genetic confounder, IVs picked up at a less stringent thresholds may be associated to the confounder, hence violating the second assumption of MR. Such phenomenon is out of the scope of our paper. In such case, IVW two-sample MR estimates would be biased, and more sophisticated approaches either specifically accounting for this genetic confounding (CAUSE[22], LHC-MR[23]) or others allowing for multiple causal effects (MR-Clust[21]) would be needed.

Our approach has its own limitations. As IVW-MR estimates, our corrected effect estimates will also be biased in case of the existence of a genetic confounder through which some of the selected instruments are primarily acting on the exposure. In addition, our analytical derivation hinges on a genetic architecture of the exposure, namely assuming a spike and slab distribution of the multivariable effect sizes. Although this is a widely used and confirmed polygenic model, deviations from it could reduce the efficiency of our bias correction. Finally, our work focused on continuous traits, and we did not explore the effect of sample overlap when using case-control designs. To do this, we would need to consider that the overlap degree might differ between cases and controls.

The effect of these three sources of bias will decrease as the sample size increases, and for very large sample sizes, such as UKBB, this bias is expected to be less pronounced. However, we showed that the overall bias is also dependent on the genetic architecture of the exposure, and low heritability and/or highly polygenic traits will be more strongly affected. Estimating the corrected effect using our approach (implemented in an R-package to facilitate its use) can be performed as a sensitivity analysis: if the corrected effect does not significantly differ from the observed effect, then the IVW-MR estimate can be safely used (with the advantage of having lower SE). However, if there is a significant difference, corrected effects should be preferred as they are be less biased, independently of the sample overlap.

Acknowledgements

The authors thank Chiara Auwerx, Liza Darrous, Sven Erik Ojavee, Marion Patxot, Eleonora Porcu, Marie Sadler and Tabea Schoeler for the helpful discussions and relevant comments.
Z.K. was funded by the Swiss National Science Foundation (# 310030-189147).
Computations have been performed on the HPC cluster of the Lausanne University Hospital.

Code availability

The `MRlap` R-package is freely available under a GPL-2 License, and can be accessed, alongside with user guides and tutorials, from https://github.com/n-mounier/MRlap.

References

[1] Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statistics in medicine.

[2] Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. Genetic Epidemiology *37*, 658–665.

[3] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. American journal of human genetics.

[4] Burgess, S., Davies, N. M., and Thompson, S. G. (2016). Bias due to participant overlap in two-sample mendelian randomization. Genetic Epidemiology *40*, 597–608.

[5] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine.

[6] Burgess, S. and Thompson, S. G. (2011). Bias in causal estimates from mendelian randomization studies with weak instruments. Statistics in Medicine *30*, 1312–1323.

[7] Burgess, S. and and, S. G. T. (2011). Avoiding bias from weak instruments in mendelian randomization studies. International Journal of Epidemiology *40*, 755–764.

[8] Zheng, J., Baird, D., Borges, M.-C., Bowden, J., Hemani, G., Haycock, P., Evans, D. M., and Smith, G. D. (2017). Recent developments in mendelian randomization studies. Current Epidemiology Reports *4*, 330–345.

[9] Bowden, J., M., F. D. G., Minelli, C., Smith, G. D., Sheehan, N. A., and Thompson, J. R. (2016). Assessing the suitability of summary data for two-sample mendelian randomization analyses using MR-egger regression: the role of the i2 statistic. International Journal of Epidemiology pp. dyw220.

[10] Bowden, J., M, F. D. G., Minelli, C., Smith, G. D., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data mendelian randomization. Statistics in Medicine *36*, 1783–1802.

[11] Palmer, C. and Pe'er, I. (2017). Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. PLOS Genetics *13*, e1006916.

[12] Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics *9*, 621–634.

[13] Bowden, J., Smith, G. D., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. International Journal of Epidemiology *44*, 512–525.

[14] Bulik-Sullivan, B., , Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nature Genetics *47*, 1236–1241.

[15] Consortium, I. H. . (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

[16] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature.

[17] Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. eLife.

[18] Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., Ip, H. F., Marioni, R. E., McIntosh, A. M., Deary, I. J., et al. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nature Human Behaviour *3*, 513–525.

[19] Bulik-Sullivan, B. K., , Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics *47*, 291–295.

[20] Sohn, M.-W., Manheim, L., Chang, R., Greenland, P., Hochberg, M., Nevitt, M., Semanik, P., and Dunlop, D. (2014). Sedentary behavior and blood pressure control among osteoarthritis initiative participants. Osteoarthritis and Cartilage *22*, 1234–1240.

[21] Foley, C. N., Mason, A. M., Kirk, P. D. W., and Burgess, S. (2020). MR-clust: clustering of genetic variants in mendelian randomization with similar causal estimates. Bioinformatics.

[22] Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nature Genetics *52*, 740–747.

[23] Darrous, L., Mounier, N., and Kutalik, Z. (2020). Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. medRxiv.

Supplementary Materials

A. Derivation of the corrected effect estimate

Let X and Y denote two random variables representing two complex traits. We intend to use MR to estimate the causal effect of X on Y. We will use m Linkage Disequilibrium (LD)-independent genetic markers as instrumental variables (IVs). Their corresponding genotype data is denoted by G and its jth column by g_j (rows representing individuals). To simplify notation we assume that $E(\mathsf{X}) = E(\mathsf{Y}) = E(\mathsf{G}) = 0$ and $Var(\mathsf{X}) = Var(\mathsf{Y}) = Var(\mathsf{G}) = 1$. Let us assume that X is observed in sample A of sample size n_A, Y is measured in sample B of sample size n_B with an overlap of $n_{A\cap B}$ individuals between the two samples. The vector of realisations of Z^C is denoted by $\boldsymbol{z}^C$ for all variables ($Z = \mathsf{X}, \mathsf{Y}, \mathsf{G}, \mathsf{g}$) and samples ($C = A, B, A \cap B$). Let us assume the following models:

$$\begin{aligned} \boldsymbol{x}^A &= G^A \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^A \\ \boldsymbol{x}^B &= G^B \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^B \\ \boldsymbol{y}^B &= \alpha \cdot \boldsymbol{x}^B + G^B \cdot \boldsymbol{\gamma_y} + \boldsymbol{\epsilon}_{\boldsymbol{y}}^B \end{aligned} \tag{S1}$$

where $\boldsymbol{\gamma_x}$ are the effect sizes of the instruments on X, $\boldsymbol{\gamma_y}$ are their pleiotropic effects on Y. Assuming that there is a single environmental confounder U acting linearly on both traits (as used for simulations) the error term can split into two parts : $\boldsymbol{\epsilon}_{\boldsymbol{x}}^C = \kappa_x \cdot \boldsymbol{u}^C + \boldsymbol{\varepsilon}^C$ and $\boldsymbol{\epsilon}_{\boldsymbol{y}}^C = \kappa_y \cdot \boldsymbol{u}^C + \boldsymbol{\varepsilon}^C$, where κ_x and κ_y refer to the effect of U on X and Y respectively, ε^C is independent of the confounder and C can take the values A, B or $A \cap B$ as above.
Under the INSIDE assumption[?] (INstrument Strength Independent of Direct Effect, i.e. horizontal pleiotropic effects are independent of the direct effect), $cov(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}) = 0$ and $E[\boldsymbol{\gamma_y}] = 0$. We denote $\rho := cov(\boldsymbol{\epsilon_x}, \boldsymbol{\epsilon_y}) = \kappa_x \cdot \kappa_y$. It corresponds to the part of the phenotypic correlation (r) due to sample overlap ($r = \rho + \alpha$).

a. Inverse Variance Weighted estimate

The GWAS effect size estimates for genetic variant j are as follows:

$$\begin{aligned} (\widehat{\beta}_x^A)_j &= \frac{1}{n_A} \cdot (\boldsymbol{g}_j^A)' \cdot \boldsymbol{x}^A = \frac{1}{n_A} \cdot (\boldsymbol{g}_j^A)' \cdot \left(G^A \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^A\right) \\ (\widehat{\beta}_x^B)_j &= \frac{1}{n_B} \cdot (\boldsymbol{g}_j^B)' \cdot \boldsymbol{x}^B = \frac{1}{n_B} \cdot (\boldsymbol{g}_j^B)' \cdot \left(G^B \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^B\right) \\ (\widehat{\beta}_y^B)_j &= \frac{1}{n_B} \cdot (\boldsymbol{g}_j^B)' \cdot \boldsymbol{y}^B = \frac{1}{n_B} \cdot (\boldsymbol{g}_j^B)' \cdot \left(\alpha \cdot \boldsymbol{x}^B + G^B \cdot \boldsymbol{\gamma_y} + \boldsymbol{\epsilon}_{\boldsymbol{y}}^B\right) \end{aligned} \tag{S2}$$

where the genotype data for genetic variant j for individuals in sample A is denoted by $\boldsymbol{g}_j^A$. To simplify notation, we introduce the following notations:

$$\tau_{x_j}^A := \frac{1}{n_A} \cdot (\boldsymbol{g}_j^A)' \cdot \left((G^A)_{-j} \cdot \boldsymbol{\gamma_{x-j}} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^A\right)$$

$$\begin{aligned}\tau^B_{x_j} &:= \frac{1}{n_B}\cdot(\boldsymbol{g}^B_j)'\cdot\left((G^B)_{-j}\cdot\boldsymbol{\gamma}_{\boldsymbol{x}-j}+\boldsymbol{\epsilon}^B_{\boldsymbol{x}}\right) \\ \tau^B_{y_j} &:= \frac{1}{n_B}\cdot(\boldsymbol{g}^B_j)'\cdot\left((G^B)_{-j}\cdot\boldsymbol{\gamma}_{\boldsymbol{y}-j}+\boldsymbol{\epsilon}^B_{\boldsymbol{y}}\right)\end{aligned} \tag{S3}$$

where $(\cdot)_{-j}$ subscripts refer to the full set of data except the jth element (column in case of matrices). Now we can reformulate the effect size equation (Equation S2) as follows:

$$\begin{aligned}(\widehat{\beta}^A_x)_j &= \gamma_{x_j}+\tau^A_{x_j} \\ (\widehat{\beta}^B_x)_j &= \gamma_{x_j}+\tau^B_{x_j} \\ (\widehat{\beta}^B_y)_j &= \alpha\cdot(\gamma_{x_j}+\tau^B_{x_j})+\gamma_{y_j}+\tau^B_{y_j}\end{aligned} \tag{S4}$$

In the following, we will work out the first two moments of the τ variables. First, since $\boldsymbol{g}_j$ is orthogonal to G_{-j} and all ϵ variables, the expectation of all τ variables is zero. Their variances can be calculated as follows:

$$\begin{aligned}Var(\tau^A_{x_j}) &= Var\left(\frac{1}{n_A}\cdot(\boldsymbol{g}^A_j)'\cdot\left((G^A)_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j}+\boldsymbol{\epsilon}^A_{\boldsymbol{x}}\right)\right) \\ &= \frac{1}{n_A}\cdot Var(\boldsymbol{g}_j)\cdot Var\left(\mathsf{G}_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j}+\boldsymbol{\epsilon}^A_{\boldsymbol{x}}\right) \\ &= \frac{1-\gamma^2_{x_j}}{n_A}\end{aligned} \tag{S5}$$

$$\begin{aligned}Var(\tau^B_{x_j}) &= Var\left(\frac{1}{n_B}\cdot(\boldsymbol{g}^B_j)'\cdot\left((G^B)_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j}+\boldsymbol{\epsilon}^B_{\boldsymbol{x}}\right)\right) \\ &= \frac{1}{n_B}\cdot Var(\boldsymbol{g}_j)\cdot Var\left(\mathsf{G}_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j}+\boldsymbol{\epsilon}^B_{\boldsymbol{x}}\right) \\ &= \frac{1-\gamma^2_{x_j}}{n_B}\end{aligned} \tag{S6}$$

$$\begin{aligned}Var(\tau^B_{y_j}) &= Var\left(\frac{1}{n_B}\cdot(\boldsymbol{g}^B_j)'\cdot\left((G^B)_{-j}\cdot\boldsymbol{\gamma}_{\boldsymbol{y}-j}+\boldsymbol{\epsilon}^B_{\boldsymbol{y}}\right)\right) \\ &= \frac{1}{n_B}\cdot Var(\boldsymbol{g}_j)\cdot Var\left(\mathsf{G}_{-j}\boldsymbol{\gamma}_{\boldsymbol{y}-j}+\boldsymbol{\epsilon}^B_{\boldsymbol{y}}\right) \\ &= \frac{1-\alpha^2-\gamma^2_{y_j}}{n_B}\end{aligned} \tag{S7}$$

Expanding the expressions for the τs, Equation S3 becomes:

$$\begin{aligned}\tau^A_{x_j} &= \frac{1}{n_A}\cdot\left\{(\boldsymbol{g}^{A\setminus B}_j)'\cdot\left((G^{A\setminus B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j}+\boldsymbol{\epsilon}^{A\setminus B}_{\boldsymbol{x}}\right)+(\boldsymbol{g}^{A\cap B}_j)'\cdot\left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j}+\boldsymbol{\epsilon}^{A\cap B}_{\boldsymbol{x}}\right)\right\} \\ \tau^B_{x_j} &= \frac{1}{n_B}\cdot\left\{(\boldsymbol{g}^{B\setminus A}_j)'\cdot\left((G^{B\setminus A})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j}+\boldsymbol{\epsilon}^{B\setminus A}_{\boldsymbol{x}}\right)+(\boldsymbol{g}^{A\cap B}_j)'\cdot\left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j}+\boldsymbol{\epsilon}^{A\cap B}_{\boldsymbol{x}}\right)\right\} \\ \tau^B_{y_j} &= \frac{1}{n_B}\cdot\left\{(\boldsymbol{g}^{B\setminus A}_j)'\cdot\left((G^{B\setminus A})_{-j}\boldsymbol{\gamma}_{\boldsymbol{y}-j}+\boldsymbol{\epsilon}^{B\setminus A}_{\boldsymbol{y}}\right)+(\boldsymbol{g}^{A\cap B}_j)'\cdot\left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{y}-j}+\boldsymbol{\epsilon}^{A\cap B}_{\boldsymbol{y}}\right)\right\}\end{aligned} \tag{S8}$$

Exploiting the fact that the covariance between quantities derived from non-overlapping samples is zero, it enables us to work out the pairwise covariances as follows:

$$\begin{aligned} cov(\tau^A_{x_j}, \tau^B_{x_j}) &= E\left\{ \frac{1}{n_A} \cdot \left\{ (\boldsymbol{g}_j^{A\setminus B})' \cdot \left((G^{A\setminus B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^{A\setminus B} \right) + (\boldsymbol{g}_j^{A\cap B})' \cdot \left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^{A\cap B} \right) \right\} \times \right. \\ &\quad \left. \frac{1}{n_B} \cdot \left\{ (\boldsymbol{g}_j^{B\setminus A})' \cdot \left((G^{B\setminus A})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^{B\setminus A} \right) + (\boldsymbol{g}_j^{A\cap B})' \cdot \left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^{A\cap B} \right) \right\} \right\} \\ &= \frac{1}{n_A \cdot n_B} \cdot E\left\{ \left\{ (\boldsymbol{g}_j^{A\cap B})' \cdot \left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^{A\cap B} \right) \right\}^2 \right\} \\ &= \frac{n_{A\cap B}}{n_A \cdot n_B} \cdot Var(\boldsymbol{g}_j) \cdot Var\left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^{A\cap B} \right) \qquad \text{(S9)} \\ &= \frac{n_{A\cap B}}{n_A \cdot n_B} \cdot (1 - \gamma^2_{x_j}) \end{aligned}$$

Similarly for $cov(\tau^A_{x_j}, \tau^B_{y_j})$, using the INSIDE assumption, i.e. that $\boldsymbol{\gamma}_{\boldsymbol{x}}$ and $\boldsymbol{\gamma}_{\boldsymbol{y}}$ are uncorrelated, we have:

$$\begin{aligned} cov(\tau^A_{x_j}, \tau^B_{y_j}) &= E\left\{ \frac{1}{n_A} \cdot \left\{ (\boldsymbol{g}_j^{A\setminus B})' \cdot \left((G^{A\setminus B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^{A\setminus B} \right) + (\boldsymbol{g}_j^{A\cap B})' \cdot \left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^{A\cap B} \right) \right\} \times \right. \\ &\quad \left. \frac{1}{n_B} \cdot \left\{ (\boldsymbol{g}_j^{B\setminus A})' \cdot \left((G^{B\setminus A})_{-j}\boldsymbol{\gamma}_{\boldsymbol{y}-j} + \boldsymbol{\epsilon}_{\boldsymbol{y}}^{B\setminus A} \right) + (\boldsymbol{g}_j^{A\cap B})' \cdot \left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{y}-j} + \boldsymbol{\epsilon}_{\boldsymbol{y}}^{A\cap B} \right) \right\} \right\} \\ &= \frac{1}{n_A \cdot n_B} \cdot E\left\{ \left\{ (\boldsymbol{g}_j^{A\cap B})' \cdot \left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{x}-j} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^{A\cap B} \right) \right\} \times \left\{ (\boldsymbol{g}_j^{A\cap B})' \cdot \left((G^{A\cap B})_{-j}\boldsymbol{\gamma}_{\boldsymbol{y}-j} + \boldsymbol{\epsilon}_{\boldsymbol{y}}^{A\cap B} \right) \right\} \right\} \\ &= \frac{n_{A\cap B}}{n_A \cdot n_B} \cdot Var(\boldsymbol{g}_j) \cdot E\left(\boldsymbol{\epsilon}_{\boldsymbol{x}} \cdot \boldsymbol{\epsilon}_{\boldsymbol{y}} \right) \qquad \text{(S10)} \\ &= \frac{n_{A\cap B}}{n_A \cdot n_B} \cdot \rho \end{aligned}$$

We are now in position to compute the covariance between $(\widehat{\beta}^A_x)_j$ and $(\widehat{\beta}^B_y)_j$:

$$\begin{aligned} cov((\widehat{\beta}^A_x)_j, (\widehat{\beta}^B_y)_j) &= E\left(((\widehat{\beta}^A_x)_j - \gamma_{x_j}) \cdot ((\widehat{\beta}^B_y)_j - \alpha \cdot \gamma_{x_j} - \gamma_{y_j}) \right) \\ &= E\left(\tau^A_{x_j} \cdot (\alpha \cdot \tau^B_{x_j} + \tau^B_{y_j}) \right) \\ &= \alpha \cdot cov(\tau^A_{x_j}, \tau^B_{x_j}) + cov(\tau^A_{x_j}, \tau^B_{y_j}) \qquad \text{(S11)} \\ &= \frac{n_{A\cap B}}{n_A \cdot n_B} \cdot (\alpha \cdot (1 - \gamma^2_{x_j}) + \rho) \end{aligned}$$

Let us now consider the fixed-effect inverse-variance weighting meta-analysis for the ratio estimates for the causal effect α. Each IV j provides a ratio estimate:

$$\widehat{\alpha}_j = \frac{(\widehat{\beta}^B_y)_j}{(\widehat{\beta}^A_x)_j} \qquad \text{(S12)}$$

$$Var(\widehat{\alpha}_j) = \frac{Var((\widehat{\beta}^B_y)_j)}{(\widehat{\beta}^A_x)^2_j} = \frac{(1 - \alpha^2 - \gamma^2_{y_j})}{n_B \cdot (\widehat{\beta}^A_x)^2_j} \qquad \text{(S13)}$$

Hence the weights (w_j) of IV j for estimating the IVW causal effect are of the form:

$$w_j = \frac{1}{Var(\widehat{\alpha}_j)} = \frac{n_B \cdot (\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)} \tag{S14}$$

Finally, the estimate can be written in the following form:

$$\begin{aligned}\widehat{\alpha}_{IVW} &= \frac{\sum_{k=1}^m \widehat{\alpha}_j \cdot w_j}{\sum_{k=1}^m w_j} = \frac{\sum_{j=1}^m \frac{(\widehat{\beta}_y^B)_j}{(\widehat{\beta}_x^A)_j} \times \left(\frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)}\right)}{\sum_{k=1}^m \frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)}} = \frac{\sum_{k=1}^m \frac{(\widehat{\beta}_y^B)_j \cdot (\widehat{\beta}_x^A)_j}{(1-\alpha^2-\gamma_{y_j}^2)}}{\sum_{k=1}^m \frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)}} \\ &\approx \frac{\sum_{j=1}^m \frac{(\widehat{\beta}_y^B)_j \cdot (\widehat{\beta}_x^A)_j}{(1-\alpha^2)}}{\sum_{k=1}^m \frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2)}} = \frac{\sum_{k=1}^m (\widehat{\beta}_y^B)_j \cdot (\widehat{\beta}_x^A)_j}{\sum_{k=1}^m (\widehat{\beta}_x^A)_j^2}\end{aligned} \tag{S15}$$

Here, the last approximation is based on the realistic assumption that the individual pleiotropic effect of each SNP is very small.

b. Expectation of the estimate

In reality we select IVs based on their estimated test statistic, typically $|(\widehat{\beta}_x^A)_j| \cdot \sqrt{n_A} > T$ with $T = -\Phi^{-1}(5 \times 10^{-8}/2) \approx 5.45$, representing the genome-wide significance threshold. By denoting $S_j := \left\{|(\widehat{\beta}_x^A)_j| \cdot \sqrt{n_A} > T\right\}$, the causal effect estimate (Equation S15) changes to:

$$\widehat{\alpha}_{IVW} \approx \frac{\sum_{j=1}^M ((\widehat{\beta}_y^B)_j \,|S_j) \cdot ((\widehat{\beta}_x^A)_j \,|S_j) \cdot Pr(S_j)}{\sum_{j=1}^M ((\widehat{\beta}_x^A)_j \,|S_j)^2 \cdot Pr(S_j)} \tag{S16}$$

Note that while m denoted the number of IVs, M represents the number of genome-wide markers from which IVs are selected. The expectation of the causal effect estimate (Equation S16) can be written as:

$$\begin{aligned}E[\hat{\alpha}_{IVW}] &\approx \frac{\sum_{j=1}^m E\left[\left((\alpha \cdot \gamma_{x_j} + \gamma_{y_j}) + \left(\alpha \cdot \tau_{x_j}^B + \tau_{y_j}^B \,|S_j\right)\right) \cdot \left(\gamma_{x_j} + \left(\tau_{x_j}^A \,|S_j\right)\right)\right] \cdot Pr(S_j)}{\sum_{j=1}^m E\left[\left(\gamma_{x_j} + \left(\tau_{x_j}^A \,|S_j\right)\right)^2\right] \cdot Pr(S_j)} \\ &= \frac{\sum_{j=1}^m s_j \cdot Pr(S_j)}{\sum_{j=1}^m t_j \cdot Pr(S_j)}\end{aligned} \tag{S17}$$

Let us expand the jth term of the numerator s_j:

$$\begin{aligned}s_j &= \left(\alpha \cdot \gamma_{x_j}^2 + \gamma_{y_j} \cdot \gamma_{x_j}\right) + \left(\alpha \cdot \gamma_{x_j} + \gamma_{y_j}\right) \cdot E\left[\tau_{x_j}^A \,|S_j\right] + (\alpha \cdot \gamma_{x_j}) \cdot E\left[\tau_{x_j}^B \,|S_j\right] \\ &+ \gamma_{x_j} \cdot E\left[\tau_{y_j}^B \,|S_j\right] + \alpha \cdot E\left[\tau_{x_j}^B \cdot \tau_{x_j}^A \,|S_j\right] + E\left[\tau_{y_j}^B \cdot \tau_{x_j}^A \,|S_j\right]\end{aligned} \tag{S18}$$

Similarly for t_j we have:

$$t_j = \gamma_{x_j}^2 + 2\gamma_{x_j} \cdot E\left[\tau_{x_j}^A \,|S_j\right] + E\left[(\tau_{x_j}^A)^2 \,|S_j\right] \tag{S19}$$

In the following sections we will compute each term of s_j and t_j.

- **Switching from** $E(Z = z|S_j) \cdot Pr(S_j)$ **to** $E(Z = z|S_j^c) \cdot Pr(S_j^c)$

If we define the complement of S_j as $S_j^c := \left\{ -\frac{T}{\sqrt{n_A}} - \gamma_{x_j} \leq \tau_{x_j}^A \leq \frac{T}{\sqrt{n_A}} - \gamma_{x_j} \right\}$, any random variable Z conditional on S_j can be rewritten as follows:

$$\begin{aligned} Pr(Z = z|S_j) \cdot Pr(S_j) &= Pr(\{Z = z\} \cap S_j) = Pr(Z = z) - Pr(\{Z = z\} \cap S_j^c) \\ &= Pr(Z = z) - Pr(\{Z = z\}|S_j^c) \cdot Pr(S_j^c) \end{aligned} \tag{S20}$$

Hence:

$$\begin{aligned} E(Z|S_j) \cdot Pr(S_j) &= \int_{-\infty}^{\infty} z \cdot Pr(Z = z|S_j) \cdot Pr(S_j) dz \\ &= \int_{-\infty}^{\infty} z \cdot \left(Pr(Z = z) - Pr(\{Z = z\}|S_j^c) \cdot Pr(S_j^c) \right) dz \\ &= E[Z] - E[Z|S_j^c] \cdot Pr(S_j^c) \end{aligned} \tag{S21}$$

We first calculate $Pr(S_j^c)$ using integration by substitution ($f(t) = t\sqrt{n_A}$), which will be necessary for all further computations:

$$\begin{aligned} h_j &:= Pr(S_j^c) = \int_{-\frac{T}{\sqrt{n_A}} - \gamma_{x_j}}^{\frac{T}{\sqrt{n_A}} - \gamma_{x_j}} \frac{1}{\sqrt{2\pi \cdot n_A^{-1}}} \cdot \exp\left(-\frac{u^2}{2 \cdot n_A^{-1}} \right) du \\ &= \Phi(T - \sqrt{n_A}\gamma_{x_j}) - \Phi(-T - \sqrt{n_A}\gamma_{x_j}) \end{aligned} \tag{S22}$$

with Φ being the standard normal cumulative distribution function. In the following we will substitute $\tau_x^A, \tau_x^B, \tau_y^B, (\tau_x^B \cdot \tau_x^A)$ and $(\tau_y^B \cdot \tau_x^A)$ for Z in Equation S21 and compute each conditional expectation. Note that when no thresholding is applied ($T = 0$) we have $h_j = 0$.

- **Computation of** $E\left[\tau_{x_j}^A | S_j\right] \cdot Pr(S_j)$

$$E\left[\tau_{x_j}^A | S_j\right] \cdot Pr(S_j) = \left(E[\tau_{x_j}^A] - E[\tau_{x_j}^A|S_j^c] \cdot h_j \right) = -h_j \cdot E[\tau_{x_j}^A|S_j^c] \tag{S23}$$

Using the properties of the truncated normal distribution we get:

$$\begin{aligned} E\left[\tau_{x_j}^A | S_j\right] \cdot Pr(S_j) &= -h \cdot \frac{1}{\sqrt{n_A}} \cdot \frac{\phi(-T - \gamma_{x_j}\sqrt{n_A}) - \phi(T - \gamma_{x_j}\sqrt{n_A}))}{\Phi(T - \gamma_{x_j}\sqrt{n_A}) - \Phi(-T - \gamma_{x_j}\sqrt{n_A})} \\ &= \frac{1}{\sqrt{n_A}} \cdot (\phi(T - \gamma_{x_j}\sqrt{n_A}) - \phi(-T - \gamma_{x_j}\sqrt{n_A})) \end{aligned} \tag{S24}$$

with ϕ being the standard normal probability density function.

- **Computation of** $E\left[(\tau_{x_j}^A)^2 | S_j\right] \cdot Pr(S_j)$

$$E\left[(\tau_{x_j}^A)^2 | S_j\right] \cdot Pr(S_j) = \left(E[(\tau_{x_j}^A)^2] - E[(\tau_{x_j}^A)^2|S_j^c] \cdot h_j \right) = \frac{1}{n_A} - h \cdot E[(\tau_{x_j}^A)^2|S_j^c] \tag{S25}$$

Using the properties of the truncated normal distribution we get:

$$
\begin{aligned}
& E\left[(\tau_{x_j}^A)^2 \,|S_j\right] \cdot Pr(S_j) = \qquad\qquad (S26)\\
= \;& \frac{1}{n_A} - h_j \cdot \frac{1}{n_A} \cdot \left(1 + \frac{(-T-\gamma_{x_j}\sqrt{n_A}) \cdot \phi(-T-\gamma_{x_j}\sqrt{n_A}) - (T-\gamma_{x_j}\sqrt{n_A}) \cdot \phi(T-\gamma_{x_j}\sqrt{n_A}))}{\Phi(T-\gamma_{x_j}\sqrt{n_A}) - \Phi(-T-\gamma_{x_j}\sqrt{n_A})}\right)\\
= \;& \frac{1}{n_A}(1-h_j) - \frac{1}{n_A} \cdot \left((-T-\gamma_{x_j}\sqrt{n_A}) \cdot \phi(-T-\gamma_{x_j}\sqrt{n_A}) - (T-\gamma_{x_j}\sqrt{n_A}) \cdot \phi(T-\gamma_{x_j}\sqrt{n_A}))\right)
\end{aligned}
$$

- **Computation of** $E\left[\tau_{x_j}^B \,|S_j\right] \cdot Pr(S_j)$

We can split the error term $\tau_{x_j}^B$ into $\tau_{x_j}^A$ dependent and independent parts. We have shown above that $cov(\tau_{x_j}^A, \tau_{x_j}^B) = (1-\gamma_{x_j}^2) \cdot n_{A\cap B}/(n_A \cdot n_B)$ and $var(\tau_{x_j}^A) = (1-\gamma_{x_j}^2)/n_A$, thus we have:

$$
\begin{aligned}
\tau_{x_j}^B &= \tau_{x_j}^A \cdot \frac{cov(\tau_{x_j}^A, \tau_{x_j}^B)}{var(\tau_{x_j}^A)} + \eta_x^B \quad \text{with} \quad cov(\eta_x^B, \tau_{x_j}^A) = 0\\
&= \tau_{x_j}^A \cdot \frac{n_{A\cap B}}{n_B} + \eta_x^B \qquad\qquad (S27)
\end{aligned}
$$

This allows us to utilise the formula for $E\left[\tau_{x_j}^A \,|S_j\right] \cdot Pr(S_j)$ to derive $E\left[\tau_{x_j}^B \,|S_j\right] \cdot Pr(S_j)$ as follows:

$$
\begin{aligned}
E\left[\tau_{x_j}^B \,|S_j\right] Pr(S_j) &= \frac{n_{A\cap B}}{n_B} \cdot E\left[\tau_{x_j}^A \,|S\right] Pr(S_j) + E\left[\eta_x^B \,|S_j\right]\\
&= \frac{n_{A\cap B}}{n_B} \cdot E\left[\tau_{x_j}^A \,|S_j\right] Pr(S_j) \qquad\qquad (S28)
\end{aligned}
$$

- **Computation of** $E\left[\tau_{x_j}^B \cdot \tau_{x_j}^A \,|S\right] \cdot Pr(S_j)$

Using the split sample notation helps us to compute the expectation of the $\left(\tau_{x_j}^B \cdot \tau_{x_j}^A \,|S_j\right)$ term analogously to how we did for $\left(\tau_{x_j}^B \,|S_j\right)$:

$$
\begin{aligned}
E\left[\tau_{x_j}^B \cdot \tau_{x_j}^A \,|S_j\right] \cdot Pr(S_j) &= E\left[(\tau_{x_j}^A \cdot \frac{n_{A\cap B}}{n_B} + \eta_x^B) \cdot \tau_{x_j}^A \,|S_j\right] \cdot Pr(S_j)\\
&= \frac{n_{A\cap B}}{n_B} E\left[(\tau_{x_j}^A)^2 \,|S_j\right] \cdot Pr(S_j) \qquad\qquad (S29)
\end{aligned}
$$

- **Computation of** $E\left[\tau_{y_j}^B \,|S_j\right] \cdot Pr(S_j)$ **and** $E\left[\tau_{y_j}^B \cdot \tau_{x_j}^A \,|S_j\right] \cdot Pr(S_j)$

The decomposing $\tau_{y_j}^B = \rho\tau_{x_j}^B + \eta^B$ with $cov(\eta^B, \tau_{x_j}^B) = 0$ allows us to trace back this computation to that of $E\left[\tau_{x_j}^B \,|S_j\right]$ and $E\left[\tau_{x_j}^A \cdot \tau_{x_j}^B \,|S_j\right]$, respectively:

$$
\begin{aligned}
E\left[\tau_{y_j}^B \,|S_j\right] Pr(S_j) &= \rho \cdot E\left[\tau_{x_j}^B \,|S_j\right] Pr(S_j) \qquad\qquad (S30)\\
E\left[\tau_{y_j}^B \cdot \tau_{x_j}^A \,|S_j\right] Pr(S_j) &= \rho \cdot E\left[\tau_{x_j}^B \cdot \tau_{x_j}^A \,|S_j\right] Pr(S_j) \qquad\qquad (S31)
\end{aligned}
$$

- **Evaluation of** s_j

We are now in position to evaluate the expression for s_j:

$$\begin{aligned}
s_j &= \left(\alpha \cdot \gamma_{x_j}^2 + \gamma_{y_j} \cdot \gamma_{x_j}\right) + \left(\alpha \cdot \gamma_{x_j} + \gamma_{y_j}\right) \cdot E\left[\tau_{x_j}^A \,|S_j\right] + (\alpha \cdot \gamma_{x_j}) \cdot E\left[\tau_{x_j}^B \,|S_j\right] + \gamma_{x_j} \cdot E\left[\tau_{y_j}^B \,|S_j\right] \\
&+ \alpha \cdot E\left[\tau_{x_j}^B \cdot \tau_{x_j}^A \,|S_j\right] + E\left[\tau_{y_j}^B \cdot \tau_{x_j}^A \,|S_j\right] \\
&= \left(\alpha \cdot \gamma_{x_j}^2 + \gamma_{y_j} \cdot \gamma_{x_j}\right) + \left(\alpha \cdot \gamma_{x_j} + \gamma_{y_j}\right) \cdot E\left[\tau_{x_j}^A \,|S_j\right] + (\alpha \cdot \gamma_{x_j}) \cdot \frac{n_{A\cap B}}{n_B} \cdot E\left[\tau_{x_j}^A \,|S_j\right] \\
&+ \gamma_{x_j} \cdot \rho \cdot \frac{n_{A\cap B}}{n_B} \cdot E\left[\tau_{x_j}^A \,|S_j\right] + \alpha \cdot \frac{n_{A\cap B}}{n_B} \cdot E\left[(\tau_{x_j}^A)^2 \,|S_j\right] + \rho \cdot \frac{n_{A\cap B}}{n_B} \cdot E\left[(\tau_{x_j}^A)^2 \,|S_j\right] \qquad \text{(S32)} \\
&= \left(\alpha \cdot \gamma_{x_j}^2 + \gamma_{y_j} \cdot \gamma_{x_j}\right) + \left(\alpha \cdot \gamma_{x_j} + \gamma_{y_j} + (\alpha \cdot \gamma_{x_j}) \cdot \frac{n_{A\cap B}}{n_B} + \gamma_{x_j} \cdot \rho \cdot \frac{n_{A\cap B}}{n_B}\right) \cdot E\left[\tau_{x_j}^A \,|S_j\right] \\
&+ (\alpha + \rho) \cdot \frac{n_{A\cap B}}{n_B} \cdot E\left[(\tau_{x_j}^A)^2 \,|S_j\right]
\end{aligned}$$

Let λ' denote a quantity closely related to the cross-trait LD score regression (LDSC) [1] intercept (λ):

$$\lambda' = (\alpha + \rho) \cdot \frac{n_{A\cap B}}{n_A \cdot n_B} = \frac{\lambda}{\sqrt{n_A \cdot n_B}} \qquad \text{(S33)}$$

The expression for s_j (Equation S32) can then be turned into:

$$\begin{aligned}
s_j \cdot Pr(S_j) &= \left(\alpha \cdot \gamma_{x_j}^2 + \gamma_{y_j} \cdot \gamma_{x_j}\right) \cdot (1 - h_j) + \left(\alpha \cdot \gamma_{x_j} + \gamma_{y_j} + \gamma_{x_j} \cdot \lambda' \cdot n_A\right) \cdot E\left[\tau_{x_j}^A \,|S_j\right] \cdot Pr(S_j) \\
&+ \lambda' \cdot n_A \cdot E\left[(\tau_{x_j}^A)^2 \,|S_j\right] \cdot Pr(S_j) \\
&= \left(\alpha \cdot \gamma_{x_j}^2 + \gamma_{y_j} \cdot \gamma_{x_j}\right) \cdot (1 - h_j) \\
&+ \left(\alpha \cdot \gamma_{x_j} + \gamma_{y_j} + \gamma_{x_j} \cdot \lambda' \cdot n_A\right) \cdot \frac{1}{\sqrt{n_A}} \cdot (\phi(T - \gamma_{x_j}\sqrt{n_A}) - \phi(-T - \gamma_{x_j}\sqrt{n_A})) \\
&+ \lambda' \cdot n_A \cdot \frac{1}{n_A}(1 - h_j) \qquad \text{(S34)} \\
&+ \lambda' \cdot n_A \cdot \frac{1}{n_A} \cdot \left((T - \gamma_{x_j}\sqrt{n_A}) \cdot \phi(T - \gamma_{x_j}\sqrt{n_A}) - (-T - \gamma_{x_j}\sqrt{n_A}) \cdot \phi(-T - \gamma_{x_j}\sqrt{n_A})\right) \\
&= \left(\alpha \cdot \gamma_{x_j}^2 + \gamma_{y_j} \cdot \gamma_{x_j}\right) \cdot (1 - h_j) \\
&+ \left(\frac{\alpha\gamma_{x_j} + \gamma_{y_j} + \lambda' \cdot \sqrt{n_A} \cdot T}{\sqrt{n_A}}\right) \cdot \phi(T - \gamma_{x_j} \cdot \sqrt{n_A}) \\
&- \left(\frac{\alpha\gamma_{x_j} + \gamma_{y_j} - \lambda' \cdot \sqrt{n_A} \cdot T}{\sqrt{n_A}}\right) \cdot \phi(-T - \gamma_{x_j} \cdot \sqrt{n_A}) \\
&+ \lambda'(1 - h_j)
\end{aligned}$$

In the following, we will assume the following popular [2] genetic architecture for X :

$$\boldsymbol{\gamma_x} = \boldsymbol{\zeta_x} \odot \boldsymbol{\upsilon_x} \qquad \text{with} \quad \boldsymbol{\zeta_x} \sim \mathcal{B}(1, \pi_x) \quad \text{and} \quad \boldsymbol{\upsilon_x} \sim \mathcal{N}(0, \sigma_x^2) \qquad \text{(S35)}$$

where π_x and σ_x^2 are characteristics of the genetic architecture of trait X (respectively, a measure of the polygenicity and the per-variant heritability).

