

Causal graphs for the analysis of genetic cohort data

Oliver Hines^{1,2}, Karla Diaz-Ordaz¹, Stijn Vansteelandt^{1,3}, and Yalda Jamshidi^{2,*}

¹Department of Medical Statistics, London School of Hygiene and Tropical Medicine, UK

²Molecular and Clinical Sciences Institute, St George's, University of London, UK

³Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

***Corresponding Author:**

Dr. Yalda Jamshidi

Molecular and Clinical Sciences Institute

St George's, University of London

Cranmer Terrace, London

SW17 0RE

UK

Email: yjamshid@sgul.ac.uk

Keywords: Causal Graphs, GWAS, Mendelian Randomisation

25
26

27

Abstract

28 The increasing availability of genetic cohort data has led to many Genome Wide Association Studies (GWASs)
29 successfully identifying genetic associations with an ever-expanding list of phenotypic traits. Association,
30 however, does not imply causation and therefore methods have been developed to study the issue of causality.
31 Under additional assumptions, Mendelian Randomisation (MR) studies have proved popular in identifying causal
32 effects between two phenotypes, often using GWAS summary statistics. Given the widespread use of these
33 methods, it is more important than ever to understand, and communicate, the causal assumptions upon which
34 they are based, so that methods are transparent, and findings are clinically relevant.

35 Causal graphs can be used to represent causal assumptions graphically and provide insights into the
36 limitations associated with different analysis methods. Here we review GWAS and MR from a causal perspective,
37 to build up intuition for causal diagrams in genetic problems. We also examine issues of confounding by ancestry,
38 and comment on approaches for dealing with such confounding, as well as discussing approaches for dealing with
39 selection biases arising from study design.

40 **1 Introduction**

41 Genetic cohort data is increasingly used to look for associations between candidate genes or genome regions and
42 specific outcome measures, or else between modifiable risk factors and disease outcomes. Genome Wide
43 Association Studies (GWAS), for example, are a popular and effective approach to analysing Single Nucleotide
44 Polymorphism (SNP) data, which identifies reproducible regions of the genome associated with common traits.
45 Observed GWAS associations, however, are not necessarily indicative of causal relationship, unless one is willing
46 to make additional assumptions on the causal structure of the cohort data.

47 Mendelian Randomisation (MR) is another popular method, which uses genetic cohort data (or GWAS summary
48 statistics) to establish causal effects between two phenotypes. MR seeks to exploit random genotype allocation,
49 which occurs naturally due to Mendelian inheritance. The requisite MR assumptions are strong, and the causal
50 structure underlying the data must be carefully considered so that biases are not unwittingly introduced. Since
51 both GWAS and MR rely on genetic cohort data, it is more important than ever to understand, and communicate
52 the causal structures found in these datasets, so that findings remain clinically relevant.

53 Universal frameworks to study causal structures have emerged in the past few decades, based on potential
54 outcomes modelling[31] or causal graphs[27], contributing towards a modern causal understanding of several
55 existing techniques, such as, randomised controlled trials, instrumental variable, and observational data
56 techniques (propensity score methods and sample matching). Causal graphs may inform both the design and
57 analysis of observational studies, and have successfully been applied to problems in epidemiology[13, 14], social
58 science[4] and economics[21] to represent causal assumptions, and derive causal quantities from observed data.

59 Eliciting and defending causal assumptions requires an expert understanding of the problem at hand. Here we
60 review methods from genomics and genetic epidemiology, highlighting common causal structures which can bias
61 observed associations. We advocate the use of causal graphs, firstly as a formal tool for representing and
62 communicating the causal assumptions regarding data collection and study design, which underly analytical
63 methods, and secondly, for deriving testable implications based on those assumptions. Causal graphs have several
64 attractive properties in this regard. As a communication tool they are inherently diagrammatic and equation-free,
65 aiding interpretability, whilst as a derivation tool one may apply powerful and rigorous mathematical rules,
66 which link causal relations to statistical associations. These rules are summarised in Section 2.1.

67 We will initially introduce causal concepts which form the basis of our discussion. These are then applied to an
68 example of pleiotropy in Section 1.2. Section 2 discusses causal methods for analysing selection biases, using, as
69 an example, the analysis of case-control data for secondary trait association. Here we see the utility of causal
70 graphs in deriving associations between variables which occur under selection. Section 3 then reviews GWAS

71 assumptions, addressing issues related to population structure, while Section 4 reviews MR causal assumptions,
72 highlighting several ways in which they may be violated.

73 1.1 Introduction to Statistical Causal Inference

74 There exists rich philosophical debate on what it means for one thing to *cause* another[40], however, in the study
75 of causal inference an interventionalist definition is used[27, 13, 18]. In this way, questions of causality are
76 reduced to questions of the type: *what would happen if...?*

77 For example, for two variables A and B , we say that A **causes** B if the value that B takes would be different (or
78 different in probability) if we had intervened by setting A to some other value. In this context we might also say
79 that A **causally influences** B or that B is **causally dependent** on A . Two variables are said to be **statistically**
80 **dependent** (or associated) if knowing the value of A in some way provides some information about the value of B
81 (or vice-versa). Statistical dependence may arise due to a causal dependence between A and B , but also as a result
82 of a causal dependence of both A and B on a third variable C , as we will see in the example in Section 1.2.
83 Conversely, two variables are **statistically independent** if knowing the value of A does not provide any
84 information about the value of B (and vice-versa).

85 This notion of causality may also be graphically represented using an arrow[13, 18, 28, 29], for example, $A \rightarrow B$
86 reads as “ A causes B , but B could not possibly cause A ”. This arrow says nothing about the magnitude or direction
87 of the effect that A has on B , just that if we were to intervene on A , then something would happen to B . Using these
88 arrows one can form **paths**, which are any sequence of variables linked by arrows. For example, if A and B shared
89 a common cause, C , then one may write the path, $A \leftarrow C \rightarrow B$. All possible paths containing three variables are
90 given in Table 1. A path is *causal* if all the arrows point in the same direction. The path $A \rightarrow C \rightarrow B$, for example, is
91 causal since A causes C which causes B , therefore if we were to intervene on A , the value of B could be different.
92 Depending on the directions of the arrows, we also have additional terminology for the intermediate variable,
93 also given in the table.

Path	Description	Terminology for the variable C
$A \rightarrow C \rightarrow B$	A causes B (through C)	Mediator
$A \leftarrow C \leftarrow B$	B causes A (through C)	Mediator
$A \leftarrow C \rightarrow B$	A and B share a common cause C	Confounder
$A \rightarrow C \leftarrow B$	A and B both cause C	Collider

94 Table 1: All possible paths between three variables (A, B, C), with a brief description and additional terminology for
95 the intermediate variable C

96 On its own, a single path is of limited use, motivating a network structure to represent several paths at once. The
97 causal Directed Acyclic Graph (DAG) is such a structure, which for a set of variables, contains *all possible* paths
98 between them. Causal graphs are said to be **acyclic** if there are no causal paths from one variable back to itself. It
99 may seem obvious to say that any two variables, A and B , on a causal graph could either be linked by the arrow A
100 $\rightarrow B$, the arrow $B \rightarrow A$, or no arrow at all. Each configuration makes different assertions about the impossible
101 causal relationship between A and B . Respectively these are that B is not a direct cause of A , A is not a direct cause
102 of B , or that A and B could not possibly be direct causes of each other. In this sense the arrows which are absent,
103 and those which are present are equally important. Similarly, one must be careful to include common causes of A
104 and B , even if they are unmeasured, since to not do so is to assert that it is impossible for such variables to exist.

105 At this stage it is also useful to introduce some terminology[13, 18, 27, 28, 29], which will become important later
106 on. Firstly, a **collider** is any variable on a path which is causally dependent on the two variables adjacent to it, as
107 in the final example in Table 1. Secondly, the **ancestors** of a variable are those which causally influence it (i.e.
108 there is a causal path from each ancestor to the variable), and finally the **descendants** of a variable are those
109 which are caused by it (i.e. there is a causal path from the variable to its descendants).

