Identification of Susceptibility Variants in Prion Disease by Integrative Causal Analysis

Angelos P. Armen

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Neurodegenerative Disease

Institute of Neurology

University College London

May 25, 2017

I, Angelos P. Armen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Prion diseases are lethal neurodegenerative disorders caused by infectious proteins called prions. All known susceptibility variants in human prion disease are found in the prion protein gene (PRNP), but there is evidence that additional, non-PRNP susceptibility loci exist. Genome-wide association studies, an exome-sequencing study and an exome-array study have been conducted by the MRC Prion Unit in order to identify those loci. None of these studies have resulted in novel discoveries yet. Data integration could overcome the pitfalls of single-dataset analysis to discover novel susceptibility factors. In this project, Integrative Causal Analysis was adopted as a framework for data integration with the aim to identify all causal relationships between variants and prion disease that are consistent with all prion datasets and prior biological knowledge. Firstly, a theory of causal discovery from genetic datasets was formulated and causal discovery was applied to the datasets from the studies mentioned above. Secondly, an algorithm for causal meta-analysis of genetic datasets with overlapping sets of variants was designed and applied to a combination of the datasets in order to increase the power of learning causal relationships. Thirdly, a variant-filtering approach based on causal prior knowledge was devised as another method to increase power: Publicly available biological data and prior knowledge were integrated into a directed graph whose nodes comprise the disease and molecular entities in the cell and are associated with genomic regions, and whose edges denote causation. Candidate causal variants were subsequently identified from the ancestors (causes) of the disease in the graph. The application of the methods to prion disease resulted in a number of candidate susceptibility variants to be further investigated. The methods are also applicable to other diseases

and have the potential to lead to novel discoveries in those diseases.

Acknowledgements

First of all, I would like to express my deepest gratitude to my primary supervisor, Holger Hummerich, for trusting me with this project and supporting me in numerous ways, and thank my secondary supervisor, Parmjit Jat, for his advice and guidance. Second of all, I would like to thank Simon Mead for trusting me with the analysis of the exome-array data and supporting me with the genetics, James Uphill for helping me with processing and association analysis of genetic datasets, and Penny Norsworthy for conducting the Sanger sequencing of the exome-array hits. In addition, I am grateful to my fellow PhD students in the Prion Unit, especially Xun Yu Choong, Justin Tosh, Billy West, and Charlotte Ridler for their biological advice, suggestions, and general support. Furthermore, I would like to thank Peter D'Eustachio and Sandra Orchard for replying to my inquiries about Reactome and IMEx, respectively, and William Rayner for generating the A/B to TOP allele mapping for the NeuroX chip. I am grateful to John Collinge and the MRC for the financial support I received. I thank Robin Evans as well for his helpful suggestions while I was working on the corrections of my thesis. Last but not least, I would like to thank my family for their continued support through the years.

Contents

List of Figures

List of Tables

Acronyms

AUC area under the curve.

BBE black box event.

BN Bayesian network.

BSE bovine spongiform encephalopathy.

CBN causal Bayesian network.

CFC causal faithfulness condition.

CHARGE Cohorts for Heart and Aging Research in Genomic Epidemiology.

ChEBI Chemical Entities of Biological Interest.

CJD Creutzfeldt-Jakob disease.

CMC causal Markov condition.

COR causal odds ratio.

CPD conditionally probability distribution.

CPKG causal prior knowledge graph.

CSP constraint satisfaction problem.

DAG directed acyclic graph.

ELISA enzyme-linked immunosorbent assay.

eQTL expression quantitative trait locus.

EWAS entity with accessioned sequence.

FDR false discovery rate.

FPR false positive rate.

FWER family-wise error rate.

GEE genome-encoded entity.

GERAD Genetic and Environmental Risk in Alzheimer's Disease.

GO Gene Ontology.

GOA Gene Ontology Annotation.

GPI glycosylphosphatidyl inositol.

GSS Gerstmann-Straussler-Scheinker.

GWAS genome-wide association study.

HGNC HUGO Gene Nomenclature Committee.

iCJD iatrogenic Creutzfeldt-Jakob disease.

IMEx International Molecular Exchange.

INCA integrative causal analysis.

kb kilobases.

LD linkage disequilibrium.

LRG Locus Reference Genomic.

MAF minor allele frequency.

MAG maximal ancestral graph.

MBM meat-and-bone meal.

MTI miRNA-target interaction.

NGS next-generation sequencing.

OR odds ratio.

PAG partial ancestral graph.

PCG pairwise causal graph.

PIR pseudo-independent relation.

PMN physical module network.

PPI protein-protein interaction.

PSI Proteomics Standards Initiative.

PSI-MI Proteomics Standards Initiative - Molecular Interaction.

PSI-MOD Proteomics Standards Initiative - Protein Modification.

RISC RNA-induced silencing complex.

RLE reaction-like event.

RR relative risk.

sCJD sporadic Creutzfeldt-Jakob disease.

SEM structural equation model.

SNP single nucleotide polymorphism.

SVM support vector machine.

UTR untranslated region.

vCJD variant Creutzfeldt-Jakob disease.

Chapter 1

Introduction

1.1 Genome-wide association studies

Mendelian diseases are caused by mutations in a single gene. In contrast, *complex* diseases are caused by a combination of genetic and/or environmental factors. The completion of *The International HapMap Project* and the advent of high-throughput *genotyping chips* made it possible to conduct *genome-wide association studies (GWASs)* of complex diseases [McCarthy et al., 2008]. The International HapMap Project [Gibbs et al., 2003] genotyped common *single nucleotide polymorphisms (SNPs)*, which occur in a population with a frequency of at least 1-5%, in individuals from different human populations. Common SNPs close to each other on the same chromosome tend to be associated and form a *haplotype block*; non-randomly associated loci are said to be in *linkage disequilibrium (LD)*. Genotyping chips provide an inexpensive means to assess a large number (currently over four million)[1] of preselected SNPs in a cohort of individuals. A GWAS takes advantage of LD by using chips that assess a set of *tag* common SNPs, interspersed across the genome, and testing their association with a phenotype. To account for multiple testing, usually the *family-wise error rate (FWER)*, defined as the probability of a false positive, is controlled at 0.05.

[1] http://www.illumina.com/techniques/popular-applications/genotyping/whole-genome-genotyping.html

1.2 Identifying causal variants

Correlation does not imply causation. For example, a non-causal variant may be associated with the disease because it is in LD with a causal variant. *Conditional analysis* is usually performed for each associated variant in order to discard (some of the) variants that are definitely not causal. In conditional analysis, one variant is selected at a time and the association of each other variant with the disease conditional on the selected variant is tested; if there is no association, the other variant is not a cause of the disease[2] and is discarded. Variants that are conditionally independent of the disease given a subset of the rest variants with cardinality greater than one are also not causal; these conditional independences are not tested, however, in a conditional analysis.

In contrast to association analysis, where marginal independence tests are performed in order to detect associations, methods for *causal discovery* [Spirtes et al., 2000] use conditional independence tests in order to directly elucidate causal relationships from data. In the case of genetic data, conditional analysis is not required as a post-processing step, because causal discovery automatically discards *all* definitely-non-causal associated variants. Although causal discovery has been applied to genetic data [Han et al., 2010, 2011, Alekseyenko et al., 2011], the (causal) output of the methods is still difficult to interpret biologically, especially in the case of case–control datasets; this is because there is no characterisation of the causal models expected to be learned from genetic data. Furthermore, no multiple-testing correction was performed and no causal effects were reported in those works. While simulation studies of the performance of causal discovery algorithms have been performed using *general* benchmark causal models [Aliferis et al., 2010a], no studies have been conducted using realistic *genetic* causal models.

Causal discovery is not widely used in genetics or other areas. One possible explanation could be that most researchers are unfamiliar with methods for causal discovery; another might be the fact that it is less comprehensible than association analysis.

[2]This is true under the so-called *causal faithfulness condition* (see Section 2.2).

Irrespective of the approach used to identify candidate causal variants, *functional validation* is subsequently performed in order to determine whether the variants are truly causal.

1.3 Finding missing heritability

Despite the fact that GWASs have identified hundreds of disease susceptibility loci, only a small proportion of the estimated heritability of complex diseases can be explained. The reasons behind *missing heritability* are under discussion [Eichler et al., 2010].

Firstly, a large number of common variants may be associated with disease but go undetected because their effect size is too small relative to the available sample size. Several approaches can be used to increase the power to detect disease associations from a GWAS dataset or any other type of genetic dataset:

- *Genotype imputation* refers to using a reference dataset (e.g. a HapMap one) from the same population with a superset of the typed SNPs in order to impute the genotypes of the untyped SNPs [Marchini and Howie, 2010]. Imputed SNPs may display a stronger association with the disease than the typed SNPs.

- Increasing the sample size is the most obvious approach to increase power, but it is not always possible, especially in the case of rare diseases. A *mega-analysis* comprises concatenating datasets obtained using the same or compatible genotyping chips and testing for association in the resulting dataset. Genotype imputation may be used to impute missing genotypes in the concatenated dataset when the original datasets are defined over different sets of SNPs. When datasets are incompatible or data sharing restrictions apply, a *meta-analysis* of GWASs may be conducted, where summary measures such as p-values or effect sizes from the different studies are combined [Evangelou and Ioannidis, 2013]. As in a mega-analysis, genotype imputation may be performed. Meta-analyses have been in fact responsible for the majority of recently identified disease susceptibility variants [Evangelou and Ioannidis,

2013].

- Methods for *pathway association analysis*, which increase power by grouping variants by pathway and testing their joint association with disease have been used to implicate pathways in disease [Wang et al., 2010a].

- The error rate being controlled also affects power. In general, the choice of error rate depends on the relative cost of false positives and false negatives. When even a single false positive is unacceptable, it is appropriate to control the FWER. On the other hand, if some false positives can be tolerated in order to increase power, it is appropriate to control the *false discovery rate (FDR)*. The FDR is loosely defined as the expected proportion of false positives among the rejected hypotheses ("discoveries") (see Section 2.1 for the exact definition). Although FDR is widely used in other scientific areas, it is rarely used in genetic association analysis. This is probably due to the very high cost of functionally validating a discovery, making the possibility of false positives completely unacceptable. However, it might as well be the case that not many practitioners are familiar with the FDR. When lack of power is an issue, it is probably more appropriate to control the FDR instead of the FWER.

- The lack of power could be overcome by easing the multiple-testing burden through decreasing the number of hypotheses by *variant filtering*. One approach to variant filtering is the use of prior knowledge [Ritchie, 2011]. Existing variant-filtering methods that use prior knowledge do not provide, however, a causal interpretation of their results. When filtering the input of causal-discovery algorithms, it would be appropriate to retain candidate *causal* variants, that is, variants from which a potential causal path to the disease is suggested by prior knowledge.

Secondly, missing heritability could be explained by the existence of *epistasis* (loosely defined as interaction between variants) with weak or absent marginal effects, where a set of (common) variants is *jointly* strongly associated with the

disease, but *single* variants are weakly associated or not at all associated with the disease. When the marginal effect of a variant is weak (resp. absent), it might (resp. will) be missed when testing for association of that variant with the disease. The variant, however, might be discovered when testing for joint association of the variant and other variants with the disease. A logistic-regression-based likelihood-ratio test is usually employed to this end [Cordell, 2009]. For a typical GWAS dataset, testing for joint association of every subset of the variants with the disease is computationally intractable. In addition, higher-order tests would require unrealistically large samples to be reliable. Testing all pairs of variants in a medium-sized GWAS dataset is feasible, though [Cordell, 2009]. Beyond joint-association analysis, a large number of statistical and machine-learning methods for epistasis detection with varying computational requirements have been devised [reviewed in Cordell, 2009, Niel et al., 2015].

Thirdly, *rare* variants (with a frequency less than 1% in the population) can also confer disease susceptibility. *Next-generation sequencing (NGS)* technologies [Goodwin et al., 2016] allow for fast and increasingly inexpensive identification of both common and rare SNPs as well as short insertions and deletions (*indels*), in the whole genome (*whole-genome sequencing*), the exome (*exome sequencing*), or a target genomic region (*targeted sequencing*). Single-variant association analysis of NGS data suffers from low power, and therefore *collapsing* methods that summarise the variants within a region (e.g. a gene) before testing for association with the phenotype have been devised [Dering et al., 2011]. Another approach is to prioritise or filter variants based on their estimated deleteriousness [see Cooper and Shendure, 2011, for a review of such methods].

Other explanations of missing heritability include disease-causing *large variants* and *copy-number variations*, which are not readily identified with currently-available technologies, and *transgenerational epigenetic inheritance*, whose mechanisms have only begun to be unraveled [Eichler et al., 2010].

1.4 Data integration

Data integration is a general approach that has the potential to solve the problem of missing heritability. Data integration refers to the combination of data from various sources and providing the end user with a unified view of the data [Lenzerini, 2002]. In life sciences, data integration can be performed in numerous ways for different purposes. For example, meta-analysis is a form of data integration. Variant-filtering approaches may estimate variant deleteriousness based on multiple types of data [e.g. evolutionary, biochemical, and structural; Cooper and Shendure, 2011]. Methods for *gene prioritisation* rank genes based on their similarity to genes already implicated in disease (termed "seed" genes) and were originally developed to rank the genes in a genomic region implicated in disease by linkage analysis [Moreau and Tranchevent, 2012]; gene similarity is usually computed by integrating various data sources. Finally, data integration may also refer to combining heterogeneous high-throughput (e.g. genomic, transcriptomic, and proteomic) datasets from the same or different individuals in order to model a phenotype [Ritchie et al., 2015]. Data integration in life sciences has been reviewed from different viewpoints [Hamid et al., 2009, Lapatas et al., 2015].

From a data-analytic viewpoint, Hamid et al. [2009] view data integration as the process of combining data and biological knowledge from different sources using statistical and bioinformatics tools in order to provide a unified view of the genome; the motivation for data integration is that elucidation of the genome may require more information than is provided by one type of data. Hamid et al. [2009] provide a conceptual data-integration framework which comprises three components— posing the statistical/biological problem, data type, and stage of integration. Identifying the problem (e.g. finding genotype–phenotype associations) is the first step of the analysis. Data used in the analysis are of a similar type if they stem from the same underlying source. An example is GWAS and exome-sequencing datasets, which are both genetic datasets; another example is *RNA-seq* and *microarray* datasets, which are both gene-expression datasets. The data are of a heterogeneous type if they stem from two or more diverse sources. For example, a

dataset used for discovering *expression quantitative trait loci (eQTLs)* (loci that are associated with gene expression) contains both genotypes and gene-expression levels from the same individuals, whilst a different kind of analysis may use separate genetic and gene-expression datasets from different individuals. Finally, integration can be performed at either of three different stages— early, intermediate, or late. Merging the data before the analysis is considered early-stage integration; one example is mega-analysis, where datasets are concatenated before the analysis. If the data are transformed before merging, integration is considered intermediate-stage; when the problem is classification, for example, covariance matrices from each dataset may be combined [Hamid et al., 2009]. If results from each data source are combined, the integration is considered late-stage; meta-analysis is an example of late-stage integration, as it is summary measures from each dataset that are combined.

From a computational viewpoint, Lapatas et al. [2015] define data integration as "the computational solution allowing users, from end user (GUI) to power users (API), to fetch data from different sources, combine, manipulate and re-analyse them as well as being able to create new datasets and share these again with the scientific community". Data-integration frameworks can be classified as *eager* or *lazy*. In the eager approach, the data are copied and stored in a central data warehouse; in the lazy approach, the data are integrated on demand. Each approach has its advantages and shortcomings. For example, the eager approach allows for easy replication of an analysis; however, the data need to be manually updated (or by an automated pipeline) when needed. In the lazy approach, the most recent data are used; this does not allow, however, for an analysis to be replicated. Which approach should be adopted depends on the problem at hand and the availability of the data. Lapatas et al. [2015] identify six major data-integration methodologies across these two approaches used in biology. *Data centralisation*, *data warehousing*, and *dataset integration* are eager approaches, while *hyperlinks*, *federated databases*, and *linked data* are lazy approaches. The UniProt database (see Section 5.2.2.2 for a description) is a case of data centralisation. The *Pathway Commons* database

[Cerami et al., 2011], which collects pathways from multiple databases and allows queries on them, is an instance of data warehousing. Custom scripts accessing online databases is a way to perform dataset integration. *ExPASy* [Artimo et al., 2012], a portal to bioinformatics resources, is an example of hyperlinks. In federated databases, a translation layer exists between each (heterogeneous) database and a central query service; a case in point is *PSICQUIC* [Aranda et al., 2011], a web service for querying multiple molecular-interaction databases. Finally, linked data is a collection of best practices for publishing and linking structured data on the Web [Bizer et al., 2009]. An example of linked data is *BIO2RDF* [Belleau et al., 2008], a system that converts several biological databases into *Resource Description Format (RDF)* (a standard data model for the Web).[3]

1.5 Integrative causal analysis

Integrative causal analysis (INCA) is a data-integration approach whose goal is to induce all *causal* models that are consistent with all relevant datasets and prior knowledge [Tsamardinos et al., 2012]. INCA encompasses methods for inducing causal models from datasets generated under different experimental conditions, datasets with overlapping sets of variables, and datasets with semantically similar variables, as well as methods for inducing causal models consistent with prior knowledge [Tsamardinos et al., 2012]. In genetics, INCA can be used to identify all causal relationships between variants and disease that are consistent with all genetic datasets from the disease and prior biological knowledge. In terms of the framework of Hamid et al. [2009], INCA methods are typically late-stage approaches, as the results of causal discovery from each dataset are combined. The data which the algorithms are applied to can be of similar or heterogeneous type. Among the methodologies identified by Lapatas et al. [2015], INCA is, obviously, a dataset-integration methodology. In contrast to other data-integration approaches, INCA produces an output with a causal interpretation. In addition, INCA may allow for inferences that are impossible with non-causal methods: the INCA algorithm of

[3] https://www.w3.org/RDF/

Tsamardinos et al. [2012], for example, is able to detect dependencies between variables that are never measured together.

1.6 Prion disease

The focus of this project is the identification of variants that confer susceptibility to *prion disease*. Prion diseases are lethal neurodegenerative disorders in humans and animals that are caused by infectious agents solely composed of protein called *prions* (from "proteinaceous infectious particle" [Prusiner et al., 1982]). Specifically, prions consist of PrP^{Sc} (Sc for *Scrapie*, a prion disease in sheep), which is a misfolded form of the host-encoded *prion protein (PrP)* and induces the conversion of PrP into additional PrP^{Sc} aggregating mainly in the brain [Collinge, 2001, Wadsworth and Collinge, 2007]. Prion diseases are also referred to as *transmissible spongiform encephalopathies* because of the spongiform appearance of diseased brain tissue.

PrP is encoded by the *prion protein gene (PRNP)*, which is located on chromosome 20 in humans and expressed mainly in the central nervous system [Collinge, 2001]. PrP is post-translationally processed to have an N-terminal signal peptide and a C-terminal propeptide removed and a *glycosylphosphatidyl inositol (GPI)* anchor attached and is subsequently transported to the cell surface, where it is tethered to the plasma membrane via its GPI anchor [Collinge, 2001]. PrP then cycles between the cell surface and early endosomes [Shyng et al., 1994]. The function of PrP remains elusive; mice display normal development and behaviour if PRNP is knocked out [Büeler et al., 1992], although several minor abnormalities (e.g. demyelinating neuropathy [Bremer et al., 2010]) have been seen.

The mechanism of neurodegeneration in prion disease is unknown. As PRNP-null mice are seemingly normal, loss of PrP to PrP^{Sc} is most probably not the cause of neurodegeneration. Since PRNP-null mice also do not develop prion disease when inoculated with PrP^{Sc} [Büeler et al., 1993], conversion of PrP to PrP^{Sc} appears to be necessary for the disease to occur. Therefore it has been suggested that a toxic intermediate, termed PrP^{L} (L for lethal) is formed during the conversion [Hill et al.,

2000, Hill and Collinge, 2003]; to date, PrP^{L} has not been identified.

Scrapie, *chronic wasting disease*, *transmissible mink encephalopathy*, *feline spongiform encephalopathy*, and *bovine spongiform encephalopathy (BSE)* (colloquially known as "mad cow" disease) are prion diseases in sheep, deer, mink, felines, and cattle, respectively. An epidemic of BSE emerged in the UK in the 1980s and declined a few years later after control measures were out into place [Smith and Bradley, 2003]. BSE was linked to the *meat-and-bone meal (MBM)* used to feed the affected cattle and produced from parts of sheep, cattle, and other animals that are not suitable for human consumption. The epidemic is hypothesised to have been caused by consumption of MBM produced from parts of sheep affected by scrapie or cattle affected by a sporadic form of BSE.

Traditionally, human prion diseases have been classified into *Creutzfeldt-Jakob disease (CJD)*, *Gerstmann-Straussler-Scheinker (GSS)* disease, and *kuru* [Collinge, 2001]; aetiologically, they can be classified into *sporadic*, *inherited*, and *acquired*. *Sporadic Creutzfeldt-Jakob disease (sCJD)* is a rapidly progressive dementia which represents about 85% of the cases of human prion disease and occurs randomly in the population with an annual mortality rate of 1–2 per million.[4] The mean onset of the disease is 60 years [Brown et al., 1994] and both sexes are equally affected [Collins et al., 2006]. The aetiology of sCJD is unknown; hypotheses include spontaneous misfolding of PrP into PrP^{Sc} and somatic mutation of PRNP [Colby and Prusiner, 2011].

Inherited prion diseases represent about 15% of the cases of human prion disease and comprise *familial Creutzfeldt-Jakob disease*, *fatal familial insomnia*, and GSS. All of them are linked to specific mutations in PRNP [Mead, 2006], although the exact causal mechanism is unknown.

Acquired human prion diseases comprise *kuru*, *iatrogenic Creutzfeldt-Jakob disease (iCJD)*, and *variant Creutzfeldt-Jakob disease (vCJD)*. An epidemic of kuru occurred in the 1950s in Papua New Guinea among the participants in cannibalistic rituals and declined when cannibalism was banned by the Australian administration

[4] http://www.cjd.ed.ac.uk/documents/report23.pdf

at the time [Mead et al., 2003]. It is hypothesised that the epidemic was caused by the consumption of an individual with sCJD [Alpers and Rail, 1970]. At the peak of the epidemic in the late 1950s, 200 people were dying from kuru every year; the number declined to 6 per year in the early 1990s and 1–2 per year in the early 2000s [Collinge et al., 2006]. The disease incubation time is estimated to be from 5 to more than 50 years [Collinge et al., 2006]. Clinically, kuru is a progressive cerebellar ataxia with a disease onset between 5 and 60 years and a duration time between 3 months and 2 years [Collinge et al., 2008].

iCJD is caused by accidental exposure to human prions during clinical procedures; reported transmission routes include treatment with cadaver-derived growth hormone, implantation of dura matter grafts, transplantation of corneas, and use of contaminated neurosurgical instruments and electroencephalography electrodes [Brown et al., 2000]. The clinical manifestation of iCJD depends on the route of transmission [Wadsworth and Collinge, 2007].

Following the BSE epidemic, a new variant of CJD appeared in young people in the UK [Will et al., 1996]; transmission experiments in mice [Hill et al., 1997, Bruce et al., 1997] subsequently confirmed that *variant CJD* is caused by BSE prions, implying that the people acquired disease due to consumption of contaminated beef. The age of onset for vCJD is between 12 and 74 years and the disease duration between 6 and 39 months [Spencer et al., 2002]. The early stages of the disease are dominated by psychiatric symptoms and many patients exhibit neurological symptoms within 4 months of clinical onset [Spencer et al., 2002]. To date, there have been 176 documented cases of vCJD in the UK.[5] The epidemic reached a peak in the year 2000, when 27 cases were diagnosed; the number of cases diagnosed per year has since declined to 1–2. The wide range of incubation times seen in kuru is raising concerns about the possibility of an epidemic of vCJD in the future [Collinge et al., 2006].

Only a few susceptibility variants have been identified in human prion disease, and all of them are located in the PRNP gene. PRNP codon 129 (rs1799990), in

[5] http://www.cjd.ed.ac.uk/documents/cjdq72.pdf

particular, is a susceptibility variant in all prion diseases [Mead, 2006]; its strongest effect is found in vCJD, where all but one patients are homozygous for methionine [Kaski et al., 2009]. The fact that a third of the population exposed to bovine prions is homozygous for methionine and unaffected suggests that there may be additional susceptibility loci [Lukic and Mead, 2011]. The strongest evidence for non-PRNP loci is provided by experiments in mice, which showed that the disease incubation time is variable across mouse strains even if they have the same PRNP haplotype [Lloyd et al., 2001].

In search of additional susceptibility factors, the MRC Prion Unit performed a GWAS on multiple prion diseases and populations [Mead et al., 2012]. However, no statistically-significant associations were found apart from SNP rs1799990 and SNPs in LD with that locus. Conditional analysis showed that the association of the latter SNPs with the disease is through rs1799990. The Unit has also undertaken an exome-sequencing study in sCJD and vCJD, which has, so far, not resulted in any novel discoveries. Finally, an *exome-array* study in sCJD was undertaken, whose association-analysis stage was conducted by me (See Appendix B). Exome arrays are chips that allow for inexpensive genotyping of SNPs that were previously identified in exome-sequencing studies [Grove et al., 2013]. The few SNPs that were significantly associated with the disease did not pass post-association quality control (see Appendix B).

1.7 Aim and contributions

As it is the case with complex diseases, missing heritability is also a problem in prion disease. Although there is evidence for the existence of additional susceptibility variants, none have been identified (at least, not yet) by analysing the available genetic datasets in isolation. Therefore, co-analysing the datasets may aid the discovery of additional variants. In contrast to other approaches to data integration, the output of INCA has a causal interpretation. For that reason, INCA was adopted as a framework for data integration in prion disease. The aim of this project is therefore to *identify all causal relationships between variants and prion disease that are*

consistent with all prion datasets and prior biological knowledge.

Towards reaching that aim, the following contributions have been made in each chapter of the thesis:

- *Chapter 3:* Before using INCA to co-analyse a collection of genetic datasets, causal discovery from a single genetic dataset needs to be characterised and realistic simulations need to be conducted. In that chapter, a theory on causal discovery from genetic datasets is formulated and specialised causal discovery algorithms with biologically interpretable output are devised. Sufficient conditions are given for the odds ratio in the sampled population to equal the *causal* odds ratio in the general population. A simulation study that was conducted for one of the algorithms using a causal model learned from real genetic data demonstrates that the algorithm is capable of discovering causal variants, while discarding variants that are definitely non-causal and controlling the FDR. Finally, the algorithm was applied to the genetic datasets in prion disease. SNP rs1799990 was successfully rediscovered from the GWAS dataset while SNPs in LD with rs1799990 were discarded; several SNPs were discovered from the other two datasets.

- *Chapter 4:* Building on the theory developed in the previous chapter, INCA algorithms for causal discovery from genetic datasets with overlapping sets of variants are developed. Contrary to existing INCA algorithms, the algorithms developed here have a biologically interpretable output. In addition, one of the algorithms is able to learn all genotype–phenotype causal relationships that are consistent with the data without having to learn the genotype–genotype relationships as well; existing algorithms have to learn consistent causal models over all variables. A version of that algorithm that controls the FDR was applied to a combination of prion genetic datasets; however, no novel discoveries were made.

- *Chapter 5:* A variant-filtering approach inspired by INCA that uses causal prior knowledge is proposed. Publicly available biological data and prior

knowledge were integrated into a directed graph with nodes comprising the disease and molecular entities in the cell associated with genomic regions, and with edges denoting causation. Candidate causal variants were subsequently identified from the ancestors (causes) of the disease in the graph. The exome-sequencing datasets from prion disease were filtered using the proposed approach and a well-established gene-prioritisation tool. Causal discovery from the datasets filtered using the former approach resulted in discoveries that were more statistically significant than when using the latter tool.

The necessary background information on causal discovery and INCA is provided in Chapter 2 and a summary of this work is given and future work is discussed in Chapter 6. All theorems in this work are proved in Appendix A, while the exome-array association study in sCJD is presented in Appendix B. The code for the experiments performed can be downloaded from http://www.angelosarmen.com.

Chapter 2

Background

In this chapter, background information on association analysis, causal Bayesian networks, structure learning from single samples, and structure learning from samples with overlapping sets of variables (an example of INCA) is provided.

In the following, random variables are denoted by capital letters (e.g. X) and their observed values by the respective lowercase ones (e.g. x). Sets are in boldface (e.g. $\mathbf{S}$), graphs in blackboard bold (e.g. $\mathbb{G}$), and probability distributions in calligraphic (e.g. $\mathscr{P}$). $|\mathbf{S}|$ denotes the cardinality of set $\mathbf{S}$. The symbol $\triangleq$ stands for "is defined as". $\Pr(A)$ is the probability of event A. Finally, $\mathbf{A} \dot{\cup} \mathbf{B}$ denotes the union of disjoint sets $\mathbf{A}$ and $\mathbf{B}$.

2.1 Association analysis

In an association analysis, the association of a target variable with every other variable in a sample is tested. In genetics, the target variable is a phenotype P and the rest variables correspond to genetic variants. In a case–control study of a disease, P is a binary variable whose levels are *unaffected* and *affected*. Only autosomal variants are usually considered, if the prevalence of a disease is the same among males and females. The observations are taken to be either (1) the gametes of the individuals (hence the sample size is doubled) or (2) the individuals themselves [Clarke et al., 2011].

In the first case, the variants are logical variables I indicating the presence of the rare allele and Pearson's χ^2 test or *Fisher's exact test* are usually used [Balding,

2006]. The statistical power is increased compared to the second case but additional assumptions must be made [see Clarke et al., 2011]. In an association analysis, the *effect* of each discovery on the target variable is typically reported along with the p-value corresponding to the discovery. In a genetic association analysis, the *allelic relative risk (RR)*, which is the ratio of the probability of being affected by the disease ("risk") for carriers of the rare allele to that for non-carriers, is typically the effect of interest:

$$\begin{aligned} \mathrm{RR} &\triangleq \frac{\Pr(P = \text{affected} \mid I = \text{true})}{\Pr(P = \text{affected} \mid I = \text{false})} \\ &= \frac{\Pr(I = \text{true} \mid P = \text{affected}) / \Pr(I = \text{true})}{\Pr(I = \text{false} \mid P = \text{affected}) / \Pr(I = \text{false})} \end{aligned} \tag{2.1}$$

An allelic RR of 5 means that the risk is 5 times higher for carriers of the rare allele compared to non-carriers. In order to estimate the RR, however, a random sample from the population is needed. A related measure is the allelic *odds ratio (OR)*, which is the ratio of the odds of being affected by the disease for carriers of the rare allele to that for non-carriers:

$$\mathrm{OR} \triangleq \frac{\Pr(P = \text{affected} \mid I = \text{true}) / \Pr(P = \text{unaffected} \mid I = \text{true})}{\Pr(P = \text{affected} \mid I = \text{false}) / \Pr(P = \text{unaffected} \mid I = \text{false})} \tag{2.2}$$

$$= \frac{\Pr(I = \text{true} \mid P = \text{affected}) / \Pr(I = \text{true} \mid P = \text{unaffected})}{\Pr(I = \text{false} \mid P = \text{affected}) / \Pr(I = \text{false} \mid P = \text{unaffected})} \tag{2.3}$$

An allelic OR of 5 means that the odds of being affected by the disease are 5 times higher for carriers of the rare allele compared to non-carriers. In contrast to the RR, the OR can be estimated from a case–control sample. Suppose that the sample is a random sample from the conditional distribution of the variables given $S = \text{true}$, where S is an unobserved logical variable, referred to as the *selection* variable. If I is independent of S given P, then $\Pr(I \mid P, S = \text{true}) = \Pr(I \mid P)$ and the OR can be computed using Equation 2.3. This situation corresponds to the sample being the concatenation of a random case sample and a random control sample. When the disease is rare, $\Pr(I) \approx \Pr(I \mid P = \text{unaffected})$ and therefore RR $\approx$ OR, as it can be

seen from Equations 2.1 and 2.3. The OR is estimated by the *sample* OR:

$$\widehat{\mathrm{OR}} \triangleq \frac{\#\{P = \text{affected}, I = \text{true}\} \cdot \#\{P = \text{affected}, I = \text{false}\}}{\#\{P = \text{unaffected}, I = \text{true}\} \cdot \#\{P = \text{unaffected}, I = \text{false}\}}$$

where $\#\{\ldots\}$ denotes the number of observations taking on the enclosed values in the sample. Clearly, $\widehat{\mathrm{OR}}$ can be zero, undefined, or infinite, depending on whether its nominator and/or its denominator are zero. The logarithm of $\widehat{\mathrm{OR}}$ asymptotically follows a normal distribution with mean $\log(\mathrm{OR})$ and standard error given by:

$$\mathrm{SE} = \sqrt{\frac{1}{N_{11}} + \frac{1}{N_{10}} + \frac{1}{N_{01}} + \frac{1}{N_{00}}}$$

where $N_{11} = \#\{P = \text{affected}, I = \text{true}\}$, $N_{10} = \#\{P = \text{affected}, I = \text{false}\}$, $N_{01} = \#\{P = \text{unaffected}, I = \text{true}\}$, and $N_{00} = \#\{P = \text{unaffected}, I = \text{false}\}$ [Morris and Gardner, 1988]. Hence, the 95% confidence interval for the OR is given by

$$[\exp(\log\widehat{\mathrm{OR}} - 1.96\mathrm{SE}), \exp(\log\widehat{\mathrm{OR}} + 1.96\mathrm{SE})]$$

In the second case, the (autosomal) variants are ternary variables G with domain $\{\mathrm{AA}, \mathrm{Aa}\,\mathrm{aa}\}$, where AA, Aa, and aa stands for homozygous for the common allele, heterozygous, and homozygous for the rare allele, respectively. *Pearson's* χ^2 *test* or the *Cochran-Armitage trend test* are typically employed [Balding, 2006]. Two "genotypic" ORs are reported. The first one ("heterozygote" OR) is the ratio of the odds of the heterozygotes to those of the homozygotes for the common allele:

$$\begin{aligned}\mathrm{OR}_{\mathrm{Aa}} &\triangleq \frac{\Pr(P = \text{affected} \mid G = \mathrm{Aa}) / \Pr(P = \text{unaffected} \mid G = \mathrm{Aa})}{\Pr(P = \text{affected} \mid G = \mathrm{AA}) / \Pr(P = \text{unaffected} \mid G = \mathrm{AA})} \\ &= \frac{\Pr(P = \text{affected} \mid G = \mathrm{Aa}) \cdot \Pr(P = \text{affected} \mid G = \mathrm{AA})}{\Pr(P = \text{unaffected} \mid G = \mathrm{Aa}) \cdot \Pr(P = \text{unaffected} \mid G = \mathrm{AA})}\end{aligned}$$

The second one ("rare homozygote" OR) is the odds of the homozygotes for the

rare allele to those of the homozygotes for the common allele:

$$\begin{aligned}\mathrm{OR}_{\mathrm{aa}} &\triangleq \frac{\Pr(P = \text{affected} \mid G = \mathrm{aa}) / \Pr(P = \text{unaffected} \mid G = \mathrm{aa})}{\Pr(P = \text{affected} \mid G = \mathrm{AA}) / \Pr(P = \text{unaffected} \mid G = \mathrm{AA})} \\ &= \frac{\Pr(P = \text{affected} \mid G = \mathrm{aa}) \cdot \Pr(P = \text{affected} \mid G = \mathrm{AA})}{\Pr(P = \text{unaffected} \mid G = \mathrm{aa}) \cdot \Pr(P = \text{unaffected} \mid G = \mathrm{AA})}\end{aligned}$$

To account for multiple testing, the FWER is usually controlled at 0.05 using the *Bonferroni correction* or a fixed p-value threshold such as $5 \cdot 10^{-8}$, obtained by estimating the effective number of independent tests in a GWAS through simulations [Sham and Purcell, 2014]. As the FWER may be too strict an error rate to control, the FDR [Benjamini and Hochberg, 1995] may be controlled instead:

$$\mathrm{FDR} \triangleq \mathrm{E}\left[\frac{\mathrm{V}}{\mathrm{R} \vee 1}\right] = \mathrm{E}\left[\frac{\mathrm{V}}{\mathrm{R}} \,\middle|\, \mathrm{R} > 0\right] \Pr(\mathrm{R} > 0)$$

where V is the number of rejected true null hypotheses, R is the number of rejections, and $\mathrm{R} \vee 1$ corresponds to setting V / R to 0 when $\mathrm{R} = 0$. When the the p-values are independent, the procedure of Benjamini and Hochberg [1995] can be used to control the FDR below the desired level. When the p-values are dependent, which is usually the case, the more conservative procedure of Benjamini and Yekutieli [2001] (here referred to as the *BY procedure*) can be employed.

2.2 Causal Bayesian networks

Elucidating causal relationships is of utmost importance in life sciences and other physical sciences. The concept of causality has preoccupied philosophers and scientists for centuries and while there is still no consensus on what constitutes a cause, approaches to causality have been developed in computer science and statistics [Kleinberg and Hripcsak, 2011]. These approaches include *causal Bayesian networks (CBNs)* [Spirtes et al., 2000, Pearl, 2009], which represent the causal and probabilistic relationships among a set of variables, *structural equation models (SEMs)* [Pearl, 2009], which are related to CBNs, *Granger causality* for inferring causal relationships between time series [Granger, 1969], and an approach based on *temporal logic* [Kleinberg and Mishra, 2009] for inferring causal relationships

from temporal observations. The SEM framework is the most comprehensive approach and subsumes other approaches such as the *potential-outcome* framework of Rubin [1974] [Pearl, 2010]. The causal relationships in SEMs are expressed as deterministic equations with stochastic error terms in contrast to CBNs, were the causal relationships are stochastic; this representation allows for the computation of *counterfactual* probabilities (probabilities of events contrary to fact), which requires knowledge of the underlying process [Pearl, 2009]. As disease mechanisms are largely unknown and computation of counterfactual probabilities is not of interest in this work, the CBN approach is preferred here.

In the CBN approach to causality, the definition of causation is based on the notion of *manipulation* (also called *intervention*). Manipulating a random variable means forcing the variable to take on some value [Neapolitan, 2004]. A random variable X is said to be a *cause* of random variable Y and Y an *effect* of X if there is some manipulation of the value of X that changes the probability distribution of Y [Cooper, 1999]. X is called a *direct cause* of Y and Y a *direct effect* of X relative to a set of variables $\mathbf{V}$ when no instantiation of a subset of $\mathbf{V} \setminus \{X,Y\}$ cancels the effect of the manipulation of X [Neapolitan, 2004]. The condition of *causal transitivity* states that, if X is a cause of Y and Y is a cause of Z, then X is a cause of Z. The causal relationships among a set of variables can be represented by a *causal directed acyclic graph*.

In graph theory, a *graph* is a pair $(\mathbf{V},\mathbf{E})$ of a set of nodes (also called vertices) $\mathbf{V}$ and a set of edges $\mathbf{E}$ that connect pairs of nodes. If there is an edge between nodes X and Y, X and Y are *adjacent* and the edge between X and Y is *incident* to X and to Y. If $\{X_1,\ldots,X_n\}$ is an ordered set of nodes such that, for $2 \leq i \leq n$, X_{i-1} and X are adjacent, the corresponding set of edges is a *path* from X_1 to X_n and is denoted by $[X_1,\ldots,X_n]$; nodes $X_2,\ldots,X_{n-1}$ are called *interior* nodes on the path. Let $p = [X_1,\ldots,X_k]$ be a path. The *subpath* $[X_i,\ldots,X_j]$ of p from X_i to X_j is denoted by $p(X_i,Y_j)$. Let $p_1 = [X_1,\ldots,X_k]$ and $p_2 = [X_k,\ldots,X_{k+n-1}]$ be two paths. The *concatenation* $[X_1,\ldots,X_{k+n-1}]$ of p_1 and p_2 is denoted by $[p_1,p_2]$.

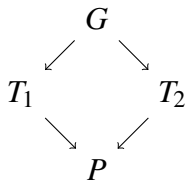
A graph is called *directed* (resp. *undirected*) when its edges are directed (resp.

undirected). If there is an edge $X \to Y$ in a directed graph, then X is a *parent* of Y and Y a *child* of X; the edge is said to be *out of* X and *into* Y. A path from X to Y is out of (into) X and out of (into) Y if the first edge of the path is out of (into) X and the last edge is out of (into) Y. A path from X to Y where all edges are directed towards X is called *directed*. If there is a directed path from X to Y or $X = Y$, then X is an *ancestor* of Y and Y a *descendant* of X. A *directed cycle* occurs if X is parent of Y and Y is an ancestor of X. A *directed acyclic graph (DAG)* is a directed graph without directed cycles.

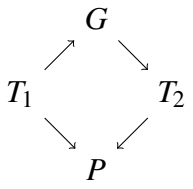
A *causal DAG* is defined as a DAG whose nodes are random variables and edge $X \to Y$ denotes that X is a direct cause of Y (and Y is a direct effect of X) relative to the variables in the DAG [Neapolitan, 2004]. Therefore, X is a direct cause (resp. a direct effect) of Y if and only if X is a parent (resp. a child) of Y. Assuming that causal transitivity holds, X is a cause of Y (and Y is an effect of X) if and only if there is a directed path from X to Y, that is, X is an ancestor of Y (and Y is a descendant of X). A directed path is then termed a *causal path*. In the rest of the thesis, the terms "node" and "variable" are used interchangeably. A causal DAG defined over a set of variables merely makes *assertions* about the causal relationships between the variables. The causal DAG which represents the *true* causal relationships between the variables is referred to as the *true* causal DAG over the variables. Figure 2.1a shows a biological causal DAG with four variables, where G is a genotype, T_1 and T_2 are the transcript (mRNA) levels of two genes, and P is a phenotype. According to the graph, G causes P through T_1 and T_2.

A causal DAG can also represent the probabilistic relationships (conditional independencies), among a set of variables. The pair $(\mathbb{G}, \mathscr{P})$ of a causal DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ and the probability distribution $\mathscr{P}$ of the variables in $\mathbf{V}$ satisfies the *causal Markov condition (CMC)* if every variable X is conditionally independent from the set $\mathbf{ND}(X)$ of its non-effects (non-descendants) given the set $\mathbf{PA}(X)$ of its direct causes (parents) [Neapolitan, 2004]:

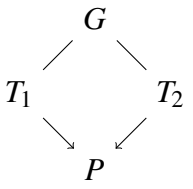
$$X \perp\!\!\!\perp \mathbf{ND}(X) \mid \mathbf{PA}(X)$$



(a) A causal DAG. G causes P through T_1 and T_2. In a CBN with this graph, G is independent from P given T_1 and T_2.



(b) A causal DAG that is Markov equivalent to the DAG in Figure 2.1a.



(c) The causal DAG pattern of the Markov equivalence class to which the causal DAGs in Figures 2.1a and 2.1b belong to.

Figure 2.1: Example of a causal DAG, a Markov-equivalent causal DAG, and a causal DAG pattern. G is the genotype at some locus, T_1 and T_2 are the transcript levels of two genes, and P is a phenotype.

In order for the CMC to be satisfied, the following conditions must hold [Neapolitan, 2004]:

1. *Causal sufficiency*: either there are no common causes of pairs of variables in $\mathbf{V}$ that are not in $\mathbf{V}$ (referred to as *hidden* common causes or *confounders*), or every such common cause is a constant.

2. There are no *causal feedback loops*: no pair of variables in $\mathbf{V}$ cause each other.

3. *Selection bias* is absent: $\mathscr{P}$ is not conditional on an instantiation of a common effect of a pair of variables in $\mathbf{V}$.

A pair $(\mathbb{G}, \mathscr{P})$ that satisfies the CMC is called a CBN; $\mathbb{G}$ is called the *causal structure* of the network. In a CBN whose causal structure is the one in Figure 2.1a, G is independent from P given T_1 and T_2 (the genotype is independent from the phenotype when the transcript levels of the two genes are known). In a CBN $(\mathbb{G}, \mathscr{P})$ over set of variables $\mathbf{V} = \{V_1, \ldots, V_n\}$, $\mathscr{P}$ is decomposed into the product of the *conditionally probability distribution (CPD)* of each node given its parents (direct causes) in $\mathbb{G}$ [Pearl, 1988]:

$$\mathscr{P}(\mathbf{v}) = \prod_{i=1}^{n} \mathscr{P}(v_i \mid \mathbf{pa}(V_i))$$

This allows for the CBN to be a compact representation of $\mathscr{P}$. Computing marginal/conditional distributions of $\mathscr{P}$ is referred to as performing *Bayesian-network inference*. The *junction tree* algorithm [Huang and Darwiche, 1996] is an efficient algorithm for that purpose.

The conditional independencies in the definition of the CMC entail additional conditional independencies. Specifically, every conditional independence entailed by the CMC corresponds to a *d-separation* in the DAG. Some definitions are needed before introducing d-separation. In a graph, a *triple* is a path with three nodes. If Z is an interior node on a path and $X \to Z \leftarrow Y$ on the path, Z is called a *collider* on the path (the triple $[X,Y,Z]$ is also referred to as a collider). A path from X to Y is *blocked* by a subset $\mathbf{Z}$ of the rest nodes if there is a collider on the path that is not an ancestor of any node in $\mathbf{Z}$, or there is a non-collider on the path that is in $\mathbf{Z}$; otherwise, the path is *active* given $\mathbf{Z}$. In Figure 2.1a, path $[G,T_1,P]$ is active given $\emptyset$ because T_1 is not a collider on the path and $T_1 \notin \emptyset$, while it is blocked by $\mathbf{Z} = \{T_1\}$ because $T_1 \in \mathbf{Z}$. $[T_1,P,T_2]$ is blocked by $\emptyset$ because P is a collider on the path and $P \notin \emptyset$, while it is active given $\mathbf{Z} = \{G,P\}$ because $P \in \mathbf{Z}$. When all paths between X and Y are blocked by $\mathbf{Z}$, X and Y are said to be *d-separated* by $\mathbf{Z}$ (denoted by $X \perp Y \mid \mathbf{Z}$); otherwise, X and Y are *d-connected* given $\mathbf{Z}$ (denoted by $X \not\perp Y \mid \mathbf{Z}$). $\mathbf{Z}$ is termed a *sepset* of X and Y. When X and Y are d-separated by the empty set, X and Y are simply said to be d-separated; when they are d-connected given the empty set, they are d-connected. In Figure 2.1a, P and G are d-separated by $\mathbf{Z} = \{T_1,T_2\}$ because both paths between P and G, $[P,T_1,G]$ and $[P,T_2,G]$, are blocked by $\mathbf{Z}$. X and Y are said to be *strictly* d-separated by $\mathbf{Z}$ if X and Y are d-separated by $\mathbf{Z}$ and are d-connected given any proper subset of $\mathbf{Z}$; $\mathbf{Z}$ is called a *minimal* sepset of G and P. Clearly, every non-minimal sepset has a minimal subset.

Let $\mathbf{I}(\mathbb{G})$ and $\mathbf{I}(\mathscr{P})$ denote the set of d-separations in DAG $\mathbb{G}$ and conditional independencies in probability distribution $\mathscr{P}$, respectively. Suppose that $(\mathbb{G},\mathscr{P})$ is a CBN. If X and Y are d-separated by $\mathbf{Z}$ in $\mathbb{G}$, then X and Y are conditionally

independent given $\mathbf{Z}$ in $\mathscr{P}$ [see Verma and Pearl, 1990, for proof]:

$$X \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{G}) \implies X \perp\!\!\!\perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathscr{P})$$

The reverse, however, is not always true, unless the *causal faithfulness condition (CFC)* is satisfied [Spirtes et al., 2000]. The CFC states that all and only conditional independencies in $\mathscr{P}$ are entailed by the CMC. Thus, when $(\mathbb{G}, \mathscr{P})$ satisfies the CFC, X and Y are d-separated by $\mathbf{Z}$ in $\mathbb{G}$ if and only if they are conditionally independent given $\mathbf{Z}$ in $\mathscr{P}$:

$$X \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{G}) \iff X \perp\!\!\!\perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathscr{P})$$

Then $\mathbb{G}$ is called a *perfect map* of $\mathscr{P}$ and $\mathbb{G}$ and $\mathscr{P}$ are said to be *faithful* to each other. If a CBN $(\mathbb{G}, \mathscr{P})$ satisfies the CFC, then the CBN is said to be *faithful*. The justification for the CFC is this: if the parameters of the CPDs in $(\mathbb{G}, \mathscr{P})$ are assigned randomly, it is unlikely that the resulting joint distribution violates the CFC [Meek, 1995]. The CFC implies causal transitivity [Neapolitan, 2004].

Before concluding this section, note that there are also *non-causal* Bayesian networks. A *Bayesian network (BN)* is defined as the pair of a DAG and a probability distribution that satisfies the Markov condition: every variable is conditionally independent from the set of its non-descendants given the set of its parents. Accordingly, the faithfulness condition states that all and only conditional independencies in the distribution are entailed by the Markov condition. Every distribution can be encoded by a BN by ordering the variables and setting, for each variable, the previous variables in the ordering as its parents in the DAG.

2.3 Causal structure learning

The goal of causal structure learning is to learn *features* of the true causal DAG over a set of random variables given a random sample from the probability distribution of the variables. The goal of *constraint-based learning* is to learn features of a DAG given the set of d-separations in the DAG. When the DAG is causal, the task is called

constraint-based causal learning. In practice, the d-separations are determined by assuming that the CFC holds and performing hypothesis tests of conditional independence on a random sample from the probability distribution of the variables (see Section 2.3.2).

Causal structure learning can typically only learn features of the true causal DAG and not the whole DAG because other DAGs may have the same d-separations. These DAGs are called *Markov equivalent* and said to belong to the same *Markov equivalence class*. The *skeleton* and the *unshielded colliders* are the same within the class. In a directed graph, a *link* is an edge without regard of direction, and the skeleton of a directed graph is the undirected graph whose edges corresponds to links in the directed graph. A triple $[X,Z,Y]$ is shielded if X and Y are adjacent. In this work, a node Z is called shielded if no unshielded triple $[X,Z,Y]$ exists. Figure 2.1b shows a causal DAG that is Markov equivalent to the one in 2.1a. A class of Markov-equivalent DAGs can be represented by a *DAG pattern*, which is a graph with two types of edges, directed and undirected; an undirected edge in the DAG pattern indicates that the orientation of the edge varies within the class. A DAG pattern that represents a Markov equivalence class of *causal* DAGs is called a *causal DAG pattern*. The causal DAG pattern that represents the Markov equivalence class of the true causal DAG over a set of variables is called the *true* causal DAG pattern over the variables. Figure 2.1c shows the causal DAG pattern of the Markov equivalence class to which the DAGs in Figures 2.1a and 2.1b belong to.

Constraint-based algorithms that learn DAG patterns consist of two phases, *skeleton identification* and *edge orientation*. In the first phase, the skeleton is identified. In the second phase, the edges of the skeleton are oriented using the sepsets. Skeleton identification (Algorithm 1) is based on the following theorem:

Theorem 2.1. *In a DAG over* $\mathbf{V}$*, nodes* X *and* Y *are adjacent if and only if there is no* $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X,Y\}$ *such that* X *and* Y *are d-separated by* $\mathbf{Z}$*.*

For each pair $\{X,Y\}$ of nodes, a search for a sepset is conducted. If no sepset is found, then an edge between X and Y is added to the skeleton. Edge orientation uses the sepsets and applies a set of four rules in order to orient the invariant edges

in the DAG pattern [see Spirtes et al., 2000]. The first orients the unshielded triples based on a result that states that if $p = [X, Y, Z]$ is an unshielded triple and $\mathbf{S}$ is a sepset of X and Z, then p is a collider if and only if $Y \notin \mathbf{S}$. The rest of the rules are then applied iteratively until no more edges can be oriented.

Algorithm 1 Basic skeleton identification. $\mathbf{V}$ is a set of random variables. $\mathbb{G}$ is a DAG over $\mathbf{V}$. $\mathbf{I}(\mathbb{G})$ is the set of d-separations in $\mathbb{G}$. $\mathbb{S}$ is an undirected graph. **Sepset** is a map from pairs of nodes to sets of nodes. $X \perp Y \mid \mathbf{Z}$ denotes that nodes X and Y are d-separated given set of nodes $\mathbf{Z}$. In the output, $\mathbb{S}$ is the skeleton of $\mathbb{G}$ and **Sepset** is a map from pairs of nodes in $\mathbf{V}$ to sets of nodes in $\mathbf{V}$ that d-separate them in $\mathbb{G}$.

Input: $\mathbf{I}(\mathbb{G})$
Output: $\mathbb{S}$ and **Sepset**

initialise $\mathbb{S}$ with the empty undirected graph over $\mathbf{V}$
initialise **Sepset** with the empty map
for each pair $\{X, Y\} \in \mathbf{V}$ **do**
    if $\exists \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$ s.t. $X \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{G})$ **then**
        Sepset$(\{X, Y\}) \leftarrow \mathbf{Z}$
    else
        add edge $X - Y$ to $\mathbb{S}$
    end if
end for

In fact, it is not required to search among all subsets of the rest variables in order to find a sepset of X and Y, owing to the following theorem [Spirtes et al., 2000]:

Theorem 2.2. *Let $\mathbb{G}$ be a DAG over $\mathbf{V}$. If X and $Y \in \mathbf{V}$ are d-separated by a subset of $\mathbf{V} \setminus \{X, Y\}$, then they are d-separated by the set of parents in $\mathbb{G}$ of X or the set of the parents in $\mathbb{G}$ of Y.*

The parents of X and Y are, of course, unknown. The skeleton-identification phase (referred to as PC–skeleton; Algorithm 2) of the prototypical PC algorithm [Spirtes et al., 2000] starts with the *complete* undirected graph, that is, the one where every node is adjacent to all the others. Then, for increasing conditioning-set cardinality k and for each undirected edge between variables X and Y, it searches for a sepset with cardinality k among the subsets of the nodes currently adjacent to X and the subsets of the nodes currently adjacent to Y. If a sepset is found, the edge is discarded.

Algorithm 2 PC–skeleton, the skeleton-identification phase of the PC algorithm [Spirtes et al., 2000]. $\mathbf{V}$ is a set of random variables. $\mathbb{G}$ is a DAG over $\mathbf{V}$. $\mathbf{I}(\mathbb{G})$ is the set of d-separations in $\mathbb{G}$. $\mathbb{S}$ is an undirected graph. **Sepset** is a map from pairs of nodes to sets of nodes. $X \perp Y \mid \mathbf{Z}$ denotes that nodes X and Y are d-separated given set of nodes $\mathbf{Z}$. In the output, $\mathbb{S}$ is the skeleton of $\mathbb{G}$ and **Sepset** is a map from pairs of nodes in $\mathbf{V}$ to sets of nodes in $\mathbf{V}$ that d-separate them in $\mathbb{G}$. $\mathbf{AD}_{\mathbb{S}}(X)$ is the set of nodes that are adjacent to X in $\mathbb{S}$.

Input: $\mathbf{I}(\mathbb{G})$
Output: $\mathbb{S}$ and **Sepset**

initialise $\mathbb{S}$ with the complete undirected graph over $\mathbf{V}$
initialise **Sepset** with the empty map
$k \leftarrow 0$
repeat
  for each $X \in \mathbf{V}$ **do**
    for each $Y \in \mathbf{AD}_{\mathbb{S}}(X)$ **do**
      if $\exists \mathbf{Z} \subseteq \mathbf{AD}_{\mathbb{S}}(X) \setminus \{Y\}$ s.t. $|\mathbf{Z}| = k$ and $X \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{G})$ **then**
        remove edge $X - Y$ from $\mathbb{S}$
        Sepset$(\{X, Y\}) \leftarrow \mathbf{Z}$
      end if
    end for
  end for
  $k \leftarrow k + 1$
until $|\mathbf{AD}_{\mathbb{S}}(X)| \leq k$ for all $X \in \mathbf{V}$

2.3.1 Local learning

It is not always of interest to learn the whole DAG pattern. *Local learning* algorithms can be used to learn local structure around a target variable T [Aliferis et al., 2010a]. Specifically, algorithms instantiating the *Generalised Local Learning - Parents and Children (GLL-PC)* template [Aliferis et al., 2010a], such as the *Max-Min Parents and Children (MMPC)* algorithm [Tsamardinos et al., 2006], can learn the set $\mathbf{AD}(T)$ of parents and children (direct causes and direct effects, in case of causal learning) of T, or, in other words, the nodes adjacent to T. For example, MMPC targeting T_1 in Figure 2.1a outputs $\mathbf{AD}(T_1) = \{G, P\}$. GLL-PC comes in two versions, *symmetric* and *non-symmetric*.

In an instantiation of non-symmetric GLL-PC (Algorithm 3), the set $\mathbf{TA}(T)$ of nodes *tentatively* adjacent to T is initialised to some subset $\mathbf{U}$ of $\mathbf{V} \setminus \{X\}$, depending on the instantiation, and the set $\mathbf{OPEN}(T)$ of nodes that are under consideration for inclusion in $\mathbf{TA}(T)$ is initialised to $\mathbf{V} \setminus (\mathbf{TA}(T) \cup \{T\})$. Then, GLL-PC alternates

between two phases. In the *insertion phase*, some subset of $\mathbf{OPEN}(T)$ is moved to $\mathbf{TA}(T)$. The choice of subset depends on the instantiation. In the *elimination* phase, every node in $\mathbf{TA}(T)$ that is d-separated by T given a subset of the rest nodes in $\mathbf{TA}(T)$ is removed from $\mathbf{TA}(T)$. In the end, $\mathbf{TA}(T)$ contains all the nodes that are not d-separated from T by a subset of the nodes adjacent to T.