It allows us to split the sums for zero effect genetic variants and non-zero effect genetic variants.

When $\gamma_{x_j} = 0$ the expression simplifies to:

$$\sum_{j \in \mathcal{M}_0} s_j \cdot Pr(S_j) \quad = \quad M \cdot (1 - \pi_x) \cdot 2\lambda' \cdot (T \cdot \phi(T) + \Phi(-T)) \tag{S36}$$

For non-zero effects we have:

$$\begin{aligned}
\frac{\sum_{j \in \mathcal{M}_1} s_j \cdot Pr(S_j)}{\pi_x \cdot M} &= \int_{-\infty}^{\infty} \frac{1}{\sigma_x} \cdot \phi(\gamma_x/\sigma_x) \cdot \left(\alpha \cdot \gamma_x^2 + \gamma_{y_j} \cdot \gamma_x + \lambda'\right) \\
&\times \left(1 - \Phi(T - \sqrt{n_A}\gamma_x) + \Phi(-T - \sqrt{n_A}\gamma_x)\right) d\gamma_x \\
&+ \int_{-\infty}^{\infty} \frac{1}{\sigma_x} \cdot \phi(\gamma_x/\sigma_x) \\
&\times \left[\left(\frac{\alpha\gamma_x + \gamma + \lambda'\sqrt{n_A}T}{\sqrt{n_A}}\right) \cdot \phi(T - \gamma_x\sqrt{n_A})\right. \\
&- \left.\left(\frac{\alpha\gamma_x + \gamma - \lambda'\sqrt{n_A}T}{\sqrt{n_A}}\right) \cdot \phi(-T - \gamma_x\sqrt{n_A})\right] d\gamma_x \\
&= 2 \cdot \int_{\infty}^{\infty} \left(\alpha \cdot \gamma_x^2 + \lambda'\right) \cdot \frac{1}{\sigma_x} \cdot \phi(\gamma_x/\sigma_x) \cdot \Phi(-T + \sqrt{n_A}\gamma_x) d\gamma_x \\
&+ \frac{\exp\left(-\frac{1}{2} \cdot \frac{T^2}{1+\sigma_x^2 \cdot n_A}\right) \cdot \sqrt{\frac{2}{\pi}} \cdot T \cdot (\alpha \cdot \sigma_x^2 + \lambda' \cdot (1 + \sigma_x^2 \cdot n_A))}{(1 + \sigma_x^2 \cdot n_A)^{(3/2)}} \\
&= 2\lambda' \cdot \Phi\left(-\frac{T}{\sqrt{1 + n_A \cdot \sigma_x^2}}\right) \\
&+ 2\alpha \cdot \int_{\infty}^{\infty} \gamma_x^2 \cdot \frac{1}{\sigma_x} \cdot \phi(\gamma_x/\sigma_x) \cdot \Phi(-T + \sqrt{n_A}\gamma_x) d\gamma_x \\
&+ \frac{\exp\left(-\frac{1}{2} \cdot \frac{T^2}{1+\sigma_x^2 \cdot n_A}\right) \cdot \sqrt{\frac{2}{\pi}} \cdot T \cdot (\alpha \cdot \sigma_x^2 + \lambda' \cdot (1 + \sigma_x^2 \cdot n_A))}{(1 + \sigma_x^2 \cdot n_A)^{(3/2)}}
\end{aligned} \tag{S37}$$

The remaining integral can be solved as follows:

$$\begin{aligned}
\frac{\partial}{\partial x}\left(x \cdot \frac{\phi(x/c)}{c} \cdot \Phi(a + b \cdot x)\right) &= \frac{\phi(x/c)}{c} \cdot \Phi(a + b \cdot x) + x \cdot \left(\frac{1}{c} \cdot \frac{-x}{c^2} \cdot \phi(x/c)\right) \cdot \Phi(a + b \cdot x) \\
&+ x \cdot \frac{\phi(x/c)}{c} \cdot (b \cdot \phi(a + b \cdot x))
\end{aligned} \tag{S38}$$

By integrating both sides w.r.t. x from $-\infty$ to ∞ we get:

$$\begin{aligned}
\left[x \cdot \frac{\phi(x/c)}{c} \cdot \Phi(a + b \cdot x)\right]_{-\infty}^{\infty} &= \int_{-\infty}^{\infty} \frac{\phi(x/c)}{c} \cdot \Phi(a + b \cdot x) dx \\
&+ \int_{-\infty}^{\infty} x \cdot \left(\frac{1}{c} \cdot \frac{-x}{c^2} \cdot \phi(x/c)\right) \cdot \Phi(a + b \cdot x) dx \\
&+ \int_{-\infty}^{\infty} x \cdot \frac{\phi(x/c)}{c} \cdot (b \cdot \phi(a + b \cdot x))\, dx
\end{aligned} \tag{S39}$$

It is easy to see that the quantity on the left hand side is zero (both limits are zero). We use

the following two well-known integral identities:

$$\int_{-\infty}^{\infty} \phi(x) \cdot \Phi(a + b \cdot x) dx = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{S40}$$

$$\int_{-\infty}^{\infty} x \cdot \phi(x) \cdot \phi(a + b \cdot x) dx = -\frac{ab \cdot \exp\left(-\frac{a^2}{2(b^2+1)}\right)}{\sqrt{2\pi}(1+b^2)^{3/2}} \tag{S41}$$

Therefore the above Equation (S39) simplifies to:

$$\begin{aligned} 0 &= \Phi\left(\frac{a}{\sqrt{1+(bc)^2}}\right) \\ &- \frac{1}{c^3} \cdot \int_{-\infty}^{\infty} x^2 \cdot \phi(x/c) \cdot \Phi(a + b \cdot x) dx \\ &- (bc) \cdot \frac{abc \cdot \exp\left(-\frac{a^2}{2((bc)^2+1)}\right)}{\sqrt{2\pi}(1+(bc)^2)^{3/2}} \end{aligned} \tag{S42}$$

Thus we have:

$$\begin{aligned} \int_{-\infty}^{\infty} x^2 \cdot \frac{\phi(x/c)}{c} \cdot \Phi(a + b \cdot x) dx &= c^2 \cdot \Phi\left(\frac{a}{\sqrt{1+(bc)^2}}\right) \\ &- a \cdot b^2 \cdot c^4 \cdot \frac{\exp\left(-\frac{a^2}{2((bc)^2+1)}\right)}{\sqrt{2\pi}(1+(bc)^2)^{3/2}} \end{aligned} \tag{S43}$$

Substituting $a = -T, b = \sqrt{n_A}, c = \sigma_x$ yields:

$$\begin{aligned} \int_{-\infty}^{\infty} x^2 \cdot \frac{\phi(x/\sigma_x)}{\sigma_x} \cdot \Phi(-T + \sqrt{n_A} \cdot x) dx &= \sigma_x^2 \cdot \Phi\left(\frac{-T}{\sqrt{1+n_A\sigma_x^2}}\right) \\ &+ T \cdot n_A \cdot \sigma_x^4 \cdot \frac{\exp\left(-\frac{T^2}{2(n_A\sigma_x^2+1)}\right)}{\sqrt{2\pi}(1+n_A\sigma_x^2)^{3/2}} \end{aligned} \tag{S44}$$

Finally, we can provide a closed form expression for $\sum_j s_j \cdot Pr(S_j)$ as follows:

$$\begin{aligned} \frac{\sum_{j \in \mathcal{M}_1} s_j \cdot Pr(S_j)}{\pi_x \cdot M} &= 2\lambda' \cdot \Phi\left(-\frac{T}{\sqrt{1+n_A \cdot \sigma_x^2}}\right) \\ &+ 2\alpha \cdot \left(\sigma_x^2 \cdot \Phi\left(\frac{-T}{\sqrt{1+n_A\sigma_x^2}}\right) + T \cdot n_A \cdot \sigma_x^4 \cdot \frac{\exp\left(-\frac{T^2}{2(n_A\sigma_x^2+1)}\right)}{\sqrt{2\pi}(1+n_A\sigma_x^2)^{3/2}}\right) \\ &+ \frac{\exp\left(-\frac{1}{2} \cdot \frac{T^2}{1+\sigma_x^2 \cdot n_A}\right) \cdot \sqrt{\frac{2}{\pi}} \cdot T \cdot (\alpha \cdot \sigma_x^2 + \lambda' \cdot (1 + \sigma_x^2 \cdot n_A))}{(1+\sigma_x^2 \cdot n_A)^{(3/2)}} \\ &= 2(\alpha\sigma_x^2 + \lambda') \cdot \Phi\left(-\frac{T}{\sqrt{1+n_A \cdot \sigma_x^2}}\right) \end{aligned} \tag{S45}$$

$$+ \quad 2T \cdot \frac{\exp\left(-\frac{T^2}{2(n_A\sigma_x^2+1)}\right)}{\sqrt{2\pi}(1+n_A\sigma_x^2)^{3/2}} \cdot \left(\alpha \cdot n_A \cdot \sigma_x^4 + \alpha\sigma_x^2 + \lambda' \cdot (1+\sigma_x^2 \cdot n_A)\right)$$

- **Evaluation of t_j**

Similarly, we can compute $t_j \cdot Pr(S_j)$:

$$\begin{aligned}
t_j \cdot Pr(S_j) &= \gamma_{x_j}^2 \cdot (1-h) + 2\gamma_{x_j} \cdot E\left[\tau_{x_j}^A \,|S_j\right] \cdot Pr(S_j) + E\left[(\tau_{x_j}^A)^2 \,|S_j\right] \cdot Pr(S_j) \\
&= \gamma_{x_j}^2 \cdot (1-h) + 2\gamma_{x_j} \cdot \frac{1}{\sqrt{n_A}} \cdot (\phi(T-\gamma_{x_j}\sqrt{n_A}) - \phi(-T-\gamma_{x_j}\sqrt{n_A})) \\
&+ \frac{1}{n_A}(1-h) - \frac{1}{n_A} \cdot \big((-T-\gamma_{x_j}\sqrt{n_A}) \cdot \phi(-T-\gamma_{x_j}\sqrt{n_A}) \\
&- (T-\gamma_{x_j}\sqrt{n_A}) \cdot \phi(T-\gamma_{x_j}\sqrt{n_A}))\big) \\
&= \left(\gamma_{x_j}^2 + \frac{1}{n_A}\right) \cdot (1-h) \\
&+ \frac{1}{n_A} \cdot \big((T+\gamma_{x_j}\sqrt{n_A}) \cdot \phi(T-\gamma_{x_j}\sqrt{n_A}) + (T-\gamma_{x_j}\sqrt{n_A}) \cdot \phi(-T-\gamma_{x_j}\sqrt{n_A}))\big)
\end{aligned} \tag{S46}$$

As before, when $\gamma_{x_j} = 0$ the expression simplifies to:

$$\sum_{j \in \mathcal{M}_0} t_j \cdot Pr(S_j) = M \cdot (1-\pi_x) \cdot \frac{2}{n_A} \cdot (\Phi(-T) + T \cdot \phi(T)) \tag{S47}$$

For non-zero effects we have:

$$\begin{aligned}
\frac{\sum_{j \in \mathcal{M}_1} t_j \cdot Pr(S_j)}{\pi_x \cdot M} &= \int_\infty^\infty (t|\gamma_x) \cdot \frac{1}{\sigma_x} \cdot \phi(\gamma_x/\sigma_x) d\gamma_x \\
&= \int_\infty^\infty \frac{1}{\sigma_x} \cdot \phi(\gamma_x/\sigma_x) \cdot \left(\gamma_x^2 + \frac{1}{n_A}\right) \cdot (1 - \Phi(T-\sqrt{n_A}\gamma_x) + \Phi(-T-\sqrt{n_A}\gamma_x))\, d\gamma_x \\
&+ \frac{1}{n_A} \cdot \int_\infty^\infty \big((T+\gamma_{x_j}\sqrt{n_A}) \cdot \phi(T-\gamma_{x_j}\sqrt{n_A}) + (T-\gamma_{x_j}\sqrt{n_A}) \cdot \phi(-T-\gamma_{x_j}\sqrt{n_A}))\big) \\
&\times \frac{1}{\sigma_x} \cdot \phi(\gamma_x/\sigma_x) d\gamma_x \\
&= 2 \cdot \int_\infty^\infty \frac{1}{\sigma_x} \cdot \phi(\gamma_x/\sigma_x) \cdot \left(\gamma_x^2 + \frac{1}{n_A}\right) \cdot \Phi(-T+\sqrt{n_A}\gamma_x) d\gamma_x \\
&+ \frac{1}{n_A} \cdot \frac{\exp\left(-\frac{1}{2} \cdot \frac{T^2}{1+\sigma_x^2 \cdot n_A}\right) \cdot \sqrt{\frac{2}{\pi}} \cdot T \cdot (2\sigma_x^2 \cdot n_A + 1)}{(1+\sigma_x^2 \cdot n_A)^{(3/2)}} \\
&= 2 \cdot \frac{1}{n_A} \cdot \Phi\left(-\frac{T}{\sqrt{1+n_A \cdot \sigma_x^2}}\right) \\
&+ 2 \cdot \int_\infty^\infty \frac{1}{\sigma_x} \cdot \phi(\gamma_x/\sigma_x) \cdot \gamma_x^2 \cdot \Phi(-T+\sqrt{n_A}\gamma_x) d\gamma_x \\
&+ \frac{1}{n_A} \cdot \frac{\exp\left(-\frac{1}{2} \cdot \frac{T^2}{1+\sigma_x^2 \cdot n_A}\right) \cdot \sqrt{\frac{2}{\pi}} \cdot T \cdot (2\sigma_x^2 \cdot n_A + 1)}{(1+\sigma_x^2 \cdot n_A)^{(3/2)}}
\end{aligned} \tag{S48}$$

Plugging in the expression for the integral we already computed for s_j (Equation S44) we have:

$$\begin{aligned}
\frac{\sum_{j\in\mathcal{M}_1} t_j \cdot Pr(S_j)}{\pi_x \cdot M} &= 2 \cdot \frac{1}{n_A} \cdot \Phi\left(-\frac{T}{\sqrt{1+n_A\cdot\sigma_x^2}}\right) \\
&+ 2 \cdot \left(\sigma_x^2 \cdot \Phi\left(\frac{-T}{\sqrt{1+n_A\sigma_x^2}}\right) + T \cdot n_A \cdot \sigma_x^4 \cdot \frac{\exp\left(-\frac{T^2}{2(n_A\sigma_x^2+1)}\right)}{\sqrt{2\pi}(1+n_A\sigma_x^2)^{3/2}}\right) \\
&+ \frac{1}{n_A} \cdot \frac{\exp\left(-\frac{1}{2}\cdot\frac{T^2}{1+\sigma_x^2\cdot n_A}\right)\cdot\sqrt{\frac{2}{\pi}}\cdot T\cdot(2\sigma_x^2\cdot n_A+1)}{(1+\sigma_x^2\cdot n_A)^{(3/2)}} \\
&= 2\cdot\left(\sigma_x^2+\frac{1}{n_A}\right)\cdot\Phi\left(-\frac{T}{\sqrt{1+n_A\cdot\sigma_x^2}}\right) \\
&+ 2T\cdot\left(n_A\cdot\sigma_x^4+2\sigma_x^2+\frac{1}{n_A}\right)\cdot\frac{\exp\left(-\frac{T^2}{2(n_A\sigma_x^2+1)}\right)}{\sqrt{2\pi}(1+n_A\sigma_x^2)^{3/2}}
\end{aligned} \tag{S49}$$

- **The final formula**

To simplify the notations, let's define:

$$\begin{aligned}
a &= \Phi\left(-\frac{T}{\sqrt{1+n_A\cdot\sigma_x^2}}\right) \\
b &= 2T\cdot\frac{\exp\left(-\frac{T^2}{2(n_A\sigma_x^2+1)}\right)}{\sqrt{2\pi}(1+n_A\sigma_x^2)^{3/2}} \\
c &= \Phi(-T)+T\cdot\phi(T)
\end{aligned} \tag{S50}$$

This leads us to the expectation of the causal effect estimation, $E[\hat{\alpha}_{IVW}]$:

$$\begin{aligned}
&\frac{\pi_x\cdot\left(2(\alpha\sigma_x^2+\lambda')\cdot a+b\cdot\left(\alpha\cdot n_A\cdot\sigma_x^4+\alpha\sigma_x^2+\lambda'\cdot(1+\sigma_x^2\cdot n_A)\right)\right)+(1-\pi_x)\cdot 2\lambda'\cdot c}{\pi_x\cdot\left(2\cdot\left(\sigma_x^2+\frac{1}{n_A}\right)\cdot a+b\cdot\left(n_A\cdot\sigma_x^4+2\sigma_x^2+\frac{1}{n_A}\right)\right)+(1-\pi_x)\cdot\frac{2}{n_A}\cdot c} \\
= \;& \alpha\cdot\frac{\left(\pi_x\cdot\sigma_x^2\right)\cdot\left(2\cdot a+b\cdot(1+n_A\cdot\sigma_x^2)\right)}{d(n_A,T,\pi_X,\sigma_x^2)} \\
&+\lambda'\cdot\frac{\pi_x\cdot\left(2\cdot a+b\cdot(1+n_A\cdot\sigma_x^2)\right)+\left(1-\pi_x\right)\cdot 2\cdot c}{d(n_A,T,\pi_x,\sigma_x^2)}
\end{aligned} \tag{S51}$$

with :

$$\begin{aligned}
d(n_A,T,\pi_x,\sigma_x^2) &= \pi_x\cdot\left(2\cdot\left(\sigma_x^2+\frac{1}{n_A}\right)\cdot a+b\cdot\left(n_A\cdot\sigma_x^4+2\sigma_x^2+\frac{1}{n_A}\right)\right) \\
&+(1-\pi_x)\cdot\frac{2}{n_A}\cdot c
\end{aligned} \tag{S52}$$

To estimate this expectation, one needs to first obtain the per-(active)variant-heritability (σ_x^2) and polygenicity (π_x) of the exposure (See Supplementary section B). Note that the exposure heritability is simply $\pi_x\cdot M\cdot\sigma_x^2$. Then a cross-trait LDSC would inform us about the value of λ'. Finally, the threshold T is decided in the MR analysis.

c. Estimation of the corrected effect

From Equation S51 , we can derive a corrected effect ($\hat{\alpha}_c$) for the causal effect :

$$\hat{\alpha}_c = \frac{\hat{\alpha}_{IVW} \cdot d(n_A, T, \pi_X, \sigma_X^2) - \lambda' \cdot (\pi_X \cdot \big(2 \cdot a + b \cdot (1 + \sigma_X^2 \cdot n_A)\big) + \big(1 - \pi_X\big) \cdot 2 \cdot c)}{\big(\pi_X \cdot \sigma_X^2\big) \cdot \big(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_X^2)\big)} \quad \text{(S53)}$$

While n_A, n_B and T are known parameters, several other quantities in the correction needs to be estimated in the above equation. The LDSC intercept can be obtained via LD score regression, giving us $\widehat{\lambda}$. The estimation of the two remaining parameters $\widehat{\pi}_x$ and $\widehat{\sigma}_x^2$ are described below.

B. Estimation of the genetic architecture

Parameters describing the genetic architecture of the exposure, the per-(active)variant-heritability (σ_x^2) and polygenicity (π_x), are needed to estimate the corrected effect $\hat{\alpha}_c$ (Equation S53). To estimate those, we are taking advantage of the equivalence between the Equations S15 and S17. The denominators of these two formulas are equal, leading us to:

$$\begin{aligned} \sum_{j=1}^{m} (\widehat{\beta}_x^A)_j^2 &= \sum_{j=1}^{m} t_j \cdot Pr(S_j) \\ \sum_{k=1}^{m} (\widehat{\beta}_x^A)_k^2 &= M \cdot d(n_A, T, \pi_X, \sigma_X^2) \end{aligned} \tag{S54}$$

The left-hand side of Equation S54 can be estimated from the data, and n_A and T are known. By plugging in the heritability estimate from LDSC regression ($\widehat{h}_x^2$) and defining $\sigma_x^2 = \frac{h_x^2}{M \cdot \pi_x}$, the only unknown parameter is π_x. We can then use an optimisation approach to minimise the difference between $\sum_{k=1}^{m} (\widehat{\beta}_x^A)_k^2$ and $M \cdot d(n_A, T, \pi_x, \sigma_x^2)$ in order to estimate π_x and the corresponding σ_x^2. These estimates, $\widehat{\pi}_x$ and $\widehat{\sigma}_x^2$ can then be plugged into Equation S53 estimate the corrected effects.

C. Estimation of the standard error of the corrected effect estimate and the covariance between observed and corrected effects

We chose a sampling strategy to estimate the variance of the corrected causal effect, computed according to Equation S53. As a first step, we simulate each parameter included in the formula s times. n_A, n_B, T are known and can be directly used. We start by simulating s observed effects to account for the variability of $\hat{\alpha}_{IVW}$:

$$\hat{\boldsymbol{\alpha}}_{1...s} \sim \mathcal{N}\big(\hat{\alpha}_{IVW}, Var(\hat{\alpha}_{IVW})\big) \tag{S55}$$

We also simulate s cross-trait LDsc intercepts and heritability estimates:

$$\widehat{\boldsymbol{\lambda}}_{1...s} \sim \mathcal{N}\big(\widehat{\lambda}, Var(\widehat{\lambda})\big) \tag{S56}$$

$$\left(\widehat{\boldsymbol{h}}^{\boldsymbol{2}}_{\boldsymbol{x}}\right)_{1...s} \sim \mathcal{N}\big(\widehat{h}^2_x, Var(h^2_x)\big) \tag{S57}$$

To account for the variability of $\widehat{\pi}_x$ we simulate s set of IVs. Let us consider $\hat{\boldsymbol{\beta}}^{\boldsymbol{A}}_{\boldsymbol{x}}$ the observed effects of the m IVs on the exposure. The effects of the s set of IVs on $\boldsymbol{X}$ are simulated using a multivariate normal distribution:

$$\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{A}}_{\boldsymbol{x}}\right)_{1...s} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{A}}_{\boldsymbol{x}}, \frac{I}{n_A}\right) \tag{S58}$$

For each set of IVs, we use the approach described in Supplementary section B to estimate $(\widehat{\boldsymbol{\pi}}_{\boldsymbol{x}})_{1...s}$ and $\left(\widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{x}}\right)_{1...s}$. Finally, using $(\hat{\boldsymbol{\alpha}})_{1...s}$, $\left(\widehat{\boldsymbol{\lambda}}\right)_{1...s}$, $(\widehat{\boldsymbol{\pi}}_{\boldsymbol{x}})_{1...s}$ and $\left(\widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{x}}\right)_{1...s}$ and Equation S53 we estimate s corrected effects $((\widehat{\boldsymbol{\alpha}_{\boldsymbol{c}}})_{1...s})$ to approximate the distribution of the corrected effect. The standard error of the corrected effect is estimated by taking the standard deviation of the s corrected effects estimates, i.e. $\widehat{Var(\widehat{\alpha}_c)} = \sum_{j=1}^{s}((\widehat{\boldsymbol{\alpha}_{\boldsymbol{c}}})_j - \widehat{\alpha}_c)^2/(s-1)$.

We validated this approach by comparing the standard error obtained as described above using $s = 10,000$ to the one that can be derived from the 100 simulated datasets (standard settings). For each threshold and each overlap, we calculated the standard error for the corrected effect of each dataset using our approach, and compared the mean of these standard errors to the standard deviation of the corrected effects estimates across the 100 datasets (Figure S10) and observed a very good agreement.

In addition, we can use the s simulated observed and corrected effects ($(\hat{\boldsymbol{\alpha}})_{1...s}$ and $(\widehat{\boldsymbol{\alpha}_{\boldsymbol{c}}})_{1...s}$) to estimate the covariance between them. This covariance estimate can be used to test if there is a difference between observed and corrected effects. Namely, the test statistic we use is

$$\frac{\widehat{\alpha} - \widehat{\alpha}_c}{\sqrt{Var(\widehat{\alpha}) + Var(\widehat{\alpha}_c) - 2 \cdot cov(\widehat{\alpha}, \widehat{\alpha}_c)}} \sim \mathcal{N}(0,1)$$

D. Simplification under the null

In Burgess *et al.* [3], a formula for the expected bias under the null is proposed. In order to compare our results with the ones from this paper, we had to make a few additional simplifications. In their simulation design, all genetic variants were considered to be causal ($\pi_x = 1$) and were used (no threshold used to select IVs).

When setting $\alpha = 0$ and $\pi_x = 1$ in Equation S51 we obtain:

$$\begin{aligned} E[\hat{\alpha}_{IVW}] &= \lambda' \cdot \frac{\left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_X^2)\right)}{d(n_A, T, \pi_x, \sigma_x^2)} \\ &= \lambda' \cdot \frac{\left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\right)}{2 \cdot a \cdot (\sigma_x^2 + \frac{1}{n_A}) + b \cdot (n_A \cdot \sigma_x^2 + 2 \cdot \sigma_x^2 + \frac{1}{nA})} \\ &= \lambda' \cdot \frac{\left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\right)}{d(n_A, T, \pi_x, \sigma_x^2)} \\ &= \lambda' \cdot \frac{\left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\right)}{(\sigma_x^2 + \frac{1}{n_A}) \cdot (2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)} \\ &= \frac{\lambda'}{\sigma_x^2 + \frac{1}{n_A}} \end{aligned} \tag{S59}$$

We can then use Equation S33 to replace λ':

$$E[\hat{\alpha}_{IVW}] = \rho \cdot \frac{n_{A \cap B}}{n_A \cdot n_B} \cdot \frac{1}{\sigma_x^2 + \frac{1}{n_A}} \tag{S60}$$

The simulation design proposed in their paper (Burgess *et al.* [3], Section 3 - "SIMULATION STUDY—CONTINUOUS OUTCOME") relies on non-standardised genotypic and phenotypic effects. In order to be compared, the expectation of the causal effect estimate we derived needs to be on the same scale and it can be done using the phenotypic variances of X and Y (respectively $Var(x)$ and $Var(y)$):

$$E[\hat{\alpha}_{IVW}] = \frac{\sqrt{Var(x)}}{\sqrt{Var(y)}} \cdot \rho \cdot \frac{n_{A \cap B}}{n_A \cdot n_B} \cdot \frac{1}{\sigma_x^2 + \frac{1}{n_A}} \tag{S61}$$

The expectation of the causal effect estimate proposed by Burgess *et al.* [3] is as follows:

$$E[\hat{\alpha}_{IVW}]_B = \text{OLS estimate} \cdot \frac{\%\text{ of overlap}}{100} \cdot \frac{1}{F} \tag{S62}$$

with:

$$\begin{aligned} \text{OLS estimate} &= \frac{cov(x, y)}{Var(x)} \\ &= \frac{\sqrt{Var(x)}}{\sqrt{Var(y)}} \cdot \rho \text{ since } \rho \text{ is the correlation between } X \text{ and } Y \end{aligned} \tag{S63}$$

$$\frac{\%\text{ of overlap}}{100} = \frac{n_{A \cap B}}{n_B} \tag{S64}$$

$$F = \frac{n_A - m - 1}{m} \cdot \frac{R^2}{1 - R^2} \quad \text{(S65)}$$

(R^2 being the coefficient of determination)

leading to:

$$E[\hat{\alpha}_{IVW}]_B = \frac{\sqrt{Var(x)}}{\sqrt{Var(y)}} \cdot \rho \cdot \frac{n_{A \cap B}}{n_B} \cdot \frac{1}{\frac{n_A - m - 1}{m} \cdot \frac{R^2}{1 - R^2}} \quad \text{(S66)}$$

By defining $\sigma_x^2 = \frac{R^2}{m}$, it can be shown that when $n_A >> m$ and R^2 is small:

$$\frac{1}{n_A} \cdot \frac{1}{\sigma_x^2 + \frac{1}{n_A}} \approx \frac{1}{\frac{n_A - m - 1}{m} \cdot \frac{R^2}{1 - R^2}} \quad \text{(S67)}$$

In this case, the expectation of the causal effect estimate we derived (Equation S61) is equivalent to the one proposed by Burgess *et al.* [3] (Equation S66).

We compared causal effect estimate expectations for all the scenarios without a causal effect they presented (Burgess *et al.* [3], Table 3). To do so, we calculated the theoretical values for $Var(x)$, $Var(y)$, OLS estimate and R^2 based on the parameters reported in the table and plugged them into Equations S61 and S66. The causal effect estimate expectations can be compared to the observed effects they reported. The expectation of the causal effect estimate we derived give results that are very close to the reported effect for all scenarios while the one proposed by Burgess *et al.* [3] suffer from a slight overestimation (Supplementary Table S16). It can be due to the fact that we used the theoretical values of R^2 which are lower than the observed ones because of overfitting.

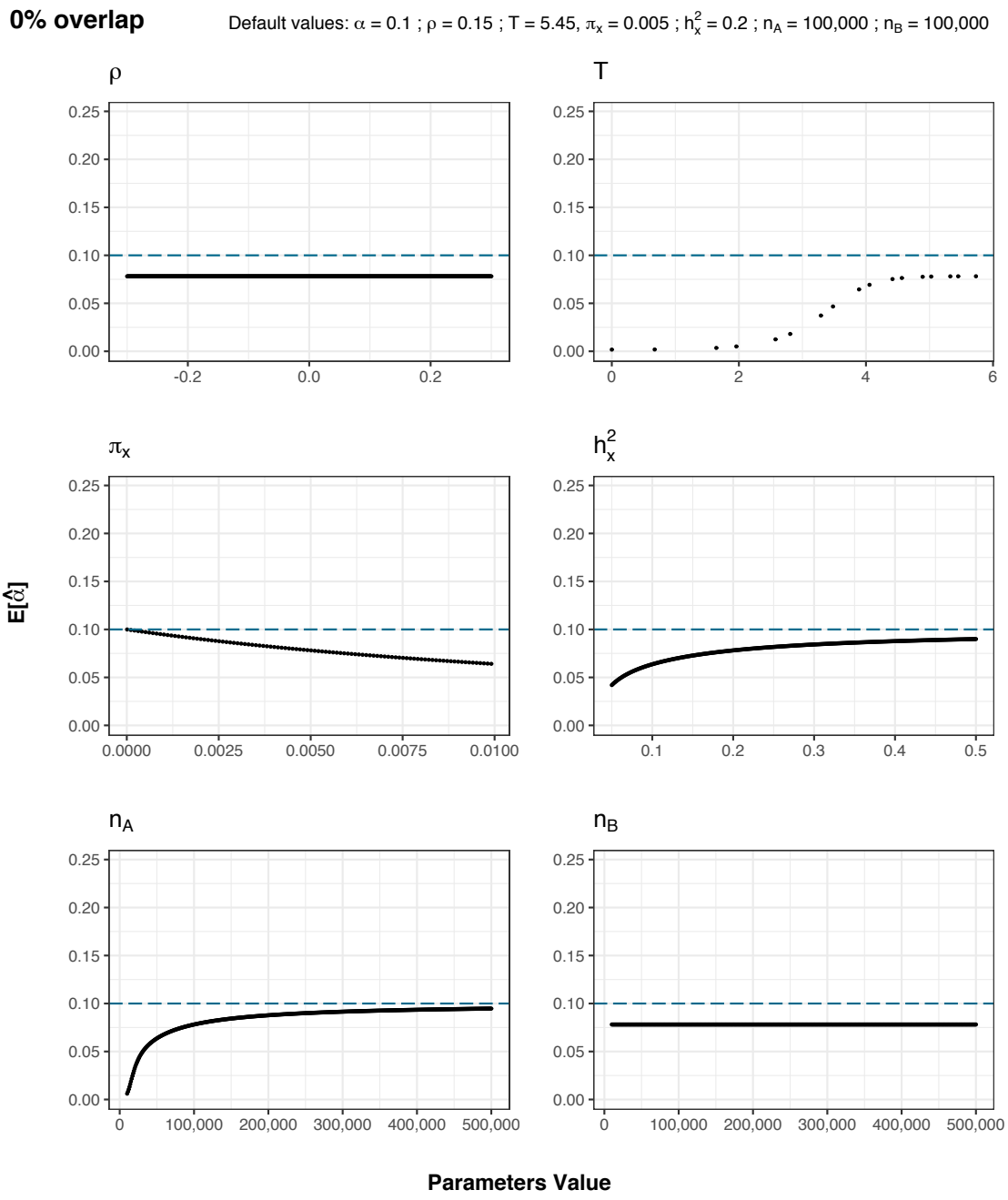


Figure S1: Effect of the different parameters when overlap = 0%
For each parameter affecting $E[\hat{\alpha}]$, we tested a wide range of values, while keeping the other parameters constant. The true causal effect value is indicated by the blue dotted line.
Defaults parameters are $n_A = 100,000, n_B = 100,000, \pi_x = 0.005, h_x^2 = 0.2, \rho = 0.15, \alpha = 0.1$

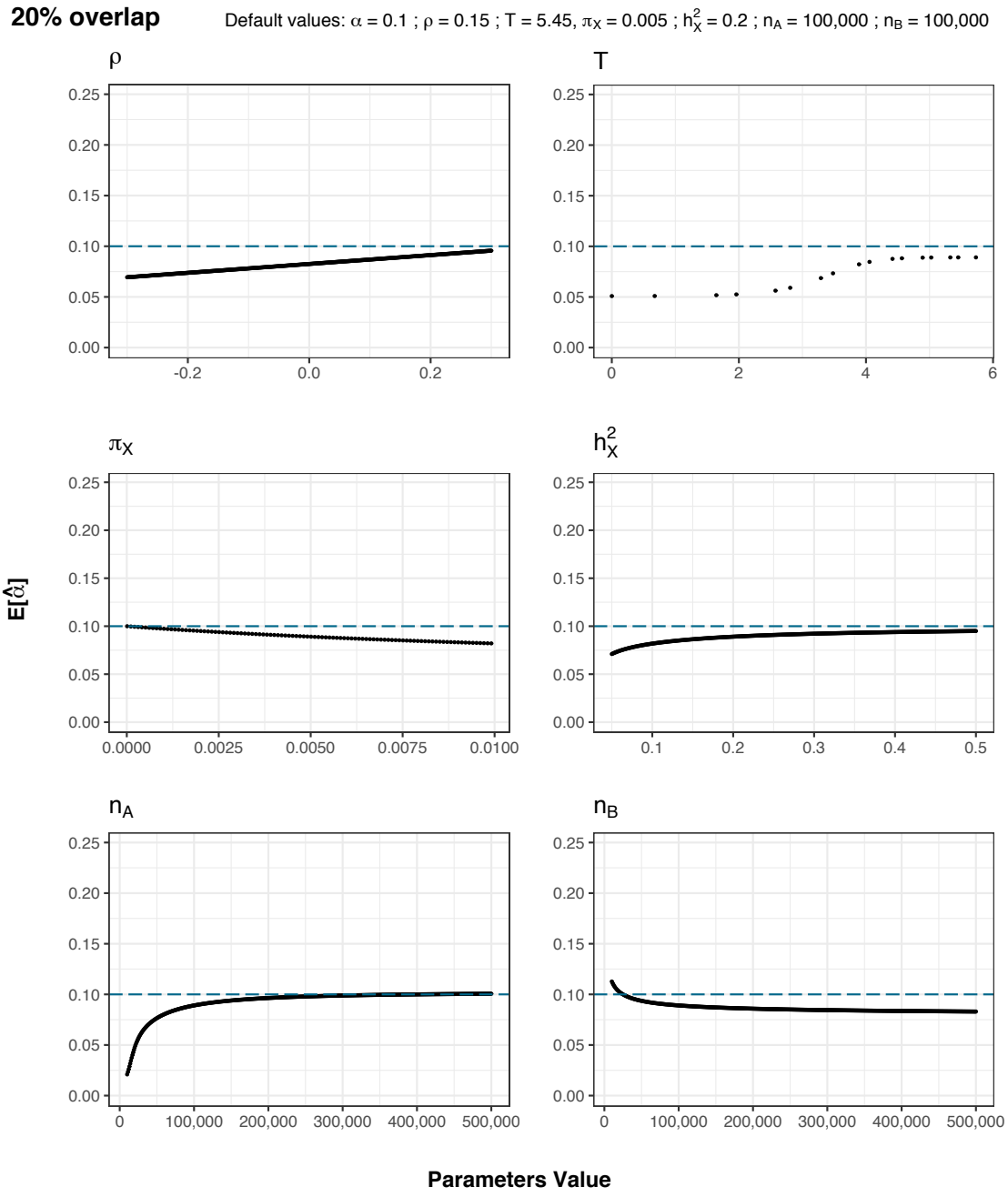


Figure S2: Effect of the different parameters when overlap = 20%
For each parameter affecting $E[\hat{\alpha}]$, we tested a wide range of values, while keeping the other parameters constant. The true causal effect value is indicated by the blue dotted line.
Defaults parameters are $n_A = 100,000, n_B = 100,000, \pi_x = 0.005, h_x^2 = 0.2, \rho = 0.15, \alpha = 0.1$

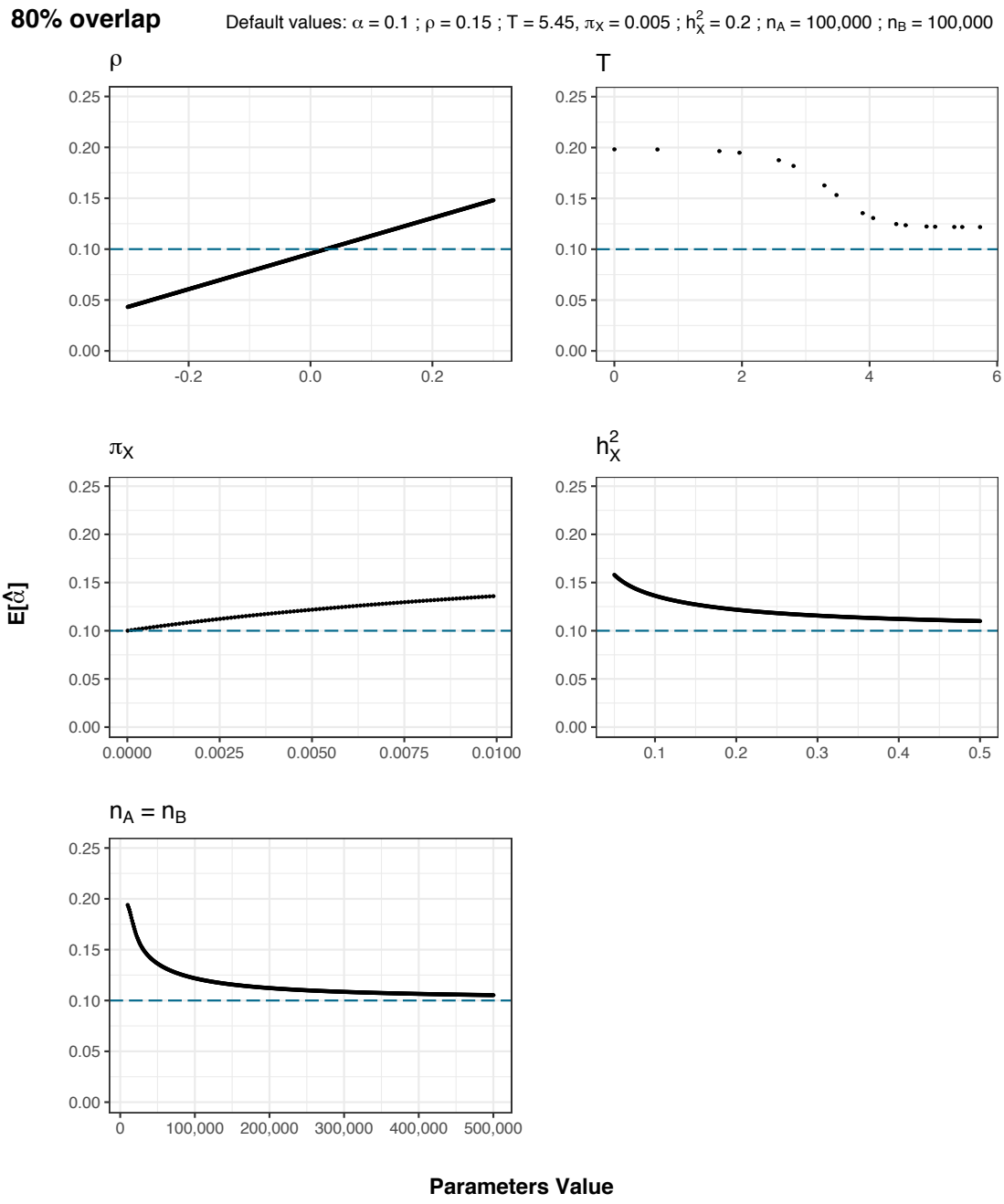


Figure S3: Effect of the different parameters when overlap = 80%
For each parameter affecting $E[\hat{\alpha}]$, we tested a wide range of values, while keeping the other parameters constant. The true causal effect value is indicated by the blue dotted line.
Defaults parameters are $n_A = 100,000, n_B = 100,000, \pi_x = 0.005, h_x^2 = 0.2, \rho = 0.15, \alpha = 0.1$
Note that for simplicity, n_A and n_B are assumed to be equal, to ensure that an overlap of 80% can be observed.