110 1.2 Example using Pleiotropy

111 Our first example is inspired by a recent discussion of pleiotropy of the fat mass and obesity-related gene
 112 (*FTO*)[12]. Consider a Single Nucleotide Polymorphism (SNP) in the *FTO* gene, such as rs1421085, which has been
 113 found to be associated with adiposity and brain function[8]. Suppose that a genetic cohort study has been
 114 conducted where, for each individual in the study population, an investigator measures body mass index (BMI), *B*,
 115 cerebral blood flow, *C*, and genotype rs1421085 in the *FTO* gene, denoted by *F* and coded as 0,1 or 2.

116 The original authors suggested that reduced cerebral blood flow in the medial prefrontal cortex may effect
 117 impulse control and hence BMI [12]. As an illustration, we will attempt to refute the null hypothesis, that there is
 118 no causal relationship between cerebral blood flow and BMI by (1) positing the causal relationships that we
 119 believe hold amongst the variables involved; (2) representing these causal relationships using a causal graph; and
 120 (3) examining the graph, using formal operations, to derive testable assumptions.

121 Since a person's genome is assigned before their BMI or cerebral blood flow is determined, we argue that it is safe
 122 to assume that *B* and *C* could not possibly cause *F*. This assumption, however, says nothing about whether *F*
 123 causes *B* or *C*. Since it is possible that *F* causes *B* and *C* we must include the arrows $F \rightarrow B$ and $F \rightarrow C$ in our causal
 124 graph. For the purposes of illustration, we will additionally make the strong assumption that no other measured
 125 or unmeasured variables causally influence both *B* and *C*.

126 The causal graph in Fig.1 represents the causal assumptions posited between *F*, *B* and *C* under the null hypothesis
 127 that there is no causal relationship *B* and *C*. These assumptions are unnecessarily strong for the purpose of
 128 illustration, since additional variables might be included such as age or physical activity level, which are common
 129 causes of both *B* and *C*. Other violations of our assumption, which could arise due to population structure, are
 130 discussed in Section 3. We remark that while the causal graph in this example is perhaps oversimplified, such
 131 assumptions are not uncommon, and by using a causal graph representation we are required to be transparent
 132 about them.

133 [Figure 1 Here]

134 In the graph in Fig.1, there is no causal path between *B* and *C*, but that does not mean that they are statistically
 135 independent. In fact one might expect a negative correlation between BMI and cerebral blood flow since those
 136 who inherit the *FTO* variant are likely to have a higher BMI and also a lower cerebral blood flow. This statistical
 137 dependency can be read off the graph in the form of the possible path: $B \leftarrow F \rightarrow C$. It is a general rule that two
 138 variables will be statistically independent if all paths between them that contain colliders. For this reason, we can
 139 refer to paths that do not contain a collider as *open paths* and those that do as *closed paths*.

140 Using our causal graph, we may derive testable assumptions in an attempt to falsify our null hypothesis. Imagine,
 141 for example, that we are told the value of *B* for a particular patient, and are asked to predict their value of *C*. The
 142 value of *B* may inform our prediction since *B* and *C* may be statistically dependent (due to confounding by *F*). If,
 143 however, we are subsequently told the patient's *FTO* variant then, under our causal assumptions, a new
 144 prediction based on *F* and *B* is no better than a prediction based on *F* alone, since *B* only informed our prediction
 145 in so much as it may have conferred some information about *F*.

146 This important observation is an example of how one may *block* open paths, such as $B \leftarrow F \rightarrow C$, by *conditioning* on
 147 an intermediate variable (*F*). Conditioning on a variable can be done either by stratifying by that variable or by
 148 including it as an independent variable in a regression model for *B* or *C*. These conditional independences are
 149 essential as they allow us to falsify our causal assumptions.

150 In practice, this means that if one were to stratify our imaginary study population by their *FTO* gene variant, then,
 151 under our causal assumptions, no association between *B* and *C* should be observed within strata. An association
 152 between *B* and *C* within strata is, therefore, evidence that our assumptions are invalid. This could be because our
 153 null hypothesis does not hold, and *B* and *C* are causally related, or else because the relationship between them is
 154 confounded by some other variables, which we have not accounted for.

155 2 Selection Bias

156 Due to the considerable cost of obtaining original genetic cohort data, it is common for case-control data to be
 157 repurposed for analysis of a secondary trait, such as human height[16, 44], obesity[25], or plasma lipid
 158 concentration[45]. Methods that fail to account for the case-control study design, are known to result in inflated
 159 error rates when testing for null association using GWAS [23]. Indeed it has been argued that epidemiological
 160 data analysis depends as much on study design and background information, as on the data itself[30].

161 Gene-phenotype associations, induced as a consequence of study design, are problematic in GWAS analyses
 162 because they are indistinguishable from underlying causal associations in GWAS results. Using causal graphs we
 163 may gain some insight into how the non-random selection of individuals to the study cohort propagates to non-
 164 randomness in our variables of interest. We will consider an illustrative example, inspired by a real study on the
 165 effect of Sex Hormone Binding Globulin (SHBG) on Type 2 diabetes in women[11]. Consider that the study cohort
 166 was recruited on a case-control basis and consists of women with a recent Type 2 diabetes diagnosis ($D = 1$) and
 167 controls ($D = 0$), with genotyping carried out for all women. We shall examine the issues which arise when this
 168 cohort is used to conduct a GWAS analysis, with SHBG as the outcome of interest.

169 SHBG is a glycoprotein, produced in the liver, and the level of SHBG in an individual's blood plasma will be
 170 denoted by H . The original authors found that high levels of SHBG were associated with a lower risk of Type 2
 171 diabetes and for this example we shall assume that diabetes status does not causally influence SHBG level.
 172 Imagine also a specific SNP, G , which does not causally influence SHBG, but does causally influence diabetes
 173 diagnosis by some other mechanism. As with the example in Section 1.2 we shall make the "no unobserved
 174 confounding" assumption, i.e. that there are no common causes of H , G , or D that we have not accounted for.

175 Due to the case-control design, diabetes status D causally influences selection to cohort, S . By definition $S = 1$ for
 176 all women in the cohort and $S = 0$ for all other women in the population as a whole. Our causal assumptions are
 177 represented by the causal graph in Fig.2a.

178 [Figure 2 Here]

179 Under these assumptions, G and H are statistically independent as there are no open paths between them. One
 180 would expect, therefore, to observe no association between G and H for women sampled from the population. Our
 181 cohort, however, is not randomly sampled from the population, but instead we observe only those for whom $S = 1$.
 182 This is equivalent to an unavoidable stratification by S , which allows us to observe only the $S = 1$ stratum. In this
 183 stratum, a "spurious" association between G and H may be induced, which we demonstrate by first examining the
 184 $D = 1$ and $D = 0$ strata separately.

185 In the cases group ($D = 1$) an association between G and H would be observed, since, if an individual's genotype
 186 suggests they are unlikely to have diabetes, then their diabetes status is more likely due to a low level of SHBG,
 187 and vice-versa. For women in the control group ($D = 0$) an association between G and H would be observed, since
 188 women in this group are less likely to carry the genotype associated with diabetes and are also more likely to
 189 have high SHBG.

190 We see, therefore, that G and H are associated in both the $D = 0$ and $D = 1$ strata and that this association must be
 191 induced by the stratification process, since G and H are not associated in the population. Worse than this,
 192 however, is that stratifying by S also induces associations between G and H because the proportions of each D
 193 strata in our cohort are not representative of the population as a whole. For selection problems such as these we
 194 have no choice but to consider only the strata $S = 1$.