Algorithm 3 Non-symmetric Generalised Local Learning - Parents and Children (GLL-PC) [Aliferis et al., 2010a]. $\mathbf{V}$ is a set of random variables. $\mathbb{G}$ is a DAG over $\mathbf{V}$. $\mathbf{I}(\mathbb{G})$ is the set of d-separations in $\mathbb{G}$. $T \in \mathbf{V}$ is the target variable. **Sepset** is a map from pairs of nodes to sets of nodes. $X \perp Y \mid \mathbf{Z}$ denotes that nodes X and Y are d-separated given set of nodes $\mathbf{Z}$. In the output, $\mathbf{TA}(T)$ is a superset of the set of nodes in $\mathbf{V}$ that are adjacent to T in $\mathbb{G}$ and **Sepset** is a map from pairs of nodes in $\mathbf{V}$ to sets of nodes in $\mathbf{V}$ that d-separate them in $\mathbb{G}$.

Input: $\mathbf{I}(\mathbb{G})$ and T
Output: $\mathbf{TA}(T)$ and **Sepset**

▷ Initialisation
$\mathbf{TA}(T) \leftarrow \mathbf{U}$ for some $\mathbf{U} \subseteq \mathbf{V} \setminus \{T\}$
$\mathbf{OPEN}(T) \leftarrow \mathbf{V} \setminus (\mathbf{TA}(T) \cup \{T\})$
initialise **Sepset** with the empty map
repeat
    ▷ Insertion phase
    $\mathbf{TA}(T) \leftarrow \mathbf{TA}(T) \cup \mathbf{W}$ for some $\mathbf{W} \subseteq \mathbf{OPEN}(T)$
    $\mathbf{OPEN}(T) \leftarrow \mathbf{OPEN}(T) \setminus \mathbf{W}$
    ▷ Elimination phase
    for each $Y \in \mathbf{TA}(T)$ s.t. $\exists \mathbf{Z} \subseteq \mathbf{TA}(T) \setminus \{Y\}$ s.t. $T \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{G})$ **do**
        $\mathbf{TA}(T) \leftarrow \mathbf{TA}(T) \setminus \{Y\}$
        $\mathbf{Sepset}(\{T, Y\}) \leftarrow \mathbf{Z}$
    end for
until $\mathbf{OPEN}(T) = \emptyset$

The non-symmetric MMPC algorithm (Algorithm 4) is a popular instantiation of non-symmetric GLL-PC. MMPC starts with an empty $\mathbf{TA}(T)$. In the insertion phase, nodes in $\mathbf{OPEN}(T)$ that are d-separated from T by a subset of $\mathbf{TA}(T)$ are removed from $\mathbf{OPEN}(T)$. Then the node W of $\mathbf{OPEN}(T)$ with the maximal (among all nodes in $\mathbf{OPEN}(T)$) minimal (among all subsets $\mathbf{Z}$ of $\mathbf{TA}(T)$) association with T given $\mathbf{Z}$ (typically taken to be the negative p-value of the corresponding conditional independence test) is moved to $\mathbf{TA}(T)$ (hence the "max-min" in the name of the algorithm). The elimination phase is the same as in Algorithm 3.

In the output of Algorithm 3, there may nodes in $\mathbf{TA}(T)$ that are d-separated

Algorithm 4 The non-symmetric Max-Min Parents and Children (MMPC) algorithm [Tsamardinos et al., 2006]. $\mathbf{V}$ is a set of random variables. $\mathbb{G}$ is a DAG over $\mathbf{V}$. $\mathbf{I}(\mathbb{G})$ is the set of d-separations in $\mathbb{G}$. $T \in \mathbf{V}$ is the target variable. **Sepset** is a map from pairs of nodes to sets of nodes. $\text{assoc}(X, Y \mid \mathbf{Z})$ denotes the association of variables X and Y given set of variables $\mathbf{Z}$. $X \perp Y \mid \mathbf{Z}$ denotes that nodes X and Y are d-separated given set of nodes $\mathbf{Z}$. In the output, $\mathbf{TA}(T)$ is a superset of the set of nodes in $\mathbf{V}$ that are adjacent to T in $\mathbb{G}$ and **Sepset** is a map from pairs of nodes in $\mathbf{V}$ to sets of nodes in $\mathbf{V}$ that d-separate them in $\mathbb{G}$.

Input: $\mathbf{I}(\mathbb{G})$ and T
Output: $\mathbf{TA}(T)$ and **Sepset**

▷ Initialisation
$\mathbf{TA}(T) \leftarrow \emptyset$
$\mathbf{OPEN}(T) \leftarrow \mathbf{V} \setminus \{T\}$
initialise **Sepset** with the empty map
repeat
    ▷ Insertion phase
    for each $Y \in \mathbf{OPEN}(T)$ s.t. $\exists \mathbf{Z} \subseteq \mathbf{TA}(T)$ s.t. $T \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{G})$ **do**
        $\mathbf{OPEN}(T) \leftarrow \mathbf{OPEN}(T) \setminus \{Y\}$
        $\mathbf{Sepset}(\{T, Y\}) \leftarrow \mathbf{Z}$
    end for
    $W \leftarrow \arg\max_{W \in \mathbf{OPEN}(T)} \min_{\mathbf{Z} \subseteq \mathbf{TA}(T)} \text{assoc}(T, W \mid \mathbf{Z})$
    $\mathbf{TA}(T) \leftarrow \mathbf{TA}(T) \cup \{W\}$
    $\mathbf{OPEN}(T) \leftarrow \mathbf{OPEN}(T) \setminus \{W\}$
    ▷ Elimination phase
    for each $Y \in \mathbf{TA}(T)$ s.t. $\exists \mathbf{Z} \subseteq \mathbf{TA}(T) \setminus \{Y\}$ s.t. $T \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{G})$ **do**
        $\mathbf{TA}(T) \leftarrow \mathbf{TA}(T) \setminus \{Y\}$
        $\mathbf{Sepset}(\{T, Y\}) \leftarrow \mathbf{Z}$
    end for
until $\mathbf{OPEN}(T) = \emptyset$

from T by a subset of the nodes adjacent to them. In symmetric GLL-PC (Algorithm 5), non-symmetric GLL-PC is initially applied to T and subsequently the *symmetry correction* is performed: non-symmetric GLL-PC is also applied to all nodes in $\mathbf{TA}(T)$, and a node X is only inserted into $\mathbf{AD}(T)$ if X is in $\mathbf{TA}(T)$ and T is in $\mathbf{TA}(X)$. The MMPC algorithm is symmetric GLL-PC with non-symmetric MMPC as the instantiation of non-symmetric GLL-PC.

GLL-PC can be also used to learn the whole skeleton through *local-to-global learning (LGL)* [Aliferis et al., 2010b]. In LGL (Algorithm 6), non-symmetric GLL-PC is first applied to each variable X; then a link between X and Y is created if and only if Y is in $\mathbf{TA}(X)$ and X is in $\mathbf{TA}(Y)$. Both skeleton identification and

Algorithm 5 Symmetric Generalised Local Learning - Parents and Children (GLL-PC) [Aliferis et al., 2010a]. $\mathbf{V}$ is a set of random variables. $\mathbb{G}$ is a DAG over $\mathbf{V}$. $\mathbf{I}(\mathbb{G})$ is the set of d-separations in $\mathbb{G}$. $T \in \mathbf{V}$ is the target variable. $\mathbf{AD}(T)$ is a set of nodes in $\mathbf{V}$. **Sepset** is a map from pairs of nodes to sets of nodes. In the output, $\mathbf{AD}(T)$ is the set of nodes in $\mathbf{V}$ that are adjacent to T in $\mathbb{G}$ and **Sepset** is a map from pairs of nodes in $\mathbf{V}$ to sets of nodes in $\mathbf{V}$ that d-separate them in $\mathbb{G}$.

Input: $\mathbf{I}(\mathbb{G})$ and T
Output: $\mathbf{AD}(T)$ and **Sepset**

1: let $\mathbf{TA}(T)$ and $\mathbf{Sepset}_T$ be the output of Algorithm 3 with $\mathbf{I}(\mathbb{G})$ and T as input.
2: $\mathbf{AD}(T) \leftarrow \mathbf{TA}(T)$
3: **Sepset** $\leftarrow \mathbf{Sepset}_T$
4: **for each** $Y \in \mathbf{AD}(T)$ **do**
5: let $\mathbf{TA}(Y)$ and $\mathbf{Sepset}_Y$ be the output of Algorithm 3 with $\mathbf{I}(\mathbb{G})$ and Y as input.
6: **if** $T \notin \mathbf{TA}(Y)$ **then**
7: $\mathbf{AD}(T) \leftarrow \mathbf{AD}(T) \setminus \{Y\}$
8: $\mathbf{Sepset}(\{T,Y\}) \leftarrow \mathbf{Sepset}_Y(\{T,Y\})$
9: **end if**
10: **end for**

parent-and-children learning may be referred to as instances of *link identification*, where the existence of certain links of interest is determined.

In a set of variables $\mathbf{V}$, a *Markov blanket* of $X \in \mathbf{V}$ is a subset $\mathbf{A}$ of $\mathbf{V} \setminus \{X\}$ such that for every subset $\mathbf{B}$ of $\mathbf{V} \setminus \{X\} \setminus \mathbf{A}$, $\mathbf{B}$ is conditionally independent from X given $\mathbf{A}$ [Pearl, 1988]. That means that a Markov blanket of X renders all the subsets of the variables in $\mathbf{V}$ that are not in the Markov blanket of X independent from X. A *Markov boundary* is a minimal Markov blanket (no proper subset of a Markov boundary is a Markov blanket) and is an optimal solution to the problem of *variable (or feature) selection*, where the goal is to find a subset of the variables that are relevant for predicting the value of the response variable in classification or regression models [Tsamardinos and Aliferis, 2003]. In a faithful CBN, each node X has a unique Markov boundary, equal to the set of parents (causes), children (effects), and parents of children (causes of effects) of X [Aliferis et al., 2010a]. Algorithms instantiating the *Generalised Local Learning - Markov Blanket (GLL-MB)* template [Aliferis et al., 2010a] can learn the Markov boundary $\mathbf{MB}(T)$ of a target node T.

Algorithm 6 Local-to-global learning (LGL) [Aliferis et al., 2010b]. **V** is a set of random variables. $\mathbb{G}$ is a DAG over **V**. $\mathbf{I}(\mathbb{G})$ is the set of d-separations in $\mathbb{G}$. $\mathbb{S}$ is an undirected graph. **Sepset** is a map from pairs of nodes to sets of nodes. In the output, $\mathbb{S}$ is the skeleton of $\mathbb{G}$ and **Sepset** is a map from pairs of nodes in **V** to sets of nodes in **V** that d-separate them in $\mathbb{G}$.

Input: $\mathbf{I}(\mathbb{G})$
Output: $\mathbb{S}$ and **Sepset**

1: **for each** $X \in \mathbf{V}$ **do**
2: let $\mathbf{TA}(X)$ and $\mathbf{Sepset}_X$ be the output of Algorithm 3 with $\mathbf{I}(\mathbb{G})$ and X as input.
3: **end for**
4: initialise $\mathbb{S}$ with the complete undirected graph over **V**
5: initialise **Sepset** with the empty map
6: **for each** pair $\{X,Y\} \in \mathbf{V}$ **do**
7: **if** $X \in \mathbf{TA}(Y)$ and $Y \in \mathbf{TA}(X)$ **then**
8: add edge $X - Y$ to $\mathbb{S}$
9: **else if** $X \notin \mathbf{TA}(Y)$ **then**
10: $\mathbf{Sepset}(\{X,Y\}) \leftarrow \mathbf{Sepset}_Y(\{X,Y\})$
11: **else**
12: $\mathbf{Sepset}(\{X,Y\}) \leftarrow \mathbf{Sepset}_X(\{X,Y\})$
13: **end if**
14: **end for**

2.3.2 Hypothesis tests of conditional independence

As mentioned above, the d-separations in the true causal DAG are typically determined by assuming that the CFC holds and performing hypothesis tests of conditional independence on a random sample from the probability distribution of the variables. When all variables are categorical, the G test is usually used [Tsamardinos et al., 2006, Aliferis et al., 2010a]. The test uses the G statistic, which asymptotically follows the χ^2 distribution with df degrees of freedom when the null hypothesis is true. In the absence of *structural zeros* (zeros in cells of the contingency table that correspond to zeros in the probability distribution), the degrees of freedom corresponding to the hypothesis of conditional independence of X and Y given **Z** are given by the following equation [Spirtes et al., 2000]:

$$df = (|\mathbf{D}_X| - 1)(|\mathbf{D}_Y| - 1) \prod_{Z \in \mathbf{Z}} |\mathbf{D}_Z|$$

where $\mathbf{D}_X$ is the domain of X. If the row where $X = x$ or the column where $Y = y$ in the conditional contingency table where $\mathbf{Z} = \mathbf{z}$ is comprised of structural zeros (that is, if $X \neq x$ or $Y \neq y$ when $\mathbf{Z} = \mathbf{z}$), the cardinality of the domain (number of levels) of X or Y, respectively, corresponding to that table must be reduced by one. Tsamardinos et al. [2006] reduce the number of levels of X or Y corresponding to each conditional contingency table by one for each all-zero row or column, respectively, of the table:

$$df = \sum_{\mathbf{z} \in \mathbf{D}_\mathbf{Z}} \max\{|\mathbf{D}_X| - 1 - \sum_{x \in \mathbf{D}_X} [1 - I(N_{x\mathbf{z}})], 0\} \cdot \max\{|\mathbf{D}_Y| - 1 - \sum_{y \in \mathbf{D}_Y} [1 - I(N_{y\mathbf{z}})], 0\}$$

where $N_{x\mathbf{z}}$ ($N_{y\mathbf{z}}$) is the number of observations with $X = x$ ($Y = y$) and $\mathbf{Z} = \mathbf{z}$ in the sample, and function $I(x)$ is 1 when $x > 0$ and 0 otherwise. This practice is referred to as the *degrees of freedom adjustment heuristic*. Tsamardinos et al. [2006] ignore a test with $df = 0$. Armen and Tsamardinos [2014] showed that this can only happen in the case of *deterministic* relations (when the value of a variable is exactly determined by the values of other variables) or the *appearance* of deterministic relations due to insufficient sample size. They set the p-value of such a test to one, which is referred to as *determinism detection*. Armen and Tsamardinos [2014] found that determinism detection results in greatly reduced execution times and more accurate estimation and control of the FDR (see Section 2.3.3) in some cases.

A test is usually performed only if it is reliable according to some *reliability criterion*. A test is considered reliable if it satisfies the assumptions about the distribution of the statistic used and has sufficient power [Fast, 2010]. When a test is unreliable, a *default decision* is made. In MMPC, the default decision is independence when the conditioning set is empty; otherwise, it is dependence. [see Tsamardinos et al., 2006, Section 1.1, for a detailed justification].

For categorical variables, Fienberg [1977] recommends that, on average, at least five observations per cell of the contingency table occur for the test to be reliable. The lower limit on the average number of observations per cell is called the *heuristic power size* (denoted by *h-ps*) and the corresponding reliability criterion is referred to as the *heuristic power rule*.

2.3.3 Multiple testing

Beyond structure learning, it is important to assess the confidence on the learnt structure. This can be achieved by viewing link identification as multiple hypothesis testing, each null hypothesis being the absence of a link, and controlling an appropriate error rate.

Let $(\mathbb{G}, \mathscr{P})$ be a CBN over $\mathbf{V}$ and $X, Y \in \mathbf{V}$. The hypothesis of absence of a link between X and Y in $\mathbb{G}$ is equivalent to the union of the hypotheses of d-separation of X and Y given each subset of $\mathbf{V} \setminus \{X, Y\}$ in $\mathbb{G}$; the latter is equivalent to the union of the hypotheses of d-separation of X and Y given each set in a collection of subsets of $\mathbf{V} \setminus \{X, Y\}$ in $\mathbb{G}$ that would include a sepset of X and Y if X and Y were not adjacent. In the following, such a collection is called *separation-sufficient* for X and Y in $\mathbb{G}$. Clearly, the powerset of $\mathbf{V} \setminus \{X, Y\}$, a collection of subsets of $\mathbf{V} \setminus \{X, Y\}$ that includes all subsets of the nodes adjacent to X and all subsets of the nodes adjacent to Y in $\mathbb{G}$, and a collection of subsets of $\mathbf{V} \setminus \{X, Y\}$ that includes a sepset of X and Y when X and Y are actually not adjacent, are all separation-sufficient for X and Y in $\mathbb{G}$. Under the CFC, d-separations in $\mathbb{G}$ are equivalent to conditional independences in $\mathscr{P}$. The p-value of the hypothesis of absence of a link between X and Y in $\mathbb{G}$ is then upper-bounded by the maximal among the p-values corresponding to the hypotheses of independence of X and Y given each set in a collection that is separation-sufficient for X and Y in $\mathbb{G}$, based on the following theorem by Tsamardinos and Brown [2008]:

Theorem 2.3. *Let* $\mathbf{V}$ *be a set of random variables,* $X, Y \in \mathbf{V}$*, and* $\mathbf{S}$ *be a collection of subsets of* $\mathbf{V} \setminus \{X, Y\}$*. The p-value corresponding to the hypothesis that there is no set in* $\mathbf{S}$ *that renders* X *and* Y *independent is upper-bounded by the maximal among the p-values corresponding to the hypotheses that* X *and* Y *are independent given* $\mathbf{Z}$ *for each* $\mathbf{Z} \in \mathbf{S}$*.*

Suppose that Algorithm 1, 2, 5, or 6 is applied to the set of d-separations in $\mathbb{G}$ determined by performing hypothesis tests of conditional independence to a random sample from $\mathscr{P}$ in order to determine the existence of a link between X and Y (among others) and that (1) all tests considered by the algorithm are reliable and

(2) performed tests never produce a type II error. Then, it is not hard to see that the algorithm identifies a link between X and Y if such a link exists. In addition, Algorithm 1 performed the tests of conditional independence of X and Y given each subset $\mathbf{V} \setminus \{X, Y\}$ when a link between X and Y is discovered and given a sepset of X and Y when a link between X and Y is not discovered (and therefore X and Y are not adjacent in $\mathbb{G}$), while Algorithms 2, 5, and 6 performed the tests of conditional independence of X and Y given each subset of the nodes adjacent to X (other than Y if X and Y are adjacent) and each subset of the nodes adjacent to Y (other than X if X and Y are adjacent) in $\mathbb{G}$ when a link between X and Y is discovered and given a sepset of X and Y when a link between X and Y is not discovered. As the collection of conditioning sets of the tests of conditional independence of X and Y performed by each algorithm is separation-sufficient for X and Y in $\mathbb{G}$, Theorem 2.3 implies that the p-value corresponding to the hypothesis of absence of a link between X and Y is upper-bounded by the maximal among the p-values from those tests. Condition (2) above may seem unrealistic, but the type II error rate of likelihood-ratio tests such as the G test actually approaches zero as the sample size approaches infinity when the significance level of the test is fixed [see Li and Wang, 2009, Appendix B].

Under the conditions above and the additional (unrealistic) condition that the link-absence p-values are independent, Algorithms 1, 2, 5, or 6 control the *false positive rate (FPR)* among the identified links. This is because a link-absence hypothesis is accepted once a conditional-independence p-value exceeds α, the significance level of the test, or, equivalently, if the maximal among the p-values of the conditional-independence tests performed for the link exceeds α. When it is of interest to have mostly true positives among the rejected hypotheses, it is appropriate to control the FDR. Tsamardinos and Brown [2008] proposed performing parent-and-children learning with significance level α and then estimating the FDR among the learned parents and children using the maximal conditional-independence p-values. Armen and Tsamardinos [2014] adapted the approach of Tsamardinos and Brown [2008] to skeleton identification and unified it with FDR control, thus

proposing to perform skeleton identification with significance level α and then to estimate or control the FDR among the identified links. This approach can be adapted to any kind of link identification, including parent-and-children learning: perform link identification with significance level α and estimate or control the FDR among the identified links. Since the link-absence p-values are dependent, an appropriate way to control the FDR is to apply the BY procedure.

After an appropriate error-controlling procedure has been applied, the sepsets corresponding to the links that were discarded by the procedure need to be set. Such a sepset can be simply set to the conditioning set corresponding to the maximal conditional-independence p-value for the link.

2.3.4 Dealing with violations of the CFC

A violation of the CFC called *triangle unfaithfulness* occurs when the probability distribution is not faithful to the true causal DAG but is faithful to some other causal DAG. When this is the case, violations of the CFC are undetectable; the opposite is true when *triangle faithfulness* holds [see Zhang and Spirtes, 2008, for the exact definition]. A causal DAG $\mathbb{G}$ satisfies the *causal minimality condition* with probability distribution $\mathscr{P}$ if $\mathbb{G}$ satisfies the CMC with $\mathscr{P}$ and no longer satisfies the CMC with $\mathscr{P}$ if an edge is removed from $\mathbb{G}$. It can be proved that this is equivalent to requiring that for each edge $X \to Y$ of $\mathbb{G}$, Y is dependent to X given the set $\mathbf{PA}(Y) \setminus \{X\}$ of the parents of Y other than X. The CFC implies the causal minimality condition [Neapolitan, 2004]. Lemeire et al. [2012] showed that, when triangle faithfulness and the causal minimality condition are satisfied, violations of the CFC are of three types, namely *pseudo-independent relations (PIRs)*, *information equivalences*, and *2-1 conditional independence (CI) patterns*. The first two types are discussed below. 2-1 CI patterns are quite complicated and are not discussed here [see Lemeire et al., 2012, for the exact definition].

Suppose that X and Y are adjacent in a causal DAG over $\mathbf{V}$ that satisfies the causal minimality condition with some probability distribution. The edge between X and Y is called a PIR if there is a subset $\mathbf{S}$ of $\mathbf{V} \setminus \{X, Y\}$ such that X and Y are independent given $\mathbf{S}$ and X and Y are strictly d-separated by $\mathbf{S}$ in the causal DAG

without the edge [Lemeire et al., 2012]. Owing to the causal minimality condition, if X is a parent of Y, X and Y are dependent given the parents of Y other than X, and if Y is a parent of X, X and Y are dependent given the parents of X other than Y. Thus, there is at least one subset $\mathbf{D}$ of $\mathbf{V} \setminus \{X, Y\}$ such that X and Y are dependent given $\mathbf{D}$; if $\mathbf{D}$ is also minimal (that is, X and Y are conditional independent given any proper subset of $\mathbf{D}$), it is called a *depset* of X and Y. An example of a PIR is the logical XOR (eXclusive OR) relationship: assume that X, Y, and Z are boolean variables, $X \rightarrow Z \leftarrow Y$ is the true causal DAG over $\{X, Y, Z\}$, and Z is the XOR of X and Y. Then X and Z are marginally independent, which is a violation of the CFC, but become dependent when conditioned on Y; similarly, Y and Z are marginally independent and dependent given X. Thus, $X \rightarrow Z$ and $Y \rightarrow Z$ are PIRs with depset $\{Y\}$ and $\{X\}$, respectively. Epistasis with absent marginal effects in genetics is actually an example of a PIR (See Section 3.3).

Information equivalences occur when different sets of variables contain the same information about a target variable [Lemeire et al., 2012]. Deterministic relations are a simple form of information equivalence. Suppose that $X \rightarrow Y \rightarrow Z$ is a causal DAG and Y is a function of X (X exactly determines Y). Then X and Y contain the same information for Z, and Y and Z are independent given X, which a violation of the CFC. In the presence of information equivalences, Markov boundaries are not necessarily unique: in the previous example, both $\{X\}$ and $\{Y\}$ are Markov boundaries of Z. Algorithms instantiating the *Target Information Equivalence (TIE*)* algorithm template [Statnikov et al., 2013] can learn all Markov boundaries of a target node in a CBN with information equivalences.

2.4 Structure learning under causal insufficiency and selection bias

Causal sufficiency is the exception rather than the norm; variables in real-life samples often have hidden common causes. This means that the sample is from the *marginal* of a distribution over the set of observed variables and their hidden common causes. Furthermore, the samples themselves are biased (selection bias is

present), meaning that they are not random samples from the distribution of interest, but random samples from the *conditional* distribution of the observed variables given an instantiation of a set of hidden *selection variables*. A notable example of non-random sampling is case–control sampling. The causal and probabilistic relationships between a set of random variables in the presence of hidden common causes and selection bias can be represented by a *causal maximal ancestral graph* [Richardson and Spirtes, 2002].

A *maximal ancestral graph (MAG)* is a special type of a *mixed graph*, which is a generalisation of undirected and directed graphs. A mixed graph has two types of edge endpoints: tail (–) and arrowhead (>), and therefore three types of edges, $X \rightarrow Y$, $X \leftrightarrow Y$, and $X - Y$. The definition of parent, child, directed path, ancestor, and descendant carries on from directed graphs. If there is an edge $X \leftrightarrow Y$ in a mixed graph, then X and Y are called *spouses*. X and Y are said to be *neighbours* if there is an edge $X - Y$. The definition of *anterior path* and *anterior* in mixed graphs is a generalisation of the definition of directed path and ancestor, respectively, in directed graphs. A path from X to Y where all edges are either undirected or directed towards Y is said to be *anterior*. If there is an anterior path from X to Y or $X = Y$, then X is *anterior* to Y. An *almost directed cycle* between X and Y occurs when X is both a spouse and an ancestor of Y. A mixed graph is said to be *ancestral* (and called an *ancestral graph*) if it satisfies the following conditions [Zhang, 2008]:

1. There is no directed cycle.

2. There is no almost directed cycle.

3. There is no undirected edge $X - Y$ such that X or Y has parents or spouses.

Clearly, DAGs are ancestral graphs. The generalisation of (strict) d-separation and (strict) d-connection in mixed graphs is called *(strict) m-separation* and *(strict) m-connection*, respectively, and their definition and notation are exactly the same. The definition of collider is generalised, however. Let $*$ denote either type of edge endpoint. If Z is an interior node on a path p and $X* \rightarrow Z \leftarrow *Y$ on p, then Z is called a collider on p (the triple $[X,Y,Z]$ is also referred to as a collider). An ancestral graph

is called *maximal* if for every pair of non-adjacent nodes there is a subset of the rest nodes that m-separates them.

A MAG can be obtained from another MAG through *marginalisation and conditioning* [Richardson and Spirtes, 2002]. The marginal/conditional MAG over $\mathbf{O}$ given $\mathbf{S}$ of a MAG $\mathbb{M}$ over $\mathbf{O} \dot{\cup} \mathbf{H} \dot{\cup} \mathbf{S}$ is a MAG where

- X and Y are adjacent if and only if there is no $\mathbf{Z} \subseteq \mathbf{O} \setminus \{X, Y\}$ such that X and Y are m-separated in $\mathbb{M}$ given $\mathbf{Z} \cup \mathbf{S}$, and
- if X and Y are adjacent, then the edge between X and Y is into X ($X \leftarrow\!* Y$) if and only if X is not anterior in $\mathbb{M}$ of Y, or of any node in $\mathbf{S}$.

The nodes in $\mathbf{O}$, $\mathbf{H}$, and $\mathbf{S}$ are called *observed*, *hidden*, and *selection* nodes, respectively. As it is the case with probability distributions, marginalisation/conditioning of MAGs is commutative: Let $\mathbb{M}$ be a MAG over $\mathbf{O}_1 \dot{\cup} \mathbf{O}_2 \dot{\cup} \mathbf{H} \dot{\cup} \mathbf{S}_1 \dot{\cup} \mathbf{S}_2$. The marginal/conditional over $\mathbf{O}_1 \cup \mathbf{O}_2$ given $\mathbf{S}_1 \cup \mathbf{S}_2$ of $\mathbb{M}$ is the same as the marginal/conditional over $\mathbf{O}_2$ given $\mathbf{S}_2$ of the marginal/conditional over $\mathbf{O}_1$ given $\mathbf{S}_1$ of $\mathbb{M}$.

Let $\mathbb{G}$ be a DAG. It is clear that the edges in a marginal/conditional MAG $\mathbb{M}$ imply the following ancestral relationships between the nodes in $\mathbb{G}$:

- $X \rightarrow Y$ implies that X is a ancestor in $\mathbb{G}$ of Y or of some selection node, but Y is not an ancestor in $\mathbb{G}$ of X or any selection node.
- $X \leftrightarrow Y$ implies that X is not an ancestor in $\mathbb{G}$ of Y or of any selection node, and Y is not an ancestor in $\mathbb{G}$ of X or of any selection node.
- $X - Y$ implies that X is an ancestor in $\mathbb{G}$ of Y or of some selection node, and Y is an ancestor in $\mathbb{G}$ of X or of some selection node.

Suppose that $\mathbb{G}$ is a *causal* DAG and causal transitivity is satisfied. Then the edges in a marginal/conditional MAG $\mathbb{M}$ imply the following causal relationships between the variables, since X is a cause of Y if and only if there is a directed path from X to Y in $\mathbb{G}$:

- $X \rightarrow Y$ implies that X is a cause of Y or of some selection variable, and Y is not a cause of Y or of some selection variable.

- $X \leftrightarrow Y$ implies that X is not a cause of Y or of any selection variable, and Y is not a cause of X or of any selection variable.

- $X - Y$ implies that X is a cause of Y or of some selection variable, and Y is a cause of X or of some selection variable.

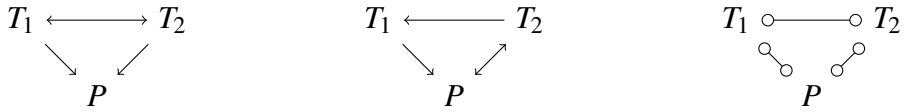
A MAG which is a marginal/conditional of a causal DAG is called a *causal MAG*. Figure 2.2a shows the causal MAG over the subset $\{T_1, T_2, P\}$ of the causal DAG in Figure 2.1a, assuming selection bias is absent. There is a bidirected edge between T_1 and T_2, because T_1 and T_2 do not cause each other or any selection variable.

It is evident that the causal interpretation of the edges in causal MAGs is not as straightforward as in causal DAGs. For example, in the absence of selection bias, edge $X \rightarrow Y$ implies that X is a cause of Y; X, however, is not necessarily a *direct* cause of Y with respect to the set of variables in the MAG. Furthermore, X and Y may or may not have a hidden common cause. Borboudakis et al. [2012] devised algorithms for distinguishing between these cases. The presence of selection bias further complicates the causal interpretation of the edges. Fortunately, the edges in the special MAGs introduced in the next chapter to describe the causal relationships between the variables in a genetic dataset have clear interpretations which follow from domain assumptions.

An important property of the marginal/conditional $\mathbb{N}$ over $\mathbf{O}$ given $\mathbf{S}$ of a MAG $\mathbb{M}$ is that X and Y are m-separated by $\mathbf{Z}$ in $\mathbb{N}$ if and only if X and Y are m-separated by $\mathbf{Z} \cup \mathbf{S}$ in $\mathbb{M}$ [Richardson and Spirtes, 2002]:

$$X \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{N}) \iff X \perp Y \mid \mathbf{Z} \cup \mathbf{S} \in \mathbf{I}(\mathbb{M})$$

Suppose that $(\mathbb{G}, \mathscr{P})$ is a CBN and let $\mathbb{M}$ be the marginal/conditional over $\mathbf{O}$ given $\mathbf{S}$ of $\mathbb{G}$. X and Y are m-separated by $\mathbf{Z}$ in $\mathbb{M}$ if and only if X and Y are



(a) The causal MAG over variables $\{T_1, T_2, P\}$ of the causal DAG in Figure 2.1a.

(b) A causal MAG that is Markov equivalent to the one in Figure 2.2a.

(c) The maximally-informative causal PAG of the Markov equivalence class to which the causal MAGs in Figures 2.2a and 2.2b belong to.

Figure 2.2: Example of a causal MAG, a Markov-equivalent causal MAG, and a maximally-informative causal PAG. T_i is the level of the transcript of gene i in the cell, and P is a phenotype.

m-separated by $\mathbf{Z} \cup \mathbf{S}$ in $\mathbb{G}$:

$$X \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{M}) \iff X \perp Y \mid \mathbf{Z} \cup \mathbf{S} \in \mathbf{I}(\mathbb{G})$$

Assuming that $(\mathbb{G}, \mathscr{P})$ satisfies the CFC, X and Y are m-separated by $\mathbf{Z} \cup \mathbf{S}$ in $\mathbb{G}$ if and only if X and Y are independent given $\mathbf{Z} \cup \mathbf{S}$ in $\mathscr{P}$:

$$X \perp Y \mid \mathbf{Z} \cup \mathbf{S} \in \mathbf{I}(\mathbb{G}) \iff X \perp\!\!\!\perp Y \mid \mathbf{Z} \cup \mathbf{S} \in \mathbf{I}(\mathscr{P})$$

Let $\mathscr{Q}$ be the marginal/conditional over $\mathbf{O}$ given $\mathbf{S} = \mathbf{s}$ of $\mathscr{P}$, where $\mathbf{s}$ is an instantiation of $\mathbf{S}$. If X and Y are independent given $\mathbf{Z} \cup \mathbf{S}$ in $\mathscr{P}$ then X and Y are independent given $\mathbf{Z}$ in $\mathscr{Q}$:

$$X \perp\!\!\!\perp Y \mid \mathbf{Z} \cup \mathbf{S} \in \mathbf{I}(\mathscr{P}) \implies X \perp\!\!\!\perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathscr{Q})$$

If we assume that the converse of the above is also true (which is the case for Gaussian distributions [Richardson and Spirtes, 2002]), then

$$X \perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathbb{M}) \iff X \perp\!\!\!\perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathscr{Q})$$

$\mathbb{M}$ is then said to be a *perfect map* of $\mathscr{Q}$. Equivalently, we may not assume that

$(\mathbb{G}, \mathscr{P})$ satisfies the CFC and directly assume the following instead:

$$X \perp Y \mid \mathbf{Z} \cup \mathbf{S} \in \mathbf{I}(\mathbb{G}) \iff X \perp\!\!\!\perp Y \mid \mathbf{Z} \in \mathbf{I}(\mathscr{Q})$$

This is called the *selection bias causal assumption* [Spirtes et al., 1999].

The notion of *inducing path* is central to marginalisation and conditioning of MAGs. In a mixed graph $\mathbb{M}$ over $\mathbf{O} \dot{\cup} \mathbf{H} \dot{\cup} \mathbf{S}$, a path p between X and Y is said to be *inducing with respect to* $\mathbf{H}$ *and* $\mathbf{S}$ if every interior node on p is either in $\mathbf{H}$ or a collider on p, and every collider on p is an ancestor of X, of Y, or of some selection node. An inducing path with respect to $\emptyset$ and $\emptyset$ is said to be *primitive*. The following theorem relates relates inducing paths and m-separations in a MAG to adjacencies and m-separations in a marginal/conditional MAG [Richardson and Spirtes, 2002]:

Theorem 2.4. *Let $\mathbb{M}$ be a MAG over $\mathbf{V} = \mathbf{O} \dot{\cup} \mathbf{H} \dot{\cup} \mathbf{S}$. Then the following statements are equivalent:*

- *X is adjacent to Y in the marginal of $\mathbb{M}$ over $\mathbf{O}$ given $\mathbf{S}$.*
- *There is an inducing path in $\mathbb{M}$ between X and Y with respect to $\mathbf{H}$ and $\mathbf{S}$.*
- *There is no $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{S} \cup \mathbf{H} \cup \{X, Y\})$ such that X and Y are m-separated in $\mathbb{M}$ by $\mathbf{Z} \cup \mathbf{S}$.*
- *There is no $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{S} \cup \mathbf{H} \cup \{X, Y\})$ such that X and Y are m-separated in the marginal of $\mathbb{M}$ over $\mathbf{O}$ given $\mathbf{S}$ by $\mathbf{Z}$.*

The following lemma relates primitive inducing paths to maximality [Richardson and Spirtes, 2002]:

Lemma 2.1. *An ancestral graph is maximal if and only if there is no primitive inducing path between any two non-adjacent nodes in the graph.*

Since many DAGs can have the same d-separations, many MAGs can have the same m-separations. These MAGs are called Markov equivalent and said to belong to the same Markov equivalence class. Markov equivalence for MAGs is

based on the notion of *discriminating path* [Spirtes et al., 2000]. In a MAG, a path $p = [X, \ldots, W, Z, Y]$ is said to be *discriminating* for triple $[W, Z, Y]$ if

- X is not adjacent to Y, and
- every interior node on $p(X, Z)$ is a collider on p and a parent of Y

In this work, node Z is also called discriminated if there is a discriminated triple $[W, Z, Y]$. The following theorem [Spirtes and Richardson, 1996] gives necessary and sufficient conditions for Markov equivalence of MAGs:

Theorem 2.5. *Two MAGs over the same set of variables are Markov equivalent if and only if*

- *they have the same adjacencies,*
- *they have the same unshielded colliders, and*
- *if a path is discriminating for a triple in both MAGs, then the triple is a collider on the path in one MAG if and only if it is a collider on the path in the other MAG.*

Figure 2.2b shows a causal MAG that is Markov equivalent to the one in Figure 2.2a.

A Markov equivalence class of MAGs can be represented by a so-called *maximally-informative partial ancestral graph (PAG)*, which is a type of *partially-oriented mixed graph*. The latter is a graph with three types of edge endpoints: tail (–), arrowhead (>), and circle (○). A non-circle and a circle endpoint is called *oriented* and *unoriented*, respectively. A PAG for a Markov equivalence class of MAGs is a partially-oriented mixed graph that has the same skeleton as the MAGs in the class, and every non-circle endpoint is invariant within the class. If every circle in the PAG corresponds to a variant endpoint in the class, then the PAG is called the *maximally informative PAG* for the class [Spirtes et al., 2000]. A PAG for a Markov equivalence class of causal MAGs is called a *causal PAG*. A causal PAG for the Markov equivalence class of the true causal MAG over a set of variables is called a

true causal PAG over the variables. Figure 2.2c shows the maximally-informative causal PAG for the Markov equivalence class to which the causal MAGs in Figures 2.2a and 2.2b belong to.

The *Fast Causal Inference (FCI)* algorithm [Spirtes et al., 2000, Zhang, 2008] can be used to learn the maximally-informative causal PAG over a set of variables given the set of m-separations in the causal MAG over the variables. In the skeleton-identification phase, the skeleton of the PAG is first learnt as in the PC algorithm (Algorithm 2). However, it is no longer sufficient to search among the subsets of the nodes adjacent to X and the subsets of the nodes adjacent to Y in order to find a sepset. Nevertheless, false positives in the skeleton can be eliminated by searching among the subsets of the *Possible D-SEP* sets [see Spirtes et al., 2000, for definition] for a sepset. In the orientation phase, a set of 11 rules is applied to orient the edges in the PAG [Zhang, 2008]. The orientation rules are based on the following lemmas [Spirtes et al., 2000], which are used in proofs in this work as well:

Lemma 2.2. *Let $p = [X,Y,Z]$ be an unshielded triple and suppose that* $\mathbf{S}$ *is a sepset of X and Z in a MAG. p is a collider if and only if $Y \notin \mathbf{S}$.*

Lemma 2.3. *Suppose that $[X,\ldots,W,Y,Z]$ is a discriminating path for Y and* $\mathbf{S}$ *is a sepset of X and Z in a MAG. $[W,Y,Z]$ is a collider if and only if $Y \notin \mathbf{S}$.*

Analogously to structure learning under causal sufficiency and no selection bias, the d-separations are determined in practice by making the selection bias causal assumption and performing hypothesis tests of conditional independence on a random sample from the probability distribution of the variables.

2.5 Estimating causal effects

In contrast to an association effect of Y on X, which is some contrast between the distributions of Y conditional on instantiations of X, a *causal* effect is some contrast between the distributions of Y conditional on *manipulations* of X [Didelez et al., 2010]. The probability of set of variables $\mathbf{Y}$ after manipulating set of variables $\mathbf{X}$ to take on values $\mathbf{x}$ is represented by $\Pr(\mathbf{Y} \mid \mathrm{do}(\mathbf{X} = \mathbf{x}))$. The causal DAG that describes

the causal relationships between a set of variables $\mathbf{V} = \{V_1, \ldots, V_n\}$ after a subset $\mathbf{X}$ of the variables have been manipulated to take on values $\mathbf{x}$ is the pre-manipulation causal DAG with all edges into the variables in $\mathbf{X}$ removed. The post-manipulation distribution of the variables satisfies the CMC with the post-manipulation causal DAG [Pearl, 2009] and therefore factorises according to that DAG:

$$\mathscr{P}(\mathbf{v} \mid \mathrm{do}(\mathbf{X} = \mathbf{x})) = \prod_{i=1}^{n} \mathscr{P}(v_i \mid \mathbf{pa}(V_i), \mathbf{X} = \mathbf{x})$$

Thus, given a CBN, any causal effect of interest can be computed from the post-manipulation CBN by performing Bayesian-network inference. Estimating causal effects under causal insufficiency and selection bias has been the focus of recent work [e.g. see Correa and Bareinboim, 2017]. In the case of genetic case–control datasets, the allelic *causal odds ratio (COR)* is defined as follows [Didelez et al., 2010]:

$$\begin{aligned} \mathrm{COR} &\triangleq \frac{\Pr(P = \text{affected} \mid \mathrm{do}(I = \text{true})) / \Pr(P = \text{unaffected} \mid \mathrm{do}(I = \text{true}))}{\Pr(P = \text{affected} \mid \mathrm{do}(I = \text{false})) / \Pr(P = \text{unaffected} \mid \mathrm{do}(I = \text{false}))} \\ &= \frac{\Pr(P = \text{affected} \mid \mathrm{do}(I = \text{true})) \cdot \Pr(P = \text{affected} \mid \mathrm{do}(I = \text{false}))}{\Pr(P = \text{unaffected} \mid \mathrm{do}(I = \text{true})) \cdot \Pr(P = \text{unaffected} \mid \mathrm{do}(I = \text{false}))} \end{aligned}$$

The genotypic CORs are similarly defined. The following adaptation of a result by Didelez et al. [2010] gives sufficient graphical conditions, assuming a single logical selection variable, for a COR in the general population to equal the respective OR in the sampled population. The former can therefore be estimated by the sample estimate of the latter.

Theorem 2.6. *Suppose that X and Y are categorical random variables, S is a logical selection variable, $\mathscr{P}$ is the distribution of $\{X, Y, S\}$, and $\mathscr{Q}$ is the conditional of $\mathscr{P}$ given $S = \text{true}$. If the following conditions are true in every causal DAG that includes X, Y, and S:*

1. *Y is not an ancestor of X.*

2. *X and T do not have a common ancestor.*

3. X and S are d-separated by Y.

then the following holds for every pair (x_1, x_2) *of values of X and every pair* (y_1, y_2) *of values of Y:*

$$\begin{aligned}&\frac{\mathscr{P}(Y = y_1 \mid do(X = x_1))/\mathscr{P}(Y = y_2 \mid do(X = x_2))}{\mathscr{P}(Y = y_2 \mid do(X = x_1))/\mathscr{P}(Y = y_1 \mid do(X = x_2))}\\ =&\frac{\mathscr{Q}(Y = y_1 \mid X = x_1)/\mathscr{Q}(Y = y_2 \mid X = x_2)}{\mathscr{Q}(Y = y_2 \mid X = x_1)/\mathscr{Q}(Y = y_1 \mid X = x_2)}\end{aligned}$$

2.6 Structure learning from samples with overlapping sets of variables

There are datasets in many scientific areas defined over the same variables. In genetics, for example, several datasets may be available over the same set of SNPs because the same genotyping chip was used. It is often desirable to analyse these datasets together, usually as a means of increasing the power of detecting statistical associations compared to analysing the datasets individually. One way to do this is to conduct a mega-analysis of the datasets, as mentioned in the introduction. The datasets, however, may be incompatible because they were generated from different populations using different technologies and therefore follow different distributions [Tillman, 2009]. A meta-analysis circumvents this problem by using some method for combining the p-values (or other summary measures) from the single datasets. Among such methods, Fisher's method [Fisher, 1925] is considered the most reliable [Lazar et al., 2002].

The term *causal mega-analysis* could be used to refer to concatenating multiple datasets and applying constraint-based causal learning to the resulting dataset. Clearly, causal mega-analysis is affected by the same problems as association mega-analysis. The term *causal meta-analysis* is used here to refer to constraint-based causal learning from multiple datasets by combining the p-values from the single datasets. This is a more general term than *Bayesian-network meta-analysis*, introduced by Tsamardinos and Borboudakis [2010] to refer specifically to learning (not necessarily causal) BNs from multiple datasets over the same (or semantically sim-

ilar) variables. In Bayesian-network meta-analysis, structure learning is applied to the set of variables at hand as usual but each conditional independence test is performed on all datasets and the resulting p-values are combined. [Tillman, 2009] compared several methods of combining p-values for the task of Bayesian-network meta-analysis and concluded that Fisher's method performs best.

It is not uncommon to have datasets defined not over exactly the same set of variables but over *overlapping* sets of variables. For example, GWAS datasets from the same disease may have been generated using different versions of a genotyping chip, each assessing a slightly different set of SNPs; a GWAS and an exome-sequencing dataset from the same disease share the phenotype and some exonic SNPs. If a set of datasets with the same phenotype are identically distributed, they can be concatenated and association analysis can be conducted on the concatenated dataset; each test of association of a SNP with the phenotype can use the observations of the SNP and the phenotype that have no values missing for the SNP. In practice, genotype imputation is performed (see Introduction). If the datasets are not identically distributed, it is possible to conduct meta-analysis instead.

In causal discovery, analysing datasets with overlapping sets of variables is more complicated. Datasets can be concatenated if they are identically distributed. In the concatenated dataset, however, observations of variables never observed together in the original datasets would always have some of their values missing, rendering conditional-independence testing and therefore, constraint-based learning, impossible. As Danks et al. [2008] explain, imputation procedures may not be applied before constraint-based learning from the concatenated dataset. These procedures assume some underlying model, estimate its parameters using the available data, and set the missing values of each observation to their expected values given the available values of the observation. Then conditional independencies in the imputed dataset that involve only variables never observed together in the original datasets are based solely on assumptions and may be incorrect.[1] The application of

[1]Nevertheless, the application of genotype imputation to a concatenated genetic dataset may still be justified, at it is based on a separate reference dataset that contains all the SNPs in the concatenated dataset.

the standard Bayesian algorithm for learning BNs from datasets with missing values, *Bayesian Structural EM* [Friedman, 1998], to a concatenated dataset is also not justified, as the algorithm assumes that the data are missing in random. Danks et al. [2008] showed that the algorithm is indeed highly unsuccessful when applied to datasets resulting from concatenation of datasets with overlapping sets of variables. Owing to these problems, applying causal discovery to a concatenated dataset is not a viable option. An alternative option is to learn all the causal relationships of interest that are consistent with the datasets.

The goal of constraint-based causal learning with overlapping sets of variables is to learn features of a causal MAG given the sets of m-separations in marginals over overlapping sets of variables of the causal MAG. For each set of variables that are never observed together, it is unknown which m-separations that involve all the variables in the set hold. Therefore, there are many possible sets of m-separations over all variables and, subsequently, many possible Markov equivalence classes of causal MAGs over the variables, each represented by a different maximally-informative causal PAG. Each of these causal MAGs and maximally-informative causal PAGs are said to be *consistent* with the marginals of the original causal MAG:

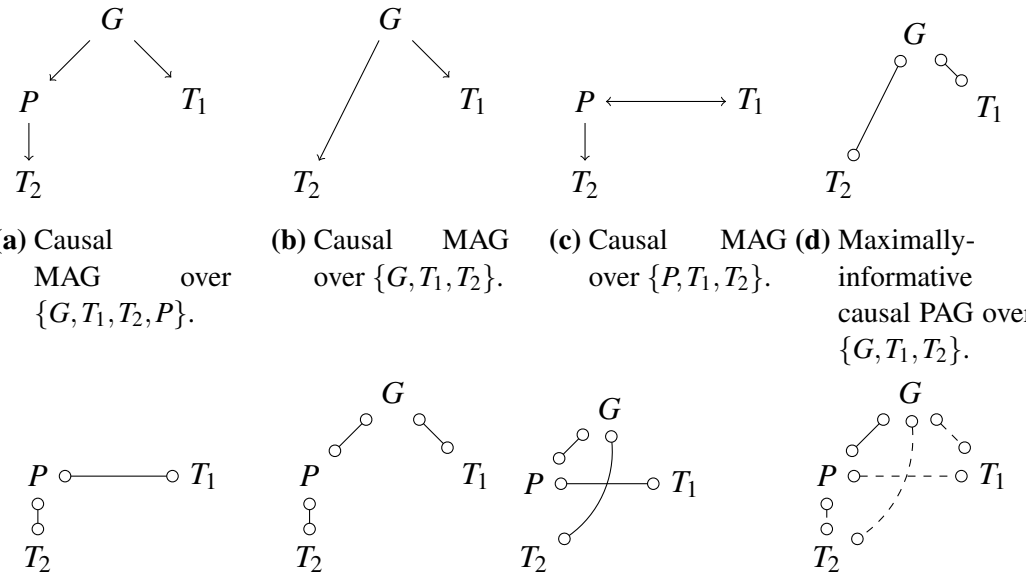
Definition 2.1 (Consistent causal MAG)**.** *Let* $\mathbb{M}$ *be a causal MAG over* $\mathbf{V}$*,* $\mathbf{O}_1, \ldots, \mathbf{O}_n \subseteq \mathbf{V}$ $(n \geq 1)$*, and* $\mathbb{M}_k$ *be the marginal of* $\mathbb{M}$ *over* $\mathbf{O}_k$ $(1 \leq k \leq n)$*. A causal MAG* $\mathbb{N}$ *over* $\mathbf{V}$ *is said to be* consistent *with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$ *if for each k, the marginal of* $\mathbb{N}$ *over* $\mathbf{O}_k$ *is Markov equivalent to* $\mathbb{M}_k$*.*

Definition 2.2 (Consistent maximally-informative causal PAG)**.** *Let* $\mathbb{M}$ *be a causal MAG over* $\mathbf{V}$*,* $\mathbf{O}_1, \ldots, \mathbf{O}_n \subseteq \mathbf{V}$ $(n \geq 1)$*, and* $\mathbb{M}_k$ *be the marginal of* $\mathbb{M}$ *over* $\mathbf{O}_k$ $(1 \leq k \leq n)$*. A maximally-informative causal PAG* $\mathbb{Q}$ *over* $\mathbf{V}$ *is called* consistent *with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$ *if the members of the class of causal MAGs represented by* $\mathbb{Q}$ *are consistent with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$*.*

The brute-force approach to identifying consistent maximally-informative PAGs is to create every possible partially-oriented mixed graph over all variables and check whether (a) it is a maximally-informative PAG and (b) the marginals of

a MAG in the Markov equivalence class represented by the PAG are Markov equivalent to the marginals of the original MAG. Algorithms for identifying consistent maximally-informative PAGs improve on this approach. Such algorithms are *sound* if they only identify consistent maximally-informative PAGs and *complete* if they identify all consistent maximally-informative PAGs.

Figure 2.3 shows an example of constraint-based causal learning with overlapping sets of variables [adapted from Tsamardinos et al., 2012]. Figure 2.3a shows an example of a causal DAG, where genotype G is a direct cause of disease P and a direct cause of the transcript level T_1 of gene 1, while P is a direct cause of the transcript level T_2 of gene 2. Assume that one research group is interested in discovering eQTLs. The group measures variables $\{G, T_1, T_2\}$ (among others) in healthy subjects. Figure 2.3b and 2.3d shows the causal MAG and the maximally-informative causal PAG, respectively, over $\{G, T_1, T_2\}$. The group finds that G is an eQTL for genes 1 and 2. Further assume that another research group is interested in genes that are differentially expressed in disease P and measures variables $\{P, T_1, T_2\}$ (among others) in healthy and affected subjects. Figures 2.3b and 2.3d show the causal MAG and the maximally-informative causal PAG, respectively, over $\{P, T_1, T_2\}$. The group finds genes 1 and 2 to be differentially expressed in disease. Figure 2.3f shows the maximally-informative PAGs consistent with the MAGs in Figures 2.3b and 2.3c. Finally, Figure 2.3g shows the pairwise causal graph over all variables. The *pairwise causal graph (PCG)* corresponding to the set of maximally-informative causal PAGs consistent with a set of MAGs is a graph over all variables with two kinds of links, solid (—) and dashed (- -), and three kinds of edge endpoints, tail (–), arrowhead (>), and circle (○). A link in the PCG is solid if it is present in all consistent PAGs; otherwise, the link is dashed. Non-circle and circle endpoints correspond to variant and invariant edge endpoints, respectively, within the subset of consistent MAGs where the link is present. PCGs were introduced by Triantafilou et al. [2010] and referred to as *summary graphs* in Triantafillou and Tsamardinos [2015]. The PCG in Figure 2.3g clearly shows that G and P are dependent, even if they were never measured together.



(a) Causal MAG over $\{G, T_1, T_2, P\}$.

(b) Causal MAG over $\{G, T_1, T_2\}$.

(c) Causal MAG over $\{P, T_1, T_2\}$.

(d) Maximally-informative causal PAG over $\{G, T_1, T_2\}$.

(e) Maximally-informative causal PAG over $\{P, T_1, T_2\}$.

(f) The set of maximally-informative causal PAGs consistent with the causal MAGs over $\{G, T_1, T_2\}$ and $\{P, T_1, T_2\}$.

(g) The PCG corresponding to the PAGs in Figure 2.3f.

Figure 2.3: Example of constraint-based causal learning with overlapping sets of variables [adapted from Tsamardinos et al., 2012]. G is the genotype at some locus, T_i is the transcript level of gene i in the cell, and P is a phenotype.

The first algorithm for identifying consistent maximally-informative PAGs was the *Integration of Overlapping Networks (ION)* algorithm [Danks et al., 2008]. The algorithm receives the set of maximally-informative PAGs over the overlapping sets of variables (learned by an algorithm such as FCI) as input, and outputs the set of consistent PAGs. The input PAGs are assumed to be correct. However, when the input PAGs are learnt from data, they often represent conflicting (in)dependencies among their shared variables and ION fails. The *Integration of Overlapping Datasets (IOD)* algorithm [Tillman and Spirtes, 2011], which directly accepts datasets, solves this problem by applying each conditional-independence test to all datasets that contain the test variables and combining the p-values using Fisher's method (Algorithm 7). IOD also conducts a more efficient search procedure than ION. Unfortunately, neither ION or IOD scale beyond a few variables. In Chapter 4, an algorithm inspired by IOD is presented that is specialised for genetic datasets. Although that algorithm is also impractical, it is used as a basis for a

practical local-learning algorithm that targets the phenotype.

Tillman and Spirtes [2011] did not use a reliability criterion for the conditional independence tests in their experiments with IOD. Tsamardinos and Borboudakis [2010] generalised the heuristic power rule for the task of Bayesian-network meta-analysis. A test is now considered reliable if the sum of the average number of observations per cell of the contingency table in each dataset is at least *h-ps*. This criterion can be used with datasets with overlapping sets of variables by considering only the datasets that contain all test variables.

Algorithm 7 Meta-analysis conditional-independence test [Tillman and Spirtes, 2011]. The independence of X and Y given $\mathbf{Z}$ is tested by performing the test on all samples that contain the variables and combining the p-values using Fisher's method. $\mathbf{D} = \{D^1, \ldots, D^n\}$ is a set of samples with corresponding sets of variables $\mathbf{V}_1, \ldots, \mathbf{V}_n$. $p^i_{X \perp\!\!\!\perp Y|\mathbf{Z}}$ is the p-value from the test of conditional independence of X and Y given $\mathbf{Z}$ in sample $\mathbf{D_i}$. $F(x, df)$ is the value of the cumulative $\mathcal{X}^2$ distribution with df degrees of freedom at x.

Input: $\mathbf{D}$, X, Y, $\mathbf{Z}$, α
Output: true or false

$k \leftarrow 0$
for each $D_i \in \mathbf{D}$ **do**
  if $\{X, Y\} \cup \mathbf{Z} \subseteq \mathbf{V_i}$ **then**
    $p^i \leftarrow p^i_{X \perp\!\!\!\perp Y|\mathbf{Z}}$
    $k \leftarrow k + 1$
  else
    $p^i \leftarrow 1$
  end if
end for
if $F(-2\sum_{i=1}^{n} \log p^i, 2k) > \alpha$ **then**
  return true
else
  return false
end if

The cSAT+ algorithm [Triantafilou et al., 2010] converts the problem of identifying consistent maximally-informative PAGs to a *constraint satisfaction problem (CSP)*, and uses a general-purpose solver to solve it efficiently. cSAT+ is more efficient that ION, but also assumes that the input is correct. If the input PAGs are learnt from data, cSAT+ may generate conflicting constraints and the resulting CSP may be unsolvable. The recently-developed *COmbINE (Causal discovery*

from Overlapping INtErventions) algorithm [Triantafillou and Tsamardinos, 2015] directly accepts datasets over overlapping sets of variables that may as well come from different experimental conditions (interventions). After applying FCI to the datasets, COmbINE converts the resulting PAGs to appropriate constraints on the form of the consistent PAGs. The constraints are then ranked using a score derived from the p-values of the conditional independence tests performed, and a constraint is not added to the CSP if it conflicts with another constraint higher in the ranking. The resulting CSP is thus guaranteed to be solvable. COmbINE scales up to 100 variables for sparse networks and was successfully applied on four mass-cytometry datasets with overlapping sets of variables under three interventions [Triantafillou and Tsamardinos, 2015].