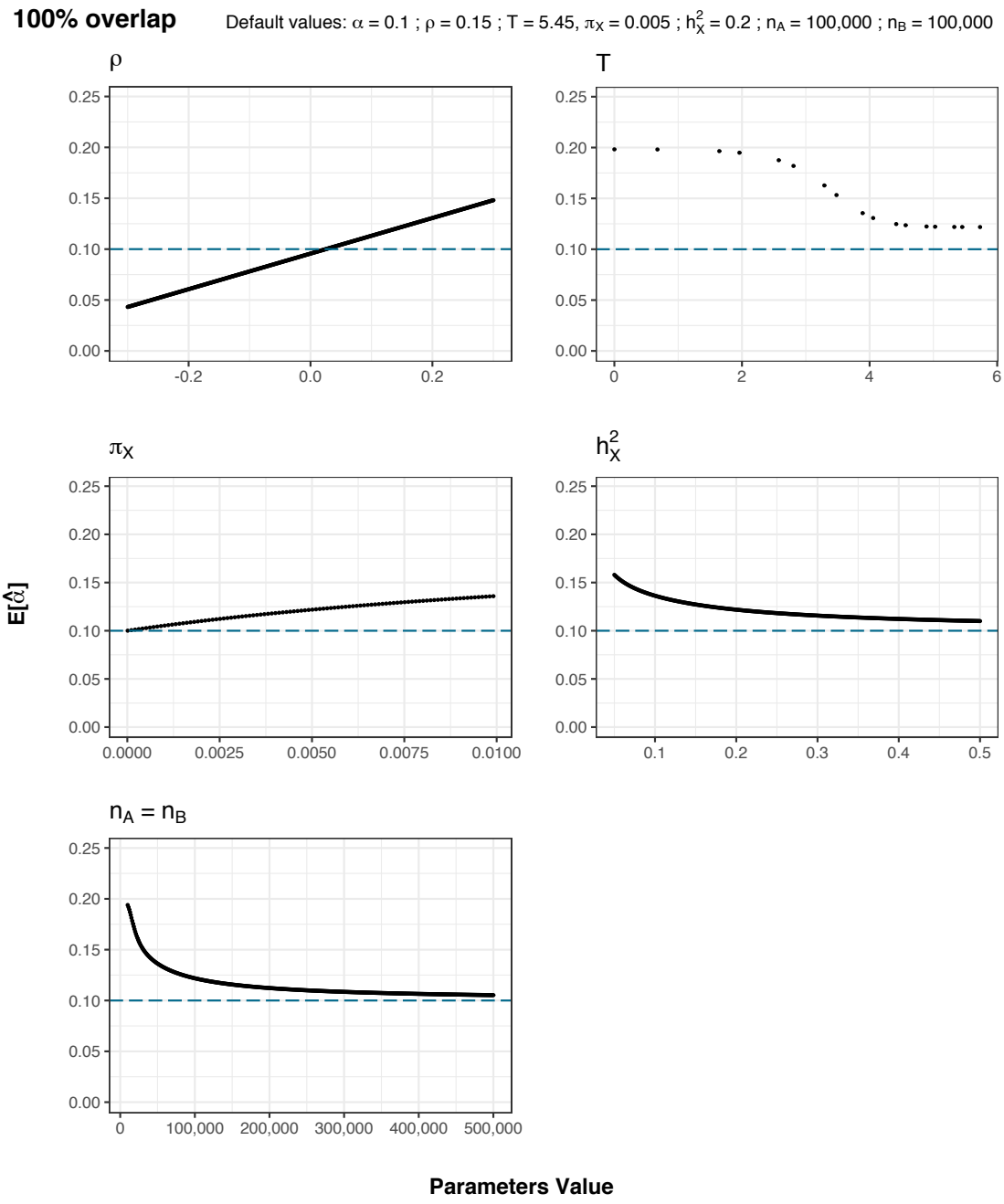


Figure S4: Effect of the different parameters when overlap = 100%
For each parameter affecting $E[\hat{\alpha}]$, we tested a wide range of values, while keeping the other parameters constant. The true causal effect value is indicated by the blue dotted line.
Defaults parameters are $n_A = 100,000, n_B = 100,000, \pi_x = 0.005, h_x^2 = 0.2, \rho = 0.15, \alpha = 0.1$
Note that for simplicity, n_A and n_B are assumed to be equal, to ensure that an overlap of 100% can be observed.

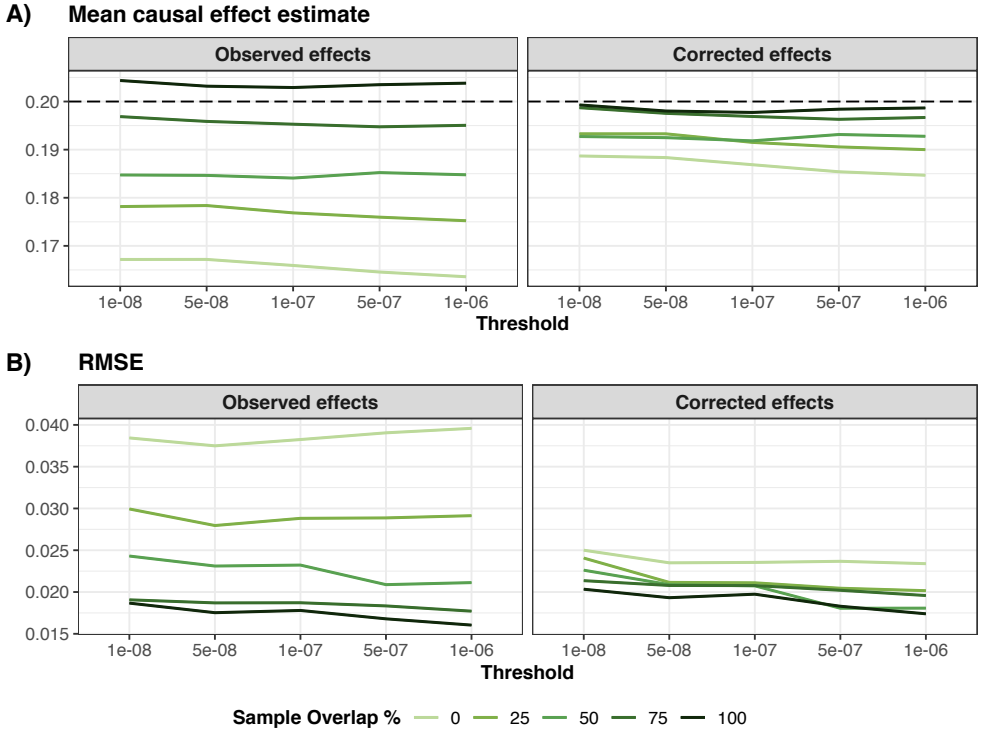


Figure S5: Simulation results for a scenario with a weaker environmental confounder $n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.15, \kappa_y = 0.3, \alpha = 0.2$
Panel A) shows the mean observed and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represent the true causal effect). Panel B) shows the mean RMSE obtained for observed and corrected effect for each overlap and threshold.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	50.5	0.000376	0.0218	58.0	0.000476	0.00198	4.16
5e-08	64.3	0.000317	0.0202	63.5	0.000403	0.00159	3.94
1e-07	71.2	0.000307	0.0214	69.9	0.000389	0.00197	5.06
5e-07	90.8	0.000267	0.0234	87.7	0.000340	0.00259	7.61
1e-06	101.0	0.000247	0.0252	102.1	0.000316	0.00308	9.72

Table S3: Analysis of variance for a scenario with a weaker environmental confounder $n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.15, \kappa_y = 0.3, \alpha = 0.2$
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.

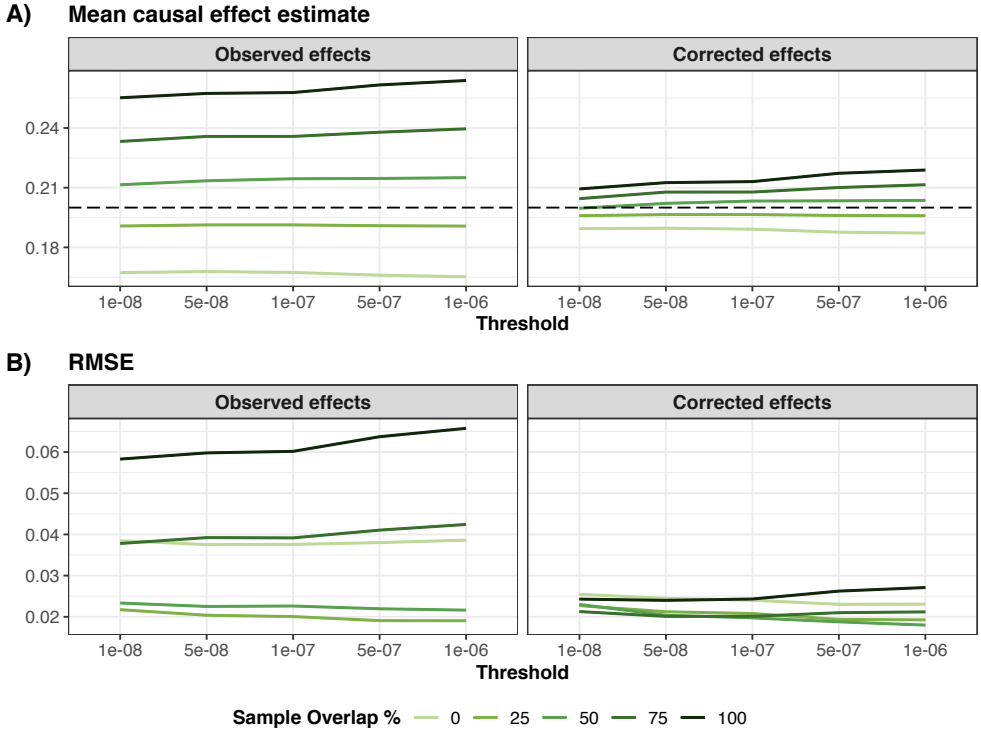


Figure S6: Simulation results for a scenario with a stronger environmental confounder $n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.5, \kappa_y = 0.8, \alpha = 0.2$
Panel A) shows the mean observed and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represent the true causal effect). Panel B) shows the mean RMSE obtained for observed and corrected effect for each overlap and threshold.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	50.2	0.000381	0.119	313	0.000505	0.00592	11.7
5e-08	64.0	0.000320	0.125	391	0.000425	0.00818	19.2
1e-07	70.7	0.000305	0.127	417	0.000409	0.00883	21.6
5e-07	90.5	0.000272	0.142	521	0.000365	0.01343	36.8
1e-06	100.8	0.000258	0.151	587	0.000349	0.01557	44.6

Table S5: Analysis of variance for a scenario with a stronger environmental confounder $n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.5, \kappa_y = 0.8, \alpha = 0.2$
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.

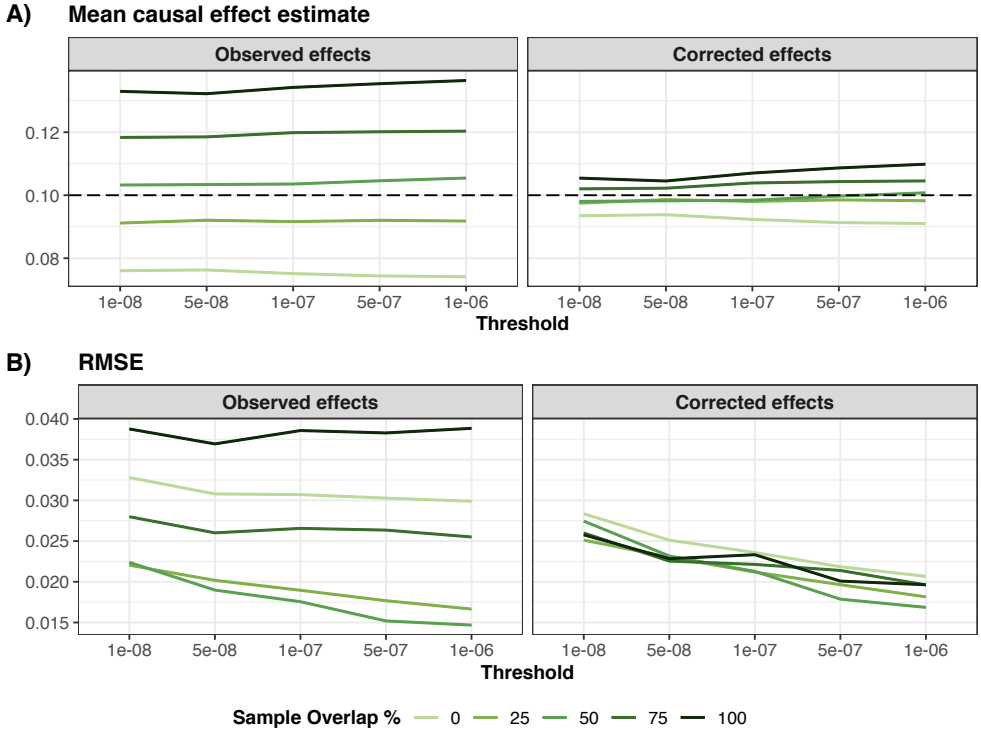


Figure S7: Simulation results for realistic settings
$n_A = n_B = 100,000, \pi_x = 0.005, h_x^2 = 0.2, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0.1$
Panel A) shows the mean observed and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represent the true causal effect). Panel B) shows the mean RMSE obtained for observed and corrected effect for each overlap and threshold.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	54.4	0.000457	0.0498	109	0.000695	0.00209	3.00
5e-08	76.5	0.000351	0.0479	137	0.000537	0.00168	3.14
1e-07	89.3	0.000310	0.0537	173	0.000477	0.00324	6.80
5e-07	129.5	0.000246	0.0565	229	0.000378	0.00425	11.24
1e-06	152.7	0.000210	0.0587	280	0.000325	0.00495	15.24

Table S9: Analysis of variance for realistic settings
$n_A = n_B = 100,000, \pi_x = 0.005, h_x^2 = 0.2, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0.1$
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	27.3	0.001218	0.0104	8.55	0.00215	0.011138	5.192
5e-08	37.9	0.001072	0.0162	15.13	0.00195	0.007121	3.647
1e-07	44.3	0.001023	0.0193	18.90	0.00189	0.005613	2.967
5e-07	65.3	0.000719	0.0277	38.52	0.00140	0.003135	2.247
1e-06	79.0	0.000661	0.0368	55.68	0.00131	0.000763	0.583

Table S11: Analysis of variance for the effect of BMI on SBP
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.
Since only results for a threshold of 5e-8 are discussed in the paper, results for all other thresholds have been greyed out.

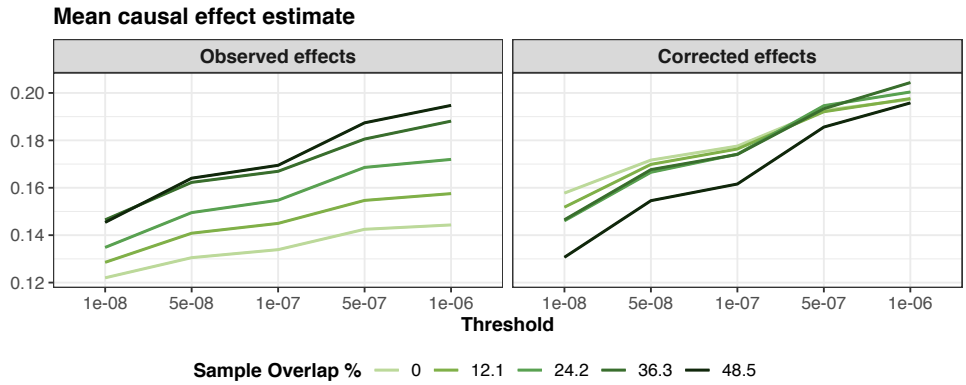


Figure S8: Effect of BMI on smoking
This figure shows the mean observed and corrected effect for each overlap and threshold obtained from 100 different sampled datasets.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	27.7	0.00225	0.0112	4.99	0.00392	0.01014	2.583
5e-08	38.2	0.00192	0.0203	10.55	0.00350	0.00453	1.295
1e-07	44.3	0.00178	0.0223	12.53	0.00331	0.00412	1.245
5e-07	66.0	0.00151	0.0339	22.45	0.00290	0.00124	0.426
1e-06	79.2	0.00134	0.0437	32.60	0.00264	0.00115	0.436

Table S13: Analysis of variance for effect of BMI on smoking
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.
Since only results for a threshold of 5e-8 are discussed in the paper, results for all other thresholds have been greyed out.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	12.9	0.00463	0.0255	5.51	0.00887	0.001424	0.1605
5e-08	17.4	0.00384	0.0360	9.39	0.00765	0.000199	0.0261
1e-07	20.5	0.00354	0.0408	11.53	0.00720	0.000422	0.0587
5e-07	30.8	0.00248	0.0570	23.03	0.00529	0.001673	0.3165
1e-06	37.5	0.00199	0.0680	34.19	0.00437	0.003815	0.8724

Table S15: Analysis of variance for the effect of SBP on BMI
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.
Since only results for a threshold of 5e-8 are discussed in the paper, results for all other thresholds have been greyed out.

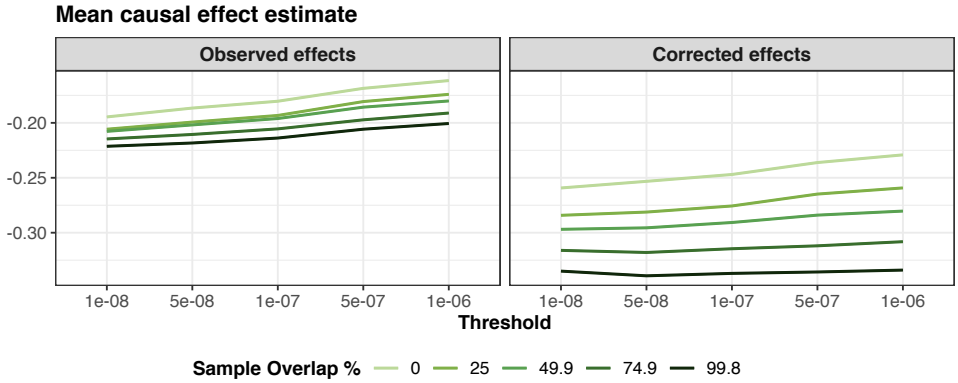


Figure S9: Effect of BMI on alcohol
This figure shows the mean observed and corrected effect for each overlap and threshold obtained from 100 different sampled datasets.

		Observed effects			Corrected effects		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	28.4	0.000803	0.0101	12.5	0.00146	0.0846	58.0
5e-08	38.7	0.000747	0.0144	19.3	0.00141	0.1098	77.7
1e-07	44.9	0.000725	0.0161	22.2	0.00140	0.1207	86.4
5e-07	66.3	0.000591	0.0209	35.4	0.00122	0.1521	125.0
1e-06	79.2	0.000482	0.0228	47.4	0.00101	0.1677	166.6

Table S17: Analysis of variance for the effect of BMI on alcohol
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.
Since only results for a threshold of 5e-8 are discussed in the paper, results for all other thresholds have been greyed out.

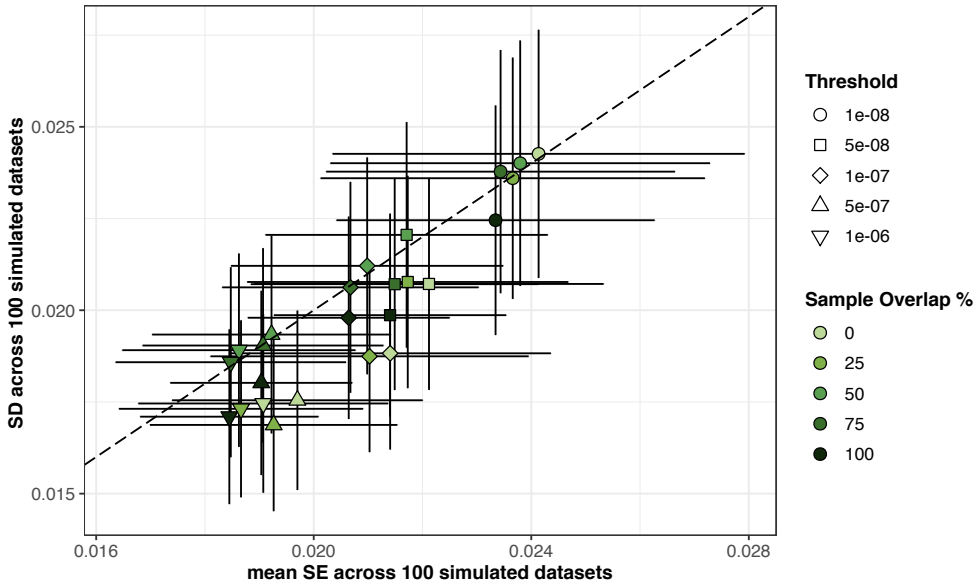


Figure S10: Standard error estimation for the corrected effects
We used $s = 100$ (simulations number used to approximate the corrected effect distribution) and estimated the corrected effect SE for 100 datasets simulated using standard settings. $n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0.2$ We reported the mean standard error (SE) estimated across the 100 simulated datasets and the observed standard deviation (SD) across the 100 datasets, as well as their 95% confidence intervals, for different overlaps and thresholds. The dashed line corresponds to the identity line.

References

[1] Bulik-Sullivan, B., , Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nature Genetics *47*, 1236–1241.

[2] Zeng, J., de Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nature Genetics *50*, 746–753.

[3] Burgess, S., Davies, N. M., and Thompson, S. G. (2016). Bias due to participant overlap in two-sample mendelian randomization. Genetic Epidemiology *40*, 597–608.

Appendix D

Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics

This article (Darrous et al., 2020) is presented in Chapter 2.

Note that this is a preprint, this paper is currently under review in *Nature Communications*.

Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics

Liza Darrous[1,2,*], **Ninon Mounier**[1,2,*], **Zoltán Kutalik**[1,2,3,†]

Abstract

Mendelian Randomisation (MR), an increasingly popular method that estimates the causal effects of risk factors on complex human traits, has seen several extensions that relax its basic assumptions. However, most of these extensions suffer from two major limitations; their under-exploitation of genome-wide markers, and sensitivity to the presence of a heritable confounder of the exposure-outcome relationship. To overcome these limitations, we propose a Latent Heritable Confounder MR (LHC-MR) method applicable to association summary statistics, which estimates bi-directional causal effects, direct heritabilities, and confounder effects while accounting for sample overlap. We demonstrate that LHC-MR outperforms several existing MR methods in a wide range of simulation settings and apply it to summary statistics of 13 complex traits. Besides several concordant results, LHC-MR unravelled new mechanisms (how being diagnosed for certain diseases might lead to improved lifestyle) and revealed new causal effects (e.g. HDL cholesterol being protective against high systolic blood pressure), hidden from standard MR methods due to a heritable confounder of opposite direction. Phenome-wide MR search suggested that the confounders indicated by LHC-MR for the birth weight-diabetes pair are likely to be obesity traits. Finally, LHC-MR results indicated that genetic correlations are predominantly driven by bi-directional causal effects and much less so by heritable confounders.

[1]University Center for Primary Care and Public Health, University of Lausanne, 1010, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland

[3]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, UK

[*]These authors contributed equally to this work.

[†]Correspondence should be addressed to zoltan.kutalik@unil.ch

1 Introduction

The identification of frequent risk factors and the quantification of their impact on common diseases is a central quest for public health policy makers. Epidemiological studies aim to address this issue, but they are most often based on observational data due to its abundance over the years. Despite major methodological advances, a large majority of such studies have inherent limitations and suffer from confounding and reverse causation[1, 2]. For these reasons, many of the reported associations found in classical epidemiological studies are mere correlates of disease risk, rather than causal factors directly involved in disease progression. Due to this limitation, additional evidence is required before developing public health interventions in a bid to reduce the future burden of diseases. While well-designed and carefully conducted randomised control trials (RCTs) remain the gold standard for causal inference, they are exceedingly expensive, time-consuming, may not be feasible for ethical reasons, and have high failure rates[3, 4].

Mendelian randomisation (MR), a natural genetic counterpart to RCTs, is an instrumental variable (IV) technique used to infer the strength of a causal relationship between a risk factor (X) and an outcome (Y)[5]. To do so, it uses genetic variants (G) as instruments and relies on three major assumptions (see Figure S1): (1) Relevance – G is robustly associated with the exposure. (2) Exchangeability – G is not associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – G is independent of the outcome conditional on the exposure and all confounders of the exposure-outcome relationship (i.e. the only path between the instrument and the outcome is via the exposure).

The advantage of the MR approach is that for most heritable exposures, dozens (if not hundreds) of genetic instruments are known to date thanks to well-powered genome-wide association studies (GWASs). Each instrument can provide a causal effect estimate, which can be combined with others, by using an inverse variance-weighting (IVW) scheme (e.g. Burgess *et al.*[6]). However, the last assumption is particularly problematic, because genetic variants tend to be pleiotropic, i.e. exert effect on multiple traits independently. Still, it can be shown that if the instrument strength is independent of the direct effect on the outcome (InSIDE assumption) and the direct effects are on average zero, IVW-based methods will still yield consistent estimates. Methods, such as MR-Egger[7], produce consistent estimates even if direct effects are allowed to have a non-zero offset. The third assumption can be further reduced to assuming that $> 50\%$ of the instruments (or in terms of their weight) are valid (median-based estimators[8]) or that zero-pleiotropy instruments are the most frequent (mode-based estimators[9]).

The InSIDE assumption (i.e. horizontal pleiotropic effects ($G \to Y$) are independent of the direct effect ($G \to X$)) is reasonable if the pleiotropic path $G \to Y$ does not branch off to X. However, if there is such a branching off, the variable representing the split is a confounder of the $X - Y$ relationship and we fall back on the violation of the second assumption (exchangeability), making it the most problematic. Therefore, in this paper, we extend the standard MR model to incorporate the presence of a latent (i.e. unmeasured) heritable confounder (U) and estimate its contribution to traits X and Y, while simultaneously estimating the bi-directional causal effect between the two traits. Standard MR methods are vulnerable to such heritable confounders, since any genetic marker directly associated with the confounder may be selected as an instrument for the exposure. However, such instruments will have a direct effect on the outcome that is correlated to their instrument strength, violating the InSIDE assumption and biasing the causal effect estimate.

The outline of the paper is as follows: first, the extended MR model is introduced and the likelihood function for the observed genome-wide summary statistics (for X and Y) is derived. We then test and compare the method against conventional and more advanced (such as CAUSE

[10] and MR RAPS [11]) MR approaches through extensive simulation settings, including several violations of the model assumptions. Finally, the approach is applied to association summary statistics (based on the UK Biobank and meta-analysis studies) of 13 complex traits to re-assess all pairwise bi-directional causal relationships between them.

2 Methods

2.1 The underlying structural equation model

Let X and Y denote continuous random variables representing two complex traits. Let us assume (for simplicity) that there is one heritable confounder U of these traits. To simplify notation we assume that $E(X) = E(Y) = E(U) = 0$ and $Var(X) = Var(Y) = Var(U) = 1$. The genome-wide sequence data for M sequence variants is denoted by $G = (G_1, G_2, \ldots, G_M)$. The aim of our work is to dissect the effects of the heritable confounding factor U from the bi-directional causal effects of these two traits (X and Y). For this we consider a model (see Figure 1) defined by the following equations:

$$\begin{aligned} X &= q_x \cdot U + \alpha_{y\to x} Y + G \cdot \boldsymbol{\gamma_x} + \boldsymbol{e_x} \quad \text{with} \quad \boldsymbol{e_x} \sim \mathcal{N}(0, \nu_x^2) \\ Y &= q_y \cdot U + \alpha_{x\to y} X + G \cdot \boldsymbol{\gamma_y} + \boldsymbol{e_y} \quad \text{with} \quad \boldsymbol{e_y} \sim \mathcal{N}(0, \nu_y^2) \\ U &= G \cdot \boldsymbol{\gamma_u} + \boldsymbol{e_u} \quad \text{with} \quad \boldsymbol{e_u} \sim \mathcal{N}(0, \nu_u^2) \end{aligned}$$

where $\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}, \boldsymbol{\gamma_u} \in \mathcal{R}^M$ denote the (true multivariable) direct effect of all M genetic variants on X, Y and U, respectively. All error terms ($\boldsymbol{e_x}, \boldsymbol{e_y}$ and $\boldsymbol{e_u}$) are assumed to be independent of each other and normally distributed with variances ν_x^2, ν_y^2 and ν_u^2, respectively.

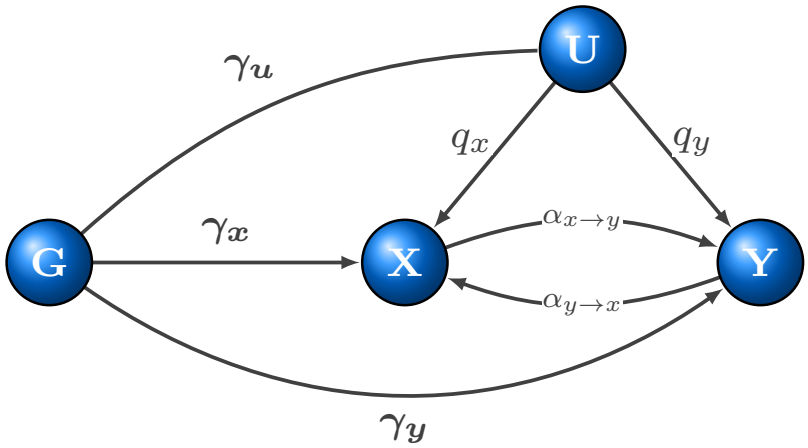


Figure 1: Schematic representation of the extended structural equation model (SEM). X and Y are two complex traits under scrutiny with a latent (heritable) confounder U with causal effects q_x and q_y on them. G represents a genetic instrument, with effects $\boldsymbol{\gamma_x}$, $\boldsymbol{\gamma_y}$ and $\boldsymbol{\gamma_u}$, respectively. Traits X and Y have causal effects on each other, which are denoted by $\alpha_{x\to y}$ and $\alpha_{y\to x}$.

Note that we do not include in the model reverse causal effects on the confounder ($X \to U$ and $Y \to U$). The reason for this is the following: Let s_x and s_y denote those causal effect of X and Y on U. We can see that by reparameterising the original model to $\alpha'_{x\to y} := \alpha_{x\to y} + s_x \cdot q_y$, $\alpha'_{y\to x} := \alpha_{y\to x} + s_y \cdot q_x$ and $q'_x := q_x/(1 - q_x \cdot s_x)$, $q'_y := q_y/(1 - q_y \cdot s_y)$, the genetic effects produced by the extended model with reverse causal effects on U and the simpler model (Figure 1) with the updated parameters are indistinguishable. Thus these extra parameters are not identifiable

and the reparameterisation means that $\alpha_{x\to y}$ and $\alpha_{y\to x}$ in our model represent the total causal effects, some of which may be mediated by U.

Note that the model cannot be represented by classical directed acyclic graphs, as the bi-directional causal effects could form a cycle. However, the equations can be reorganised to avoid recursive formulation as follows:

$$\begin{aligned}
X &= q_x \cdot U + \alpha_{y\to x} \cdot (q_y \cdot U + \alpha_{x\to y} X + G \cdot \boldsymbol{\gamma_y} + \boldsymbol{e_y}) + G \cdot \boldsymbol{\gamma_x} + \boldsymbol{e_x} \\
Y &= q_y \cdot U + \alpha_{x\to y} \cdot (q_x \cdot U + \alpha_{y\to x} Y + G \cdot \boldsymbol{\gamma_x} + \boldsymbol{e_x}) + G \cdot \boldsymbol{\gamma_y} + \boldsymbol{e_y} \\
U &= G \cdot \boldsymbol{\gamma_u} + \boldsymbol{e_u}
\end{aligned}$$

Regrouping the terms gives

$$\begin{aligned}
(1 - \alpha_{y\to x}\alpha_{y\to x}) \cdot X &= (q_x + \alpha_{y\to x} \cdot q_y) \cdot U + \alpha_{y\to x}(G \cdot \boldsymbol{\gamma_y}) + G \cdot \boldsymbol{\gamma_x} + (\boldsymbol{e_x} + \alpha_{y\to x} \cdot \boldsymbol{e_y}) \\
(1 - \alpha_{y\to x}\alpha_{y\to x}) \cdot Y &= (q_y + \alpha_{x\to y} \cdot q_x) \cdot U + \alpha_{x\to y}(G \cdot \boldsymbol{\gamma_x}) + G \cdot \boldsymbol{\gamma_y} + (\boldsymbol{e_y} + \alpha_{x\to y} \cdot \boldsymbol{e_x}) \\
U &= G \cdot \boldsymbol{\gamma_u} + \boldsymbol{e_u}
\end{aligned}$$

Substituting U into the first two equations yields

$$\begin{aligned}
X &= \frac{q_x + \alpha_{y\to x} \cdot q_y}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot (G \cdot \boldsymbol{\gamma_u}) + \frac{\alpha_{y\to x}}{1 - \alpha_{y\to x}\alpha_{y\to x}}(G \cdot \boldsymbol{\gamma_y}) + \frac{1}{1 - \alpha_{y\to x}\alpha_{y\to x}}(G \cdot \boldsymbol{\gamma_x}) + \boldsymbol{\epsilon_x} \\
Y &= \frac{q_y + \alpha_{x\to y} \cdot q_x}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot (G \cdot \boldsymbol{\gamma_u}) + \frac{\alpha_{x\to y}}{1 - \alpha_{y\to x}\alpha_{y\to x}}(G \cdot \boldsymbol{\gamma_x}) + \frac{1}{1 - \alpha_{y\to x}\alpha_{y\to x}}(G \cdot \boldsymbol{\gamma_y}) + \boldsymbol{\epsilon_y}
\end{aligned}$$

with

$$\begin{aligned}
\boldsymbol{\epsilon_x} &:= \frac{\boldsymbol{e_x} + \alpha_{y\to x} \cdot \boldsymbol{e_y} + (q_x + \alpha_{y\to x} \cdot q_y) \cdot \boldsymbol{e_u}}{1 - \alpha_{y\to x}\alpha_{y\to x}} \sim \mathcal{N}(0, i_x) \\
\boldsymbol{\epsilon_y} &:= \frac{\boldsymbol{e_y} + \alpha_{x\to y} \cdot \boldsymbol{e_x} + (q_y + \alpha_{x\to y} \cdot q_x) \cdot \boldsymbol{e_u}}{1 - \alpha_{y\to x}\alpha_{y\to x}} \sim \mathcal{N}(0, i_y)
\end{aligned}$$

where $i_x := (\nu_x^2 + \alpha_{y\to x}^2\nu_y^2 + (q_x + \alpha_{y\to x}q_y)\nu_u^2)/(1 - \alpha_{x\to y}\alpha_{y\to x})^2$ and $i_y := (\nu_y^2 + \alpha_{x\to y}^2\nu_x^2 + (q_y + \alpha_{x\to y}q_x)\nu_u^2)/(1 - \alpha_{x\to y}\alpha_{y\to x})^2$. Note that i_x is equivalent to the LD score regression intercept[12].