195 In this simple example we were able to reason that selection bias may influence our results, however, in other
 196 examples it may not be so clear. Causal graphs may go some way to elucidate selection biases. It is a general rule
 197 that conditioning on a collider, or the descendants of a collider, induces statistical dependencies between the
 198 ancestors of the collider. In our case-control example D was a collider on the path: $G \rightarrow D \leftarrow H$ and we were forced
 199 to condition on S , which is a descendant of D . This conditioning resulted in a statistical dependency between G
 200 and H (the ancestors of D). This induced dependency is represented by the dashed line on the causal graph in
 201 Fig.2b.

202 In Section 1.2 we saw how open paths on causal graphs could be blocked by conditioning on intermediate
 203 variables. In this example, however, conditioning has the opposite effect. By unintentionally conditioning on
 204 colliders, we are effectively unblocking a path that was otherwise closed, thereby inducing associations. Several
 205 solutions have been proposed, which allow case-control data to be used for secondary trait analysis in association
 206 studies. Example analysis strategies include analysing the cases and controls separately, re-weighting the data
 207 using additional models, or including case-control status as a covariate [38, 33].

208 Biases introduced by conditioning on colliders are generally referred to as *collider stratification biases*[2]. The
 209 inclusion of selection variables in causal graphs, like the variable S in the case-control example, can also be useful
 210 for expressing selection and retention assumptions which suffer from similar collider stratification biases[26].
 211 The UK Biobank is an example of a cross-sectional cohort study ($n \approx 500,000$) self-selected from a population of 9
 212 million individuals invited to participate. The resultant cohort contains a lower proportion of current smokers
 213 (11% in the UK Biobank, vs approximately 19% in the general population), with a similar discrepancy observed in
 214 educational qualification attainment. For a highly self-selected cohort, such as the UK Biobank, causal graphs may
 215 be useful in exposing subtle biases induced by this self-selection.

216 2.1 D-separation

217 The rules discussed in Sections 1.1 and 2 are collectively known as the rules of d-separation (statistical
 218 dependence separation). These rules describe statistical dependencies implied by causal graphs before and after
 219 conditioning on variables. Table 2 gives a summary of these rules for all possible paths of three variables. To
 220 consider longer, more complex paths one must ‘chain together’ these triplets, and to consider the statistical
 221 dependence between variables on the whole causal graph, one must consider all possible paths.

222 For complex, multivariate causal graphs this could result in a laborious manual analysis. Fortunately, however,
 223 the tool www.dagitty.net [20] may be used to examine statistical dependence on causal graphs using an online
 224 web tool or R package.

Path	Before conditioning on C	After conditioning on C
$A \rightarrow C \rightarrow B$	open	closed
$A \leftarrow C \leftarrow B$	open	closed
$A \leftarrow C \rightarrow B$	open	closed
$A \rightarrow C \leftarrow B$	closed	open

225 Table 2: Summary of the rules of d-separation for all possible paths containing three variables. The two additional
 226 columns describe the statistical dependence of A and B before and after conditioning on the intermediate variable
 227 C .

228 3 Causal Graphs for Genome Wide Association Studies

229 GWAS studies are a popular and effective approach to analysing SNP data, which identifies reproducible regions
 230 of the genome associated with common traits. As of February 2020, the GWAS Catalogue contains 4439
 231 publications and 175870 associations[6]. Despite their popularity, it is important to remember that the
 232 associations discovered by GWAS are not necessarily causal unless one is willing to make additional assumptions.
 233 In this section, we use causal graphs to make these assumptions explicit. Genetic relatedness between individuals
 234 in the study population poses an additional, well-known challenge that results in individuals with shared ancestry
 235 inheriting similar common variants. Heterogeneous study populations, therefore, complicate the task of
 236 separating the contributions of individual genetic variants toward phenotypes of interest. We refer to the
 237 problem of heterogeneous ancestry as confounding by ancestry, since this more closely aligns with the language
 238 of causal inference. It is also referred to as population structure or population stratification, when at the
 239 population level, and kinship, at the familial level.

240 As an illustrative example, we will use Carotid Intima-Media Thickness (CIMT) as a phenotype of interest Y . In its
 241 most basic form, one assumes that the study population is in Hardy-Weinberg Equilibrium (HWE), that is, for each

242 individual, the value of their value of a particular SNP of interest, G , is drawn from a binomial distribution with
 243 some fixed minor allele frequency for the population.

244 Common practice is to model a continuous phenotype, Y , using a model which is linear in G , and other relevant
 245 variables, such as age and sex, denoted by the 'Environmental' vector, E . When Y is a binary outcome, generalised
 246 linear models such as the logistic model, are often used. The linear model for a continuous phenotype, Y , may be
 247 written as

$$Y = \alpha G + \sum_{j=1}^p \beta_j E_j + \epsilon \quad (1)$$

248 where ϵ is a noise term, with constant mean given G and E , and β is a vector of parameters associated with the p
 249 environmental variables contained in the vector E . The unknown model parameters, α and β , may be estimated by
 250 Ordinary Least Squares (OLS). Ideally we would like to interpret the α parameter as *a parameter which quantifies*
 251 *the influence that the gene of interest has on the phenotype*, however, to do so is to make a causal assertion,
 252 requiring an examination of causal assumptions. We note that for a discussion of causal assumptions, the exact
 253 form of the regression model is not important. Instead, from a causal perspective, we are concerned with the
 254 variables which are and are not included in the regression model.

255 One possible causal graph for the basic GWAS analysis, which gives the α parameter the desired causal
 256 interpretation is given in Fig.3a. This graph is not unique since it is not strictly required that G and E are
 257 independent. Using the running example, the key features of this graph required to interpret α causally are
 258

- 259 1. CIMT does not influence the gene of interest, but the reverse may be true.
- 260 2. CIMT does not influence age or sex, but the reverse may be true.
- 261 3. There are no variables (observed or otherwise), which are common causes of CIMT and the gene of interest,
 262 or of CIMT and age or sex.

263 The first of these assumptions is justified through the biological understanding that G is assigned before
 264 phenotypes are determined, hence reverse causation is not possible. Likewise, the second assumption is
 265 reasonable from a biological perspective. Assumption 3, however, is where the basic model breaks down. Under
 266 modern theories of Mendelian inheritance, the gene of interest depends on an individual's parental genotypes, or
 267 more generally on their ancestry. Along with the gene of interest, each individual inherits many other genetic
 268 variants, G^* , each of which could also have a causal influence over Y . The ancestry of an individual is therefore a
 269 confounder as it may be a common cause of both G and Y .

270 This effect is, however, negated if one assumes that Y is monogenic, so is causally affected by only one single SNP.
 271 Conversely the effect is amplified for polygenic traits, such as CIMT, which are thought to be affected by multiple
 272 genetic variants.

273 [Figure 3 Here]

274 To adequately adjust for confounding by ancestry, the basic GWAS graph Fig.3a must be updated to reflect
 275 Mendelian inheritance assumptions. Fig.3b shows a causal graph, modified to include an unmeasured ancestry
 276 variable, C , which affects the phenotype of interest through both the gene of interest, G , and other inherited
 277 variants, G^* . In this updated causal graph, we see that there are two open paths by which the gene of interest is
 278 associated with CIMT, specifically the $G \rightarrow Y$ causal path and the $G \leftarrow C \rightarrow G^* \rightarrow Y$ non-causal path. If one were able
 279 to block the non-causal path, then, the remaining association between G and Y must be due to the causal path.

280 One strategy for blocking the path is to condition on ancestry by stratification. Since C is unmeasured, one must
 281 assume that the population consists of one strata, which is homogeneous in ancestry with a random mating
 282 scheme and no natural selection. Under these assumptions, the HWE model is recovered, whereby G is drawn
 283 from the same distribution for all individuals, hence G and Y are not confounded by ancestry.

284 The causal graph in Fig.3b made several additional assumptions regarding the ancestry variable, C . The first is
 285 that there is no direct path $C \rightarrow Y$. Modern epigenetic theory, however, does permit such paths through
 286 ‘imprinting’ mechanisms, whereby an individual inherits DNA of the same sequence, whose function is altered by
 287 the presence of additional methyl groups.