In terms of the data-integration framework provided by Hamid et al. [2009], identifying consistent maximally-informative PAGs is a late-stage approach, as the results of causal discovery (PAGs) from each dataset are combined. The data which the algorithms are applied to can be of similar or heterogeneous type: each dataset may contain data of similar or heterogeneous type, and the variables in common contain data of a similar type. Among the data-integration methodologies identified by Lapatas et al. [2015], identifying consistent maximally-informative PAGs is, obviously, a dataset-integration methodology.

Chapter 3

Causal discovery from genetic datasets

In this chapter, a theory on causal discovery from genetic datasets that are random samples from the population of interest is presented first. Owing to certain assumptions about the causal relationships between the variables in such datasets, causal MAGs over the variables are of a certain form. This enables the clear interpretation of the edges in the MAG and the development of specialised causal-discovery algorithms. Learning from case–control samples is considered next. A local-learning algorithm is devised for learning only the genotype–phenotype links in the causal MAG that describes the causal relationships in such a sample. Specifically, the algorithm learns the genotypes that are causes, indicators of a hidden cause, and potential causes of the phenotype. An algorithm that learns the FDR-controlled genotype–phenotype links is then designed. A simulation study of the algorithm's performance is conducted using realistic simulated genetic case–control datasets. Finally, the algorithm is applied to datasets from prion disease.

3.1 Genetic random samples

A *genetic set of variables* is a set of genotypes and a phenotype. In addition to causal transitivity, the following four assumptions are made regarding the *causal* relationships between the variables in such a set:

Assumption 3.1. *The phenotype is not a cause of any genotype.*

Assumption 3.2. *No genotype is a cause of another.*

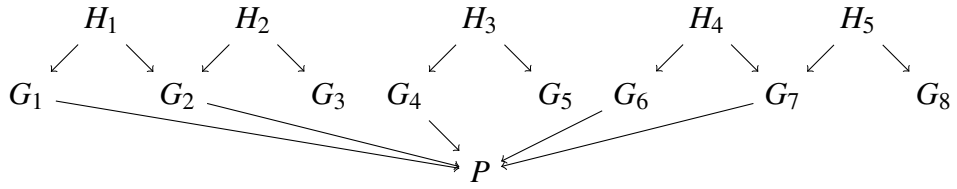
Assumption 3.3. *If two genotypes have a common cause, then they are on the same chromosome.*

Assumption 3.4. *If a variable is a common cause of a genotype and the phenotype, then every causal path from the variable to the phenotype is through some genotype.*

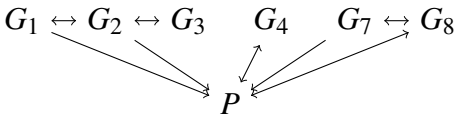
Assumption 3.1 follows from the *central dogma of molecular biology*, according to which information does not flow from the phenotype to the genotype. When selection bias is absent, Assumption 3.2 implies that two genotypes are in LD if and only if they have a hidden common cause. No assumption is made about the nature of the common cause; e.g., it could be the haplotype block where the variants reside. In light of Assumption 3.2, Assumption 3.3 follows from the principle of *independent assortment*, according to which genotypes on different chromosomes are independent. In light of Assumptions 3.1 and 3.2, Assumption 3.4 implies that, when selection bias is absent, a genotype is associated with the phenotype but is not a cause of the phenotype if and only if the genotype is in LD with another genotype which is a cause of the phenotype. Assumption 3.4 does not place any restriction on the causal relationships between the variables but is made in order to aid the causal interpretation of PAGs learned from a genetic dataset.

A causal DAG over the union of a genetic set of variables and another set of variables is called a *genetic causal DAG*. Assuming that causal transitivity holds, a genetic causal DAG that satisfies Assumptions 3.1–3.4 is said to be *plausible*. Obviously, the true genetic causal DAG over a set of variables is plausible if the assumptions hold. Clearly, under causal transitivity, a genetic causal DAG is plausible if and only if the following conditions are satisfied:

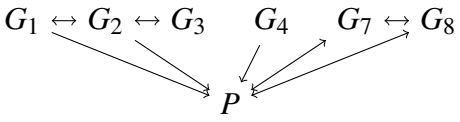
1. The phenotype is not an ancestor of any genotype.
2. No genotype is an ancestor of another.
3. Any two genotypes that have a common ancestor are on the same chromosome.



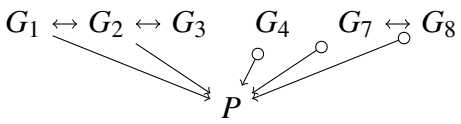
(a) A plausible genetic causal DAG. G_1, G_2, G_4, G_6, and G_7 are causes of P. G_1 and G_2, G_2 and G_3, G_4 and G_5, G_6 and G_7, and G_7 and G_8 are on the same chromosome and have a common cause H_1, H_2, H_3, H_4, and H_5 respectively. $\{G_1, G_2\}$, $\{G_2, G_3\}$, $\{G_1, G_2, G_3\}$, $\{G_4, G_5\}$, $\{G_6, G_7\}$, $\{G_7, G_8\}$, and $\{G_6, G_7, G_8\}$ are genetic chains relative to $\{G_1, G_2, G_3, G_4, G_7, G_8, P\}$.



(b) The plausible genetic causal MAG which is the marginal over $\{G_1, G_2, G_3, G_4, G_7, G_8, P\}$ of the DAG in Figure 3.1a. G_1, G_2, and G_7 are causes of P, G_4 is a proxy of a hidden cause of P, G_8 is an indicator of a hidden cause of P, and G_1 and G_2, G_2 and G_3, and G_7 and G_8 have a hidden common cause. $[G_3, G_2, G_1, P]$ is a genetic discriminating path for G_1.



(c) A plausible genetic causal MAG which is Markov equivalent to the one in Figure 3.1b. G_2, G_3, and G_4 are the unshielded genotypes and each of them is a parent of P in the other MAG if it is a parent of P in this one. G_1 is genetically discriminated in both MAGs and is a parent of P in both MAGs. G_4 is a spouse of P in the other MAG and a parent of P in this one. G_7 is a parent of P in the other MAG and a spouse of P in this one.



(d) The maximally-informative plausible genetic causal PAG representing the class of plausible genetic causal MAGs of which the MAGs in Figures 3.1b and 3.1c are members.

Figure 3.1: Example of a plausible genetic causal DAG, plausible genetic causal MAGs, and a plausible genetic causal PAG. G_i ($1 \leq i \leq 8$) is a genotype and P is a phenotype.

Figure 3.1a shows an example plausible genetic causal DAG.

The term *genetic dataset* is used here to refer to a dataset over a genetic set of variables. A genetic dataset that is a random sample from an unconditional distribution is called a *genetic random sample*. A MAG that is the marginal over a subset of the genotypes and the phenotype of a (plausible) genetic causal DAG is called a

(plausible) genetic causal MAG. The following definitions are central to the causal interpretation of plausible genetic causal MAGs. P is used to denote a phenotype; G is used to denote a genotype and $\mathbf{G}$ a set of genotypes.

Definition 3.1 (Genetic chain)**.** *A* genetic chain *from G_1 to G_2 relative to $\mathbf{G} \cup \{P\}$ is an ordered set $\{G^1, \ldots, G^n\}$ ($n \geq 2$) where $G^1 = G_1$, $G^n = G_2$, G^i $(2 \leq i \leq n-1)$ is a cause of P in $\mathbf{G}$ and G^i and G^{i+1} $(1 \leq i \leq n-1)$ have a common cause not in $\mathbf{G}$. $G^2, \ldots, G^{n-1}$ are called* interior *genotypes in the genetic chain. G_1 and G_2 may or may not be in $\mathbf{G}$.*

G_1 and G_2 are said to be *genetically chained* relative to $\mathbf{G} \cup \{P\}$ if there is a genetic chain between G_1 and G_2 relative to $\mathbf{G} \cup \{P\}$. In Figure 3.1a, $\{G_1, G_2\}$, $\{G_2, G_3\}$, $\{G_1, G_2, G_3\}$, $\{G_4, G_5\}$, $\{G_6, G_7\}$, $\{G_7, G_8\}$, and $\{G_6, G_7, G_8\}$ are genetic chains relative to $\{G_1, G_2, G_3, G_4, G_7, G_8, P\}$.

Definition 3.2 (Indicator of a hidden cause of a phenotype)**.** *G_1 is an* indicator of a hidden cause *of P relative to $\mathbf{G} \cup \{P\}$ if G_1 is in $\mathbf{G}$, not a cause of P, and genetically chained relative to $\mathbf{G} \cup \{P\}$ to some cause G_2 of P not in $\mathbf{G}$. It is then said that* the presence of G_2 is indicated by G_1.

Definition 3.3 (Proxy of a hidden cause of a phenotype)**.** *G_1 is a* proxy of a hidden cause *of P relative to $\mathbf{G} \cup \{P\}$ if G_1 is in $\mathbf{G}$, not a cause of P, and has a common cause not in $\mathbf{G}$ with some cause G_2 of P not in $\mathbf{G}$.*

Clearly, if G_1 is a proxy of a hidden cause of P, then G_1 is an indicator of a hidden cause of P. The following theorem gives necessary and sufficient conditions for the existence of each type of edge in a plausible genetic causal MAG. In the forward direction, it states the causal interpretation of the edges; in the reverse direction, it provides a way of constructing a plausible genetic causal MAG given the causal relationships among a genetic set of variables. The proof of the theorem, as well as the proofs of all other propositions in this work, can be found in Appendix A.

Theorem 3.1. *In a plausible genetic causal MAG over a set of variables $\mathbf{G} \cup \{P\}$ there are three types of edges: $G \rightarrow P$, $G \leftrightarrow P$, and $G_1 \leftrightarrow G_2$, and*

- *$G \to P$ exists if and only if G is a cause of P,*
- *$G \leftrightarrow P$ exists if and only if G is an indicator of a hidden cause of P,*
- *$G_1 \leftrightarrow G_2$ exists if and only if G_1 and G_2 have a hidden common cause.*

The following corollary of Theorem 3.1 gives sufficient conditions for an observed genotype to be a proxy of a hidden cause the phenotype. Figure 3.1b shows an example of a plausible genetic causal MAG, which is the marginal of the DAG in Figure 3.1a.

Corollary 3.1. *In a plausible genetic causal MAG over a set of variables $\mathbf{G} \cup \{P\}$, if G_1 is a spouse of P and there is no G_2 such that G_2 is a parent of P and G_2 and G_1 are adjacent, then G_1 is a proxy of a hidden cause of P.*

A *genetic mixed graph* is a mixed graph over a genetic set of variables. A *potential plausible genetic causal MAG* is a genetic mixed graph which is to be interpreted causally if it is a MAG, in which case it would be a genetic causal MAG by definition. The following theorem gives necessary and sufficient conditions for a potential plausible genetic causal MAG to be a plausible genetic causal MAG.

Theorem 3.2. *A potential plausible genetic causal MAG over a set of variables $\mathbf{G} \cup \{P\}$ is a plausible genetic causal MAG if and only if the following conditions are satisfied:*

1. *Edges incident to P are into P.*
2. *Genotype–genotype edges are bidirected.*
3. *Adjacent genotypes are on the same chromosome.*

Markov equivalence of plausible genetic causal MAGs is based on the notion of *genetic discriminating path*.

Definition 3.4. *In a plausible genetic causal MAG over $\mathbf{G} \cup \{P\}$, a path $p = [G_n, \ldots, G_2, G_1, P]$ is called a* genetic discriminating path *for genotype G_1 if the following two conditions are satisfied:*

- *G_n is not adjacent to P.*
- *G_i $(2 \leq i \leq n-1)$ is adjacent to P and the edge between G_i and P is out of G_i.*

A genotype for which there is genetic discriminating path is said to be *genetically-discriminated*. In Figure 3.1b, $[G_3, G_2, G_1, P]$ is a genetic discriminating path for G_1. The following theorem characterises Markov-equivalent plausible genetic causal MAGs.

Theorem 3.3. *Suppose that $\mathbb{M}_1$ and $\mathbb{M}_2$ are plausible genetic causal MAGs over the same set of variables. $\mathbb{M}_1$ and $\mathbb{M}_2$ are Markov equivalent if and only if the following conditions are satisfied:*

1. *$\mathbb{M}_1$ and $\mathbb{M}_2$ have the same skeleton.*
2. *Each unshielded genotype is a parent of the phenotype in $\mathbb{M}_1$ if and only if it is a parent of the phenotype in $\mathbb{M}_2$.*
3. *Each genetically-discriminated genotype is a parent of the phenotype in $\mathbb{M}_1$ if and only if it is a parent of the phenotype in $\mathbb{M}_2$.*

Figure 3.1c shows a plausible genetic causal MAG which is Markov equivalent to the one in Figure 3.1b.

A *(plausible) genetic causal PAG* is defined as a PAG for a Markov equivalence class of (plausible) genetic causal MAGs. A partially-oriented mixed graph over a genetic set of variables is called a *partially-oriented genetic mixed graph*. A *potential plausible genetic causal PAG* is a partially-oriented genetic mixed graph which is a genetic causal PAG if it is a PAG. The following theorem gives necessary and sufficient conditions for a potential plausible genetic causal PAG to be a maximally-informative plausible genetic causal PAG.

Theorem 3.4. *A potential plausible genetic causal PAG is a maximally-informative plausible genetic causal PAG if and only if the following conditions are satisfied:*

1. *Edges incident to P are into P.*

2. *Genotype–genotype edges are bidirected.*

3. *Adjacent genotypes are on the same chromosome.*

4. *Endpoints at G are oriented if and only if G is unshielded or genetically-discriminated.*

Figure 3.1d shows the maximally-informative plausible genetic causal PAG representing the class of plausible genetic causal MAGs of which the MAGs in Figures 3.1b and 3.1c are members.

3.1.1 Skeleton identification

In order to learn the skeleton of the true genetic causal MAG, PC–skeleton (Algorithm 2) or local-to-global learning (Algorithm 6) can be used, followed by a search for sepsets in the Possible D-SEP sets (see Section 2.4). Under Assumptions 3.1–3.4, however, the true genetic causal MAG is a *plausible* genetic causal MAG whose special structure (Theorem 3.2) can be exploited by a specialised algorithm such as Algorithm 8. The correctness of Algorithm 8 follows by the next theorem, which characterises m-separation in a plausible genetic causal MAG:

Theorem 3.5. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*:*

- G_1 *and* P *are adjacent if and only if they are not m-separated by any subset of the genotypes adjacent to* P *on the same chromosome.*

- G_1 *and* G_2 *are adjacent if and only if they are not m-separated.*

Owing to Theorem 3.5, it is sufficient to apply non-symmetric GLL-PC (Algorithm 3) targeting the phenotype independently for each chromosome in order to learn the genotype–phenotype links (lines 1–9 of Algorithm 8) and to determine the pairwise m-separation of all genotypes on the same chromosome in order to learn the genotype–genotype links (lines 10–18).

3.1.2 Edge orientation

After the skeleton of a plausible genetic causal PAG is identified, the 11 orientation rules of the FCI algorithm [Zhang, 2008] can be applied in order to obtain the

Algorithm 8 Skeleton identification for plausible genetic causal PAGs. $\mathbf{G}$ is a set of genotypes and P is a phenotype. $\mathbb{M}$ is a plausible genetic causal MAG over $\mathbf{G} \cup \{P\}$. $\mathbf{I}(\mathbb{M})$ is the set of m-separations in $\mathbb{M}$. $\mathbb{S}$ is an undirected graph. **Sepset** is a map from pairs of nodes to sets of nodes. In the output, $\mathbb{S}$ is the skeleton of $\mathbb{M}$ and **Sepset** is a map from pairs of nodes in $\mathbf{G} \cup \{P\}$ to sets of nodes in $\mathbf{G} \cup \{P\}$ that m-separate them in $\mathbb{M}$. n is the number of chromosomes. $\mathbf{G}_k$ is the subset of $\mathbf{G}$ on the k-th chromosome. $\mathbf{I}(\mathbb{M})[_{\mathbf{G}_k \cup \{P\}}$ is the subset of $\mathbf{I}(\mathbb{M})$ over $\mathbf{G}_k \cup \{P\}$. $X \perp Y \mid \mathbf{Z}$ denotes that nodes X and Y are m-separated given set of nodes $\mathbf{Z}$.

Input: $\mathbf{I}(\mathbb{M})$
Output: $\mathbb{S}$ and **Sepset**

for each $1 \leq k \leq n$ **do**
  ▷ Learn genotype–phenotype links
  let $\mathbf{TA}_k(P)$ and $\mathbf{Sepset}_k^P$ be the output of Algorithm 3 with $\mathbf{I}(\mathbb{M})[_{\mathbf{G}_k \cup \{P\}}$ and P as input.
  for each $G \in \mathbf{TA}_k(P)$ **do**
    add edge $G - P$ to $\mathbb{S}$
  end for
  for each $G \in \mathbf{G}_k \setminus \mathbf{TA}_k(P)$ **do**
    $\mathbf{Sepset}(\{G, P\}) \leftarrow \mathbf{Sepset}_k^P(\{G, P\})$
  end for
  ▷ Learn genotype–genotype links
  for each $\{G_1, G_2\} \subseteq \mathbf{G}_k$ **do**
    if $G_1 \perp G_2 \notin \mathbf{I}(\mathbb{M})$ **then**
      add edge $G_1 - G_2$ to $\mathbb{S}$
    else
      $\mathbf{Sepset}(\{G_1, G_2\}) \leftarrow \emptyset$
    end if
  end for
end for

maximally-informative genetic causal PAG. However, it can be shown that not all 11 rules are applicable. Furthermore, it is the maximally-informative *plausible* genetic causal PAG which is of interest here. Algorithm 9, which uses 4 orientation rules, returns the maximally-informative plausible genetic causal PAG (Theorem 3.6).

Theorem 3.6 (Correctness of Algorithm 9)**.** *If the input of Algorithm 9 is* $\mathbb{P}$ *and* **Sepset***, then in the output of Algorithm 9,* $\mathbb{P}$ *is the maximally-informative plausible genetic causal PAG for the same Markov equivalence class as in the input.*

Algorithm 9 Edge orientation for plausible genetic causal PAGs. $\mathbf{G}$ is a set of genotypes and P is a phenotype. In the input, $\mathbb{P}$ is an unoriented plausible genetic causal PAG over $\mathbf{G} \cup \{P\}$ and **Sepset** is a map from pairs of nodes in $\mathbf{G} \cup \{P\}$ to sets of nodes in $\mathbf{G} \cup \{P\}$ that m-separate them in a MAG in the Markov equivalence class of $\mathbb{P}$. $\text{Adj}_{\mathbb{P}}(X,Y)$ denotes that nodes X and Y are adjacent in $\mathbb{P}$. In the output, $\mathbb{P}$ is maximally-informative plausible genetic causal PAG over $\mathbf{G} \cup \{P\}$.

Input: $\mathbb{P}$ and **Sepset**
Output: $\mathbb{P}$

for each edge $G_1 \ast\!\!-\!\!\ast P$ in $\mathbb{P}$ **do**
    ▷ Rule (1)
    orient $G_1 \ast\!\!-\!\!\ast P$ as $G \ast\!\!\rightarrow P$
    ▷ Rule (2)
    if $\exists G_2 \in \mathbf{G} \setminus \{G_1\}$ s.t. $\text{Adj}_{\mathbb{P}}(G_2, G_1)$ and $\neg\text{Adj}_{\mathbb{P}}(G_2, P)$ **then**
        if $G_1 \in$ **Sepset**$(\{G_2, P\})$ **then**
            orient $G_1 \ast\!\!\rightarrow P$ as $G_1 \rightarrow P$
        else
            orient $G_1 \ast\!\!\rightarrow P$ as $G_1 \leftrightarrow P$
        end if
    end if
end for
▷ Rule (3)
for each edge $G_1 \ast\!\!-\!\!\ast G_2$ in $\mathbb{P}$ **do**
    orient $G_1 \ast\!\!-\!\!\ast G_2$ as $G_1 \leftrightarrow G_2$
end for
while more edges can be oriented in $\mathbb{P}$ **do**
    ▷ Rule (4)
    for each edge $G_1 \circ\!\!-\!\!\ast P$ in $\mathbb{P}$ **do**
        if $\exists$ genetic discriminating path $[G_n, \ldots, G_2, G_1, P]$ for G_1 in $\mathbb{P}$ **then**
            if $G_1 \in$ **Sepset**$(\{G_n, P\})$ **then**
                orient $G_1 \circ\!\!-\!\!\ast P$ as $G_1 -\!\!\ast P$
            else
                orient $G_1 \circ\!\!-\!\!\ast P$ as $G_1 \leftarrow\!\!\ast P$
            end if
        end if
    end for
end while

3.1.3 Local learning

It is usually not of interest to learn the whole plausible genetic causal PAG. If only the nodes adjacent to P are of interest, it is sufficient to apply non-symmetric GLL-PC (Algorithm 3) targeting P. If the orientations of the edges incident to P are also of interest, it is clear that it is only needed to learn the PAG over P, nodes adjacent to P, and nodes adjacent to a node adjacent to P (that is, nodes at distance ≤ 2 from P). The topic of local learning is not pursued any further in the unconditional case since the main focus of this work is learning from the case–control datasets from prion disease.

3.2 Conditional genetic random samples

It is impractical to obtain a random sample from the population of interest for the purpose of identifying disease-susceptibility variants, especially for rare diseases. For example, a random sample for sCJD, which has a prevalence of about one in a million, would have to contain about one million controls for each sCJD case. Therefore, most genetic datasets are case–control samples, containing about the same number of cases and controls. In this section, in addition to causal transitivity and Assumptions 3.1–3.4, the following assumption is made:

Assumption 3.5. *A variable is a cause of some selection variable if and only if the variable is a cause of the phenotype.*

In a genetic case–control dataset, the observations are included based solely on the phenotype. Therefore, it can be assumed that there is a single logical selection variable which is an effect of the phenotype and that every causal path from a variable to the selection variable is through the phenotype. The more general Assumption 3.5 is used for mathematical convenience.

A causal DAG over the union of a genetic set of variables, a set of selection variables, and a set of other variables is called a *genetic causal DAG with selection nodes*. Assuming that causal transitivity holds, a genetic causal DAG with selection nodes that satisfies Assumptions 3.1–3.5 is said to be *plausible*. Obviously, the true genetic causal DAG with selection nodes over a set of variables is plausible if the

assumptions hold. It is easy to see that a genetic causal DAG with selection nodes is plausible if and only if the following conditions are satisfied:

1. The phenotype is not an ancestor of any genotype.
2. No genotype is an ancestor of another.
3. Any two genotypes that have a common ancestor are on the same chromosome.
4. A node is an ancestor of the selection variable if and only if the node is an ancestor of the phenotype.

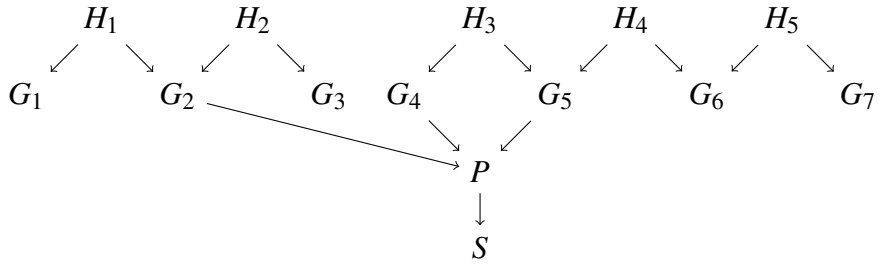
Figure 3.2a shows an example of a plausible genetic causal DAG with selection nodes.

A genetic dataset that is a random sample from the conditional distribution given an instantiation of the selection variables is referred to as a *conditional genetic random sample*. A MAG that is the marginal/conditional over a subset of the genotypes and the phenotype of a (plausible) genetic causal DAG with selection nodes given the selection variables is called a *(plausible) conditional genetic causal MAG*.

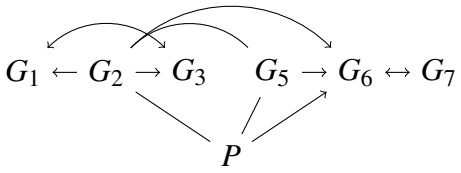
The following theorem gives necessary and sufficient conditions for the existence of each type of edge in a plausible conditional genetic causal MAG. In the forward direction, it states the causal interpretation of the edges; in the reverse direction, it provides a way of constructing a plausible conditional genetic causal MAG given the causal relationships among a genetic set of variables.

Theorem 3.7. *In a plausible conditional genetic causal MAG, there are five types of edges:* $G - P$, $G \leftarrow P$, $G_1 \rightarrow G_2$, $G_1 \leftrightarrow G_2$, *and* $G_1 - G_2$, *and*

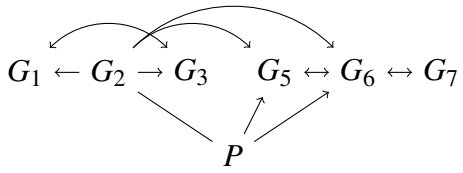
- $G - P$ *exists if and only if* G *is a cause of* P,
- $G \leftarrow P$ *exists if and only if* G *is an indicator of a hidden cause of* P,
- $G_1 - G_2$ *exists if and only if* G_1 *is a cause of* P *and* G_2 *is a cause of* P,



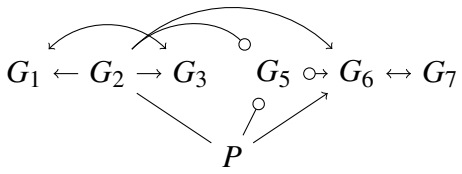
(a) A plausible genetic causal DAG with selection nodes. G_2, G_4, and G_5 are causes of P. G_1 and G_2, G_2 and G_3, G_4 and G_5, G_5 and G_6, and G_6 and G_7 are on the same chromosome and have a common cause H_1, H_2, H_3, H_4, and H_5 respectively. S is a selection variable, and P is a cause of S. $\{G_1, G_2\}$, $\{G_2, G_3\}$, $\{G_1, G_2, G_3\}$, $\{G_5, G_6\}$, $\{G_4, G_5, G_6\}$, and $\{G_6, G_7\}$ are genetic chains relative to $\{G_1, G_2, G_3, G_5, G_6, G_7, P\}$.



(b) The plausible conditional genetic causal MAG which is the marginal/conditional over $\{G_1, G_2, G_3, G_5, G_6, G_7, P\}$ given $\{S\}$ of the DAG in Figure 3.2a. G_2 and G_5 are causes of P, G_6 is an indicator of a hidden cause of P, and G_1 and G_2, G_2 and G_3, G_1 and G_3, and G_6 and G_7 are genetically chained.



(c) A plausible conditional genetic causal MAG which is Markov equivalent to the one in Figure 3.2b. G_2 and G_6 are the unshielded genotypes that are adjacent to P and the edges incident to each of them are into the genotype in the other MAG if they are into the genotype in this one. The edges incident to G_5 are into G_5 in the other MAG and out of G_5 in this one.



(d) The maximally-informative plausible conditional genetic causal PAG representing the class of plausible conditional genetic causal MAGs of which the MAGs in Figures 3.2b and 3.2c are members.

Figure 3.2: Example of a plausible genetic causal DAG with selection nodes, plausible conditional genetic causal MAGs, and a maximally-informative plausible conditional genetic causal PAG. G_i ($1 \leq i \leq 7$) is a genotype and P is a phenotype.

- $G_1 \to G_2$ *exists if and only if* G_1 *is a cause of* P*,* G_2 *is not a cause of* P*, and either* G_2 *is an indicator of a hidden cause of* P *or* G_1 *and* G_2 *are genetically chained, and*

- *$G_1 \leftrightarrow G_2$ exists if and only if G_1 is not a cause of P, G_2 is not a cause of P, and either G_1 and G_2 are genetically chained, or G_1 is an indicator of a hidden cause of P and G_2 is an indicator of a hidden cause of P.*

The following corollary of Theorem 3.7 gives sufficient conditions for an observed genotype to be a proxy of a hidden cause of the phenotype. Unfortunately, the conditions are much stronger than in the unconditional case. Figure 3.2b shows an example of a plausible conditional genetic causal MAG, which is the marginal/conditional of the DAG in Figure 3.2a.

Corollary 3.2. *In a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$, if G_1 is a child of P and P has no neighbours, then G_1 is a proxy of a hidden cause of P.*

A *potential plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given* $\mathbf{S}$ is a genetic mixed graph over $\mathbf{G} \cup \{P\}$ which is a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$ if it is a MAG. The following theorem gives necessary and sufficient conditions for a potential plausible conditional genetic causal MAG to be a plausible conditional genetic causal MAG.

Theorem 3.8. *A potential plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$ is a plausible conditional genetic causal MAG if and only if the following conditions are satisfied:*

1. *Edges incident to P are out of P.*
2. *Edges incident to G adjacent to P are either all out of G or all into G.*
3. *Edges incident to G not adjacent to P are into G.*
4. *Genotypes adjacent to P are adjacent.*
5. *Each triple $[G_1, G_3, G_2]$ such that G_1 and G_2 are not adjacent to P, G_3 is adjacent to P, and the edge between G_3 and P is out of G_3 is shielded.*
6. *Every pair of adjacent genotypes that are not both adjacent to P are on the same chromosome.*

The following theorem characterises Markov-equivalent plausible conditional genetic causal MAGs.

Theorem 3.9. *Two plausible conditional genetic causal MAGs* $\mathbb{M}_1$ *and* $\mathbb{M}_2$ *over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$ *are Markov equivalent if and only if the following two conditions are satisfied:*

1. $\mathbb{M}_1$ *and* $\mathbb{M}_2$ *have the same skeleton.*
2. *Edges incident to each unshielded G adjacent to P are into G in* $\mathbb{M}_1$ *if and only if they are into G in* $\mathbb{M}_2$.

Figure 3.2c shows a plausible conditional genetic causal MAG which is Markov equivalent to the one in Figure 3.2b.

A *(plausible) conditional genetic causal PAG* is a PAG for a Markov equivalence class of (plausible) conditional genetic causal MAGs. A *potential plausible conditional genetic causal PAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$ is a partially-oriented genetic mixed graph over $\mathbf{G} \cup \{P\}$ which is a plausible conditional genetic causal PAG over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$ if it is a PAG. The following theorem gives necessary and sufficient conditions for a potential plausible conditional genetic causal PAG to be a maximally-informative plausible conditional genetic causal PAG.

Theorem 3.10. *A potential plausible conditional genetic causal PAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$ *is a maximally-informative plausible conditional genetic causal PAG if and only if the following conditions are satisfied:*

1. *Edges incident to P are out of P.*
2. *Edges incident to G adjacent to P are either all out of G or all into G.*
3. *Edges incident to G not adjacent to P are into G.*
4. *Genotypes adjacent to P are adjacent.*
5. *Each triple* $[G_1, G_3, G_2]$ *such that* G_1 *and* G_2 *are not adjacent to P,* G_3 *is adjacent to P, and the edge between* G_3 *and P is out of* G_3 *is shielded.*

6. *Every pair of adjacent genotypes that are not both adjacent to P are on the same chromosome.*

7. *Endpoints at G are oriented if and only if G is unshielded.*

Figure 3.2d shows the maximally-informative plausible conditional genetic causal PAG representing the class of plausible conditional genetic causal MAGs of which the MAGs in Figures 3.2b and 3.2c are members.

3.2.1 Skeleton identification

As in the unconditional case, the special structure of plausible conditional genetic MAGs (Theorem 3.8) is exploited by a specialised skeleton-identification algorithm (Algorithm 10). The correctness of the algorithm follows by the next theorem, which concerns m-separation and adjacencies in a plausible conditional genetic MAGs:

Theorem 3.11. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*:*

- *G and P are adjacent if and only if they are not m-separated by any subset of the genotypes adjacent to P on the same chromosome.*

- *If* G_1 *and* G_2 *are adjacent to P, then* G_1 *and* G_2 *are adjacent.*

- *If* G_1 *and P are strictly m-separated by* $\mathbf{Z}$*, then* G_1 *and* G_2 *are adjacent if and only if* $G_2 \in \mathbf{Z}$ *or* G_1 *and* G_2 *are not m-separated by* $\mathbf{Z}$*.*

- *If* G_1 *and P are strictly m-separated by* $\mathbf{Z}_1$ *and* G_2 *and P are strictly m-separated by* $\mathbf{Z}_2$*, then* G_1 *and* G_2 *are adjacent if and only if* G_1 *and* G_2 *are not m-separated by the smallest among* $\mathbf{Z}_1$ *and* $\mathbf{Z}_2$*.*

- *If* G_1 *and* G_2 *are adjacent and* G_1 *and P are not adjacent, then* G_1 *and* G_2 *are on the same chromosome.*

- *If* G_1 *is not adjacent to P or to a node adjacent to P, then* G_1 *and* G_2 *are adjacent if and only if they are not m-separated.*

Owing to Theorem 3.11, it is sufficient to apply non-symmetric GLL-PC (Algorithm 3) targeting the phenotype independently for each chromosome in order to learn the genotype–phenotype links (lines 1–10 of Algorithm 10), as in the unconditional case. Genotype nodes adjacent to (at distance 1 from) the phenotype are then adjacent (lines 12–15). All other genotype–genotype links are between genotypes on the same chromosome. In order to determine whether a genotype at distance 2 from the phenotype (that is, a genotype that is not adjacent to the phenotype but is adjacent to a node adjacent to the phenotype) is adjacent to a genotype adjacent to the phenotype, it is sufficient to check whether the latter genotype is in the sepset of the former genotype and the phenotype, and if this is not the case, determine whether the genotypes are m-separated given the sepset (lines 17–26). In order to determine whether two genotypes at distance 2 from the phenotype are adjacent, it suffices to determine whether they are m-separated given the smallest among the sepsets of each of the two genotypes and the phenotype (lines 27–39). Finally, in order to determine whether a genotype at distance ≥ 3 from the phenotype (that is, a genotype not adjacent to the phenotype or to some genotype adjacent to the phenotype) is adjacent to another genotype, it is sufficient to determine whether the genotypes are m-separated (part 3), as in the unconditional case.

3.2.2 Edge orientation

As in the unconditional case, it can be shown that not all 11 orientation rules of the FCI algorithm are applicable to a plausible conditional genetic causal PAG. Furthermore, it is the maximally-informative *plausible* conditional genetic causal PAG which is of interest. Algorithm 11, which uses three orientation rules, returns the maximally-informative plausible genetic causal PAG (Theorem 3.12).

Theorem 3.12 (Correctness of Algorithm 11)**.** *If the input of Algorithm 9 is* $\mathbb{P}$ *and* **Sepset***, then in the output of Algorithm 11,* $\mathbb{P}$ *is the maximally-informative plausible conditional genetic causal PAG for the same Markov equivalence class as in the input.*

Algorithm 10 Skeleton identification for plausible conditional genetic causal PAGs — part 1 out of 3. $\mathbf{G}$ is a set of genotypes and P is a phenotype. $\mathbb{M}$ is a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$. $\mathbf{I}(\mathbb{M})$ is the set of m-separations in $\mathbb{M}$. $\mathbb{S}$ is an undirected graph. **Sepset** is a map from pairs of nodes to sets of nodes. In the output, $\mathbb{S}$ is the skeleton of $\mathbb{M}$ and **Sepset** is a map from pairs of nodes in $\mathbf{G} \cup \{P\}$ to sets of nodes in $\mathbf{G} \cup \{P\}$ that m-separate them in $\mathbb{M}$. n is the number of chromosomes. $\mathbf{G}_k$ is the subset of $\mathbf{G}$ on the k-th chromosome. $\mathbf{I}(\mathbb{M})[_{\mathbf{G}_k \cup \{P\}}$ is the subset of $\mathbf{I}(\mathbb{M})$ over $\mathbf{G}_k \cup \{P\}$. $\mathrm{Adj}_{\mathbb{S}}(X,Y)$ denotes that nodes X and Y are adjacent in $\mathbb{S}$.

Input: $\mathbf{I}(\mathbb{M})$
Output: $\mathbb{S}$ and **Sepset**

for each $1 \leq k \leq n$ **do**
  ▷ Learn genotypes adjacent to (at distance 1 from) the phenotype
  let $\mathbf{TA}_k(P)$ and $\mathbf{Sepset}_k^P$ be the output of Algorithm 3 with $\mathbf{I}(\mathbb{M})[_{\mathbf{G}_k \cup \{P\}}$ and P as input.
  for each $G \in \mathbf{TA}_k(P)$ **do**
    add edge $G - P$ to $\mathbb{S}$
  end for
  for each $G \in \mathbf{G}_k \setminus \mathbf{TA}_k(P)$ **do**
    $\mathbf{Sepset}(\{G,P\}) \leftarrow \mathbf{Sepset}_k^P(\{G,P\})$
  end for
end for
▷ Learn links between the genotypes adjacent to (at distance 1 from)
▷ the phenotype
for each $\{G_1, G_2\}$ s.t. $\mathrm{Adj}_{\mathbb{S}}(G_1,P)$ and $\mathrm{Adj}_{\mathbb{S}}(G_2,P)$ **do**
  add edge $G_1 - G_2$ to $\mathbb{S}$
end for

3.2.3 Local learning

As in the unconditional case, it is sufficient to apply non-symmetric GLL-PC (Algorithm 3) targeting P if only the nodes adjacent to P in the plausible conditional genetic causal MAG are of interest and learn the PAG over the nodes at distance ≤ 2 from P if the orientations of the edges incident to P are also of interest. In the conditional case, learning these orientations is even simpler. For each G_1 adjacent to P, it is sufficient to find a G_2 not adjacent to P such that G_2 is in the sepset of G_1 and P in order to orient the edge out of G_1, and if such G_2 does not exist, it is sufficient to find a G_2 not adjacent to P and adjacent to G_1 in order to orient the edge into G_1. Algorithm 12, whose correctness follows from Theorems 3.11 and 3.12, learns the neighbours, spouses, and *potential neighbours* of the phenotype in a

Algorithm 10 Skeleton identification for conditional genetic causal PAGs — part 2 out of 3. $X \perp Y \mid \mathbf{Z}$ denotes that nodes X and Y are m-separated given set of nodes $\mathbf{Z}$. $\mathbf{AD}^d_{\mathbb{S}}(X)$ is the set of nodes at distance d from node X in $\mathbb{S}$.

16: **for each** $1 \leq k \leq n$ **do**
17: ▷ Learn genotypes at distance 2 from the phenotype
18: **for each** $G_1 \in \mathbf{G}_k \cap \mathbf{AD}_{\mathbb{S}}(P)$ **do**
19: **for each** $G_2 \in \mathbf{G}_k \setminus \mathbf{AD}_{\mathbb{S}}(P)$ **do**
20: **if** $G_1 \in \mathbf{Sepset}(\{G_2, P\})$ or $G_1 \perp G_2 \mid \mathbf{Sepset}(\{G_2, P\}) \notin \mathbf{I}(\mathbb{M})$ **then**
21: add edge $G_1 - G_2$ to $\mathbb{S}$
22: **else**
23: $\mathbf{Sepset}(\{G_1, G_2\}) \leftarrow \mathbf{Sepset}(\{G_2, P\})$
24: **end if**
25: **end for**
26: **end for**
27: ▷ Learn links between the genotypes at distance 2 from the phenotype
28: **for each** $\{G_1, G_2\} \subseteq \mathbf{G}_k \cap \mathbf{AD}^2_{\mathbb{S}}(P)$ **do**
29: **if** $|\mathbf{Sepset}(\{G_1, P\})| \leq |\mathbf{Sepset}(\{G_2, P\})|$ **then**
30: $\mathbf{S} \leftarrow \mathbf{Sepset}(\{G_1, P\})$
31: **else**
32: $\mathbf{S} \leftarrow \mathbf{Sepset}(\{G_2, P\})$
33: **end if**
34: **if** $G_1 \perp G_2 \mid \mathbf{S} \notin \mathbf{I}(\mathbb{M})$ **then**
35: add edge $G_1 - G_2$ to $\mathbb{S}$
36: **else**
37: $\mathbf{Sepset}(\{G_1, G_2\}) \leftarrow \mathbf{S}$
38: **end if**
39: **end for**

plausible conditional genetic causal PAG. Node X is a *potential neighbour* of node Y in a mixed graph if edge $X \circ\!\!- Y$ exists.

Algorithm 12, like Algorithm 8, accepts a set of m-separations that is assumed to be the set of m-separations in a plausible conditional genetic causal MAG. In practice, however, m-separations are determined by performing hypothesis tests of conditional independence on a random sample from the distribution of the variables, as explained in Chapter 2. Algorithm 13 directly accepts a conditional genetic random sample and controls the FDR of the nodes adjacent to the phenotype. Since it not clear what error rate to control for the orientations and how to control it, differentiating the nodes adjacent to the phenotype is not attempted by Algorithm 13.

Algorithm 10 Skeleton identification for conditional genetic causal PAGs — part 3 out of 3. n is the number of chromosomes.

▷ Learn genotypes at distance 3 from the phenotype
for each $G_1 \in \mathbf{G}_k \cap \mathbf{AD}^2_{\mathbb{S}}(P)$ **do**
  for each $G_2 \in \mathbf{G}_k \setminus \mathbf{AD}^{\leq 2}_{\mathbb{S}}(P)$ **do**
    if $G_1 \perp G_2 \notin \mathbf{I}(\mathbb{M})$ **then**
      add edge $G_1 - G_2$ to $\mathbb{S}$
    else
      Sepset$(\{G_1, G_2\}) \leftarrow \emptyset$
    end if
  end for
end for
▷ Learn links between genotypes at distance ≥ 3 from the phenotype
for each $\{G_1, G_2\} \subseteq \mathbf{G}_k \setminus \mathbf{AD}^{\leq 2}_{\mathbb{S}}(P)$ **do**
  if $G_1 \perp G_2 \notin \mathbf{I}(\mathbb{M})$ **then**
    add edge $G_1 - G_2$ to $\mathbb{S}$
  else
    Sepset$(\{G_1, G_2\}) \leftarrow \emptyset$
  end if
end for
end for

Theorem 3.13 (Correctness of Algorithm 13)**.** *Let $\mathbb{G}$ be a genetic causal DAG with selection nodes over $\mathbf{V} = \mathbf{G} \cup \{P\} \dot{\cup} \mathbf{H} \dot{\cup} \mathbf{S}$, $\mathbb{M}$ be the marginal/conditional of $\mathbb{G}$ given $\mathbf{S}$, $\mathscr{P}$ be the probability distribution of the variables in $\mathbf{V}$, and $\mathscr{M}$ be the marginal/conditional of $\mathscr{P}$ over $\mathbf{G} \cup \{P\}$ given $\mathbf{S} = \mathbf{s}$. Suppose that Algorithm 13 is applied to a sample D from $\mathscr{M}$ with FDR threshold q. The FDR among the nodes in $\widehat{\mathbf{AD}}(P)$ is not greater than q if the following conditions are satisfied:*

1. *$\mathbb{G}$ and $\mathscr{M}$ satisfy the selection bias causal assumption.*
2. *All tests considered by the algorithm are reliable.*
3. *Performed tests never produce a type II error.*

3.2.4 Estimating genetic causal effects

It would be appropriate to accompany the (upper bound of the) link-absence p-value corresponding to each genotype discovered by Algorithm 13 with some *causal* effect. The following theorem is based on Theorem 2.6 and gives sufficient *causal*

Algorithm 11 Edge orientation for conditional genetic causal PAGs. $\mathbf{G}$ is a set of genotypes and P is a phenotype. $\mathbb{P}$ is an unoriented plausible genetic causal PAG over $\mathbf{G} \cup \{P\}$ in the input and a partially-oriented mixed graph over $\mathbf{G} \cup \{P\}$ in the output. **Sepset** is a map from pairs of variables in $\mathbf{G} \cup \{P\}$ to subsets of variables in $\mathbf{G} \cup \{P\}$. $\text{Adj}_{\mathbb{P}}(X, Y)$ denotes that nodes X and Y are adjacent in $\mathbb{P}$

Input: $\mathbb{P}$ and **Sepset**
Output: $\mathbb{P}$

for each $G_1 \in \mathbf{G}$ **do**
  if $\text{Adj}_{\mathbb{P}}(G_1, P)$ **then**
    ▷ Rule (1)
    orient $G_1 \ast\!\!-\!\!\ast P$ as $G_1 \ast\!\!- P$
    ▷ Rule (2)
    if $\exists G_2 \in \mathbf{G} \setminus \{G_1\}$ s.t. $\text{Adj}_{\mathbb{P}}(G_2, G_1)$ and $\neg\text{Adj}_{\mathbb{P}}(G_2, P)$ **then**
      if $G_1 \in$ **Sepset**$(\{G_2, P\})$ **then**
        for each edge $G_1 \ast\!\!-\!\!\ast X$ **do**
          orient $G_1 \ast\!\!-\!\!\ast X$ as $G_1 -\!\!\ast X$
        end for
      else
        for each edge $G_1 \ast\!\!-\!\!\ast X$ **do**
          orient $G_1 \ast\!\!-\!\!\ast X$ as $G_1 \leftarrow\!\!\ast X$
        end for
      end if
    end if
  else
    ▷ Rule (3)
    for each edge $G_1 \ast\!\!-\!\!\ast G_2$ **do**
      orient $G_1 \ast\!\!-\!\!\ast G_2$ as $G_1 \leftarrow\!\!\ast G_2$
    end for
  end if
end for

conditions for the genotypic CORs in the general population to equal the respective genotypic ORs in the sampled population when there is a single logical selection variable.

Theorem 3.14. *Let S be a logical selection variable, $\mathbb{M}$ be a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given $\{S\}$, $\mathscr{P}$ be the distribution of $\mathbf{G} \cup \{P\}$, $\mathscr{Q}$, $\mathscr{Q}$ be the conditional of $\mathscr{P}$ given $S = true$, and $G_1 \in \mathbf{G}$. The genotypic CORs for G_1 in $\mathscr{P}$ equal the respective genotypic ORs in $\mathscr{Q}$ if the following conditions are true:*

1. *G_1 does not have a common cause with some cause G_2 of P.*

Algorithm 12 Learn the neighbours, spouses, and potential neighbours of the phenotype in a plausible conditional genetic causal PAG. $\mathbf{G}$ is a set of genotypes and P is a phenotype. $\mathbb{M}$ is a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$. $\mathbf{I}(\mathbb{M})$ is the set of m-separations in $\mathbb{M}$. $\mathbf{NE}(P)$, $\mathbf{CH}(P)$, and $\mathbf{PNE}(P)$ are the neighbours, spouses, and potential neighbours, respectively, of P in $\mathbb{M}$. n is the number of chromosomes. $\mathbf{G}_k$ is the subset of $\mathbf{G}$ on the k-th chromosome. $\mathbf{I}(\mathbb{M})[_{\mathbf{G}_k \cup \{P\}}$ is the subset of $\mathbf{I}(\mathbb{M})$ over $\mathbf{G}_k \cup \{P\}$. $X \perp Y \mid \mathbf{Z}$ denotes that nodes X and Y are m-separated given set of nodes $\mathbf{Z}$.

Input: $\mathbf{I}(\mathbb{M})$
Output: $\mathbf{NE}(P)$, $\mathbf{CH}(P)$, and $\mathbf{PNE}(P)$

$\mathbf{NE}(P) \leftarrow \emptyset$
$\mathbf{CH}(P) \leftarrow \emptyset$
$\mathbf{PNE}(P) \leftarrow \emptyset$
for each $1 \leq k \leq n$ **do**
    ▷ Learn nodes adjacent to P
    let $\mathbf{TA}_k(P)$ and $\mathbf{Sepset}_k^P$ be the output of Algorithm 3 with $\mathbf{I}(\mathbb{M})[_{\mathbf{G}_k \cup \{P\}}$ and P as input.
    ▷ Differentiate nodes adjacent to P
    for each $G_1 \in \mathbf{TA}_k(P)$ **do**
        if $\exists G_2 \in \mathbf{G} \setminus \mathbf{TA}_k(P)$ s.t. $G_1 \in \mathbf{Sepset}_k^P(\{G_2, P\})$ **then**
            $\mathbf{NE}(P) \leftarrow \mathbf{NE}(P) \cup \{G_1\}$
        else if $\exists G_2 \in \mathbf{G}_k \setminus \mathbf{TA}_k(P)$ s.t. $G_1 \perp G_2 \mid \mathbf{Sepset}_k^P(\{G_2, P\}) \notin \mathbf{I}(\mathbb{M})$ **then**
            $\mathbf{CH}(P) \leftarrow \mathbf{CH}(P) \cup \{G_1\}$
        else
            $\mathbf{PNE}(P) \leftarrow \mathbf{PNE}(P) \cup \{G_1\}$
        end if
    end for
end for

2. *G_1 is not a cause of S.*

3. *S is not a cause of G_1.*

4. *G_1 and S do not have a common cause.*

3.3 Epistasis with absent marginal effects

Epistasis with weak or absent marginal effects is considered as a possible explanation for the lack of discoveries in genetic studies. However, the literature is inconsistent when it comes to the exact definition of epistasis. Here, a general definition is adopted: epistasis is the scenario in which the effect of a variable X on another variable Y depends on a set $\mathbf{Z}$ of other variables [McKinney et al., 2006]. If the

Algorithm 13 Estimate the genotype–phenotype links from a conditional genetic random sample. $\mathbf{G}$ is a set of genotypes and P is a phenotype. D is a conditional random sample over $\mathbf{G} \cup \{P\}$. $0 < \alpha < 1$ is the significance level for the hypothesis tests of conditional independence performed by the algorithm. $0 < q < 1$ is an FDR threshold. $\mathbf{I}(D)$ is the set of m-separations among the variables in D as determined by performing conditional independence tests on D. $\widehat{\mathbf{AD}}(P)$ is the estimated set of genotypes adjacent of P in the true causal MAG over $\mathbf{G} \cup \{P\}$. n is the number of chromosomes. $\mathbf{G}_k$ is the subset of $\mathbf{G}$ on the k-th chromosome. $\mathbf{I}(D)\lfloor_{\mathbf{G}_k \cup \{P\}}$ is the subset of $\mathbf{I}(D)$ over $\mathbf{G}_k \cup \{P\}$.

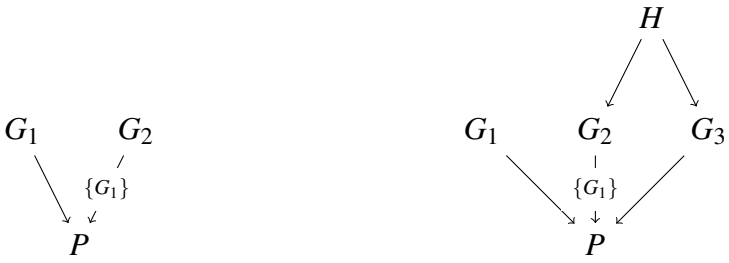
Input: D, α, q
Output: $\widehat{\mathbf{AD}}(P)$

1: **for each** $1 \leq k \leq n$ **do**
2: let $\mathbf{TA}_k(P)$ be the first output of Algorithm 3 with $\mathbf{I}(D)\lfloor_{\mathbf{G}_k \cup \{P\}}$ and P as input.
3: **end for**
4: let $\widehat{\mathbf{AD}}(P)$ be the result of applying an appropriate FDR-controlling procedure to the maximal conditional-independence p-values corresponding to the genotypes in $\bigcup_{1 \leq k \leq n} \mathbf{TA}_k(P)$ to control the FDR of the genotypes below q

marginal effect of X to Y is weak, X might be overlooked in a statistical or machine-learning analysis that does not take $\mathbf{Z}$ into account; if the effect is *absent*, X *will* be overlooked. X might be discovered, however, if $\mathbf{Z}$ is considered as well.

In structure learning, the effect of a variable X on another variable Y may be defined as the union of the effects of X on Y conditional on subsets of the rest variables. If some conditional effects are weak (resp. absent), structure learning might (resp. will) miss a link between X and Y. Clearly, epistasis with absent marginal effects is a PIR with a marginal independence. In this work, dealing with PIRs, and unfaithfulness in general, is not attempted. Nevertheless, a few examples of PIRs are provided in order to illustrate cases where the algorithms developed in this chapter will fail to discover nodes adjacent to the phenotype. Figures 3.3a and 3.3b show an example of a PIR with a marginal and a conditional independence, respectively, in a plausible genetic causal DAG.

The definition of PIR can be extended to MAGs, where it is possible for a PIR to be a bidirected edge. After marginalisation of a causal DAG, it is possible for a PIR to be replaced by a set of new PIRs with potentially different depsets (Figure 3.4), no longer be a PIR (Figure 3.5), or become undetectable (Figure 3.6).



(a) G_2 is marginally independent from P but conditioning on G_1 renders them dependent; this is an example of epistasis with absent marginal effects.

(b) G_2 is marginally dependent to P (through H and G_3), conditioning on G_3 renders them independent, and conditioning on $\{G_3, G_1\}$ renders them dependent again.

Figure 3.3: Example of a PIR with a (a) marginal and a (b) conditional independence in a plausible genetic causal DAG. An edge labelled with **D** denotes a PIR with depset **D**. G_i ($1 \leq i \leq 3$) is a genotype, H is a common cause of G_2 and G_3, and P is a phenotype.

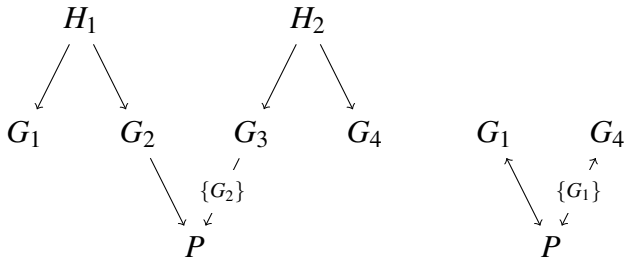


Figure 3.4: Example of a PIR that is replaced by another with a different depset after marginalisation of a plausible genetic causal DAG. The graph on the left is a plausible genetic causal DAG, while the graph on the right is the plausible genetic causal MAG over $\{G_1, G_4, P\}$. An edge labelled with **D** denotes a PIR with depset **D**. G_i ($1 \leq i \leq 4$) is a genotype, H_1 is a common cause of G_1 and G_2, H_2 is a common cause of G_3 and G_4, and P is a phenotype. PIR $G_3 \to P$ with depset $\{G_2\}$ in the DAG is replaced by PIR $G_4 \leftrightarrow P$ with depset $\{G_1\}$ in the MAG.

In a genetic association study, *pure* epistasis among a set of variants means that every variant is marginally independent from the disease but the variants are jointly dependent to the disease; in the worst case, namely *strict* epistasis, no proper subset of the variants is dependent to the disease [Jiang and Neapolitan, 2012]. Both types of epistasis can be explained in terms of PIRs. In pure epistasis, one or more variants has a PIR with a marginal independence with the disease and the depset of each PIR is a subset of the variants with the rest PIRs. In strict epistasis, each variant in a set of variants has a PIR with a marginal independence with the disease with all the

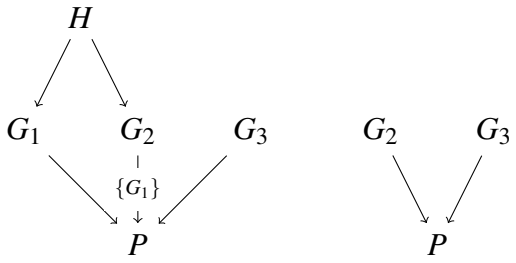


Figure 3.5: Example of a PIR that is no longer a PIR after marginalisation of a plausible genetic causal DAG. The graph on the left is a plausible genetic causal DAG, while the graph on the right is the plausible genetic causal MAG over $\{G_1, G_2, P\}$. G_i $(1 \leq i \leq 3)$ is a genotype, H is a common cause of G_1 and G_2, and P is a phenotype. An edge labelled with **D** denotes a PIR with depset **D**. PIR $G_2 \rightarrow P$ in the DAG is not a PIR in the MAG because G_1, the only variable that renders G_2 and P independent in the DAG, is not present in the MAG.

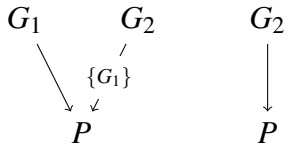
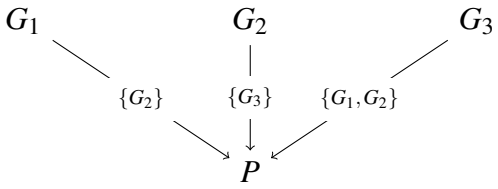


Figure 3.6: Example of a PIR that becomes undetectable after marginalisation of a plausible genetic causal DAG. The graph on the left is a plausible genetic causal DAG, while the graph on the right is the plausible genetic causal MAG over $\{G_2, P\}$. G_1 and G_2 are genotypes and P is a phenotype. An edge labelled with **D** denotes a PIR with depset **D**. PIR $G_2 \rightarrow P$ in the DAG is undetectable in the MAG: G_2 and P are independent but there is no depset in the MAG to render them dependent again.