We model the genetic architecture of these direct effects with a spike-and-slab distribution, assuming that only $0 \le \pi_x, \pi_y, \pi_u \le 1$ proportion of the genome have a direct effect on X, Y, U, respectively and these direct effects come from a Gaussian distribution. Namely,

$$\begin{aligned}
\boldsymbol{\gamma_x} &= \boldsymbol{\zeta_x} \odot \boldsymbol{\kappa_x} \quad \text{with} \quad \boldsymbol{\kappa_x} \sim \mathcal{N}(0, \sigma_x^2 \cdot I) \quad \text{and} \quad \boldsymbol{\zeta_x} \sim \mathcal{B}_m(1, \pi_x) \\
\boldsymbol{\gamma_y} &= \boldsymbol{\zeta_y} \odot \boldsymbol{\kappa_y} \quad \text{with} \quad \boldsymbol{\kappa_y} \sim \mathcal{N}(0, \sigma_y^2 \cdot I) \quad \text{and} \quad \boldsymbol{\zeta_y} \sim \mathcal{B}_m(1, \pi_y) \\
\boldsymbol{\gamma_u} &= \boldsymbol{\zeta_u} \odot \boldsymbol{\kappa_u} \quad \text{with} \quad \boldsymbol{\kappa_u} \sim \mathcal{N}(0, \sigma_u^2 \cdot I) \quad \text{and} \quad \boldsymbol{\zeta_y} \sim \mathcal{B}_m(1, \pi_u)
\end{aligned}$$

Here, $\odot$ denotes element-wise multiplication and $\mathcal{B}_m(1, q)$ the m dimensional independent Bernoulli distribution. Further, we assume that all $\boldsymbol{\kappa_x}, \boldsymbol{\kappa_y}, \boldsymbol{\kappa_u}$s are independent of each other and so are all $\boldsymbol{\zeta_x}, \boldsymbol{\zeta_y}, \boldsymbol{\zeta_u}$s. We can refer to $h_x^2 := M \cdot \pi_x \cdot \sigma_x^2$ as the *direct heritability* of X, i.e. independent of the genetic basis of U and Y. Similar notation is adapted for U ($h_u^2 := M \cdot \pi_u \cdot \sigma_u^2$) and Y ($h_y^2 := M \cdot \pi_y \cdot \sigma_y^2$). Note that when $q_x = 0$ and $q_y \neq 0$ (or vice versa), this means that there is no confounder U present, but the genetic architecture of Y (or X) can be better described by a three component Gaussian mixture distribution.

We assume that the correlation (across markers) between the direct effects of a genetic variant on X, Y and U is zero, i.e. $cov(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}) = cov(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_u}) = cov(\boldsymbol{\gamma_u}, \boldsymbol{\gamma_y}) = 0$. Note that this

assumption still allows for a potential correlation between the total effect of G on X and its horizontal pleiotropic effect on Y, but only due to the confounder U and through the reverse causal effect $Y \to X$. As we argued above, this is a reasonable assumption, since the most plausible reason (apart from outcome-dependent sampling, which is out of the scope of this paper) for the violation of the InSIDE assumption may be one or more heritable confounder(s).

For simplicity, we also assume that the set of genetic variants with direct effects on each trait overlap only randomly, i.e. the fraction of the genome directly associated with both X and Y is $\pi_x \cdot \pi_y$, etc. This assumption is in line with recent observation that the bulk of observed pleiotropy can be explained by extreme polygenicity with random overlap between trait loci[13]. Note that uncorrelated effects (e.g. $cov(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}) = 0$) do not ensure that the active variant sets overlap randomly, this is a slightly stronger assumption.

2.2 The observed association summary statistics

Let us now assume that we observe univariable association summary statistics for these two traits from two (potentially overlapping) finite samples N_x and N_y of size n_x, n_y, respectively. In the following, we will derive observed summary statistics in sample N_x and then we will repeat the analogous exercise for sample N_y. Let the realisations of X, Y and U be denoted by $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{u} \in \mathcal{R}^{n_x}$. The genome-wide genetic data is represented by $\mathsf{G}_x \in \mathcal{R}^{n_x \times M}$ and the genetic data for a single nucleotide polymorphism (SNP) k tested for association is $\boldsymbol{g_k} \in \mathcal{R}^{n_x}$. Note the distinction between the k-th column of G_x, which is the k-th sequence variant, in contrast to $\boldsymbol{g_k}$, which is the k-th SNP tested for association in the GWAS. We assume that all SNP genotypes have been standardised to have zero mean and unit variance. The marginal effect size estimate for SNP k of trait X can then be written as $\widehat{\beta}_k^x = \boldsymbol{g}_k' \cdot \boldsymbol{x}/n_x$, which is a special case of univariable standard normal linear regression when both the outcome and the predictor is standardised to have zero mean and unit variance[12]. Note that $\boldsymbol{x}'$ denotes the transposed of the column vector $\boldsymbol{x}$. This can be further transformed as

$$\begin{aligned}
\widehat{\beta}_k^x &= \boldsymbol{g}_k' \cdot \boldsymbol{x}/n_x \\
&= \frac{q_x + \alpha_{y\to x} \cdot q_y}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot \boldsymbol{g}_k' \cdot \mathsf{G}_x \cdot \boldsymbol{\gamma_u}/n_x + \frac{\alpha_{y\to x}}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot \boldsymbol{g}_k' \cdot \mathsf{G}_x \cdot \boldsymbol{\gamma_y}/n_x \\
&+ \frac{1}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot \boldsymbol{g}_k' \cdot \mathsf{G}_x \cdot \boldsymbol{\gamma_x}/n_x + \boldsymbol{g}_k' \cdot \boldsymbol{\epsilon_x}/n_x
\end{aligned}$$

By denoting the linkage disequilibrium (LD) between variant k and all markers in the genome with $\boldsymbol{\rho_k} = \mathsf{G}_x' \cdot \boldsymbol{g}_k/n_x$ we get

$$\widehat{\beta}_k^x = \frac{q_x + \alpha_{y\to x} \cdot q_y}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot \boldsymbol{\rho_k}' \cdot \boldsymbol{\gamma_u} + \frac{\alpha_{y\to x}}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot \boldsymbol{\rho_k}' \cdot \boldsymbol{\gamma_y} + \frac{1}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot \boldsymbol{\rho_k}' \cdot \boldsymbol{\gamma_x} + \eta_k^x$$

with $\eta_k^x := \boldsymbol{g}_k' \cdot \boldsymbol{\epsilon_x}/n_x \sim \mathcal{N}(0, i_x/n_x)$. Given the above-defined genetic effect size distribution the equation becomes

$$\begin{aligned}
\widehat{\beta}_k^x &= \frac{q_x + \alpha_{y\to x} \cdot q_y}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot \underbrace{\boldsymbol{\rho_k}' \cdot (\boldsymbol{\zeta_u} \odot \boldsymbol{\kappa_u})}_{z_k^{(u)}} + \frac{\alpha_{y\to x}}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot \underbrace{\boldsymbol{\rho_k}' \cdot (\boldsymbol{\zeta_y} \odot \boldsymbol{\kappa_y})}_{z_k^{(y)}} \\
&+ \frac{1}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot \underbrace{\boldsymbol{\rho_k}' \cdot (\boldsymbol{\zeta_x} \odot \boldsymbol{\kappa_x})}_{z_k^{(x)}} + \eta_k^x \\
&= \frac{q_x + \alpha_{y\to x} \cdot q_y}{1 - \alpha_{y\to x}\alpha_{y\to x}} \cdot z_k^{(u)} + \frac{\alpha_{y\to x}}{1 - \alpha_{x\to y}\alpha_{y\to x}} \cdot z_k^{(y)} + \frac{1}{1 - \alpha_{x\to y}\alpha_{y\to x}} z_k^{(x)} + \eta_k^x
\end{aligned}$$

Similarly, assuming that the LD structures ($\boldsymbol{\rho_k}$) in the two samples are comparable, for $\widehat{\beta}_k^y$ estimated in the other sample (N_y) we obtain

$$\widehat{\beta}_k^y = \frac{\alpha_{x\to y}\cdot q_x + q_y}{1-\alpha_{x\to y}\alpha_{y\to x}}\cdot z_k^{(u)} + \frac{\alpha_{x\to y}}{1-\alpha_{x\to y}\alpha_{y\to x}}\cdot z_k^{(x)} + \frac{1}{1-\alpha_{x\to y}\alpha_{y\to x}} z_k^{(y)} + \eta_k^y$$

with $\eta_k^y \sim \mathcal{N}(0, i_y/n_y)$.

Therefore, the joint effect size estimates can be written as

$$\begin{pmatrix}\widehat{\beta}_k^x\\ \widehat{\beta}_k^y\end{pmatrix} = \frac{1}{1-\alpha_{x\to y}\alpha_{y\to x}}\left(\begin{pmatrix}(\alpha_{y\to x}\cdot q_y + q_x)\\ (\alpha_{x\to y}\cdot q_x + q_y)\end{pmatrix} z_k^{(u)} + \begin{pmatrix}1\\ \alpha_{x\to y}\end{pmatrix} z_k^{(x)} + \begin{pmatrix}\alpha_{y\to x}\\ 1\end{pmatrix} z_k^{(y)}\right) + \begin{pmatrix}\eta_k^x\\ \eta_k^y\end{pmatrix}$$

Following the same rational as the cross-trait LD score regression[14], the noise term distribution is readily obtained

$$\begin{pmatrix}\eta_k^x\\ \eta_k^y\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix} i_x/n_x & \frac{n_{x\cap y}}{n_x\cdot n_y}\cdot r_{x,y}\\ \frac{n_{x\cap y}}{n_x\cdot n_y}\cdot r_{x,y} & i_y/n_y\end{pmatrix}\right)$$

where $r_{x,y}$ is the observational correlation between variables X and Y and $n_{x\cap y}$ is the size of the overlapping samples for X and Y. Since both $n_{x\cap y}$ and $r_{x,y}$ cannot be estimated, we simply denote $i_{x,y} := r_{x,y}\cdot\frac{n_{x\cap y}}{\sqrt{n_x\cdot n_y}}$ as the only estimated parameter and parameterise the covariance term as $\frac{i_{x,y}}{\sqrt{n_x\cdot n_y}}$. Note that $i_{x,y}$ is the cross-trait LD score regression intercept.

The bivariate probability density function (PDF) of these summary statistics cannot be obtained analytically, but in the following we demonstrate that the characteristic function can be derived. Let us first compute the characteristic function of this two-dimensional random variable, knowing that $z_k^{(x)}, z_k^{(u)}, z_k^{(y)}$ and (η_k^x, η_k^y) are independent, hence the characteristic function can be factorised:

$$\begin{aligned}
\varphi_{(\widehat{\beta}_k^x,\widehat{\beta}_k^y)}(v,w) &= E\left[\exp\left(i\cdot(v\cdot\widehat{\beta}_k^x + w\cdot\widehat{\beta}_k^y\right)\right]\\
&= E\left[\exp\left(i\cdot\left(v\cdot\left(\frac{z_k^{(x)} + (\alpha_{y\to x}\cdot q_y + q_x)\cdot z_k^{(u)} + \alpha_{y\to x}\cdot z_k^{(y)}}{1-\alpha_{x\to y}\alpha_{y\to x}} + \eta_k^x\right) +\right.\right.\right.\\
&+ \left.\left.\left. w\cdot\left(\frac{z_k^{(y)} + (\alpha_{x\to y}\cdot q_x + q_y)\cdot z_k^{(u)} + \alpha_{x\to y}\cdot z_k^{(x)}}{1-\alpha_{x\to y}\alpha_{y\to x}} + \eta_k^y\right)\right)\right)\right]\\
&= E\left[\exp\left(i\cdot z_k^{(u)}\cdot\frac{v\cdot(\alpha_{y\to x}\cdot q_y + q_x) + w\cdot(\alpha_{x\to y}\cdot q_x + q_y)}{1-\alpha_{x\to y}\alpha_{y\to x}}\right)\right]\\
&\times E\left[\exp\left(i\cdot z_k^{(x)}\cdot\frac{v+\alpha_{x\to y}\cdot w}{1-\alpha_{x\to y}\alpha_{y\to x}}\right)\right]\cdot E\left[\exp\left(i\cdot z_k^{(y)}\cdot\frac{w+\alpha_{y\to x}\cdot v}{1-\alpha_{x\to y}\alpha_{y\to x}}\right)\right]\\
&\times E\left[\exp\left(i\cdot\left(v\cdot\eta_k^x + w\cdot\eta_k^y\right)\right)\right]\\
&= \varphi_{z_k^{(u)}}\left(\frac{v\cdot(\alpha_{y\to x}\cdot q_y + q_x) + w\cdot(\alpha_{x\to y}\cdot q_x + q_y)}{1-\alpha_{x\to y}\alpha_{y\to x}}\right)\\
&\times \varphi_{z_k^{(x)}}\left(\frac{v+\alpha_{x\to y}\cdot w}{1-\alpha_{x\to y}\alpha_{y\to x}}\right)\cdot\varphi_{z_k^{(y)}}\left(\frac{w+\alpha_{y\to x}\cdot v}{1-\alpha_{x\to y}\alpha_{y\to x}}\right)\cdot\varphi_{(\eta_k^x,\eta_k^y)}(v,w)
\end{aligned}$$

In the following we will work out each of the characteristic functions on the right hand side.

2.3 Characteristic functions of $z_k^{(u)}$, $z_k^{(x)}$, $z_k^{(y)}$ and (η_k^x, η_k^y)

It is reasonable to assume that linkage disequilibrium (LD) fades off beyond 1Mb distance. Thus, without loss of generality we can assume that non-zero LD does not extend beyond m_0 markers around the focal variant. Hence we can assume that the length of $\boldsymbol{\rho}_k$ is m_0 and only consider $\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}$ and $\boldsymbol{\gamma_u}$ to be of length m_0 instead of m. Let us first approximate the distribution of $\boldsymbol{\rho}_k$ values following a spike and slab Gaussian mixture, i.e. proportion π_k of the m_0 SNPs have non-zero LD, coming from a Gaussian distribution $\mathcal{N}(0, \sigma_k^2)$ and the remaining $(1-\pi_k)$ fraction of the LD values is zero. In mathematical notation

$$\boldsymbol{\rho}_k = \boldsymbol{r}_k \odot \boldsymbol{\kappa}_k \quad \text{with} \quad \boldsymbol{r_k} \sim \mathcal{N}(0, \sigma_k^2 \cdot I) \quad \text{and} \quad \boldsymbol{\kappa}_k \sim \mathcal{B}_{m_0}(1, \pi_k)$$

Therefore $z_k^{(u)}$ can be written of the form

$$\begin{aligned} z_k^{(u)} &= \boldsymbol{\rho_k'} \cdot (\boldsymbol{\zeta_u} \odot \boldsymbol{\kappa_u}) = (\boldsymbol{r}_k \odot \boldsymbol{\kappa}_k)' \cdot (\boldsymbol{\zeta_u} \odot \boldsymbol{\kappa_u}) = (\boldsymbol{r}_k \odot \boldsymbol{\zeta_u})' \cdot (\underbrace{\boldsymbol{\kappa}_k \odot \boldsymbol{\kappa_u}}_{\boldsymbol{\kappa_{k,u}} \sim \mathcal{B}(1, \pi_k \cdot \pi_u)}) \\ &= \sum_{j=1}^{m_0} (\boldsymbol{r}_k \odot \boldsymbol{\zeta_u})_j \cdot (\boldsymbol{\kappa_{k,u}})_j \end{aligned}$$

The PDF of the product of two zero-mean Gaussians ($\boldsymbol{r}_k$ and $\boldsymbol{\zeta_u}$) is a modified Bessel function of the second kind of order zero ($K_0(\omega)$) [15], more precisely

$$f_{(\boldsymbol{r}_k \odot \boldsymbol{\zeta_u})_j}(t) = \frac{1/\pi}{\sigma_u \cdot \sigma_k} \cdot K_0\left(\frac{|t|}{\sigma_u \cdot \sigma_k}\right)$$

and its characteristic function [16,17] is

$$\varphi_0(t) = E(\exp(i \cdot t \cdot (\boldsymbol{r}_k \odot \boldsymbol{\zeta_u})_j)) = \frac{1}{\sqrt{\sigma_u^2 \cdot \sigma_k^2 \cdot t^2 + 1}}$$

Next, the characteristic function of the product of $(\boldsymbol{r}_k \odot \boldsymbol{\zeta_u})_j$ and a Bernoulli distributed $(\boldsymbol{\kappa_{k,u}})_j$ is

$$\begin{aligned} \varphi_1(t) &= E\left(\exp\left(i \cdot t \cdot (\boldsymbol{r}_k \odot \boldsymbol{\zeta_u})_j\right) \cdot (\boldsymbol{\kappa_{k,u}})_j\right)) \\ &= \pi_k \cdot \pi_u \cdot E(\exp(i \cdot t \cdot (\boldsymbol{r}_k \odot \boldsymbol{\zeta_u})_j)) + (1 - \pi_k \cdot \pi_u) \cdot E(\exp(i \cdot t \cdot 0)) \\ &= \pi_k \cdot \pi_u \cdot \varphi_0(t) + (1 - \pi_k \cdot \pi_u) \\ &= \frac{\pi_k \cdot \pi_u}{\sqrt{\sigma_u^2 \cdot \sigma_k^2 \cdot t^2 + 1}} + (1 - \pi_k \cdot \pi_u) \end{aligned}$$

Hence the characteristic function of the sum of m_0 independent random variables is the product of them, we have

$$\varphi_{z_k^{(u)}}(t) = \left(\frac{\pi_k \cdot \pi_u}{\sqrt{\sigma_u^2 \cdot \sigma_k^2 \cdot t^2 + 1}} + (1 - \pi_k \cdot \pi_u)\right)^{m_0}$$

Finally, we apply a first order Taylor series approximation (around 1) of the log of the characteristic function in order to speed up computation and improve numerical accuracy

$$\begin{aligned}
\log(\varphi_{z_k^{(u)}}(t)) &= m_0 \cdot \log\left(\frac{\pi_k \cdot \pi_u}{\sqrt{\sigma_u^2 \cdot \sigma_k^2 \cdot t^2 + 1}} + (1 - \pi_k \cdot \pi_u)\right) \\
&= m_0 \cdot \log\left(1 - \pi_k \cdot \pi_u \cdot \left(1 - \frac{1}{\sqrt{\sigma_u^2 \cdot \sigma_k^2 \cdot t^2 + 1}}\right)\right) \\
&\approx -m_0 \cdot \pi_k \cdot \pi_u \cdot \left(1 - \frac{1}{\sqrt{\sigma_u^2 \cdot \sigma_k^2 \cdot t^2 + 1}}\right)
\end{aligned}$$

Analogously, the approximation of the logarithm of the characteristic functions of $z_k^{(x)}$ and $z_k^{(y)}$ is

$$\begin{aligned}
\log(\varphi_{z_k^{(x)}}(t)) &\approx -m_0 \cdot \pi_k \cdot \pi_x \cdot \left(1 - \frac{1}{\sqrt{\sigma_x^2 \cdot \sigma_k^2 \cdot t^2 + 1}}\right) \\
\log(\varphi_{z_k^{(y)}}(t)) &\approx -m_0 \cdot \pi_k \cdot \pi_y \cdot \left(1 - \frac{1}{\sqrt{\sigma_y^2 \cdot \sigma_k^2 \cdot t^2 + 1}}\right)
\end{aligned}$$

Since the characteristic function of a centred multivariate Gaussian with variance-covariance matrix Σ is $\exp(-(1/2) \cdot \boldsymbol{t}' \cdot \Sigma \cdot \boldsymbol{t})$ we have

$$\log\left(\varphi_{(\eta_k^x, \eta_k^y)}(v, w)\right) = -\frac{1}{2} \cdot \left(\frac{i_x}{n_x} \cdot v^2 + 2 \cdot \frac{i_{x,y}}{\sqrt{n_x \cdot n_y}} \cdot v \cdot w + \frac{i_y}{n_y} \cdot w^2\right)$$

2.4 From characteristic function to probability density function

The final form of the logarithm of the joint characteristic function of the transformed summary statistics is

$$\begin{aligned}
\log\left(\varphi_{\left(\widehat{\beta}_k^x,\widehat{\beta}_k^y\right)}(v,w)\right) &= \log\left(\varphi_{z_k^{(x)}}\left(\frac{v+\alpha_{x\to y}w}{1-\alpha_{x\to y}\alpha_{y\to x}}\right)\right)+\log\left(\varphi_{z_k^{(y)}}\left(\frac{w+\alpha_{y\to x}v}{1-\alpha_{x\to y}\alpha_{y\to x}}\right)\right)\\
&+ \log\left(\varphi_{z_k^{(u)}}\left(\frac{v\cdot(\alpha_{y\to x}\cdot q_y+q_x)+w\cdot(\alpha_{x\to y}\cdot q_x+q_y)}{1-\alpha_{x\to y}\alpha_{y\to x}}\right)\right)\\
&+ \log\left(\varphi_{\left(\eta_k^x,\eta_k^y\right)}(v,w)\right)\\
&\approx -m_0\cdot\pi_k\cdot\pi_x\cdot\left(1-\frac{1}{\sqrt{\frac{\sigma_x^2\cdot\sigma_k^2\cdot(v+\alpha_{x\to y}w)^2}{(1-\alpha_{x\to y}\alpha_{y\to x})^2}+1}}\right) \qquad (1)\\
&- m_0\cdot\pi_k\cdot\pi_y\cdot\left(1-\frac{1}{\sqrt{\frac{\sigma_y^2\cdot\sigma_k^2\cdot(w+\alpha_{y\to x}v)^2}{(1-\alpha_{x\to y}\alpha_{y\to x})^2}+1}}\right)\\
&- m_0\cdot\pi_k\cdot\pi_u\cdot\left(1-\frac{1}{\sqrt{\frac{\sigma_u^2\cdot\sigma_k^2\cdot(v\cdot(\alpha_{y\to x}\cdot q_y+q_x)+w\cdot(\alpha_{x\to y}\cdot q_x+q_y))^2}{(1-\alpha_{x\to y}\alpha_{y\to x})^2}+1}}\right)\\
&- \frac{1}{2}\cdot\left(\frac{i_x}{n_x}\cdot v^2+2\cdot\frac{i_{x,y}}{\sqrt{n_x\cdot n_y}}\cdot v\cdot w+\frac{i_y}{n_y}\cdot w^2\right)
\end{aligned}$$

Using the inversion theorem for characteristic functions we can express the joint distribution of $\left(\widehat{\beta}_k^x,\widehat{\beta}_k^y\right)$ as

$$f_{\left(\widehat{\beta}_k^x,\widehat{\beta}_k^y\right)}(x,y) = \left(\frac{1}{2\pi}\right)^2\cdot\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(-i\cdot(x\cdot v+y\cdot w))\cdot\varphi_{\left(\widehat{\beta}_k^x,\widehat{\beta}_k^y\right)}(v,w)\;dv\;dw$$

This integral can be efficiently computed by Fast Fourier Transformation (FFT, see [18] and references within). To speed up computation, we bin SNPs according to their π_k and σ_k values (10×10 bins with equidistant centres) and for SNPs in the same bin the PDF function is evaluated over a fine grid ($2^7\times 2^7$ combinations) using the FFT.

To reduce the number of parameters we define $t_x := \sigma_u\cdot q_x$ and $t_y := \sigma_u\cdot q_y$ since σ_u and q_x are separately not identifiable, but only their product is. Similarly π_u is unidentifiable, and is set to an arbitrary value of 0.1. For improved interpretability, we slightly reparameterise the likelihood function by using $h_x^2 := \pi_x\cdot M\cdot\sigma_x^2, h_y^2 := \pi_y\cdot M\cdot\sigma_y^2$. Since different SNPs are correlated we have to estimate the over-counting of each SNP. We choose the same strategy as LD score regression[12] and weigh each SNP by the inverse of its restricted LD score, i.e. $w_k = 1/\sum_{j=1}^{m_0} r_{jk}^2$, where r_{jk} is the correlation between GWAS SNPs k and j. The log-likelihood function is, thus, of the form

$$\log\left(\mathcal{L}\left(\boldsymbol{\theta}|\begin{pmatrix}\widehat{\boldsymbol{\beta}}^{\boldsymbol{x}}\\ \widehat{\boldsymbol{\beta}}^{\boldsymbol{y}}\end{pmatrix}\right)\right)\propto\sum_{k=1}^{K}w_k\cdot f_k\left(\widehat{\beta}_k^x,\widehat{\beta}_k^y\right) \qquad (2)$$

where $f_k\left(\widehat{\beta}_k^x,\widehat{\beta}_k^y\right)$ is the log-likelihood function value for SNP k. Parameters $\{n_x,n_y,m,\sigma_{k=1,\ldots,K},\pi_{k=1,\ldots,K}\}$ are known and the other 11 parameters

$$\boldsymbol{\theta} = \{\pi_x,\pi_y,h_x^2,h_y^2,t_x,t_y,\alpha_{x\to y},\alpha_{y\to x},i_x,i_y,i_{x,y}\}$$

are to be estimated from the observed association summary statistics. In order to speed up computation, we can estimate the 11 parameters in two separate steps: the first estimates for each trait the parameters π_x, i_x and π_y, i_y (SNP polygenicity and LD-score intercept) and the total heritability (unlike the direct heritability obtained by the full-model of LHC-MR) by using a simplified model with only the trait of interest, without a second trait or confounder, e.g. we fit only π_x, h_x^2 and i_x using $\widehat{\boldsymbol{\beta}}^{\boldsymbol{x}}$ and assume that π_x and i_x do not change when two traits are taken into account. Note that π_x may change slightly (decreasing from the total- to direct polygenicity), but its value has little impact on the likelihood function. The estimates from the first step can then be fixed for the parameter estimation of trait pairs. Since only π_x, i_x and π_y, i_y are fixed, the remaining parameters to estimate are now:

$$\boldsymbol{\theta} = \{h_x^2, h_y^2, t_x, t_y, \alpha_{x\to y}, \alpha_{y\to x}, i_{x,y}\}$$

It is key to note that our approach does not aim to estimate individual (direct or indirect) SNP effects, as these are handled as random effects. By replacing U with $-U$ we swap the signs of both t_x and t_y, therefore these parameters are unique only if the sign of one of them is fixed. Thus, we will have the following restrictions on the parameter ranges: h_x^2, h_y^2, t_x are in $[0, 1]$, $t_y, \alpha_{x\to y}, \alpha_{y\to x}, i_{x,y}$ are in $[-1, 1]$.

2.5 Likelihood maximisation and standard error calculation

Our method, termed *Latent Heritable Confounder Mendelian Randomisation (LHC-MR)*, maximises this likelihood function to obtain the maximum likelihood estimate (MLE). Due to the complexity of the likelihood surface, we initialise the maximisation using 50 different starting points, where they come from a uniform distribution within the parameter-specific ranges mentioned above. We then choose parameter estimates corresponding to the highest likelihood of the 50 runs. Run time depends on the number of iterations during the maximisation procedure, and is linear with respect to the number of SNPs. It takes ~ 0.25 CPU-minute to fit the complete model to 50,000 SNPs with a single starting point.

Given the particular nature of the underlying directed graph, two different sets of parameters lead to an identical fit of the data, resulting in two global optima. The reason for this is the difficulty in distinguishing the ratio of the confounder effects (t_y/t_x) from the causal effect ($\alpha_{x\to y}$), as illustrated in Figure S2 by the slopes belonging to different SNP-clusters. More rigorously, it can be show that if $\{h_x, h_y, \alpha_{x\to y}, \alpha_{y\to x}, t_x, t_y\}$ is an optimum, then so will be $\{h'_x, h'_y, \alpha'_{x\to y}, \alpha'_{y\to x}, t'_x, t'_y\}$, where

$$\begin{aligned}
h'_x &= t_x + t_y \cdot \alpha_{y\to x} \\
h'_y &= h_y \\
\alpha'_{x\to y} &= \frac{\alpha_{x\to y} + w}{1 + \alpha_{y\to x} \cdot w} \\
\alpha'_{y\to x} &= \alpha_{y\to x} \\
t'_x &= h_x \cdot (1 + \alpha_{y\to x} \cdot w) \\
t'_y &= -h_x \cdot w
\end{aligned}$$

with $w = t_y/t_x$ (for further derivations, see Supplementary Section 1.1). This allows us to directly obtain both optima, even if the optimisation only revealed one of them. It happens very often that one of these parameter sets are outside of the allowed ranges and hence can be automatically excluded. If not, we keep track of both parameter estimates maximising the likelihood function. Note that, we call the one for which the direct heritability is larger than the indirect one, i.e. $h_x^2 > t_x^2$, the primary solution. We show that for real data application this

solution is far more plausible than the alternative optimum. Finally, note that such bimodality can be observed at different levels: (i) For one given data generation, using multiple starting points leads to different optima; (ii) LHC-MR applied to multiple different data generations for a fixed parameter setting can yield different optima. Both of these situations are signs of the same underlying phenomenon and most often co-occur.

We implemented the block jackknife procedure that is also used by LD score regression to calculate the standard errors. For this we split the genome into 200 jackknife blocks and compute MLE in a leave-one-block-out fashion yielding $\widehat{\boldsymbol{\theta}}^{(-i)}, i = 1, \ldots, 200$ estimates. The variance of the full SNP MLE is then defined as $Var(\widehat{\theta}) := \frac{m - m\cdot(1/200)}{m\cdot(1/200)} \cdot \frac{1}{200-1} \sum_{i=1}^{200} (\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}})^2 = \sum_{i=1}^{200} (\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}})^2$.

2.6 Decomposition of genetic correlation

Given the starting equations for X and Y we can calculate their genetic correlation. Denoting the total (multivariate) genetic effect for X and Y as $\boldsymbol{\delta_x}$ and $\boldsymbol{\delta_y}$, we can express them as follows

$$\begin{aligned} \boldsymbol{\delta_x} &= q_x \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x}\boldsymbol{\delta_y} + \boldsymbol{\gamma_x} \\ \boldsymbol{\delta_y} &= q_y \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y}\boldsymbol{\delta_x} + \boldsymbol{\gamma_y} \end{aligned}$$

Substituting the second equation to the first yields

$$\begin{aligned} \boldsymbol{\delta_x} &= q_x \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x}(q_y \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y}\boldsymbol{\delta_x} + \boldsymbol{\gamma_y}) + \boldsymbol{\gamma_x} \\ &= (q_x + \alpha_{y\to x}q_y) \cdot \boldsymbol{\gamma_u} + (\alpha_{y\to x}\alpha_{x\to y})\boldsymbol{\delta_x} + \alpha_{y\to x}\boldsymbol{\gamma_y} + \boldsymbol{\gamma_x} \\ &= ((q_x + \alpha_{y\to x}q_y) \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x}\boldsymbol{\gamma_y} + \boldsymbol{\gamma_x}) / (1 - \alpha_{y\to x}\alpha_{x\to y}) \end{aligned}$$

Similarly,

$$\boldsymbol{\delta_y} = ((q_y + \alpha_{x\to y}q_x) \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y}\boldsymbol{\gamma_x} + \boldsymbol{\gamma_y}) / (1 - \alpha_{y\to x}\alpha_{x\to y})$$

Thus the genetic covariance is

$$\begin{aligned} E[\boldsymbol{\delta_x} \cdot \boldsymbol{\delta_y}] &= ((q_x + \alpha_{y\to x}q_y) \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x}\boldsymbol{\gamma_y} + \boldsymbol{\gamma_x}) ((q_y + \alpha_{x\to y}q_x) \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y}\boldsymbol{\gamma_x} + \boldsymbol{\gamma_y}) / (1 - \alpha_{y\to x}\alpha_{x\to y})^2 \\ &= \left((q_x + \alpha_{y\to x}q_y)(q_y + \alpha_{x\to y}q_x)h_u^2 + \alpha_{y\to x}h_y^2 + \alpha_{x\to y}h_x^2\right) / (1 - \alpha_{y\to x}\alpha_{x\to y})^2 \\ &= \left((t_x + \alpha_{y\to x}t_y)(t_y + \alpha_{x\to y}t_x) + \alpha_{y\to x}h_y^2 + \alpha_{x\to y}h_x^2\right) / (1 - \alpha_{y\to x}\alpha_{x\to y})^2 \end{aligned}$$

and the heritabilities are

$$\begin{aligned} E[\boldsymbol{\delta_x^2}] &= \left((t_x + \alpha_{y\to x}t_y)^2 + \alpha_{y\to x}^2 h_y^2 + h_x^2\right) / (1 - \alpha_{y\to x}\alpha_{x\to y})^2 \\ E[\boldsymbol{\delta_y^2}] &= \left((t_y + \alpha_{x\to y}t_x)^2 + \alpha_{x\to y}^2 h_x^2 + h_y^2\right) / (1 - \alpha_{y\to x}\alpha_{x\to y})^2 \end{aligned}$$

Therefore the genetic correlation takes the form

$$corr(\boldsymbol{\delta_x}, \boldsymbol{\delta_y}) = \frac{(t_x + \alpha_{y\to x}t_y)(t_y + \alpha_{x\to y}t_x) + \alpha_{y\to x}h_y^2 + \alpha_{x\to y}h_x^2}{\sqrt{\left((t_x + \alpha_{y\to x}t_y)^2 + \alpha_{y\to x}^2 h_y^2 + h_x^2\right)\left((t_y + \alpha_{x\to y}t_x)^2 + \alpha_{x\to y}^2 h_x^2 + h_y^2\right)}} \quad (3)$$

These values can be compared to those obtained by LD score regression.

2.7 Computation of the LD scores

We first took 4,773,627 SNPs with info (imputation certainty measure) ≥ 0.99 present in the association summary files from the second round of GWAS by the Neale lab[19]. This set was restricted to 4,650,107 common, high-quality SNPs, defined as being present in both UK10K and UK Biobank, having MAF $> 1\%$ in both data sets, non-significant ($P_{diff} > 0.05$) allele frequency difference between UK Biobank and UK10K and residing outside the HLA region (chr6:28.5-33.5Mb). For these SNPs, LD scores and regression weights were computed based on 3,781 individuals from the UK10K study[20]. To estimate the local LD distribution for each SNP (k), characterised by π_k, σ_k^2, we fitted a two-component Gaussian mixture distribution to the observed local correlations (focal SNP +/− 2'500 markers with MAF$\geq 0.5\%$ in the UK10K): (1) one Gaussian component corresponding to zero correlations, reflecting only measurement noise (whose variance is proportional to the inverse of the reference panel size) and (2) a second component with zero mean and a larger variance than the first component (encompassing measurement noise plus non-zero LD).

2.8 Simulation settings

First, we tested LHC-MR using realistic parameter settings with a mild violation of the classical MR assumptions. These standard parameter settings consisted of simulating $m = 234{,}000$ SNPs for two non-overlapping cohorts of equal size (for simplicity) of $n_x = n_y = 50{,}000$ for each trait. X, Y and U were simulated with moderate polygenicity ($\pi_x = 5 \times 10^{-3}, \pi_y = 1 \times 10^{-2}, \pi_u = 5 \times 10^{-2}$), and considerable direct heritability ($h_x^2 = 0.25, h_y^2 = 0.2, h_u^2 = 0.3$). U had a confounding effect on the two traits as such, $q_x = 0.3, q_y = 0.2$ (resulting in $t_x = 0.16, t_y = 0.11$), and X had a direct causal effect on Y ($\alpha_{x \to y} = 0.3$), while the reverse causal effect from Y to X was set to null. Note that in this setting the total heritability of each of these traits is principally driven by direct effects and less than 10% of the total heritability is through a confounder and in case of Y less than an additional 8% of its total heritability is through X.
It is important to note that for each tested parameter setting, we generated 50 different data sets, and each data generation underwent a likelihood maximisation of Eq. 2 using 50 starting points, and produced estimated parameters corresponding to the highest likelihood (simplified schema in Figure S3).

In the following simulations, we changed various parameters of these standard settings to test the robustness of the method. We explored how increased sample size ($n_x = n_y = 500{,}000$) or differences in sample sizes ($(n_x, n_y) = (50{,}000, 500{,}000)$ and $(n_x, n_y) = (500{,}000, 50{,}000)$) influence causal effect estimates of LHC-MR and other MR methods. We also simulated data with no causal effect (or with no confounder) and then examined how LHC-MR estimates those parameters. Next, we varied our causal effects between the two traits by lowering $\alpha_{x \to y}$ to 0.1, and in another setting by introducing a reverse causal effect ($\alpha_{y \to x} = -0.1$). In addition, we tried to create extremely unfavourable conditions for all MR analyses by varying the confounding effects. We did this in several ways: (i) increasing q_x and q_y ($q_x = 0.75, q_y = 0.50$), (ii) having a confounder with causal effects of opposite signs on X and Y ($q_x = 0.3, q_y = -0.2$). We also drastically increased the proportion of SNPs with non-zero effect on traits X, Y and U (π_x, π_y and $\pi_u = 0.1, 0.15, 0.2$ respectively). We also simulated data whereby the confounder has lower ($\pi_u = 0.01$) polygenicity than the two focal traits.

Finally, we explored various violations of the assumptions of our model (see Section 2). First, we introduced two confounders in the simulated data, once with causal effects on X and Y that were concordant ($t_x^{(1)} = 0.16, t_y^{(1)} = 0.11, t_x^{(2)} = 0.22, t_y^{(2)} = 0.16$) in sign, and another with discordant effects ($t_x^{(1)} = 0.16, t_y^{(1)} = 0.11, t_x^{(2)} = 0.22, t_y^{(2)} = -0.16$), while still fitting the model with only one U. Second, we breached the assumption that the non-zero effects come from a

Gaussian distribution. By design, the first three moments of the direct effects are fixed: they have zero mean, their variance is defined by the direct heritabilities and they must have zero skewness because the effect size distribution has to be symmetrical. Therefore, to violate the normality assumption, we varied the kurtosis (2, 3, 5, and 10) of the distribution drawn from the Pearson's distribution family. Third, we tested the assumption of the direct effects on our traits coming from a two-component Gaussian mixture by introducing a third component and observing how the estimates were effected. In this simulation scenario we introduced a large effect third component for X while decreasing the polygenicity of U ($\pi_{x1} = 1 \times 10^{-4}, \pi_{x2} = 1 \times 10^{-2}, h^2_{x1} = 0.15, h^2_{x2} = 0.1, \pi_u = 1 \times 10^{-2}$).