288 Furthermore, Fig.3b assumes that C and E are independent. This may not be true, however, for a global study,
 289 where individuals from different ethnic groups, may have been brought up in different geographical locations,
 290 and hence, different meteorological and socio-economic conditions. It is reasonable, therefore, to posit a $C \rightarrow Y$
 291 path through some unobserved environmental variables. We emphasise again that the arrows absent from a
 292 causal graph are important as they represent causal relationships which are assumed not to exist, whilst the
 293 arrows represent causal relationships which may exist.

294 3.1 Using Principal Components to Adjust for Ancestral Confounding

295 Examining the causal graph in Fig.3b, we discussed how the non-causal path: $G \leftarrow C \rightarrow G^* \rightarrow Y$ may be blocked by
 296 conditioning on C when one assumes the study population is homogeneous. For heterogeneous populations,
 297 however, stratification by C is not possible because it is unmeasured. Instead, the non-causal path can be blocked
 298 by conditioning on the remaining observed SNPs, G^* . This involves using G^* in a regression model for Y , or using
 299 G^* for stratification.

300 Intuitively, conditioning on G and G^* removes any dependency between C and Y since, if the full genotype of an
 301 individual is used to predict their phenotype, then knowledge of their ancestral genotypes provides no new
 302 information to improve our prediction. Using the full genotype in a regression model for Y requires careful
 303 consideration, since the number of covariates (SNPs), p , may exceed the number of individuals in the study, $n < p$.
 304 Such ‘high-dimensional’ problems require alternative models and estimation techniques.

305 Due to the high-dimensionality, modifying the linear model in Eq.1 to include the remaining genes as covariates
 306 would result in a model which is impossible to fit by OLS. One very common solution is to drastically reduce the
 307 dimensionality of the genetic information, using Principal Components (PCs).

308 PCs are used in several ways within genomic analysis: (i) PCs can be used to cluster individuals, either by
 309 excluding anomalous individuals from the dataset [1], or else clustering the data for use in a Structured
 310 Association analysis, (ii) some PC values may be included as fixed effects in a GWAS analysis, thereby accounting
 311 for some of the phenotype variation, which can be explained by the remaining SNPs, and (iii) PCs may be included
 312 as random effects in the GWAS analysis, an approach which is equivalent to using a Linear Mixed Model (LMM)
 313 [19].

314 Method (i) may be causally interpreted as stratifying the population into one or more sub-populations, for which
 315 we believe that HWE holds. Analysis of each sub-population may be conducted using a basic GWAS analysis.
 316 Limitations of this method are that confounding by ancestry is not accounted for within strata and it is not clear
 317 how to tune the stratification process.

318 The linear model for methods (ii) and (iii) may be written as

$$319 \quad Y = \alpha G + \sum_{j=1}^p \beta_j E_j + \sum_{j=1}^q \gamma_j P_j + \epsilon \quad (2)$$

320 where P is the vector of q principal components, summarising the genetic data of a particular individual, each
 321 component of which has a coefficient given by the γ parameter vector, and where ϵ has constant mean given G, E
 322 and P . In the fixed effect model (method ii), the q -dimensional parameter vector, γ is treated as a fixed covariate,
 323 which may be estimated using conventional methods such as by OLS.

324 Alternatively, one may treat the parameters γ_j as random effects (method iii), by assuming a normally distributed
 325 prior for γ , resulting in a LMM. The use of LMMs in genomic data is not restricted to GWAS analyses. They are

326 frequently applied to phenotype prediction, heritability estimation, and rare-variant analysis [24]. One key
 327 feature of LMMs is that the random effect (given by $\sum_{j=1}^q \gamma_j P_j$ above) may be written in terms of a ‘genetic
 328 similarity matrix’, which is used to model the covariance between any pair of individuals in the cohort. A more
 329 detailed discussion of LMMs and methods for measuring genetic similarity can be found in Appendix A.

330 4 Causal Graphs for Mendelian Randomisation

331 Mendelian Randomisation (MR) studies also make use of genetic SNP data, or GWAS summary statistics, with the
 332 aim of inferring the effect of a genetically modified exposure (e.g. alcohol consumption) on another phenotype
 333 (e.g. cardiovascular disease). GWAS results from multiple cohorts may be used to conduct Two- Sample MR
 334 analysis. MR base which is a database of GWAS statistics for conducting Two-Sample MR, contained associations
 335 from 1673 GWAS, as of May 2018[17]. Another systematic review estimates a 10-fold increase in published MR
 336 studies between 2004 and 2015, with the majority (51%) in the fields of cardiovascular disease and diabetes[37].
 337 MR is therefore increasing in popularity, most likely due to the increasing availability of GWAS summary statistics
 338 and large cohorts with genetic and phenotypic data.

339 This section provides an overview of the technique, from the statistical causal inference framework. We refer the
 340 interested reader to [40, 7, 32].

341 4.1 Instrumental Variable Methods

342 MR exploits the idea that a particular genotype affects the phenotype of interest only indirectly, through the
 343 exposure of interest, and that this genotype is assigned randomly (given the parents’ genes) at meiosis,
 344 independently of the possible confounding factors. This is essentially using the genotype as a so-called
 345 *instrumental variable* (IV) for the effect of the exposure on the outcome [9]. This is appealing, as it allows to
 346 estimate causal effects even in the presence of exposure-outcome unobserved confounding. Nevertheless, MR
 347 makes a number of causal assumptions, known as IV assumptions, which are not always carefully stated and
 348 evaluated in applications and are separate from any parametric modelling assumptions, which may also be
 349 required.

350 For illustration, we consider a specific example [22] where the interest is to investigate the causal effect of the
 351 level of C-reactive Protein (CRP) on CIMT by exploiting random assignment of a genetic variant, G , associated with
 352 CRP. Here CRP is referred to as the exposure, X , CIMT as the outcome, Y , and G as the instrumental variable (or
 353 instrumental gene).

354 [Figure 4 Here]

355 Note that the IV causal graph permits unmeasured variables that may influence both the exposure CRP and the
 356 outcome CIMT, here denoted by U . The IV assumptions encoded by the causal graph in Fig.4a can be written
 357 formally as follows

- 358 1. CIMT does not influence CRP, but the reverse may be true.
- 359 2. Relevance: The instrumental gene is associated with the level of CRP.
- 360 3. Exclusion restriction: The instrumental gene may affect CIMT only through its effect on CRP.
- 361 4. Unconfoundedness: There is no variable, observed or otherwise, which is a common cause of the
 362 instrumental gene and CIMT.

363 For assumption 1, domain specific knowledge is generally required to defend the $X \rightarrow Y$ causal relationship over
 364 the alternative, $Y \rightarrow X$. For this example, it is usually assumed that proteins causally influence disease outcomes,

rather than the other way round. Collectively, assumptions 2 to 4 are known as the IV assumptions as they describe the relationship between the IV and the variables U , X and Y . In a randomised control trial (RCT), where the IV is the randomly assigned treatment group, these assumptions are more simple to justify, since the randomisation process is known, and we can engineer the randomised treatment so that it is (a) associated with the exposure, and (b) does not influence the outcome except through the exposure, although in some settings justification of the exclusion restriction remains challenging.

In the MR setting, we justify the relevance condition (assumption 2) by choosing instrumental genes following a GWAS analysis. In practice, several candidate instrumental genes are often used to support or discredit the evidence of a single one. The exclusion restriction (assumption 3) is, however, more problematic as genetic variants may have independent pleiotropic effects on multiple phenotypes. Pleiotropic effects violate the exclusion restriction by introducing alternative paths of the type $G \rightarrow Y$.

Recent developments in MR do allow for some limited pleiotropy, such as MR-Egger[3], which permits a direct path from $G \rightarrow Y$ in Two-Sample studies (under specific assumptions), and the MRGxE method[36], which allows for pleiotropic ‘Gene-by-environment’ interactions provided they reside on the $G \rightarrow X$ path. Selection of instrumental genes in MR is, however, an open topic of debate, both in terms of statistical and biological considerations[37]. Recent statistical work considers variable selection methods, such as the Lasso, to select IVs[46]. Whilst the exclusion restriction cannot be proven, it may sometimes be possible to show that they are inconsistent with prior evidence. Methods for doing so include leveraging prior causal assumptions, identifying modifying subgroups, or by use of instrument inequality tests[15].