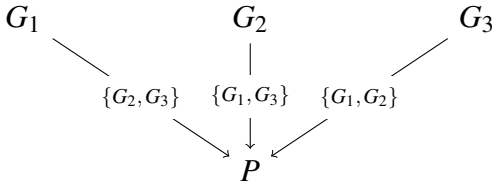
other variants in the set as its depset. Figures 3.7a and 3.7b show an example of pure and strict epistasis, respectively.

3.4 Simulation study

Using standard benchmark BNs , Aliferis et al. [2010a] demonstrated that instantiations of symmetric GLL-PC (Algorithm 5), including MMPC, are superior (in terms of power and FPR) to non-causal feature-selection algorithms in the task of parent-and-children learning. Also using standard benchmarks, Armen and Tsamardinos [2014] showed that FDR control of the output of the instantiation of LGL (Algorithm 6) with MMPC is achieved with sufficient sample size when the CFC is satisfied. Nevertheless, it would be of great interest to specifically evaluate the performance of Algorithm 13 using realistic simulated genetic case–control datasets.



(a) Example of pure epistasis. Each genotype in $\{G_1, G_2, G_3\}$ is marginally independent from P, but conditioning on G_2, G_3, and $\{G_1, G_2\}$ renders G_1, G_2, and G_3, respectively, dependent to P.



(b) Example of strict epistasis. Each genotype in $\{G_1, G_2, G_3\}$ is marginally independent from P, but conditioning on the rest genotypes renders them dependent to P.

Figure 3.7: Example of (a) pure and (b) strict epistasis in a plausible genetic causal MAG. G_i ($1 \leq i \leq 3$) is a genotype and P is a phenotype. An edge labelled with **D** denotes a PIR with depset **D**.

There are several genome-simulation packages available [Ritchie and Bush, 2010], which usually simulate *population* samples; in order to simulate *case–control* samples, a disease model is independently specified and used to assign disease status to the observations in a population sample. Clearly, this is grossly inefficient for rare diseases; for a disease as rare as sCJD, one million observations have to be generated, on average, to obtain a case. Another drawback of the available methods is that the true causal MAG over the observed variables is not available. These methods therefore cannot be used to evaluate the performance of the algorithms presented in this chapter, in general, because the learned structure needs to be compared with the true causal MAG; it is not hard to see that only Algorithm 13 can be evaluated and only when all causes of the disease are observed (although this is the case in this simulation study anyway). To overcome these drawbacks, a different approach is followed here: a CBN with hidden variables is learned from a real genetic dataset and a disease node is attached to it; case–control samples are then generated from the network using Bayesian-network inference to compute the conditional distribution of the variants given each disease status. Us-

ing this approach, case–control samples from a disease of any prevalence can be efficiently generated and the MAG over a subset of the variants and the disease can be obtained via marginalisation of the DAG. The approach was used to perform a simulation study of the performance of Algorithm 13, whose details are explained in the following sections. The study was conducted in MATLAB using additional tools for some computations. For Bayesian-network inference, the junction tree algorithm was used.

3.4.1 Learning a causal Bayesian network from a real genetic sample

Currently, there are no large exome-sequencing datasets available from a certain population (for example, only about 100 individuals are sequenced per population in *The 1000 Genomes Project*).[1] For this reason, a combined autosomal exome-sequencing dataset of 1013 prion disease, Alzheimer's disease, Huntington's disease, frontotemporal dementia, and glaucoma cases with a white British background was obtained from the Prion Unit and assumed to be a random sample of the white British population. The dataset was post-processed as follows. Only exonic non-synonymous variants were retained using *ANNOVAR* [Wang et al., 2010b] and *GATK* [McKenna et al., 2010]. Genomic coordinates were converted from build (version) 18 of the reference human genome to build 38 using *liftover*[2] and variants in non-reference chromosomes, duplicate variants (defined as variants with the same genomic coordinates), variants with all values missing in the dataset, monomorphic variants (that is, variants with only one allele specified), and constant variants (that is, variants with only one allele in the dataset) were removed. The resulting dataset had 89030 variants.

For each autosome, the skeleton of the MAG over the variants was learned by performing hypothesis tests of marginal independence between pairs of variants within 500 *kilobases (kb)* of each other; variants on different chromosomes were assumed to be independent (Assumptions 3.2 and 3.3), and variants on the same

[1] http://browser.1000genomes.org/index.html
[2] https://genome.ucsc.edu/cgi-bin/hgLiftOver

chromosome but further than 500 kb apart were assumed to be independent due to the observation that LD decays considerably within 500 kb in humans [e.g. see Figure 3 in Ke et al., 2004]; both assumptions help to reduce false negatives and the computational burden of the method. The G test with the degrees of freedom adjustment heuristic and determinism detection was used for the tests and the heuristic power rule with $h\text{-}ps = 5$ was used as the reliability criterion for the tests. The FDR of the combined skeleton over all autosomes was controlled below 5% using the permutation method of Storey and Tibshirani [2001] with 100 permutations. When it is possible to simulate null p-values (as is the case here, by independently permuting the observations of each variant), the method of Storey and Tibshirani [2001] provides a less conservative alternative to the BY procedure for controlling the FDR under p-value dependence.

In an undirected graph, a *clique* is a set of nodes that are all adjacent to each other; a clique is *maximal* if it is not a subset of another. Owing to Assumption 3.2, variants that are adjacent in the skeleton have a hidden common cause. In order to reduce computational complexity, it was assumed that all variants that form a maximal clique in the skeleton have the *same* hidden common cause. For each autosome, the maximal cliques of the skeleton were identified using the modification of the *Bron–Kerbosch* algorithm by Eppstein et al. [2010]. A DAG over the variants and a hidden variable for each maximal clique was then created with each hidden variable set as a parent of the variants in the corresponding maximal clique. The number of parents of each variant varied between 0 and 7928. Learning the CPD of a node with 7928 parents is impossible, so for each variant with more than 3 parents, the parents were sorted in descending order based on the mean association of the variant with the other children of the parent and only the top 3 parents were retained.

To simplify parameter learning, the hidden variables were assumed to be categorical. Ideally, the number of levels of each hidden variable would be optimised as in Chen et al. [2012]. In order to reduce computational time, however, the hidden variables were assumed to be ternary. Moreover, parameter learning was per-

formed for chromosome 22 only using the *Expectation–Maximisation (EM)* algorithm within the random-restart scheme of Maxwell Chickering and Heckerman [1997] with 8 restarts only (instead of 64). Learning was performed independently for each *weakly-connected component* of the DAG (that is, sets of nodes that are mutually reachable by violating edge directions).

3.4.2 Attaching a disease node

Following the approach of the *SIMLA* simulator [Schmidt et al., 2005], a binary disease variable P was modelled as a logistic regression in functions $f_1(g_1),\ldots,f_n(g_n)$ of the values $g_1,\ldots,g_n$ of causal genotypes $G_1,\ldots,G_n$:

$$\mathscr{P}(P=\text{affected} \mid G_1=g_1,\ldots,G_n=g_n)=\frac{\exp(\beta_0+\sum_{i=1}^{n}\ln(rr_i)f_i(g_i))}{1+\exp(\beta_0+\sum_{i=1}^{n}\ln(rr_i)f_i(g_i))} \quad (3.1)$$

where β_0 is the intercept, rr_i $(1\leq i\leq n)$ is the rare homozygous RR for g_i, and $f_i(g_i)$ is defined as follows:

$$f_i(g_i)\triangleq\begin{cases}0 & \text{if } g_i=\text{AA}\\ w_i & \text{if } g_i=\text{Aa}\\ 1 & \text{if } g_i=\text{aa}\end{cases}$$

where w_i specifies the *mode of inheritance* for genotype G_i. For example, setting $w_i=0$ specifies a *recessive* model, where a single copy of the rare allele has no effect on the disease, setting $w_i=0$ specifies a *dominant* model, where a single copy has the same effect as two copies, and $w_i=(\ln(rr_i)-\ln(2))/\ln(rr_i)$ specifies an *additive* model, where a single copy has half of the effect of two copies. If the prevalence of the disease

$$\mathscr{P}(P=\text{affected})=\sum_{g_1\in\{AA,Aa,aa\}}\cdots\sum_{g_n\in\{AA,Aa,aa\}}$$
$$\mathscr{P}(P=\text{affected} \mid G_1=g_1,\ldots,G_n=g_n)\mathscr{P}(G_1=g_1,\ldots,G_n=g_n)$$

and the joint distribution $\mathscr{P}(G_1,\ldots,G_n)$ of the causal genotypes is specified, equation 3.1 can be solved for the intercept β_0 by finding the root of the following function using a numerical method such as Newton's:

$$g(\beta_0) \triangleq \mathscr{P}(P = \text{affected}) - \sum_{g_1\in\{AA,Aa,aa\}} \cdots \sum_{g_n\in\{AA,Aa,aa\}} \frac{\exp(\beta_0 + \sum_{i=1}^{n} \ln(rr_i) f_i(g_i))}{1+\exp(\beta_0 + \sum_{i=1}^{n} \ln(rr_i) f_i(g_i))} \mathscr{P}(G_1,\ldots,G_n)$$

First, the marginal distribution $\mathscr{P}(G_i)$ of each genotype was computed from the CBN and the *minor allele frequency (MAF)* maf_i of each variant was calculated using the formula $\text{maf}_i = \mathscr{P}(G_i = \text{aa}) + \mathscr{P}(G_i = \text{Aa})/2$. For each of three MAF ranges $(0, 0.01)$, $[0.01, 0.05)$, and $(0.05, 1)$, one variant was randomly selected as a causal variant. Table 3.1 contains information about the causal variants. For all of them, the rare homozygous RR was set to 10 and the heterozygous RR was set to 5. The disease prevalence was set to $1/1000000$ and the joint distribution of the causal genotypes was computed from the CBN. After solving for the intercept, a disease node with the resulting CPD was added to the network as a child of the causal variant.

#	RS#	Position	MAF	RR	FM	#CC	nCC	IN	nLD	PD (CD)	PD (AA)
1	rs192851961	44786481	0.0060	10	0.006	3663	1	0	0	0.01	0.15
2	rs202154713	50120597	0.0316	10	0.003	4055	91	3	3	0.74	1.00
3	rs7235	20319733	0.4265	10	0.023	458	34	2	3	0.99	1.00

Table 3.1: Causal variants in the simulation study of the performance of Algorithm 13, ordered by MAF. *RS#* is the RS number of the variant in dbSNP. Position is on chromosome 22. *RR (aa)* is the rare homozygous RR. *FM*denotes the missing-value frequency. *#CC* is the number of weakly-connected component. *nCC* denotes the number of variants in the component. *IN* is the number of parents of the variant. *nLD* denotes the number of variants in LD with the causal variant, which is equal to the number of siblings of the variant (that is, variants that share a parent with the variant). *PD (CD)* and *PD (AA)* denote the probability of discovery in causal discovery and association analysis, respectively (see Section 3.4.4).

3.4.3 Sampling

Clearly, only the distribution of the variants in the weakly-connected component of the DAG that contains the disease node differs between cases and controls; these

variants are said to be *relevant* to the disease. The distribution of the rest variants is unaffected by the disease status and is equal to the population distribution (which is encoded by the CBN) in both cases and controls. Therefore, 100 samples with 1000 observations each were first generated from the network. The conditional distribution of the relevant variants given each disease status (the distribution in controls and cases, respectively) was then computed from the network, and 1000 samples with 500 observations were generated from each distribution. The three sets of samples were subsequently combined to create a set of 100 case–control samples.

Real genetic datasets contain missing values. To learn the CBN, data were implicitly assumed to be *missing completely at random*. Under the same assumption, for each variant, the same proportion of missing values as in the original dataset was randomly set to missing in every case–control sample.

3.4.4 Results

Algorithm 13 was applied to each sample using non-symmetric MMPC (Algorithm 4) as the instantiation of Algorithm 3, the G test as in Section 3.4.1, and the approach of Armen and Tsamardinos [2014] with the BY procedure in order to control the FDR of the nodes adjacent to the disease below 5%. The *average power* (the expected proportion of rejected true null hypotheses among the true null hypotheses) and the FDR were computed by approximating expectations by the respective means across the samples and the probability of discovery of each causal variant (listed in Table 3.1) was approximated by the respective relative frequency in the samples. The average power was 0.58 and the FDR was 0, implying successful control. While the common variant, rs7235, is almost always discovered, the rare variant, rs192851961, is almost always not. The other variant, rs202154713, is discovered with probability 0.74. Although rs7235 and rs202154713 are in LD with 3 non-causal variants each, the latter are not discovered, as expected.

Algorithm 13 was also applied with a maximal conditioning-set cardinality of zero, making it equivalent to association analysis with the same test and reliability criterion, followed by application of the BY procedure. The average power and the

FDR *with regard to causal variants* was 0.75 and 0.092, respectively. The average power of association analysis is higher than that of causal discovery, which is expected, as only marginal tests are performed for each variant. The nonzero FDR implies that additional, non-causal variants are discovered. It turns out that the FDR is solely due to the discovery with probability 0.28 of a variant in LD with rs7235.

The results of this simulation study show that Algorithm 13 is capable of discovering causal variants and discarding definitely-non-causal ones while controlling the FDR, in expense of reduced average power compared to association analysis.

3.5 Application to prion disease

Three autosomal datasets from the GWAS, two autosomal datasets from the exome-sequencing study, and three autosomal datasets from the exome-array study in prion disease were post-processed as in Section 3.4.1; selection of exonic non-synonymous variants was only performed for the exome-sequencing datasets. Table 3.2 contains information about the post-processed datasets. Algorithm 13 was applied to each dataset as in Section 3.4.4. Table 3.3 lists the discoveries made from each dataset.

Only PRNP codon 129 (rs1799990) was discovered from the GWAS datasets. This is expected, as the SNPs in LD with rs1799990 discovered in the GWAS were found to be independent from the disease given rs1799990 in the subsequent conditional analysis. Two SNPs were discovered (rs73460769 and rs755431752) from the UK sCJD and 22 discoveries were made from the UK vCJD exome-sequencing dataset, including rs1799990 but neither rs73460769 or rs755431752. In the sCJD dataset, rs1799990 was rendered independent from the disease by three SNPs that were not discovered either. There were one, five, and seven discoveries from the UK, German, and US sCJD exome-array dataset, respectively. All but three SNPs were also discovered in the exome-array study but failed post-association quality control (see Section B.3). SNPs rs45458098, rs200992486, and rs41311333 were discovered by Algorithm 13 but not discovered in the exome-array study. SNP rs45458098 was discovered from the German and the US dataset; rs200992486

and rs41311333 were discovered from the German and US dataset, respectively. Although the significance level was lower in the exome-array study (10^{-6} vs. $5.54 \cdot 10^{-9}$ and $1.87 \cdot 10^{-8}$ in the sCJD and vCJD datasets, respectively, as determined by the FDR-controlling procedure) and thresholding was applied to association, not link-absence p-values, these three SNPs did not achieve significance in the exome-array study. In that study, however, the observations were taken to be the gametes of the individuals, not the individuals themselves, and association analysis was performed using Fisher's exact test, not the G test.

The experiment was repeated with a maximal conditioning-set cardinality of zero. There were 5, 3, 2, 469, 6407, 3, 7, and 16 discoveries from the respective datasets in Table 3.2, highlighting the utility of causal-discovery methods in discarding definitely-non-causal variants.

For each dataset, the SNPs discovered by Algorithm 13 were used as features in training a *support vector machine (SVM)* [Cristianini and Shawe-Taylor, 2000] using the default parameters in MATLAB. Since SVMs work with continuous features, *1-of-k encoding* (here $k = 3$) was used to convert genotypes to features: three features were constructed per SNP, with AA, Aa, and aa represented as $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$, respectively. Individuals with any of these genotypes missing were removed, because SVMs cannot handle missing data. The number of cases, controls, and individuals in each resulting dataset and the corresponding mean *area under the curve (AUC)* [Fawcett, 2003] across 5 cross-validation folds is listed in Table 3.4. The low AUCs indicate that the discovered SNPs are not good predictors of the disease, which indicates that additional causal variants remain to be discovered.

3.6 Related work

Three other works on constraint-based learning from genetic datasets are discussed below. All of them assume causal sufficiency; the first two assume that the CFC holds and the third one deals with information equivalences. Unlike Algorithm 13, the algorithms in those works are not applied to each chromosome independently

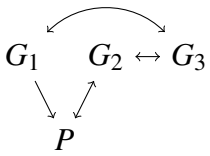
#	Pop.	Dis.	Type	Controls	b	n1	n2	n	m
1	UK	sCJD	GWAS	WTCCC2	18	524	5200	5724	499653
2	UK	vCJD	GWAS	WTCCC2	18	125	5200	5325	499545
3	DE	sCJD	GWAS	KORA-gen	18	634	815	1449	479379
4	UK	sCJD	ES	HD+FTD+Glaucoma	19	224	663	887	381354
5	UK	vCJD	ES	HD+FTD+Glaucoma	19	97	663	760	337069
6	UK	sCJD	EA	GERAD	19	622	822	1444	84496
7	DE	sCJD	EA	KORA-gen	19	719	2757	3476	103638
8	US	sCJD	EA	Coriell	19	814	840	1654	91109

Table 3.2: Genetic case–control datasets from prion disease on which Algorithm 13 was applied. Pop. and Dis. stands for population and disease, respectively. *ES* and *EA* stand for exome sequencing and exome array, respectively. The controls used are indicated in the respective column. *WTCCC2* refers to controls from *Wellcome Trust Case Control Consortium 2* [Burton et al., 2007], *KORA-gen* to controls from the *KORA-gen* resource [Wichmann et al., 2005]. *HD+FTD+Glaucoma* to Huntington's disease, frontotemporal dementia, and glaucoma cases from the Prion Unit, *GERAD* to controls from the *GERAD Consortium (Genetic and Environmental Risk in Alzheimer's Disease)*, and *Coriell* to controls from the *Coriell Institute for Medical Research*. All cases are from the Prion Unit. *b* denotes the build of the human genome used in the datasets before conversion to build 38 (which is the successor of build 19). *n1*, *n2*, *n*, and *m* denote the number of cases, controls, individuals, and variants, respectively.

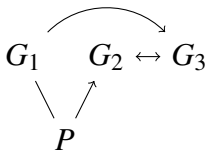
and no multiple-testing correction is performed. Finally, no theory was formulated about the causal relationships between the variables in the genetic datasets, causal effects were not discussed, and no simulation was performed.

Like non-symmetric MMPC used here, *DASSO-MB (Detection of ASSOciations using Markov Blanket)* [Han et al., 2010] is an instantiation of non-symmetric GLL-PC; however, no theoretical justification is given for the lack of symmetry correction. Applied to simulated datasets with 100 SNPs and 2000 individuals from three disease models with two causal SNPs, the output of DASSO-MB was correct (all true positives and no false positives were identified) more often than the output of well-established methods *Multifactor Dimensionality Reduction (MDR)*, *step-PLR*, and *Bayesian epistasis association mapping (BEAM)*, and a *Support Vector Machine (SVM)* approach. The algorithm was also applied to a real case–control GWAS dataset from *Age-related Macular Degeneration (AMD)* with 91,495 SNPs and 146 individuals from which association analysis had identified two associated SNPs; one of the SNPs and some other SNP were identified.

The *FEPI-MB (Fast EPistatic Interactions detection using Markov Blanket)* algorithm [Han et al., 2011] assumes that the phenotype has no children. When this is the case, it follows from the proof of Theorem 2.2 that, if the phenotype P and genotype G are d-separated by a subset of the rest genotypes, then P and G are d-separated by the set of parents of P. Therefore, non-symmetric GLL-PC can be used and in the elimination phase it is sufficient to check whether $T \perp Y \mid \mathbf{TA}(T) \setminus \{Y\} \in \mathbf{I}(\mathbb{G})$ in order to remove Y from $\mathbf{TA}(T)$. FEPI-MB (Algorithm 14) is an instantiation of this variation of non-symmetric GLL-PC. Based on the results of this chapter, FEPI-MB is incorrect for both unconditional and conditional genetic samples. Consider the plausible genetic causal MAG of Figure 3.8a. If G_2 enters $\mathbf{TA}(T)$ first, then G_1 and G_3 would follow and would never be removed from $\mathbf{TA}(T)$. Therefore, FEPI-MB would incorrectly discover G_3. The same is true for the plausible conditional genetic causal MAG of Figure 3.8b. This shows the importance of a solid theoretical basis for algorithms learning from real data. In any case, FEPI-MB was shown to outperform BEAM, SVM, and MDR in the same experimental setup as in Han et al. [2010] using one more model with three loci and epistasis with weak marginal effects. Applied to an AMD dataset with 97,327 SNPs, FEPI-MB discovered one GWAS hit and one more SNP.



(a) Example of a plausible genetic causal MAG where the output of the FEPI-MB algorithm may be incorrect.



(b) Example of a plausible conditional genetic causal MAG where the output of the FEPI-MB algorithm may be incorrect.

Figure 3.8: Example of a plausible genetic causal MAG (a) and a plausible conditional genetic causal MAG (b) where the output of the FEPI-MB algorithm (Algorithm 14) may be incorrect. G_i ($1 \leq i \leq 3$) is a genotype and P is a phenotype. If G_2 enters $\mathbf{TA}(T)$ first, then G_1 and G_3 would follow and would never be removed from $\mathbf{TA}(T)$.

Alekseyenko et al. [2011] applied an instantiation of TIE* to a case–control GWAS dataset in *rheumatoid arthritis* with 490,073 SNPs and 2044 individuals.

Algorithm 14 The FEPI-MB algorithm [Han et al., 2011]. $\mathbf{V}$ is a set of random variables. $\mathbb{G}$ is a DAG over $\mathbf{V}$. $\mathbf{I}(\mathbb{G})$ is the set of d-separations in $\mathbb{G}$. $T \in \mathbf{V}$ is the target variable. $X \perp Y \mid \mathbf{Z}$ denotes that nodes X and Y are d-separated given set of nodes $\mathbf{Z}$. $\text{assoc}(X, Y \mid \mathbf{Z})$ denotes the association of variables X and Y given set of variables $\mathbf{Z}$.

Input: $\mathbf{I}(\mathbb{G})$ and T
Output: $\mathbf{TA}(T)$

▷ Initialisation
$\mathbf{TA}(T) \leftarrow \emptyset$
$\mathbf{OPEN}(T) \leftarrow \mathbf{V} \setminus \{T\}$
repeat
    ▷ Insertion phase
    for each $Y \in \mathbf{OPEN}(T)$ s.t. $T \perp Y \mid \mathbf{TA}(T) \in \mathbf{I}(\mathbb{G})$ **do**
        $\mathbf{OPEN}(T) \leftarrow \mathbf{OPEN}(T) \setminus \{Y\}$
    end for
    $W \leftarrow \arg\max_{W \in \mathbf{OPEN}(T)} \text{assoc}(T, W \mid \mathbf{TA}(T))$
    $\mathbf{TA}(T) \leftarrow \mathbf{TA}(T) \cup \{W\}$
    $\mathbf{OPEN}(T) \leftarrow \mathbf{OPEN}(T) \setminus \{W\}$
    ▷ Elimination phase
    for each $Y \in \mathbf{TA}(T)$ s.t. $T \perp Y \mid \mathbf{TA}(T) \in \mathbf{I}(\mathbb{G})$ **do**
        $\mathbf{TA}(T) \leftarrow \mathbf{TA}(T) \setminus \{Y\}$
    end for
until $\mathbf{OPEN}(T) = \emptyset$

They discovered two five-SNP Markov boundaries of the phenotype sharing four SNPs and having six SNPs in total; the unique SNPs in each Markov boundary were in perfect LD with each other, that is, their relation was deterministic. Five of the SNPs are from a locus already implicated in the disease, and four of the SNPs had achieved significance in a meta-analysis of the disease. The SNPs were subsequently used to build a predictive model with an AUC of 0.81. Finally, the SNPs rendered all other previously SNPs known to be associated with the disease independent from the disease.

3.7 Summary and future work

The theoretical work in this chapter culminated in the development of a specialised algorithm (Algorithm 13) for learning FDR-controlled genotype–phenotype links from a genetic case–control sample. A Bayesian-network-based simulator of realistic genetic case–control datasets was created and used to evaluate the algorithm's

performance before its application to datasets from prion disease. In the case of the GWAS datasets, the algorithm successfully discarded SNPs that are not causal of the disease but associated with it through PRNP codon 129 (rs1799990). The algorithms developed in the next chapter for learning from genetic datasets with overlapping sets of variants are based on the theory developed in this chapter. Note that this theory is applicable, with minor modifications, to other types of *cross-sectional* datasets (datasets whose variables are measured at a specific point in time) as well, assuming that there is no instantaneous causation. For example, a case–control gene-expression dataset is defined over a set of gene expressions and a phenotype. Since expression is measured at a specific point in time for all genes, it may be assumed that no gene expression is a cause of another.

Future work includes adapting Algorithm 13 to output the orientation of genotype–phenotype edges after performing appropriate multiple-testing corrections and developing algorithms that deal with PIRs and information equivalences in genetic sets of variables. In addition, a CBN could be learned for *all* autosomes from the exome-sequencing dataset of Section 3.4.1 performing more than 8 restarts of the EM algorithm this time, in order to ensure a better fit. A network could be also learned from a GWAS dataset. Further simulation studies would use samples of various sizes from these networks and explore different disease models, where some of the causal variants may be hidden. For additional realism, a molecular entity of the cell may be chosen to be disrupted in the disease and only variants in genes whose products have a known causal path to the entity may be selected as causal. For this purpose the graph developed in Chapter 5 may be used.

D#	RS#	C#	Position	M(A)	M(U)	GF (A)	GF (U)	P-value	OR (Aa)	OR (Aa) 95% CI	OR (aa)	OR (aa) 95% CI
1	rs1799990	20	4699605	0.311	0.349	0.584/0.209/0.207	0.428/0.446/0.126	$2.10 \cdot 10^{-18}$	0.34	[0.27, 0.43]	1.20	[0.95, 1.53]
2	rs1799990	20	4699605	0.004	0.349	0.992/0.008/0.000	0.428/0.446/0.126	$7.63 \cdot 10^{-12}$	0.01	[0.00, 0.06]	0.00	[0.00, N/A]
3	rs1799990	20	4699605	0.232	0.335	0.697/0.142/0.161	0.450/0.431/0.119	$6.30 \cdot 10^{-23}$	0.21	[0.16, 0.28]	0.87	[0.64, 1.19]
4	rs73460769	11	48365783	0.396	0.470	0.295/0.619/0.086	0.068/0.923/0.009	$7.59 \cdot 10^{-10}$	0.15	[0.09, 0.26]	2.20	[0.64, 7.48]
4	rs755431752	16	56451153	0.124	0.020	0.752/0.248/0.000	0.961/0.039/0.000	$9.72 \cdot 10^{-11}$	8.06	[4.90, 13.24]	N/A	[N/A, N/A]
5	rs149796618	1	12847431	0.500	0.042	0.000/1.000/0.000	0.918/0.080/0.002	$3.44 \cdot 10^{-17}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
5	N/A	2	95944647	0.500	0.045	0.000/1.000/0.000	0.911/0.089/0.000	$6.48 \cdot 10^{-72}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
5	rs149290349	2	43224818	0.000	0.487	1.000/0.000/0.000	0.184/0.658/0.158	$1.03 \cdot 10^{-13}$	0.00	[0.00, N/A]	0.00	[0.00, N/A]
5	rs62269817	3	75937535	0.443	0.039	0.115/0.885/0.000	0.923/0.075/0.002	$5.45 \cdot 10^{-22}$	94.78	[47.43, 189.39]	0.00	[0.00, N/A]
5	rs73714233	7	100952498	0.500	0.053	0.000/1.000/0.000	0.895/0.105/0.000	$4.81 \cdot 10^{-67}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
5	rs73592211	8	51820668	0.467	0.038	0.066/0.934/0.000	0.924/0.076/0.000	$2.31 \cdot 10^{-68}$	171.18	[71.09, 412.16]	N/A	[N/A, N/A]
5	rs201906409	9	64411223	0.494	0.052	0.012/0.988/0.000	0.896/0.104/0.000	$6.08 \cdot 10^{-35}$	722.04	[98.24, 5306.68]	N/A	[N/A, N/A]
5	rs3127819	10	46550598	0.476	0.234	0.106/0.835/0.059	0.551/0.430/0.019	$1.79 \cdot 10^{-8}$	10.12	[4.58, 22.37]	16.39	[2.75, 97.53]
5	rs75650204	11	48366273	0.500	0.044	0.000/1.000/0.000	0.913/0.087/0.000	$9.46 \cdot 10^{-11}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
5	rs78810484	13	36438729	0.471	0.072	0.059/0.941/0.000	0.856/0.144/0.000	$6.91 \cdot 10^{-31}$	95.33	[36.73, 247.45]	N/A	[N/A, N/A]
5	rs201041085	13	45596584	0.473	0.065	0.272/0.511/0.217	0.898/0.074/0.028	$1.69 \cdot 10^{-7}$	22.79	[12.03, 43.20]	25.87	[10.66, 62.77]
5	N/A	14	19434140	0.763	0.217	0.115/0.244/0.641	0.771/0.025/0.204	$1.42 \cdot 10^{-7}$	63.86	[17.88, 228.14]	21.01	[9.35, 47.20]
5	rs2445603	15	21431800	0.701	0.076	0.037/0.524/0.439	0.896/0.055/0.049	$6.46 \cdot 10^{-23}$	231.63	[67.15, 798.94]	220.36	[62.91, 771.88]
5	rs9921162	16	15001818	0.135	0.412	0.729/0.271/0.000	0.193/0.790/0.017	$1.36 \cdot 10^{-7}$	0.09	[0.05, 0.17]	0.00	[0.00, N/A]
5	rs150910818	17	16165082	0.500	0.044	0.000/1.000/0.000	0.912/0.088/0.000	$1.28 \cdot 10^{-72}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
5	rs79341062	19	8895694	0.387	0.031	0.227/0.773/0.000	0.938/0.062/0.000	$1.14 \cdot 10^{-8}$	51.72	[29.23, 91.52]	N/A	[N/A, N/A]
5	rs10439604	20	30388547	0.500	0.043	0.000/1.000/0.000	0.914/0.086/0.000	$6.24 \cdot 10^{-47}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
5	rs1799990	20	4699605	0.005	0.331	0.990/0.010/0.000	0.455/0.429/0.116	$9.82 \cdot 10^{-9}$	0.01	[0.00, 0.08]	0.00	[0.00, N/A]
5	rs4818891	21	44229660	0.842	0.411	0.079/0.158/0.763	0.233/0.713/0.055	$1.06 \cdot 10^{-9}$	0.65	[0.16, 2.66]	41.19	[11.54, 147.03]
5	rs200870155	22	41221250	0.443	0.030	0.264/0.586/0.149	0.942/0.055/0.003	$5.03 \cdot 10^{-22}$	38.07	[20.97, 69.09]	174.65	[37.22, 819.43]
5	rs267606255	22	40246088	0.320	0.025	0.373/0.613/0.013	0.953/0.043/0.004	$1.02 \cdot 10^{-9}$	36.42	[19.53, 67.89]	9.50	[0.84, 107.95]
5	rs62231276	22	18847708	0.080	0.634	0.851/0.138/0.011	0.321/0.092/0.588	$5.31 \cdot 10^{-9}$	0.57	[0.23, 1.38]	0.01	[0.00, 0.05]
6	rs142631461	17	29249184	0.023	0.000	0.953/0.047/0.000	1.000/0.000/0.000	$5.54 \cdot 10^{-9}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
7	rs1058065	2	27364857	0.016	0.001	0.968/0.032/0.000	0.999/0.001/0.000	$2.08 \cdot 10^{-11}$	30.34	[9.08, 101.33]	N/A	[N/A, N/A]
7	rs45458098	6	34858817	0.032	0.013	0.968/0.000/0.032	0.974/0.026/0.000	$3.21 \cdot 10^{-16}$	0.00	[0.00, N/A]	∞	[N/A, ∞]
7	rs200992486	16	72787445	0.015	0.006	0.985/0.000/0.015	0.989/0.011/0.000	$3.25 \cdot 10^{-10}$	0.00	[0.00, N/A]	∞	[N/A, ∞]
7	rs142631461	17	29249184	0.017	0.000	0.967/0.032/0.001	1.000/0.000/0.000	$6.36 \cdot 10^{-13}$	∞	[N/A, ∞]	∞	[N/A, ∞]
7	rs200542656	19	46495344	0.009	0.000	0.982/0.018/0.000	1.000/0.000/0.000	$1.71 \cdot 10^{-8}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
8	rs114501427	2	138568946	0.021	0.000	0.958/0.042/0.000	1.000/0.000/0.000	$4.67 \cdot 10^{-9}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
8	rs144218313	4	989606	0.038	0.000	0.924/0.076/0.000	1.000/0.000/0.000	$5.44 \cdot 10^{-17}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
8	rs41311333	5	90619108	0.095	0.052	0.905/0.000/0.095	0.898/0.101/0.001	$1.59 \cdot 10^{-31}$	0.00	[0.00, N/A]	78.99	[10.96, 569.43]
8	rs45458098	6	34858817	0.031	0.014	0.969/0.000/0.031	0.971/0.029/0.000	$1.87 \cdot 10^{-8}$	0.00	[0.00, N/A]	∞	[N/A, ∞]
8	rs151199705	14	105392632	0.028	0.000	0.944/0.056/0.000	1.000/0.000/0.000	$2.68 \cdot 10^{-12}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
8	rs142631461	17	29249184	0.000	0.027	1.000/0.000/0.000	0.945/0.055/0.000	$3.94 \cdot 10^{-11}$	0.00	[0.00, N/A]	N/A	[N/A, N/A]
8	rs115310908	22	37069081	0.023	0.001	0.954/0.046/0.000	0.999/0.001/0.000	$5.27 \cdot 10^{-9}$	40.32	[5.52, 294.55]	N/A	[N/A, N/A]

Table 3.3: Discoveries resulted from the application of Algorithm 13 to the datasets in Table 3.2. *D#* is the dataset number. *RS#* is the RS number of the variant in dbSNP, with *N/A* denoting that the variant is not in dbSNP. *C#* is the chromosome number. *M(A/U)* is the MAF in cases/controls. *M(A/U)* is the MAF in cases/controls. *GF (A/U)* are the relative frequencies of homozygotes for the common allele, heterozygotes, and homozygotes for the rare allele in cases/controls. *P-value* is (an upper bound of) the link-absence p-value. *OR (Aa)* and *OR (aa)* is the heterozygote and rare homozygote, respectively, OR, with *N/A* denoting an undefined OR. *OR (Aa) 95% CI* and *OR (aa) 95% CI* are the corresponding 95% confidence intervals. The ORs equal the respective CORs under the conditions of Theorem 3.14.

D#	n1	n2	n	m	$\overline{\text{AUC}}$
1	522	5197	5719	1	0.50
2	125	5197	5322	1	0.50
3	633	812	1445	1	0.63
4	139	439	578	2	0.62
5	7	3	10	22	N/A
6	622	822	1444	1	0.52
7	719	2757	3476	5	0.58
8	795	840	1635	7	0.68

Table 3.4: Datasets from which SVMs were trained and corresponding mean AUC across 5 cross-validation folds, with *N/A* denoting that the mean AUC could not be computed due to the number of cases or controls being less than 5. The datasets were constructed from the ones in Table 3.2 by selecting the SNPs discovered by Algorithm 13, removing individuals with missing values for those SNPs, and performing 1-of-3 encoding of the genotypes. *D#* is the dataset number in Table 3.2. *n1*, *n2*, *n*, and *m* denote the number of cases, controls, individuals, and SNPs discovered by Algorithm 13, respectively.

Chapter 4

Causal discovery from genetic datasets with overlapping sets of variants

In this chapter, an algorithm is devised that, given a set of conditional genetic random samples with the same phenotype and overlapping sets of variants, identifies all conditional genetic causal PAGs over all variables that are consistent with the samples. Since the causal relationships between the variants and the phenotype are of primary interest, an algorithm is developed that identifies all causal relationships between variants and phenotype that are consistent with the samples. Finally, an algorithm that learns the FDR-controlled genotype–phenotype links that are consistent with the samples is presented and applied to a combination of datasets from prion disease.

4.1 Learning consistent plausible conditional genetic causal MAGs

Given a set of conditional genetic random samples with the same phenotype and overlapping sets of variants, an algorithm such as IOD can be used to learn the PAGs that are consistent with the samples. However, it is the set of consistent plausible conditional genetic causal PAGs that is of interest here. Consistent plausible conditional genetic causal MAGs and PAGs are formally defined as follows.

Definition 4.1 (Consistent plausible conditional genetic causal MAG)**.** *Let* $\mathbb{M}$ *be a plausible conditional genetic causal MAG defined over a set of variables* $\mathbf{G} \cup \{P\}$*,* $\mathbf{G}_1, \ldots, \mathbf{G}_n \subseteq \mathbf{G}$ $(n \geq 1)$*, and* $\mathbb{M}_k$ *is the marginal of* $\mathbb{M}$ *over* $\mathbf{G}_k \cup \{P\}$ $(1 \leq k \leq n)$*. A plausible conditional genetic causal MAG* $\mathbb{N}$ *over* $\mathbf{G} \cup \{P\}$ *is said to be* consistent *with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$ *if for each* $1 \leq k \leq n$ *the marginal of* $\mathbb{N}$ *over* $\mathbf{G}_k \cup \{P\}$ *is Markov equivalent to* $\mathbb{M}_k$*.*

Definition 4.2 (Consistent maximally-informative plausible conditional genetic causal PAG)**.** *Let* $\mathbb{M}$ *be a plausible conditional genetic causal MAG defined over a set of variables* $\mathbf{G} \cup \{P\}$*,* $\mathbf{G}_1, \ldots, \mathbf{G}_n \subseteq \mathbf{G}$ $(n \geq 1)$*, and* $\mathbb{M}_k$ *is the marginal of* $\mathbb{M}$ *over* $\mathbf{G}_k \cup \{P\}$ $(1 \leq k \leq n)$*. A maximally-informative plausible conditional genetic causal PAG* $\mathbb{Q}$ *over* $\mathbf{G} \cup \{P\}$ *is said to be* consistent *with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$ *if the members of the class of plausible conditional genetic causal MAGs represented by* $\mathbb{Q}$ *are consistent with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$*.*

Note that, by definition of plausible conditional genetic causal MAG and commutativity of MAG marginalisation, the marginal of a plausible conditional genetic causal MAG over a subset of the genotypes and the phenotype is itself a plausible conditional genetic causal MAG. One way to obtain the consistent plausible conditional genetic causal PAGs would be to filter the output of IOD. Another approach would be to directly search for them. Algorithm 15 is inspired by IOD and proved to be sound (Theorem 4.1) and complete (Theorem 4.2), in the sense that it returns all and only consistent plausible conditional genetic causal PAGs. Each part of the algorithm is described below.

- *Part 1:* Algorithms 10 and 11 are first used to learn the plausible conditional genetic causal PAGs $\mathbb{P}_1, \ldots, \mathbb{P}_n$ over the sets of overlapping sets of variables and then a partially-oriented mixed graph $\mathbb{L}$ over all variables is created which provably contain a superset of the links and a subset of the orientations in every consistent plausible conditional genetic causal PAG. Notably, edge $G - P$ exists in $\mathbb{L}$ if it exists in some $\mathbb{P}_k$ $(1 \leq k \leq n)$.

- *Part 2:* Edges that violate Condition (6) of Theorem 3.10 and edges whose

existence is provably equivalent to the existence of an inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ between non-existent genotype–phenotype links in $\mathbb{P}_k$ are removed from $\mathbb{L}$.

- *Part 3:* A set $\mathbf{R}_1$ of "removable" genotype–phenotype links and a set $\mathbf{R}_2$ of "removable" genotype–phenotype links are created; the other, "fixed" links in $\mathbb{L}$ are provably present in every consistent plausible conditional genetic causal PAG. It turns out that all $G - P$ edges in $\mathbb{L}$ are fixed.

- *Parts 4 and 5:* Every subgraph $\mathbb{S}$ of $\mathbb{L}$ containing the fixed links, a subset of the links in $\mathbf{R}_1$, and a subset of the links in $\mathbf{R}_2$ such that Condition (4) of Theorem 3.10 is not violated is considered; subgraphs that violate conditions (5)–(7) of Theorem 3.10 are ignored and additional orientations are performed in order to satisfy conditions (3) and (5) of Theorem 3.10 and prevent edges whose existence is provably equivalent to the existence of an inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ between non-existent genotype–phenotype links in $\mathbb{P}_k$. At the end of part 5, $\mathbb{S}$ provably contains the orientations at P, the orientations at a *subset* of the unshielded genotypes adjacent to P, the orientations at the genotypes not adjacent to P, and no orientations at the remaining genotypes.

- *Part 6:* All orientations at the *remaining* unshielded genotypes adjacent to P that satisfy Condition (2) of Theorem 3.10 are considered, each provably resulting in a plausible conditional genetic causal PAG $\mathbb{Q}$. $\mathbb{Q}$ does not contain any inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ corresponding to a non-existent genotype–phenotype link in $\mathbb{P}_k$ or, as it turns out, any inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ corresponding to a non-existent genotype–genotype link in $\mathbb{P}_k$. $\mathbb{Q}$ already contains an inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ corresponding to every edge $G - P$ in $\mathbb{P}_k$, since the edge also exists in $\mathbb{Q}$. If there is an inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{Q}$ corresponding to every edge $G \leftarrow P$ or $G \circ\!\!- P$ in $\mathbb{P}_k$, it turns out that $\mathbb{Q}$ also contains an inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ corresponding to each genotype–

genotype link in $\mathbb{P}_k$. Thus, owing to Theorem 2.4, if $\mathbb{N}$ is a member of the class of plausible conditional genetic causal MAGs represented by $\mathbb{Q}$, then the marginal $\mathbb{N}_k$ of $\mathbb{N}$ over $\mathbf{G}_k \cup \{P\}$ has the same skeleton as $\mathbb{M}_k$. In addition, $\mathbb{N}_k$ provably has the same orientations as $\mathbb{M}_k$ at unshielded genotypes adjacent to P. Theorem 3.9 therefore implies that $\mathbb{Q}$ is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$; thus, it is added to the output of the algorithm.

Correctness of Algorithm 15 follows directly from the two theorems below.

Theorem 4.1 (Soundness of Algorithm 15)**.** *In the output of Algorithm 15,* $\mathbf{Q}$ *is a set of maximally-informative plausible conditional genetic PAG over* $\mathbf{O}$ *that are consistent with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$.

Theorem 4.2 (Completeness of Algorithm 15)**.** *In the output of Algorithm 15,* $\mathbf{Q}$ *is a superset of the set of maximally-informative plausible conditional genetic PAG over* $\mathbf{O}$ *that are consistent with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$.

As a specialised algorithm, Algorithm 15 is expected to be faster than IOD and cSAT+/COmbINE. A performance comparison, however, is not attempted here. The reason for developing the algorithm was to get useful insights for devising the local-learning algorithm of the next section.

4.2 Learning consistent genotype–phenotype relationships

It is not of much interest to learn whole consistent plausible conditional genetic causal PAGs; only the consistent genotype–phenotype relationships can be learned instead. As discussed above, $G - P$ edges in $\mathbb{L}$ at line 25 of Algorithm 15 are fixed: they exist in every consistent plausible conditional genetic causal PAG. As it turns out (see proof of Theorem 4.3), "removable" $G \leftarrow P$ and $G \circ\!\!- P$ edges in $\mathbb{L}$ at line 25 (that is, edges between genotype–phenotype pairs in $\mathbf{R}_1$ constructed at line 42) are truly removable in the sense that, for each such edge, some consistent plausible conditional genetic causal PAGs contain the edge while the others do not. Moreover, fixed $G \leftarrow P$ and $G \circ\!\!- P$ edges in $\mathbb{L}$ at line 25 (that is, edges between

Algorithm 15 Learn consistent plausible conditional genetic causal PAGs — part 1 out of 6. $\mathbf{G}$ is a set of genotypes and P is a phenotype, $\mathbf{O} = \mathbf{G} \cup \{P\}$, and $\mathbb{M}$ is a plausible conditional genetic causal MAG defined over $\mathbf{O}$. $\mathbf{G}_1, \ldots, \mathbf{G}_n \subseteq \mathbf{G}$ $(n \geq 1)$. $\mathbf{O}_k = \mathbf{G}_k \cup \{P\}$, $\mathbb{M}_k$ is the marginal of $\mathbb{M}$ over $\mathbf{G}_k \cup \{P\}$ and $\mathbf{I}(\mathbb{M}_k)$ is the set of m-separations in $\mathbb{M}_k$ $(1 \leq k \leq n)$. In the output, $\mathbf{Q}$ is the set of maximally-informative plausible conditional genetic PAGs over $\mathbf{O}$ that are consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. $\mathrm{Adj}_{\mathbb{G}}(X,Y)$ denotes that nodes X and Y are adjacent in graph $\mathbb{G}$. $\mathbf{AD}_{\mathbb{G}}(X)$ is the set of nodes adjacent to node X in graph $\mathbb{G}$. $\mathrm{chr}(G)$ is the chromosome of genotype G.

Input: $\mathbf{I}(\mathbb{M}_k), \ldots, \mathbf{I}(\mathbb{M}_k)$
Output: $\mathbf{Q}$

for each $1 \leq k \leq n$ **do**
    let $\mathbb{P}_k$ be the output of Algorithm 11 applied to the output of Algorithm 10 with $\mathbf{I}(\mathbb{M}_k)$ as input
end for
let $\mathbb{L}$ be the empty partially-oriented mixed graph over $\mathbf{O}$
for each $\{X,Y\} \subseteq \mathbf{O}$ s.t. $\forall\, 1 \leq k \leq n : \{X,Y\} \subseteq \mathbf{O}_k \implies \mathrm{Adj}_{\mathbb{P}_k}(X,Y)$ **do**
    add edge $X \circ\!\!-\!\!\circ Y$ to $\mathbb{L}$
end for
for each $G \in \mathbf{AD}_{\mathbb{L}}(P)$ **do**
    ▷ Apply orientation Rule (1)
    orient $G *\!\!-\!\!\circ P$ as $G *\!\!- P$
    ▷ Transfer orientations from $\mathbb{P}_1, \ldots, \mathbb{P}_n$
    if $\exists k$ s.t. $G \in \mathbf{G}_k$ and $G -\!\!* P$ in $\mathbb{P}_k$ **then**
        for each edge $G \circ\!\!-\!\!* X$ in $\mathbb{L}$ **do**
            orient $G \circ\!\!-\!\!* X$ as $G -\!\!* X$
        end for
    else if $\exists k$ s.t. $G \in \mathbf{G}_k$ and $G \leftarrow\!\!* P$ in $\mathbb{P}_k$ **then**
        for each edge $G \circ\!\!-\!\!* X$ in $\mathbb{L}$ **do**
            orient $G \circ\!\!-\!\!* X$ as $G \leftarrow\!\!* X$
        end for
    end if
end for
▷ Apply orientation Rule (3)
for each edge $G_1 \circ\!\!-\!\!* G_2$ in $\mathbb{L}$ s.t. $\neg\mathrm{Adj}_{\mathbb{L}}(G_1, P)$ **do**
    orient $G_1 \circ\!\!-\!\!* G_2$ as $G_1 \leftarrow\!\!* G_2$
end for

Algorithm 15 Learn consistent plausible conditional genetic causal PAGs — part 2 out of 6. $\text{chr}(G)$ is the chromosome of genotype G. IP stands for inducing path.

▷ Prevent violations of Condition (6) of Theorem 3.10
for each edge $G_1 \ast\!\!-\!\!\ast\, G_2$ in $\mathbb{L}$ **do**
  if $\{G_1, G_2\} \not\subseteq \mathbf{AD}_{\mathbb{L}}(P)$ and $\text{chr}(G_1) \neq \text{chr}(G_2)$ **then**
    remove $G_1 \ast\!\!-\!\!\ast\, G_2$ from $\mathbb{L}$
  end if
end for
▷ Prevent IPs corresponding to genotype–phenotype non-adjacencies
for each edge $G_1 \ast\!\!-\, G_2$ in $\mathbb{L}$ s.t. $G_1 \notin \mathbf{AD}_{\mathbb{L}}(P)$ **do**
  if $\exists k$ s.t. $G_1 \in \mathbf{G}_k$, $G_2 \notin \mathbf{G}_k$, and $G_1 \notin \mathbf{AD}_{\mathbb{P}_k}(P)$ **then**
    remove $G_1 \ast\!\!-\, G_2$ from $\mathbb{L}$
  end if
end for

Algorithm 15 Learn consistent conditional genetic causal PAGs — part 3 out of 6. $\mathbf{NE}_{\mathbb{G}}(P)$, $\mathbf{CH}_{\mathbb{G}}(P)$, and $\mathbf{PNE}_{\mathbb{G}}(P)$ is the set of neighbours, children, and potential neighbours, respectively, of node X in partially-oriented mixed graph $\mathbb{G}$. $\mathbf{chrENE}^{G}_{\mathbb{L}}(P)$ is the set of neighbours and potential neighbours (extended neighbours) of P that are on the same chromosome as G in $\mathbb{L}$.

$\mathbf{R}_1 \leftarrow \emptyset$
▷ Find removable genotype–phenotype links
for each $G \in \mathbf{CH}_{\mathbb{L}}(P) \cup \mathbf{PNE}_{\mathbb{L}}(P)$ **do**
  if $\forall\, 1 \leq k \leq n : \{G\} \cup \mathbf{chrENE}^{G}_{\mathbb{L}}(P) \not\subseteq \mathbf{G}_k$ **then**
    add $\{G, P\}$ to $\mathbf{R}_1$
  end if
end for
▷ Find removable genotype–genotype links
$\mathbf{R}_2 \leftarrow \emptyset$
for each $\{G_1, G_2\} \subseteq \mathbf{AD}_{\mathbb{L}}(P)$ s.t. $\{G_1, P\} \subseteq \mathbf{R}_1$ or $\{G_2, P\} \subseteq \mathbf{R}_1$ **do**
  add $\{G_1, G_2\}$ to $\mathbf{R}_2$
end for
for each $\{G_1, G_2\} \subseteq \mathbf{G}$ s.t. $\{G_1, G_2\} \not\subseteq \mathbf{AD}_{\mathbb{L}}(P)$ and $\text{Adj}_{\mathbb{L}}(G_1, G_2)$ **do**
  if $\forall\, 1 \leq k \leq n : \{G_1, G_2\} \not\subseteq \mathbf{G}_k$ or $\{G_1, G_2\} \subseteq \mathbf{AD}_{\mathbb{P}_k}(P)$ **then**
    add $\{G_1, G_2\}$ to $\mathbf{R}_2$
  end if
end for

Algorithm 15 Learn consistent conditional genetic causal PAGs — part 4 out of 6.

55: **for each** $\mathbf{E_1} \in 2^{\mathbf{R}_1}$ **do**
56: $\quad\mathbf{E_3} \leftarrow \{\{G_1, G_2\} \subseteq \mathbf{AD}_{\mathbb{L}}(P)$ s.t. $\{G_1, P\} \notin \mathbf{E_1}$ and $\{G_2, P\} \notin \mathbf{E_1}\}$
57: $\quad$**for each** $\mathbf{E_2} \in 2^{\mathbf{R}_2 \setminus \mathbf{E_3}}$ **do**
58: $\quad\quad$Let $\mathbb{S}$ be the subgraph of $\mathbb{L}$ without edges between pairs in $\mathbf{E_1} \cup \mathbf{E_2}$
59: $\quad\quad flag \leftarrow true$
60: $\quad\quad\triangleright$ Detect violations of Condition (5) of Theorem 3.10
61: $\quad\quad$**for each** edge $G_3 \in \mathbf{NE}_{\mathbb{S}}(P)$ **do**
62: $\quad\quad\quad$**if** $\exists \{G_1, G_2\} \subseteq (\mathbf{G} \setminus \mathbf{AD}_{\mathbb{S}}(P)) \cap \mathbf{AD}_{\mathbb{S}}(G_3)$ s.t. $\neg \mathrm{Adj}_{\mathbb{S}}(G_1, G_2)$ **then**
63: $\quad\quad\quad\quad flag \leftarrow false$
64: $\quad\quad\quad\quad$**break**
65: $\quad\quad\quad$**end if**
66: $\quad\quad$**end for**
67: $\quad\quad$**if** $\neg flag$ **then**
68: $\quad\quad\quad$**continue**
69: $\quad\quad$**end if**
70: $\quad\quad\triangleright$ Detect violations of Condition (6) of Theorem 3.10
71: $\quad\quad$**for each** edge $G_1 \ast\!\!-\!\!\ast\, G_2$ in $\mathbb{S}$ **do**
72: $\quad\quad\quad$**if** $\{G_1, G_2\} \not\subseteq \mathbf{AD}_{\mathbb{S}}(P)$ and $\mathrm{chr}(G_1) \neq \mathrm{chr}(G_2)$ **then**
73: $\quad\quad\quad\quad flag \leftarrow false$
74: $\quad\quad\quad\quad$**break**
75: $\quad\quad\quad$**end if**
76: $\quad\quad$**end for**
77: $\quad\quad$**if** $\neg flag$ **then**
78: $\quad\quad\quad$**continue**
79: $\quad\quad$**end if**
80: $\quad\quad\triangleright$ Detect violations of Condition (7) of Theorem 3.10
81: $\quad\quad$**if** $\exists$ shielded $G \in \mathbf{NE}_{\mathbb{S}}(P) \cup \mathbf{SP}_{\mathbb{S}}(P)$ **then**
82: $\quad\quad\quad$**continue**
83: $\quad\quad$**end if**

genotype–phenotype pairs not in $\mathbf{R}_1$) are truly fixed in the sense that they exist in every consistent plausible conditional genetic causal PAG. Algorithm 16 is based on these results and outputs the *fixed neighbours*, *fixed children*, *fixed potential neighbours*, *removable children*, and *removable potential neighbours* of the phenotype in a plausible conditional genetic causal MAG. The *fixed neighbours*, *fixed children*, and *fixed potential neighbours* of the phenotype are neighbours, children, and potential neighbours, respectively, in every consistent plausible conditional genetic causal PAG. The *removable children* and *removable potential neighbours* of the phenotype are children and potential neighbours, respectively, in some but not all

Algorithm 15 Learn consistent conditional genetic causal PAGs — part 5 out of 6.

▷ Apply orientation Rule (3) to satisfy Condition (3) of Theorem 3.10
for each edge $G_1 \circ\!\!-\!\!* G_2$ in $\mathbb{S}$ s.t. $G_1 \notin \mathbf{AD}_{\mathbb{S}}(P)$ **do**
  orient $G_1 \circ\!\!-\!\!* G_2$ as $G_1 \leftarrow\!\!* G_2$
end for
▷ Prevent IPs corresponding to genotype–phenotype non-adjacencies
for each edge $G_1 *\!\!-\!\!\circ G_2$ in $\mathbb{S}$ s.t. $G_1 \notin \mathbf{AD}_{\mathbb{L}}(P)$ **do**
  if $\exists k$ s.t. $G_1 \in \mathbf{G}_k$, $G_2 \notin \mathbf{G}_k$, and $G_1 \notin \mathbf{AD}_{\mathbb{P}_k}(P)$ **then**
    for each edge $G_2 \circ\!\!-\!\!* X$ in $\mathbb{S}$ **do**
      orient $G_2 \circ\!\!-\!\!* X$ as $G_2 \leftarrow\!\!* X$
    end for
  end if
end for
▷ Prevent violations of Condition (5) of Theorem 3.10
for each edge $G_3 \in \mathbf{PNE}_{\mathbb{S}}(P)$ **do**
  if $\exists\{G_1, G_2\} \subseteq (\mathbf{G} \setminus \mathbf{AD}_{\mathbb{S}}(P)) \cap \mathbf{AD}_{\mathbb{S}}(G_3)$ s.t. $\neg\mathrm{Adj}_{\mathbb{S}}(G_1, G_2)$ **then**
    for each edge $G_3 \circ\!\!-\!\!* X$ in $\mathbb{S}$ **do**
      orient $G_3 \circ\!\!-\!\!* X$ as $G_3 \leftarrow\!\!* X$
    end for
  end if
end for

consistent plausible conditional genetic causal PAGs.

Theorem 4.3 (Correctness of Algorithm 16)**.** *Let* $\mathbb{M}$ *be a plausible conditional genetic causal MAG defined over a set of variables* $\mathbf{G} \cup \{P\}$, $\mathbf{G}_1, \ldots, \mathbf{G}_n \subseteq \mathbf{G}$ $(n \geq 1)$, *and* $\mathbb{M}_k$ *is the marginal of* $\mathbb{M}$ *over* $\mathbf{G}_k \cup \{P\}$. *If* $\mathbf{I}(\mathbb{M}_k), \ldots, \mathbf{I}(\mathbb{M}_k)$ *is the input of Algorithm 16, then in the output,* $\mathbf{fNE}(P)$, $\mathbf{fCH}(P)$, $\mathbf{fPNE}(P)$, $\mathbf{rCH}(P)$, *and* $\mathbf{rPNE}(P)$ *are the fixed neighbours, fixed children, fixed potential neighbours, removable children, and removable potential neighbours, respectively, of P in the conditional genetic causal PAGs consistent with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$.