2.9 Application to real summary statistics

Once we demonstrated favourable performance of our method on simulated data, we went on to apply LHC-MR to summary statics obtained from the UK Biobank and other meta-analytic studies (Table S1) in order to estimate pairwise bi-directional causal effect between 13 complex traits. The traits varied between conventional risk factors (such as low education, high body mass index (BMI), dislipidemia) and diseases (including diabetes and coronary artery disease among others). SNPs with imputation quality greater than 0.99, and minor allele frequency (MAF) greater than 0.5% were selected. Moreover, SNPs found within the human leukocyte antigen (HLA) region on chromosome 6 were removed due to the abundance of SNPs associated with autoimmune and infectious diseases as well as the complicated LD structure present in that region. For traits with total heritability below 2.5%, the outgoing causal effect estimates were ignored since instrumenting such barely heritable traits is questionable.

In order to perform LHC-MR between trait pairs, a set of overlapping SNPs was used as input for each pair. The effects of these overlapping SNPs were then aligned to the same effect allele in both traits. To decrease computation time further (while only minimally reducing power), we selected every 10th QC-filtered SNP as input for the analysis. We calculated regression weights using the UK10K panel, which may be sub-optimal for summary statistics not coming from the UK Biobank, but we have previously shown[21] that estimating LD in a ten-times larger data set (UK10K) outweighs the benefit of using smaller, but possibly better-matched European panel (1000 Genomes[22]).

We also ran IVW for each trait pair in both directions to estimate bi-directional causal effects as well as LD score regression to get the cross trait intercept term. We then added uniformly distributed ($\sim U(-0.1, 0.1)$) noise to these pre-estimated parameters to generate starting points for the second step of the likelihood optimisation. These closer-to-target starting points did not change the optimisation results, simply sped up the likelihood maximisation and increased the chances to converge to the same (primary) optimum. The LHC-MR procedure was run for each pair of traits 100 times, each using a different set of randomly generated starting points within the ranges of their respective parameters. For the optimisation of the likelihood function (Eq. 2), we used the R function 'optim' from the 'stats' R package[23]. Once we fitted this *complete* model estimating 11 parameters in two steps $\{i_x, i_y, \pi_x, \pi_y, h^2_x, h^2_y, t_x, t_y, \alpha_{x\to y}, \alpha_{y\to x}, i_{xy}\}$, we then ran block jackknife to obtain the SE of the parameters estimated in the second step: $\{h^2_x, h^2_y, t_x, t_y, \alpha_{x\to y}, \alpha_{y\to x}, i_{xy}\}$.

To support the existence of the confounders identified by LHC-MR, we used EpiGraphDB[24, 25] to systematically identify those potential confounders. The database provided for each potential confounder of a causal relationship, a causal effect on trait X and Y ($r1$, and $r3$ in their notation), the sign of the ratio of which ($sign(r_3/r_1)$) was compared to the sign of the LHC-MR estimated t_y/t_x values representing the strength of the confounder acting on the two traits. We restricted our comparison to the sign only, since the $r1, r3$ values reported in EpiGraphDB are

not necessarily on the same scale.

2.10 Comparison against conventional MR methods and CAUSE

We compared the causal parameter estimates of the LHC-MR method to those of five conventional MR approaches (MR-Egger, weighted median, IVW, mode MR, and weighted mode MR) using a Z-test[26]. The 'TwoSampleMR' R package[27] was used to get the causal estimates for all the pairwise traits as well as their standard errors from the above-mentioned MR methods. The same set of genome-wide SNPs that were used by LHC-MR, were used as input for the package. SNPs associated with the exposure were selected to various degrees (for simulation we selected SNPs over a range of thresholds: absolute P-value $< 5 \times 10^{-4}$ to $< 5 \times 10^{-8}$), and SNPs more strongly associated with the outcome than with the exposure (P-value < 0.05 in one-sided t-test) were removed. The default package settings for the clumping of SNPs ($r^2 = 0.001$) were used and the analysis was run with no further changes. We tested the agreement between the significance and direction of our estimates and that of standard MR methods, with the focus being on finding differences in statistical conclusions regarding causal effect sizes.

We compared our causal estimates from all our simulation settings to the causal estimates obtained by running MR-RAPS [11] also using the 'TwoSampleMR' R package, once by using the entire set of SNPs, and another by filtering for SNPs with a significance threshold of $< 5 \times 10^{-4}$. We also compared both our simulation as well as real data results against those of CAUSE[10]. We first generated simulated data under the LHC model and used them as input to estimate the causal effect using CAUSE. We then generated simulated data using the CAUSE framework and inputted them to LHC-MR (as well as standard MR methods) to estimate the causal parameters. Lastly, we compared causal estimates obtained for the 78 trait pairs (156 bi-directional causal effects) from LHC-MR to those obtained when running CAUSE.

3 Results

3.1 Overview of the method

We fitted an 11-parameter structural equation model (SEM) (Figure 1) to genome-wide summary statistics of two studied complex traits in order to estimate bi-directional causal effects between them (for details see Methods). Additional model parameters represent direct heritabilities for X and Y, confounder effects, cross-trait and individual trait LD score intercepts and the polygenicity for X and Y. All SNPs associated with the heritable confounder (U) are indirectly associated with X and Y with effects that are proportional (ratio t_y/t_x). SNPs that are directly associated with X (and not with U) are also associated with Y with proportional effects (ratio $1/\alpha_{x\to y}$). Finally, SNPs that are directly Y-associated are also X-associated with a proportionality ratio of $1/\alpha_{y\to x}$. These three groups of SNPs are illustrated on the β_x-vs-β_y scatter plot (Figure S2). In simple terms, the aim of our method is to identify the different clusters, estimate the slopes and distinguish which corresponds to the causal- and confounder effects. In this paper, we focus on the properties of the maximum likelihood estimates (and their variances) for the bi-directional causal effects arising from our SEM.

3.2 Simulation results

We started off with a realistic simulation setting of 234,000 SNPs on chromosome 10 (LD patterns used from the UK10K panel) and 50,000 samples for both traits. Traits X, Y and confounder U had average polygenicity ($\pi_x = 5 \times 10^{-3}, \pi_y = 1 \times 10^{-2}, \pi_u = 5 \times 10^{-2}$), with substantial direct heritability for X and Y ($h_x^2 = 0.25, h_y^2 = 0.2$), mild confounding ($t_x = 0.16, t_y = 0.11$) and a

causal effect between X and Y ($\alpha_{x\to y} = 0.3, \alpha_{y\to x} = 0$). Note that with these settings, SNPs associated with U would violate the InSIDE assumption but might still be used by conventional MR methods. Under this standard setting, there were no genome-wide significant SNPs for standard MR methods, and estimates derived using SNPs with a p-value $< 5 \times 10^{-6}$ showed a downward bias for all MR methods (Figure 2 panel *a*). MR-RAPS using filtered SNPs (p-value $< 5 \times 10^{-4}$) was similarly downward biased whereas MR-RAPS using the entire set of SNPs was upward biased with the least amount of variance compared to all methods including LHC-MR. LHC-MR in this scenario slightly over estimated the causal effect in comparison but had the smallest RMSE after MR-RAPS (0.13 vs 0.06, Supplementary Table S2).

We ran all our simulation scenarios with a smaller and a larger sample size (50,000 and 500,000) and observed that the relative performance of the methods were in some cases sample size specific. Smaller sample sizes often meant that standard MR methods had little to no IVs reaching genome-wide (GW) significance and hence we were forced to use IVs from less stringent thresholds ($< 5 \times 10^{-4}$ and $< 5 \times 10^{-6}$). Therefore, the causal effects were estimated with a substantial downward bias due to weak instrument bias (and winner's curse). LHC-MR in these cases was able to estimate the causal effect with less bias but with a larger variance compared to most standard MR methods – still outperforming them in terms of RMSE in most settings. In the larger sample size setting, standard MR methods had IVs for every threshold cutoff. However, a pattern also observed with smaller sample sizes, but to a lesser extent, the causal estimates of some methods changed (either in mean or in variance, most noticeably observed in weighted median and IVW) as the threshold became more stringent. This is of particular concern and highlights that while in this simulation setting the 5×10^{-8} threshold may have optimally cancelled out the different biases for IVW (downward bias due to winner's curse and weak instrument bias, upward bias due to genetic confounding), its estimate remains strongly setting-dependent. LHC-MR has performed reasonably well, exhibiting lower RMSE than most other methods, except for IVW and MR-RAPS for the 5×10^{-4} threshold (Figure S4 panel *a*). However, we observed that the performance of MR-RAPs is particularly setting- and threshold dependent.

Furthermore, unequal sample sizes for the two traits showed an underestimation of the causal effects for almost all MR methods, while LHC-MR remained the most accurate in the case where n_x (50,000) was smaller than n_y (500,000). However, the performances in the reverse scenario, where n_x was larger in size, were akin to the large sample size standard setting, where only IVW and filtered MR-RAPS ($< 5 \times 10^{-4}$) showed superior performance to LHC-MR both in terms of bias and variance (see Figure S5).

When testing scenarios in the absence of a causal- or a confounder effect (imitating the classical MR assumptions), with a smaller causal effect ($\alpha_{x\to y} = 0.1$), or with both forward- and reverse causal effects, we note that LHC-MR outperforms the standard MR methods as well as MR-RAPS in all these scenarios.

When there was no causal effect ($\alpha_{x\to y} = 0$), LHC-MR had the smallest bias out of all the methods in both sample sizes (0.004 in both, Figure S6 panel *a* and Figure S7 panel *a*). The variance of the LHC-MR estimates in the larger sample size was much lower (0.0001 vs 0.01), similarly the other methods had a smaller variance in larger sample sizes and had more clearly seen upward biased estimates. The increased upward bias of standard MR methods is due to the fact that confounder-associated SNPs could only be detected in larger sample size and those lead to positive bias (due to the concordant effect of the confounder on the two traits). Note that the variance of standard MR methods are low simply because, in these settings, we were forced to lower the instrument selection threshold, hence artificially included many (potentially invalid) instruments, which lowers the estimator variance while increasing bias. MR-RAPS greatly overestimates the causal effects when the sample size is larger.

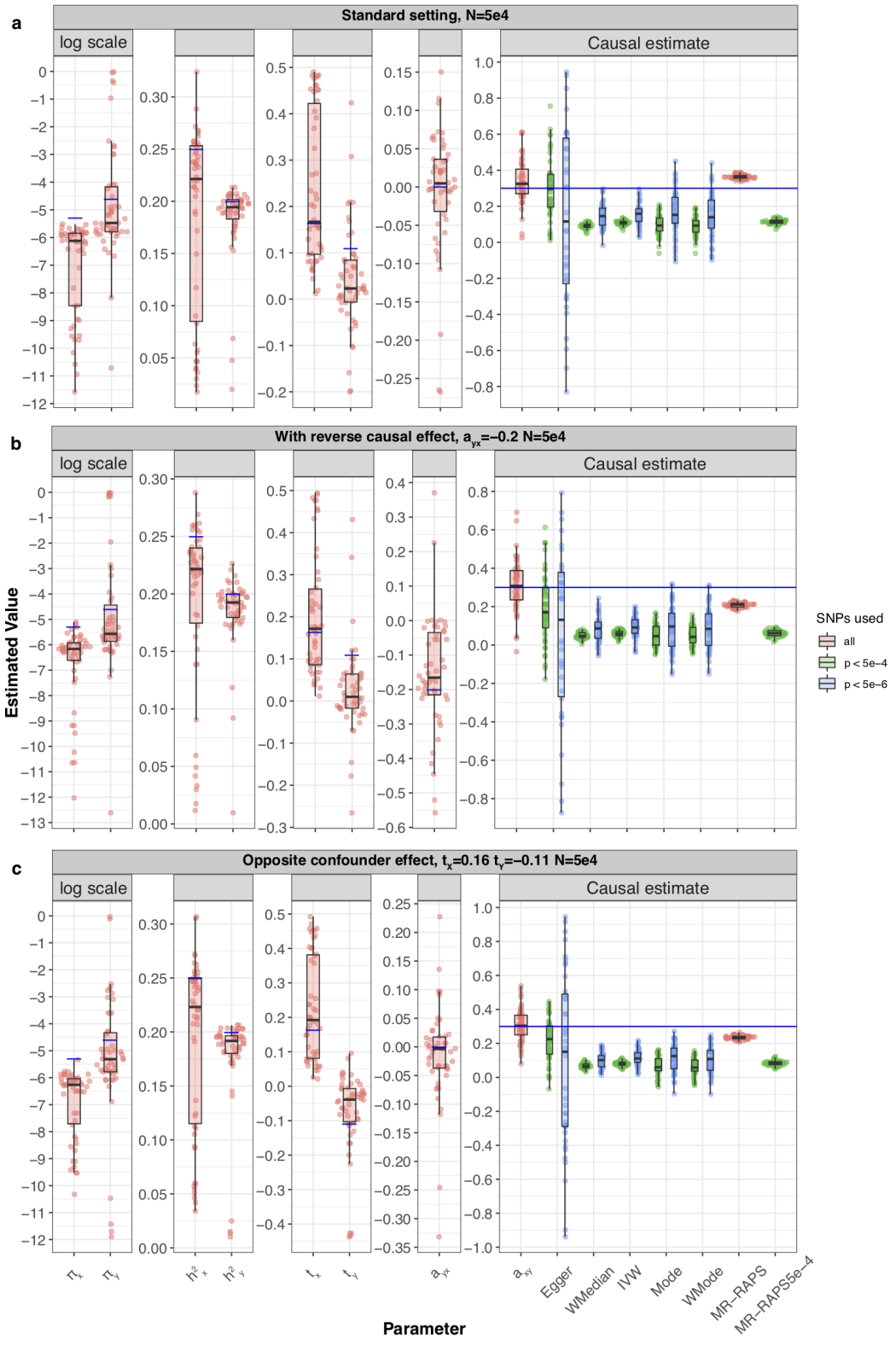


Figure 2: Simulation results under various scenarios. These Raincloud boxplots[28] represent the distribution of parameter estimates from 50 different data generations under various conditions. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** Estimation under standard settings ($\pi_x = 5 \times 10^{-3}, \pi_y = 1 \times 10^{-2}, \pi_u = 5 \times 10^{-2}, h^2_x = 0.25, h^2_y = 0.2, h^2_u = 0.3, t_x = 0.16, t_y = 0.11$). **b** Addition of a reverse causal effect $\alpha_{y \to x} = -0.2$. **c** Confounder with opposite causal effects on X and Y ($t_x = 0.16, t_y = -0.11$).

In the absence of a confounder effect, there is not much of a difference between the two sample sizes; standard MR methods have a large variance and are downward biased, LHC-MR is less biased compared to them but MR-RAPS performs best with the least bias and variance when all the SNPs are used as instruments (Figure S6 panel b and Figure S7 panel b). Trying a smaller causal effect led to an upward bias for all MR methods including both filterings of MR-RAPS in the larger sample size. Alternately, when $n_x = n_y = 50,000$, the MR methods are downward biased (Figures S6 panel c and Figures S7 panel c). Lastly, when a (negative) reverse causal effect is introduced, all MR methods and MR-RAPS are negatively biased in their estimation of the causal effect (see Figure 2 panel b). LHC-MR has a much smaller bias for the forward causal effect estimate in this case, and a generally small bias for the reverse causal effect in both sample sizes (0.05 for $n = 50,000$ and 0.03 for $n = 500,000$, Figure S4 panel b).

Increasing the indirect genetic effects, by intensifying the contribution of the confounder to X and Y ($t_x = 0.41, t_y = 0.27$), led to a general over estimation of the causal effects by all methods including LHC-MR, though more drastically seen in standard MR methods and MR-RAPS in larger sample sizes, when there is sufficient power to pick up these confounder-associated SNPs. The causal effect estimates of standard MR methods in the smaller sample size were much less affected by the presence of a strong confounder compared to LHC-MR and MR-RAPS (Figure S8). The reason for this is that the confounder-associated SNPs remain undetectable at lower sample size and hence instruments will not violate the classical MR assumptions.
Further testing the effects of the confounder trait on the causal estimation, we tested the impact of confounders with opposite effects on X and Y. We observe a major underestimation of the causal effects for standard MR methods as well as MR-RAPS, whereas LHC-MR performs better for both sample sizes (RMSE = 0.01 and 0.1 for larger and smaller n respectively), see Figures 2 panel c and S4 panel c.

Our LHC-MR method is influenced by the unlikely scenario of extreme polygenicity for traits X, Y and U, and it suffers from increased bias and variance regardless of sample size (see Figure S9). Standard MR methods as well as filtered MR-RAPS underestimated the causal effect when $n = 50,000$. Some also underestimated $\alpha_{x\to y}$ when $n = 500,000$, with the exception of IVW, Mode and filtered MR-RAPS, that outperformed the rest. Decreasing the proportion of confounder-associated SNPs to 1% only, does not seem to affect our method and shows similar results to the standard setting (Figure S10).

Furthermore, we simulated summary statistics, where (contrary to our modelling assumptions) the $X - Y$ relationship has two confounders, U_1 and U_2. When the ratio of the causal effects of these two confounders on X and Y ($q_x^{(1)}/q_y^{(1)}$ and $q_x^{(2)}/q_y^{(2)}$ respectively) agreed in sign, the corresponding causal effects of standard MR methods were over-estimated in larger sample sizes and, conversely, underestimated in smaller sample sizes (Figures S11 and S12, panels a). LHC-MR and weighted median performed better however in larger sample sizes and had a bias of 0.03 and 0.07 respectively. However, when the signs were opposite ($q_x^{(1)} = 0.3, q_y^{(1)} = 0.2$ for U_1 and $q_x^{(2)} = 0.3, q_y^{(2)} = -0.2$ for U_2), conventional MR methods and MR-RAPS in this case almost all underestimated the causal effect regardless of sample size. LHC-MR outperformed them both in the larger sample size (bias of 0.007) and in the smaller sample size (bias of -0.003), see Figures S11 and S12, panels b.

Finally, we explored how sensitive our method is to different violations of our modelling assumptions. First, we simulated summary statistics when the underlying non-zero effects come from a non-Gaussian distribution. Interestingly, we observed that, for smaller sample sizes, the variance of the causal effect estimate was dependent on the kurtosis for most MR methods. LHC-MR estimations yielded slightly more pronounced upward bias than IVW, while still exhibiting the lowest RMSE among all methods (Figure 3 panel a). Similar results are seen in larger sample

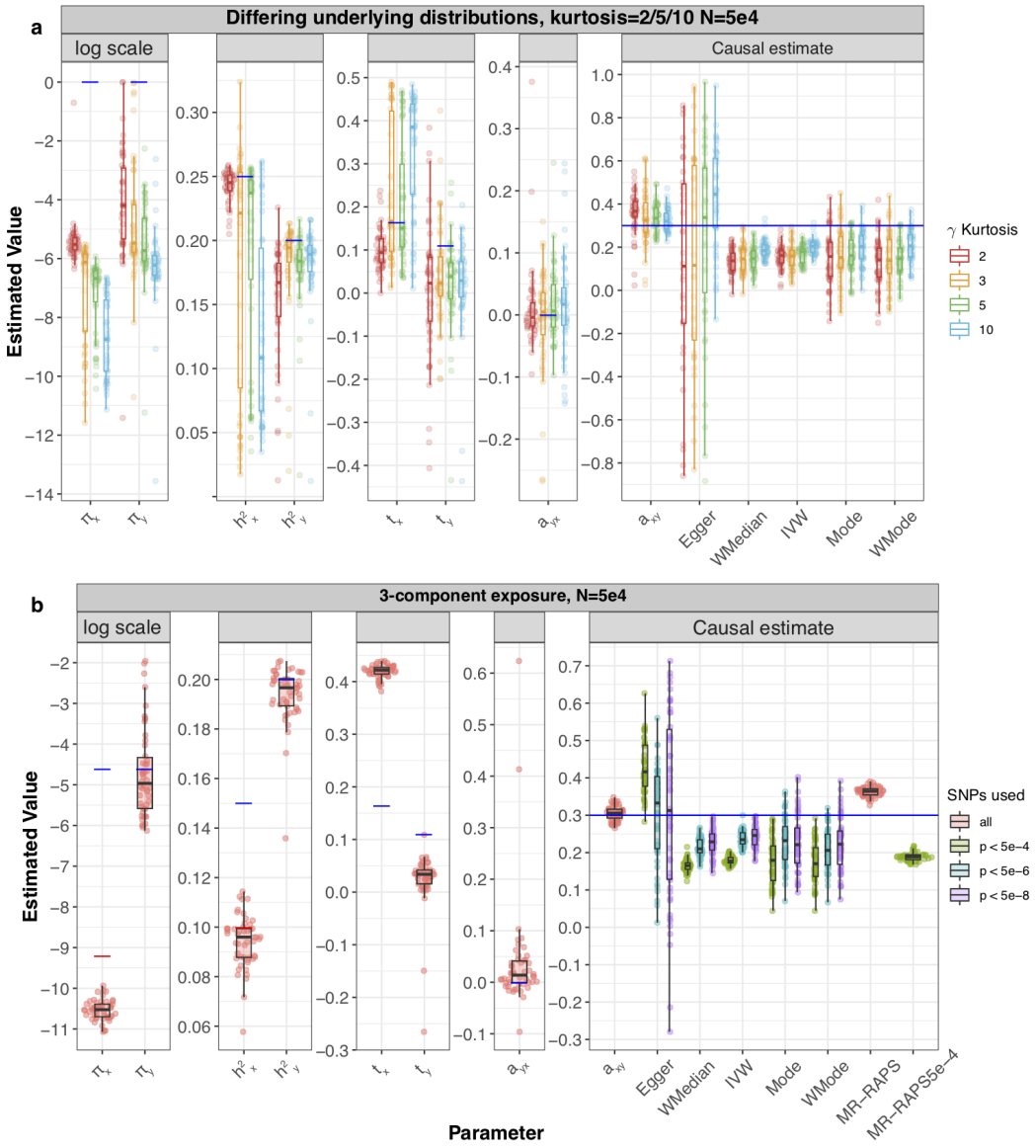


Figure 3: Simulation results under various scenarios. These Raincloud boxplots[28] represent the distribution of parameter estimates from 50 different data generations under various conditions. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** The different coloured boxplots represent the underlying non-normal distribution used in the simulation of the three $\gamma_x, \gamma_x, \gamma_u$ vectors associated to their respective traits. The Pearson distributions had the same zero mean and skewness, however their kurtosis ranged between 2 and 10, including the kurtosis of 3, which corresponds to a normal distribution assumed by our model. The standard MR results reported had IVs selected with a p-value threshold of 5×10^{-6}. **b** Addition of a third component for exposure X, while decreasing the strength of U. True parameter values are in colour, blue and red for each component ($\pi_{x1} = 1 \times 10^{-4}, \pi_{x2} = 1 \times 10^{-2}, h^2_{x1} = 0.15, h^2_{x2} = 0.1$).

size with smaller variance for all methods under all degrees of kurtosis except for IVW, which showed a better performance than LHC-MR (Figure S13 panel *a*). Second, we simulated effect sizes coming from a three-component Gaussian mixture distribution (null/small/large effects), instead of the classical spike-and-slab assumption of our model. The smaller sample size estimates mirror those of the standard setting with n also equal to $50,000$ (see Figure 3 panel *b*). However, in the larger sample size, LHC-MR overestimates the causal effect. This bias could be due to the merging of true effect estimates with confounder effect leading to an overestimation of $\alpha_{x\to y}$ (Figure S13 panel *b*). MR-Egger, IVW and filtered MR-RAPS have the smallest RMSE in this case.

3.2.1 Comparing CAUSE and LHC-MR

When running CAUSE on data simulated using the LHC-MR model framework in order to estimate a causal effect (γ in their notation), we investigated three different scenarios, each with multiple data generations: one where the underlying model has a shared factor/confounder with effect on both exposure and outcome only, another where the underlying model has a causal effect of 0.3 only, a third where the underlying model has both a causal effect and a shared factor. The data generated using the LHC-MR model was done under the standard settings ($\pi_x = 5\times10^{-3}, \pi_y = 1\times10^{-2}, \pi_u = 5\times10^{-2}$, $h^2_x = 0.25, h^2_y = 0.2, h^2_u = 0.3$, $t_x = 0.16, t_y = 0.11$, $\alpha_{x\to y} = 0.3, \alpha_{y\to x} = 0$, $m = 234,000, n_x = n_y = 50,000$). For each setting, 50 different replications were investigated.

In the case of an underlying shared effect only, CAUSE preferred the sharing model 100% of the time, and thus there was no causal estimation, however it underestimated both *eta* and q. When there was an underlying causal effect only, CAUSE preferred the causal model only 4% of the times, where it slightly underestimated the causal effect ($\widehat{\gamma} = 0.241$). Although the true values of η and q are null in this scenario, the sharing model returned estimates for these two parameters overestimating them both (probably driven by their priors), as seen in Figure S14. In the third case, and in the presence of both, CAUSE preferred the sharing model in 48 of the 50 simulations, yet it underestimated η (corresponding to t_y/t_x for our model) but overestimated q ($t^2_x/(t^2_x + h^2_x)$ in our model) (mean of 0.566 and 0.222 respectively where the true values are 0.667 and 0.097) showing a similar estimation pattern to the second case. Interestingly, in larger sample sizes, CAUSE selects the correct model 100% of the time, but still underestimates γ, see Figure S15.

In the reverse situation, where data was generated using the CAUSE framework (with parameters $h_1 = h_2 = 0.25, m = 97,450, N1 = N2 = 50,000$) and LHC-MR was used to estimate the causal effect, we saw the following results (see Figure S16). First, when we generated data in the absence of causal effect ($\gamma = 0, \eta = \sqrt{0.05}, q = 0.1$), CAUSE does extremely well in estimating a null causal effect 100% of the time. Standard MR methods yield a slight overestimation of the (null) causal effect with varying degrees of variance, whereas LHC-MR shows both a greater variance and an upward bias – still leading to a causal effect compatible with zero. Second, in the absence of a confounder combined with non-zero causal effect ($\gamma = \sqrt{0.05} = 0.22, \eta = 0, q = 0$), CAUSE underestimates the causal effect ($\widehat{\gamma} = 0.18$) compared to LHC-MR which overestimates the causal effect: the mean of the estimates was 0.38 (over the 50 runs). Finally, in the presence of both a confounder and a causal effect ($\gamma = \sqrt{0.05}, \eta = \sqrt{0.05}, q = 0.1$), CAUSE slightly underestimates the causal effect ($\widehat{\gamma} = 0.20$), whereas LHC-MR overestimates the effects and shows estimates reaching the boundaries 11 out of 50 times (mean of the converged $\widehat{\gamma} = 0.39$ over the 39 data simulations, see Figure S16 panel *c*) – indicating that this setting of the CAUSE model is not compatible with the LHC-MR model framework. Interestingly, classical MR methods outperform CAUSE in this case. Note that in the interest of run time we used less

SNPs (than usual) for parameter estimations. The analysis was repeated for a larger sample size of $500,000$ (Figure S17), with more favourable results for LHC-MR. In the absence of a causal effect, we had similar results to smaller sample sizes, whereas in the absence of a shared effect, LHC-MR estimates the causal effect accurately with a mean of 0.22, CAUSE underestimates it and the rest of the MR methods are less biased. In the presence of both causal and shared factor, CAUSE recovers the causal effect. IVW, unlike the other MR methods and CAUSE, is more affected by the presence of the confounder, while LHC-MR exhibits upward bias with a mean estimate of 0.27.

3.3 Application to association summary statistics of complex traits

We applied our LHC-MR and other MR methods to estimate all pairwise causal effects between 13 complex traits (156 causal relationships in both directions). Our results are presented as a heatmap in Figure 4 (and are detailed in Supplementary Table S3). Further, we calculated the alternate set of estimated parameters that naturally results from our model (for reference see Sections 2.2 and Supplementary materials 1.1). Among trait pairs for which the exposure had sufficient heritability ($> 2.5\%$), the alternate parameters of a 102 trait pairs were within the possible ranges mentioned in methods (i.e. the confounder and the exposure are interchangeable). However, for all of these pairs, the alternative parameter optima lead to lower direct- than indirect heritability, which we deem unrealistic. Therefore, we report only the primary set of estimated optimal parameters in the main results and provide the alternative parameters in the Supplementary Table S4. The comparison of the results obtained by LHC-MR and standard MR methods is detailed below and more extensively in Supplementary Tables S5-S6. In summary, LHC-MR provided reliable causal effect estimates for 132 out of 156 exposure traits (i.e. those exposures had an estimated total heritability greater than 2.5%). These estimates were compared to five different MR methods. Seventy-four causal relationships were deemed significant by LHC-MR. Furthermore, for 117 out of those 132 comparable causal relationships, our LHC-MR causal effect estimates were concordant (not significantly different) with at least two out of five standard MR methods' estimates.

By simply comparing the significance status and the direction of the causal effects between the methods, we see that LHC-MR agrees in sign and significance (or the lack there of) with at least 3 MR methods 77 times. For 31 relationships, LHC-MR results lead to different conclusions than those of standard MR methods. For 28 of those, LHC-MR identified a causal effect missed by all standard MR methods. For the other three, we observed a disagreement in sign: LDL has a negative effect on BMI according to weighted mode and weighted median, whereas we show a positive effect, HDL and LDL show a negative bi-directional causal effect for weighted mode but a positive bi-directional effect with LHC-MR. Despite the conflicting evidence for the causal relationship of LDL on BMI, studies have shown that the relationship between them is non-linear[29], possibly explaining the discrepancy between the results.

LHC-MR agreed with most MR estimates and confirmed many previous findings, such as increased BMI leading to elevated blood pressure[30,31], diabetes mellitus[32,33] (DM), myocardial infarction[34] (MI) and coronary artery disease[35] (CAD). Furthermore, we confirmed previous results[36] that diabetes increases SBP ($\hat{\alpha}_{x\to y} = 0.39 \ -P = 1.70 \times 10^{-9}$).

Interestingly, it revealed that higher BMI increases smoking intensity, concordant with other studies[37,38]. It has also shown the protective effect of education against a range of diseases (e.g. CAD and diabetes[39,40]) and risk factors such as smoking[41,42], in agreement with previous observational and MR studies. Probably reflecting lifestyle change recommendations by medical doctors upon disease diagnosis, statin use is greatly increased when being diagnosed with CAD, (systolic) hypertension, dislipidemia, and diabetes as is shown by both LHC-MR and standard MR methods.

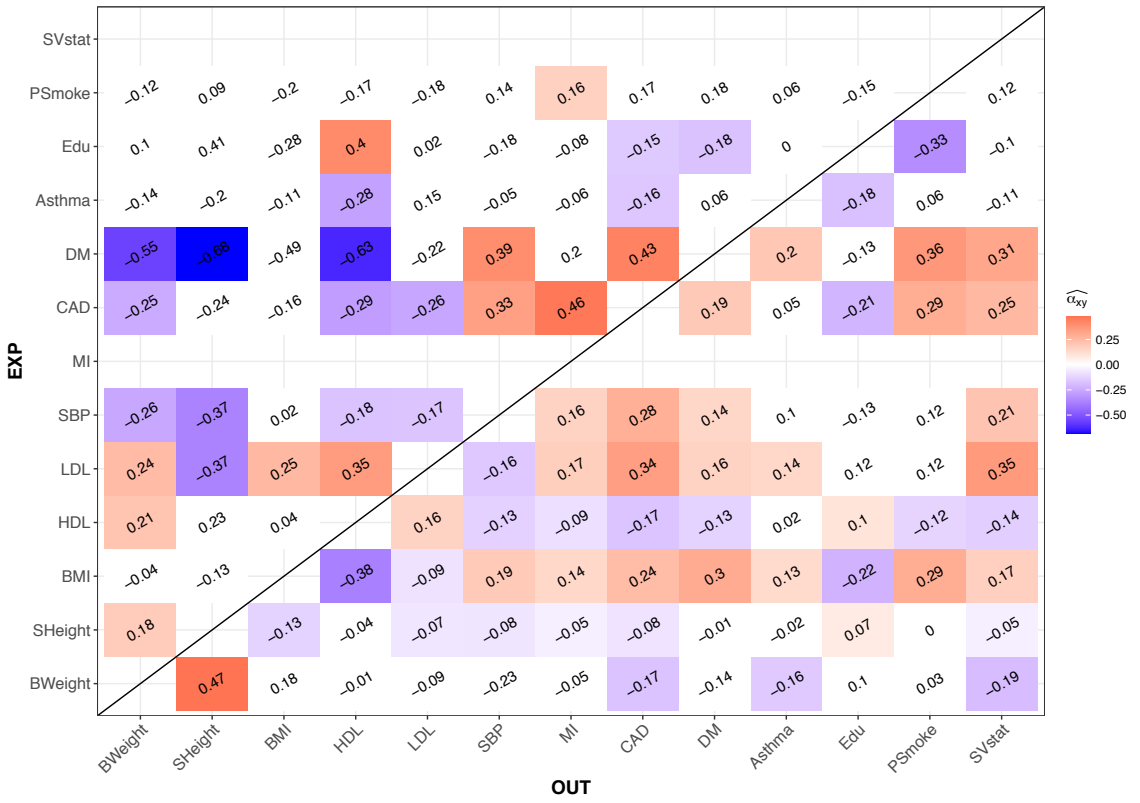


Figure 4: Heatmap representing the bi-directional causal relationship between the 13 UK Biobank traits. The causal effect estimates in coloured tiles all have a significant p-values surviving Bonferroni multiple testing correction with a threshold of 3.2×10^{-4}. We did not report an estimated causal effects for exposures with an estimated total heritability less than 2.5%. White tiles show an absence of a significant causal effect estimate.
Abbreviations: BMI: Body Mass Index, BWeight: Birth Weight, CAD: coronary artery disease, DM: Diabetes Mellitus, Edu: Years of Education, HDL: High-density Lipoprotein LDL: Low-density Lipoprotein, MI: Myocardial Infarction, PSmoke: # of Cigarettes Previously Smoked, SBP: Systolic Blood Pressure, SHeight: Standing Height, SVstat: Medication Simvastatin

Furthermore, causal effects of height on CAD, DM and SBP have been previously examined in large MR studies[43, 44]. LHC-MR, agreeing with these claims, did not find significant evidence to support the effect of height on DM, but did find a significant protective effect on CAD and SBP. However, unlike the first two, the relationship between height and SBP also revealed the existence of a confounder with causal effects 0.14 ($P = 9.2 \times 10^{-11}$) and 0.11 ($P = 3.39 \times 10^{-8}$) on height and SBP respectively. Another example of a trait pair for which LHC-MR found an opposite sign confounder effect is HDL and its protective effect on SBP. The confounder had a positive effect ratio of $t_y/t_x = 0.84$, opposing the negative causal effect of $\hat{\alpha}_{x\to y} = -0.13$ supported by observational studies[45]. This causal effect was not found by any other MR method.

It is important to note that while the effects of parental exposures on offspring outcomes can be seen as genetic confounding, LHC-MR would not be able to distinguish parental and offspring causal effects, because the LHC-MR model assumes that there is no correlation between the genetic effects on the exposure and the genetic effects on the confounder (which is not the case of parental vs offspring traits). Thus, LHC-MR causal effect estimates are just as likely to reflect parental effects as any other MR method [46]. This may be the case, for example,

for the detrimental effect of increased (parental) BMI on education (supported by longitudinal studies[47]), the positive effect of (parental) height on birth weight[48], or on education[49]. There are also some associations identified only by LHC-MR that might reflect parental effects: the negative causal effect of CAD on education or on birth weight, the positive impact of HDL on birth weight, or DM reducing height. All these pair associations uniquely found by LHC-MR are examples of LHC-MR's use of whole-genome SNPs instead of GW-significant SNPs only, as our estimates are of larger magnitude than those found by standard MR. Interestingly, for the CAD→birth weight relationship, LHC-MR revealed a confounder of opposite causal effects, which could have masked/mitigated the causal effect of standard MR methods.

A systematic comparison between IVW and LHC-MR has shown generally good agreement between the two methods, which is illustrated in Figure 5. To identify discrepancies between our causal estimates and those of the standard MR results, we grouped the estimates into several categories, either non-significant p-value for both or either, significant with an agreeing sign for the causal estimate, or significant with a disagreeing sign. The diagonal (seen in Figure 5) representing the agreement in significance status and sign between the two methods, is heavily populated. On the other hand, 34 pairs have causal links that are significantly non-zero according to LHC-MR, but are non-significant for IVW, while the opposite is true for seven pairs. We believe that many of these seven pairs may be false positives, since four of them are picked up by no other MR method, two are confirmed by only one other method and the last one by two methods. Further comparisons of significance between LHC-MR estimates and the remaining standard MR methods can be found in Table S7.

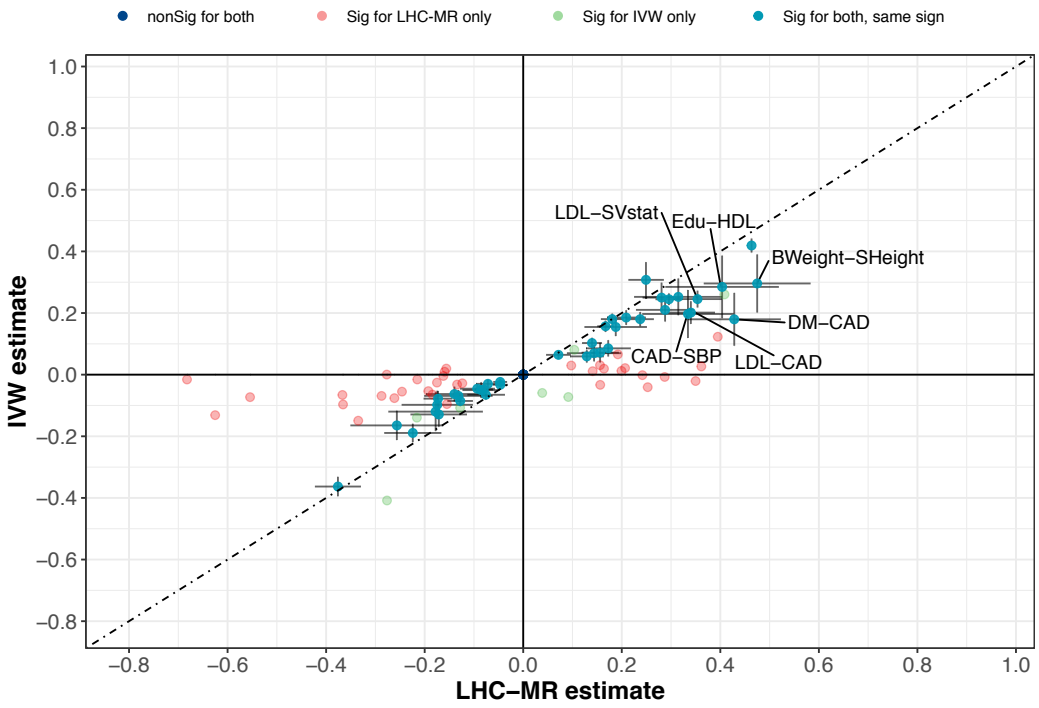


Figure 5: A scatter plot of the causal effect estimates between LHC-MR and IVW. To improve visibility, non-significant estimates by both methods are placed at the origin, while significant estimates by both methods appear on the diagonal with 95% CI error bars.