Unconfoundedness (assumption 4) prohibits edges of the type $U \rightarrow G$, which is reasonably well justified on the basis of Mendelian inheritance. As in Section 3, however confounding by ancestry violates this assumption, since unobserved ancestry variables, C , may causally influence the outcome through their effect on other genetic variants as well as causally influencing the instrumental gene itself. Ancestrally heterogeneous populations are therefore known to violate the unconfoundedness in MR, and practitioners are recommended where possible to use homogeneous cohorts, thought to be in HWE.

A modified causal graph, which relaxes the IV assumptions to allow for confounding by ancestry, and limited pleiotropic effects, can be seen in Fig.4b. This graph represents a more general set of causal assumptions, to emphasise the assumptions of the IV graph. The standard IV graph may be recovered by removing arrows from the modified causal graph, or in other words, by assuming certain null causal relationships.

If only the $G \rightarrow Y$ arrow is removed from the causal graph in Fig.4b (i.e. G has no pleiotropic effect on Y) then G may be used as a *conditional instrumental variable*, assuming one collects adequate data on the other genetic variants G^* . In a *conditional instrumental variable* analysis, the gene G acts as an instrumental variable after conditioning on G^* in the models for X and for Y . This conditioning has the effect of blocking the open paths: $G \leftarrow C \rightarrow G^* \rightarrow X$ and $G \leftarrow C \rightarrow G^* \rightarrow Y$. Once blocked, unconfoundedness is no longer violated so G again acts as an instrument, allowing for valid MR analysis with ancestrally heterogeneous cohorts. Conditioning on G^* may be achieved using the methods in Section 3.1.

Violation of any of the IV assumptions would result in invalid causal estimates. We refer the interested reader to [41] for a comprehensive discussion of the challenges faced by MR studies when justifying the IV assumptions and on how to conduct sensitivity analyses.

4.2 Survivor Bias in Mendelian Randomisation

One setting where causal graphs are especially useful for evaluating MR assumptions is in the use of genetic instruments to assess survival biases. Here we consider the example given in [42], namely where an MR analysis of the effect of vitamin D levels on mortality is performed using a cohort of ancestrally homogenous, genotyped

408 individuals between the ages of 40 and 71 years old. Using causal graphs, we show how survivor bias may be
 409 introduced because recruitment to the cohort depends on an individual having survived long enough to be
 410 eligible for recruitment.

411 Selection to the cohort depends on T , the lifetime of an individual, being larger than some index time, T_0 . By
 412 definition, an index time is actually assigned only to individuals in the cohort (who are indexed at some point
 413 between the ages of 40 and 71), however, we could imagine that individuals outside the cohort could also be
 414 given an index time, for example by sampling from the birth register. As before, we will denote selection to the
 415 cohort by the variable S , with $S = 1$ for all individuals in the cohort.

416 Let D be the level of vitamin D at index and assume that it captures the effect on lifetime of an individual's entire
 417 exposure to vitamin D since birth. This assumption is implicit in all MR studies, since to not assume it would
 418 generally violate the exclusion assumption, in the sense that we could imagine an additional variable (e.g.
 419 adolescent vitamin D level) which causally influences the vitamin D level recorded at index, as well as the lifetime
 420 of the individual directly.

421 Finally, we shall assume that an appropriate genetic instrument (e.g. filaggrin genotype) has been recorded, which
 422 we shall denote, G , and assume is randomised by Mendelian inheritance, since the cohort is homogenous. As with
 423 the standard MR causal graph, we shall permit unmeasured confounding variables which might causally influence
 424 both vitamin D level and lifetime. Our causal assumptions for this example are represented by the causal graph in
 425 Fig.5a. In this example, S , is a variable which we have no choice but to condition on, hence we must be very careful
 426 to consider collider stratification biases, as discussed in Section 2.

427 [Figure 5 Here]

428 We see that S is a descendent of D , due to the $D \rightarrow T \rightarrow S$ path, and that D is also a collider on the path $G \rightarrow D \leftarrow U$.
 429 Hence, by selecting only individuals who have survived, the ancestors of D (namely G and U) become associated.
 430 This violates the exclusion assumption, since association between G and T may arise from either the causal path G
 431 $\rightarrow D \rightarrow T$ or from the path $U \rightarrow T$, where U is associated with G .

432 The association induced by conditioning on selection is illustrated by the dashed line in Fig.5b. Recent work
 433 proposes various strategies for MR estimation under survivor bias, using a semi-parametric additive hazard
 434 model[42], similar to the canonical Cox proportional hazards model. This relates to similar work on MR for
 435 censored survival outcomes[39].

436 Interestingly, however, this problem of survivor bias disappears when testing the null hypothesis that D has no
 437 causal influence on T . Under this null hypothesis, there is, by definition, no $D \rightarrow T$ arrow, hence G is not an
 438 ancestor of T and no association between G and U is induced.

439 5 Conclusion

440 We have demonstrated, through examples of the most common analytical techniques employed in genetic studies,
 441 that a causal inference framework, and in particular the use of causal graphs, allows the analyst to (i) to represent
 442 their knowledge of the causal relationships involved in the question at hand, and (ii) use the rules of d-separation,
 443 to query the assumptions under which popular genetic analysis methods lead to causal interpretations.

444 Causal graphs may also inform intuition regarding the advantages and limitations of different analytical
 445 techniques from the outset and are useful in deciding which variables should (and should not) be conditioned on
 446 to avoid subtle confounding and selection biases, arising from study design or data collection methods.
 447 Recognising these biases is necessary so that unbiased estimates of causal effects may be obtained.

448 Despite their utility, causal inference methods, and in particular causal graphs, do have limitations. Unavoidably,
 449 expert knowledge is still required to elicit and defend causal assumptions, and it is recommended that sensitivity
 450 analyses be conducted to explore the consequences that departures from causal assumptions have on estimates

451 of interest. Moreover, even in situation where causal assumptions may be well justified, correct specification of
 452 regression models remains an issue. These regression models may be required to adequately block open paths. In
 453 Section 3.1, we saw that specification of regression models is especially difficult in genomic applications, where
 454 dimensionality reduction strategies are required to condition on high-dimensional genetic information. These
 455 strategies come with their own model validity assumptions, separate from the causal ones we have discussed.

456 We reiterate that causal graphs are not the only framework for representing causal assumptions and deriving
 457 statistical dependencies, and that this can be done within other causal frameworks, for example[30]. We hope this
 458 review may, however, contribute to the discourse of GWAS and MR analyses by allowing causal assumptions to be
 459 explicitly acknowledged and communicated in a transparent and intuitive manner. Finally, since causal graphs are
 460 common in the communication and development of novel analytical methods, we hope to have contributed to a
 461 better understanding of them, thus helping the adoption of new analytical methods in the future.