Algorithm 16, as Algorithm 15, accepts a set of sets of m-separations that are assumed to be the sets of m-separations in marginals of a plausible conditional genetic causal MAG. In contrast, Algorithm 17 directly accepts a set of conditional genetic random samples and controls the FDR of the genotypes whose adjacency with the phenotype is consistent with the samples. Compared to learning from single genetic datasets, it is even less clear how to assess the confidence on the orientations. Learning the orientations is, therefore, not attempted by Algorithm 17.

Algorithm 15 Learn consistent conditional genetic causal PAGs — part 6 out of 6.

```
104:          U ← {unshielded G ∈ PNE_𝕊(P)}
105:          for each W ∈ 2^U do
106:              ℚ ← 𝕊
107:              for each edge G ∘—∗ X in ℚ s.t. G ∈ W do
108:                  orient G ∘—∗ X as G —∗ X
109:              end for
110:              for each edge G ∘—∗ X in ℚ s.t. G ∈ U \ W do
111:                  orient G ∘—∗ X as G ←∗ X
112:              end for
113:              ▷ Ensure there are IPs corresponding to G ← P and G ∘— P edges
114:              flag ← true
115:              for each 1 ≤ k ≤ n do
116:                  for each G_1 ∈ CH_{ℙ_k}(P) ∪ PNE_{ℙ_k}(P) do
117:                      if G_1 ∉ AD_Q(P) and ∄ G_1 ∗— G_2 s.t. G_2 ∉ G_k in ℚ then
118:                          flag ← false
119:                          break
120:                      end if
121:                  end for
122:                  if ¬flag then
123:                      break
124:                  end if
125:              end for
126:              if flag then
127:                  Q ← Q ∪ ℚ
128:              end if
129:          end for
130:      end for
131:  end for
```

Since Algorithm 17 uses the meta-analytic test in Algorithm 7, there is a potential for increased power.

Theorem 4.4 (Correctness of Algorithm 17)**.** *Let $\mathbb{G}$ be a genetic causal DAG with selection nodes defined over a set of variables $\mathbf{V} = \mathbf{G} \cup \{P\} \dot{\cup} \mathbf{H} \dot{\cup} \mathbf{S}$, $\mathbb{M}$ be the marginal/conditional of $\mathbb{G}$ given $\mathbf{S}$, $\mathbf{G}_1, \ldots, \mathbf{G}_n \subseteq \mathbf{G}$ $(n \geq 1)$, $\mathbb{M}_k$ be the marginal of $\mathbb{M}$ over $\mathbf{G} \setminus \mathbf{G}_k$, $\mathscr{P}$ be the probability distribution of the variables in $\mathbf{V}$, $\mathscr{M}$ be the marginal/conditional of $\mathscr{P}$ over $\mathbf{G} \cup \{P\}$ given $\mathbf{S} = \mathbf{s}$, and $\mathscr{M}_k$ be the marginal of $\mathscr{M}$ over $\mathbf{G} \setminus \mathbf{G}_k$. Suppose that Algorithm 17 is applied to $D_1, \ldots, D_n$, where D_k is a sample from $\mathscr{M}_k$, with FDR threshold q. The FDR among the nodes in $\widehat{\mathbf{cAD}}(P)$ is not greater than q if the following conditions are satisfied:*

Algorithm 16 Learn the fixed neighbours, fixed children, fixed potential neighbours, removable children, and removable potential neighbours of the phenotype. $\mathbf{G}$ is a set of genotypes and P is a phenotype, and $\mathbb{M}$ is a plausible conditional genetic causal MAG defined over $\mathbf{G} \cup \{P\}$. $\mathbf{G}_1, \ldots, \mathbf{G}_n \subseteq \mathbf{G}$ $(n \geq 1)$. $\mathbb{M}_k$ is the marginal of $\mathbb{M}$ over $\mathbf{G}_k \cup \{P\}$ and $\mathbf{I}(\mathbb{M}_k)$ is the set of m-separations in $\mathbb{M}_k$ $(1 \leq k \leq n)$. $\mathbf{fNE}(P)$, $\mathbf{fCH}(P)$, $\mathbf{fPNE}(P)$, $\mathbf{rCH}(P)$, and $\mathbf{rPNE}(P)$ is the set of fixed neighbours, fixed children, fixed potential neighbours, removable children, and removable potential neighbours, respectively, of P in the conditional genetic causal PAGs consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. $\mathbf{G}^k$ is the subset of $\mathbf{G}$ on the k-th chromosome.

Input: $\mathbf{I}(\mathbb{M}_1), \ldots, \mathbf{I}(\mathbb{M}_n)$
Output: $\mathbf{fNE}(P)$, $\mathbf{fCH}(P)$, $\mathbf{fPNE}(P)$, $\mathbf{rCH}(P)$, and $\mathbf{rPNE}(P)$

for each $1 \leq k \leq n$ **do**
    let $\mathbf{NE}_k(P)$, $\mathbf{CH}_k(P)$, and $\mathbf{PNE}_k(P)$ be the output of Algorithm 12 with $\mathbf{I}(\mathbb{M}_k)$ as input
end for
▷ Learn genotypes whose adjacency with P is consistent
$\mathbf{cAD}(P) \leftarrow \emptyset$
for each $G \in \mathbf{G}$ **do**
    if $\forall\, 1 \leq k \leq n : G \in \mathbf{G}_k \implies G \in \mathbf{NE}_k(P) \cup \mathbf{CH}_k(P) \cup \mathbf{PNE}_k(P)$ **then**
        $\mathbf{cAD}(P) \leftarrow \mathbf{cAD}(P) \cup \{G\}$
    end if
end for
▷ Learn fixed neighbours of P; these are all the consistent neighbours of P
$\mathbf{fNE}(P) \leftarrow \bigcup_{1 \leq k \leq n} \mathbf{NE}_k(P)$
▷ Get consistent children of P
$\mathbf{cCH}(P) \leftarrow \mathbf{cAD}(P) \cap \bigcup_{1 \leq k \leq n} \mathbf{CH}_k(P)$
▷ Get consistent potential neighbours of P
$\mathbf{cPNE}(P) \leftarrow \mathbf{cAD}(P) \setminus \mathbf{fNE}(P) \setminus \mathbf{cCH}(P)$
▷ Get consistent extended neighbours (neighbours and potential neighbours) of P
$\mathbf{cENE}(P) \leftarrow \mathbf{fNE}(P) \cup \mathbf{cPNE}(P)$
▷ Differentiate between fixed and removable children of P
$\mathbf{rCH}(P) \leftarrow \{G \in \mathbf{cCH}(P) \text{ s.t. } \forall\, 1 \leq k \leq n : (\{G\} \cup \mathbf{cENE}(P)) \cap \mathbf{G}^{\mathrm{chr}(G)} \nsubseteq \mathbf{G}_k\}$
$\mathbf{fCH}(P) \leftarrow \mathbf{cCH}(P) \setminus \mathbf{rCH}(P)$
▷ Differentiate between fixed and removable potential neighbours of P
$\mathbf{rPNE}(P) \leftarrow \{G \in \mathbf{cPNE}(P) \text{ s.t. } \forall\, 1 \leq k \leq n : (\{G\} \cup \mathbf{cENE}(P)) \cap \mathbf{G}^{\mathrm{chr}(G)} \nsubseteq \mathbf{G}_k\}$
$\mathbf{fPNE}(P) \leftarrow \mathbf{cPNE}(P) \setminus \mathbf{rPNE}(P)$

Algorithm 17 Estimate the genotype–phenotype links consistent with a set of conditional genetic random samples with overlapping sets of variants. $\mathbf{G}$ is a set of genotypes and P is a phenotype. $\mathbf{G}_1, \ldots, \mathbf{G}_n \subseteq \mathbf{G}$ $(n \geq 1)$. D_i is a conditional random sample defined over $\mathbf{G}_i \cup \{P\}$. $0 < \alpha < 1$ is the significance level for the hypothesis tests of conditional independence performed by the algorithm. $0 < q < 1$ is an FDR threshold. $\widehat{\mathbf{cAD}}(P)$ is the estimated set of genotypes whose adjacency with P in the true causal MAG over $\mathbf{G} \cup \{P\}$ is consistent with $D_1, \ldots, D_n$. m is the number of chromosomes. $\mathbf{G}_i^j$ is the subset of $\mathbf{G}_i$ on the j-th chromosome. $\mathbf{I}(D)$ is the set of m-separations among the variables in dataset D as determined by Algorithm 7 applied to D with significance level α. $\mathbf{I}(D_i)\lfloor_{\mathbf{G}_i^j \cup \{P\}}$ is the subset of $\mathbf{I}(D_i)$ over $\mathbf{G}_i^j \cup \{P\}$.

Input: $D_1, \ldots, D_n$, α, q
Output: $\widehat{\mathbf{cAD}}(P)$

for each $1 \leq i \leq n$ **do**
  for each $1 \leq j \leq m$ **do**
    let $\mathbf{TA}_i^j(P)$ be the first output of Algorithm 3 with $\mathbf{I}(D_i)\lfloor_{\mathbf{G}_i^j \cup \{P\}}$ and P as input
  end for
  $\mathbf{TA}_i(P) = \bigcup_{1 \leq j \leq m} \mathbf{TA}_i^j(P)$
end for
$\widehat{\mathbf{cAD}}(P) \leftarrow \emptyset$
for each $G \in \mathbf{G}$ **do**
  if $\forall\, 1 \leq i \leq n : G \in \mathbf{G}_i \implies G \in \mathbf{TA}_i(P)$ **then**
    $\widehat{\mathbf{cAD}}(P) \leftarrow \widehat{\mathbf{cAD}}(P) \cup \{G\}$
  end if
end for
apply an appropriate FDR-controlling procedure to the maximal conditional-independence p-values corresponding to the genotypes in $\widehat{\mathbf{cAD}}(P)$ to control the FDR of the genotypes at q

1. *$\mathbb{G}$ and $\mathscr{M}$ satisfy the selection bias causal assumption.*

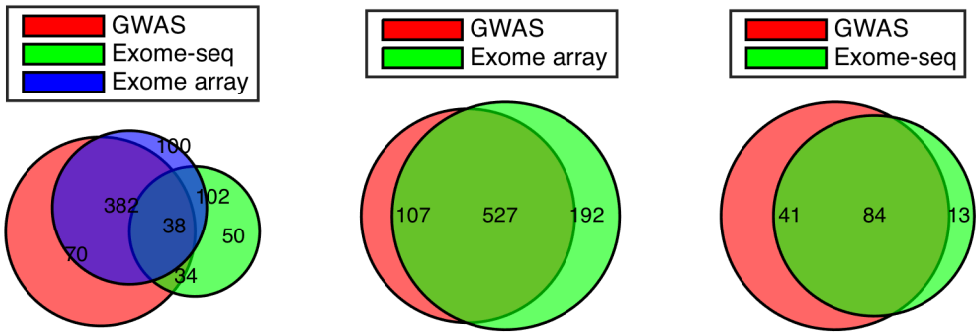
2. *All tests considered by the algorithm are reliable.*

3. *Performed tests never produce a type II error.*

4.3 Application to prion disease

Algorithm 17 cannot be directly applied to any combination of the datasets in Table 3.2, because Fisher's method (used in Algorithm 7) assumes that the datasets are independent and the datasets in Table 3.2 of each type from a certain population

and disease happen to share cases. Therefore, duplicate cases across the datasets must be removed prior to causal discovery. Figure 4.1 visualises the overlap of cases between datasets of each type in Table 3.2 for each population and disease. There is no value in co-analysing German sCJD or UK vCJD datasets since the overlap between cases is quite large. Furthermore, it is preferable to co-analyse the UK sCJD GWAS dataset with the UK sCJD exome-array dataset over the UK sCJD exome-sequencing dataset, since the former contains 202 unique cases and the latter contains 152. Thus, Algorithm 17 was applied to the UK sCJD GWAS, German sCJD GWAS, UK sCJD exome-array, and US sCJD exome-array datasets after removing cases from the UK sCJD exome-array dataset that are also present in the UK sCJD GWAS dataset. The parameters used were the same as in Section 3.5. The list of discoveries can be found in Table 4.1.



(a) UK sCJD **(b)** DE sCJD **(c)** UK vCJD

Figure 4.1: Visualisation of the overlap of cases between datasets of each type in Table 3.2 for each population and disease. Exome-seq stands for exome sequencing.

Causal meta-analysis of the datasets resulted in one less discovery than causal discovery from the single datasets and in no novel discoveries. All discoveries were somewhat expected: rs114501427, rs41311333, rs45458098, and rs151199705 were discovered from the US exome-array dataset by Algorithm 13 and are absent from the UK exome-array dataset and the GWAS datasets. Therefore, tests involving these SNPs were only performed on the US exome-array dataset by Algorithm 7. Similarly, rs142631461 was discovered from both exome-array datasets by Algorithm 13 and is absent from the GWAS datasets. Finally,

rs1799990 was discovered from every GWAS dataset by Algorithm 13 and is absent from the exome-array datasets. rs144218313 was discovered from the US GWAS dataset by Algorithm 13 but was not discovered in the meta analysis. This is because rs144218313 was rendered independent from the disease by the set $\{\text{rs78441178}, \text{rs148843120}\}$. rs78441178 is present in both exome-array datasets but rs148843120 is present in the UK dataset only. Therefore, the test with conditioning set $\{\text{rs78441178}, \text{rs148843120}\}$ could not be performed by Algorithm 13 on the US exome-array dataset and render rs144218313 independent from the disease.

RS#	C#	Position
rs114501427	2	138568946
rs41311333	5	90619108
rs45458098	6	34858817
rs151199705	14	105392632
rs142631461	17	29249184
rs1799990	20	4699605

Table 4.1: Discoveries resulting from the application of Algorithm 17 to the set of datasets in sCJD described in the text. *RS#* is the RS number of the variant in dbSNP. *C#* is the chromosome number.

4.4 Summary and future work

The main contribution of this chapter is an algorithm that learns the FDR-controlled genotype–phenotype links that are consistent with overlapping conditional genetic samples (Algorithm 17). The algorithm has a potential for increased power due to its use of a test that conducts meta-analysis of the samples (Algorithm 7). However, when applied to a combination of datasets from prion disease, the algorithm did not result in novel discoveries. As in case of single datasets, the algorithms developed here may be applied, with minor modifications, to other types of cross-sectional datasets as well (e.g. gene-expression datasets).

It would be of interest to conduct a simulation study of the performance of Algorithm 17 like the one conducted for Algorithm 13. For such a study, a Bayesian network would have to be learned from either a whole-genome dataset or a dataset resulting from the concatenation of a GWAS and an exome-sequencing dataset from

the same individuals. Sampling over tag SNPs would then be performed to generate simulated GWAS datasets and over exonic SNPs to generate simulated exome-sequencing and exome-array datasets.

It would also be compelling to find out how applying Algorithm 17 to a combination of datasets compares to applying Algorithm 13 to a dataset resulting from the concatenation of the datasets, followed by genotype imputation. Although the output of the latter approach is easier to interpret, it might be erroneous due to errors in imputation. Accuracy might be even lower when combining exome-sequencing and exome-array datasets, as imputation of rare variants is problematic [Evangelou and Ioannidis, 2013]. Applying Algorithm 17 is the only choice between the two approaches when no reference dataset for genotype imputation is available.

Chapter 5

Variant filtering using causal prior knowledge

The lack of power in association analysis or causal discovery could be overcome by easing the multiple-testing burden through decreasing the number of hypotheses. In genetic studies, this corresponds to performing *variant filtering*. Variant filtering also decreases the execution time of an analysis and may result in previously intractable analyses becoming tractable [e.g. searching for N-locus epistasis; Ritchie, 2011]. One approach to performing variant filtering is to use prior knowledge. Methods that use prior knowledge are biased towards well-studied genes and are dependent on the quality of the knowledge; prior knowledge can assist, however, with the biological interpretation of the results [Ritchie, 2011]. In this chapter, a variant-filtering method inspired by INCA is presented. Publicly available biological data and prior knowledge are integrated into a special type of directed graph. The nodes of the graph comprise a phenotype and molecular entities in the cell and are associated with genomic regions, while the edges denote causation. Candidate causal variants are identified from the ancestors (causes) of the phenotype.

5.1 Method

In this section, graphs for representing causal prior knowledge are discussed and their use in variant filtering is described.

5.1.1 Causal prior knowledge graphs

Following Borboudakis and Tsamardinos [2012], it is assumed that publicly-available biological data and prior knowledge can be expressed as a set of statements of the form "X is a cause of Y" or "X is not a cause of Y". The causal prior knowledge can be represented by a directed graph which is called *causal prior knowledge graph (CPKG)* in this work. The graph has two types of edges. Edge $X \to Y$ is said to be *positive* and denotes that there is at least one piece of evidence that X is a cause of Y; edge $X \not\to Y$ is called *negative* and denotes that there is at least one piece of evidence that X is not a cause of Y. Each pair of nodes X and Y can be connected with at most two edges ($X \to Y$ or $X \not\to Y$ and $X \leftarrow Y$ or $X \not\leftarrow Y$). Directed cycles (defined as in directed graphs) are allowed. The graph is accompanied by a set of references (bibliography). A reference is typically comprised of the name of a biological database and the identifier of an entry in the database. Each edge is associated with a nonempty set of citations from the bibliography, so that the user of the graph can track the evidence supporting the edge.

5.1.2 CPKGs for variant filtering

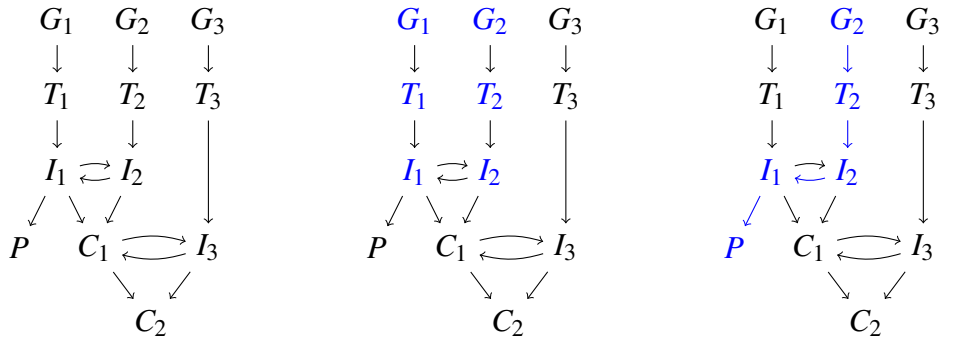
A CPKG used for variant filtering is study-specific and represents known causal relationships between the phenotype P and the levels of molecular entities in a cell of the cell type of interest in the population of interest. For example, in a case–control disease study, the phenotype is the affected/unaffected status, the cell type of interest is an affected cell type, and the population of interest is the union of the population of healthy individuals and the individuals affected by the disease. In the following, both a molecular entity and its level in the cell are referred to by the name of the entity. In the case of a gene, its "level" is the number of copies of the gene.

A CPKG for variant filtering only contains positive edges. Every node in the CPKG is associated with a set of genomic regions, which are then used to filter the dataset(s) at hand. For example, a protein node is associated with the coding regions of the corresponding transcript, a transcript node with the exons the transcript comprises, and a gene node with the gene. Entities that are not genes or gene products (e.g. the phenotype) are associated with an empty set of regions. Variants within the

regions associated with a node are considered to be *candidate causes of the effects* of the node. For example, a variant within a gene is a candidate cause of the level of the transcripts of the gene; for instance, the variant may disrupt the transcription of the gene. Once a CPKG is built, the ancestors of P are identified. Assuming causal transitivity, these are the known causes of P. Variants that fall within the associated regions form the set of candidate causal variants for P and are the ones included in the filtered dataset. In the end, the CPKG can aid the biological interpretation of any discovery made from the filtered dataset. For every such discovery, *every directed (causal) path from an associated node to the phenotype node suggests a possible causal biological mechanism*.

Figure 5.1 illustrates an example of using a CPKG for variant filtering. Figure 5.1a shows an example of a CPKG for variant filtering. G_1, G_2, and G_3 are genes, T_1, T_2, and T_3 are transcripts of the genes, I_1, I_2, and I_3 are protein isoforms translated from the transcripts, C_1 and C_2 are protein complexes, and P is the phenotype. The ancestors of P are highlighted in Figure 5.1b. The regions associated with them are used to filter a genetic dataset with P. Suppose that a variant in G_2 is discovered from the filtered dataset. Figure 5.1c shows a causal path (the only one in this example) from G_2 to P. A potential causal mechanism is that the variant disrupts the transcription of G_2, resulting in the absence of T_2 and I_2. Subsequently, abnormally levels of I_1, which no longer interacts with I_2 to produce C_1, affect P.

If multiple cell types are relevant to the phenotype (e.g. both neurons and glial cells are affected in prion disease), a separate CPKG must be built for each cell type, and the causes of P must be found separately from each graph. The associated regions can be subsequently merged and used for filtering. For each discovery, causal paths to the phenotype have to be found separately in each CPKG. Having separate graphs is necessary in order to avoid violations of causal transitivity. For example, suppose that X is a cause of Y in cell type A and Y is a cause of Z in cell type B; then X is not necessarily a cause of Y in cell types A and B. This is, of course, true for different conditions in general. If X is a cause of Y in controls and Y is a cause of Z in cases in a case–control study, X is not necessarily a cause of Y



(a) A CPKG for variant filtering.

(b) The CPKG with ancestors of P highlighted.

(c) The CPKG with a causal path from G_2 to P highlighted.

Figure 5.1: Example of variant filtering using a CPKG. G_1, G_2, and G_3 are genes, T_1, T_2, and T_3 are transcripts of the genes, I_1, I_2, and I_3 are protein isoforms translated from the transcripts, C_1 and C_2 are protein complexes, and P is the phenotype.

in the general population. It must be assumed that this does not occur if the method is to be used at all for filtering case–control datasets.

If the causal variants of interest are not all included in the dataset to be filtered, the filtering regions should be expanded to include variants that may be not causal themselves but are in LD with causal variants inside the unexpanded regions but not in the dataset. This is needed in order to capture the latter variants. For example, if the available dataset originates from a GWAS and the causal variants of interest are the common SNPs, the filtering regions should be expanded so that untyped common SNPs inside the unexpanded regions is captured by typed common SNPs outside of them. If however, the dataset originates from exome sequencing and the causal variants of interest are exonic SNPs and short indels, the filtering regions should be left intact; this is because the dataset contains, in principle, all exonic SNPs and short indels.

Variant filtering using a directed graph such as a CPKG, should result in higher specificity than using an undirected one [e.g. BioGraph; Liekens et al., 2011]. For example, in directed graph $P \leftarrow P_1 \rightarrow P_2$, only P_1 is an ancestor of P; P and P_2 only share a common ancestor. In contrast, in the corresponding undirected graph $P - P_1 - P_2$, both P_1 and P_2 are (directly or indirectly) connected to P. In addition, a CPKG is, arguably, more helpful with the interpretation of the results than a non-

causal graph.

5.2 Implementation

In this section, the implementation of CPKG-based variant-filtering for prion disease is discussed in detail. For each *conceptual* data source, *real* data sources are used to build a CPKG. The CPKGs are then merged into a *general* CPKG. Post-processing results into a *post-processed* general CPKG. Finally, adaptation to prion disease results in a *prion-disease* CPKG to be used for variant filtering. Figure 5.2 illustrates the CPKG-building process, which was implemented in MATLAB. Among the data-integration methodologies identified by Lapatas et al. [2015], the methodology used to build the CPKG was data warehousing. On the downside, several biological databases (some of them quite big) had to be downloaded and installed locally; on the plus side, the analysis can be replicated at any time.

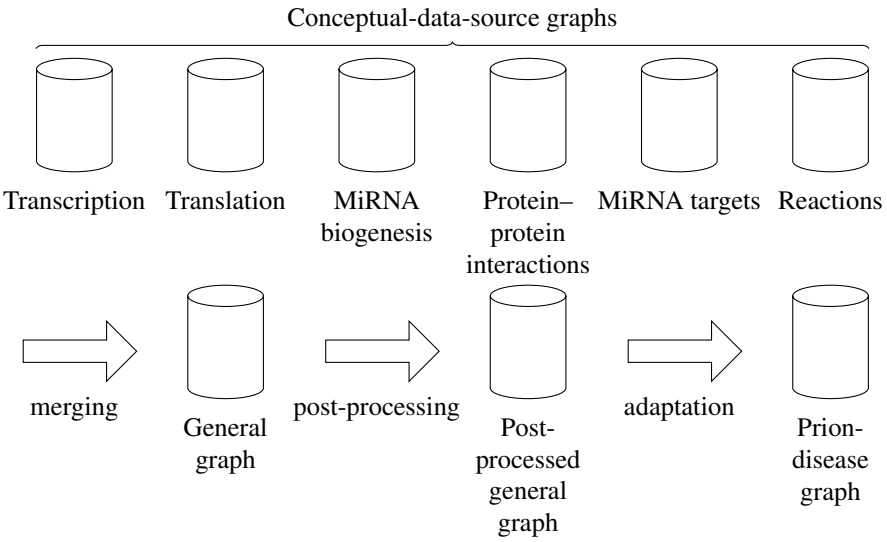


Figure 5.2: Illustration of the process of building a CPKG.

5.2.1 Conceptual data sources

A *conceptual data source* can be thought of as a table or view corresponding to a relationship between molecular entities or between molecular entities and phenotypes in an imaginary centralised biological relational database. Every molecular entity is associated with a species, a cell type, and a cellular compartment. Protein isoforms

are also associated with a set of post-translational modifications. Every relationship is associated with a certain condition. The condition may be, for example, the phase of the cell or the affected/unaffected status of the individual for some disease.

The conversion of a conceptual data source to a CPKG is described in Algorithm 18. For each conceptual data source, a CPKG is created that only contains the phenotype node. For each record of the source, every object in the represented relationship is first converted to a CPKG node. In the case of a molecular entity, the node is identified by the name of the external table and the key of the external record. In the case of a phenotype, the node is simply identified as the phenotype, since there is only one such node per variant-filtering CPKG. Subsequently, a reference is created from the key of the source record. Edges between the nodes are then created in a manner that depends on the data source. Every edge cites the newly created reference. Finally, the nodes, edges, reference, and citations created are then merged with the current CPKG.

The conceptual data sources used in this work and the conversion to causal knowledge is discussed below, where it is assumed that the data sources are already filtered so that they only contain relationships in humans between human molecular entities in the cell type of interest under the condition(s) of interest:

- *Transcription:* Each record denotes that gene G has transcript T. Either G is located in the nucleoplasm and T is located in the cytosol, or both G and T are located in the mitochondrial matrix. The corresponding causal knowledge is that G is a cause of T.

- *Translation:* Each record denotes that transcript T is translated to protein isoform I. Either both T and I are located in the cytosol, or both T and I are located in the mitochondrial matrix. Since I is newly translated, it is not post-translationally modified. The corresponding causal knowledge is that T is a cause of I.

- *MiRNA biogenesis:* MicroRNAs (miRNAs) are small, single-stranded non-coding RNAs that regulate gene expression in animals and plants by binding

Algorithm 18 Conversion of a conceptual data source to a CPKG. **R** is a set of records representing a conceptual data source. **V**, **E**, and **B** is the set of nodes, edges, and references, respectively, in the CPKG. **C** is a map from edges in **E** to sets of references in **B** representing the citations in the CPKG.

Input: **R**
Output: **V**, **E**, **B**, **C**
1: $\mathbf{V} \leftarrow \{P\}$
2: $\mathbf{E} \leftarrow \emptyset$
3: $\mathbf{B} \leftarrow \emptyset$
4: Initialise **C** with the empty map
5: **for each** $R \in \mathbf{R}$ **do**
6: $\quad \mathbf{V}_R \leftarrow \emptyset$
7: $\quad$ **for each** object X in R **do**
8: $\quad\quad$ Convert X to node V_x
9: $\quad\quad \mathbf{V}_R \leftarrow \mathbf{V}_R \cup \{V_x\}$
10: $\quad$ **end for**
11: $\quad$ Create reference B_R from R
12: $\quad$ Convert causal knowledge in R to set of edges $\mathbf{E}_R$ among the nodes in $\mathbf{V}_R$
13: $\quad \mathbf{V} \leftarrow \mathbf{V} \cup \mathbf{V}_R$
14: $\quad \mathbf{E} \leftarrow \mathbf{E} \cup \mathbf{E}_R$
15: $\quad \mathbf{B} \leftarrow \mathbf{B} \cup \{B_R\}$
16: $\quad$ **for each** E in $\mathbf{E}_R$ **do**
17: $\quad\quad$ **if** $\mathbf{C}(E)$ is undefined **then**
18: $\quad\quad\quad \mathbf{C}(E) \leftarrow \{B_R\}$
19: $\quad\quad$ **else**
20: $\quad\quad\quad \mathbf{C}(E) \leftarrow \mathbf{C}(E) \cup \{B_R\}$
21: $\quad\quad$ **end if**
22: $\quad$ **end for**
23: **end for**

at target sites of mRNAs, leading to mRNA cleavage or translational repression [Bartel, 2004]. The following discussion pertains to animals, where the target sites are located in the $3'$ *untranslated region (UTR)*. The primary transcript of a microRNA gene has a stem-loop ("hairpin") structure with long single-stranded tails. The tails are subsequently cleaved, resulting in the miRNA *precursor*. The precursor is then transported to the cytosol, where the loop is cleaved; the result is an RNA duplex termed the *miRNA-miRNA** duplex. One of the strands of each individual duplex ends up as the *mature* miRNA and is loaded into a *RNA-induced silencing complex (RISC)*, where it performs its regulatory functions on target mRNA. The other strand, called

*miRNA**, is degraded. It is also possible that both strands of the duplex end up as a mature miRNA: only one strand becomes the miRNA each time but with similar frequency [Bartel, 2004]. A class of miRNAs called *mirtrons* result from the processing of introns that have a structure similar to that of a miRNA precursor [Ruby et al., 2007].

Each record in the miRNA biogenesis data source denotes that transcript T of a miRNA gene is processed into miRNA MI. T and MI are both located in the cytosol. The corresponding causal knowledge is that T is a cause of MI.

- *Protein–protein interactions:* A *protein-protein interaction (PPI)* commonly refers to physical contact between proteins that occurs in vivo. Each record in the conceptual PPI database denotes that proteins (when the exact isoform is unknown) or protein isoforms $I_1, \ldots, I_n$ $(n \geq 1)$ participate in an interaction indicated in experiment E. A participant may be post-translationally modified. If E was performed in vitro and the proteins were purified from the contents of lysed cells, the subcellular location of the proteins is unknown and has to be somehow estimated; it is even possible that the interaction does not happen at all in vivo. Assuming that the proteins are localised, the conversion of a PPI to causal knowledge is dictated by E and n. PPI experiments can be classified into five classes according to their information content:

  1. *Experiments that show that two proteins bind to each other:* These experiments are performed in vitro and involve two pre-selected proteins I_1 and I_2 that are the only molecules in solution. Interaction is concluded because some measurement indicates that the proteins are bound. An example of such an experiment is *enzyme-linked immunosorbent assay (ELISA)* [Gan and Patel, 2013], where one protein is attached to the bottom of a well and a solution of the other protein is added, resulting in a colour change if the two proteins bind to each other. The corresponding causal knowledge is that I_1 a cause of I_2 and I_2 a cause of I_1.

  2. *Experiments that show that a bait protein binds to complexes composed*

of unknown subsets of prey proteins: These experiments are also performed in vitro and involve a pre-selected *bait* protein I_1 immobilised on a column and used to "fish" free *prey* proteins $\mathbf{P} = I_2, \ldots, I_n$ from a mixture, usually the contents of lysed cells. The prey proteins are then identified by methods such as mass spectroscopy. An example of such experiment is a *pull-down assay* [Vikis and Guan, 2004, Sambrook and Russell, 2006], where a *tagged* protein is first immobilised on a column coated with material that binds to the tag. These experiments cannot tell the components of the complexes formed by the prey proteins before binding to the bait protein. However, for each such complex C_i with unknown components $\mathbf{P_i} \subseteq \mathbf{P}$, every $P_{ij} \in \mathbf{P_i}$ is a cause of C_i, C_i is a cause of I_1, and I_1 is a cause of C_i. Therefore, each of $I_2, \ldots, I_n$ is a cause of I_1, and, if $n = 2$, then also I_1 is a cause of I_2.

3. *Experiments that show that a bait protein participates in complexes with unknown subsets of prey proteins:* The difference between this class and the previous class of experiments is that the bait protein I_1 is not immobilised on the column but forms complexes with other (prey) proteins $I_2, \ldots, I_n$ in the mixture, and the material on the column is used to "fish" the bait protein and its bound prey from the mixture. In *co-immunoprecipitation*, for example, the column is coated with an antibody specific for the bait [Phizicky and Fields, 1995]. The causal knowledge corresponding to such an experiment is that unknown subsets of $I_1, \ldots, I_n$ that include I_1 are causes of unknown complexes $C_1, \ldots, C_m$, which is not very informative. If $n = 2$, however, I_1 a cause of I_2 and I_2 a cause of I_1.

4. *Experiments that show that a set of proteins forms a complex:* These experiments usually conclude that proteins $I_1, \ldots, I_n$ form a complex C because they were detected in the same place after an attempt to separate the contents of a mixture. In *gel electrophoresis*, for example, an electric field is applied to a mixture of proteins in order to force them

to move through a gel [Wittig and Schägger, 2009]. Shorter molecules move faster, resulting in a set of discrete bands on the gel, each with different composition. If a set of proteins are identified in the same band, participation in the same complex is concluded, as the electric field did not manage to separate them. The causal knowledge corresponding to an experiment in this class is that each of $I_1, \ldots, I_n$ is a cause of C. The causal relationships between $I_1, \ldots, I_n$ are unknown because the steps of the complex formation are unknown. The former causal relationships are not of interest as an edge from every component of a complex node to the complex node is added automatically in a post-processing step of the CPKG (see section 5.2.10.1). However, I_1 a cause of I_2 and I_2 a cause of I_1 if $n = 2$.

5. *Experiments that show that two proteins participate in the same complex:* In *two-hybrid* systems, two pre-selected proteins I_1 and I_2 are expressed in cultured cells (usually yeast) [Legrain and Selig, 2000]. One protein is fused to the first of two domains of a transcription factor of a *reporter gene* and the other is joined with the second one. If expression of the reporter gene is detected, it must be because I_1 and I_2, and subsequently, the two domains of the transcription factor, came close and the transcription factor initiated transcription of the gene. It is therefore concluded that I_1 and I_2 interact. There may be, however, other proteins in the complex formed by I_1 and I_2. Thus, two-hybrid experiments only prove that two proteins participate in the same complex. The corresponding causal knowledge is that I_1 and I_2 are causes of some unknown complex C. As explained above, this knowledge is not useful here.

The three *modes of conversion to causal knowledge* encountered above are summarised as follows:

0. no causal relationship is entailed
1. $n = 2$ and I_1 a cause of I_2 and I_2 a cause of I_1

2. $n \geq 2$, I_1 is bait, $I_2, \ldots, I_n$ are prey, and each of $I_2, \ldots, I_n$ is a cause of I_1.

An experiment with non-zero mode of conversion is said to be *eligible* for conversion to causal knowledge.

- *MiRNA targets:* Each record is the interaction of miRNA *MI* with target gene *G* indicated in experiment *E*. The interaction is called a *miRNA-target interaction (MTI)*. The corresponding causal knowledge depends on *E*. Experiments for identifying MTIs can be classified in the following classes according to their information content:

  - *Experiments that show that a miRNA forms a complex with an mRNA:* In a *reporter assay*, a reporter gene G', which the fusion of the $3'$ UTR of a candidate target gene *G* and the gene of a fluorescent protein, is inserted into the genome of cultured cells [Thomson et al., 2011]. The level of a miRNA *MI* is manipulated in the cells, and the change in fluorescence is measured. A change in fluorescence indicates that *MI* affects the expression of G', implying that *MI* also affects the expression of *G*. Consequently, *MI* is a cause of *T*, for each transcript *T* of *G* (assuming that the $3'$ UTR of G' is found in all transcripts of *G*).

  - *Experiments that show that differential expression of a miRNA results in differential levels of an mRNA in RISC precipitates:* Co-immunoprecipitation can be used to "fish" RISC components and bound miRNA–mRNA pairs from the contents of lysed cells [Hendrickson et al., 2008]; for every candidate target gene, the corresponding mRNA levels can then be assessed in the precipitates using e.g. microarrays. If a certain miRNA *MI* is differentially expressed between two sets of cultured cells, the mRNA levels of its targets will be different in the RISC precipitates from each set of cells, implying that they are different in the cells. The corresponding causal knowledge is that *MI* is a cause of *T*, for each transcript *T* with differential levels in the RISC precipitates.

  - *Experiments that identify miRNA binding sites:* In techniques such as

HITS-CLIP, *PAR-CLIP*, and *CLASH*, co-immunoprecipitation of RISC components is also performed but subsequently treatments are applied that cut the parts of the mRNA that are not bound to miRNA and the remaining mRNA is sequenced [Thomson et al., 2011, Vlachos et al., 2015]; target genes have to be predicted from the sequenced miRNA binding sites using bioinformatics approaches. As predictions are not of interest here, these experiments are ignored.

- *Experiments that show that differential expression of a miRNA results in differential expression of a gene:* In this class of experiments, the effect of manipulating the level of a single miRNA *MI* on the level of a single or multiple mRNAs is assessed using e.g. microarrays. The corresponding causal knowledge is that *MI* is a cause of T, for each transcript T whose level is altered in the experiment.
- *Experiments that show that differential expression of a miRNA results in differential levels of a protein:* In this class of experiments, the effect of manipulating the level of a single miRNA *MI* on the level of a single or multiple *proteins* is assessed using e.g. western blotting. The corresponding causal knowledge is that *MI* is a cause of I, for each protein (if the exact isoform is unknown) or protein isoform I whose level is altered in the experiment.

In summary, E dictates one of the following modes of conversion:

0. no causal relationship is entailed
1. *MI* is a cause of T for some transcript T of G
2. *MI* is a cause of I for some protein isoform I of G

MI and T are located in the cytosol. I may be post-translationally modified and located in any compartment, but this knowledge is usually not available from E. Therefore, it is assumed that I refers to the newly-translated form of the isoform, which is located in the cytosol and is unmodified. An edge

from each newly translated isoform to each of each of its derivatives is added automatically in a post-processing step of the CPKG (see section 5.2.10.1).

- *Biochemical reactions:* Each record denotes a biochemical reaction $I_1 + ... + I_n \longrightarrow O_1 + ... + O_n$, where each input I_i and each output O_i may be located in any compartment and, in the case of protein isoforms, be post-translationally modified. The corresponding causal knowledge is that I_i is a cause of I_j for each pair (I_i, I_j) of inputs, and I_i is a cause of O_j for each input/output pair (I_i, O_j).

5.2.2 Real data sources

The real data sources used to build the CPKG are presented below.

5.2.2.1 Ensembl

Ensembl [Yates et al., 2016] is a multi-species reference-genome database which uses MySQL for data storage. There are four databases for each species: *Core* stores sequences and genes, *Compara* (Comparative genomics) genomic alignments, *Variation* genetic variation, and *Funcgen* (Regulation) regulatory features. A Perl API is provided for each database that offers programmatic access to an installation of the database. In this work only the Core database is used.

Core stores *coordinates systems* (e.g. "chromosome") and each coordinate system contains *sequence regions*. The sequence regions within the "chromosome" coordinate system are the chromosomes. Sequence regions contain *genes*, which have a *status* (e.g. known or predicted), a *biological type* (e.g. protein coding, miRNA), and zero or more *transcripts*. Transcripts also have a status and a biological type. If a transcript is translatable, it has a canonical *translation*. A transcript usually has a gene specified but this is not necessary, as transcripts can be stored independently of genes. Both genes and transcripts are assigned a *stable ID*, are contained in a sequence region, and have *coordinates* (start position, end position, and strand) in the sequence region. Translations are also assigned a stable ID. Sequence regions, genes, transcripts, and translations may have *attributes*. For sequence regions, the "nonReference" and "lrg" attributes are of interest here. The former denotes that

the region is non-reference; the latter that the region is *Locus Reference Genomic (LRG)*.[1] Only reference, non-LRG regions in the "chromosome" coordinate system are used in this work; genes and transcripts in these regions are said to be *standard*. For transcripts, the "TSL" attribute is of interest. TSL stands for *Transcript Support Level* and is a measure of how well-supported a transcript is.[2] The TSL attribute takes takes values 1–5, 1 indicating best support, or N/A if the transcript was not analysed because it is a pseudogene annotation, a human leukocyte antigen transcript, an immunoglobin gene transcript, a T-cell receptor transcript, or a single-exon transcript. Finally, Core contains mappings between Ensembl (gene, transcript, or translation) stable IDs and external IDs in several major databases.

5.2.2.2 UniProt

UniProt [The UniProt Consortium, 2014] is a multi-species protein database. UniProt consists of *Swiss-Prot*, which contains manually curated and reviewed entries, and *TrEMBL*, which contains automatically-annotated, unreviewed entries. Only Swiss-Prot is used in this work.

Each *protein* has at least one *accession*, a (canonical) *sequence*, zero or more *features*, and zero or more *isoforms*. Features are annotations of the sequence, have types and location (start and end position on the sequence), and may have an identifier. A position may be exact, approximate (preceded by $<$ or $>$), or unknown. Of interest here are features of type *initiator methionine*, *signal peptide*, *propeptide*, *transit peptide*, *genetic chain*, and *peptide*, which are collectively referred to as *molecule-processing features*. Isoforms have an ID, which is the accession of the protein followed by a dash and a number. The sequence of an isoform is described with respect to the reference sequence using features of type *variable sequence*. If a protein has isoforms, one of the isoforms is the *canonical sequence*, and all feature locations refer to it.

UniProt also provides mappings between UniProt (protein or isoform) IDs and external IDs in several major databases.

---

[1] See http://www.ensembl.org/Help/Faq?id=300 for explanation.
[2] See http://www.ensembl.org/Help/Glossary?id=492 for details.

5.2.2.3 Gene Ontology

Gene Ontology (GO) [The Gene Ontology Consortium, 2015] is a set of three ontologies (called *aspects*): *molecular function*, *cellular component*, and *biological process*. Each ontology describes "is a" and "part of" relationships between molecular functions, cellular components, and biological processes, respectively, in a generic cell. In this work, only the cellular component aspect is used.

5.2.2.4 Gene Ontology Annotation

Gene Ontology Annotation (GOA) [Huntley et al., 2015] is a database of GO annotations for proteins in UniProt, complexes in *IntAct Complex Portal*, and RNA in *RNACentral*. Only protein annotations are used here.

Each annotation has a *database*, a *database object ID*, a possibly empty set of *qualifiers*, a *GO ID*, an *aspect*, a *database object type*, one or two *taxa*, and up to one *gene product form ID*. The database of the annotation is the database that contains the annotated object. The database object ID is the ID of the *canonical* version of the annotated object in the database; if a *variant* version of an object is being annotated, the gene product form ID is the ID of the variant version. The qualifiers can be "not", "contributes to", and "co-localizes with". The GO ID is the GO term the object is annotated with. The aspect is biological process, molecular function, or cellular component. The database object type is the type of object being annotated (e.g. protein, complex, RNA). The first taxon is the species that encodes the annotated object. If a second taxon is specified, it is the other species in a cross-species interaction. For proteins, the database is UniProt, the database object ID is a UniProt protein ID, and the gene product form ID is a Uniprot isoform ID.

Note that GOA mostly includes computationally inferred annotations [Škunca et al., 2012], and their use is the only time that predictions are used, instead of knowledge, for building the CPKG. Using GOA is necessary in order to localise the proteins in protein-protein interactions (see Section 5.2.8.4).

5.2.2.5 miRBase

miRBase [Kozomara and Griffiths-Jones, 2014] is a multi-species miRNA database. The main type of entry in miRBase is the *hairpin*, which is a *predicted* hairpin portion of a miRNA transcript. Each hairpin has an *accession number*, a *name*, and one or two mature *miRNAs*. Each miRNA also has a name and an accession, as well as a *start* and an *end position* on the hairpin.

5.2.2.6 The Molecular Interaction ontology

The *Proteomics Standards Initiative - Molecular Interaction (PSI-MI)* ontology was developed by the *Proteomics Standards Initiative (PSI)* of the *Human Proteome Organization* [Kerrien et al., 2007] for the annotation of molecular interaction experiments. As in GO, a relationship between two terms is either "part of" or "is a".

Ontology terms include "interaction detection method" (MI:0001), "interaction type" (MI:0190), "interactor type" (MI:0313), "experimental role" (MI:0495), "biological role" (MI:0500), "feature type" (MI:0116), and "feature range status" (MI:0333). An interaction detection method is the method used to determine an interaction. An interaction type can be an "association" (MI:0914), "colocalization" (MI:0403), a "genetic interaction" (MI:0208), and "predicted interaction" (MI:1110). Only associations are of interest here. An association denotes that the interaction participants are components of one or more complexes. A "physical association" (MI:0915) is an association where the participants are components of the same complex, but they are not necessary in direct contact with each other in the complex. Finally, a "direct interaction" (MI:0407) is a physical association where the participants are in direct contact with each other. There are also subclasses of direct interaction but they are not of interest here. When describing PPIs, the interactor type is either "protein" (MI:0326) or "peptide" (MI:0327). An experimental role is the role of a participant in a an experiment, and is usually "bait" (MI:0496), "prey" (MI:0498), "fluorescence acceptor" (MI:0584), "fluorescence donor" (MI:0583), "neutral component" (MI:0497), or "unspecified role" (MI:0499). A biological role is the role of a participant in its cell of origin. Biolog-

ical roles of interest here are “enzyme” (MI:0501), “enzyme target” (MI:0502), and “unspecified role” (MI:0499). A feature type is the type of a participant sequence feature, and can be a “biological feature” (MI:0252) or a “experimental feature” (MI:0505). Experimental features are not used in this work. Biological features of interest are “mutation” (MI:0118), “polyprotein fragment” (MI:0828), and “variant” (MI:1241). A feature range status assesses the certainty on the location of a feature. Examples of feature range statuses are “certain sequence position” (MI:0335) and “greater-than” (MI:0336).

5.2.2.7 The Protein Modification ontology

The *Proteomics Standards Initiative - Protein Modification (PSI-MOD)* ontology [Montecchi-Palazzi et al., 2008], also developed by the PSI, is an ontology of protein chemical modifications.

5.2.2.8 IMEx

There are several protein–protein interaction databases, which use different curation strategies and file formats, and often contain redundant interactions. The *International Molecular Exchange (IMEx)* consortium [Orchard et al., 2012] is a collaboration between providers of molecular interaction data with the goal of making a *non-redundant* and *deeply* and *homogeneously* curated set of molecular interactions available in a *standard file format*. The unit of curation is the publication; each paper is only found once in the IMEx dataset, and it is curated in full. IMEx currently focusses on PPIs.

The IMEx dataset is available in two formats developed by the PSI, *PSI-MI*,[3] an XML format, and *MITAB*, a simplified tabular format [Kerrien et al., 2007]. Both formats use the PSI-MI ontology. MITAB can only describe binary interactions and does not contain all information needed for conversion to causal knowledge; therefore, it not used here. The PSI has developed Java parsers for both formats.[4]

The central class in the data model of the PSI-MI format is the *entry*. Entries are associated with a list of *experiments*, a list of *interactors*, and a list of *interac-*

[3] http://www.psidev.info/node/60
[4] https://code.google.com/archive/p/psimi/

tions. Each experiment uses an *interaction detection method*, which is specified by a subclass of the namesake term (MI:0001) from the PSI-MI ontology. Interactors have an *interactor type* and may have an *organism*. The interactor type is specified by a subclass of the namesake term (MI:0313) from the PSI-MI ontology. An organism is linked to an NCBI Taxonomy ID and may have a *compartment* specified by a GO cellular component term. Interactions have a list of *participants*, may be associated with a list of experiments and a list of interaction types, and may be *modelled*, *intramolecular*, and/or *negative*:

- A modelled interaction has been inferred from another species.
- An intramolecular interaction has only one participant.
- A negative interaction has been shown *not* to occur in the experiments.

Participants reference an interactor and may have a list of *experimental roles* (each in a certain experiment), a list of *features*, and a *biological role*. An *experimental role* and a *biological role* is specified by a subclass of the namesake term (MI:0495 and MI:0500, respectively) from the PSI-MI ontology. Each feature has a *feature type*, a list of *feature ranges*, and is associated with a list of experiments where it is present. The feature type is either a subclass of the namesake term (MI:0116) from the PSI-MI ontology or a PSI-MOD term. A feature range is specified by a *start status* and either a start position or a start interval, and an *end status* and either an end position or an end interval. A start status or end status is specified by a subclass of "feature range status" (MI:0333).

IMEx has adopted a "deep" curation model, whose goal is to provide all the details of an interaction experiment [Orchard et al., 2012]. As these details are essential for the conversion of the interaction to causal knowledge, the IMEx dataset was chosen instead of the other databases. The IMEx curation rules can be found online.[5] The following rules are relevant:

- Negative interactions are out of scope.

[5] http://www.imexconsortium.org/sites/imexconsortium.org/files/documents/imex_curation_rules.pdf

- The interaction type is a subclass of "direct interaction" (MI:0407) only if the interaction occurs in vitro and the number of participants is 2. An interaction which has 1 bait and whose interaction detection method is a subclass of "affinity chromatography technology" (MI:0004) must have interaction type "physical association" (MI:0915) if it has 1 prey and "association" (MI:0914) if it has more than 1 prey.

- Participants must have a biological role and exactly one experimental role per experiment.

- When a participant is a protein that has isoforms, if the curated paper does not clearly specify which isoform is the interactor, the UniProt canonical isoform must be used in the annotation; otherwise, the specified isoform must be used.

- A post-translational modification required for an interaction to occur must be added in the annotation.

5.2.2.9 Chemical Entities of Biological Interest

Chemical Entities of Biological Interest (ChEBI) [Hastings et al., 2013] is an ontology of small chemical compounds that are relevant to biology and are not encoded directly by the genome.

5.2.2.10 miRTarBase

miRTarBase [Kozomara and Griffiths-Jones, 2014] stores validated MTIs for multiple species. Each MTI has a miRBase miRNA name, an Entrez gene ID, a set of free-text experiment names, and a support type. The support type is "Functional MTI", "Functional MTI (Weak)", "Non-functional MTI", or "Non-Functional MTI (Weak)", depending on whether the evidence suggests that the interaction occurs or does not occur ("functional" vs. "non-functional") and whether the evidence *strongly* suggests or opposes a *direct* interaction of the miRNA with the target.

5.2.2.11 Reactome

Reactome [Fabregat et al., 2016] is a pathway database. The central classes in the Reactome data model[6] are *reaction-like event (RLE)* and *physical entity*.

Each RLE has *inputs*, *outputs*, and *catalysts*. Inputs, outputs, and catalysts are physical entities (see next paragraph). Some input components (e.g. protein domains) may be as flagged as *required*. An RLE is associated with a set of *species*, and possibly a set of *diseases*. Finally, an RLE can be *chimeric* (that is, involving entities from more than one species), *computationally inferred*, and/or *inferred from another species*. Each RLE is either a *reaction*, a *black box event (BBE)*, a *polymerisation*, or a *depolymerisation*:

- A reaction represents an actual chemical reaction, one that has balanced inputs and outputs.
- A BBE is used to describe a reaction that is unbalanced for some reason, or a more complex process that is not understood entirely or has intermediate steps that need not be described.
- A polymerisation describes the addition of a unit to a polymer and a depolymerisation describes the removal of a unit to a polymer. Polymerisation and depolymerisation are inherently unbalanced processes.

A physical entity is any entity that can interact with other entities in the cell, or a set of such entities; sets are used to prevent combinatorial explosion. A physical entity can be a *complex*, *entity set*, *genome-encoded entity (GEE)*, *other entity*, *polymer*, or *simple entity*:

- A complex has a set of *components*, which are physical entities.
- An entity set has a set of *members*. An entity set can be either a *defined set*, a *candidate set*, or an *open set*.
  - The members of a defined set are physical entities with interchangeable function.

[6] http://www.reactome.org/pages/documentation/data-model/

  - A candidate set has both members and *candidate members*; the latter are hypothesised to perform the function performed by the former.
  - An open set is a set of entities that cannot be counted in practice, e.g. mRNA; the members of an open set are merely examples.

- A GEE can be a gene, protein, or RNA molecule. An *entity with accessioned sequence (EWAS)* is a GEE with a *reference entity*. A reference entity represents an entry in some database and has an identifier and a reference database. An EWAS represents a gene, protein, protein isoform (when the specific isoform is unknown), and RNA molecule when its reference entity is a *reference DNA sequence*, *reference gene product*, *reference isoform*, and *reference RNA sequence*, respectively. An EWAS can also be a fragment and/or a modification of the accessioned sequence. Fragments are specified by a *start coordinate* and an *end coordinate* in the coordinate system of the reference database. Modifications are specified by a set of *abstract modified residues* (see below).

- An other entity is used to describe a complex structure in the cell that cannot or needs not be described on the molecular level.

- A polymer has a set of *repeated units*, which are physical entities.

- A simple entity is a molecule not encoded by the genome (e.g. ATP). Each simple entity has a reference entity of class *reference molecule*. The reference database is always ChEBI.

An abstract modified residue is either a *genetically modified residue* or a *translational modification*. The former is not of interest here, since it is found in disease RLEs, which are not used (see Section 5.2.8.6). A translational modification has a *coordinate* and a *PSI-MOD term* that describes the translational modification. A translational modification can be a *modified residue*, a *group-modified residue*, or a *cross-linked residue*:

- A modified residue has no additional properties.

- A group-modified residue describes the attachment of a chemical group or a polymer. Because this cannot be represented by a PSI-MOD term alone, the PSI-MOD term is used to describe the link, and an additional *modification* property is used to describe the attached group or polymer. The modification is a *reference group*, a subclass of reference entity, if the attached entity is a chemical group. The reference database is always ChEBI. The modification is a polymer if the attached entity is a polymer.

- A cross-linked residue describes a cross-link within a protein or with other proteins. The PSI-MOD term is used to describe half of the link. The *second coordinates* property is the list of coordinates, in the same or another protein, of the residues at the other side of the link. A cross-linked residue can be either an *intrachain cross-linked residue* or an *interchain cross-linked residue*.

  - An intrachain cross-linked residue has only one second coordinate in the same protein.
  - An interchain cross-linked residue has an additional *second reference sequences* property that specifies the linked proteins (in UniProt), and the second coordinates are in those proteins.

5.2.3 Cellular compartments

As mentioned in Section 5.2.1, the molecular entities in the conceptual database, and therefore the corresponding CPKG nodes, are localised. In this implementation, each localised CPKG node is associated with a compartment from (an edited version of) the set of GO cellular compartments in Reactome (123 compartments in total). The compartments are listed in Table 1 of *Supplementary Information I – Cellular Compartments* and are meant to be non-overlapping;[7] however, it turns out that this not the case. For example, both "mitochondrion" (GO:0005739) and "mitochondrial matrix" (GO:0005759) are in the set. Nevertheless, each localised Reactome entity was converted to a CPKG node with the same compartment. The "nucleoplasm" (GO:0005654), "cytosol" (GO:0005829), and "mitochondrial matrix"

[7] http://wiki.reactome.org/index.php/Glossary_Data_Model#EntityCompartment.3D

(GO:0005759) terms in the Reactome set where used for Ensembl, UniProt, and miRBase objects. Compartment overlaps had to be taken into account when assigning compartments to the participants of an IMEx interaction (see Section 5.2.8.4). A drawback of using Reactome's compartments is that the CPKG is confined to the biology covered by Reactome.

5.2.4 CPKG node types

The objects of real data sources are converted to CPKG nodes of certain types. Table 5.1 lists the *general* CPKG node types, that is, the ones that are not specific to any data source. Note that sometimes the GO cellular compartment is used in the name of a node and sometimes is not. In the latter case the entity is always localised in one compartment (e.g. a miRNA in the cytosol); therefore, the compartment is not required to uniquely identify the node. In order for isoforms and proteins to be uniquely identified, their *post-translational modifications* need to be specified. A post-translational modification is a pair of a coordinate and a *modified residue*. Table 5.2 lists the general modified-residue types. Table 5.3 and 5.4 lists the Reactome-specific node types and modified-residue types, respectively. Finally, Table 5.5 lists the IMEx-specific node types. There are no node types or modified-residue types specific to the rest real data sources used.

5.2.5 CPKG reference types

The edges in the CPKG of each conceptual data source have a different reference type, shown in Table 5.6. The edges created in each post-processing and adaptation step also have a different reference type, shown in Table 5.7 and 5.8, respectively.

5.2.6 ID mapping

The various ID mappings that are performed when building the CPKG are discussed below.