LHC-MR identified a confounder for 16 trait pairs out of the possible 78. In order to support

these findings, we used EpiGraphDB[24, 25] to systematically identify those potential confounders. EpiGraphDB could identify reliable confounders for ten out of the 16 trait pairs. Notably, for the birth weight - diabetes pair, the average epigraph confounder-effect ratio (r_3/r_1) clearly agreed in sign with our t_y/t_x ratio, indicating that the characteristics of the confounder(s) evidenced by LHC-MR agree with those found in an exhaustive confounder search, and are mainly obesity-related traits (Figure S18 panel *a*). Six other trait pairs showed mixed signs of different confounders, indicating the possibility of having heterogeneous confounders (Figure S18 panels *b-e*). Finally, three trait pairs showed a disagreement between our estimated confounder effect ratio and the bulk of those found by epighraphDB as seen in Supplementary Figure S18 panels *f-j*. However, at least one of the top ten potential confounders showed effects that are in agreement with our ratio for each of these pairs. Note that since the reported causal effects of the confounders on X and Y reported in EpiGraphDB are not necessarily on the same scale, we do not expect the magnitudes to agree.

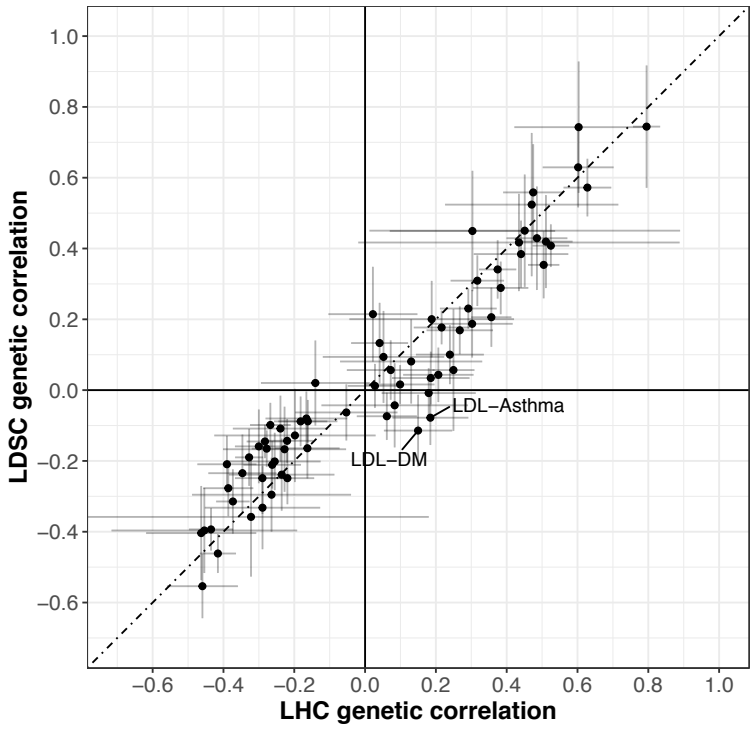


Figure 6: Scatter plot comparing the genetic correlation for each trait obtained from LDSC against the value calculated using parameter estimates from the LHC-MR model. A 95% CI is shown for each point. Values from both methods are reported in Supplementary Table S9.

As described in the methods (Eq. 3), genetic correlation can be computed from our estimated model parameters. To verify that the fitted LHC-MR model leads to a genetic correlation similar to the one obtained from LD score regression[50] (LDSC), we compared whether the two approaches produce similar genetic correlation estimates. We did this by taking the estimated parameters obtained from the 200 block jackknife to estimate the genetic correlations between traits (and their standard errors), and plotted them against LD score regression values as seen in Figure 6. As expected, we observe an overall good agreement between the estimates of the two methods, with only six trait pairs differing in sign. Of these six, only 2 were nominally significantly different between the two methods (LDL→Asthma and LDL→DM). Further decomposition of the genetic covariance into heritable confounder-led or causal effect-led revealed

that most of the genetic covariance between traits can be attributed to bi-directional causal effects. A reason for this could be that confounders would need to have very strong effects to substantially contribute to the genetic correlation ($\approx t_x \cdot t_y$) compared to the bi-directional causal effects ($\approx \alpha^2_{x\to y} \cdot h^2_x + \alpha^2_{y\to x} \cdot h^2_y$).

As for the comparison of LHC-MR against CAUSE for real trait pairs, we ran CAUSE on all 156 trait pairs (bi-directional), and extracted the parameter estimates that corresponded to the methods winning model. The p-value threshold was corrected for multiple testing and was equivalent to 0.05/156. Based on that threshold, the p-value that compared between the causal and the sharing model of CAUSE was used to choose one of the two. Then the parameters estimated from the winning model, γ (only for causal model), η and q, were compared to their counterparts in LHC-MR. A visual comparison of LHC-MR's causal estimates and those of CAUSE can be seen in Figure S19.

Whenever the causal effect estimates were significant both for CAUSE and LHC-MR (30 causal relationships), they always agreed in sign (Table S8) with a high Pearson correlation of 0.592. Calculating the correlation for their estimates regardless of significance yielded a smaller value of 0.377. When compared to the causal effect estimate from IVW, LHC-MR was strongly correlated (0.585), whereas CAUSE had a slightly weaker correlation (0.471) using all estimates.

Similarly, the significant confounder effect ratio of LHC-MR (t_y/t_x) can be compared to the significant confounder effect estimate of CAUSE (η) when a sharing model is chosen. These 12 confounding quantities by CAUSE and LHC-MR disagreed in sign for all but one trait pair (Height→MI), with a Pearson correlation compatible with zero (-0.357 (95% CI [-0.77, 0.27])) .

4 Discussion

We have developed a structural equation (mixed effect) model to account for a latent heritable confounder (U) of an exposure (X) - outcome (Y) relationship in order to estimate bi-directional causal effects between the two traits (X and Y). The method, termed LHC-MR, fits this model to association summary statistics of genome-wide genetic markers to estimate various global characteristics of these traits, including bi-directional causal effects, confounder effects, direct heritabilities, polygenicities, and population stratification.

We first demonstrated through simulations that in most scenarios, the method produces causal effect estimates with substantially less bias and variance (in larger sample sizes) than other MR tools. The direction and magnitude of the bias of classical MR approaches varied across scenarios and sample sizes. This bias was mainly influenced by two often opposite forces: downward bias resulting from winner's curse and weak instruments, and upward bias due to a positive confounder of the $X - Y$ relationship, evident in larger sample sizes. In the scenario lacking a confounder (thus respecting all MR assumptions), MR methods were distinctly underestimating the causal effect, except for LHC-MR and to a better extent MR-RAPS. However, under standard settings with an added small heritable confounder and no reverse causality present, all classical MR methods still slightly underestimated the causal effect in smaller sample sizes, except for the MR-RAPS estimate which was now overestimated. For the same standard setting scenario but in a larger sample size where confounder effects were more detectable, IVW had an estimation that was close to the true causal value chosen ($\alpha_{x\to y} = 0.3$) due to the opposite biases cancelling out. However, when the causal effect was set to be smaller ($\alpha_{x\to y} = 0.1$), the estimates of IVW became biased. More substantial violations of classical MR assumptions, such as the presence of negative-effect confounder or a negative reverse causal effect, led to more substantial biases that impacted all methods (including MR-RAPS) except LHC-MR.
Interestingly, in smaller sample sizes, standard MR methods showed a slight decreasing trend in the variance of the causal effect estimate as the kurtosis of the underlying effect size distribution went up from 2 to 10. On the other hand, LHC-MR did not show a similar trend with growing kurtosis, and estimated the causal effect with a smaller bias. As confounder causal effects (q_x, q_y) increased, classical MR methods (except weighted ones) were prone to produce overestimated causal effects with at least twice the bias than that of LHC-MR, especially in larger sample sizes where the confounder-associated SNPs make it to the set of GW-significant instruments for all methods. Furthermore, mode-based estimators were robust to the presence of two concordant confounders, yet their bias was still 10-fold higher than LHC-MR's, and they did not perform as well in the presence of discordant confounders. In summary, LHC-MR was robust to a wide range of violations of the classical MR assumptions and was less impacted than standard MR methods. Thus it outperformed all MR methods in virtually all tested scenarios, many of which violated even its own modelling assumptions.

We then applied our method to summary statistics of 13 complex traits from large studies, including the UK Biobank. We observed a general trend in our results that (in agreement with epidemiological studies) higher BMI and LDL are risk factors for most diseases such as diabetes and CAD. We also note the protective effect HDL has on these same diseases. Moreover, we observe many disease traits increasing the intake of lipid-lowering medication (simvastatin), reflecting the recommendation/treatment of medical personnel following the diagnosis.

LHC-MR can have discordant results compared to other MR methods for many possible reasons. The positive causal effect of smoking on MI, diabetes on asthma, the protective impact of higher birth weight on asthma, or higher education on smoking intensity, all of which were missed by standard MR could reflect the increased power of LHC-MR with its use of full-genome SNPs as opposed to genome-wide significant SNPs of classical MR approaches. Estimates from classical

MR methods could also be impacted by sample overlap between the exposure and outcome datasets, whereas LHC-MR takes this into account. However, when using large sample sizes, the bias due to sample overlap is expected to be very small, and therefore not sufficient to explain any discrepancy in the results[51]. Another possible reason for the discrepancy between our findings and those of standard MR methods is the presence of a significant heritable confounder found by LHC-MR with opposite effect to the estimated causal effect between the pair. These two opposite forces lead to association summary statistics that may be compatible with reduced (or even null) causal effect when the confounder is ignored. Possible examples of this scenario can be observed for when (parental) traits, e.g. diabetes and CAD, act on birth weight. These pairs have a confounder of opposite effects, possibly related to (parental) obesity. Similarly, standard MR methods show little evidence for a causal effect of SBP on height, while our LHC-MR estimate is -0.37 ($P = 4.81 \times 10^{-8}$) which most probably reflects parental (maternal) effects as seen in previous studies[52, 53]. The protective effect of HDL on SBP is another example where a confounder of opposite sign to that of the causal effect allows it to be uniquely found by LHC-MR. LHC-MR assumes no genetic correlation between the confounder and the direct effects on the exposure, which may be violated when the confounder is the same trait as the exposure, but in the parent. Such parental effects can mislead most MR methods [54], including ours, and hence we may observe biased results for traits such as BMI→education and HDL→birth weight.

Sixteen trait pairs showed a strong confounder effect, in the form of significant t_x and t_y estimates. These pairs were investigated for the presence of confounders using EpiGraphDB, and 10 of them returned possible confounders. The bulk of such pairs returned confounders with both agreeing and disagreeing effect directions on X and Y, making it difficult to pinpoint a group of concordant and dominant confounders. However, for the birth weight-DM pair, where LHC-MR identifies a negative reverse causal effect and a confounder with effects $t_x = 0.10 (P = 6.77 \times 10^{-8})$ and $t_y = 0.15 (P = 3.13 \times 10^{-7})$ on birth weight and DM respectively, EpigraphDB confirmed several confounders related to body fat distribution and weight that matched in sign with our estimated confounder effect (Figure S18 panel *a*). Note that EpiGraphDB causal estimates are not necessarily on the scale of SD outcome difference upon 1 SD exposure change scale, hence they are not directly comparable with the t_y/t_x ratio, but are rather indicative of the sign of the causal effect ratio of the confounder. Furthermore, if EpigraphDB does not find a causal relationship between the trait pair in either directions, then it does not return any possible confounders of the two, a reason why only 10 out of 16 confounder-associated trait pairs returned any hits.

Lastly, our comparison of the genetic correlations calculated from our estimated parameters against those calculated from LD score regression showed good concordance, confirming that the detailed genetic architecture proposed by our model is compatible with the observed genetic covariance. The major difference between the genetic correlation obtained by LD score regression *vs* LHC-MR is that our model approximates all existing confounders by a single latent variable, which may be inaccurate when multiple ones exist with highly variable t_y/t_x ratios. Furthermore, LHC-MR decomposed the observed genetic correlation into confounder and bi-directional causality driven components, revealing that most genetic correlations are primarily driven by bi-directional causal effects. Note that we have much higher statistical power to detect situations when the confounder effects are of opposite sign compared to the causal effects, because opposing genetic components are more distinct.

To our knowledge only two recent papers use similar models and genome-wide summary statistics. The LCV approach[55] is a special case of our model, where the causal effects are not included in the model, but they estimate the confounder effect mixed with the causal effect to

estimate a quantity of genetic causality proportion (GCP). In agreement with others[10,56], we would not interpret non-zero GCP as evidence for causal effect. Moreover, in other simulation settings, LCV has shown very low power to detect causal effects (by rejecting GCP=0) (Fig S15 in Howey et al.[57]). Another very recent approach, CAUSE[10], proposes a structural equation mixed effect model similar to ours. However, there are several differences between LHC-MR and CAUSE: (a) we allow for bi-directional causal effects and model them simultaneously, while CAUSE is fitted twice for each direction of causal effect; (b) they first use an adaptive shrinkage method to integrate out the multivariable SNP effects and then go on to estimate other model parameters, while we fit all parameters at once; (c) CAUSE estimates the correlation parameter empirically; (d) we assume that direct effects come from a two-component Gaussian mixture, while they allow for larger number of components; (e) their likelihood function does not explicitly model the shift between univariate vs multivariate effects (i.e. the LD); (f) CAUSE adds a prior distribution for the causal/confounder effects and the proportion π_u, while LHC-MR does not; (g) to calculate the significance of the causal effect they estimate the difference in the expected log point-wise posterior density and its variance through importance sampling, whereas we use a simple block jackknife method. Because of point (a), the CAUSE model can be viewed as a special case of ours when there is no reverse causal effect. We have the advantage of fitting all parameters simultaneously, while they only approximate this procedure. Although they allow for more than a two-component Gaussian mixture, for most traits with realistic sample sizes we do not have enough power to distinguish whether two or more components fit the data better. Therefore, we believe that a two component Gaussian is a reasonable simplification. Due to the more complicated approach described in points (e-g), CAUSE is computationally more intense than LHC-MR, taking up to 1.25 CPU-hours in contrast to our 2.5 CPU-minute run time for a single starting point optimisation (which is massively parallelisable).

When we compared the performance of CAUSE and LHC-MR, we found that for large sample sizes both LHC-MR and CAUSE performed well not only when applied to data simulated by their own model, but also by the model of the other method. For smaller sample sizes, both methods performed poorly when applied to data generated by the other model. However, LHC-MR was less biased when applied to data generated by its own model than CAUSE was on data simulated based on its own model, where it provided rather conservative estimates. This is somewhat expected, since the primary aim of CAUSE is model selection and it is less geared towards parameter estimation, especially for settings where both sharing and causal effects are present (leading to very broad estimates). Also, CAUSE parameter estimates have shown to be somewhat sensitive to the choice of the prior.

Finally, when applying both LHC-MR and CAUSE to 156 complex trait pairs, we observed that the causal effects are reasonably well correlated (0.38 for all estimates, 0.59 for significant estimates) and agree in sign for trait pairs deemed significantly causal by either or both methods. In addition, LHC-MR causal estimates were more similar to those of IVW than the estimates provided by CAUSE. Surprisingly, when a confounding factor was identified by both methods, the confounder effects (LHC-MR t_y/t_x ratio and CAUSE η parameter) were uncorrelated. There are two possible explanations for this: (i) CAUSE may confuse/merge the confounder with the reverse causal effect, since it does not explicitly model the latter one. (ii) The two models assume different marginal effect size distributions, hence when multiple heterogeneous confounders exist, one method may detect one of the confounders, while the other method picks up the other confounder, depending on which has more similar genetic architecture to the assumed one.

Our approach has its own limitations, which we list below. Like any MR method, LHC-MR provides biased causal effect estimates if the input summary statistics are flawed (e.g. not corrected for complex population stratification, parental/dynasty effects). As mentioned in the Methods

section, our model is strictly-speaking unidentifiable and two distinct sets of parameters fit the data equally well, if the alternate set of parameters fall within the parameter ranges. As opposed to classical MR methods that give a single (biased) causal effect estimate, ours can detect and calculate the competing model. Due to biological considerations, from these competing models, we chose the one which yielded larger direct heritability than confounder-driven (indirect) heritability. Additional pointers to decide which parameter optimum we choose can be to pick the one with smaller magnitude of causal effects (large causal effects are unrealistic) or pick the one that includes causal effects that agree better with those of other MR methods.

LHC-MR is not an optimal solution for traits whose genetic architecture substantially deviates from a two-component Gaussian mixture of effect sizes. Also, for traits with low heritability ($< 2.5\%$), it is particularly important to compare the causal effect estimates to those from standard MR methods as results from LHC-MR may be less robust. In addition, trait pairs with multiple confounders with heterogeneous effect ratios can violate the single confounder assumption of the LHC model and can lead to biased causal effect estimates. Finally, LHC-MR, like other methods, is not immune to parental effects that are correlated with offspring effects. In such cases, the parental effect is grouped with the exposure (due to their strong genetic correlation) and not viewed as a confounder of the exposure-outcome relationship.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 16389. LD scores were calculated based on the UK10K data resource (EGAD00001000740, EGAD00001000741). Z.K. was funded by the Swiss National Science Foundation (31003A_169929, 310030_189147 and 32003B_173092). For computations we used the CHUV HPC cluster.
We would like to thank Jack Bowden, Valentin Rousson, Matthew Robinson, George Davey Smith, Thomas Richardson and Eleonora Porcu for their valuable feedback and comments on this manuscript.

Author contributions

Z.K. devised and directed the project. Z.K., N.M., and L.D. contributed to the mathematical derivations, design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Data availability

The origin of the summary statistics data used is referenced in Table S1. UK Biobank summary statistics can be downloaded from `http://www.nealelab.is/uk-biobank`. CAD GWAS summary statistics data from `http://www.cardiogramplusc4d.org/data-downloads/`

Code availability

The source code for this work can be found on `https://github.com/LizaDarrous/LHC-MR_v2/`

References

[1] Fewell, Z., Davey Smith, G., and Sterne, J. A. C. (2007). The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. American journal of epidemiology *166*, 646–655.

[2] Pingault, J.-B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F., and Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. Nature reviews. Genetics *19*, 566–580.

[3] Barter, P. J., Caulfield, M., Eriksson, M., Grundy, S. M., Kastelein, J. J. P., Komajda, M., Lopez-Sendon, J., Mosca, L., Tardif, J.-C., Waters, D. D., et al. (2007). Effects of torcetrapib in patients at high risk for coronary events. The New England journal of medicine *357*, 2109–2122.

[4] Fordyce, C. B., Roe, M. T., Ahmad, T., Libby, P., Borer, J. S., Hiatt, W. R., Bristow, M. R., Packer, M., Wasserman, S. M., Braunstein, N., et al. (2015). Cardiovascular drug development: is it dead or just hibernating? Journal of the American College of Cardiology *65*, 1567–1582.

[5] Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statistics in medicine *27*, 1133–1163.

[6] Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. Genetic epidemiology *37*, 658–665.

[7] Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. International journal of epidemiology *44*, 512–525.

[8] Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. Genetic epidemiology *40*, 304–314.

[9] Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. International journal of epidemiology *46*, 1985–1998.

[10] Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nature Genetics *52*, 740–747.

[11] Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. The Annals of Statistics *48*, 1742 – 1769.

[12] Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics *47*, 291–295.

[13] Jordan, D. M., Verbanck, M., and Do, R. (2019). The landscape of pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. bioRxiv.

[14] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nature genetics *47*, 1236–1241.

[15] Nadarajah, S. and Pogány, T. K. (2016). On the distribution of the product of correlated normal random variables. Comptes Rendus Mathematique *354*, 201–204.

[16] McNolty, F. (1973). Some probability density functions and their characteristic functions.

[17] Bateman, H. (1953). Volume i.

[18] Heideman, M. T., Johnson, D. H., and Burrus, C. S. (1985). Gauss and the history of the fast fourier transform. Archive for History of Exact Sciences *34*, 265–277.

[19] Neale Lab (2018). UK BioBank. `http://www.nealelab.is/uk-biobank/`.

[20] Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., et al. (2015). The uk10k project identifies rare variants in health and disease. Nature *526*, 82–90.

[21] Rüeger, S., McDaid, A., and Kutalik, Z. (2018). Evaluation and application of summary statistic imputation to discover new height-associated loci. PLoS genetics *14*, e1007371.

[22] 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

[23] R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.

[24] MRC IEU (2019). EpiGraphDB. `http://epigraphdb.org/`.

[25] Liu, Y., Elsworth, B., Erola, P., Haberland, V., Hemani, G., Lyon, M., Zheng, J., and Gaunt, T. R. (2020). Epigraphdb: A database and data mining platform for health data science. bioRxiv.

[26] Clogg, C. C., Petkova, E., and Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. American Journal of Sociology *100*, 1261–1293.

[27] Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. eLife *7*, e34408.

[28] Allen, M., Poggiali, D., Whitaker, K., Marshall, T., and Kievit, R. (2019). Raincloud plots: a multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. Wellcome Open Research *4*.

[29] Laclaustra, M., Lopez-Garcia, E., Civeira, F., Garcia-Esquinas, E., Graciani, A., Guallar-Castillon, P., Banegas, J. R., and Rodriguez-Artalejo, F. (2018). Ldl cholesterol rises with bmi only in lean individuals: Cross-sectional u.s. and spanish representative data. Diabetes Care *41*, 2195–2201.

[30] Drøyvold, W. B., Midthjell, K., Nilsen, T. I. L., and Holmen, J. (2005). Change in body mass index and its impact on blood pressure: a prospective population study. International Journal of Obesity *29*, 650–655.

[31] Lee, M.-R., Lim, Y.-H., and Hong, Y.-C. (2018). Causal association of body mass index with hypertension using a mendelian randomization design. Medicine (Baltimore) *97*, e11252.

[32] Corbin, L. J., Richmond, R. C., Wade, K. H., Burgess, S., Bowden, J., Smith, G. D., and Timpson, N. J. (2016). Bmi as a modifiable risk factor for type 2 diabetes: Refining and understanding causal estimates using mendelian randomization. Diabetes *65*, 3002–3007.

[33] Narayan, K., Boyle, J. P., Thompson, T. J., Gregg, E. W., and Williamson, D. F. (2007). Effect of bmi on lifetime risk for diabetes in the u.s. Diabetes Care *30*, 1562–1566.

[34] Yusuf, S., Hawken, S., Ounpuu, S., Bautista, L., Franzosi, M. G., Commerford, P., Lang, C. C., Rumboldt, Z., Onen, C. L., Lisheng, L., et al. (2005). Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study. Lancet *366*, 1640–1649.

[35] Riaz, H., Khan, M. S., Siddiqi, T. J., Usman, M. S., Shah, N., Goyal, A., Khan, S. S., Mookadam, F., Krasuski, R. A., and Ahmed, H. (2018). Association Between Obesity and Cardiovascular Outcomes: A Systematic Review and Meta-analysis of Mendelian Randomization Studies. JAMA Network Open *1*, e183788–e183788.

[36] Sun, D., Zhou, T., Heianza, Y., Li, X., Fan, M., Fonseca, V. A., and Qi, L. (2019). Type 2 diabetes and hypertension. Circulation research *124*, 930–937.

[37] Tomeo, C. A., Field, A. E., Berkey, C. S., Colditz, G. A., and Frazier, A. L. (1999). Weight concerns, weight control behaviors, and smoking initiation.

[38] Cawley, J., Markowitz, S., and Tauras, J. (2004). Lighting up and slimming down: the effects of body weight and cigarette prices on adolescent smoking initiation.

[39] Cao, M. and Cui, B. (2020). Association of educational attainment with adiposity, type 2 diabetes, and coronary artery diseases: A mendelian randomization study. Frontiers in Public Health *8*, 112.

[40] Loucks, E. B., Buka, S. L., Rogers, M. L., Liu, T., Kawachi, I., Kubzansky, L. D., Martin, L. T., and Gilman, S. E. (2012). Education and coronary heart disease risk associations may be affected by early-life common prior causes: a propensity matching analysis. Ann Epidemiol *22*, 221–232.

[41] Gage, S. H., Bowden, J., Davey Smith, G., and Munafò, M. R. (2018). Investigating causality in associations between education and smoking: a two-sample Mendelian randomization study. International Journal of Epidemiology *47*, 1131–1140.

[42] Sanderson, E., Davey Smith, G., Bowden, J., and Munafò, M. R. (2019). Mendelian randomisation analysis of the effect of educational attainment and cognitive ability on smoking behaviour. Nature Communications *10*, 2949.

[43] Marouli, E., Del Greco, M. F., Astley, C. M., Yang, J., Ahmad, S., Berndt, S. I., Caulfield, M. J., Evangelou, E., McKnight, B., Medina-Gomez, C., et al. (2019). Mendelian randomisation analyses find pulmonary factors mediate the effect of height on coronary artery disease. Communications biology *2*, 119.

[44] Tan, L. E., Llano, A., Aman, A., Dominiczak, A. F., and Padmanabhan, S. (2018). A18709 mendelian randomization study of causal relationship of height on blood pressure and arterial stiffness. Journal of Hypertension *36*.

[45] Laaksonen, D. E., Niskanen, L., Nyyssönen, K., Lakka, T. A., Laukkanen, J. A., and Salonen, J. T. (2008). Dyslipidaemia as a predictor of hypertension in middle-aged men. Eur Heart J *29*, 2561–2568.

[46] Davies, N. M., Howe, L. J., Brumpton, B., Havdahl, A., Evans, D. M., and Davey Smith, G. (2019). Within family Mendelian randomization studies. Human Molecular Genetics *28*, R170–R179.

[47] Benson, R., von Hippel, P. T., and Lynch, J. L. (2018). Does more education cause lower bmi, or do lower-bmi individuals become more educated? evidence from the national longitudinal survey of youth 1979. Soc Sci Med *211*, 370–377.

[48] Witter, F. R. and Luke, B. (1991). The effect of maternal height on birth weight and birth length. Early Human Development *25*, 181–186.

[49] Tyrrell, J., Jones, S. E., Beaumont, R., Astley, C. M., Lovell, R., Yaghootkar, H., Tuke, M., Ruth, K. S., Freathy, R. M., Hirschhorn, J. N., et al. (2016). Height, body mass index, and socioeconomic status: mendelian randomisation study in uk biobank. BMJ *352*.

[50] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nature Genetics *47*, 1236–1241.

[51] Mounier, N. and Kutalik, Z. (2021). Correction for sample overlap, winner's curse and weak instrument bias in two-sample mendelian randomization. bioRxiv.

[52] Thomas, D., Strauss, J., and Henriques, M.-H. (1991). How does mother's education affect child height? The Journal of Human Resources *26*, 183–211.

[53] Warrington, N. M., Beaumont, R. N., Horikoshi, M., Day, F. R., Helgeland, Ø., Laurin, C., Bacelis, J., Peng, S., Hao, K., Feenstra, B., et al. (2019). Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. Nat Genet *51*, 804–814.

[54] Brumpton, B., Sanderson, E., Hartwig, F. P., Harrison, S., Vie, G. Å., Cho, Y., Howe, L. D., Hughes, A., Boomsma, D. I., Havdahl, A., et al. (2019). Within-family studies for mendelian randomization: avoiding dynastic, assortative mating, and population stratification biases. bioRxiv.

[55] O'Connor, L. J. and Price, A. L. (2018). Distinguishing genetic correlation from causation across 52 diseases and complex traits. Nature genetics *50*, 1728–1734.

[56] Brown, B. C. and Knowles, D. A. (2020). Phenome-scale causal network discovery with bidirectional mediated mendelian randomization. bioRxiv.

[57] Howey, R., Shin, S.-Y., Relton, C., Smith, G. D., and Cordell, H. J. (2019). Bayesian network analysis incorporating genetic anchors complements conventional mendelian randomization approaches for exploratory analysis of causal relationships in complex data. bioRxiv.

Supplementary Materials & Methods

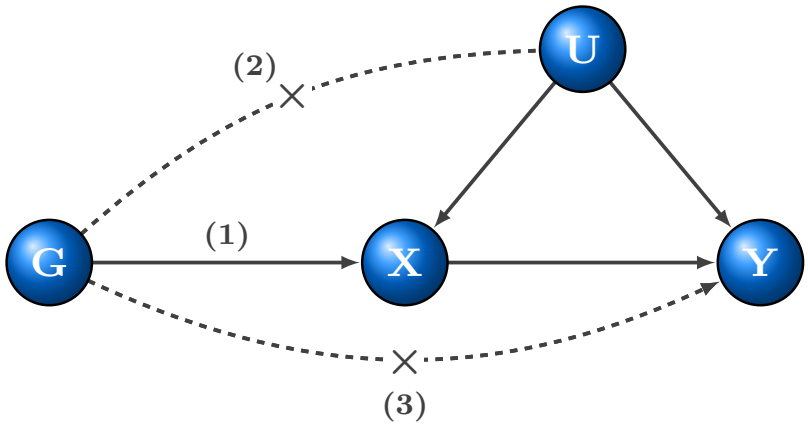


Figure S1: Basic assumptions of Mendelian randomisation. (1) Relevance – genetic data, denoted by G, is robustly associated with the exposure. (2) Exchangeability – G is not associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – G is independent of the outcome conditional on the exposure and all confounders of the exposure-outcome relationship (i.e. the only path between the instrument and the outcome is via the exposure).

1.1 Likelihood function identifiability

The likelihood function is symmetric around U, but for simplicity we will consider the general case where the variables of U and X are flipped, although the same can be said for the variables of U and Y. The likelihood function is partially identifiable such that there exists for any given model parameters, another model with different parameters but with the exact same likelihood function.

Proof: given that the SNPs effects between trait X and the confounder U are flipped, the new parameters follow the following structure:

$$\begin{aligned}
h'_x &= t_x + t_y \cdot \alpha_{y\to x} \\
h'_y &= h_y \\
\alpha'_{y\to x} &= \alpha_{y\to x} \\
\alpha'_{x\to y} &= \frac{q_x \cdot \alpha_{x\to y} + q_y}{q_x + q_y \cdot \alpha_{y\to x}} \\
&= \frac{q_x(\alpha_{x\to y} + \frac{q_y}{q_x})}{q_x(1 + \frac{q_y}{q_x} \cdot \alpha_{y\to x})} \\
&= \frac{\alpha_{x\to y} + \frac{q_y}{q_x}}{1 + \frac{q_y}{q_x} \cdot \alpha_{y\to x}}
\end{aligned}$$

through inverse transformation,

$$\alpha_{x\to y} = \frac{\alpha'_{x\to y} + \frac{q'_y}{q'_x}}{1 + \frac{q'_y}{q'_x} \cdot \alpha_{y\to x}}$$

Plugging in $\alpha'_{x\to y}$ in the above equation, and simplifying $\frac{ty}{tx}$ by w and $\frac{ty'}{tx'}$ by w' to get the confounding ratio:

$$\begin{aligned}
\alpha_{x\to y} &= \frac{\alpha'_{x\to y} + w'}{1 + w' \cdot \alpha_{y\to x}} \\
\alpha_{x\to y} + \alpha_{x\to y} \cdot w' \cdot \alpha_{y\to x} &= \alpha'_{x\to y} + w' \\
\alpha_{x\to y} - \alpha'_{x\to y} &= w' - \alpha_{x\to y} \cdot w' \cdot \alpha_{y\to x} \\
\alpha_{x\to y} - \alpha'_{x\to y} &= w'(1 - \alpha_{x\to y} \cdot \alpha_{y\to x}) \\
w' &= \frac{\alpha_{x\to y} - \alpha'_{x\to y}}{1 - \alpha_{x\to y} \cdot \alpha_{y\to x}}
\end{aligned}$$

inserting the complete form of $\alpha'_{x\to y}$,

$$\begin{aligned}
w' &= \frac{\alpha_{x\to y} - \frac{\alpha_{x\to y}+\frac{q_y}{q_x}}{1+\frac{q_y}{q_x}\cdot\alpha_{y\to x}}}{1-\alpha_{x\to y}\cdot\alpha_{y\to x}} \\
&= \frac{\alpha_{x\to y}(1+w\cdot\alpha_{y\to x}) - w - \alpha_{x\to y}}{(1-\alpha_{x\to y}\cdot\alpha_{y\to x})(1+w\cdot\alpha_{y\to x})} \\
&= \frac{\alpha_{x\to y}\cdot w\cdot\alpha_{y\to x} - w}{(1-\alpha_{x\to y}\cdot\alpha_{y\to x})(1+w\cdot\alpha_{y\to x})} \\
&= \frac{w(\alpha_{x\to y}\cdot\alpha_{y\to x} - 1)}{(1-\alpha_{x\to y}\cdot\alpha_{y\to x})(1+w\cdot\alpha_{y\to x})} \\
&= \frac{-w}{1+w\cdot\alpha_{y\to x}}
\end{aligned}$$

In order to obtain t'_y and t'_x, we use the equations of h'_x, $\alpha'_{x\to y}$ and by using the inverse transformation of $\alpha'_{y\to x} = \alpha_{y\to x}$, $\alpha_{x\to y}$ as well as w' as follows:

$$\begin{aligned}
t'_y &= \frac{-t'_x\cdot w}{1+w\cdot\alpha_{y\to x}} \\
h_x &= t'_x + t'_y\cdot\alpha_{y\to x} \\
&= t'_x + \frac{-t'_x\cdot w}{1+w\cdot\alpha_{y\to x}}\cdot\alpha_{y\to x} \\
&= \frac{t'_x + t'_x\cdot w\cdot\alpha_{y\to x} - t'_x\cdot w\cdot\alpha_{y\to x}}{1+w\cdot\alpha_{y\to x}} \\
&= \frac{t'_x}{1+w\cdot\alpha_{y\to x}} \\
t'_x &= h_x(1+w\cdot\alpha_{y\to x})
\end{aligned}$$

Replacing t'_x in h_x to get t'_y:

$$t'_y = h_x\cdot w$$

Under these two models with equal likelihood, there are three slopes obtained from the observed data: two are the correlation of effect sizes ($\alpha_{x\to y}$ and $1/\alpha_{y\to x}$), where one of them is greater than, and the other is within the parameter bounds. The third is the correlation of the confounder $\frac{\alpha_{x\to y}+\frac{q_y}{q_x}}{1+\frac{q_y}{q_x}\cdot\alpha_{y\to x}}$.

More often than not, only one slope is recovered within the boundaries of the parameters set for LHC-MR. However, given the now known re-parameterisation, the second (and if found, third) slope can be simply calculated if not found by the likelihood function minimisation. It is reasonable to assume that the direct heritability of each trait is larger than the indirect heritability, hence we report parameter sets where $h_x^2 > t_x^2$ or $h_y^2 > t_y^2$.

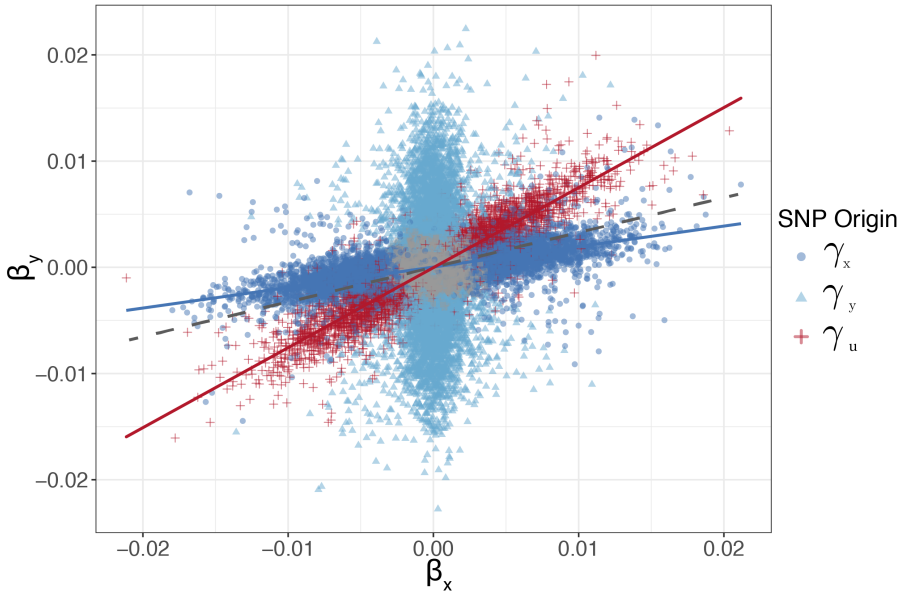


Figure S2: An illustration of a scatter plot showing simulated observed SNP effects on traits X and Y, coloured by the strongest effect between the three vectors $\gamma_x, \gamma_y, \gamma_u$. SNPs in grey are those with no effect on any of the traits. This illustration shows the distinct clusters that could arise in the presence of a confounder. The dark blue cluster of SNPs represents those that are not in violation of any of the MR assumption, and hence its slope reflects the true causal effect of X on Y, while the red cluster of SNPs are those associated with the confounder. The steeper slope of the red cluster of SNPs causes a typical regression line - shown in grey - that represents the causal effect (estimated using conventional MR methods) to be overestimated.

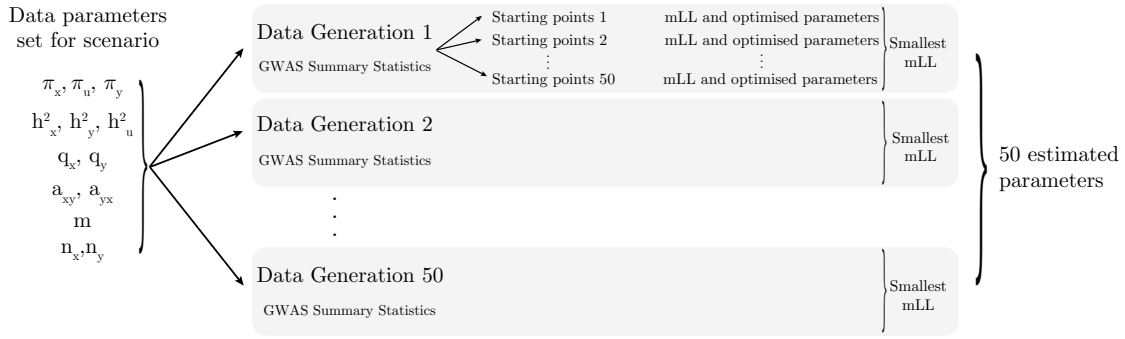


Figure S3: A schema showing the workflow of the simulation results. For a single set of parameter settings, 50 different data generations of GWAS summary statistics are created for trait X and Y. The summary statistics of a single data generation, as well as the sample size, SNP number and SNP-based LD structure are used in the likelihood optimisation function that is run with 100 different random starting points in order to explore the likelihood surface. A single maximum likelihood and its corresponding estimated parameters are selected to represent the estimates of that data generation. And this is repeated for the other generations. The results for several data generation are often represented in boxplots throughout the paper.