462 References

- 463 [1] Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and
 464 Zondervan, K. T. Data quality control in genetic case-control association studies. *Nature Protocols* 5, 9 (2010),
 465 1564–1573.
- 466 [2] Bareinboim, E., Tian, J., and Pearl, J. Recovering from Selection Bias in Causal and Statistical Inference Elias.
 467 In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014), AAAI Press, pp. 2410 –
 468 2416.
- 469 [3] Bowden, J., Smith, G. D., and Burgess, S. Mendelian randomization with invalid instruments: Effect
 470 estimation and bias detection through Egger regression. *International Journal of Epidemiology* 44, 2 (2015),
 471 512–525.
- 472 [4] Brady, H. E. *Oxford Handbooks Online Causation and Explanation in Social Science 1 Causality*. No. April 2017.
 473 2013.
- 474 [5] Browning, B. L., and Browning, S. R. A fast, powerful method for detecting identity by descent. *American*
 475 *Journal of Human Genetics* 88, 2 (2011), 173–182.
- 476 [6] Buniello, A., MacArthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J.,
 477 Mountjoy, E., Sollis, E., Suveges, D., Vrousou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S.,
 478 Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F., and Parkinson, H.
 479 The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary
 480 statistics 2019. *Nucleic Acids Research* 47, D1 (2019), D1005–D1012.
- 481
- 482 [7] Burgess, S., and Thompson, S. G. *Mendelian Randomization: Methods for Using Genetic Variants in Causal*
 483 *Estimation*. Chapman & Hall/CRC, 2015.
- 484 [8] Chuang, Y. F., Tanaka, T., Beason-Held, L. L., An, Y., Terracciano, A., Sutin, A. R., Kraut, M., Singleton, A. B.,
 485 Resnick, S. M., and Thambisetty, M. FTO genotype and aging: Pleiotropic longitudinal effects on adiposity,
 486 brain function, impulsivity and diet. *Molecular Psychiatry* 20, 1 (2015), 133–139.
- 487 [9] Didelez, V., Meng, S., and Sheehan, N. A. Assumptions of IV methods for observational epidemiology.
 488 *Statistical Science* 25, 1 (2010), 22–40.
- 489 [10] Didelez, V., and Sheehan, N. A. Mendelian randomization as an instrumental variable approach to causal
 490 inference. *Statistical Methods in Medical Research* 16, 4 (2007), 309–330.
- 491 [11] Ding, E. L., Song, Y., Manson, J. E., Hunter, D. J., Lee, C. C., Rifai, N., Buring, J. E., Gaziano, J. M., and Liu, S.
 492 Sex HormoneBinding Globulin and Risk of Type 2 Diabetes in Women and Men. *New England Journal of*
 493 *Medicine* 361, 12 (sep 2009), 1152–1163.

- 494 [12] Ganef, I. M. M., Bos, M. M., van Heemst, D., and Noordam, R. BMI-associated gene variants in
495 FTO and cardiometabolic and brain disease: obesity or pleiotropy? . *Physiological Genomics* 51, 8 (2019),
496 311–322.
- 497 [13] Glymour, M. M. Using causal diagrams to understand common problems in social epidemiology. In *In*
498 *Methods in Social Epidemiology* (2006), John Wiley and Sons, pp. 393–428.
- 499 [14] Glymour, M. M., and Spiegelman, D. Evaluating public health interventions: 5. Causal inference in public
500 health research-do sex, race, and biological factors cause health outcomes? *American Journal of Public Health*
501 107, 1 (2017), 81–85.
- 502 [15] Glymour, M. M., Tchetgen, E. J., and Robins, J. M. Credible mendelian randomization studies:
503 Approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology* 175, 4
504 (2012), 332–339.
- 505 [16] Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem,
506 P., Thorlacius, S., Gylfason, A., Steinberg, S., Helgadóttir, A., Ingason, A., Steinthorsdóttir, V., Olafsdóttir, E. J.,
507 Olafsdóttir, G. H., Jonsson, T., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Pedersen, O., Aben, K.
508 K., Witjes, J. A., Swinkels, D. W., Heijer, M. D., Franke, B., Verbeek, A. L., Becker, D. M., Yanek, L. R., Becker,
509 L. C., Tryggvadóttir, L., Rafnar, T., Gulcher, J., Kiemeneý, L. A., Kong, A., Thorsteinsdóttir, U., and Stefansson,
510 K. Many sequence variants affecting diversity of adult human height. *Nature Genetics* 40, 5 (2008), 609–615.
511
- 512 [17] Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J.,
513 Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M.,
514 Davey Smith, G., Gaunt, T. R., and Haycock,
515 P. C. The mr-base platform supports systematic causal inference across the human phenome. *eLife* 7 (may
516 2018), e34408.
- 517 [18] Hernan, M., and Robins, J. *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC, 2020.
- 518 [19] Hoffman, G. E. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and
519 Extensions. *PLoS ONE* 8, 10 (oct 2013), e75707.
- 520 [20] Holland, P. W. Statistics and Causal Inference. *Journal of the American Statistical Association* 81, 396 (dec
521 1986), 945–960.
- 522 [21] Imbens, G. W. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical
523 practice in economics. *NBER Working Paper No.w26104* (2019).
- 524 [22] Kivimaki, M., Lawlor, D. A., Smith, G. D., Kumari, M., Donald, A., Britton, A., Casas, J. P., Shah, T., Brunner,
525 E., Timpson, N. J., Halcox, J. P., Miller, M. A., Humphries, S. E., Deanfield, J., Marmot, M. G., and Hingorani, A.
526 D. Does high C-reactive protein concentration increase atherosclerosis? The Whitehall II study. *PLoS ONE* 3,
527 8 (2008), 1–8.
- 528 [23] Lin, D. Y., and Zeng, D. Proper analysis of secondary phenotype data in case-control association studies.
529 *Genetic Epidemiology* 33, 3 (apr 2009), 256–265.
- 530 [24] Lippert, C., Quon, G., Kang, E. Y., Kadie, C. M., Listgarten, J., and Heckerman, D. The benefits of selecting
531 phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports* 3 (may 2013),
532 1815.
- 533 [25] Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Hua Zhao, J., Prokopenko, I., Inouye, M., Freathy, R. M.,
534 Attwood, A. P., Beckmann, J. S., Berndt, S. I., Bergmann, S., Bennett, A. J., Bingham, S. A., Bochud, M., Brown,
535 M., Cauchi, S., Connell, J. M., Cooper, C., Davey Smith, G., Day, I., Dina, C., De, S., Dermitzakis, E. T., Doney,
536 A. S., Elliott, K. S., Elliott, P., Evans, D. M., Sadaf Farooqi, I., Froguel, P., Ghorí, J., Groves, C. J., Gwilliam, R.,
537 Hadley, D., Hall, A. S., Hattersley, A. T., Hebebrand, J., Heid, I. M., Herrera, B., Hinney, A., Hunt, S. E., Jarvelin,
538 M. R., Johnson, T., Jolley, J. D., Karpe, F., Keniry, A., Khaw, K. T., Luben, R. N., Mangino, M., Marchini, J.,
539 McArdle, W. L., McGinnis, R., Meyre, D., Munroe, P. B., Morris, A. D., Ness, A. R., Neville, M. J., Nica, A. C.,