Name	Description	Identifier(s)	Genomic regions
chemical entity	localised non-genome-encoded molecule	ChEBI term and GO cellular compartment term	none
gene	gene	Ensembl gene stable ID	gene
isoform	localised (fragment of) (post-translationally-modified) protein isoform	UniProt isoform ID, fragment coordinates, set of post-translational modifications, GO cellular compartment term	isoform-fragment regions of each gene
miRNA	miRNA	miRBase accession	miRNA region of each miRNA gene
phenotype	the phenotype	"phenotype"	none
protein	set of localised (fragments of) (post-translationally-modified) protein isoforms	UniProt protein ID, fragment coordinates on the canonical sequence, set of post-translational modifications of the canonical sequence (see Table 5.2), GO cellular compartment term	genomic regions of the isoforms
transcript	localised transcript	Ensembl transcript stable ID, GO cellular compartment term	Ensembl transcript

Table 5.1: General node types.

Name	Description	Identifier(s)
group-linked residue	residue linked to a chemical group	PSI-MOD term and ChEBI term
interchain cross-linked residue	residue cross-linked to residues in other proteins	PSI-MOD term, UniProt IDs, coordinates on the corresponding UniProt canonical sequences
intrachain cross-linked residue	residue cross-linked to another residue in the same protein	PSI-MOD term, coordinate of the other residue
singly-modified residue	residue undergone a single modification	PSI-MOD term

Table 5.2: General modified-residue types.

5.2.6.1 Mapping UniProt isoform IDs to sets of Ensembl transcript stable IDs

A mapping from a UniProt isoform ID to a set of Ensembl transcript stable IDs is performed in order to identify the genomic regions associated with an isoform node. Both the UniProt isoform ID and the corresponding UniProt protein ID are first mapped to a set of Ensembl transcript stable IDs using the UniProt ID mappings. IDs corresponding to translatable standard transcripts with TSL = 1 or N/A whose translation sequence matches that of the isoform are then returned.

5.2.6.2 Mapping Ensembl transcript stable IDs to sets of UniProt isoform IDs

When building the Translation CPKG (see Section 5.2.8.2), mapping an Ensembl transcript stable ID to a set of UniProt isoform IDs is necessary for identifying the isoforms that correspond to a transcript. Using the UniProt ID mappings, the Ensembl transcript stable ID is first mapped to a set of UniProt IDs. For each UniProt ID that is a valid protein ID, the IDs of the protein's isoforms whose sequence matches that of the translation are returned. UniProt IDs that correspond to isoforms whose sequence matches that of the translation are also returned.

Name	Description	Genomic regions
Reactome candidate set	Reactome candidate set with only candidate members	none
Reactome complex	Reactome complex	genomic regions of the components
Reactome EWAS	Reactome EWAS with an invalid database or invalid ID in a valid database	none
Reactome entity set	Reactome defined set or candidate set with at least one non-candidate member	genomic regions of the members
Reactome GEE	Reactome GEE	none
Reactome isoform	Reactome EWAS whose reference entity is a reference isoform and either with coordinates that do not correspond to UniProt molecule-processing features or with invalid modified-residue coordinates	isoform-fragment regions of each gene
Reactome miRNA hairpin	Reactome EWAS with valid miRBase accession but invalid (miRNA) coordinates	miRNA gene(s)
Reactome open set	Reactome open set	none
Reactome other entity	Reactome other entity	none
Reactome polymer	Reactome polymer	genomic regions of the repeated units
Reactome protein	Reactome EWAS whose reference entity is a reference gene product and either with coordinates that do not correspond to UniProt molecule-processing features or with invalid modified-residue coordinates	genomic regions of the isoforms
Reactome simple entity	Reactome simple entity with invalid ChEBI term	none

Table 5.3: Reactome-specific node types. The node identifier is always the Reactome stable identifier.

5.2.6.3 Mapping UniProt protein IDs to sets of Ensembl gene stable IDs

In order to delete nodes by gene expression as a post-processing step of the graph (see Section 5.2.11.3), each gene-product node is associated with a set of genes. A

Name	Description	Identifier(s)
Reactome polymer-linked residue	residue linked to a Reactome polymer	PSI-MOD term and Reactome stable identifier

Table 5.4: Reactome-specific modified-residue types.

Name	Description
IMEx isoform	IMEx participant with valid UniProt isoform ID but with either undetermined coordinates or invalid feature coordinates
IMEx participant	IMEx participant with invalid UniProt ID
IMEx protein	IMEx participant with valid UniProt protein ID but with either undetermined coordinates or invalid feature coordinates

Table 5.5: IMEx-specific node types. The node identifiers are always IMEx ID, participant ID, and GO cellular compartment term.

mapping from a UniProt protein ID to a set of Ensembl gene stable IDs is performed when associating a protein node with a set of genes. First, the UniProt protein ID

Conceptual data source	Reference type	Reference identifier(s)
Transcription	transcription reference	Ensembl transcript stable ID
Translation	translation reference	Ensembl transcript stable ID
miRNA biogenesis	miRNA-biogenesis reference	Ensembl transcript stable ID
miRNA targets	miRTarBase experiment miRNA-target interaction (MTI) reference	miRTarBase ID, miRTarBase experiment name
Reactions	Reactome reaction-like event reference	Reactome stable identifier

Table 5.6: Reference type for each conceptual data source.

Post-processing step	Reference type	Reference identifier(s)
addition of derivation edges	derivation reference	name of the newly-translated-isoform node, name of the derivative node
addition of set membership edges	set membership reference	name of the member node, name of the set node
addition of subset relationship edges	subset relationship reference	name of the subset node, name of the set node

Table 5.7: Reference types for each post-processing step.

Adaptation step	Reference type	Reference identifier(s)
addition of protein-phenotype edges	protein phenotype causation reference	protein-node name

Table 5.8: Reference types for each adaptation step.

is mapped to a set of Ensembl gene stable IDs using the UniProt ID mappings. Subsequently, IDs corresponding to standard genes are returned.

5.2.6.4 Mapping Ensembl gene stable IDs to sets of miRBase hairpin accessions

When building the miRNA biogenesis CPKG (see Section 5.2.8.3), mapping an Ensembl gene stable ID to a set of miRBase hairpin accessions is needed in order to identify the hairpins that correspond to a miRNA gene. The miRBase hairpin accessions that correspond to an Ensembl gene stable ID are first identified using the Ensembl Core ID mappings. Accessions that correspond to hairpins whose genomic coordinates and sequence match those of the gene are then returned.

5.2.6.5 Mapping Entrez gene IDs to sets of Ensembl gene stable IDs

Mapping Entrez gene IDs to sets of Ensembl gene stable IDs is performed when building the MiRNA Target CPKG (see Section 5.2.8.5). Using the Ensembl Core ID mappings, the Ensembl gene stable IDs that correspond to the Entrez gene ID are identified and IDs corresponding to standard genes are returned.

5.2.6.6 Mapping miRBase hairpin accessions to sets of Ensembl gene stable IDs

In order to identify the genomic regions and genes of a miRNA or Reactome miRNA hairpin node, a mapping from a miRBase hairpin accession to a set of Ensembl gene stable IDs is performed. The Ensembl gene stable IDs corresponding to a miRBase hairpin accession are first found using the Ensembl Core ID mappings. IDs corresponding to standard genes whose coordinates and sequence match those of the hairpin are subsequently returned.

5.2.6.7 Mapping HGNC symbols to sets of Ensembl gene stable IDs

The *HUGO Gene Nomenclature Committee (HGNC)* assigns unique symbols to human genes. Mapping HGNC symbols to Ensembl gene stable IDs is performed when removing nodes from the CPKG based on gene expression (see Section 5.2.11.3). The HGNC symbol is first mapped to a HGNC ID using the mappings from the HGNC website.[8] If the mapping is unsuccessful, alias and previous symbols are also searched. If a symbol is found, the HGNC ID is mapped to a set of Ensembl gene stable IDs using the Ensembl Core ID mappings and the IDs corresponding to standard genes are returned.

5.2.7 Conversion of real-data-source objects to CPKG nodes

For each real data source, the details of the conversion of its objects to CPKG nodes are described below.

5.2.7.1 Ensembl records

Each gene and transcript is converted to a gene and transcript node, respectively. Gene nodes of genes and transcript nodes of transcripts on the "MT" sequence region are associated with GO cellular compartment "mitochondrial matrix" (GO:0005759); gene nodes of genes and transcript nodes of transcripts on other sequence regions are associated with GO cellular component "nucleoplasm" (GO:0005654) and "cytosol" (GO:0005829), respectively.

5.2.7.2 miRBase entries

Each miRNA is converted to a miRNA node.

5.2.7.3 IMEx-interaction participants

In order for a participant in an IMEx interaction to be converted to a CPKG node, a GO cellular compartment must be supplied. The assignment of participant compartments is discussed in Section 5.2.8.4. The conversion depends on the form of the UniProt ID of the participant's interactor:

- *The ID has the form of a protein or peptide ID:* If the ID is invalid, the par-

[8] http://www.genenames.org/

ticipant is converted to an IMEx participant node. Otherwise, the participant coordinates are determined as follows:

- *Proteins:* If the participant has a feature of type "polyprotein fragment" (MI:0828), the coordinates are set to the feature range if it corresponds to molecule-processing features in UniProt, otherwise they are left undetermined. If the participant has no such feature, the coordinates are set to the location of the sole genetic chain in UniProt, if such a genetic chain exists, otherwise they are left undetermined.
- *Peptides:* The coordinates are set to the location of the corresponding molecule-processing feature in Uniprot.

If the participant coordinates cannot be determined or there are invalid PSI-MOD feature ranges, the participant is converted to an IMEx protein node. Otherwise, the PSI-MOD features of the participant are converted to singly-modified residues, and the participant is converted to a protein node.

- *The ID has the form of an isoform ID:* If the ID is invalid, the participant is converted to an IMEx participant node. Otherwise, the participant coordinates are set to the isoform translation of the location of the sole genetic chain in UniProt, if such a genetic chain exists, and are left undetermined otherwise (no polyprotein-fragment features were found for isoform participants). If the participant coordinates cannot be determined or there are invalid PSI-MOD feature ranges, the participant is converted to an IMEx isoform node. Otherwise, the PSI-MOD features of the participant are converted to singly-modified residues, and the participant is converted to an isoform node. The experimental features of the participants are ignored.

5.2.7.4 Reactome entities

Each complex is converted to a Reactome complex node and each defined set is converted to a Reactome entity set node. Each candidate set is converted to a Reactome entity set node if it has at least one non-candidate member; otherwise, it

is converted to a Reactome candidate set node. For each EWAS, the conversion depends on the class of the reference entity:

- *The reference entity is a reference DNA sequence:* The EWAS is converted to a gene node if the reference database is human Ensembl gene and the reference entity identifier is a valid Ensembl gene stable ID; otherwise, the EWAS is converted to a Reactome EWAS node.

- *The reference entity is a reference gene product:* If the reference entity identifier is not a valid Ensembl protein ID, the EWAS is converted to a Reactome EWAS node; otherwise, the EWAS coordinates and the modified-residue coordinates are checked. If the EWAS coordinates do not correspond to UniProt molecule-processing features or there are invalid modified-residue coordinates, the EWAS is converted to a Reactome protein node; otherwise, the EWAS is converted to a protein node. The EWAS coordinates are required to match UniProt molecule-processing features as the former are not versioned; thus, they may refer to an older version of UniProt and no longer be valid.

- *The reference entity is a reference isoform:* If the reference entity identifier is not a valid Ensembl isoform ID, the EWAS is converted to a Reactome EWAS node; otherwise, the EWAS coordinates and the modified-residue coordinates are checked. If the EWAS coordinates do not correspond to UniProt molecule-processing features or there are invalid modified-residue coordinates, the EWAS is converted to a Reactome isoform node; otherwise, the EWAS is converted to an isoform node.

- *The reference entity is a reference RNA sequence:* The conversion depends on the reference database:

  - *The reference database is human Ensembl transcript:* The EWAS is converted to a transcript node if the reference entity identifier is a valid Ensembl transcript stable ID; otherwise, the EWAS is converted to a Reactome EWAS node.

- *The reference database is miRBase:* The EWAS is converted to a miRNA node if the reference entity identifier is a valid miRBase hairpin accession and the EWAS coordinates match the coordinates of a miRNA on the miRBase hairpin. If the reference entity identifier is a valid miRBase hairpin accession but the EWAS coordinates do not match the coordinates of any miRNA on the miRBase hairpin, the EWAS is converted to a Reactome miRNA hairpin node. Finally, if the reference entity identifier is not a valid miRBase hairpin accession, the EWAS is converted to a Reactome EWAS node.
- *The reference database is neither human Ensembl transcript nor miRBase:* The EWAS is converted to a Reactome EWAS node.

Each polymer is converted to a Reactome polymer node. Each simple entity is converted to a simple entity node if the reference entity identifier is valid ChEBI term; otherwise, it is converted to a Reactome simple entity node. Finally, each open set, GEE, and other entity, is converted to a Reactome open set node, Reactome GEE node, and Reactome other entity node, respectively.

The abstract modified residues are converted to CPKG modified residues as follows. Each modified residue is converted to a singly modified residue. Each group-modified residue is converted to a group-linked residue and to a Reactome polymer-linked residue if the modification is a reference group and a polymer, respectively. Finally, each interchain cross-linked residue and intrachain cross-linked residue is converted to the namesake CPKG modified residue.

5.2.8 Conversion of conceptual data sources to CPKGs

In this section, the conversion of conceptual data sources to CPKGs is discussed. Table 5.9 lists the real data sources used for each conceptual data source.

5.2.8.1 Transcription

Ensembl is used as the sole data source for building the transcription CPKG. For each known standard gene, the known (standard) transcripts with TSL = 1 or N/A are identified. If at least one transcript is identified, a transcription reference is

Conceptual data source	Real data source
Transcription	**Ensembl**
Translation	**Ensembl**, UniProt
MiRNA biogenesis	**miRBase**, Ensembl
MiRNA targets	**miRTarBase**, Ensembl, miRBase, Uniprot
Protein-protein interactions	**IMEx**, MI, PSI-MOD, GO, GOA, Uniprot, Ensembl
Reactions	**Reactome**, Ensembl, miRBase, Uniprot, ChEBI, PSI-MOD, GO

Table 5.9: Real data sources used for each conceptual data source. For each conceptual data source, the main real data source used is in boldface.

created, the gene and the transcripts are converted to nodes, and an edge citing the reference is created from the gene node to each transcript node. The resulting graph contained 113157 nodes and 68699 edges.

5.2.8.2 Translation

The translation CPKG is built using Ensembl and UniProt. For each known translatable standard transcript with TSL = 1 or N/A in Ensembl, the corresponding UniProt isoforms are identified. If at least one isoform is identified, a translation reference is created, the transcript and the isoforms are converted to nodes, and an edge citing the reference is created from the transcript node to each isoform node. The resulting graph contained 51175 nodes and 27241 edges.

5.2.8.3 MiRNA biogenesis

miRBase and Ensembl are the data sources used for building the miRNA-biogenesis CPKG. For each known miRNA standard gene in Ensembl, the corresponding miRBase hairpins are identified using the Ensembl gene stable ID to miRBase accession mappings. If at least one hairpin is identified, the transcript corresponding to the gene and the miRNAs corresponding to the hairpin are identified, a miRNA-biogenesis reference is created, the transcript and the miRNAs are converted to nodes, and an edge citing the reference is created from the transcript node to each miRNA node. The resulting graph contained 3730 nodes and 2314 edges.

5.2.8.4 Protein-protein interactions

IMEx is the main data source used for building the PPI CPKG. The IMEx dataset is not currently available as a single download. Therefore, the IMEx interactions in the *IntAct* database and its hosted databases [Orchard et al., 2013] were filtered from the file in PSI-MI format that is available to download from the IntAct website[9] and contains all interactions in those databases. The *Database of Interacting Proteins (DIP)* [Salwinski et al., 2004] offers its IMEx subset as a single file in PSI-MI format.[10] *MatrixDB* [Launay et al., 2014], a database specialised in extracellular matrix interactions, is, unfortunately, only available to download in MITAB format and was not used. The files were processed in MATLAB using the Java parser provided by the PSI and the MATLAB Java interface. An interaction is eligible for conversion to causal knowledge if it is neither negative nor intramolecular, it has at least two participants, all participants are eligible, and at least one experiment is eligible for the interaction. A participant is eligible if all the following conditions are satisfied:

- The interactor type is "bioactive entity" (MI:1100), defined as "molecules showing activity in a living system but not encoded by a genomic sequence", or the interactor organism is human.
- The participant has a biological role that is a term in the PSI-MI ontology.
- The participant has exactly one experimental role per experiment.
- All feature types are terms in the PSI-MOD or PSI-MI ontology.
- No feature has type "mutation" (MI:0118) or "variant" (MI:1241).

An experiment is eligible for an interaction if the mode of conversion to causal knowledge is nonzero. The mode of conversion depends on the experimental detection method, the number of participants, their experimental and biological roles, and the interaction type. For each experimental detection method, only the *standard* protocol is considered. For example, in a pull-down assay, one participant is

[9] http://www.ebi.ac.uk/intact/downloads
[10] http://dip.doe-mbi.ucla.edu/dip/Download.cgi?SM=10

expected to be bait, the rest participants are expected to be prey, and the participants have no special biological roles. If any of these conditions are false, the experiment is not a standard pull-down experiment and the mode of conversion is zero. The mode of conversion is determined as follows:

- *The experimental detection method is a subclass of "display technology" (MI:0034):* Typically, this means that the binding of a single bait protein to complexes composed of unknown subsets of a set of prey proteins was demonstrated. If one participant of the interaction is bait and the rest are prey and all participants have unspecified biological role, the mode of conversion is 1 if there are only two participants and the interaction is direct and 2 otherwise. In all the other cases the mode of conversion is 0. Note that the definition of "pull down" (MI:0096) includes experiments where the bait protein is either immobilised on the column or "fished" from the mixture. Since it is not possible to distinguish between the two cases without consulting the respective publication, the method is treated as one that shows that a single bait protein participates in complexes composed of unknown subsets of a set of prey proteins (see below).

- *The experimental detection method is a subclass of "pull down" (MI:0096), a subclass of "coimmunoprecipitation" (MI:0019), or "tandem affinity purification" (MI:0676):* This regularly implies that a single bait protein was shown to participate in complexes with unknown subsets of a set of prey proteins. If there are only two participants in the interaction, one is bait and the other is pray, both have unspecified biological role, and the interaction is direct, the mode of conversion is 1; otherwise, the mode of conversion is 0.

- *The experimental detection method is a subclass of "surface plasmon resonance" (MI:0107), a subclass of "solid phase assay" (MI:0892), "peptide array" (MI:0081), "protein array" (MI:0089), or "competition binding" (MI:0405):* Normally, this means that the binding of a single bait protein to a single prey protein was demonstrated. If there are only two participants in the

interaction, one is bait and the other is pray, both have unspecified biological role, and the interaction is direct, the mode of conversion is 1; otherwise, the mode of conversion is 0.

- *The experimental detection method is a subclass of "enzymatic study" (MI:0415):* Typically, what has been shown is that a single enzyme binds to a single enzyme target. If there are only two participants in the interaction, one is enzyme and the other is enzyme target, the experimental role of each participant is either neutral or unspecified, and the interaction is direct, the mode of conversion is 1; otherwise, the mode of conversion is 0.

- *The experimental detection method is a subclass of "cross-linking" (MI:0030), a subclass of "cosedimentation" (MI:0027), a subclass of "electrophoretic mobility-based method" (MI:0982, a subclass of "comigration in gel electrophoresis", MI:0807), a subclass of "ion exchange chromatography" (MI:0226), or "molecular sieving" (MI:0071):* What has been normally shown is that a set of proteins participate in the same complex because they were detected in the same place after an attempt to separate the contents of a mixture. If there are only two participants, the experimental role of each participant is either neutral or unspecified, their biological role is unspecified, and the interaction is direct, the mode of conversion is 1; otherwise, the mode of conversion is 0.

- *The experimental detection method is a subclass of "light scattering" (MI:0067), a subclass of "classical fluorescence spectroscopy" (MI:0017), "isothermal titration calorimetry" (MI:0065), or "fluorescence polarization spectroscopy" (MI:0053):* This commonly implies that two proteins that were the only proteins in solution in an in-vitro experiment directly interact because some measurement indicated that the proteins must be bound. If there are only two participants in the interaction, the experimental role of each participant is either neutral or unspecified, their biological role is unspecified, and the interaction is direct, the mode of conversion is 1; otherwise, the mode

of conversion is 0.

- *The experimental detection method is "fluorescent resonance energy transfer" (MI:0055):* Typically, this means that two proteins, a fluorescence donor and a fluorescence acceptor, came really close together either in vitro or in vivo. If there are only two participants in the interaction, one is fluorescence donor and the other is fluorescence acceptor, their biological role is unspecified, and the interaction is direct (indicating that the in vitro version of the experiment was performed), the mode of conversion is 1; otherwise, the mode of conversion is 0.

- *The experimental detection method is a subclass of "protein complementation assay" (MI:0090) (a superclass of "two hybrid", MI:0018), or "luminescence based mammalian interactome mapping" (MI:0729):* These experiments are ignored because they can normally only show that two proteins participate in the same complex [see Blasche and Koegl, 2013, for a description of luminescence based mammalian interactome mapping].

- *The experimental detection method is a subclass of "imaging technique" (MI:0428), "x-ray crystallography" (MI:0114), or "nuclear magnetic resonance" (MI:0077):* These experiments are ignored because, even if a binary interaction is deemed to be direct, there could be other proteins in the complex.

- *The experimental detection method is a subclass of "affinity chromatography technology" (MI:0004) or "fluorescence technology" (MI:0051):* These experiments are ignored because the experimental detection method is too general.

- *The experimental detection method is "proximity ligation assay" (MI:0813):* These experiments are ignored because they normally only show that two proteins came close in vivo [Weibrecht et al., 2014].

- *The experimental detection method is none of the above:* Experiments with

other experimental detection methods are ignored because there are too few (less than 100) interactions for each of the methods.

For each eligible experiment of each eligible IMEx interaction, the mode is checked. If the mode is 1, the binary interaction between the two participants is processed with mode 1 (see below). If the mode is 2, the binary interaction between the bait and each prey protein is processed with mode 2. The resulting graph contained 4659 nodes and 12673 edges.

Processing of binary interactions. A binary interaction between participants X and Y is processed with mode M as follows. First, a set of cellular compartment pairs is identified as explained in the next paragraph. Then, for each pair $\{C_x, C_y\}$ of compartments, X and Y are converted to CPKG nodes V_x and V_y with compartment C_x and C_y, respectively. If the nodes are distinct (they may not be for distinct participants), an IMEx interaction reference is created. If $M = 1$, edges $V_x \rightarrow V_y$ and $V_y \rightarrow V_x$ citing the reference are created. If $M = 2$, only edge $V_y \rightarrow V_x$ citing the reference is created, where Y is prey and X is bait.

Identification of interaction compartments. Experimental methods for detecting molecular interactions can rarely identify the compartments of the participants. Indeed, only just a few interactor organisms have a compartment specified in the IMEx dataset. However, CPKG nodes need to be localised. Therefore, if PPIs are to be converted to a CPKG at all, compartments for the participants must be somehow identified. The protein GO-cellular-component annotations in GOA are used to achieve this. First, the GOA annotations without any qualifiers and with a cellular component aspect, a protein database object type, and a single human taxon are selected. For each participant, the form of the interactor UniProt ID is checked. If the ID has the form a protein or peptide ID, the annotations with a database object ID matching the protein ID or the protein ID corresponding to the peptide ID, respectively, are identified. If the ID has the form of an isoform ID, the annotations with a matching gene product form ID *or* a database object ID matching the protein ID corresponding to the isoform ID are identified. The GO IDs of the annotations have then to be mapped to the set of Reactome compartments. However,

as mentioned in Section 5.2.3, there are overlapping compartments in that set. In addition, some compartments are too general (e.g. "cytoplasm"; GO:0005737) or too specific (e.g. "integrin complex"; GO:0008305). Furthermore, there are non-human (e.g. "plastid"; GO:0009536) and disease-related (e.g. "host cell cytosol"; GO:0044164) compartments. Finally, the inclusion or non-inclusion of organelle, membrane, and lumen terms of membrane-bounded organelles is inconsistent in the set. For example, "late endosume lumen" (GO:0031906) and "late endosome membrane" (GO:0031902) are in the set, but "late endosome" (GO:0005770) is not; "azurophil granule lumen" (GO:0035578) is in the set, but "azurophil granule membrane" (GO:0035577) and "azurophil granule" (GO:0042582) are not. In order for the set of Reactome compartments to be used for interaction-participant localisation, compartments that are general, too specific, non-human, or disease-related were removed. For each membrane-bounded organelle such that the corresponding membrane, lumen, or organelle term is in the set, the other terms were added as well. Then the set was split into a set of non-overlapping (*specific*) compartments and a set of overlapping (*general*) compartments, which are listed in Table 2 and 3, respectively, of *Supplementary Information I – Cellular Compartments*. The general compartments are used to facilitate the merging of IMEx nodes with Reactome nodes. A specific compartment is assigned to a participant if the compartment is in the set of participant compartments identified earlier or it is a superterm of some compartment in that set; a general compartment is assigned to a participant if the compartment is in the set of participant compartments. After both participants have compartments assigned, the cartesian product of the compartments can be used as the set of interaction compartment pairs. However, some of the pairs could be biologically implausible (e.g. nucleoplasm and extracellular region). Alternatively, participants can be required to be in the same compartment. However, interactions that happen in adjacent compartments will be missed. Therefore, a set of *compartment adjacencies* was created based on basic understanding of cell biology and used to filter the cartesian product of the participant compartments. The adjacencies are listed in Table 4, *Supplementary Information I – Cellular Compartments*.

5.2.8.5 MiRNA targets

The main data source used for building the miRNA-target CPKG is miRTarBase. Ignoring experiments with < 100 MTIs and ChIP-seq experiments, the experiments in miRTarBase can be classified according to their information content as in Section 5.2.1. Table 5.10 lists indicative experiment names in miRTarBase for the experiments of each class, as well as the mode of conversion to causal knowledge. ChIP-seq experiments are ignored as their aim is to identify transcription factors regulating the expression of genes (in this case, miRNA genes) [Yang et al., 2013].

Experiment class	**Indicative experiment names in miRTarBase**	**Mode**
reporter assays	B-globin reporter assay, EGFP reporter assay, GFP reporter assay, GLuc reporter assay, GUS reporter assay, LacZ reporter assay, luciferase reporter assay, reporter assay	1
immunoprecipitation of RISC components	Coimmunoprecipitation, immunoprecipitation	1
identification of miRNA binding sites	CLASH, HITS-CLIP, PAR-CLIP	0
techniques for measuring mRNA levels	In situ hybridization, microarray, Next Generation Sequencing, northern blot, qPCR, qRT-PCR, RT-PCR, semi-qRT-PCR, Sequencing	1
techniques for measuring protein levels	ELISA, flow cytometry, immunocytochemistry, immunofluorescence, immunohistochemistry, immunostaining, proteomics, pSILAC, quantitative proteomic approach, SILAC, Western blot (immunoblot)	2

Table 5.10: Experiment classes, indicative experiment names in miRTarBase, and mode of conversion to causal knowledge (see Section 5.2.1).

For each MTI, the following conditions must be satisfied in order for conversion to causal knowledge to be considered:

- The support type is “Functional MTI” or “Functional MTI (Weak)”.
- There is at least one eligible experiment.
- The miRBase miRNA name is valid.

- The Entrez gene ID maps to at least one Ensembl gene stable ID of a standard gene.

- At least one of the mapped standard genes has at least one known (standard) transcript with TSL = 1 or N/A.

If all of the above are true, the miRNA is converted to a miRNA node, the transcripts are converted to transcript nodes if any experiment has mode 1, and the UniProt isoforms corresponding to each transcript are identified and converted to isoform nodes if any experiment has mode 2. For each experiment, a miRTarBase MTI reference is created. If the experiment has mode 1, an edge citing the reference is created from the miRNA node to each transcript node. If the experiment has mode 2 and there is at least one isoform, an edge citing the reference is created from the miRNA node to each isoform node. The resulting graph contained 23598 nodes and 54532 edges.

It is worth appreciating the fact that even "weak" MTIs can be converted to causal knowledge, as it is the information content of the experiment that matters and not the strength of the evidence regarding a direct MTI. In contrast, non-causal graph-based methods [e.g. see Emily et al., 2009] may discard "weak" relationships in order increase confidence in the resulting graph.

5.2.8.6 Reactions

The reaction CPKG is built using Reactome as the main data source. Catalysts are considered to be both inputs and outputs. Non-human, chimeric, computationally-inferred, and/or disease reaction-like events are ignored (only prion-disease disease RLEs would be of interest, and there is no one in Reactome). Finally, BBEs are ignored because treating them like reactions may lead to incorrect causal knowledge. For example, assume that $\mathrm{X}+\mathrm{Y}+\mathrm{Z} \longrightarrow \mathrm{U}+\mathrm{V}$ is a BBE. Then, $Z \rightarrow X$, $Z \rightarrow Y$, and $Z \rightarrow U$ will be among the edges created in the CPKG. However, the BBE may actually be a shortcut for two reactions, $\mathrm{X}+\mathrm{Y} \longrightarrow \mathrm{U}+\mathrm{W}$ and $\mathrm{W}+\mathrm{Z} \longrightarrow \mathrm{V}$. Clearly, edges $Z \rightarrow X$, $Z \rightarrow Y$, and $Z \rightarrow U$ would not be created from these reactions.

An RLE involving entity sets is equivalent to a set of "concrete" RLEs,

i.e, RLEs involving concrete entities. For example, RLE $\mathbf{A} + \mathbf{B} + \mathrm{C} \longrightarrow \mathbf{D}$, where $\mathbf{A} = \{A_1, A_2\}$, $\mathbf{B} = \{B_1, B_2\}$, and $\mathbf{D} = \{D_1, D_2, D_3, D_4\}$, is equivalent to RLEs $\mathrm{A}_1 + \mathrm{B}_1 + \mathrm{C} \longrightarrow \mathrm{D}_1$, $\mathrm{A}_1 + \mathrm{B}_2 + \mathrm{C} \longrightarrow \mathrm{D}_2$, $\mathrm{A}_2 + \mathrm{B}_1 + \mathrm{C} \longrightarrow \mathrm{D}_3$, and $\mathrm{A}_2 + \mathrm{B}_2 + \mathrm{C} \longrightarrow \mathrm{D}_4$. However, automatically resolving the concrete RLEs is not possible since it is not possible to know which concrete outputs correspond to which concrete inputs without manual inspection. In the previous example, it is not known that D_1 corresponds to A_1 and B_1, unless **D**, **A**, and **B** are manually inspected. Also, resolving the concrete RLEs will lead to combinatorial explosion. Therefore, entity sets are used as is, and the definition of causation is extended from single variables to sets of variables:

Definition 5.1. *A set* **A** *is said to be a* cause *of another set* **B** *when there is a variable A in* **A** *and a variable B in* **B** *such that A is a cause of B.*

Since it is unknown to which isoform(s) a protein EWAS corresponds to, it is assumed that it corresponds to *all* isoforms in UniProt and protein EWAS are treated as sets of isoform EWAS. Finally, candidate members of candidate sets are ignored. The resulting graph contained 12661 nodes and 50245 edges.

5.2.9 Merging the graphs

The graphs from each conceptual data source are merged in order to create a general graph, which contained 156186 nodes and 215677 edges.

5.2.10 General-graph post-processing

The general graph goes through three post-processing steps in order to further further its connectivity, resulting in the processed general graph.

5.2.10.1 Addition of derivation edges

In order to make sure that all gene-product nodes are connected to their corresponding gene nodes, edges are added from each newly-translated-isoform node (which is already connected to the corresponding gene node) to each of its *derivative* nodes. A derivative of a newly-translated isoform is defined as another (post-translationally modified and/or transported) version of (part of) the isoform, or a complex, poly-

mer, or set that contains some version of the isoform. 65647 edges, each citing a different derivation reference, were added, resulting in a graph with 281324 edges.

5.2.10.2 Addition of set-membership edges

Set variables are instant effects of their members. Therefore, edges are added to each set node from each of its member nodes in the graph. 2619 edges, each citing a different set-membership reference, were added, resulting in a graph with 283943 edges.

5.2.10.3 Addition of subset-relationship edges

Set variables are instant effects of their subsets. Therefore, edges are added to each set node from each of its subset nodes in the graph. 615 edges, each citing a different subset-relationship reference, were added, resulting in a graph with 284558 edges. It is worth noting that set-membership and subset-relationship edges are not found even in Reactome's pathway browser.

5.2.11 Adaptation of the post-processed general graph to prion disease

The post-processed general graph is cell-type-agnostic and the phenotype node is an orphan. The graph is adapted for variant filtering in prion disease in three steps.

5.2.11.1 Addition of protein-phenotype edges

Gene knockout experiments in model organisms can elucidate genotype–phenotype relationships. The causal knowledge corresponding to an experiment showing that knocking gene G out affects phenotype P is that G is a cause of P. This is not very useful, though, since the connectivity of G in the graph is limited. Therefore, is it assumed that I is a cause of P for each protein-isoform version I of G.

As mentioned in the introduction, knocking PRNP out in mice prevents prion disease. Therefore, edges are added from each protein node corresponding to a version of PrP to the phenotype node. Three edges, each citing a different protein-phenotype-causation reference, were added, resulting in a graph with 284561 edges.

5.2.11.2 Deletion of nodes by compartment

Although prion disease affects multiple cell types, including neurons and glial cells, the focus of this work is on neurons. Therefore, localised nodes with a compartment not found in neurons were removed. Table 5 and 6 in *Supplementary Information I – Cellular Compartments* lists the compartments in the general CPKG and the removed compartments, respectively. Reactome complexes and Reactome polymers are removed if any of their components and repeated units, respectively, are removed. Reactome entity sets are removed if all their members are removed. 190 nodes were removed in total, resulting in a graph with 155996 nodes and 283496 edges.

5.2.11.3 Deletion of nodes by gene expression

A set of 131 RNA-seq datasets, each measuring the gene expression of 22085 genes in a different healthy human cortical neuron, was obtained from the *Gene Expression Omnibus* database (accession number: GSE67835). The neurons were obtained from tissue that was deemed to be healthy during brain surgery of patients with medical refractory seizures [Darmanis et al., 2015]. Following the recommendation of the *edgeR* software's manual,[11] a gene was deemed to be non-expressed if it had a CPM (counts per million) value of 1.[12] On average, 16560 genes were deemed non-expressed in each neuron. 1261 genes were deemed non-expressed in all neurons. Their HGNC symbols in the datasets were converted to 1140 Ensembl gene stable IDs that were used to delete nodes from the CPKG. Gene products (transcripts, proteins, protein isoforms, miRNA hairpins, and miRNA) are removed if all their associated Ensembl gene stable IDs are among the Ensembl gene stable IDs of the non-expressed genes. Reactome complexes, Reactome polymers, and Reactome entity sets are removed as described in the previous section. 3196 nodes and 36283 edges were removed, resulting in a the final CPKG having 152800 nodes and 247213 edges. Ideally, case–control gene-expression datasets from human prion

[11] http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

[12] The state-of-the-art method of Hart et al. [2013] for detecting non-expressed genes could not be used because it uses FPKM (fragments per kilobase of exon per million reads mapped) values, which are not available.

disease should have been used, as a gene may be not expressed in controls but expressed in cases. However, no such datasets are available.

5.3 Evaluation

In order to evaluate a variant-filtering approach, some performance measure must be chosen. It is important that the method is able to identify *novel* disease variants, not merely re-discover known ones. Therefore, the performance measure of interest is the increase in power of discovering novel disease variants from a filtered dataset compared to the original, unfiltered dataset. There are two main approaches to estimating this measure. The first approach is to apply the method to a disease or set of diseases using the most recent data sources and perform functional studies to determine which of the discovered novel variants are true positives [Aerts et al., 2006]. This is, of course, laborious and expensive. The second approach is to "mimic" novel discoveries by using *archived* data sources and determine whether filtering enables the discovery of disease variants reported *afterwards* in the literature. This can be done prospectively [Börnigen et al., 2012] or retrospectively [Aerts et al., 2006].

In order for a disease to be used for the evaluation of the CPKG-based variant-filtering method using the second approach, the disease must satisfy two requirements: It must be influenced by at least one molecular entity, because otherwise the phenotype would have no ancestors in the CPKG, and have some recently-reported disease variants, because otherwise data sources used in building the CPKG may be unavailable. Prion disease cannot be used because there is only one known disease variant, PRNP codon 129, which was reported back in 1991 as a susceptibility variant for sCJD [Palmer et al., 1991]. In contrast, Parkinson's disease satisfies both requirements, as it is is influenced by the levels of the α-*synuclein* protein [Martin et al., 2011] and has a number of variants reported relatively recently in the *ClinVar* database.[13] Unfortunately, evaluation using genetic datasets from Parkinson's disease could not be performed due to data-access restrictions. In any case, evalu-

[13] http://www.ncbi.nlm.nih.gov/clinvar/

ation can be performed in the future by studying the function of the novel variants discovered from the filtered datasets from prion disease (see Section 5.5).

5.4 Suggesting causal mechanisms

Once a discovery is made from the filtered dataset, the causal paths from the associated nodes to the phenotype can be inspected in the CPKG (see Section 5.1.2); every such path suggests a causal biological mechanism. However, finding all paths between two nodes in a large directed graph is computationally intractable [specifically, the problem is #P-complete; Valiant, 1979] and finding a arbitrary set of $k \geq 1$ paths is not satisfactory. Another approach is to inspect *sequences of edges* from each associated node to the phenotype; a sequence of edges is the multiset of edges that corresponds to an ordered multiset of nodes such that, for $2 \leq i \leq n$, X_{i-1} and X are adjacent. Clearly, a path is a sequence of edges whose corresponding multiset of nodes is a set. The nodes on a sequence of edges from X to Y are easily identifiable, since they comprise the intersection of the descendants of X and the ancestors of Y. However, sequences of edges are hardly useful compared to paths: in the sequence of edges $X \to Y \to Z \to Y \to P$, where X is a node associated with the hit and P is the phenotype, Z is irrelevant since path $X \to Y \to P$ already suggests a causal mechanism. Therefore, highlighting sequences of edges in the CPKG will result in a lot of noise and make the graph uninterpretable. Owing to these problems, investigation of hits using the CPKG was not attempted in this work.

5.5 Results

The prion-disease CPKG had 33343 weakly-connected components. This means that the graph actually consisted of 33343 isolated graphs. Most of these graphs were gene-transcript-protein graphs generated from Ensembl, with none of the entities participating in Reactome reactions. There were 68494 nodes in the weakly connected component that contained the phenotype. 33897 (49.49%) of those nodes (22.18% of all nodes) were ancestors of the phenotype. The fact that not all nodes weakly connected to the phenotype were its ancestors demonstrates that using a directed graph results in higher specificity than using an undirected one.

The genomic regions associated with the ancestors of the phenotype were used to filter the exome-sequencing datasets in Table 3.2. PRNP codon 129 was retained in both datasets. Table 5.11 contains the results of filtering. The reduction in the number of variants was $\approx 70\%$. In contrast, filtering using the genomic regions associated with all nodes in the CPKG (that is, all regions for which there is some prior knowledge) resulted in a reduction of only $\approx 6\%$. This demonstrates that filtering using the causal ancestors of the phenotype is not trivial, as it results in a much smaller dataset than simply using all regions for which there is some prior knowledge.

A well-established gene-prioritisation tool, *Endeavour* [Aerts et al., 2006], was also used to filter the datasets. Endeavour accepts a set of seed genes and a set of candidate genes as its input. For each of several data sources, the seed genes are used to train a model for ranking candidate genes according to their similarity to the seed genes which is then used to rank the candidate genes. The rankings from each data source are combined using order statistics in order to create the overall ranking. Endeavour was used to identify candidate genomic regions as follows. PRNP and the known standard genes in Ensembl (56091 in total) were given as the sole seed gene and the set of candidate genes, respectively. 19515 genes were considered valid and ranked by Endeavour. Because the output of Endeavour consists of HGNC symbols, the symbols were mapped back to Ensembl gene stable IDs as described in Section 5.2.6.7. 18789 genes had an Ensembl gene stable ID, and the top 10% of them was selected (1879 genes in total). The corresponding gene regions were subsequently obtained from Ensembl and used to filter the datasets, resulting, as expected, in a reduction of $\approx 89\%$.

Algorithm 13 was applied to the datasets filtered by the CPKG and Endeavour. Table 5.12 and 5.13 lists the discoveries made from each dataset filtered by the CPKG and Endeavour, respectively. In sCJD, there were no discoveries from the CPKG-filtered dataset and four discoveries from the Endeavour-filtered dataset. rs1071727 was discovered from both the vCJD and sCJD Endeavour-filtered datasets. In vCJD, there were 17 discoveries from the CPKG-filtered

dataset and 16 discoveries from the Endeavour-filtered dataset. rs201076736 and rs11558171 were not discovered in the original dataset but were discovered from both filtered datasets. rs78810484 was discovered from the original and the CPKG-filtered dataset. Finally, rs150910818 was discovered from the original and both filtered datasets. Overall, there was little overlap between the discoveries from the three datasets. However, the discoveries made from the CPKG-filtered dataset were, on average, more significant than the ones made from the Endeavour-filtered dataset, and slightly less significant, on average, than the ones made from the original dataset: the mean discovery link-absence p-value was $2.24 \cdot 10^{-8}$, $5.82 \cdot 10^{-8}$, and $2.85 \cdot 10^{-7}$ in the original, CPKG-filtered, and Endeavour-filtered vCJD dataset, respectively. Note that the link-absence p-value of the same discovery varies across the datasets, as it depends on the other variables in the dataset. A link-absence p-value is smaller in datasets with fewer variables, as fewer tests are performed. This does not explain, however, the smaller p-values in the CPKG-filtered vCJD dataset, as it is much bigger than the Endeavour-filtered vCJD dataset.

#	Dis.	m	m (C)	r (C)	m (CA)	r (CA)	m (E)	r (E)
1	sCJD	381354	113088	70.35%	359923	5.62%	41993	88.99%
2	vCJD	337069	99188	70.57%	317862	5.70%	36457	89.18%

Table 5.11: Exome-sequencing case–control datasets from prion disease filtered by the CPKG, the CPKG using all nodes, and Endeavour. Dis. stands for disease. *m* denotes the number of variants in the original dataset. *m (X)* denotes the number of variants in the dataset filtered by method X. C refers to the CPKG, CA refers to the CPKG using all nodes, and E refers to Endeavour. *r (X)* denotes the reduction in the number of variants in the dataset filtered by method X compared to the original dataset.

5.6 Related work

Any gene-prioritisation method can be used for the purpose of variant filtering in the same way as Endeavour was used in the previous section. Methods that prioritise or filter variants based on their estimated deleteriousness can be used alternatively or as a post-processing step after candidate genomic regions have been identified by the CPKG-based or some other variant-filtering method. Another body of work [Pattin and Moore, 2008, Bush et al., 2009, Emily et al., 2009, Ritchie, 2011] focusses

on using biological databases to filter *pairs* of variants before epistasis detection; a pair of variants is retained if a "strong" connection between the corresponding genes is supported by the databases, where the strength of the connection is calculated based on some scoring scheme. In contrast to the links in a CPKG, these connections do not have a causal interpretation. In addition, as mentioned in Section 5.2.8.5, methods like these may ignore "weak" connections that have a causal interpretation nonetheless. Other works are related to certain aspects of CPKG-based variant filtering. *ComPPI* [Veres et al., 2014] is a PPI database created by filtering binary PPIs from other databases based on the compartments of the interacting proteins. The latter are obtained from several sources, but they are mapped to only six major compartments (cytosol, nucleus, mitochondrion, secretory-pathway, membrane, extracellular). Furthermore, only interactions with participants in the same compartment are retained. In at least one publication [Jonsson et al., 2006] PPIs are filtered using a list of adjacent compartments, but only 6 compartments are used (Extracellular, Intracellular, Cytoplasm, Nucleus, Mitochondrion, Membrane). Schaefer et al. [2013] devised a method for adapting a general PPI network to a certain context (e.g. disease, tissue) using gene expression, functional and disease annotations, and pathways. Finally, Novershtern et al. [2011] introduced *physical module networks (PMNs)* and devised a search-and-score approach to learn them from data. A PMN is the combination of a *module network* and a *physical interaction network* that are consistent with each other. The former is a BN over sets of co-expressed genes (referred to as *modules*). The latter is a graph over genes and proteins with three types of edges: (1) undirected protein-protein edges corresponding to protein-protein interactions, (2) directed edges from proteins to genes corresponding to protein-DNA interactions, and (3) directed edges from genes to their protein products corresponding to transcription interactions. The two are consistent if for each gene G in a module $\mathbf{M}_1$ that is a parent of module $\mathbf{M}_2$ in the BN there is a path from the protein product of G to a transcription factor for all genes in $\mathbf{M}_2$ in the physical interaction network. This is reminiscent to retaining a variant if there is a (causal) path from a node associated with the variant to the phenotype in

the CPKG. Also note that the transcription edges in the physical interaction network are shortcuts to the transcription and translation paths in the CPKG.

5.7 Summary and future work

A variant-filtering method based on the use of a CPKG was developed. The main strength of the method is the causal interpretation of the results: for each variant in the filtered dataset, there is at least one possible disease-causing mechanism, corresponding to a causal path from a node associated with the variant to the phenotype. The method was used to filter two exome-sequencing datasets from prion disease and causal discovery was subsequently applied to the filtered datasets as well as to the datasets filtered by the Endeavour gene-prioritisation tool. Notably, the discoveries made from the former datasets were more significant than the discoveries made from the latter datasets.

Future work includes devising a method for identifying a suitable subset of the causal paths between a node associated with a variant and the phenotype, as finding all such paths is intractable in practice. In addition, CPKG-based variant-filtering needs to evaluated using either of the approaches discussed in Section 5.3 in order to prove its effectiveness.

D#	RS#	C#	Position	M(A)	M(U)	GF (A)	GF (U)	P-value	OR (Aa)	OR (Aa) 95% CI	OR (aa)	OR (aa) 95% CI
2	rs72909030	1	1719393	0.462	0.042	0.076/0.924/0.000	0.916/0.084/0.000	$4.33 \cdot 10^{-36}$	132.62	[58.27, 301.81]	N/A	[N/A, N/A]
2	rs3749073	2	230910379	0.065	0.393	0.882/0.106/0.012	0.276/0.663/0.061	$1.46 \cdot 10^{-12}$	0.05	[0.02, 0.11]	0.06	[0.01, 0.48]
2	rs2911272	3	195783902	0.553	0.143	0.085/0.723/0.191	0.788/0.138/0.074	$6.99 \cdot 10^{-12}$	48.40	[22.17, 105.65]	24.07	[9.73, 59.56]
2	rs200075071	4	102905612	0.488	0.041	0.023/0.977/0.000	0.918/0.082/0.000	$2.46 \cdot 10^{-10}$	469.15	[111.85, 1967.76]	N/A	[N/A, N/A]
2	rs2015018	5	130185433	0.700	0.220	0.082/0.435/0.482	0.709/0.142/0.149	$4.55 \cdot 10^{-15}$	26.43	[10.39, 67.24]	27.95	[11.10, 70.42]
2	rs201076736	8	100709589	0.342	0.032	0.316/0.684/0.000	0.936/0.064/0.000	$1.77 \cdot 10^{-14}$	31.44	[18.40, 53.74]	N/A	[N/A, N/A]
2	rs79075295	11	108244022	0.379	0.029	0.242/0.758/0.000	0.942/0.058/0.000	$7.76 \cdot 10^{-8}$	50.78	[26.35, 97.85]	N/A	[N/A, N/A]
2	rs78562467	12	11031123	0.494	0.055	0.012/0.988/0.000	0.891/0.109/0.000	$1.59 \cdot 10^{-7}$	683.79	[93.40, 5005.94]	N/A	[N/A, N/A]
2	rs78810484	13	36438729	0.471	0.072	0.059/0.941/0.000	0.856/0.144/0.000	$6.91 \cdot 10^{-31}$	95.33	[36.73, 247.45]	N/A	[N/A, N/A]
2	rs201152813	13	75289109	0.322	0.037	0.356/0.644/0.000	0.925/0.075/0.000	$5.35 \cdot 10^{-9}$	22.35	[13.05, 38.26]	N/A	[N/A, N/A]
2	rs757569576	13	19432576	0.000	0.566	1.000/0.000/0.000	0.434/0.000/0.566	$2.41 \cdot 10^{-7}$	N/A	[N/A, N/A]	0.00	[0.00, N/A]
2	rs75554647	15	72752488	0.330	0.054	0.340/0.660/0.000	0.894/0.105/0.002	$4.10 \cdot 10^{-8}$	16.57	[10.16, 27.03]	0.00	[0.00, N/A]
2	rs11635870	15	78771801	0.352	0.039	0.443/0.409/0.148	0.936/0.048/0.015	$1.24 \cdot 10^{-8}$	17.88	[9.37, 34.12]	20.44	[7.36, 56.82]
2	rs62028647	16	55828779	0.368	0.032	0.263/0.737/0.000	0.936/0.064/0.000	$1.97 \cdot 10^{-10}$	41.27	[23.73, 71.77]	N/A	[N/A, N/A]
2	rs62050978	16	70148487	0.335	0.026	0.330/0.670/0.000	0.949/0.051/0.000	$3.71 \cdot 10^{-7}$	37.46	[21.70, 64.68]	N/A	[N/A, N/A]
2	rs11558171	16	58734225	0.065	0.449	0.913/0.043/0.043	0.164/0.774/0.062	$8.22 \cdot 10^{-8}$	0.01	[0.00, 0.04]	0.13	[0.03, 0.64]
2	rs150910818	17	16165082	0.500	0.044	0.000/1.000/0.000	0.912/0.088/0.000	$1.39 \cdot 10^{-51}$	∞	[N/A, ∞]	N/A	[N/A, N/A]

Table 5.12: Discoveries resulting from the application of Algorithm 13 to the datasets in Table 5.11 filtered by the CPKG. *D#* is the dataset number. *RS#* is the RS number of the variant in dbSNP. *C#* is the chromosome number. *M(A/U)* is the minor allele frequency in cases/controls. *GF (A/U)* are the relative frequencies of homozygotes for the common allele, heterozygotes, and homozygotes for the rare allele in cases/controls. *P-value* is (an upper bound of) the link-absence p-value. *OR (Aa)* and *OR (aa)* is the heterozygote and rare homozygote, respectively, OR, with *N/A* denoting an undefined OR. *OR (Aa) 95% CI* and *OR (aa) 95% CI* are the corresponding 95% confidence intervals. The ORs equal the respective CORs under the conditions of Theorem 3.14.

D#	RS#	C#	Position	M(A)	M(U)	GF (A)	GF (U)	P-value	OR (Aa)	OR (Aa) 95% CI	OR (aa)	OR (aa) 95% CI
1	rs200496974	1	155868733	0.007	0.107	0.986/0.014/0.000	0.785/0.215/0.000	$5.87 \cdot 10^{-9}$	0.05	[0.02, 0.16]	N/A	[N/A, N/A]
1	rs764447374	3	10046720	0.101	0.024	0.797/0.203/0.000	0.952/0.048/0.000	$2.60 \cdot 10^{-7}$	5.05	[2.89, 8.82]	N/A	[N/A, N/A]
1	rs752229614	4	150315602	0.129	0.283	0.748/0.248/0.005	0.441/0.553/0.006	$3.03 \cdot 10^{-7}$	0.26	[0.19, 0.38]	0.45	[0.05, 4.07]
1	rs1071727	12	124911942	0.478	0.184	0.473/0.097/0.430	0.788/0.056/0.156	$3.40 \cdot 10^{-7}$	2.90	[1.21, 6.98]	4.59	[2.69, 7.83]
2	rs7579903	2	188997371	0.151	0.451	0.733/0.233/0.035	0.187/0.724/0.089	$1.40 \cdot 10^{-7}$	0.08	[0.04, 0.15]	0.10	[0.03, 0.36]
2	rs73205842	3	196803142	0.134	0.435	0.732/0.268/0.000	0.129/0.871/0.000	$9.05 \cdot 10^{-14}$	0.05	[0.03, 0.09]	N/A	[N/A, N/A]
2	rs12330369	3	10048039	0.412	0.147	0.175/0.825/0.000	0.705/0.295/0.000	$1.97 \cdot 10^{-6}$	11.26	[6.50, 19.52]	N/A	[N/A, N/A]
2	rs77470308	3	196803055	0.165	0.012	0.670/0.330/0.000	0.975/0.025/0.000	$9.94 \cdot 10^{-8}$	19.44	[10.08, 37.48]	N/A	[N/A, N/A]
2	N/A	4	139730430	0.704	0.048	0.296/0.000/0.704	0.951/0.002/0.048	$3.96 \cdot 10^{-7}$	0.00	[0.00, N/A]	47.62	[25.50, 88.91]
2	N/A	6	29942916	0.041	0.547	0.942/0.035/0.023	0.145/0.615/0.240	$1.22 \cdot 10^{-10}$	0.01	[0.00, 0.03]	0.01	[0.00, 0.06]
2	rs201076736	8	100709589	0.342	0.032	0.316/0.684/0.000	0.936/0.064/0.000	$7.19 \cdot 10^{-20}$	31.44	[18.40, 53.74]	N/A	[N/A, N/A]
2	rs1071727	12	124911942	0.758	0.184	0.079/0.326/0.596	0.788/0.056/0.156	$2.60 \cdot 10^{-12}$	58.78	[22.31, 154.84]	38.19	[16.31, 89.42]
2	rs747489126	13	32332846	0.242	0.017	0.515/0.485/0.000	0.966/0.034/0.000	$3.84 \cdot 10^{-7}$	26.70	[14.91, 47.82]	N/A	[N/A, N/A]
2	rs11558171	16	58734225	0.065	0.449	0.913/0.043/0.043	0.164/0.774/0.062	$8.22 \cdot 10^{-8}$	0.01	[0.00, 0.04]	0.13	[0.03, 0.64]
2	rs150910818	17	16165082	0.500	0.044	0.000/1.000/0.000	0.912/0.088/0.000	$5.49 \cdot 10^{-48}$	∞	[N/A, ∞]	N/A	[N/A, N/A]
2	rs200691513	18	31545954	0.320	0.033	0.361/0.639/0.000	0.934/0.066/0.000	$1.73 \cdot 10^{-21}$	24.92	[14.89, 41.71]	N/A	[N/A, N/A]
2	rs200621862	18	23939340	0.278	0.032	0.443/0.557/0.000	0.935/0.065/0.000	$4.73 \cdot 10^{-13}$	18.08	[10.90, 29.98]	N/A	[N/A, N/A]
2	rs56131056	19	603634	0.111	0.442	0.811/0.156/0.033	0.253/0.610/0.136	$5.63 \cdot 10^{-8}$	0.08	[0.04, 0.16]	0.08	[0.02, 0.27]
2	rs202094581	21	14381534	0.153	0.026	0.693/0.307/0.000	0.947/0.053/0.000	$1.15 \cdot 10^{-6}$	7.98	[4.49, 14.20]	N/A	[N/A, N/A]
2	rs200593805	22	39670868	0.167	0.016	0.667/0.333/0.000	0.968/0.032/0.000	$2.78 \cdot 10^{-7}$	14.90	[7.68, 28.93]	N/A	[N/A, N/A]

Table 5.13: Discoveries resulting from the application of Algorithm 13 to the datasets in Table 5.11 filtered by Endeavour. *D#* is the dataset number. *RS#* is the RS number of the variant in dbSNP, with *N/A* denoting that the variant is not in dbSNP. *C#* is the chromosome number. *M(A/U)* is the minor allele frequency in cases/controls. *GF (A/U)* are the relative frequencies of homozygotes for the common allele, heterozygotes, and homozygotes for the rare allele in cases/controls. *P-value* is (an upper bound of) the link-absence p-value. *OR (Aa)* and *OR (aa)* is the heterozygote and rare homozygote, respectively, OR, with *N/A* denoting an undefined OR. *OR (Aa) 95% CI* and *OR (aa) 95% CI* are the corresponding 95% confidence intervals. The ORs equal the respective CORs under the conditions of Theorem 3.14.

Chapter 6

Summary and Future Work

Although GWASs resulted in the discovery of thousands of variant-disease associations, the vast majority of genetic causes of disease remain to be found. Genetic studies in prion disease have, so far, not lead to any discoveries other than the known PRNP codon 129. A general approach to discovering additional susceptibility variants is to perform data integration. In this work, INCA was adopted as a data-integration framework for identifying all causal relationships between variants and prion disease that are consistent with all genetic datasets from prion disease and prior biological knowledge. Towards that goal, a theory of causal discovery from genetic datasets was formulated, algorithms for causal discovery from single and multiple genetic datasets were devised, and a variant-filtering method based on a CPKG that represents causal relationships from several biological data sources was developed. The methods were applied to datasets from prion disease, resulting in the discovery of variants to be further investigated.

The FDR-controlled local-learning algorithm devised for conditional genetic samples does not currently output the orientations of the genotype–phenotype edges. As differentiating between causal variants, potentially-causal variants and variants that are merely indicators of hidden causal variants is of utmost importance for prioritising the variants for further investigation, it is highly desirable to extend the algorithm to perform this function while taking multiple testing into account. Furthermore, PIRs, which are a case of epistasis with absent marginal effects, and information equivalences, which are both violations of the CFC, are not taken into

account by the algorithm, possibly resulting in links missing from the output. Future extensions of the algorithm should be able to handle these violations and may uncover additional susceptibility variants in prion disease and other diseases.