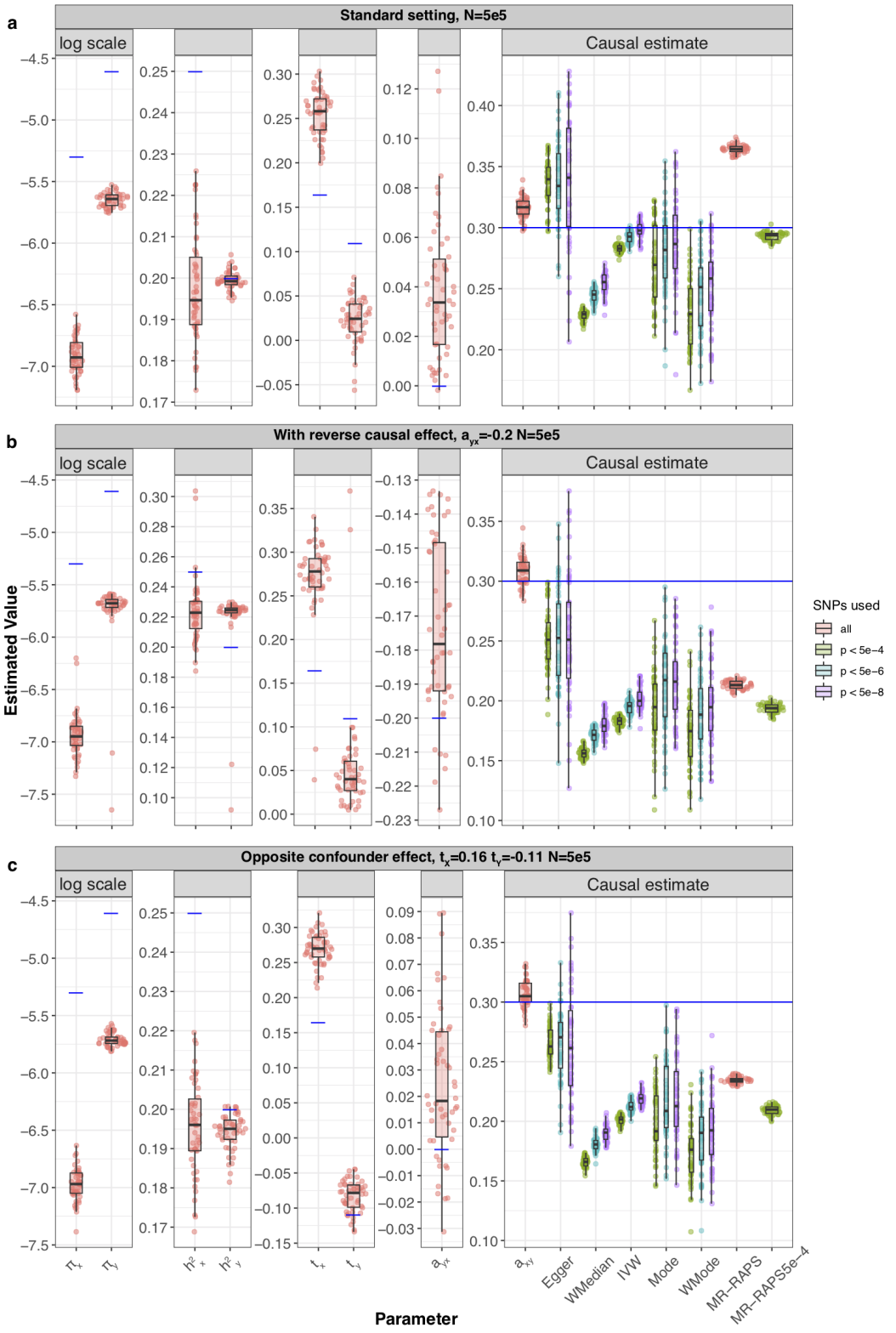


Figure S4: Simulation results under various scenarios. These Raincloud boxplots[28] represent the distribution of parameter estimates from 50 different data generations under various conditions. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** Estimation under standard settings ($\pi_x = 5 \times 10^{-3}, \pi_y = 1 \times 10^{-2}, \pi_u = 5 \times 10^{-2}, h^2_x = 0.25, h^2_y = 0.2, h^2_u = 0.3, t_x = 0.16, t_y = 0.11$). **b** Addition of a reverse causal effect $\alpha_{y \to x} = -0.2$. **c** Confounder with opposite causal effects on X and Y ($t_x = 0.16, t_y = -0.11$).

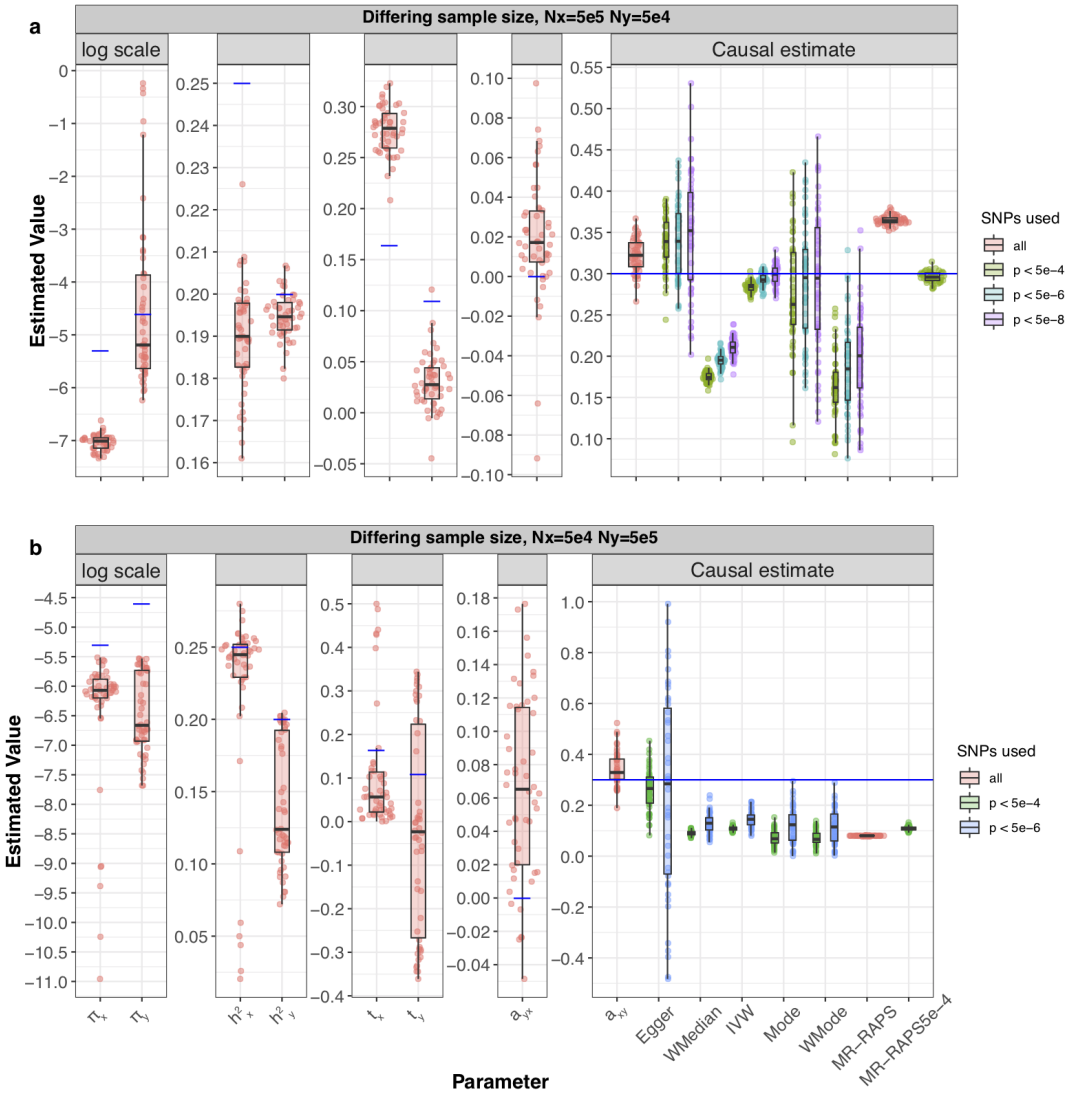


Figure S5: Simulation results showing varying sample sizes for the two exposure and outcome samples. Raincloud boxplots representing the distribution of parameter estimates from 50 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. In this figure, samples sizes for the two traits differ as such $n_x = 500{,}000$ and $n_y = 50{,}000$ for **a**, and $n_x = 50{,}000$ and $n_y = 500{,}000$ for **b**.

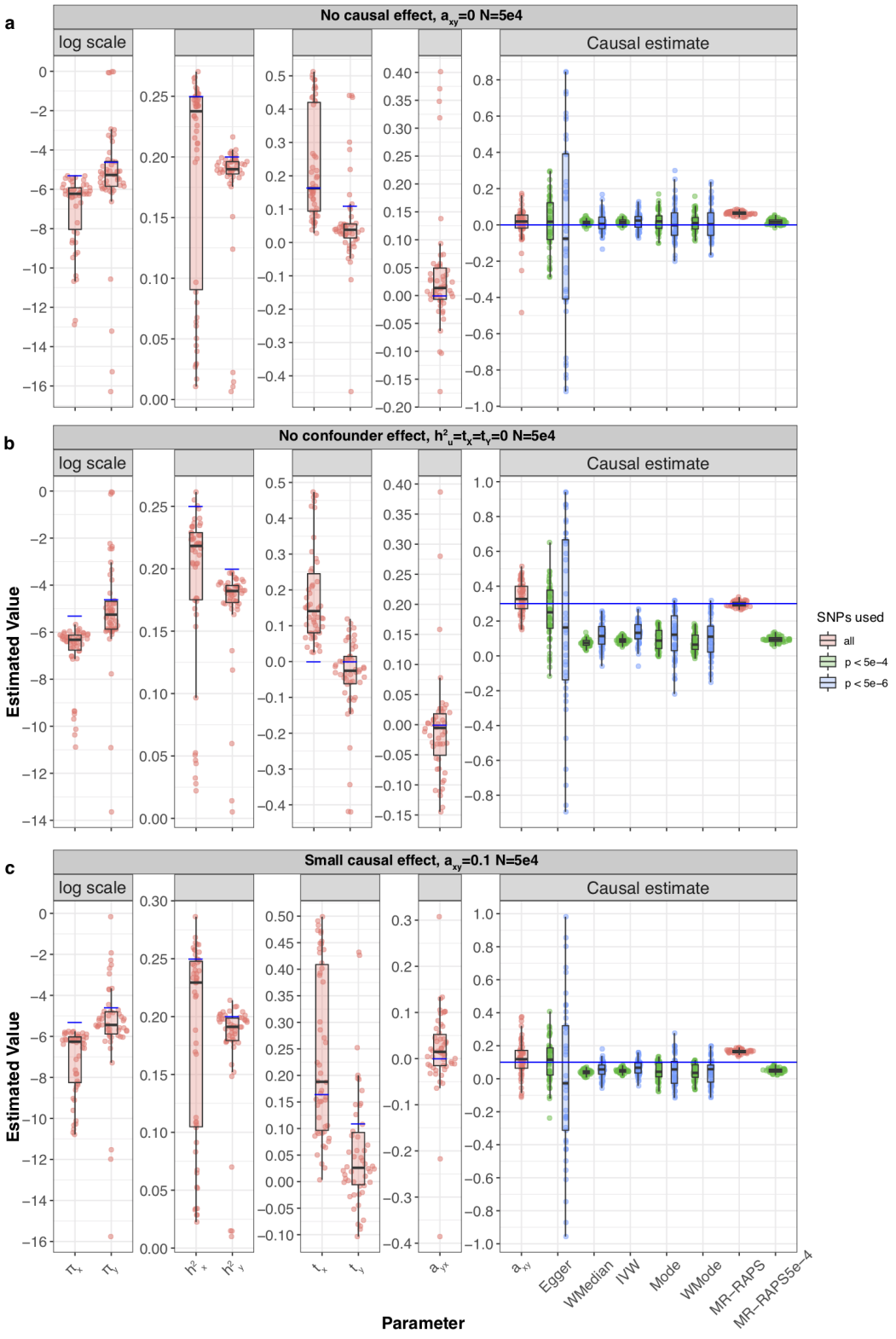


Figure S6: Simulation results under various scenarios. These Raincloud boxplots[28] represent the distribution of parameter estimates from 50 different data generations under various conditions. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** The data simulated had no causal effect in either direction. **b** The data simulated had no confounder effect with π_u, t_x, and $t_y = 0$. **c** This model had a small causal effect of $\alpha_{x \to y} = 0.1$.

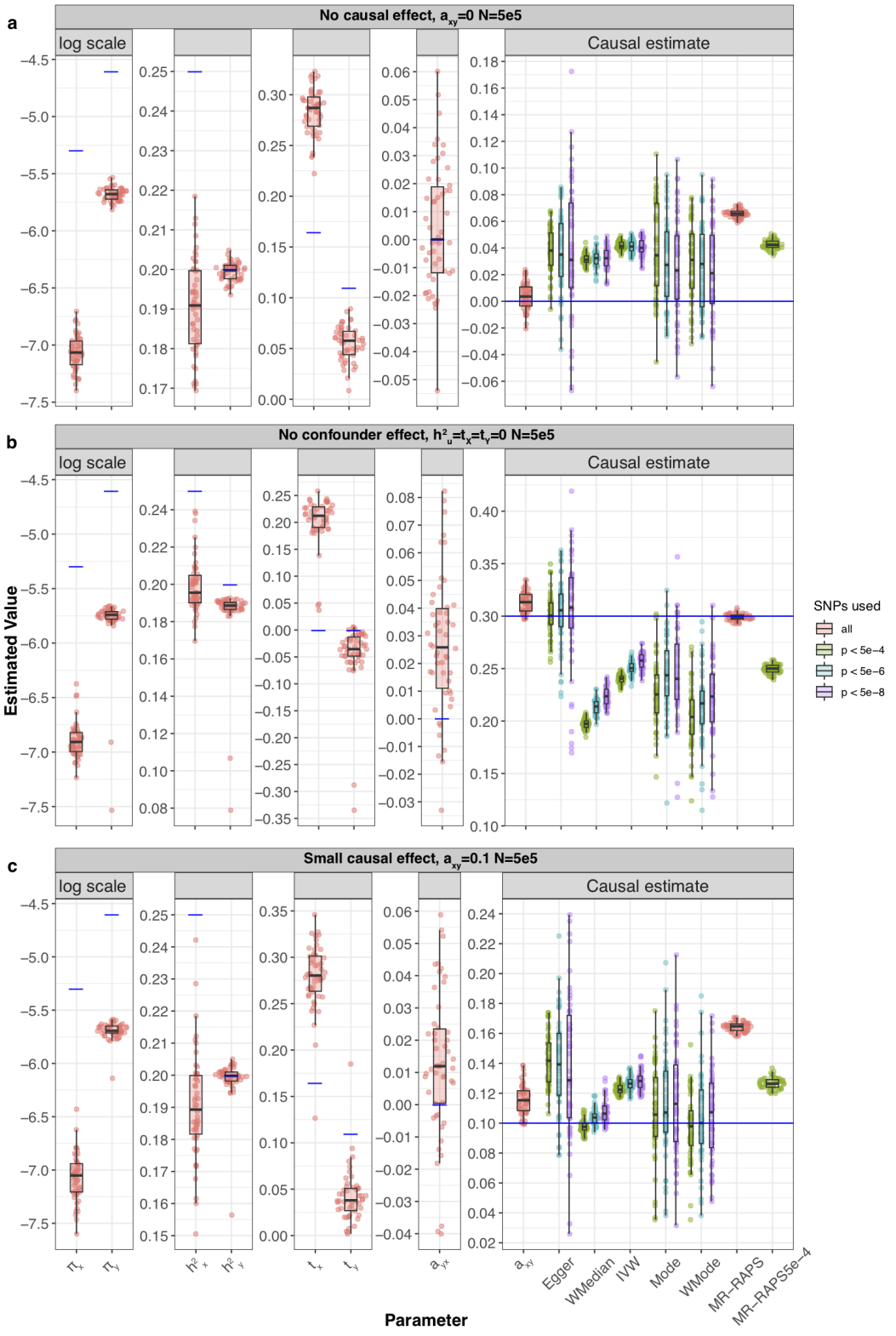


Figure S7: Simulation results under various scenarios. These Raincloud boxplots[28] represent the distribution of parameter estimates from 50 different data generations under various conditions. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** The data simulated had no causal effect in either direction. **b** The data simulated had no confounder effect with π_u, t_x, and $t_y = 0$. **c** This model had a small causal effect of $\alpha_{x \to y} = 0.1$.

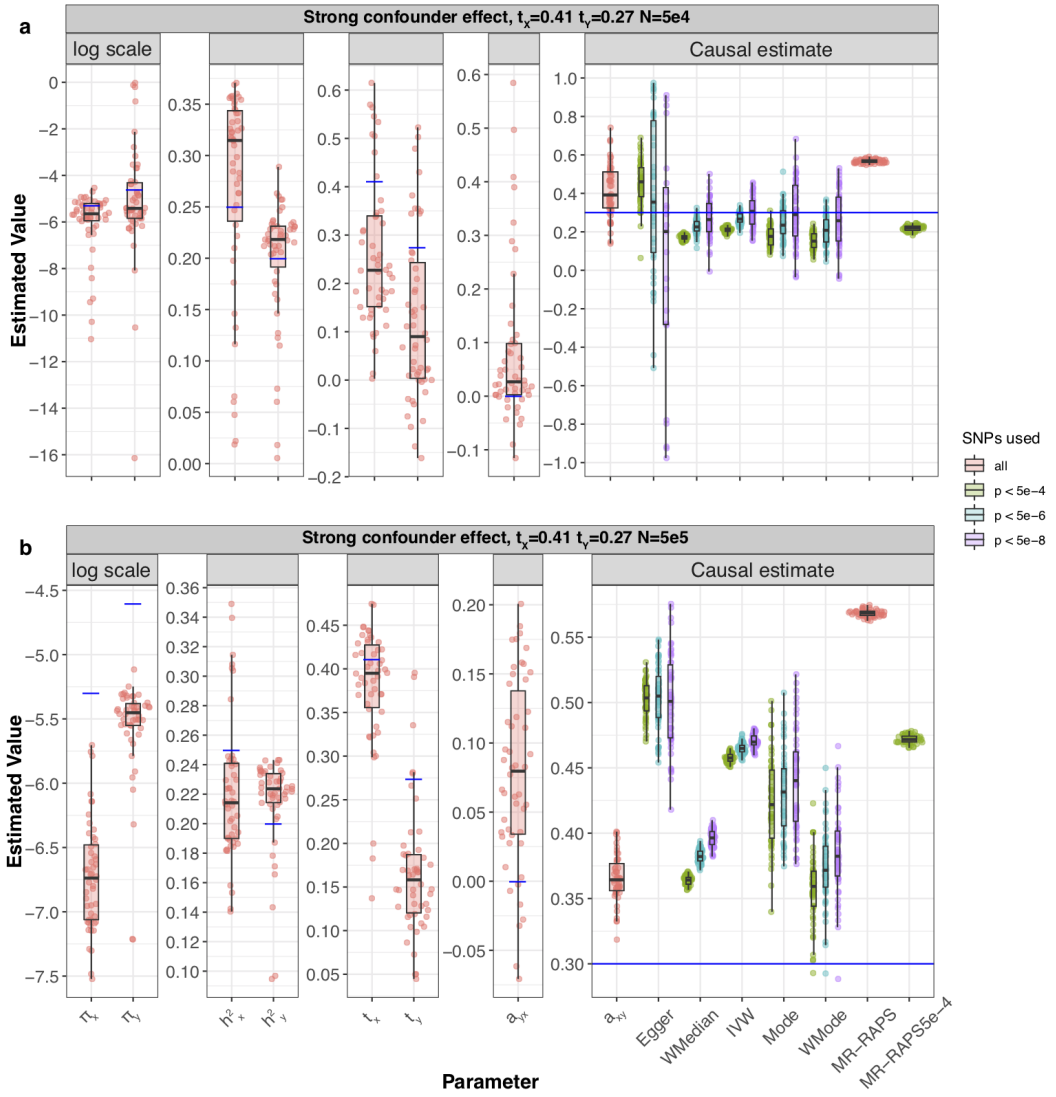


Figure S8: Simulation results under various scenarios. These Raincloud boxplots[28] represent the distribution of parameter estimates from 50 different data generations under various conditions. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** The data simulated shows the increased effect of U on X and Y through $t_x = 0.41, t_y = 0.27$ instead of the standard setting $t_x = 0.16, t_y = 0.11$. **b** This panel show the same thing but with a larger sample size of $n_x = n_y = 500,000$

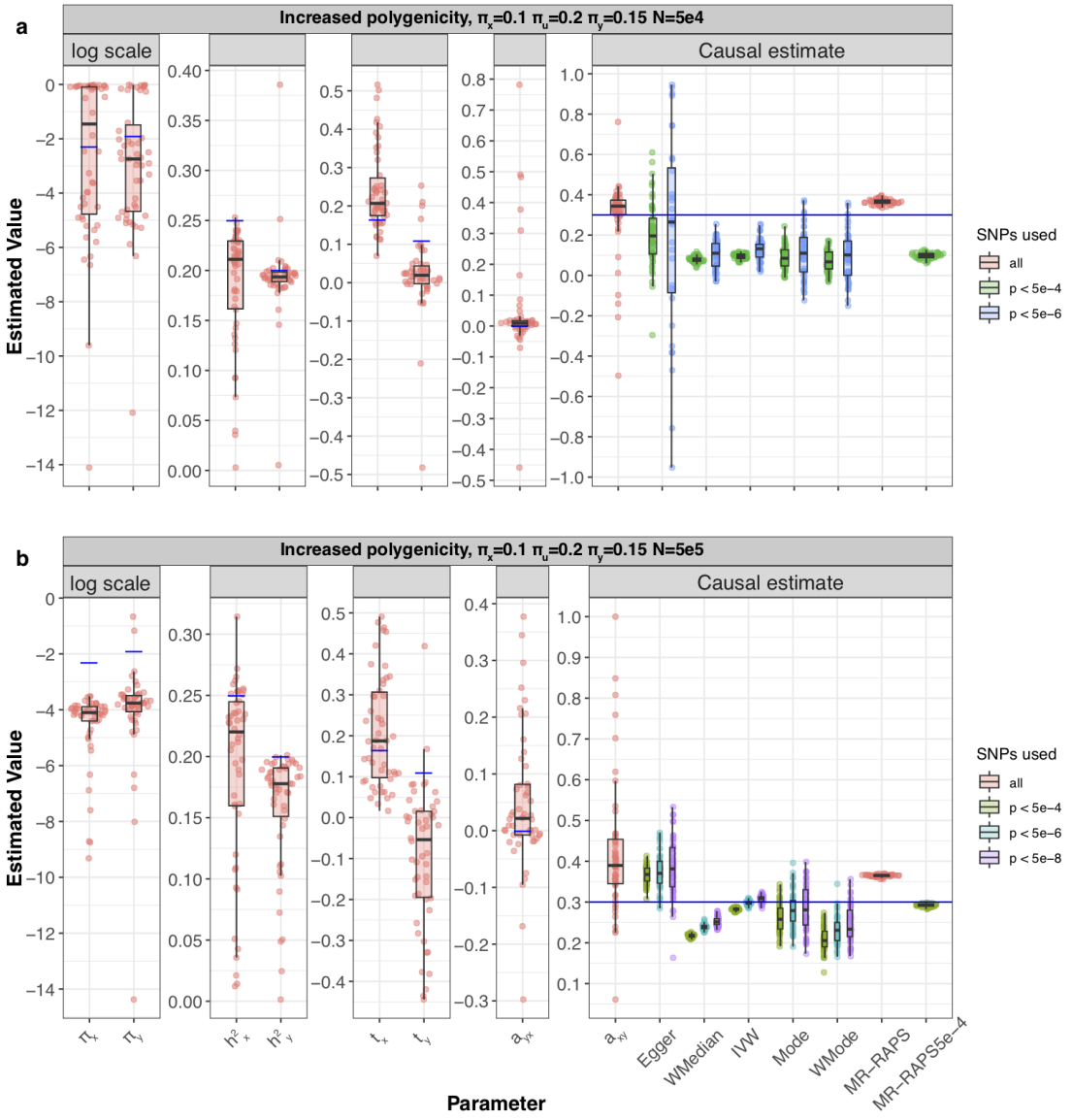


Figure S9: Simulation results where there is an increased polygenicity for all traits. Boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. The proportion of effective SNPs that make up the spike-and-slab distributions of the γ vectors in this setting is $10\%, 15\%, and 20\%$ for traits X, Y and U respectively. **a** Results for smaller sample size of $n_x = n_y = 50,000$. **b** Results for larger sample size of $n_x = n_y = 500,000$.

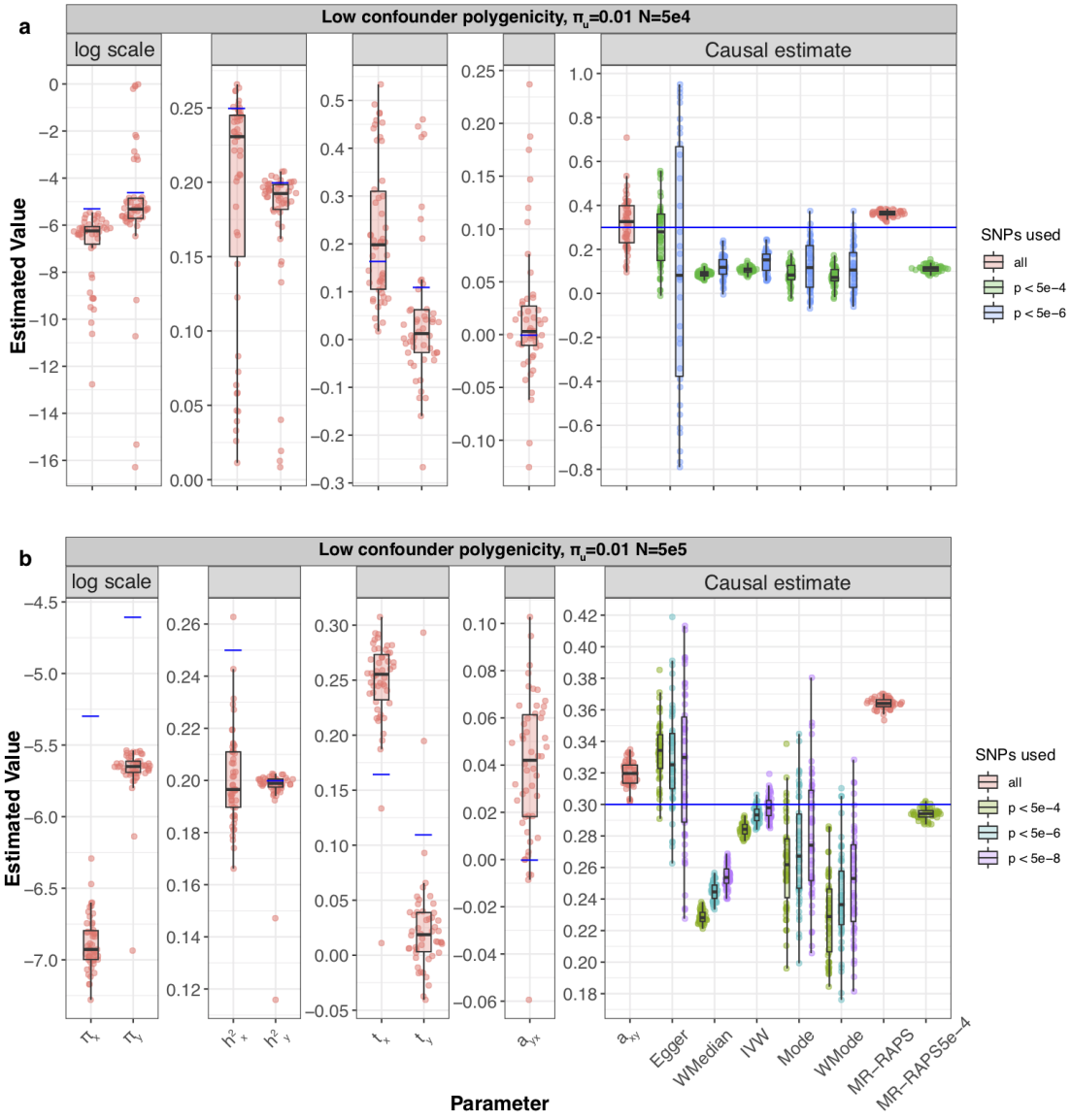


Figure S10: Simulation results where the polygenicity of the confounder is reduced. Boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. In this figure, the polygenicity for U is decreased in the form of lower $\pi_u = 0.01$. **a** Results for smaller sample size of $n_x = n_y = 50,000$. **b** Results for larger sample size of $n_x = n_y = 500,000$.

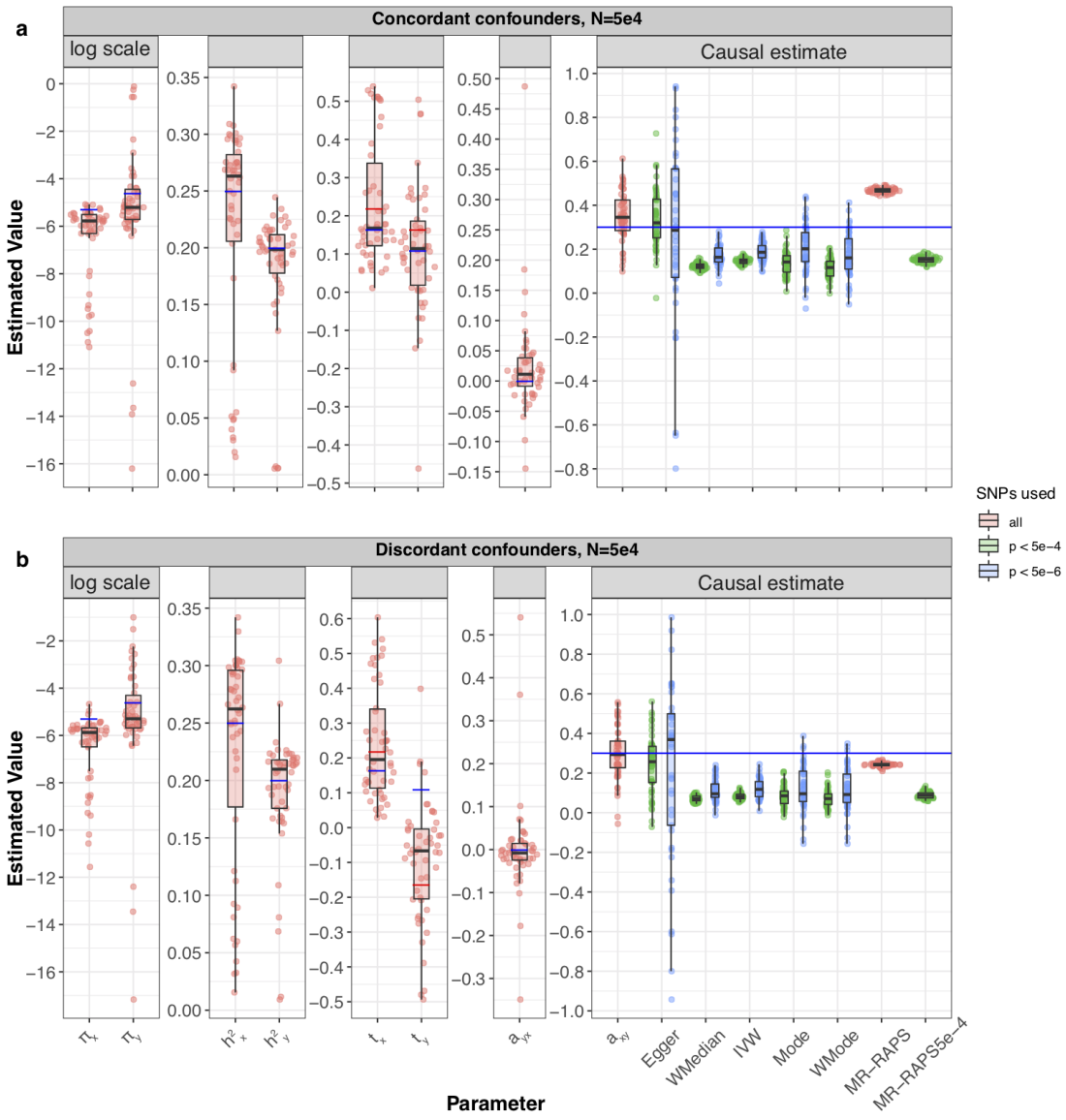


Figure S11: Simulation results where there are two underlying confounders, once with concordant and another with discordant effects on the exposure-outcome pair. Boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** The underlying data generations have two concordant heritable confounders U_1 and U_2 with positive effects on traits X and Y. **b** The data generations have two discordant heritable confounders with $t_x^{(1)} = 0.16, t_y^{(1)} = 0.11$ shown as blue dots and $t_x^{(2)} = 0.22, t_y^{(2)} = -0.16$ shown as red dots.

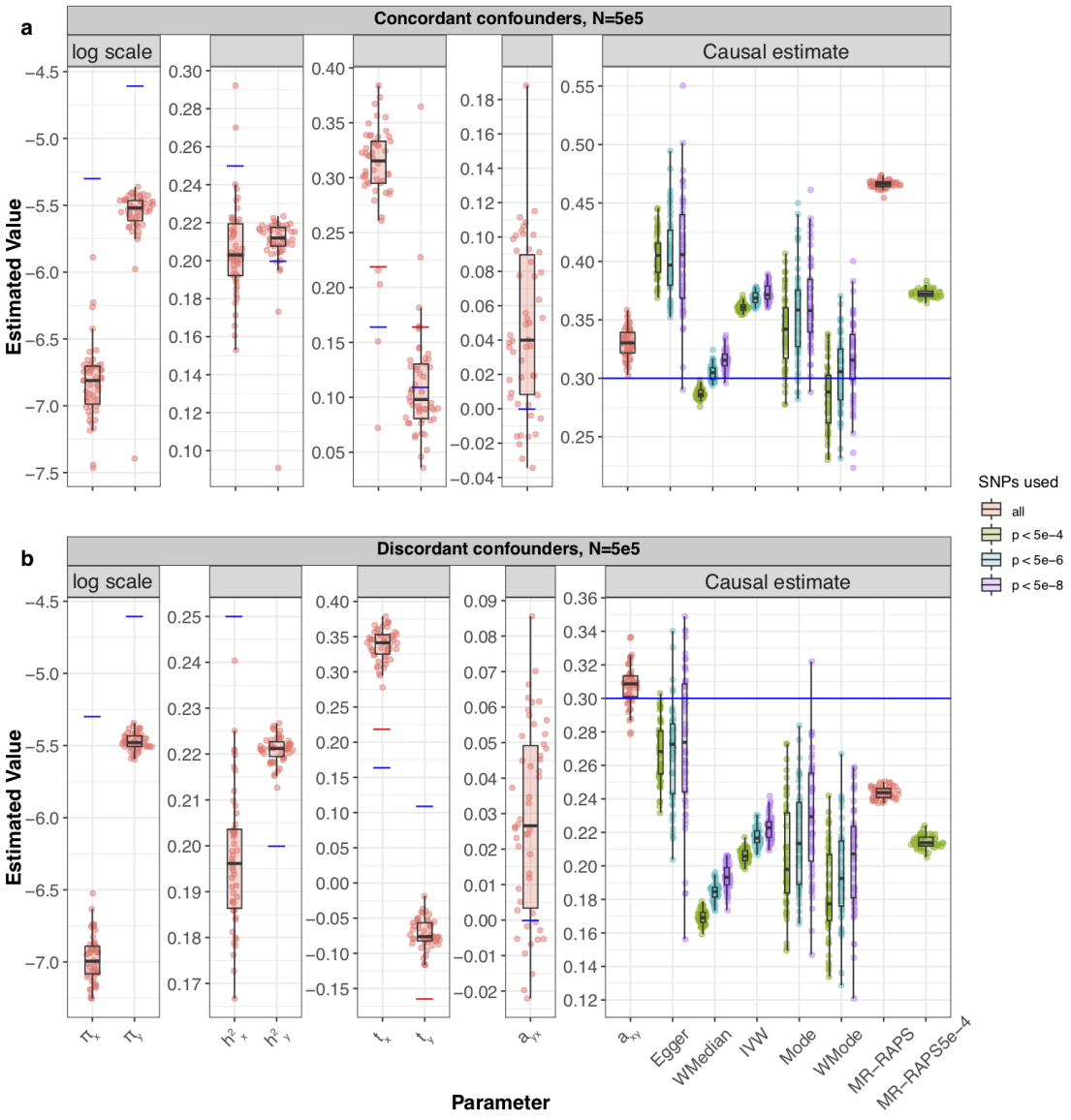


Figure S12: Simulation results where there are two underlying confounders, once with concordant and another with discordant effects on the exposure-outcome pair. Boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** The underlying data generations have two concordant heritable confounders U_1 and U_2 with positive effects on traits X and Y. **b** The data generations have two discordant heritable confounders with $t_x^{(1)} = 0.16, t_y^{(1)} = 0.11$ shown as blue dots and $t_x^{(2)} = 0.22, t_y^{(2)} = -0.16$ shown as red dots.