- 540 Ong, K. K., O’Rahilly, S., Owen, K. R., Palmer, C. N., Papadakis, K., Potter, S., Pouta, A., Qi, L., Randall, J. C.,
541 Rayner, N. W., Ring, S. M., Sandhu, M. S., Scherag, A., Sims, M. A., Song, K., Soranzo, N., Speliotes, E. K.,
542 Syddall, H. E., Teichmann, S. A., Timpson, N. J., Tobias, J. H., Uda, M., Ganz Vogel, C. I., Wallace, C.,
543 Waterworth, D. M., Weedon, M. N., Willer, C. J., Wraight, V. L., Yuan, X., Zeggini, E., Hirschhorn, J. N.,
544 Strachan, D. P., Ouwehand, W. H., Caulfield, M. J., Samani, N. J., Frayling, T. M., Vollenweider, P., Waeber, G.,
545 Mooser, V., Deloukas, P., McCarthy, M. I., Wareham, N. J., Barroso, I., Jacobs, K. B., Chanock, S. J., Hayes, R.
546 B., Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H. E., Kraft, P., Hankinson, S. E., Hunter, D. J., Hu,
547 F. B., Lyon, H. N., Voight, B. F., Ridderstrale, M., Groop, L., Scheet, P., Sanna, S., Abecasis, G. R., Albai, G.,
548 Nagaraja, R., Schlessinger, D., Jackson, A. U., Tuomilehto, J., Collins, F. S., Boehnke, M., and Mohlke, K. L.
549 Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics* 40, 6
550 (2008), 768–775.
- 551
- 552 [26] Munafo, M. R., Tilling, K., Taylor, A. E., Evans, D. M., and Davey Smith, G. Collider scope: when selection bias
553 can substantially influence observed associations. *International Journal of Epidemiology* 47, 1 (feb 2018),
554 226–235.
- 555 [27] Pearl, J. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence* 29, 3 (sep 1986), 241–
556 288.
- 557 [28] Pearl, J. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.
- 558 [29] Pearl, J. *Causality: Models, Reasoning and Inference*. 2000.
- 559 [30] Robins, J. M. Data, design, and background knowledge in etiologic inference. *Epidemiology* 12, 3 (2001), 313–
560 320.
- 561 [31] Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American*
562 *Statistical Association* 100, 469 (2005), 322–331.
- 563 [32] Sheehan, N. A., and Didelez, V. Human Genetics Epidemiology, Genetic Epidemiology and Mendelian
564 Randomisation: more need than ever to attend to detail? Epidemiology, Genetic Epidemiology and Mendelian
565 Randomisation: more need than ever to attend to detail? *Human Genetics*, 0123456789 (2018).
- 566 [33] Song, X., Ionita-Laza, I., Liu, M., Reibman, J., and Wei, Y. A general and robust framework for secondary
567 traits analysis. *Genetics* 202, 4 (2016), 1329–1343.
- 568 [34] Speed, D., and Balding, D. J. Relatedness in the post-genomic era : is it still useful ? *Nature Publishing Group*,
569 November (2014), 1–12.
- 570 [35] Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. Improved heritability estimation from genome-wide
571 SNPs. *American Journal of Human Genetics* 91, 6 (2012), 1011–1021.
- 572 [36] Spiller, W., Slichter, D., Bowden, J., and Davey Smith, G. Detecting and correcting for bias in Mendelian
573 randomization analyses using Gene-by-Environment interactions. *International Journal of Epidemiology*
574 (2018), 1–11.
- 575 [37] Swerdlow, D. I., Kuchenbaecker, K. B., Shah, S., Sofat, R., Holmes, M. V., White, J., Mindell, J. S., Kivimaki, M.,
576 Brunner, E. J., Whittaker, J. C., Casas, J. P., and Hingorani, A. D. Selecting instruments for Mendelian
577 randomization in the wake of genome-wide association studies. *International Journal of Epidemiology* 45, 5
578 (2016), 1600–1616.
- 579
- 580 [38] Tchetgen Tchetgen, E. J. A general regression framework for a secondary outcome in case-control studies.
581 *Biostatistics* 15, 1 (2014), 117–128.
- 582 [39] Tchetgen Tchetgen, E. J., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M.
583 Instrumental Variable Estimation in a Survival Context. *Epidemiology* 26, 3 (may 2015), 402–410.

- 584 [40] Vandenbroucke, J. P., Broadbent, A., and Pearce, N. Causality and causal inference in epidemiology: The need
585 for a pluralistic approach. *International Journal of Epidemiology* 45, 6 (2016), 1776–1786.
- 586 [41] VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., and Kraft, P. Methodological challenges in mendelian
587 randomization. *Epidemiology (Cambridge, Mass.)* 25, 3 (2014), 427.
- 588 [42] Vansteelandt, S., Dukes, O., and Martinussen, T. Survivor bias in Mendelian randomization analysis.
589 *Biostatistics* 19, 4 (2018), 426–443.
- 590 [43] Vilhjalmsón, B. J., and Nordborg, M. The nature of confounding in genome-wide association studies. *Nature*
591 *Reviews Genetics* 14, 1 (2013), 1–2.
- 592 [44] Weedon, M. N., Lettre, G., Freathy, R. M., M, C., Voight, B. F., Perry, J. R. B., Elliott, K. S., Guiducci, C., Shields,
593 B., Zeggini, E., Lango, H., Lyssenko, V., Timpson, N. J., Burt, N. P., Rayner, N. W., Ardlie, K., Tobias, J. H.,
594 Ness, A. R., and Ring, S. M. UKPMC Funders Group UKPMC Funders Group Author Manuscript A common
595 variant of HMG2 is associated with adult and childhood height in the general population. *October* 39, 10
596 (2011), 1245–1250.
- 597 [45] Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J.,
598 Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A.
599 J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D.,
600 Chen, W. M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor,
601 D. A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J.,
602 Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., and Abecasis, G.
603 R. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature*
604 *Genetics* 40, 2 (2008), 161–169.
- 605 [46] Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. On the Use of the Lasso for Instrumental
606 Variables Estimation with Some Invalid Instruments. *Journal of the American Statistical Association* 1459
607 (2018).
- 608 [47] Zhou, X., and Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association
609 studies. *Nature Methods* 11, 4 (apr 2014), 407–409.

612 List of Figures

- 613 1 Causal graph representing the causal assumptions between a patients FTO gene variant, F, body mass index, B,
614 and cerebral blood flow, C. [Page 3]
- 615 2 (a) Causal graph representing the causal assumptions between a specific gene of interest, G, Type 2 diabetes
616 status, D, SHBG level, H, and selection to the cohort, S. (b) Causal graph when considering only individuals
617 in the cohort ($S = 1$). The selection variable has been conditioned on, indicated by the box around it. The induced
618 association between G and H is represented by the dashed line. [Page 4]
- 619 3 Causal graphs for GWAS analysis. Graph (a) shows the basic causal GWAS model, where the phenotype of
620 interest, Y , is dependent on the gene of interest, G, and some other environmental factors, E. Graph (b)
621 accounts for confounding by the ancestry of the individual, C, which affects the gene of interest, and the
622 remaining genes, G^* . This modified graph assumes that a polygenic trait, Y , depends on both the gene of
623 interest, and the remaining genes. By convention, unobserved (or latent) variables, such as the ancestry
624 variable, C, are circled. [Page 7]
- 625 4 Causal graphs for MR analysis. Graph (a) shows the traditional IV causal graph, where the gene, G, acts as an IV
626 for the $X \rightarrow Y$ relationship of interest, itself confounded by the unmeasured variable, U. Graph (b) shows
627 modifications to graph (a) which relax assumptions by allowing for confounding by ancestry, and some
628 pleiotropic effects. [Page 9]
- 629 5 Causal graphs for MR analysis of a survival outcome. Graph (a) shows the instrumental gene, G, acts as an IV
630 for the $D \rightarrow T$ relationship of interest where D is vitamin D level and T is lifetime. Graph (b), however, shows

631 that conditioning on selection to the cohort, S , which depends on an individual surviving to index time T_0 ,
 632 introduces associations between G and U which violate the IV exclusion assumption. [Page 11]

633

634 **Appendix A Linear Mixed Models**

635 Consider again the linear model in Eq.2. When the model parameters are estimated by OLS, one effectively makes
 636 no prior assumptions about the parameter values, other than that they are fixed to some true unknown value.
 637 Considering P as a random effect, however, we impose, in a Bayesian sense, a normally distributed prior for
 638 $\gamma \sim \mathcal{N}_p(0, \sigma_g^2 I_p)$, where I_p is a p by p identity matrix, σ_g^2 is a hyper parameter and $\mathcal{N}_p(\mu, \Sigma)$ is a p -multivariate
 639 normal distribution with mean μ and variance Σ .

640 By making this prior assumption we arrive at a LMM, which may be written as a model for the full n -dimensional
 641 observed phenotype vector, Y . Here bold notation is used to refer to vector (or matrix) quantities with n entries
 642 (or rows), each representing a single individual in the cohort. Again I_n is the n by n identity matrix,

$$643 \quad Y \sim \mathcal{N}_n(\alpha G + E\beta, \sigma_g^2 K + \sigma_e^2 I_n) \quad (3)$$

644 where $K = PP^T$ and P is an n by q matrix where each row represents the vector of PCs for a particular individual.
 645 The n by n matrix, K is referred to as the genetic similarity matrix, since the entry K_{ij} is a measure of the genetic
 646 similarity between the i^{th} and j^{th} individuals in the cohort, obtained by comparing their PCs. In general one is not
 647 restricted to using PCs to define the genetic similarity matrix. In fact several different methods can be expressed
 648 by the LMM equation above, using different measures of genetic similarity [19].