The performance of the algorithm was evaluated using simulated exome-sequencing datasets of a fixed size over chromosome 22 with all causal variants being observed. Future simulation studies should use exome-sequencing and GWAS datasets of various sizes over all autosomes, generated using disparate disease models and excluding some of the causal variants.

As in the single-dataset case, the FDR-controlled local-learning algorithm for multiple conditional genetic samples can be extended to estimate the orientation of the edges and deal with unfaithfulness. A simulation study of the performance of the algorithm should be conducted. It would be also interesting to compare the performance of the multiple-dataset algorithm to that of the single-dataset algorithm applied to the concatenation of the datasets, followed by genotype imputation.

The CPKG-based variant-filtering approach was shown to result in more significant discoveries from the filtered datasets compared to using the well-established Endeavour gene-prioritisation tool. Although the approach is able to suggest possible disease-causing mechanisms for the discovered variants in principle by finding all causal paths from the nodes associated with a discovery to the phenotype, finding all such paths is intractable in practice. Therefore, additional work is needed on identifying a suitable subset of the paths. Moreover, the filtering approach was not evaluated by either performing functional studies of the discovered variants or mimicking novel discoveries using archived data sources. In order to prove the effectiveness of the approach, either type of evaluation should be conducted.

The causal-discovery algorithms developed here can be applied to any phenotype, not just prion disease. The algorithms may be also applied, with minor modifications, to other types of cross-sectional datasets as well (e.g. gene-expression datasets). Furthermore, the process of adapting the causal-discovery theory and algorithms to genetic sets of variables can serve as an example for developing domain-specific algorithms in other domains. In addition, the CPKG-based filtering method

can be applied to any disease with at least one known molecular cause, e.g. Parkinson's disease. It is hoped that the methods devised here, and causal discovery in general, will play a role not only in the discovery and interpretation of susceptibility variants in prion disease and other diseases, but in elucidating genotype–phenotype relationships in general.

Appendix A

Proofs

Chapter 3

Theorem 3.1

The proof of Theorem 3.1 is based on the following lemmas. The first two lemmas give necessary and sufficient conditions for the existence of the different types of links in a plausible genetic causal MAG. The term "underlying causal structure" is used to refer to any plausible genetic causal DAG whose marginal is the plausible genetic causal MAG at hand. Inducing paths in the underlying causal structure are with respect to the nodes in the underlying causal structure that are not in the plausible genetic causal MAG, and $\emptyset$. Genetic chains are relative to the variables of the plausible genetic causal MAG. In a causal DAG over $\mathbf{O}\dot{\cup}\mathbf{H}$, a *hidden-cause path* between X and Y ($X, Y \in \mathbf{O}\dot{\cup}\mathbf{H}$) relative to $\mathbf{O}$ is a path of the form $X \leftarrow \cdots \leftarrow H \rightarrow \cdots \rightarrow Y$ for some $H \in \mathbf{H}$.

Lemma A.1. *Let $\mathbb{M}$ be a plausible genetic causal MAG over $\mathbf{G} \cup \{P\}$ and $G \in \mathbf{G}$. G and P are adjacent in $\mathbb{M}$ if and only if one of the following holds:*

1. *G is a cause of P.*
2. *G is an indicator of a hidden cause of P.*

Proof.

Forward direction: Due to Theorem 2.4, there is an inducing path p between G and P in the underlying causal structure. There are two cases:

1. *There are no colliders on p:* Then either G is a cause of P, P is a cause of G, or P and G have a hidden common cause. The second possibility is ruled out by Assumption 3.1. If P and G have a hidden common cause H, then there is some genotype on $p(H,P)$ due to Assumption 3.4. Therefore, unless Condition (1) is satisfied, Condition (2) is satisfied.

2. *There are $m \geq 1$ colliders on p:* It will be first proved by induction on $1 \leq i \leq m$ that, for each collider $C^{(i)}$ on p, where $C^{(1)}$ is the collider closer to G, either Condition (1) or (2) is satisfied, (3) $C^{(i)}$ is an ancestor of G, or (4) G is not a cause of P and $C^{(i)}$ is an ancestor of some observed cause $G^{(k)}$ $(k \geq 1)$ of P that is genetically chained to G.

   Base case: Since selection bias is absent, $C^{(1)}$ is an ancestor of either G or P. If $C^{(1)}$ is an ancestor of G, then Condition (3) is satisfied for $C^{(1)}$. Suppose that $C^{(1)}$ is an ancestor of P and let $p^{(1)}$ be a directed path from $C^{(1)}$ to P. If p is out of G, then $[p(G,C^{(1)}),p^{(1)}]$ is a directed path from G to P. That is, Condition (1) is satisfied. If p is into G, then G and P have some hidden common cause $H^{(1)}$ that is an interior node on $p(G,C^{(1)})$. Owing to Assumption 3.4, there is some $G^{(1)}$ on $[p(H^{(1)},C^{(1)}),p^{(1)}]$. If $G^{(1)}$ is hidden, then Condition (2) is satisfied unless Condition (1) is satisfied. If $G^{(1)}$ is observed, then $G^{(1)}$ is on $p^{(1)}$. Therefore, Condition (4) is satisfied for $C^{(1)}$, unless Condition (1) is satisfied.

   Inductive step: Since selection bias is absent, $C^{(i)}$ is an ancestor of either G or P. If $C^{(i)}$ is an ancestor of G, then Condition (3) is satisfied for $C^{(i)}$. Suppose that $C^{(i)}$ is an ancestor of P and let $p^{(i)}$ be a directed path from $C^{(i)}$ to P. If Condition (3) is satisfied for $C^{(i-1)}$, then G and P have some hidden common cause $H^{(i)}$ that is an interior node on $p(C^{(i-1)},C^{(i)})$. Owing to Assumption 3.4, there is some $G^{(1)}$ on $[p(H^{(i)},C^{(i)}),p^{(i)}]$. If $G^{(1)}$ is hidden, then Condition (2) is satisfied unless Condition (1) is satisfied. If $G^{(1)}$ is observed, then $G^{(1)}$ is on $p^{(i)}$. Therefore, Condition (4) is satisfied for $C^{(i)}$ unless Condition (1) is satisfied. If Condition (4) is satisfied for $C^{(i-1)}$, then $G^{(k)}$ and P have some hidden common cause $H^{(k+1)}$ that is an interior node on $p(C^{(i-1)},C^{(i)})$. Due

to Assumption 3.4, there is some $G^{(k+1)}$ on $[p(H^{(k+1)}, C^{(i)}), p^{(i)}]$. If $G^{(k+1)}$ is hidden, then Condition (2) is satisfied. If $G^{(k+1)}$ is observed, then $G^{(k+1)}$ is on $p^{(i)}$. Therefore, Condition (4) is satisfied for $C^{(i)}$.

If Condition (3) is satisfied for $C^{(m)}$, then G and P have some hidden common cause $H^{(1)}$ that is an interior node on $p(C^{(m)}, P)$ due to Assumption 3.1. Assumption 3.4 therefore implies that there is some hidden genotype on $p(H^{(1)}, P)$, which means that Condition (2) is satisfied unless Condition (1) is satisfied. If Condition (4) is satisfied for $C^{(m)}$, then $G^{(k)}$ and P have some hidden common cause $H^{(k+1)}$ that is an interior node on $p(C^{(m)}, P)$. Owing to Assumption 3.4, there is some hidden genotype on $p(H^{(k+1)}, P)$. Therefore, Condition (2) is satisfied.

Reverse direction: There are two cases:

1. *Condition (1) is satisfied:* Let p be a directed path from G to P in the underlying causal structure. Owing to Assumption 3.2, there cannot be any interior nodes on p that are genotypes, which means that all interior nodes on p are hidden. Therefore, p is an inducing path between G and P. Theorem 2.4 then says that G and P are adjacent in $\mathbb{M}$.

2. *Condition (2) is satisfied:* Let G' be a hidden cause of P whose presence is indicated by G, $\{G^{(1)}, \ldots, G^{(n)}\}$ be a genetic chain between G_1 to G', where $G^{(1)} = G$ and $G^{(n)} = G'$, $p^{(i)}$ $(1 \leq i \leq n-1)$ be a hidden-cause path from G_i to G_{i+1} in the underlying causal structure, and p_n be a hidden-cause path from $G^{(n)}$ to P in the underlying causal structure. There cannot be any interior nodes on $p^{(i)}$ that are genotypes due to Assumption 3.2, which means that all interior nodes on $p^{(i)}$ are hidden. Owing to Assumption 3.1, there cannot be any interior nodes on p_n that are genotypes, which implies that all interior nodes on p_n are hidden too. Let $X^{(i)}$ $(2 \leq i \leq n-1)$ be the first node on $p^{(i-1)}$ that is also on $p^{(i)}$. Then $[p_1(G, X^{(2)}), \ldots, p_{n-1}(X^{(n-1)}, G^{(n)}), p_n]$ is an inducing path between G and P and Theorem 2.4 implies that G and P are adjacent in $\mathbb{M}$.

□

Lemma A.2. *Let* $\mathbb{M}$ *be a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *and* $G_1, G_2 \in \mathbf{G}$. *Then* G_1 *and* G_2 *are adjacent in* $\mathbb{M}$ *if and only if they have a hidden common cause.*

Proof.

Forward direction: Owing to Theorem 2.4, there is an inducing path p between G_1 and G_2 in the underlying causal structure. There are three cases:

- *There are no colliders on p:* Then either G_1 is a cause of G_2, G_2 is a cause of G_1, or G_1 and G_2 have a hidden common cause. The first two possibilities are ruled out by Assumption 3.2.

- *There are colliders on p and all of them are ancestors of* G_1*:* Let X be the collider closest to G_2 on p and p_1 be a directed path from X to G_1. Then p is into G_2 owing to Assumption 3.2, there is some hidden common cause of G_1 and G_2 that is an interior node on $p(X, G_2)$, and path $[p(G_2, X), p_1]$ satisfies the condition of the lemma.

- *There are colliders on p and some of them are not ancestors of* G_1*:* Let X be the collider closest to G_1 on p that is not an ancestor of G_1. If there are no colliders on p between G_1 and X, let $p_1 = p(G_1, X)$; otherwise, let Y be the collider preceding X on p, p_2 a directed path from Y to G_1, and $p_1 = [p_2(G_1, Y), p(Y, X)]$. Since selection bias is absent, X is an ancestor of G_2. Let p_3 be a directed path from X to G_2. Then path $[p_1, p_3]$ satisfies the condition.

Reverse direction: Let p be a hidden-cause path from G_1 to G_2 in the underlying causal structure. There are no interior nodes on p that are genotypes due to Assumption 3.2. Therefore, all interior nodes on p are hidden, which means that p is an inducing path between G_1 and G_2. Hence, G_1 and G_2 are adjacent in $\mathbb{M}$ due to Theorem 2.4. □

The next two lemmas concern the orientation of the edges in a plausible genetic causal MAG.

Lemma A.3. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*, edges incident to* P *are into* P*.*

Proof. The proof follows from Assumption 3.1 and the absence of selection bias. □

Lemma A.4. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*, genotype–genotype edges are bidirected.*

Proof. The proof follows from Assumption 3.2 and the absence of selection bias. □

The next three lemmas give necessary and sufficient conditions for the existence of the different types of edges in a plausible genetic causal MAG.

Lemma A.5. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*, edge* $G \to P$ *exists if and only if* G *is a cause of* P*.*

Proof.

Forward direction: $G \to P$ implies that G is a cause of P or of some selection variable. The latter possibility is ruled out by the absence of selection bias.

Reverse direction: If G is cause of P, then G and P are adjacent due to Lemma A.1. Owing to Lemma A.3, the edge between G and P is oriented as $G \to P$. □

Lemma A.6. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*, edge* $G \leftrightarrow P$ *exists and only if* G *is an indicator of a hidden cause of* P*.*

Proof.

Forward direction: $G \leftrightarrow P$ implies that G is not a cause of P or of any selection variable. Since selection bias is absent, G is not a cause of P. Owing to Lemma A.1, G is an indicator of a hidden cause of P.

Reverse direction: If G is an indicator of a hidden cause of P, Lemma A.1 says that G and P are adjacent. The edge between G and P is oriented as $G \leftrightarrow P$ due to Lemma A.3. □

Lemma A.7. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*, edge* $G_1 \leftrightarrow G_2$ *exists if and only if* G_1 *and* G_2 *have a hidden common cause.*

Proof.

Forward direction: The proof follows directly from Lemma A.2.

Reverse direction: If G_1 and G_2 have a hidden common cause, then G_1 and P are adjacent due to Lemma A.2. Owing to Lemma A.4, the edge between G_1 and P is bidirected. □

Proof of Theorem 3.1. The proof follows from Lemmas A.3–A.7. □

Proof of Corollary 3.1. If G_1 is a spouse of P, then G_1 is an indicator of a hidden cause of P due to Theorem 3.1. If G_1 is not a proxy of a hidden cause of P, then there is an observed G_2 such that G_2 is a cause of P and G_2 and G_1 have a hidden common cause. Theorem 3.1 therefore implies that G_2 is a parent of P and G_2 and G_1 are adjacent. □

Theorem 3.2

The proof of Theorem 3.2 requires the following lemmas and definition. The first lemma pertains to the genomic location of adjacent genotypes in a plausible genetic causal MAG.

Lemma A.8. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*, if* G_1 *and* G_2 *are adjacent, then* G_1 *and* G_2 *are on the same chromosome.*

Proof. If G_1 and G_2 are adjacent, then G_1 and G_2 have a hidden common cause due to Lemma A.2. Assumption 3.3 therefore implies that G_1 and G_2 are on the same chromosome. □

The second lemma gives sufficient conditions for a potential plausible genetic causal MAG to be a genetic causal MAG.

Lemma A.9. *Let* $\mathbb{M}$ *be a potential plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*.* $\mathbb{M}$ *is a genetic causal MAG if the following two conditions are satisfied:*

1. *Edges incident to P are out of P.*

2. *Genotype–genotype edges are bidirected.*

Proof. Suppose that $\mathbb{M}$ satisfies the conditions above. There are no directed paths from P to any genotype or from a genotype to another in $\mathbb{M}$, since there are no directed genotype–genotype edges or edges out of P. Therefore, there are no directed cycles, almost directed cycles, or primitive inducing paths between any two non-adjacent nodes in $\mathbb{M}$. Furthermore, there are no undirected edges incident to nodes with parents or spouses in $\mathbb{M}$, since there no undirected edges. Thus, $\mathbb{M}$ is a genetic causal MAG due to Lemma 2.1. □

The *canonical genetic causal DAG* of a genetic causal MAG is defined as follows.

Definition A.1 (Canonical genetic causal DAG)**.** *Suppose that $\mathbb{M}$ is a genetic causal MAG over $\mathbf{G} \cup \{P\}$. Then the* canonical genetic causal DAG $\mathbb{D}(\mathbb{M})$ *is the genetic causal DAG that is obtained from $\mathbb{M}$ by the following procedure:*

1. *Initialise $\mathbb{D}(\mathbb{M})$ with the nodes in $\mathbb{M}$.*

2. *For each edge $G \to P$ in $\mathbb{M}$, create edge $G \to P$ in $\mathbb{D}(\mathbb{M})$ and let $p_G = [G, P]$.*

3. *For each edge $G \leftrightarrow P$ in $\mathbb{M}$, create node X_G and path $p_G = G \leftarrow X_G \to P$ in $\mathbb{D}(\mathbb{M})$.*

4. *For each edge $G_1 \leftrightarrow G_2$ in $\mathbb{M}$, create node Y_{G_1,G_2} and path $q_{G_1,G_2} = G_1 \leftarrow Y_{G_1,G_2} \to G_2$ in $\mathbb{D}(\mathbb{M})$.*

The following lemma gives sufficient conditions for a genetic causal MAG to be the marginal of its canonical genetic causal DAG.

Lemma A.10. *Let $\mathbb{M}$ be a genetic causal MAG over $\mathbf{G} \cup \{P\}$. Then $\mathbb{M}$ is the marginal of $\mathbb{D}(\mathbb{M})$ over $\mathbf{G} \cup \{P\}$ if it satisfies the following two conditions:*

1. *Edges incident to P are into P.*

2. *Genotype–genotype edges are bidirected.*

Proof. Let $\mathbf{H}$ be the set of nodes in $\mathbb{D}(\mathbb{M})$ that are not in $\mathbb{M}$. It will be first proved that for each edge between X and Y in $\mathbb{M}$, there is an inducing path between X and Y with respect to $\mathbf{H}$ and $\emptyset$ in $\mathbb{D}(\mathbb{M})$ and the edge endpoints match the ancestral relationships in $\mathbb{D}(\mathbb{M})$.

- For each edge between G and P in $\mathbb{M}$, p_G is an inducing path with respect to $\mathbf{H}$ and $\emptyset$. Condition (1) says that the edge is into P, which agrees with the fact that P is not an ancestor of G in $\mathbb{D}(\mathbb{M})$. If the edge is out of G, then G is an ancestor of P in $\mathbb{D}(\mathbb{M})$; otherwise, G is not an ancestor of P in $\mathbb{D}(\mathbb{M})$.

- For each edge between G_1 and G_2 in $\mathbb{M}$, q_{G_1,G_2} is an inducing path between G_1 and G_2 with respect to $\mathbf{H}$ and $\emptyset$ in $\mathbb{D}(\mathbb{M})$. Owing to Condition (1), the edge is bidirected. This is in accordance with G_1 not being an ancestor of G_2, and G_2 not being an ancestor of G_1 in $\mathbb{D}(\mathbb{M})$.

It will now be proved that for each pair of nodes X and Y that are not adjacent in $\mathbb{M}$, there is no inducing path between X and Y with respect to $\mathbf{H}$ and $\emptyset$ in $\mathbb{D}(\mathbb{M})$.

- If G and P are not adjacent in $\mathbb{M}$, suppose that there is an inducing path p between G and P with respect to $\mathbf{H}$ and $\emptyset$ in $\mathbb{D}(\mathbb{M})$. p is of the form $[G^{(1)},H^{(1)},\ldots,G^{(i)},H^{(i)},\ldots,G^{(n)},H^{(n)},P]$, where $n \geq 2$, $G^{(i)}$ $(2 \leq i \leq n)$ is a collider on p, $G^{(1)} = G$, and $H^{(i)} \in \mathbf{H}$ $(1 \leq i \leq n)$ is a non-collider on p. $G^{(n)}$ is not an ancestor of G or P in $\mathbb{D}(\mathbb{M})$, which is a contradiction. Therefore, there is no inducing path between G and P with respect to $\mathbf{H}$ and $\emptyset$ in $\mathbb{D}(\mathbb{M})$.

- If G_1 and G_2 are not adjacent in $\mathbb{M}$, suppose that there is an inducing path p between G and P with respect to $\mathbf{H}$ and $\emptyset$ in $\mathbb{D}(\mathbb{M})$. There are two cases:

  1. p is of the form $[G^{(1)},H^{(1)},\ldots,G^{(i)},H^{(i)},\ldots,G^{(n)},H^{(n)},P,H^{(n+1)},G^{(n+1)},\ldots,H^{(j)},G^{(j)},\ldots,H^{(n+m)},G^{(n+m)}]$, where $n \geq 1$, $m \geq 1$, $G^{(i)}$ $(1 \leq i \leq n)$, $G^{(j)}$ $(n+1 \leq j \leq n+m)$, and P are colliders on p, $G^{(1)} = G_1$, $G^{(n+m)} = G_2$, and $H^{(i)} \in \mathbf{H}$ $(1 \leq i \leq n)$ and $H^{(j)} \in \mathbf{H}$ $(n+1 \leq j \leq n+m)$ are non-colliders on p. P is not an ancestor of $G^{(1)}$ or $G^{(n+m)}$ in $\mathbb{D}(\mathbb{M})$, which is a contradiction. Therefore, p is not of this form.

2. p is of the form $[G^{(1)}, H^{(1)}, \ldots, G^{(i)}, H^{(i)}, \ldots, G^{(n)}, H^{(n)}, G_2]$, where $n \geq 2$, $G^{(i)}$ $(2 \leq i \leq n)$ is a collider on p, $G^{(1)} = G_1$, and $H^{(i)} \in \mathbf{H}$ $(1 \leq i \leq n)$ is a non-collider on p. G^2 is not an ancestor of $G^{(1)}$ or G_2 in $\mathbb{D}(\mathbb{M})$, which is a contradiction. Thus, p is not of this form either. Hence, there is no inducing path between G and P with respect to $\mathbf{H}$ and $\emptyset$ in $\mathbb{D}(\mathbb{M})$. The proof therefore follows from Theorem 2.4.

□

The following lemma gives a sufficient condition for a canonical genetic causal DAG to be plausible.

Lemma A.11. *Let $\mathbb{M}$ be a genetic causal MAG over $\mathbf{G} \cup \{P\}$. If adjacent genotypes in $\mathbb{M}$ are on the same chromosome, then $\mathbb{D}(\mathbb{M})$ is plausible.*

Proof. There are no directed paths from P to any genotype or from a genotype to another in $\mathbb{D}(\mathbb{M})$. Therefore, $\mathbb{D}(\mathbb{M})$ is a genetic causal DAG that satisfies Assumptions 3.1 and 3.2. If two genotypes G_1 and G_2 have a common ancestor in $\mathbb{D}(\mathbb{M})$, then G_1 and G_2 are adjacent in $\mathbb{M}$. By hypothesis, G_1 and G_2 are on the same chromosome. Thus, $\mathbb{D}(\mathbb{M})$ satisfies Assumption 3.3. Since Assumption 3.4 places no restriction on $\mathbb{D}(\mathbb{M})$, $\mathbb{D}(\mathbb{M})$ is plausible. □

Proof of Theorem 3.2.

Forward direction: Condition (1), (2), and (3) is satisfied for plausible genetic causal MAGs due to Lemma A.3, A.4, and A.8, respectively.

Reverse direction: Owing to Lemma A.9, $\mathbb{M}$ is a genetic causal MAG. Lemma A.11 therefore implies that $\mathbb{D}(\mathbb{M})$ is a plausible genetic causal DAG. Due to Lemma *A*.10, $\mathbb{M}$ is the marginal of $\mathbb{D}(\mathbb{M})$ over $\mathbf{G} \cup \{P\}$. Thus, $\mathbb{M}$ is a plausible genetic causal MAG. □

Theorem 3.3

The proof of Theorem 3.3 is based on the following lemmas. According to the first one, the notion of genetic discriminating path is equivalent to that of discriminating path in a plausible genetic causal MAG:

Lemma A.12. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$, $p = [X_n, \ldots, X_3, X_2, X_1]$ *is a genetic discriminating path for* X_2 *if and only if* p *is a discriminating path for* $[X_3, X_2, X_1]$.

Proof.

Forward direction: If $p = [G_n, \ldots, G_2, G_1, P]$ is a genetic discriminating path for G_1, then G_i $(2 \leq i \leq n-1)$ is a collider on p due to Lemma A.4 and G_i is a parent of P due to Lemma A.3. Therefore, p is a discriminating path for $[G_2, G_1, P]$.

Reverse direction: If $p = [X_n, \ldots, X_3, X_2, X_1]$ is a discriminating path for triple $[X_3, X_2, X_1]$, then X_i $(3 \leq i \leq n-1)$ is a parent of X_1. Therefore, $X_i \in \mathbf{G}$ $(3 \leq i \leq n-1)$ and $X_1 = P$ due to Theorem 3.1. Thus, $X_n, X_2 \in \mathbf{G}$ as well. Hence, p is a genetic discriminating path for X_2. □

The next three lemmas pertain to Markov equivalence of plausible genetic causal MAGs.

Lemma A.13. *If plausible genetic causal MAGs* $\mathbb{M}_1$ *and* $\mathbb{M}_2$ *are Markov equivalent, then an unshielded genotype is a parent of the phenotype in* $\mathbb{M}_1$ *if and only if it is a parent of the phenotype in* $\mathbb{M}_2$.

Proof. The proof follows from Theorems 3.2 and 2.5. □

Lemma A.14. *If two plausible genetic causal MAGs are Markov equivalent, then they have the same genetic discriminating paths.*

Proof. Suppose that plausible genetic causal MAGs $\mathbb{M}_1$ and $\mathbb{M}_2$ are Markov equivalent. If $[G_n, \ldots, G_2, G_1, P]$ is a genetic discriminating path for G_1 in $\mathbb{M}_1$, then there is a path $[G_n, \ldots, G_2, G_1, P]$, G_n is not adjacent to P, and every node G on $p(G_{n-1}, G_2)$ is adjacent to P in $\mathbb{M}_2$ due to Theorem 2.5. $[G_n, \ldots, G_2, G_1, P]$ will be proved to be a genetic discriminating path for G_1 in $\mathbb{M}_2$ by induction on $1 \leq i \leq n-2$:

Base case: G_{n-1} is a parent of P due to Lemma A.13. Therefore, $[G_n, G_{n-1}, G_{n-2}, P]$ is a genetic discriminating path for G_{n-2} in $\mathbb{M}_2$.

Inductive step: Suppose $[G_n, \ldots, G_i, P]$ is a genetic discriminating path for G_i in $\mathbb{M}_2$. Since $[G_n, \ldots, G_i, P]$ is a genetic discriminating path for G_i and G_i is parent

of P in $\mathbb{M}_1$, G_i is parent of P in $\mathbb{M}_2$ due to Lemma A.12 and Theorem 2.5. Therefore, $[G_n, \ldots, G_{i+1}, P]$ is a genetic discriminating path for G_i in $\mathbb{M}_2$. □

Lemma A.15. *If plausible genetic causal MAGs $\mathbb{M}_1$ and $\mathbb{M}_2$ are Markov equivalent, then a genetically-discriminated genotype is a parent of the phenotype in $\mathbb{M}_1$ if and only if it is a parent of the phenotype in $\mathbb{M}_2$.*

Proof. The proof follows from Lemmas A.4 and A.14. □

Proof of Theorem 3.3. The proof follows from Theorem 2.5 and Lemmas A.3, A.4, A.13, A.12, and A.15. □

Theorem 3.4

Proof of Theorem 3.4.

Forward direction: Conditions (1)-(3) are satisfied for maximally-informative plausible genetic causal PAGs due to Theorem 3.2. Condition (4) is satisfied due to Theorem 3.3.

Reverse direction: Let $\mathbb{P}$ be a potential plausible genetic causal PAG that satisfies Conditions (1)–(4). Owing to Theorem 3.2, orienting the genotype–phenotype edges incident to shielded and non-genetically-discriminated genotypes in $\mathbb{P}$ in either way results in a plausible genetic causal MAG. Let $\mathbb{M}_1$ and $\mathbb{M}_2$ be two members of the class of plausible genetic causal MAGs represented by $\mathbb{P}$. Owing to Theorem 3.3, $\mathbb{M}_1$ and $\mathbb{M}_2$ are Markov equivalent, which means that $\mathbb{P}$ is a maximally-informative plausible genetic causal PAG. □

Theorem 3.5

The proof of Theorem 3.5 is based on the following lemmas. The first one gives a necessary condition for strict m-separation.

Lemma A.16. *Suppose that X and Y are strictly m-separated by $\mathbf{Z}$ in a MAG. Then for every $Z \in \mathbf{Z}$ there is a path p that satisfies the following conditions:*

- *Z is a noncollider on p.*
- *There is no $Z' \neq Z$ such that Z' is a noncollider on p.*

- *Every collider on p is in* $\mathbf{Z}$.

Proof.

Suppose that there is some $Z_0 \in \mathbf{Z}$ such that no path that satisfies the conditions above exists. Then X and Y are m-separated by $\mathbf{Z} \setminus \{Z_0\}$, which is a contradiction. □

The requirement for Z to lie on a path between X and Y in Lemma A.16 is called the *necessary path condition* and is used in a variant of the PC algorithm as a means to reduce false negatives [Steck and Tresp, 1999].

The lemmas below characterise m-separation in a plausible genetic causal MAG.

Lemma A.17. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*, if* G_1 *and* P *are strictly m-separated by* $\mathbf{Z}$*, then* $\mathbf{Z}$ *is a subset of the parents of* P *on the same chromosome.*

Proof. If G_1 and P are strictly m-separated by $\mathbf{Z}$, G_1 and P are not adjacent and for each $G_2 \in \mathbf{Z}$ there is a path p from G_1 to P such that G_2 is a noncollider on p due to Lemma A.16. Owing to Theorem 3.2, p is of the form $G_1 \leftrightarrow \cdots \leftrightarrow G_2 \rightarrow P$ and $G_1, \ldots, G_2$ are on the same chromosome. □

Lemma A.18. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*,* G_1 *and* P *are adjacent if and only if they are not m-separated by any subset of the genotypes adjacent to* P *on the same chromosome.*

Proof.

Forward direction: If G_1 and P are adjacent, then they are not m-separated by any subset of the rest variables.

Reverse direction: If G_1 and P are not m-separated by any subset of the genotypes adjacent to P on the same chromosome, then they are not strictly m-separated by any subset of the parents of P on the same chromosome. Owing to Lemma A.17, G_1 and P are not m-separated by any subset of the rest variables, which implies that G_1 and P are adjacent. □

Lemma A.19. *In a plausible genetic causal MAG over* $\mathbf{G} \cup \{P\}$*,* G_1 *and* G_2 *are adjacent if and only if they are not m-separated.*

Proof.

Forward direction: If G_1 and G_2 are adjacent, then they are not m-separated by any subset of the rest variables.

Reverse direction: If G_1 and G_2 are not m-separated, then there is a path without colliders from G_1 to G_2. Theorem 3.2 therefore implies that edge $G_1 \leftrightarrow G_2$ exists. □

Proof of Theorem 3.5. The proof follows from Lemmas A.18 and A.19. □

Theorem 3.6

It will be first proved that Algorithm 9 is sound, in the sense that it only creates orientations in the Markov equivalence class of plausible genetic causal MAGs.

Lemma A.20 (Soundness of Algorithm 9)**.** *If the input of Algorithm 9 is* $\mathbb{P}$ *and* **Sepset***, then in the output of Algorithm 9,* $\mathbb{P}$ *is a plausible genetic causal PAG for the same Markov equivalence class as in the input.*

Proof. The soundness of each rule in Algorithm 9 is proved separately. The proof follows by induction.

Rule (1): Soundness follows directly from Lemma A.3.

Rule (2): When $G_1 \notin \mathbf{Sepset}(\{G_2, P\})$, edge $G_1 \ast\!\!-\!\!\ast\, P$ must be oriented as $G_1 \leftarrow\!\!\ast\, P$ due to Lemma 2.2. Otherwise, triple $[G_2, G_1, P]$ must be oriented either as $G_2 \ast\!\!\rightarrow G_1 \ast\!\!-\!\!\ast\, P$, $G_2 \ast\!\!- G_1 \leftarrow\!\!\ast\, P$, or $G_2 \ast\!\!- G_1 \ast\!\!-\!\!\ast\, P$, again due to Lemma 2.2. The second and third case are ruled out by Assumption 3.2.

Rule (3): Soundness follows directly from Lemma A.4.

Rule (4): Soundness follows from Lemmas A.12 and 2.3. □

The following lemma is used in the proof of Theorem 3.6.

Lemma A.21. *In the output of Algorithm 9, the endpoint at the genotype end of a genotype–phenotype edge is oriented if and only if the genotype is unshielded or discriminated.*

Proof. There are no $P \circ\!\!-\!\!* G$ or $G_1 \circ\!\!-\!\!* G_2$ edges in the output PAG, since they are all oriented by Rules 1 and 3, respectively. Therefore, the only circle endpoints are in $G \circ\!\!\rightarrow P$ edges. G is shielded because otherwise $G \circ\!\!\rightarrow P$ would have been oriented by Rule (2), and non-discriminated because otherwise Rule (2) and successive applications of Rule (4) would have resulted in the orientation of $G \circ\!\!\rightarrow P$. □

Proof of Theorem 3.6. The proof follows from Theorem A.20, Lemma A.21, and Theorem 3.4. □

Theorem 3.7

The proof of Theorem 3.7 is based on the following lemmas. The first two lemmas give necessary and sufficient conditions for the existence of the different types of links in a plausible conditional genetic causal MAG. The term "underlying causal structure" is used to refer to any plausible genetic causal DAG whose marginal and conditional is the plausible conditional genetic causal MAG at hand. Inducing paths in the underlying causal structure are with respect to the nodes in the underlying causal structure that are not in the plausible conditional genetic causal MAG, and **S**. Genetic chains are relative to the variables of the plausible conditional genetic causal MAG.

Lemma A.22. *Let* $\mathbb{M}$ *be a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *and* $G \in \mathbf{G}$. *G and P are adjacent in* $\mathbb{M}$ *if and only if either of the following conditions are satisfied:*

1. *G is a cause of P.*

2. *G is an indicator of a hidden cause of P.*

Proof.

Forward direction: Owing to Assumption 3.5, if C is a collider on some inducing path between G and P in the underlying causal structure, then C is an ancestor of either G or P in the underlying causal structure. The proof is therefore the same as in Lemma A.1.

Reverse direction: The proof is the same as in Lemma A.1. □

Lemma A.23. *Let $\mathbb{M}$ be a plausible genetic causal MAG over $\mathbf{G} \cup \{P\}$ and $G_1, G_2 \in \mathbf{G}$. G_1 and G_2 are adjacent in $\mathbb{M}$ if and only either of the following conditions are satisfied:*

1. *G_1 and G_2 are genetically chained.*
2. *G_1 is a cause of P or an indicator of a hidden cause of P, and G_2 is a cause of P or an indicator of a hidden cause of P.*

Proof.

Forward direction: Owing to Theorem 2.4, there is an inducing path p between G_1 and G_2 in the underlying causal structure. There are two cases:

- *There are no colliders on p:* Then either G_1 is a cause of G_2, G_1 is a cause of G_2, or G_1 and G_2 have a hidden common cause. The first two possibilities are ruled out by Assumption 3.2. Therefore, Condition (1) is satisfied.

- *There are $m \geq 1$ colliders on p:* It will be first proved by induction on $1 \leq i \leq m$ that, for each collider $C^{(i)}$ on p, where $C^{(1)}$ is the collider closer to G_1, either Condition (1) is satisfied, (3) G_1 is a cause of P or an indicator of a hidden cause of P, (4) $C^{(i)}$ is an ancestor of G_1, or (5) $C^{(i)}$ is an ancestor of an observed cause $G^{(k)}$ of P genetically chained to G_1.

  Base case: Owing to Assumption 3.5, $C^{(1)}$ is an ancestor of G_1, G_2, or P. If $C^{(1)}$ is an ancestor of G_1, then Condition (4) is satisfied for $C^{(1)}$. Suppose that $C^{(1)}$ is an ancestor of G_2 and p_1 be a directed path from $C^{(1)}$ to G_2. Then p is into G_1 because otherwise Assumption 3.2 would be violated. Therefore, Condition (1) is satisfied for $C^{(1)}$. Suppose that $C^{(1)}$ is an ancestor of P. If p is out of G_1, then Condition (3) is satisfied for $C^{(1)}$. If p is into G_1, then G_1 and P have a hidden common cause $H^{(1)}$ that is an interior node on $p(G_1, C^{(1)})$. Let p_1 be a directed path from $C^{(1)}$ to P. Owing to Assumption 3.4, there is some $G^{(1)}$ that is an interior node on $[p(H^{(1)}, C^{(1)}), p_1]$. If $G^{(1)}$ is hidden, then Condition (3) is satisfied for $C^{(1)}$. If $G^{(1)}$ is observed, then $G^{(1)}$ is on p_1. Therefore, Condition (5) is satisfied for $C^{(1)}$.

Inductive step: If Condition (1) or (3) is satisfied for $C^{(i-1)}$, then it is satisfied for $C^{(i)}$ as well. Owing to Assumption 3.5, $C^{(i)}$ is an ancestor of G_1, G_2, or P. If $C^{(i)}$ is an ancestor of G_1, then Condition (4) is satisfied for $C^{(i)}$. Suppose that $C^{(1)}$ is an ancestor of G_2. If Condition (4) or (5) is satisfied for $C^{(i-1)}$, then Condition (1) is satisfied for $C^{(i)}$. Suppose that $C^{(i)}$ is an ancestor of P and let $p^{(i)}$ be a directed path from $C^{(i)}$ to P. If Condition (4) is satisfied for $C^{(i-1)}$, then G_1 and P have a hidden common cause $H^{(1)}$ that is an interior node on $p(C^{(i-1)},C^{(i)})$. Due to Assumption 3.4, there is some $G^{(1)}$ on $[p(H^{(1)},C^{(i)}),p^{(i)}]$. If $G^{(1)}$ is hidden, then Condition (3) is satisfied for $C^{(i)}$. If $G^{(1)}$ is observed, then $G^{(1)}$ is on $p^{(i)}$. Therefore, Condition (5) is satisfied for $C^{(i)}$. If Condition (5) is satisfied for $C^{(i-1)}$, then $G^{(k)}$ and P have a hidden common cause $H^{(k+1)}$ that is an interior node on $p(C^{(i-1)},C^{(i)})$. Due to Assumption 3.4, there is some $G^{(k+1)}$ on $[p(H^{(k+1)},C^{(i)}),p^{(i)}]$. If $G^{(k+1)}$ is hidden, then Condition (3) is satisfied for $C^{(i)}$. If $G^{(k+1)}$ is observed, then $G^{(k+1)}$ is on $p^{(i)}$. Therefore, Condition (5) is satisfied for $C^{(i)}$.

If Condition (1) or (3) is satisfied for $C^{(m)}$, then it is satisfied overall. If Condition (4) or (5) is satisfied for $C^{(m)}$, then Condition (1) is satisfied overall. Therefore, either Condition (1) or (3) is satisfied. Applying the proof in the reverse direction (that is, when $C^{(1)}$ is the collider closer to G_2), it is concluded that either Condition (1) is satisfied or 6) G_2 is a cause of P or an indicator of a hidden cause of P. Therefore, either Condition (1) or both conditions 4 and 6 are satisfied. Thus, either Condition (1) or (2) is satisfied.

Reverse direction: There are two cases:

1. *Condition (1) is satisfied.* Let $\{G^{(1)},\ldots,G^{(n)}\}$, where $G^{(1)} = G_1$ and $G^{(n)} = G_2$, be a genetic chain between G_1 and G_2 , and $p^{(i)}$ be a hidden-cause path between $G^{(i)}$ and $G^{(i+1)}$ $(1 \leq i \leq n-1)$ in the underlying causal structure. There are no interior nodes on $p^{(i)}$ that are genotypes due to Assumption 3.2. Therefore, all interior nodes on $p^{(i)}$ are hidden. Let $X^{(i)}$ $(2 \leq i \leq n)$ the first node on $p^{(i-1)}$ that is also on $p^{(i)}$. $[p^{(1)}(G^{(1)},X^{(2)}),\ldots,p^{(n-1)}(X^{(n-1)},X^{(n)}),p_2(X^{(n)},G^{(n)})]$ is an inducing path

between G_1 and G_2. Hence, G_1 and G_2 are adjacent in $\mathbb{M}$ due to Theorem 2.4.

2. *Condition (2) is satisfied.* If G_1 is a cause of P, let p_1 be a directed path from G_1 to P in the underlying causal structure. If G_1 is an indicator of a hidden cause of P, let G' be a hidden cause of P whose presence is indicated by G_1, $\{G^{(1)},\ldots,G^{(n)}\}$ be a genetic chain between G_1 to G_2 , where $G^{(1)} = G_1$ and $G^{(n)} = G'$, $r^{(i)}$ $(1 \leq i \leq n-1)$ be a hidden-cause path from $G^{(i)}$ to $G^{(i+1)}$ in the underlying causal structure, $r^{(n)}$ be a directed path from $G^{(n)}$ to P in the underlying causal structure, $X^{(i)}$ $(2 \leq i \leq n-1)$ be the first node on $r^{(i-1)}$ that is also on $r^{(i)}$, $p_1 = [r^{(1)}(G_1, X^{(2)}),\ldots,r^{(n-1)}(X^{(n-1)}, G'), r^{(n)}]$, p_2 be the path corresponding to p_1 for G_2, X the first node on p_1 that is also on p_2, and $p = [p_1(G_1, X), p_2(X, G_2)]$. Then p is an inducing path between G_1 and G_2. Therefore, G_1 and G_2 are adjacent in $\mathbb{M}$ due to Theorem 2.4.

□

The next lemma follows from the previous two.

Lemma A.24. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, if* G_1 *and* G_2 *are adjacent to P, then* G_1 *and* G_2 *are adjacent.*

Proof. If G_1 and G_2 are adjacent to P, then G_1 is a cause of P or an indicator of a hidden cause of P, and G_2 is a cause of P or an indicator of a hidden cause of P due to Lemma A.22. Therefore, G_1 and G_2 are adjacent due to Lemma A.23. □

The next five lemmas pertain to the orientation of the edges in a plausible genetic causal MAG.

Lemma A.25. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, edges incident to P are out of P.*

Proof. The proof follows directly from Assumption 3.5. □

Lemma A.26. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, edge* $G -\!\!* X$ *implies that G is cause of P.*

Proof. $G -\!\!* X$ implies that G is a cause of X or of some selection variable. If $X = P$, then G is a cause of P due to Assumption 3.5. If $X \in \mathbf{G}$, then G is a cause of some selection variable due to Assumption 3.2. Due to Assumption 3.5, G is a cause of P. □

Lemma A.27. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, edge* $G \leftarrow\!\!* X$ *implies that* G *is not a cause of* P*.*

Proof. $G \leftarrow\!\!* X$ implies that G is not a cause of X or of any selection variable. Owing to Assumption 3.5, G is not a cause of P. □

Lemma A.28. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, if* $X *\!\!- G *\!\!-\!\!* Y$ *is a triple, then the edge between* G *and* Y *is out of* G*.*

Proof. Due to Lemma A.26, G is a cause of P. Suppose that the edge between G and Y is into G. Then, due to Lemma A.27, G is not a cause of P, which is a contradiction. Therefore, the edge between G and Y is out of G. □

Lemma A.29. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, if* G_1 *and* G_2 *are adjacent and* G_1 *and* P *are not adjacent, then the edge between* G_1 *and* G_2 *is into* G_1*.*

Proof. Suppose that the edge between G_1 and G_2 is out of G_1. Then G_1 is a cause of P due to Lemma A.26. Lemma A.23 therefore implies that G_1 and P are adjacent, which is a contradiction. Therefore, the edge is into G_1. □

The next five lemmas give necessary and sufficient conditions for the existence of the different types of edges in a plausible conditional genetic causal MAG.

Lemma A.30. *Let* $\mathbb{M}$ *be a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *and* $G \in \mathbf{G}$*. Edge* $G - P$ *exists if and only if* G *is a cause of* P*.*

Proof.

Forward direction: The proof follows from Lemma A.26.

Reverse direction: If G is cause of P, then G and P are adjacent due to Lemma A.22. Owing to Assumption 3.5, the edge between G and P is undirected. □

Lemma A.31. *Let $\mathbb{M}$ be a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ and $G \in \mathbf{G}$. Edge $G \leftarrow P$ exists if and only if G is an indicator of a hidden cause of P.*

Proof.

Forward direction: The proof follows from Lemmas A.22 and A.27.

Reverse direction: If G is an indicator of a hidden cause of P, then Lemma A.22 says that G and P are adjacent. The edge between G and P is into G due to Lemma A.26 and out of P due to Lemma A.25. □

Lemma A.32. *Let $\mathbb{M}$ be a plausible genetic causal MAG over $\mathbf{G} \cup \{P\}$ and $G_1, G_2 \in \mathbf{G}$. Edge $G_1 - G_2$ exists if and only if G_1 and G_2 are both causes of P.*

Proof.

Forward direction: The proof follows from Lemma A.26.

Reverse direction: If G_1 and G_2 are both causes of P, then G_1 and G_2 are adjacent due to Lemma A.23. Owing to Lemma A.27, the edge between G_1 and G_2 is undirected. □

Lemma A.33. *Let $\mathbb{M}$ be a plausible genetic causal MAG over $\mathbf{G} \cup \{P\}$ and $G_1, G_2 \in \mathbf{G}$. Edge $G_1 \rightarrow G_2$ exists if and only if the following three conditions are satisfied:*

1. *G_1 is a cause of P.*
2. *G_2 is not a cause of P.*
3. *Either G_2 is an indicator of a hidden cause of P or G_1 and G_2 are genetically chained .*

Proof.

Forward direction: If edge $G_1 \rightarrow G_2$ exists, then G_1 is a cause of P and G_2 is not a cause of P due to Lemma A.26 and A.27, respectively. The proof therefore follows from Lemma A.23.

Reverse direction: If G_1 is a cause of P and either G_2 is an indicator of a hidden cause of P or G_1 and G_2 are genetically chained, then G_1 and G_2 are adjacent to due

to Lemma A.23. Owing to Lemma A.27, the edge between G_1 and G_2 is out of G_1. If also G_2 is not a cause of P, then the edge between G_1 and G_2 is into G_2 due to Lemma A.26. □

Lemma A.34. *Let $\mathbb{M}$ be a plausible genetic causal MAG over $\mathbf{G} \cup \{P\}$ and $G_1, G_2 \in \mathbf{G}$. Edge $G_1 \leftrightarrow G_2$ exists if and only if the following three conditions are satisfied:*

1. *G_1 is not a cause of P.*
2. *G_2 is not a cause of P.*
3. *either (a) G_1 and G_2 are genetically chained or (b) G_1 is an indicator of a hidden cause of P and G_2 is an indicator of a hidden cause of P.*

Proof.

Forward direction: Owing to Lemma A.27, G_1 and G_2 are not causes of P. The proof therefore follows from Lemma A.23.

Reverse direction: If either (a) G_1 and G_2 are genetically chained or (b) G_1 is an indicator of a hidden cause of P and G_2 is an indicator of a hidden cause of P, then G_1 and G_2 are adjacent to due to Lemma A.23. If also G_1 and G_2 are not causes of P, then Lemma A.26 implies that the edge between G_1 and G_2 is bidirected. □

Proof of Theorem 3.7. The proof follows from Lemmas A.25, A.29–A.34. □

Proof of Corollary 3.2. If G_1 is a child of P, then G_1 is an indicator of a hidden cause of P due to Theorem 3.7. If G_1 is not a proxy of a hidden cause of P, then there is an observed cause G_2 of P. Owing to Theorem 3.7, G_2 is a neighbour of P. □

Theorem 3.8

In order to prove Theorem 3.8, the following lemmas and definition are needed. The first lemma gives sufficient conditions for two adjacent genotypes in a plausible conditional genetic causal MAG to be genetically chained.

Lemma A.35. *In a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$, if G_1 and G_2 are adjacent and G_1 and P are not adjacent, then G_1 and G_2 are genetically chained .*

Proof. If G_1 and P are not adjacent, then Lemma A.22 implies that G_1 is neither a cause of P nor an indicator of a hidden cause of P. If also G_1 and G_2 are adjacent, then G_1 and G_2 are genetically chained to due to Lemma A.23. □

The next lemma follows from the above.

Lemma A.36. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, if* $[G_1, G_3, G_2]$ *is a triple such that* G_1 *and* G_2 *are not adjacent to* P*,* G_3 *is adjacent to* P*, and the edge between* G_3 *and* P *is out of* G_3*, then the triple is shielded.*

Proof. If $[G_1, G_3, G_2]$ is a triple such that G_1 and G_2 are not adjacent to P, then G_1 and G_3 are genetically chained and G_3 and G_2 are genetically chained due to Lemma A.35. Let $C_1 = \{G_1, G^{(1)}, \ldots, G^{(k)}, G_3\}$ $(k \geq 0)$ be a genetic chain from G_1 to G_3 and $C_2 = \{G_3, G^{(k+1)}, \ldots, G^{(n)}, G_2\}$ $(n \geq k)$ be a genetic chain from G_3 to G_2. If edge $G_3 -\!\!* P$ exists, then Lemma A.26 says that G_3 is a cause of P. Therefore, ordered set $\{G_1, G^{(1)}, \ldots, G^{(k)}, G_3, G^{(k+1)}, \ldots, G^{(n)}, G_2\}$ is a genetic chain from G_1 to G_2. Thus, G_1 and G_2 are adjacent due to Lemma A.23. □

The next lemma pertains to the genomic location of genetically-chained genotypes.

Lemma A.37. *If* G_1 *and* G_2 *are genetically chained, then they are on the same chromosome.*

Proof. Let $\{G^{(1)}, \ldots, G^{(n)}\}$, where $n \geq 2$, $G^{(1)} = G_1$, and $G^{(n)} = G_2$, be a genetic chain from G_1 to G_2. The fact that G_1 and $G^{(i)}$ $(1 \leq i \leq n)$ are on the same chromosome will proved by induction on i:

Base case: $G^{(1)}$ and G^2 are on the same chromosome due to Assumption 3.3.

Inductive step: Suppose that $G^{(i-1)}$ $(2 \leq i \leq n)$ is on the same chromosome as G_1. Owing to Assumption 3.3, $G^{(i-1)}$ and $G^{(i)}$ are on the same chromosome. Therefore, G_1 and $G^{(i)}$ are on the same chromosome by Assumption 3.3.

Thus, G_1 and G_2 are on the same chromosome. □

The lemma below concerns the genomic location of adjacent genotypes in a plausible conditional genetic causal MAG.

Lemma A.38. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, if* G_1 *and* G_2 *are adjacent and* G_1 *and* P *are not adjacent, then* G_1 *and* G_2 *are on the same chromosome.*

Proof. The proof follows from Lemmas A.35 and A.37. □

The next lemma gives sufficient conditions for a potential plausible conditional genetic causal MAG to be a conditional genetic causal MAG.

Lemma A.39. *Let* $\mathbb{M}$ *be a potential plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*.* $\mathbb{M}$ *is a conditional genetic causal MAG if the following two conditions are satisfied:*

1. *Edges incident to P are out of P.*
2. *Edges incident to G are either all out of G or all into G.*

Proof. There are no directed cycles, almost directed cycles, undirected edges incident to nodes with parents or spouses in $\mathbb{M}$. Therefore, $\mathbb{M}$ is an ancestral graph. In addition, there are no primitive inducing paths between any two non-adjacent nodes in $\mathbb{M}$, since there are no edges out of a genotype with incoming edges or edges into P and therefore no colliders in $\mathbb{M}$. Thus, $\mathbb{M}$ is a MAG due to Lemma 2.1. Hence, $\mathbb{M}$ is a conditional genetic causal MAG. □

The *canonical genetic causal DAG with selection nodes* of a conditional genetic causal MAG is defined below.

Definition A.2 (Canonical genetic causal DAG with selection nodes)**.** *Suppose that* $\mathbb{M}$ *is a conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*. The* canonical genetic causal DAG with selection nodes $\mathbb{D}(\mathbb{M})$ *is the genetic causal DAG with selection nodes that is obtained from* $\mathbb{M}$ *by the following procedure:*

1. *Initialise* $\mathbb{D}(\mathbb{M})$ *with the nodes in* $\mathbb{M}$ *and* $\mathbf{S}$
2. *For each* $S \in \mathbf{S}$*, create edge* $P \to S$ *in* $\mathbb{D}(\mathbb{M})$
3. *For each edge* $G - P$ *in* $\mathbb{M}$*, create edge* $G \to P$ *in* $\mathbb{D}(\mathbb{M})$ *and let* $p_G = [G, P]$

4. *For each edge $G \leftarrow P$ in $\mathbb{M}$, create node H_G and path $p_G = G \leftarrow H_G \rightarrow P$ in $\mathbb{D}(\mathbb{M})$*

5. *For each edge between G_1 and G_2 such that G_1 and G_2 are not both adjacent to P in $\mathbb{M}$, create node Y_{G_1,G_2} and path $q_{G_1,G_2} = G_1 \leftarrow Y_{G_1,G_2} \rightarrow G_2$ in $\mathbb{D}(\mathbb{M})$*

The following lemma gives sufficient conditions for a conditional genetic causal MAG to be the marginal of its canonical genetic causal DAG with selection nodes.

Lemma A.40. *Let $\mathbb{M}$ be a conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$. $\mathbb{M}$ is the marginal/conditional of $\mathbb{D}(\mathbb{M})$ over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$ if it satisfies the following four conditions:*

1. *Edges incident to P are out of P.*

2. *Edges incident to each G not adjacent to P are into G.*

3. *Genotypes adjacent to P are adjacent.*

4. *Each triple $[G_1, G_3, G_2]$ such that G_1 and G_2 are not adjacent to P, G_3 is adjacent to P, and the edge between G_3 and P is out of G_3 is shielded.*

Proof. Let $\mathbf{H}$ be the set of nodes in $\mathbb{D}(\mathbb{M})$ that are not in $\mathbb{M}$ or $\mathbf{S}$. It will be first proved that for each edge between X and Y in $\mathbb{M}$, there is an inducing path between X and Y with respect to $\mathbf{H}$ and $\mathbf{S}$ and the edge endpoints match the ancestral relationships in $\mathbb{D}(\mathbb{M})$.

- For each edge between G and P in $\mathbb{M}$, p_G is an inducing path with respect to $\mathbf{H}$ and $\mathbf{S}$. Owing to Condition (1), the edge is out of P. This is in agreement with P being an ancestor of of all nodes in $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$. If the edge is out of G, then G is an ancestor of all nodes in $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$; otherwise, G is not an ancestor of P or any node in $\mathbb{S}$.

- For each edge between G_1 and G_2 in $\mathbb{M}$:

- If G_1 and G_2 are adjacent to P in $\mathbb{M}$, then $[p_{G_1}, p_{G_2}(P, G_2)]$ is an inducing path between G_1 and G_2 with respect to $\mathbf{H}$ and $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$.
- If G_1 is not adjacent to P in $\mathbb{M}$, then q_{G_1,G_2} is an inducing path between G_1 and G_2 with respect to $\mathbf{H}$ and $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$. Condition (2) says that the edge is into G_1.

If the edge is out of G_1, then G_1 and G_2 are adjacent to P in $\mathbb{M}$ and G_1 is an ancestor of all nodes in $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$; otherwise, G_1 is not an ancestor of G_2 or any node in $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$.

It will now be proved that for each pair of nodes X and Y that are not adjacent in $\mathbb{M}$, there is no inducing path between X and Y with respect to $\mathbf{H}$ and $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$.

- If G and P are not adjacent in $\mathbb{M}$, then p_G does not exist in $\mathbb{D}(\mathbb{M})$. Suppose that there is an inducing path p between G and P with respect to $\mathbf{H}$ and $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$. p is of the form $[G^{(1)}, H^{(1)}, \ldots, G^{(i)}, H^{(i)}, \ldots, G^{(n)}, H^{(n)}, P]$, where $n \geq 2$, $G^{(i)}$ $(2 \leq i \leq n)$ is a collider on p, $G^{(1)} = G$, and $H^{(i)} \in \mathbf{H}$ $(1 \leq i \leq n)$ is a non-collider on p. $G^{(n)}$ is not an ancestor of G, P, or any node in $\mathbf{S}$, which is a contradiction. Therefore, there is no inducing path between G and P with respect to $\mathbf{H}$ and $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$.

- If G_1 and G_2 are not adjacent in $\mathbb{M}$, then q_{G_1,G_2} does not exist in $\mathbb{D}(\mathbb{M})$. Suppose that there is an inducing path p between G_1 and G_2 with respect to $\mathbf{H}$ and $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$. There are three cases:

  1. $p = G_1 \rightarrow P \leftarrow G_2$*:* Then edges $G_1 - P$ and $G_2 - P$ exist in $\mathbb{M}$. Condition (3) therefore implies that G_1 and G_2 are adjacent in $\mathbb{M}$. This contradiction shows that $p \neq G_1 \rightarrow P \leftarrow G_2$.
  2. *p is of the form $[G^{(1)}, H^{(1)}, \ldots, G^{(i)}, H^{(i)}, \ldots, G^{(n)}, H^{(n)}, P, H^{(n+1)}, G^{(n+1)}, \ldots, H^{(j)}, G^{(j)}, \ldots, H^{(n+m)}, G^{(n+m)}]$, where $n \geq 1$, $m \geq 1$, $G^{(i)}$ $(2 \leq i \leq n)$, $G^{(j)}$ $(n+1 \leq j \leq n+m-1)$, and P are colliders on p, $G^{(1)} = G_1$, $G^{(n+m)} = G_2$, and $H^{(i)} \in \mathbf{H}$ $(1 \leq i \leq n)$ and $H^{(j)} \in \mathbf{H}$ $(n+1 \leq j \leq n+m)$ are non-colliders on p:* It must be the case that

$n = 1$ because otherwise $G^{(n)}$ is not an ancestor of $G^{(1)}$, $G^{(n+m)}$, or any node in $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$. Similarly, $m = 1$ must hold because otherwise $G^{(n+1)}$ is not an ancestor of $G^{(1)}$, $G^{(n+m)}$, or any node in $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$. Therefore, $p = [G_1, H^{(1)}, P, H^{(2)}, G_2]$, which means that edges $G_1 \to P$ and $G_2 \to P$ exist in $\mathbb{M}$. Condition (3) therefore says that G_1 and G_2 are adjacent in $\mathbb{M}$, which is a contradiction. Thus, p is not of this form.