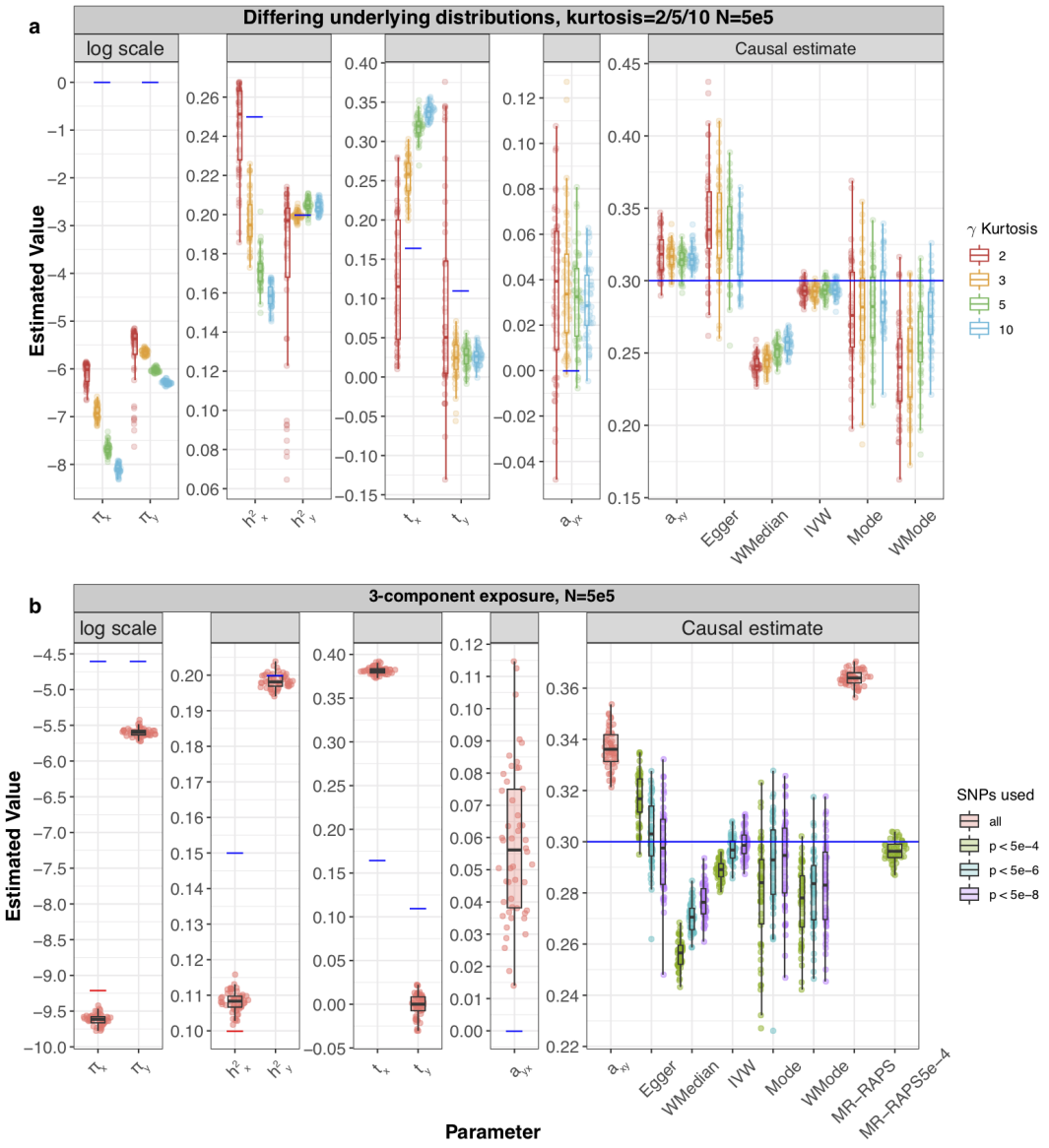


Figure S13: Simulation results under various scenarios. These Raincloud boxplots[28] represent the distribution of parameter estimates from 50 different data generations under various conditions. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** The different coloured boxplots represent the underlying non-normal distribution used in the simulation of the three $\gamma_x, \gamma_x, \gamma_u$ vectors associated to their respective traits. The Pearson distributions had the same 0 mean and skewness, however their kurtosis ranged between 2 and 10, including the kurtosis of 3, which corresponds to a normal distribution assumed by our model. The standard MR results reported had IVs selected with a p-value threshold of 5×10^{-6}. **b** Addition of a third component for exposure X, while decreasing the strength of U. True parameter values are in colour, blue and red for each component ($\pi_{x1} = 1 \times 10^{-4}, \pi_{x2} = 1 \times 10^{-2}, h^2_{x1} = 0.15, h^2_{x2} = 0.1$).

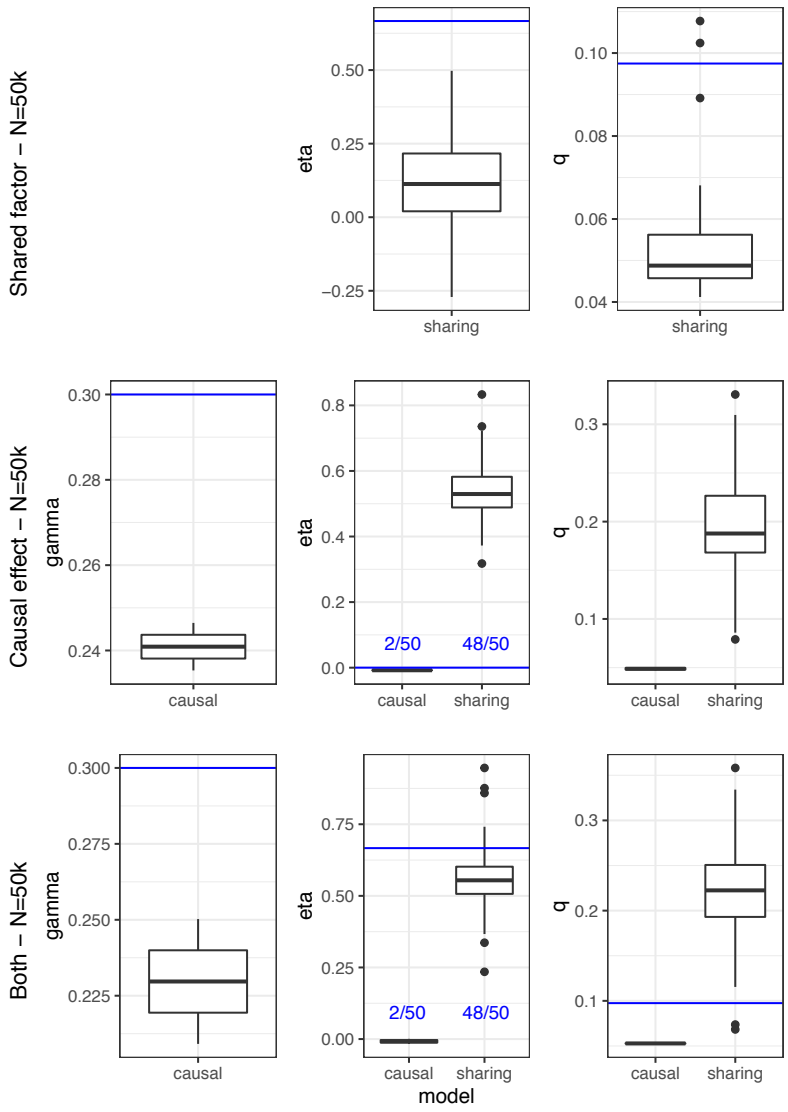


Figure S14: Running CAUSE on LHC-MR simulated data under the standard settings. Boxplots of the parameter estimation of CAUSE on LHC-simulated data ($n_x = n_y = 50,000$) under three different scenarios: presence of a shared factor only, presence of a causal effect only, presence of both. CAUSE returns two possible models with a respective p-value, the sharing and the causal model, where the causal mode is the significant of the two. When only an underlying shared factor was present in the simulated data, CAUSE had no significant causal estimates. With a true underlying causal effect, or when both an underlying causal effect and a shared factor was present, the causal model was significant only 4% of the simulations.

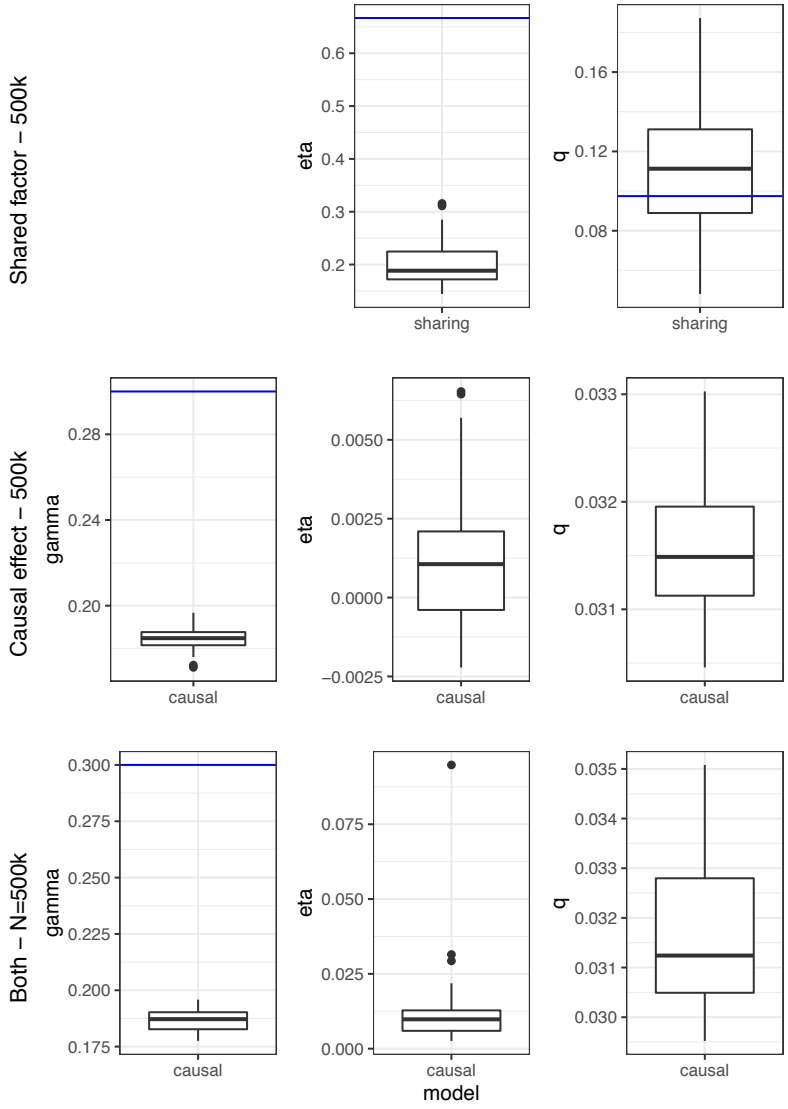


Figure S15: Running CAUSE on LHC-MR simulated data under the standard settings. Boxplots of the parameter estimation of CAUSE on LHC-simulated data ($n_x = n_y = 500,000$) under three different scenarios: presence of a shared factor only, presence of a causal effect only, presence of both. CAUSE returns two possible models with a respective p-value, the sharing and the causal model, where the causal mode is the significant of the two. When only an underlying shared factor was present in the simulated data, CAUSE had no significant causal estimates. With a true underlying causal effect, or when both an underlying causal effect and a shared factor was present, the causal model was significant 100% of the simulations.

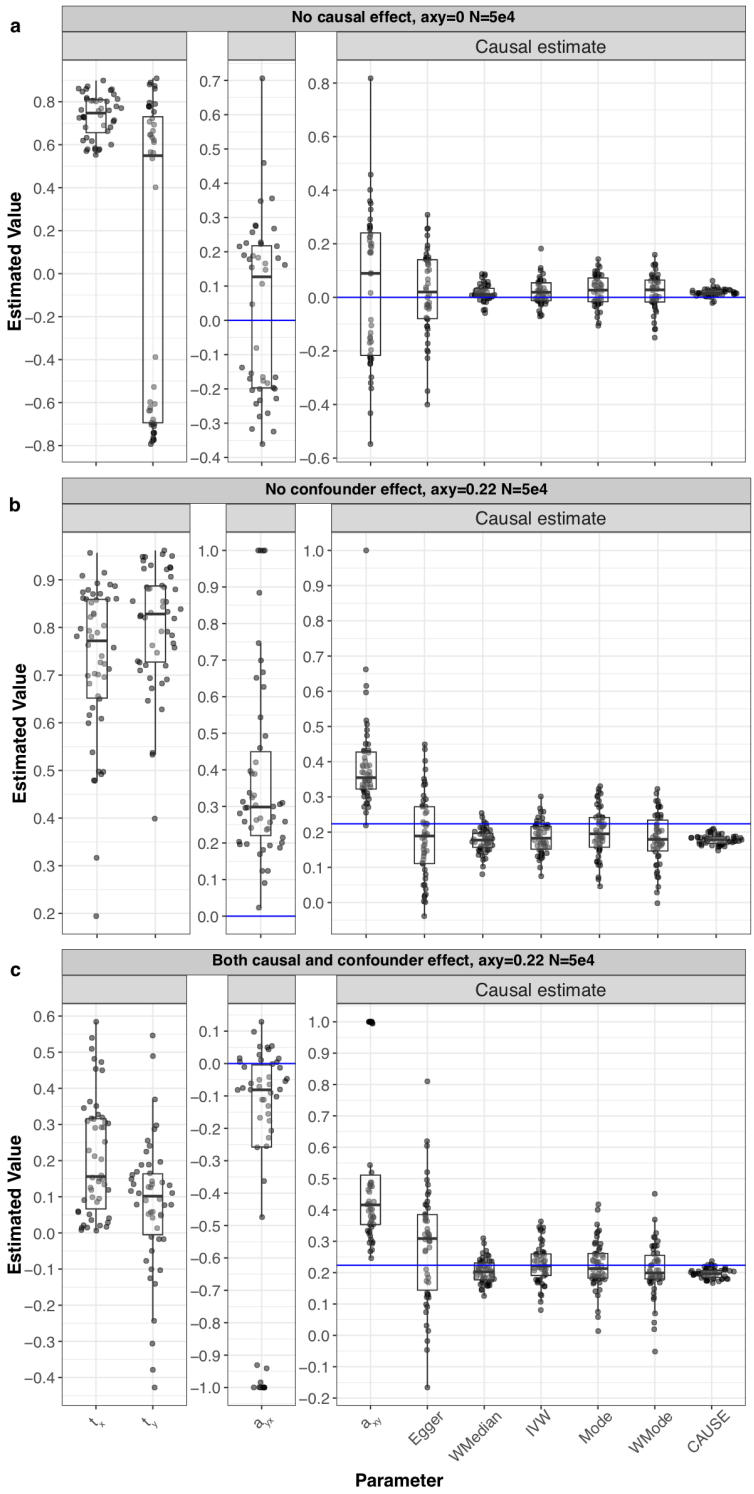


Figure S16: Running LHC-MR on CAUSE simulated data under various scenarios. Raincloud boxplots representing the distribution of parameter estimates from LHC-MR of 50 different data generations using the CAUSE framework. For each generation, standard MR methods, CAUSE as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** CAUSE data was generated with no causal effect but with a shared factor with an η value of ~ 0.22. CAUSE chooses a sharing model 100% of the time with no estimate for a causal effect. **b** CAUSE is simulated with causal effect but with no shared factor. **c** CAUSE is simulated with both a causal effect and a shared factor.

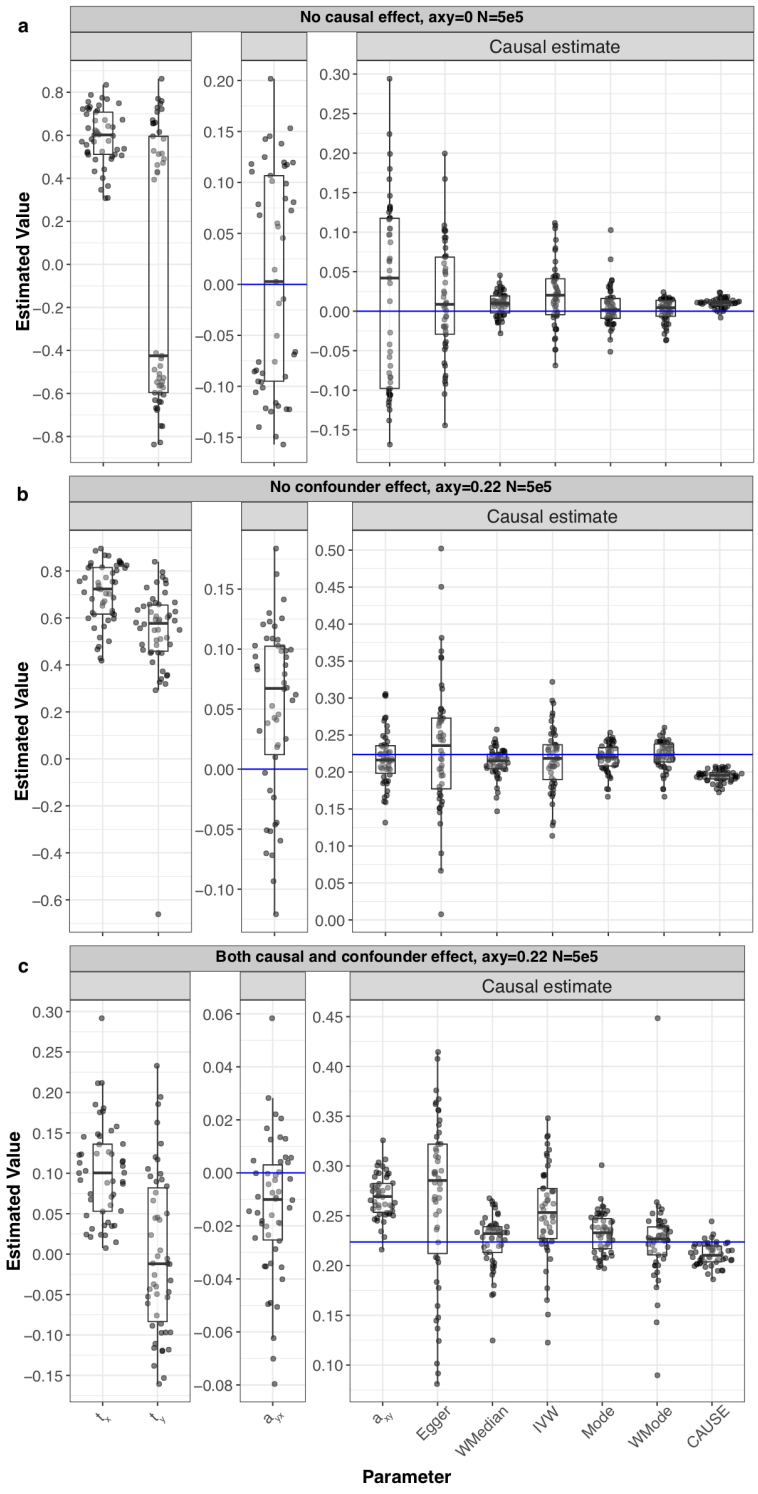
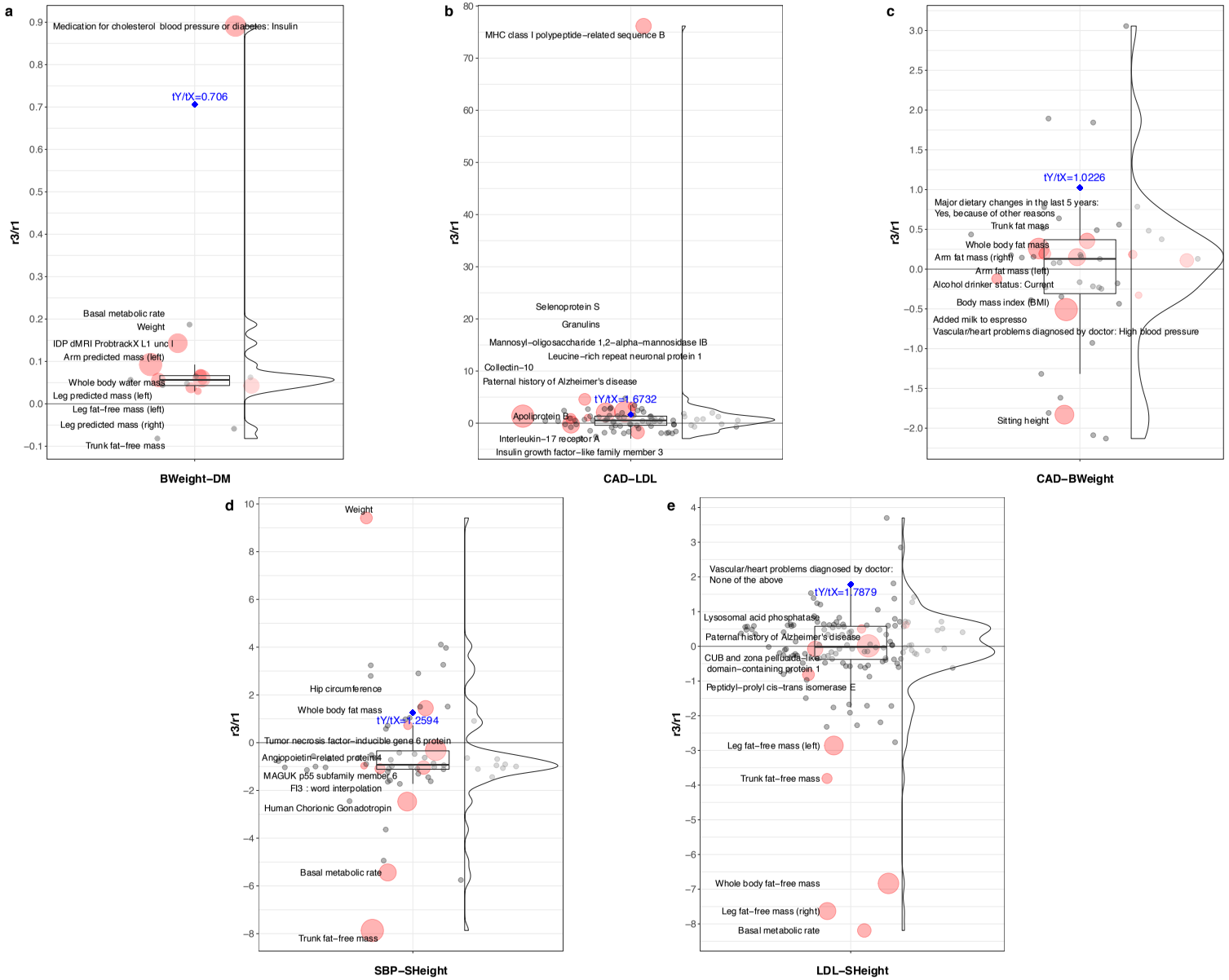


Figure S17: Running LHC-MR on CAUSE simulated data under various scenarios. Raincloud boxplots representing the distribution of parameter estimates from LHC-MR of 50 different data generations using the CAUSE framework. For each generation, standard MR methods, CAUSE as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. **a** CAUSE data was generated with no causal effect but with a shared factor with an η value of ~ 0.22. **b** CAUSE is simulated with causal effect but with no shared factor. **c** CAUSE is simulated with both a causal effect and a shared factor. LHC-MR seems to exhibit a bimodal effect at first glance, but the two peaks are not connected.


a
Medication for cholesterol blood pressure or diabetes: Insulin
tY/tX=0.706
Basal metabolic rate
Weight
IDP dMRI ProbtrackX L1 unc l
Arm predicted mass (left)
Whole body water mass
Leg predicted mass (left)
Leg fat−free mass (left)
Leg predicted mass (right)
Trunk fat−free mass
r3/r1
BWeight−DM
b
MHC class I polypeptide−related sequence B
Selenoprotein S
Granulins
Mannosyl−oligosaccharide 1,2−alpha−mannosidase IB
Leucine−rich repeat neuronal protein 1
Collectin−10
Paternal history of Alzheimer's disease
tY/tX=1.6732
Apolipoprotein B
Interleukin−17 receptor A
Insulin growth factor−like family member 3
r3/r1
CAD−LDL
c
tY/tX=1.0226
Major dietary changes in the last 5 years: Yes, because of other reasons
Trunk fat mass
Whole body fat mass
Arm fat mass (right)
Arm fat mass (left)
Alcohol drinker status: Current
Body mass index (BMI)
Added milk to espresso
Vascular/heart problems diagnosed by doctor: High blood pressure
Sitting height
r3/r1
CAD−BWeight
d
Weight
Hip circumference
Whole body fat mass
tY/tX=1.2594
Tumor necrosis factor−inducible gene 6 protein
Angiopoietin−related protein 4
MAGUK p55 subfamily member 6
FI3 : word interpolation
Human Chorionic Gonadotropin
Basal metabolic rate
Trunk fat−free mass
r3/r1
SBP−SHeight
e
Vascular/heart problems diagnosed by doctor: None of the above
tY/tX=1.7879
Lysosomal acid phosphatase
Paternal history of Alzheimer's disease
CUB and zona pellucida−like domain−containing protein 1
Peptidyl−prolyl cis−trans isomerase E
Leg fat−free mass (left)
Trunk fat−free mass
Whole body fat−free mass
Leg fat−free mass (right)
Basal metabolic rate
r3/r1
LDL−SHeight

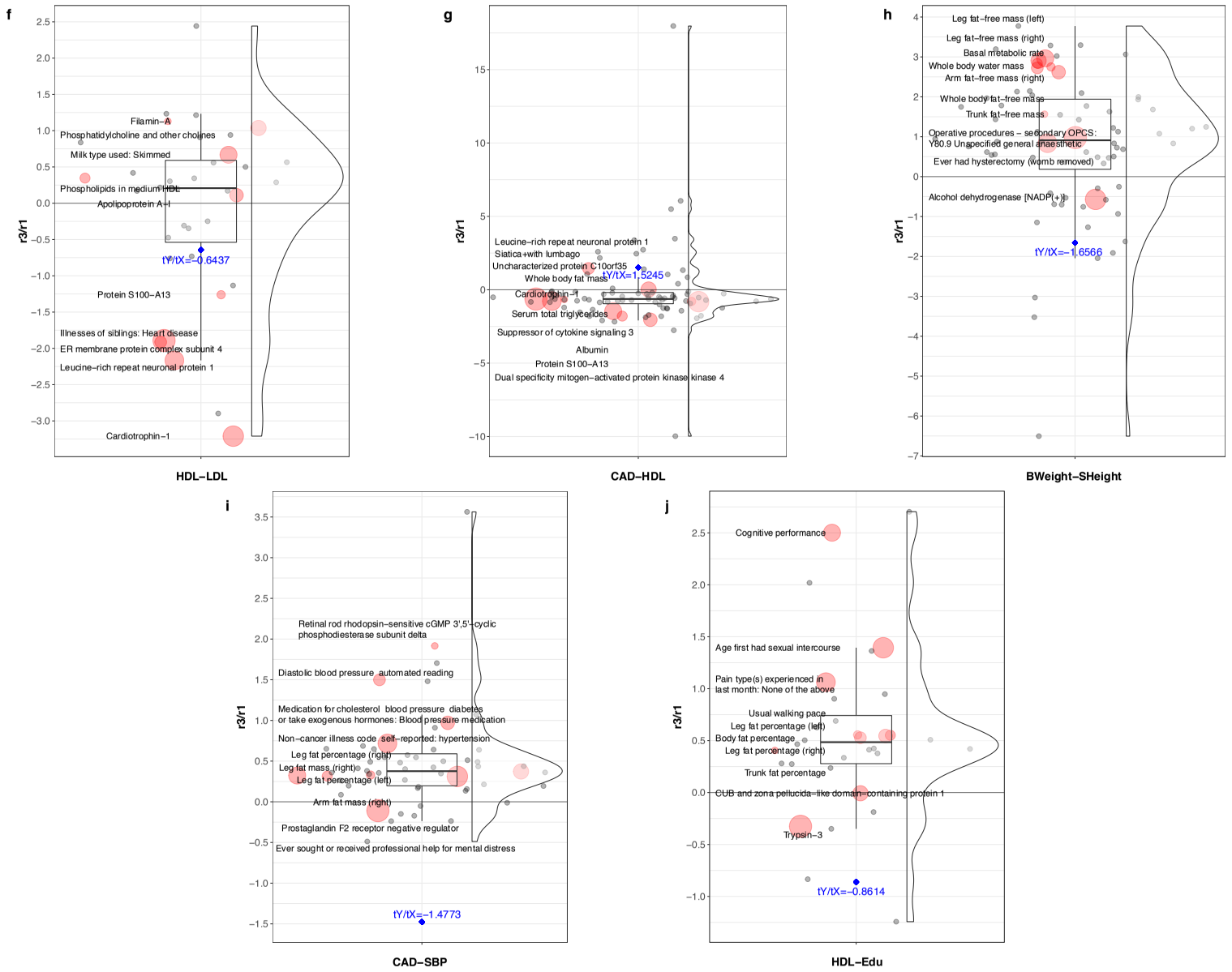


Figure S18: Confounder effects obtained from EpiGraphDB plotted as a raincloud with the r_3/r_1 ratio on the y-axis. The blue diamonds represent the t_y/t_x ratio derived from the LHC model for that trait pair, also reported in blue text. Labelled dots in red and their varying size show the ten largest confounder traits in terms of their absolute effect product on the two traits, whereas grey dots represent the rest of the confounder traits found by EpigraphDB.

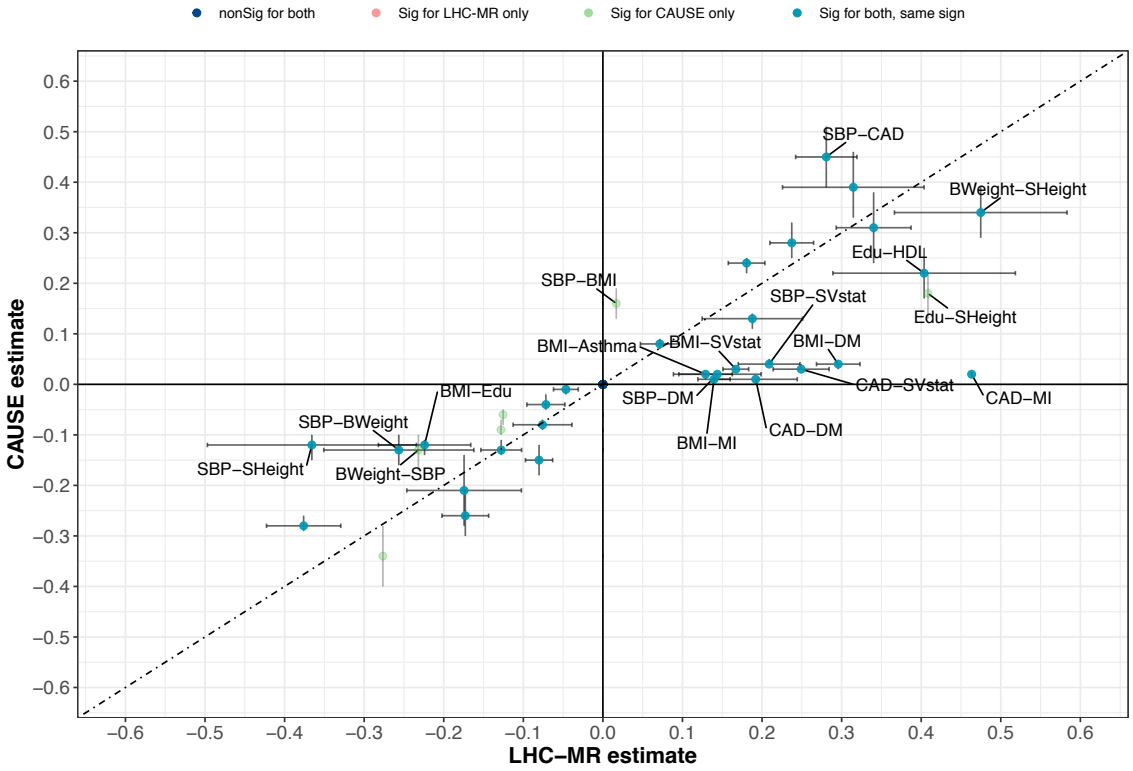


Figure S19: A scatter plot of the causal effect estimates between LHC-MR and CAUSE. To improve visibility, non-significant estimates by both methods are placed at the origin, while significant estimates by both methods appear on the diagonal with 95% CI error bars for LHC-MR estimates, and 95% credible interval error bars for CAUSE estimates. Labelled pairs are those with an estimate difference greater than 0.1.

Supplementary Tables

UKBB ID / Data Origin	Trait Name	Abbreviation	Sample Size	PMID
845	Age completed full time education	Edu	240,547	25826379
21001_irnt	Body mass index (BMI)	BMI	359,983	25826379
2443	Diabetes diagnosed by doctor	DM	360,192	25826379
20002_1075	Non-cancer illness code, self-reported: heart attack/myocardial infarction	MI	361,141	25826379
20002_1111	Non-cancer illness code, self-reported: asthma	Asthma	361,141	25826379
2887	Number of cigarettes previously smoked daily	PSmoke	84,456	25826379
20022_irnt	Birth weight	BWeight	205,475	25826379
50_irnt	Standing height	SHeight	360,388	25826379
4080	Systolic blood pressure, automated reading	SBP	340,159	25826379
20003_1140861958	Treatment/medication code: simvastatin	SVstat	361,141	25826379
30780_irnt	LDL Cholesterol	LDL	343,621	25826379
30760_irnt	HDL Cholesterol	HDL	315,133	25826379
UKBB + CARDIoGRAMplusC4D	Coronary Artery Disease	CAD	380,831	29212778

Table S1: Details of the origin study of each trait, its abbreviation used in this paper, the sample size of the study for that trait, as well as the PubMed article ID.

a

LHC-MR \ MR-Egger	Sig+	Sig−	nonSig
Sig+	10	0	27
Sig−	0	2	35
nonSig	0	2	56

b

LHC-MR \ WMedian	Sig+	Sig−	nonSig
Sig+	21	1	15
Sig−	0	12	25
nonSig	1	6	51

c

LHC-MR \ IVW	Sig+	Sig−	nonSig
Sig+	23	0	14
Sig−	0	17	20
nonSig	2	5	51

d

LHC-MR \ Mode	Sig+	Sig−	nonSig
Sig+	12	0	25
Sig−	0	2	35
nonSig	0	0	58

e

LHC-MR \ WMode	Sig+	Sig−	nonSig
Sig+	13	3	21
Sig−	0	7	30
nonSig	0	2	56

f

MR-Egger \ WMode	Sig+	Sig−	nonSig
Sig+	10	0	0
Sig−	0	1	3
nonSig	3	11	104

Table S7: Cross tables between LHC-MR and various standard MR methods comparing the significance and sign of each respective causal estimate. **f** shows a cross table between the two-least correlated MR methods in terms of their estimates.

Pair	$\alpha_{x\to y}$	**p-value**	γ	**IVW** $\alpha_{x\to y}$	**p-value**
BMI-Asthma	0.1290	4.99E-14	0.02 (0.01, 0.02)	0.0593	1.00E-08
BMI-DM	0.2958	1.07E-99	0.04 (0.03, 0.04)	0.2447	2.25E-140
BMI-SBP	0.1878	5.55E-09	0.13 (0.11, 0.14)	0.1547	1.11E-24
BMI-SVstat	0.1670	2.08E-91	0.03 (0.03, 0.03)	0.1570	4.26E-63
BMI-MI	0.1396	1.67E-41	0.01 (0.01, 0.01)	0.1027	9.16E-32
BWeight-SHeight	0.4748	9.60E-18	0.34 (0.29, 0.39)	0.2959	8.01E-10
SHeight-BWeight	0.1806	1.93E-53	0.24 (0.22, 0.25)	0.1803	7.21E-86
SBP-DM	0.1437	3.17E-07	0.02 (0.01, 0.02)	0.0697	3.69E-07
DM-SVstat	0.3147	4.11E-12	0.39 (0.33, 0.46)	0.2524	1.28E-16
SHeight-Edu	0.0715	8.42E-09	0.08 (0.07, 0.09)	0.0643	2.28E-21
SBP-SVstat	0.2089	4.84E-26	0.04 (0.04, 0.05)	0.1853	1.46E-52
Edu-HDL	0.4037	5.25E-12	0.22 (0.17, 0.27)	0.2848	4.06E-08
BMI-CAD	0.2373	2.37E-64	0.28 (0.25, 0.32)	0.1800	2.42E-53
CAD-DM	0.1920	5.92E-13	0.01 (0.01, 0.01)	0.0659	0.002455431
DM-CAD	0.4283	5.60E-19	1.95 (1.26, 2.64)	0.1796	4.15E-05
SBP-CAD	0.2807	2.86E-46	0.45 (0.39, 0.51)	0.2500	9.77E-24
CAD-SVstat	0.2491	8.82E-44	0.03 (0.03, 0.04)	0.3077	1.15E-25
CAD-MI	0.4634	0	0.02 (0.02, 0.02)	0.4191	3.07E-285
LDL-CAD	0.3402	1.17E-45	0.31 (0.24, 0.38)	0.2014	8.56E-27
BMI-Edu	-0.2241	3.74E-14	-0.12 (-0.14, -0.11)	-0.1892	6.15E-35
SHeight-BMI	-0.1278	1.40E-22	-0.13 (-0.14, -0.11)	-0.0854	9.01E-23
SBP-BWeight	-0.2565	9.85E-08	-0.13 (-0.16, -0.1)	-0.1646	1.20E-11
SBP-SHeight	-0.3657	4.81E-08	-0.12 (-0.15, -0.1)	-0.0967	0.004422636
SHeight-SBP	-0.0759	5.74E-05	-0.08 (-0.09, -0.07)	-0.0652	1.25E-15
SHeight-SVstat	-0.0465	4.76E-09	-0.01 (-0.02, -0.01)	-0.0328	6.78E-12
BMI-HDL	-0.3760	3.54E-56	-0.28 (-0.29, -0.26)	-0.3630	3.17E-111
SHeight-LDL	-0.0716	4.26E-09	-0.04 (-0.05, -0.02)	-0.0298	5.07E-06
BWeight-CAD	-0.1745	2.05E-06	-0.21 (-0.28, -0.14)	-0.0978	2.83E-05
SHeight-CAD	-0.0802	3.72E-20	-0.15 (-0.18, -0.12)	-0.0482	2.18E-12
HDL-CAD	-0.1729	7.00E-31	-0.26 (-0.3, -0.21)	-0.0778	5.45E-10

Table S8: Table comparing the causal estimates of LHC-MR, CAUSE, and IVW for trait pairs that had a significant causal effect in LHC-MR and CAUSE. The column showing the gamma (causal effect) estimate of the CAUSE method also reports its 95% credible intervals. A complete table for all the studied pairs is found in the Supplementary Table S5.