649 **Measures of Genetic Similarity**

650 Methods for measuring genetic similarity may be broadly separated into two categories: Those related to the
 651 Principal Component Analysis (Principal Components like), and those where some biologically motivated
 652 measure of genetic similarity is made. We will refer to methods of the latter type as Identity By Descent like, since
 653 they often measure similarity by finding genetic regions which are thought to be identical by descent in two
 654 individuals. A brief overview of these approaches is provided below.

655 **Principal Component like**

656 In a conventional PC analysis, the variables from which PCs are constructed (in this case the SNP values) are
 657 standardised. Variations exist, however, in how the SNPs are selected, how they are weighted in the
 658 standardisation step, and how the resultant PCs are selected. These include:

- 659 1. Selection of which SNPs to use for PC analysis: It is possible to include all available SNPs, however, it has
 660 been suggested that only variants thought to be causally related to the phenotype of interest should be
 661 included [43, 24], since these are the ones which lie on the causal pathway between C and Y . The process of
 662 selecting SNPs is known as pruning or thinning.
- 663 2. The choice of SNP dependent scaling constant before constructing PCs: The intuition behind scaling the SNP
 664 value is that sharing a rare variant is greater evidence of common ancestry than sharing a common variant.
 665 Scaling values are often estimates of the SNP standard deviation. This may be estimated by the sample
 666 standard deviation or using the standard deviation under the Hardy-Weinberg equilibrium model.
 667 It has also been suggested that, rather than pruning SNPs, SNPs should be weighted according to their
 668 degree of LD, to account for replication of causal information by neighbouring, imputed, SNPs in LD [35].

669 Their proposal uses weights, chosen such that SNPs with high LD are down-weighted. This is implemented
670 in their LDAK software package.

671 3. The number of PC dimensions chosen for inclusion in the linear model: This is often determined using
672 heuristic measures. Each successive PC accounts for a smaller amount of genetic variation in the chosen
673 SNPs. Most methods use estimates for the proportion of variance explained by each PC, for example
674 selecting PCs to exceed some threshold of the total proportion of variance explained, or else choosing an
675 arbitrary number of PCs.

676 In the LMM, it is possible to include all PCs. This is the choice made in the GEMMA software package [47].
677 This approach is equivalent to measuring the covariance between two individuals based on all chosen SNPs.

678 **Identity By Descent like**

679 Traditional measures for relatedness pre-date modern genomic study and were originally used to study trait
680 inheritance within pedigrees. Using known pedigree information one can construct the probabilities that genomic
681 regions of two individuals are identical-by-descent (IBD) from a recent common ancestor ('recent' in so far as it is
682 assumed that there is no intermediate mutation or recombination event).

683 Pedigree based relatedness measures are broadly obsolete in modern genomic analysis for several reasons [34]:
684 (i) When studying natural populations pedigree information is often unavailable or insufficient to account for
685 population structure. (ii) Even when pedigree information is available, it is usually unrealistic to assume that
686 pedigree founders have zero genetic similarity. (iii) The relatedness of any two individuals tends towards one, as
687 the size of the pedigree is increased.

688 Rather than using pedigree information to estimate IBD probabilities, modern theories instead measure IBD by
689 appealing to SNP data itself. These methods generally examine the length and frequencies of similar genomic
690 regions in two individuals and are based on biochemical theories regarding the process by which gametes divide
691 and recombine from two parents. Examples include: FastIBD [5], which estimates the frequencies of shared
692 haplotype distributions; and shared segment detection in PLINK [1]. Reviewing these methods is beyond the
693 scope of this review.

Table 1: All possible paths between three variables (A, B, C), with a brief description and additional terminology for the intermediate variable C

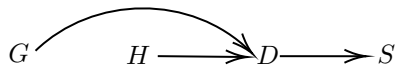
Path	Description	Terminology for the variable C
$A \rightarrow C \rightarrow B$	A causes B (through C)	Mediator
$A \leftarrow C \leftarrow B$	B causes A (through C)	Mediator
$A \leftarrow C \rightarrow B$	A and B share a common cause C	Confounder
$A \rightarrow C \leftarrow B$	A and B both cause C	Collider

Table 2: Summary of the rules of d-separation for all possible paths containing three variables. The two additional columns describe the statistical dependence of A and B before and after conditioning on the intermediate variable C .

Path	Before conditioning on C	After conditioning on C
$A \rightarrow C \rightarrow B$	open	closed
$A \leftarrow C \leftarrow B$	open	closed
$A \leftarrow C \rightarrow B$	open	closed
$A \rightarrow C \leftarrow B$	closed	open



Fig.1 Causal graph representing the causal assumptions between a patients FTO gene variant, F, body mass index, B, and cerebral blood flow, C.



(a)



(b)

Fig.2 (a) Causal graph representing the causal assumptions between a specific gene of interest, G, Type 2 diabetes status, D, SHBG level, H, and selection to the cohort, S. (b) Causal graph when considering only individuals in the cohort ($S = 1$). The selection variable has been conditioned on, indicated by the box around it. The induced association between G and H is represented by the dashed line.

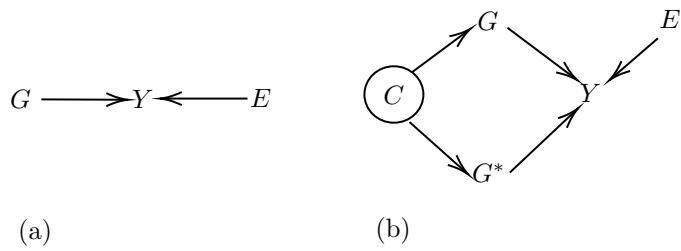
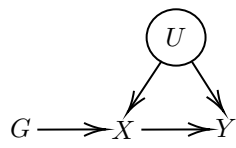
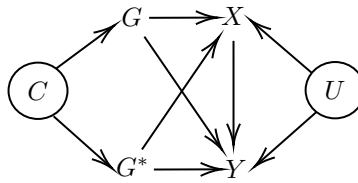


Fig.3 Causal graphs for GWAS analysis. Graph (a) shows the basic causal GWAS model, where the phenotype of interest, Y , is dependent on the gene of interest, G , and some other environmental factors, E . Graph (b) accounts for confounding by the ancestry of the individual, C , which affects the gene of interest, and the remaining genes, G^* . This modified graph assumes that a polygenic trait, Y , depends on both the gene of interest, and the remaining genes. By convention, unobserved (or latent) variables, such as the ancestry variable, C , are circled.

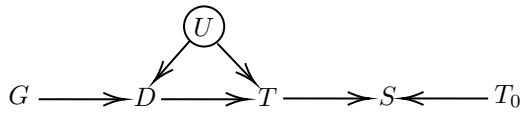


(a)

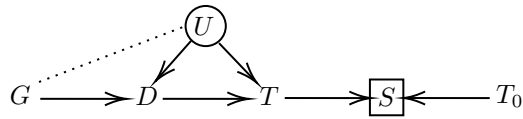


(b)

Fig.4 Causal graphs for MR analysis. Graph (a) shows the traditional IV causal graph, where the gene, G , acts as an IV for the $X \rightarrow Y$ relationship of interest, itself confounded by the unmeasured variable, U . Graph (b) shows modifications to graph (a) which relax assumptions by allowing for confounding by ancestry, and some pleiotropic effects.



(a)



(b)

Fig.5 Causal graphs for MR analysis of a survival outcome. Graph (a) shows the instrumental gene, G , acts as an IV for the $D \rightarrow T$ relationship of interest where D is vitamin D level and T is lifetime. Graph (b), however, shows that conditioning on selection to the cohort, S , which depends on an individual surviving to index time T_0 , introduces associations between G and U which violate the IV exclusion assumption.