3. *p is of the form $[G^{(1)}, H^{(1)}, \ldots, G^{(i)}, H^{(i)}, \ldots, G^{(n)}, H^{(n)}, G_2]$, where $n \geq 2$, $G^{(i)}$ $(2 \leq i \leq n)$ is a collider on p, $G^{(1)} = G_1$, and $H^{(i)} \in \mathbf{H}$ $(1 \leq i \leq n)$ is a non-collider on p:* Since no genotype is an ancestor of another in $\mathbb{D}(\mathbb{M})$, for each $2 \leq i \leq n$, edge $G^{(i)} \to P$ exists in $\mathbb{D}(\mathbb{M})$, which means that edge $G^{(i)} - P$ exists in $\mathbb{M}$. Suppose that $n > 2$. Owing to Condition (3), $G^{(2)}$ and $G^{(3)}$ are adjacent in $\mathbb{M}$. Therefore, path $q_{G^{(2)},G^{(3)}}$ does not exist in $\mathbb{D}(\mathbb{M})$, which is a contradiction. Thus, $n = 2$. Suppose that G_1 and P are adjacent in $\mathbb{M}$. Condition (3) then implies that G_1 and $G^{(2)}$ are adjacent in $\mathbb{M}$, which means that path $q_{G_1,G^{(2)}}$ does not exist in $\mathbb{D}(\mathbb{M})$. This contradiction shows that G_1 is not adjacent to P in $\mathbb{M}$. Similarly, G_2 is not adjacent to P in $\mathbb{M}$. Condition (4) therefore implies that G_1 and G_2 are adjacent in $\mathbb{M}$, which is a contradiction. Thus, p is not of this form either. Hence, there is no inducing path between G and P with respect to $\mathbf{H}$ and $\mathbf{S}$ in $\mathbb{D}(\mathbb{M})$. The proof therefore follows from Theorem 2.4.

□

The following lemma gives a sufficient condition for a canonical genetic causal DAG with selection nodes to be plausible.

Lemma A.41. *Let $\mathbb{M}$ be a conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$. $\mathbb{D}(\mathbb{M})$ is a plausible genetic causal DAG with selection nodes if every pair of adjacent genotypes that are not both adjacent to P in $\mathbb{M}$ are on the same chromosome.*

Proof. There are no directed paths from P to any genotype or from a genotype to another in $\mathbb{D}(\mathbb{M})$, since $\mathbb{D}(\mathbb{M})$ has no paths out of P or paths without colliders from

a genotype to another. Therefore, $\mathbb{D}(\mathbb{M})$ is a genetic causal DAG with selection nodes that satisfies Assumption 3.1 and Assumption 3.2. If two genotypes G_1 and G_2 have a common ancestor in $\mathbb{D}(\mathbb{M})$, then G_1 and G_2 are adjacent and at least one of them is not adjacent to P in $\mathbb{M}$. By hypothesis, G_1 and G_2 are on the same chromosome. Therefore, $\mathbb{D}(\mathbb{M})$ satisfies Assumption 3.3. In $\mathbb{D}(\mathbb{M})$, a node is an ancestor of all nodes in $\mathbf{S}$ if and only if the node is an ancestor of P. Hence, $\mathbb{D}(\mathbb{M})$ satisfies Assumption 3.5. Since Assumption 3.4 places no restriction on $\mathbb{D}(\mathbb{M})$, $\mathbb{D}(\mathbb{M})$ is plausible. □

Proof of Theorem 3.8.

Forward direction: Condition (1), (2), (3), (4), (5), and (6) is satisfied for plausible conditional genetic causal MAGs due to Lemma A.25, A.28 and A.29, A.29, A.24, A.36, and A.38, respectively.

Reverse direction: Owing to Lemma A.39, $\mathbb{M}$ is a conditional genetic causal MAG. Lemma A.40 therefore implies that $\mathbb{M}$ is the marginal/conditional of $\mathbb{D}(\mathbb{M})$ over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$. $\mathbb{D}(\mathbb{M})$ is then a plausible genetic causal DAG with selection nodes due to Lemma A.41. Thus, $\mathbb{M}$ is a plausible conditional genetic causal MAG. □

Theorem 3.9

The proof of Theorem 3.9 is based on the following two lemmas concerning Markov equivalence of plausible conditional genetic causal MAGs.

Lemma A.42. *If two plausible conditional genetic causal MAGs $\mathbb{M}_1$ and $\mathbb{M}_2$ over $\mathbf{G} \cup \{P\}$ are Markov equivalent, then the edges incident to an unshielded genotype G adjacent to P are into G in $\mathbb{M}_1$ if and only if they are into G in $\mathbb{M}_2$.*

Proof. The proof follows from Lemma A.28 and Theorem 2.5. □

Lemma A.43. *There are no discriminating paths in a plausible conditional genetic causal MAG.*

Proof. Owing to Lemma A.25, the phenotype cannot be a collider and a parent at the same time. Due to Lemma A.28, a genotype cannot be a collider and a

parent at the same time. Therefore, there are no discriminating paths in a plausible conditional genetic causal MAG. □

Proof of Theorem 3.9. The proof follows from Theorem 2.5 and Lemmas A.25, A.42, A.29, and A.43. □

Theorem 3.10

Proof of Theorem 3.10.

Forward direction: Conditions (1)–(6) are satisfied for maximally-informative plausible conditional genetic causal PAGs due to Theorem 3.8. Condition (7) is satisfied due to Theorem 3.9.

Reverse direction: Let $\mathbb{P}$ be a potential plausible conditional genetic causal PAG that satisfies conditions (1)–(7). Orienting the edges incident to each shielded genotype G in $\mathbb{P}$ either all out of G or all into G results in a plausible conditional genetic causal MAG due to Theorem 3.8. Let $\mathbb{M}_1$ and $\mathbb{M}_2$ be two members of the class of plausible conditional genetic causal MAGs represented by $\mathbb{P}$. Owing to Theorem 3.9, $\mathbb{M}_1$ and $\mathbb{M}_2$ are Markov equivalent, which means that $\mathbb{P}$ is a maximally-informative plausible conditional genetic causal PAG. □

Theorem 3.11

The proof of Theorem 3.11 is based on the following propositions characterising m-separation in a plausible conditional genetic causal MAG.

Lemma A.44. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, if edge* $G -\!\!* X$ *exists, then edge* $G - P$ *exists.*

Proof. If edge $G -\!\!* X$ exists, G is a cause of P due to Lemma A.26. Edge $G - P$ exists due to Lemma A.30. □

Lemma A.45. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, if* G_1 *and* P *are strictly m-separated by* $\mathbf{Z}$*, then for each* $G_2 \in \mathbf{Z}$*, edge* $G_1 *\!\!- G_2$ *exists.*

Proof. If G_1 and P are strictly m-separated by $\mathbf{Z}$, then for each $G_2 \in \mathbf{Z}$, there is a path p that satisfies the conditions of Lemma A.16. Suppose that there is a collider

G_3 on p. Then $G_3 \in \mathbf{Z}$. Owing to Lemma A.28, G_3 is a collider on every path between G_1 and P. Therefore, G_1 and P are m-separated, which is a contradiction. Thus, there are no colliders on p. Since there are no noncolliders on p other than G_2, p is of the form $G_1 *\!\!- G_2 - P$ by Lemma A.28. Hence, edge $G_1 *\!\!- G_2$ exists. □

Lemma A.46. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, if* G_1 *and* P *are strictly m-separated by* $\mathbf{Z}$*, then* $\mathbf{Z}$ *is a subset of the neighbours of* P *on the same chromosome.*

Proof. If G_1 and P are strictly m-separated by $\mathbf{Z}$, then G_1 and P are not adjacent by the definition of MAG and Lemma A.45 says that for each $G_2 \in \mathbf{Z}$, edge $G_1 *\!\!- G_2$ exists. Owing to Theorem A.38, G_1 and G_2 are on the same chromosome. Due to Lemma A.26, G_2 is a cause of P, and due to Lemma A.30, G_2 is a neighbour of P. □

Corollary A.1. *In a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*, G and P are adjacent if and only if they are not m-separated by any subset of the genotypes adjacent to P on the same chromosome.*

Proof. The proof follows from Lemma A.46. □

Lemma A.47. *Suppose that* G_1 *and* P *are strictly m-separated by* $\mathbf{Z}$ *in a plausible conditional genetic causal MAG over* $\mathbf{G} \cup \{P\}$ *given* $\mathbf{S}$*. If* G_1 *and* G_2 *are not adjacent, then they are m-separated by* $\mathbf{Z}$*.*

Proof. By the definition of MAG, G_1 and P are not adjacent. Therefore, there is no path $[G_1, P, \ldots, G_2]$. If G_1 and G_2 are not adjacent, there is no path $[G_1, G_2]$. Suppose that there is a path $p = [G_1, G_3, \ldots, G_2]$. If G_3 is a collider on p, then $G_3 \notin \mathbf{Z}$ due to Lemma A.45 and G_3 is not an ancestor of any node in $\mathbf{Z}$ due to Lemma A.28. Therefore, p is blocked by $\mathbf{Z}$. If G_3 is not a collider on p, then the edge between G_1 and G_3 is out of G_3 due to Lemma A.28 and edge $G_3 - P$ exists due to Lemma A.44. Therefore, $G_3 \in \mathbf{Z}$, because otherwise G_1 and P would be m-connected given $\mathbf{Z}$. Thus, p is blocked by $\mathbf{Z}$. Hence, G_1 and G_2 are m-separated by $\mathbf{Z}$. □

Lemma A.48. *Suppose that G_1 and P are strictly m-separated by $\mathbf{Z}$ and G_2 is not adjacent to P in a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$. Then G_1 and G_2 are adjacent if and only if $G_2 \in \mathbf{Z}$ or G_1 and G_2 are not m-separated by $\mathbf{Z}$.*

Proof.

Forward direction: By the definition of MAG, if G_1 and G_2 are adjacent, then they are not m-separated by $\mathbf{Z}$.

Reverse direction: There are two cases:

1. $G_2 \in \mathbf{Z}$. Owing to Lemma A.45, edge $G_1 *\!\!-\!\!* G_2$ exists. Therefore, G_1 and G_2 are adjacent.

2. $G_2 \notin \mathbf{Z}$. Due to Lemma A.47, if G_1 and G_2 are not m-separated by $\mathbf{Z}$, then G_1 and G_2 are adjacent.

□

Lemma A.49. *Suppose that G_1 and P are strictly m-separated by $\mathbf{Z}_1$ and G_2 and P are strictly m-separated by $\mathbf{Z}_2$ in a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$. G_1 and G_2 are adjacent if and only if G_1 and G_2 are not m-separated by the smallest among $\mathbf{Z}_1$ and $\mathbf{Z}_2$.*

Proof.

Forward direction: By the definition of MAG, if G_1 and G_2 are adjacent, then they are not m-separated by the smallest among $\mathbf{Z}_1$ and $\mathbf{Z}_2$.

Reverse direction: Suppose that $G_2 \in \mathbf{Z}_1$. Then edge $G_1 *\!\!-\!\!* G_2$ exists due to Lemma A.45. Lemma A.44 therefore implies that edge $G_2 - P$ exists, which is a contradiction. Therefore, $G_2 \notin \mathbf{Z}_1$. Similarly, $G_1 \notin \mathbf{Z}_2$. Owing to Lemma A.47, if G_1 and G_2 are not m-separated by the smallest among $\mathbf{Z}_1$ and $\mathbf{Z}_2$, then G_1 and G_2 are adjacent. □

Lemma A.50. *Suppose that G_1 is not adjacent to P or to a node adjacent to P in a plausible conditional genetic causal MAG over $\mathbf{G} \cup \{P\}$ given $\mathbf{S}$. G_1 and G_2 are adjacent if and only if they are not m-separated.*

Proof. By the definition of MAG, G_1 and P are strictly m-separated by some set $\mathbf{Z}$. Suppose that $\mathbf{Z} \neq \emptyset$. Lemma A.45 then says that $G_1 *\!\!- G_3$ exists. Owing to Lemma A.44, edge $G_3 - P$ exists. This contradiction shows that $\mathbf{Z} = \emptyset$. Due to Lemma A.47, G_1 and G_2 are adjacent if and only if they are not m-separated. □

Proof of Theorem 3.11. The proof follows from Corollary A.1, and Lemmas A.24, A.48, A.49, A.38, and A.50. □

Theorem 3.12

Towards proving Theorem 3.12, the soundness of Algorithm 11 is proved first.

Lemma A.51 (Soundness of Algorithm 11)**.** *If the input of Algorithm 11 is* $\mathbb{P}$ *and* **Sepset***, then in the output of Algorithm 11,* $\mathbb{P}$ *is a plausible conditional genetic causal PAG for the same Markov equivalence class as in the input.*

Proof.

Rule (1). Soundness follows directly from Lemma A.25.

Rule (2). When $G_1 \notin \mathbf{Sepset}(\{G_2, P\})$, edge $G_1 *\!\!-\!\!* P$ is oriented as $G_1 \leftarrow\!* P$ due to Lemma 2.2. Therefore, edge $G_1 *\!\!-\!\!* G_2$ is oriented as $G_1 \leftarrow\!* G_2$ due to Lemma A.28. Otherwise, triple $[G_2, G_1, P]$ is oriented either as $G_2 *\!\!\rightarrow G_1 -\!* P$, $G_2 *\!\!- G_1 \leftarrow\!* P$, or $G_2 *\!\!- G_1 -\!* P$, again due to Lemma 2.2. The first two cases are ruled out by Lemma A.28. Therefore, edge $G_1 *\!\!-\!\!* G_2$ is oriented as $G_1 -\!* G_2$.

Rule (3). Soundness follows directly from Lemma A.29.

□

The following lemma is used in the proof of Theorem 3.12.

Lemma A.52. *In the output of Algorithm 11, the endpoints at genotypes adjacent to P are oriented if and only if the genotype is unshielded.*

Proof. There are no $P *\!\!-\!\!\circ G$ edges, $G \circ\!\!-\!* X$ edges such that G is unshielded and adjacent to P, or $G_1 \circ\!\!-\!* G_2$ edges such that G_1 and P are not adjacent in the output PAG, since they are all oriented by Rule (1), Rule (2), and Rule (3), respectively. Therefore, there are only $G \circ\!\!-\!* X$ edges such that G is shielded and adjacent to P. □

Proof of Theorem 3.12. The proof follows from Lemmas A.51 and A.52, and Theorem 3.10. □

Theorem 3.13

Proof of Theorem 3.13. Owing to Conditions (1)–(3), Algorithm 13 discovers G if G is adjacent to P. When G is discovered, the algorithm performed the tests of conditional independence of P and G given each subset of the genotypes adjacent to P on the same chromosome as G (other than G if P and G are adjacent) in $\mathbb{M}$ again due to Conditions (1)–(3). When G is not discovered, the algorithm performed the tests of conditional independence of P and G given a sepset of X and Y due to Corollary A.1 and Conditions (1)–(3). Therefore, the conditioning sets of the tests of conditional independence of G and P performed by the algorithm are separation-sufficient for G and P in $\mathbb{M}$ due to Corollary A.1. Theorem 2.3 and Condition (1) therefore imply that the p-value corresponding to the hypothesis of absence of a link between P and G is upper-bounded by the maximal among the p-values from the tests of conditional independence of G and P performed by the algorithm. Thus, an appropriate FDR-controlling procedure applied to the upper bounds of the link-absence p-values corresponding to the genotypes in $\bigcup_{1\leq k\leq n}\mathbf{TA}_k(P)$ controls the FDR of the genotypes below q. □

Theorem 3.14

Proof of Theorem 3.14. Suppose that the conditions of the theorem are satisfied. P is not an ancestor of G_1 in the underlying causal structure due to Assumption 3.1. Suppose that G_1 and P have a common ancestor. Owing to Assumption 3.4, G_1 has a common cause with some cause G_2 of P, which contradicts Condition (1). Therefore, G_1 and P do not have a common ancestor. Owing to Assumption 3.5, P is a ancestor of S. Therefore, every path from G_1 to S through P is blocked by P. Suppose that there is path p from G_1 to S that is not through P. Owing to Conditions (2)–(4), there are colliders on p. Therefore, p is blocked by P. Thus, the proof follows from Theorem 2.6. □

Chapter 4

Theorem 4.1

In order to prove Theorem 4.1, the following lemmas are needed. The first two lemmas are used in the proof of the third one, which states that Algorithm 15 outputs maximally-informative plausible conditional genetic PAGs.

Lemma A.53. *In $\mathbb{L}$ at line 7 of Algorithm 15, if G_1 and G_2 are adjacent to P, then G_1 and G_2 are adjacent.*

Proof. If G_1 and G_2 are adjacent to P in $\mathbb{L}$ at line 7 of Algorithm 15, then G_1 and G_2 became adjacent to P in $\mathbb{L}$ at line 6. Therefore, G_1 and P are adjacent in $\mathbb{P}_k$ for every k such that $G_1 \in \mathbf{G}_k$ and G_2 and P are adjacent in $\mathbb{P}_k$ for every k such that $G_2 \in \mathbf{G}_k$. Therefore, for every k such that $\{G_1, G_2\} \subseteq \mathbf{G}_k$, G_1 and G_2 are adjacent in $\mathbb{P}_k$ due to Lemma A.24. Thus, G_1 and G_2 also became adjacent in $\mathbb{L}$ at line 6. □

Lemma A.54. *In $\mathbb{S}$ at line 58 of Algorithm 15, if G_1 and G_2 are adjacent to P, then G_1 and G_2 are adjacent.*

Proof. Since $\mathbb{S}$ is set to the subgraph of $\mathbb{L}$ without edges between pairs in $\mathbf{E_1} \cup \mathbf{E_2}$ at line 58, if G_1 and G_2 are adjacent to P in $\mathbb{S}$, then they are adjacent to P in $\mathbb{L}$ at line 58 and neither $\{G_1, P\}$ nor $\{G_2, P\}$ is in $\mathbf{E}_1$ constructed at line 55. Since G_1 and G_2 are adjacent to P in $\mathbb{L}$ at line 58 and no edge between genotypes adjacent to P is removed at lines 29 and 35, G_1 and G_2 are adjacent in $\mathbb{L}$ at line 56 due to Lemma A.53. Therefore, $\{G_1, G_2\} \in \mathbf{E}_3$ constructed at line 56, and $\{G_1, G_2\} \notin \mathbf{E}_2$ constructed at line 57. Thus, G_1 and G_2 are adjacent in $\mathbb{S}$. □

Lemma A.55. *In the output of Algorithm 15, $\mathbf{Q}$ is a set of maximally-informative plausible conditional genetic PAG over $\mathbf{O}$.*

Proof. Let $\mathbb{Q} \in \mathbf{Q}$. It will be proved that $\mathbb{Q}$ satisfies the conditions of Theorem 3.10. $\mathbb{Q}$ is derived by orientation of the edges in $\mathbb{S}$, which is derived by a subgraph of $\mathbb{L}$.

- *Condition (1):* Edges incident to P in $\mathbb{Q}$ were oriented out of P at line 10 of the algorithm.

- *Condition (2):* Edges incident to each unshielded G adjacent to P were all oriented either all out of G or all into G at line 14, 18, 92, 100, 108, or 111.

- *Condition (3):* Edges incident to each G not adjacent to P were oriented into G at line 24 or 86.

- *Condition (4):* Genotypes adjacent to P in $\mathbb{S}$ are adjacent in $\mathbb{S}$ due to Lemma A.54. Since no edges are removed from $\mathbb{Q}$ once it is set to $\mathbb{S}$ at line 106, genotypes adjacent to P in $\mathbb{Q}$ are adjacent in $\mathbb{Q}$.

- *Condition (5):* Owing to lines 62 and 100, there is no unshielded triple $[G_1, G_3, G_2]$ such that G_1 and G_2 are not adjacent to P, G_3 is adjacent to P, and the edge between G_3 and P is out of G_3 in $\mathbb{Q}$.

- *Condition (6):* Owing to lines 29 and 72, there are no adjacent genotypes not both adjacent to P in $\mathbb{Q}$ that are not on the same chromosome.

- *Condition (2):* There is no shielded G adjacent to P such that the endpoints at G were oriented at line 14 or 18 owing to line 81, and no shielded G adjacent to P such that the endpoints at G were oriented at lines 92, 100, 108, and 111. Therefore, endpoints at each G adjacent to P are oriented if and only if G is unshielded.

Theorem 3.10 therefore implies that $\mathbb{Q}$ is a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$.

□

The following lemmas characterise graph $\mathbb{L}$ at different lines of Algorithm 15.

Lemma A.56. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$ that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. At line 7 of Algorithm 15, $\mathbb{L}$ contains a superset of the links in $\mathbb{Q}'$.*

Proof. Let $\mathbb{N}'$ a member of the Markov equivalence class of plausible conditional genetic MAGs represented by $\mathbb{Q}'$ and $\mathbb{N}_{k'}$ be the marginal of $\mathbb{N}'$ over $\mathbf{G}_k \cup \{P\}$. For every pair $\{X, Y\} \subseteq \mathbf{O}$ such that edge $X \circ\!\!-\!\!\circ Y$ is not added to $\mathbb{L}$ at line 6 of Algorithm

15, there is some k such that X and Y are not adjacent in $\mathbb{P}_k$ and, subsequently, in $\mathbb{N}_{k'}$. There is no inducing path between X and Y with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}'$ due to Theorem 2.4. Therefore, X and Y are not adjacent in $\mathbb{N}'$ or in $\mathbb{Q}'$. □

Lemma A.57. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$ that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. If the edges incident to G in $\mathbb{L}$ are oriented out of G at line 14 of Algorithm 15, then edge $G -\!\!* P$ exists in $\mathbb{Q}'$.*

Proof. Let $\mathbb{N}'$ a member of the Markov equivalence class of plausible conditional genetic MAGs represented by $\mathbb{Q}'$ and $\mathbb{N}_{k'}$ be the marginal of $\mathbb{N}'$ over $\mathbf{G}_k \cup \{P\}$. If the edges incident to G in $\mathbb{L}$ are oriented out of G at line 14 of Algorithm 15, then there is some k such that edge $G -\!\!* P$ exists in $\mathbb{P}_k$ and, subsequently, in $\mathbb{N}_{k'}$. Owing to Lemma A.66, edge $G -\!\!* P$ exists in $\mathbb{N}'$, and, subsequently, in $\mathbb{Q}'$. □

Lemma A.58. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$ that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. If the edges incident to G in $\mathbb{L}$ are oriented out of G at line 14 of Algorithm 15, then the edges incident to G in $\mathbb{Q}'$ are out of G.*

Proof. If the edges incident to G in $\mathbb{L}$ are oriented out of G at line 14 of Algorithm 15, then edge $G -\!\!* P$ exists in $\mathbb{Q}'$ due to Lemma A.57. Lemma A.28 therefore implies that the edges incident to G in $\mathbb{Q}'$ are out of G. □

Lemma A.59. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$ that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. If the edges incident to G in $\mathbb{L}$ are oriented into G at line 18 of Algorithm 15, then the edges incident to G in $\mathbb{Q}'$ are into G.*

Proof. Let $\mathbb{N}'$ be a member of the Markov equivalence class of plausible conditional genetic MAGs represented by $\mathbb{Q}'$ and $\mathbb{N}_{k'}$ be the marginal of $\mathbb{N}'$ over $\mathbf{G}_k \cup \{P\}$. If the edges incident to G in $\mathbb{L}$ are oriented into G at line 18 of Algorithm 15, then there is some k such that edge $G \leftarrow\!\!* P$ exists in $\mathbb{P}_k$ and, subsequently, in $\mathbb{N}_{k'}$. Owing to Lemma A.66, edge $G -\!\!* P$ does not exist in $\mathbb{N}'$. If edge $G \leftarrow\!\!* P$ exists in $\mathbb{N}'$ and, therefore, in $\mathbb{Q}'$, then the edges incident to G are into G in $\mathbb{Q}'$ due to Lemma A.28.

If G and P are not adjacent in $\mathbb{N}'$, and, subsequently, in $\mathbb{Q}'$, then the edges incident to G are into G in $\mathbb{Q}'$ due to Lemma A.29. □

Lemma A.60. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over* **O** *that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. At line 25 of Algorithm 15, $\mathbb{L}$ contains a superset of the links, the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at a subset of the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes.*

Proof. Lemma A.56 says that $\mathbb{L}$ at line 7 of Algorithm 15 contains a superset of the links in $\mathbb{Q}'$. At line 25, orientations in $\mathbb{L}$ have been performed at lines 10, 14, 18, and 24. It will be shown that these orientations are present in $\mathbb{Q}'$.

- At line 10, edges incident to P in $\mathbb{L}$ are oriented out of P. Owing to Lemma A.25, edges incident to P in $\mathbb{Q}'$ are out of P.
- If the edges incident to G in $\mathbb{L}$ are oriented out of G at line 14, then the edges incident to G in $\mathbb{Q}'$ are out of G due to Lemma A.58.
- If the edges incident to G in $\mathbb{L}$ are oriented into G at line 18, then Lemma A.59 says that the edges incident to G in $\mathbb{Q}'$ are into G.
- At line 24, edges incident to each genotype G not adjacent to P are oriented into G in $\mathbb{L}$. Since these genotypes are also not adjacent to P in $\mathbb{Q}'$, Lemma A.29 implies the edges incident to the genotypes in $\mathbb{Q}'$ are into the genotypes.

Since no links are removed from $\mathbb{L}$ at lines 7–25, $\mathbb{L}$ also contains a superset of the links in $\mathbb{Q}'$ at line 25. □

Lemma A.61. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over* **O** *that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. At line 37 of Algorithm 15, $\mathbb{L}$ contains a superset of the links, the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at a subset of the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes.*

Proof. At line 25 of Algorithm 15, $\mathbb{L}$ contains a superset of the links, the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at a subset of the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes due to Lemma A.60. Let $\mathbb{N}'$ be a member of the Markov equivalence class of plausible conditional genetic MAGs represented by $\mathbb{Q}'$ and $\mathbb{N}_{k'}$ be the marginal of $\mathbb{N}'$ over $\mathbf{G}_k \cup \{P\}$. Between lines 25 and 37, edges are removed from $\mathbb{L}$ at lines 29 and 35. It will be shown that edges removed at those lines are not in $\mathbb{N}'$.

- Edge $G_1 *\!\!-\!\!* G_2$ is removed from $\mathbb{L}$ at line 29 if G_1 and G_2 are not both adjacent to P in $\mathbb{L}$ and they are not on the same chromosome. Therefore, G_1 and G_2 are not both adjacent to P in $\mathbb{N}'$. Suppose that G_1 and G_2 are adjacent in $\mathbb{N}'$. Then G_1 and G_2 are on the same chromosome due to Lemma A.38, which is a contradiction. Therefore, G_1 and G_2 are not adjacent in $\mathbb{N}'$.

- Edge $G_1 *\!\!- G_2$ is removed from $\mathbb{L}$ at line 35 if G_1 is not adjacent to P in $\mathbb{L}$ and there is some k such that G_1 is not adjacent to P in $\mathbb{P}_k$, and, subsequently, in $\mathbb{N}_{k'}$, and G_2 is not in $\mathbf{G}_k$. Suppose that edge $G_1 *\!\!- G_2$ exists in $\mathbb{N}'$. Then edge $G_2 -\!\!* P$ exists in $\mathbb{N}'$ due to Lemma A.44 and $[G_1, G_2, P]$ is an inducing path between G_1 and P with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}'$. Owing to Theorem 2.4, G_1 and P are adjacent in $\mathbb{N}'_k$, which is contradiction. Therefore, edge $G_1 *\!\!- G_2$ does not exists in $\mathbb{N}'$, and, subsequently, in $\mathbb{Q}'$.

Since the removed edges are not $\mathbb{N}'$, they are not in $\mathbb{Q}'$ either. □

The next lemma is used in the proof of the subsequent two lemmas, which characterise inducing paths with respect to $\mathbf{G}_h$ and $\emptyset$ in a plausible conditional genetic MAG over $\mathbf{G}_o \dot{\cup} \mathbf{G}_h \cup \{P\}$.

Lemma A.62. *In a plausible conditional genetic MAG over $\mathbf{G}_o \dot{\cup} \mathbf{G}_h \cup \{P\}$, if Z is an interior node on an inducing path between X and Y with respect to $\mathbf{G}_h$ and $\emptyset$, then $Z \in \mathbf{G}_h$, edges incident to Z are out of Z, and edge $Z - P$ exists.*

Proof. Let p be an inducing path between X and Y with respect to $\mathbf{G}_h$ and $\emptyset$. Suppose that Z is a collider on p. Due to Lemma A.25, $Z \neq P$. Therefore, $Z \in \mathbf{G}$. Due

to Lemma A.28, Z is not an ancestor of X or Y. Therefore, p is not an inducing path with respect to $\mathbf{G}_h$ and $\emptyset$, which is a contradiction. Hence, Z is a noncollider on p and $Z \in \mathbf{G}_h$. Edge $Z - P$ exists due to Lemma A.44. □

Lemma A.63. *In a plausible conditional genetic MAG over $\mathbf{G}_o \dot{\cup} \mathbf{G}_h \cup \{P\}$, there is an inducing path between G_1 and P with respect to $\mathbf{G}_h$ with respect to $\mathbf{G}_h$ and $\emptyset$ if and only if either G_1 and P are adjacent or edge $G_1 *\!\!- G_2$ such that $G_2 \in \mathbf{G}_h$ exists.*

Proof.

Forward direction: Let p be an inducing path between G_1 and P with respect to $\mathbf{G}_h$ and $\emptyset$. If there are no interior nodes on p, then G_1 and P are adjacent. If there are interior nodes on p, let G_2 be the node adjacent to G_1 on p. Owing to Lemma A.62, $G_2 \in \mathbf{G}_h$ and the edge between G_1 and G_2 is out of G_2.

Reverse direction: If G_1 and P are adjacent, $[G_1, P]$ is an inducing path between G_1 and P with respect to $\mathbf{G}_h$ and $\emptyset$. If edge $G_1 *\!\!- G_2$ such that $G_2 \in \mathbf{G}_h$ exists, then path $G_1 *\!\!- G_2 -\!\!* P$ exists due to Lemma A.44 and is an inducing path between G_1 and P with respect to $\mathbf{G}_h$ and $\emptyset$. □

Lemma A.64. *In a plausible conditional genetic MAG over $\mathbf{G}_o \dot{\cup} \mathbf{G}_h \cup \{P\}$, there is an inducing path between G_1 and G_2 with respect to $\mathbf{G}_h$ and $\emptyset$ if and only if G_1 and G_2 are adjacent or there is a path $G_1 *\!\!- G_3 -\!\!* G_2$ such that $G_3 \in \mathbf{G}_h$ or a path $G_1 *\!\!- G_3 - G_4 -\!\!* G_2$ such that $\{G_3, G_4\} \subseteq \mathbf{G}_h$.*

Proof.

Forward direction: Let p be an inducing path between G_1 and G_2 with respect to $\mathbf{G}_h$ and $\emptyset$. If there are no interior nodes on p, then G_1 and G_2 are adjacent. If there are interior nodes on p, let Z be such a node. Due to Lemma A.62, $Z \in \mathbf{G}_h$, all edges incident to Z are out of Z, and edge $Z - P$ exists. If G_3 is the only interior node on p, then path $G_1 *\!\!- G_3 -\!\!* G_2$ exists. If there are more than one interior nodes on p, let G_3 and G_4 be the nodes adjacent to G_1 and G_2, respectively, on p. Due to Lemmas A.24 and A.28, edge $G_3 - G_4$ exists. Therefore, path $G_1 *\!\!- G_3 - G_4 -\!\!* G_2$ exists.

Reverse direction: $G_1 *\!\!-\!\!* G_2$, $G_1 *\!\!- G_3 -\!\!* G_2$ such that $G_3 \in \mathbf{G}_h$, and $G_1 *\!\!- G_3 - G_4 -\!\!* G_2$ such that $\{G_3, G_4\} \subseteq \mathbf{G}_h$ are inducing paths between G_1 and G_2 with respect to $\mathbf{G}_h$ and $\emptyset$. □

The following lemma gives necessary and sufficient conditions for a plausible conditional genetic MAG to be consistent with a set of marginal MAGs.

Lemma A.65. *A plausible conditional genetic MAG* $\mathbb{N}$ *over* $\mathbf{O}$ *is consistent with* $\mathbb{M}_1, \ldots, \mathbb{M}_n$ *if the following conditions are satisfied for every* $\mathbb{M}_k$ $(1 \leq k \leq n)$*:*

1. *For each edge* $G -\!\!* P$ *such that* G *is unshielded in* $\mathbb{M}_k$*, edge* $G -\!\!* P$ *exists in* $\mathbb{N}$.

2. *For each edge* $G_1 *\!\!-\!\!* P$ *such that either* G_1 *is shielded or the edge is into* G_1 *in* $\mathbb{M}_k$*,* G_1 *and* P *are adjacent or edge* $G_1 *\!\!- G_2$ *such that* $G_2 \in \mathbf{G} \setminus \mathbf{G}_k$ *exists in* $\mathbb{N}$.

3. *For each edge between* G_1 *and* G_2 *in* $\mathbb{M}_k$ *such that* G_1 *and* P *are not adjacent in* $\mathbb{M}_k$*,* G_1 *and* G_2 *are adjacent in* $\mathbb{N}$.

4. *For each* G_1 *not adjacent to* P *in* $\mathbb{M}_k$*,* G_1 *and* P *are not adjacent and edge* $G_1 *\!\!- G_2$ *such that* $G_2 \in \mathbf{G} \setminus \mathbf{G}_k$ *does not exist in* $\mathbb{N}$.

5. *For each* G_1 *and* G_2 *not adjacent in* $\mathbb{M}_k$*,* G_1 *and* G_2 *are not adjacent in* $\mathbb{N}$.

6. *For each edge* $G \leftarrow\!\!* P$ *such that* G *is unshielded in* $\mathbb{M}_k$*, edge* $G -\!\!* P$ *does not exist in* $\mathbb{N}$.

Proof. Let $\mathbb{N}_k$ be the marginal of $\mathbb{N}$ over $\mathbf{G}_k \cup \{P\}$ for some k and suppose that $\mathbb{N}$ satisfies the above conditions with $\mathbb{N}_k$. It will be first proved that, for each type of edge in $\mathbb{M}_k$, there is an inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$.

- For each edge $G -\!\!* P$ such that G is unshielded in $\mathbb{M}_k$, Condition (1) implies that G and P are adjacent in $\mathbb{N}$. Therefore, $[G, P]$ is an inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$.

- For each edge $G *\!\!-\!\!* P$ such that either G is shielded or the edge is into G in $\mathbb{M}_k$, there is an inducing path between G and P with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$ due to Condition (2) and Lemma A.63.

- For each edge $G_1 - G_2$ in $\mathbb{M}_k$ such that G_1 and G_2 are unshielded, edges $G_1 - P$ and $G_2 - P$ exist in $\mathbb{M}_k$ due to Lemma A.44 and G_1 and G_2 are adjacent in $\mathbb{N}$ due to Condition (1). Lemma A.24 therefore implies that G_1 and G_2 are adjacent in $\mathbb{N}$. $[G_1, G_2]$ is then an inducing path between G_1 and G_2 with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$.

- For each edge $G_1 *\!\!- G_2$ in $\mathbb{M}_k$ such that G_1 and G_2 are adjacent to P, G_2 is unshielded, and either G_1 is shielded or the edge is into G_1, the edge between G_2 and P in $\mathbb{M}_k$ is out of G_2 due to Lemma A.28, and edge $G_2 -\!\!* P$ exists in $\mathbb{N}$ due to Condition (1). Furthermore, either edge $G_1 \leftarrow\!\!* P$ or edge $G_1 \circ\!\!-\!\!* P$ exists in $\mathbb{P}_k$ due to Lemma A.42. Owing to Condition (2), either G_1 and P are adjacent in $\mathbb{N}$ or edge $G_1 *\!\!- G_3$ such that $G_3 \in \mathbf{G} \setminus \mathbf{G}_k$ exists in $\mathbb{N}$. In the former case, Lemma A.24 says that G_1 and G_2 are adjacent in $\mathbb{Q}$. In the latter case, edge $G_3 -\!\!* P$ exists in $\mathbb{N}$ due to Lemma A.44, G_2 and G_3 are adjacent in $\mathbb{N}$ due to Lemma A.24, and the edge between G_2 and G_3 in $\mathbb{N}$ is out of G_3 due to Lemma A.28. Therefore, path $G_1 *\!\!- G_3 -\!\!* G_2$ exists in $\mathbb{N}$. Lemma A.64 therefore implies that there is an inducing path between G_1 and G_2 with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$.

- For each edge $G_1 *\!\!-\!\!* G_2$ in $\mathbb{M}_k$ such that G_1 and G_2 are adjacent to P, either G_1 is shielded or the edge is into G_1, and either G_2 is shielded or the edge is into G_2, either edge $G_1 \leftarrow\!\!* P$ or edge $G_1 \circ\!\!-\!\!* P$ and either edge $G_2 \leftarrow\!\!* P$ or edge $G_2 \circ\!\!-\!\!* P$ exists in $\mathbb{P}_k$ due to Lemma A.42. Therefore, either G_1 and P are adjacent in $\mathbb{N}$ or edge $G_1 *\!\!- G_3$ such that $G_3 \in \mathbf{G} \setminus \mathbf{G}_k$ exists in $\mathbb{N}$, and either G_2 and P are adjacent in $\mathbb{N}$ or edge $G_2 *\!\!- G_4$ such that $G_4 \in \mathbf{G} \setminus \mathbf{G}_k$ exists in $\mathbb{N}$ due to Condition (2).

  - If G_1 and P are adjacent and G_2 and P are adjacent in $\mathbb{N}$, then G_1 and G_2 are adjacent in $\mathbb{N}$ due to Lemma A.24.

  - If G_1 and P are adjacent in $\mathbb{N}$ and edge $G_2 *\!\!- G_4$ such that $G_4 \in \mathbf{G} \setminus \mathbf{G}_k$ exists in $\mathbb{N}$, then edge $G_4 -\!\!* P$ exists in $\mathbb{N}$ due to Lemma A.44, G_1 and G_4 are adjacent in $\mathbb{N}$ due to Lemma A.24, and the edge between G_1 and

G_4 in $\mathbb{N}$ is out of G_4 due to Lemma A.28. Therefore, path $G_1 *\!\!- G_4 -\!\!* G_2$ exists in $\mathbb{N}$. Thus, there is an inducing path between G_1 and G_2 with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$ due to Lemma A.64.

– Suppose that edge $G_1 *\!\!- G_3$ such that $G_3 \in \mathbf{G} \setminus \mathbf{G}_k$ and edge $G_2 *\!\!- G_4$ such that $G_4 \in \mathbf{G} \setminus \mathbf{G}_k$ exist in $\mathbb{N}$. If $G_3 = G_4$, then path $G_1 *\!\!- G_3 -\!\!* G_2$ exists in $\mathbb{N}$. Therefore, there is an inducing path between G_1 and G_2 with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$ due to Lemma A.64. If $G_3 \neq G_4$, then G_3 and G_4 are adjacent in $\mathbb{N}$ due to Lemma A.24, and the edge between G_3 and G_4 in $\mathbb{N}$ is undirected due to Lemma A.28. Therefore, path $G_1 *\!\!- G_3 - G_4 -\!\!* G_2$ exists in $\mathbb{N}$. Lemma A.64 then says that there is an inducing path between G_1 and G_2 with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$.

- For each edge $G_1 *\!\!-\!\!* G_2$ in $\mathbb{M}_k$ such that G_1 and P are not adjacent in $\mathbb{M}_k$, Condition (3) says that G_1 and G_2 are adjacent in $\mathbb{N}$. Thus, $[G_1, G_2]$ is an inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$.

It will be subsequently proved that, for each type of non-adjacency in $\mathbb{M}_k$, there is no inducing path with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$.

- For each G_1 not adjacent to P in $\mathbb{M}_k$, there is no inducing path between G_1 and P with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$ due to Condition (4) and Lemma A.63.

- If G_1 and G_2 are not adjacent in $\mathbb{M}_k$, then either G_1 or G_2 is not adjacent to P in $\mathbb{M}_k$ because otherwise G_1 and G_2 would be adjacent due to Lemma A.24. Without loss of generality, suppose that G_1 is not adjacent to P in $\mathbb{M}_k$. Condition (4) then says that there is no edge $G_1 *\!\!- G_3$ such that $G_3 \in \mathbf{G} \setminus \mathbf{G}_k$ in $\mathbb{N}$. Therefore, there is no path $G_1 *\!\!- G_3 -\!\!* G_2$ such that $G_3 \in \mathbf{G} \setminus \mathbf{G}_k$ or path $G_1 *\!\!- G_3 - G_4 -\!\!* G_2$ such that $\{G_3, G_4\} \subseteq \mathbf{G} \setminus \mathbf{G}_k$ in $\mathbb{N}$. Owing to Condition (5), G_1 and G_2 are not adjacent in $\mathbb{N}$. Lemma A.64 therefore implies that there is no inducing path between G_1 and G_2 with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}$.

Owing to Theorem 2.4, $\mathbb{M}_k$ and $\mathbb{N}_k$ have the same skeleton. It will be now shown that the edges incident to unshielded genotypes adjacent to P in $\mathbb{M}_k$ and $\mathbb{N}_k$ have the same orientation at the genotype in $\mathbb{M}_k$ and $\mathbb{N}_k$. Due to Lemma A.28, the edges at a genotype G in $\mathbb{M}_k$ are either all into or all out of G.

- For each unshielded genotype G that is adjacent to P and the edges incident to G are out of G in $\mathbb{M}_k$, Condition (1) implies that edge $G -\!\!* P$ exists in $\mathbb{N}$. Lemma A.66 therefore implies that edge $G -\!\!* P$ exists in $\mathbb{N}_k$. The other edges incident to G in $\mathbb{N}_k$ are also out of G due to Lemma A.28.

- For each unshielded genotype G that is adjacent to P and the edges incident to G are into G in $\mathbb{M}_k$, edge $G -\!\!* P$ does not exist in $\mathbb{N}$ due to Condition (6). Owing to Lemma A.66, edge $G -\!\!* P$ does not exist in $\mathbb{N}_k$. Thus, the edge between G and P in $\mathbb{N}_k$ is into G. The other edges incident to G in $\mathbb{N}_k$ are also into G due to Lemma A.28.

Therefore, $\mathbb{M}_k$ and $\mathbb{N}_k$ are Markov equivalent due to Theorem 3.9. Thus, $\mathbb{N}$ is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$.

□

The lemmas below are needed in the proof of soundness of Algorithm 15, that follows after.

Lemma A.66. *Let $\mathbb{M}$ be a plausible conditional genetic causal MAG over $\mathbf{G}_o \dot{\cup} \mathbf{G}_h \cup \{P\}$ and $\mathbb{M}_o$ be the marginal of $\mathbb{M}$ over $\mathbf{G}_o \cup \{P\}$. Edge $G -\!\!* P$ exists in $\mathbb{M}_o$ if and only if edge $G -\!\!* P$ exists in $\mathbb{M}$.*

Proof.

Forward direction: If edge $G -\!\!* P$ exists in $\mathbb{M}_o$, then G is anterior to P in $\mathbb{M}$. Therefore, some edge $G -\!\!* X$ exists in $\mathbb{M}$ (X may be P). Owing to Lemma A.44, edge $G -\!\!* P$ exists in $\mathbb{M}$.

Reverse direction: If edge $G -\!\!* P$ exists in $\mathbb{M}$, then G and P are adjacent in $\mathbb{M}_o$ due to Theorem 2.4 and the edge between G and P in $\mathbb{M}_o$ is out of G. □

Lemma A.67. *Let $\mathbb{M}$ be a plausible conditional genetic causal MAG over $\mathbf{G}_o \dot{\cup} \mathbf{G}_h \cup \{P\}$ and $\mathbb{M}_o$ be the marginal of $\mathbb{M}$ over $\mathbf{G}_o \cup \{P\}$. If G_1 and G_2 are adjacent and G_1 and P are not adjacent in $\mathbb{M}_o$, then G_1 and G_2 are adjacent in $\mathbb{M}$.*

Proof. If G_1 and P are not adjacent in $\mathbb{M}_o$, then there is no inducing path between G_1 and P with respect to $\mathbf{G}_h$ and $\emptyset$ in $\mathbb{M}$ due to Theorem 2.4. Theorem 2.4 also says that, if G_1 and G_2 are adjacent in $\mathbb{M}_o$, then there is an inducing path between G_1 and G_2 with respect to $\mathbf{G}_h$ and $\emptyset$ in $\mathbb{M}$. Owing to Lemma A.64, G_1 and G_2 are adjacent or there is a path $G_1 \ast\!\!-\!\!- G_3 -\!\!\ast G_2$ such that $G_3 \in \mathbf{G}_h$ or a path $G_1 \ast\!\!-\!\!- G_3 - G_4 -\!\!\ast G_2$ such that $\{G_3, G_4\} \subseteq \mathbf{G}_h$ in $\mathbb{M}$. Suppose that G_1 and G_2 are not adjacent in $\mathbb{M}$. Then $[G_1, G_3, P]$ is an inducing path between G_1 and P with respect to $\mathbf{G}_h$ and $\emptyset$ in $\mathbb{M}$, which is a contradiction. Therefore, G_1 and G_2 are adjacent in $\mathbb{M}$. □

Lemma A.68. *If there is some k such that edge $G -\!\!\ast P$ exists and G is unshielded in $\mathbb{M}_k$, then $G -\!\!\ast P$ exists in $\mathbb{L}$ at line 25 of Algorithm 15.*

Proof. If there is some k such that edge $G -\!\!\ast P$ exists and G is unshielded in $\mathbb{M}_k$, then edge $G -\!\!\ast P$ exists in $\mathbb{P}_k$ due to Lemma A.42, in $\mathbb{M}$ due to Lemma A.66, in every $\mathbb{M}_{k'}$ such that $G \in \mathbf{G}_{k'}$ again due to Lemma A.66, and in every $\mathbb{P}_{k'}$ such that $G \in \mathbf{G}_{k'}$ and G is unshielded in $\mathbb{P}_{k'}$ due to Lemma A.42. Therefore, G and P become adjacent in $\mathbb{L}$ at line 6 of Algorithm 15 and the edge between G and P is oriented out of G at line 14. □

Lemma A.69. *If there is some k such that edge $G \leftarrow\!\!\ast P$ exists in $\mathbb{M}_k$ and G is unshielded in $\mathbb{M}_k$, then either G and P are not adjacent or edge $G \leftarrow\!\!\ast P$ exists in $\mathbb{L}$ at line 25 of Algorithm 15.*

Proof. If there is some k such that edge $G \leftarrow\!\!\ast P$ exists and G is unshielded in $\mathbb{M}_k$, then Lemma A.42 says that edge $G \leftarrow\!\!\ast P$ exists in $\mathbb{P}_k$. Edge $G -\!\!\ast P$ does not exist in $\mathbb{M}$ due to Lemma A.66, in any $\mathbb{M}_{k'}$ such that $G \in \mathbf{G}_{k'}$ again due to Lemma A.66, and in any $\mathbb{P}_{k'}$ such that $G \in \mathbf{G}_{k'}$ and G is unshielded in $\mathbb{P}_{k'}$ due to Lemma A.42. Therefore, if G and P are adjacent in $\mathbb{L}$ at line 7 of Algorithm 15, then the edge between G and P is oriented into G at line 18. □

Lemma A.70. *Let $\mathbb{P}$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{G} \cup \{P\}$ and $\mathbb{M}$ a plausible conditional genetic MAG in the Markov equivalence class of plausible conditional genetic MAGs represented by $\mathbb{P}$. G_1 and P are adjacent or edge $G_1 \ast\!\!- G_2$ exists in $\mathbb{P}$ if and only if G_1 and P are adjacent or edge $G_1 \ast\!\!- G_2$ exists in $\mathbb{M}$.*

Proof. G_1 and P are adjacent in $\mathbb{P}$ if and only if they are adjacent in $\mathbb{M}$. If edge $G_1 \ast\!\!- G_2$ exists in $\mathbb{P}$, then edge $G_1 \ast\!\!- G_2$ exists in $\mathbb{M}$. If edge $G_1 \ast\!\!- G_2$ exists in $\mathbb{M}$, then path $G_1 \ast\!\!- G_2 -\!\!\ast P$ exists in $\mathbb{M}$ due to Lemma A.44. Owing to Lemma A.28, the corresponding path in $\mathbb{P}$ is either $G_1 \ast\!\!- G_2 -\!\!\ast P$ or $G_1 \ast\!\!-\!\!\circ G_2 \circ\!\!-\!\!\ast P$. In the former case, edge $G_1 \ast\!\!- G_2$ exists in $\mathbb{P}$. In the latter case, Lemma A.42 says that G_1 and P are adjacent in $\mathbb{P}$. □

Proof of Theorem 4.1. Owing to Lemma A.55, $\mathbf{Q}$ is a set of maximally-informative plausible conditional genetic PAGs over $\mathbf{O}$. Let $\mathbb{Q} \in \mathbf{Q}$ and $\mathbb{N}$ be MAG in the Markov equivalence class of plausible conditional genetic MAGs represented by $\mathbb{Q}$. It will be shown that $\mathbb{Q}$ satisfies the conditions of Lemma A.65 for $\mathbb{M}_k$.

- *Condition (1):* For each edge $G -\!\!\ast P$ such that G is unshielded in $\mathbb{M}_k$, edge $G -\!\!\ast P$ exists in $\mathbb{L}$ at line 25 due to Lemma A.68. Since the edge is not added to $\mathbf{R}_1$ at line 42, it exists in $\mathbb{S}$, and, subsequently, in $\mathbb{Q}$ and $\mathbb{N}$.

- *Condition (2):* For each edge $G_1 \ast\!\!-\!\!\ast P$ such that either G_1 is shielded or the edge is into G_1 in $\mathbb{M}_k$, Lemma A.42 says that either edge $G_1 \leftarrow\!\!\ast P$ or edge $G_1 \circ\!\!-\!\!\ast P$ exists in $\mathbb{P}_k$. Therefore, either G_1 and P are adjacent in $\mathbb{Q}$ or edge $G_1 \ast\!\!- G_2$ such that $G_2 \in \mathbf{G} \setminus \mathbf{G}_k$ exists in $\mathbb{Q}$ due to line 126 and in $\mathbb{N}$ due to Lemma A.70.

- *Condition (3):* For each edge between G_1 and G_2 such that G_1 and P are not adjacent in $\mathbb{M}_k$, Lemma A.67 says that G_1 and G_2 are adjacent in $\mathbb{M}$. Lemma A.61 therefore implies that G_1 and G_2 are adjacent in $\mathbb{L}$ at line 37. Furthermore, since G_1 and P are not adjacent in $\mathbb{M}_k$, and therefore in $\mathbb{P}_k$, G_1 and P do not become adjacent in $\mathbb{L}$ at line 6. Therefore, $\{G_1, G_2\}$ is not added

to $\mathbf{R}_2$ at line 48 or 52. Thus, G_1 and G_2 are adjacent in $\mathbb{S}$, and, subsequently, in $\mathbb{Q}$ and $\mathbb{N}$.

- *Condition (4):* For each G_1 not adjacent to P in $\mathbb{M}_k$, G_1 and P do not become adjacent in $\mathbb{L}$ at line 6. Therefore, G_1 and P are not adjacent in $\mathbb{Q}$, and, subsequently, in $\mathbb{N}$. Owing to lines 35 and 92, there is no edge $G_1 *\!\!-\!\!* G_2$ such that $G_2 \in \mathbf{G} \setminus \mathbf{G}_k$ in $\mathbb{Q}$, and, subsequently, in $\mathbb{N}$.

- *Condition (5):* For each G_1 and G_2 not adjacent in $\mathbb{M}_k$, G_1 and G_2 are not adjacent in $\mathbb{P}_k$. Therefore, G_1 and G_2 do not become adjacent in $\mathbb{L}$ at line 6. Thus, G_1 and G_2 are not adjacent in $\mathbb{Q}$, and, subsequently, in $\mathbb{N}$.

- *Condition (6):* For each edge $G \leftarrow\!\!* P$ such that G is unshielded in $\mathbb{M}_k$, either G and P are not adjacent or edge $G \leftarrow\!\!* P$ exists in $\mathbb{L}$ at line 25 due to Lemma A.69. Therefore, edge $G \leftarrow\!\!* P$ does not exist in $\mathbb{Q}$, and, subsequently, in $\mathbb{N}$.

Therefore, $\mathbb{N}$, and, subsequently, $\mathbb{Q}$, is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$ due to Lemma A.65.

□

Theorem 4.2

The proof of Theorem 4.2 is based on the following lemmas.

Lemma A.71. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$ that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. If edge $G -\!\!* P$ exists in $\mathbb{L}$ at line 25 of Algorithm 15, then G and P are adjacent in $\mathbb{Q}'$.*

Proof. If edge $G -\!\!* P$ exists in $\mathbb{L}$ at line 25 of Algorithm 15, then the endpoints at G in $\mathbb{L}$ were set to tails at line 14. Owing to Lemma A.57, edge $G -\!\!* P$ exists in $\mathbb{Q}'$. □

Lemma A.72. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$ that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. At line 25 of Algorithm 15, if G and P are adjacent in $\mathbb{L}$ and there is some k such that G and the neighbours and potential neighbours of P that are on the same chromosome as G in $\mathbb{L}$ are in $\mathbf{G}_k$, then G and P are adjacent in $\mathbb{Q}'$.*

Proof. If G and P are adjacent in $\mathbb{L}$ at line 25 of Algorithm 15, then G and P are adjacent in every $\mathbb{P}_k$ such that $G \in \mathbf{G}_k$. Therefore, if there is some k such that G and the neighbours and potential neighbours of P that are on the same chromosome as G in $\mathbb{L}$ are in $\mathbf{G}_k$, then G and P are not m-separated by any subset of the neighbours and potential neighbours of P that are on the same chromosome as G in $\mathbb{L}$. Owing to Lemma A.60, G and P are not m-separated by any subset of the neighbours of P that are on the same chromosome as G in $\mathbb{Q}'$. Lemma A.46 therefore implies that G and P are adjacent in $\mathbb{Q}'$. □

Lemma A.73. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$ that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. At line 58 of Algorithm 15, some $\mathbb{S}$ is considered with the same skeleton as $\mathbb{Q}'$ and the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at a subset of the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes.*

Proof. At line 37, $\mathbb{L}$ contains a superset of the links, the orientations at P, the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at a subset of the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes due to Lemma A.61.

At line 58, every subgraph $\mathbb{S}$ of $\mathbb{L}$ with the genotype–phenotype edges in $\mathbb{L}$ between pairs of nodes not in $\mathbf{R}_1$, the genotype–genotype edges in $\mathbb{L}$ between pairs of nodes not in $\mathbf{R}_2$, and the genotype–genotype edges in $\mathbb{L}$ between nodes adjacent to P in $\mathbb{S}$ is constructed. It will be first shown that links in $\mathbb{L}$ between pairs of nodes not in $\mathbf{R}_1$ or $\mathbf{R}_2$ are present in $\mathbb{Q}'$.

- If G and P are adjacent in $\mathbb{L}$ and $\{G, P\} \notin \mathbf{R}_1$, then either the edge between G and P in $\mathbb{L}$ is undirected or there is some k such that G and the neighbours and potential neighbours of P that are on the same chromosome as G in $\mathbb{L}$ are in $\mathbf{G}_k$. In the former case, G and P are adjacent in $\mathbb{Q}'$ due to Lemma A.71. In the latter case, Lemma A.72 says that G and P are adjacent in $\mathbb{Q}'$.

- If G_1 and G_2 are adjacent in $\mathbb{L}$ and $\{G_1, G_2\} \notin \mathbf{R}_2$, then either G_1 and G_2 are

both adjacent to P in $\mathbb{L}$ and $\{G_1, P\} \notin \mathbf{R}_1$ and $\{G_1, P\} \notin \mathbf{R}_1$, or G_1 and G_2 are not both adjacent to P in $\mathbb{L}$ and there is some k such that G_1 and G_2 are in $\mathbf{G}_k$ and G_1 and G_2 are not both adjacent to P in $\mathbb{P}_k$. In the former case, Lemma A.24 implies that G_1 and G_2 are adjacent to P in $\mathbb{Q}'$. In the latter case, G_1 and G_2 are adjacent to P in $\mathbb{M}_k$. Therefore, G_1 and G_2 are adjacent to P in $\mathbb{Q}'$ due to Lemma A.67.

Furthermore, if G_1 and G_2 are adjacent to P in $\mathbb{Q}'$, then G_1 and G_2 are adjacent in $\mathbb{Q}'$ due to Lemma A.24. Therefore, some $\mathbb{S}$ is considered with the same skeleton as $\mathbb{Q}'$ and the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at a subset of the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes. □

Lemma A.74. *Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$ that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. At line 106 of Algorithm 15, some $\mathbb{Q}$ is considered with the same skeleton as $\mathbb{Q}'$ and the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes.*

Proof. At line 58 of Algorithm 15, Lemma A.73 says that some $\mathbb{S}$ is considered with the same skeleton as $\mathbb{Q}'$ and the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at a subset of the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes. The condition at line 62, 72, and 81 fails due to Lemma A.36, A.38, and A.42 respectively. The orientations made at lines 86 are present in $\mathbb{Q}'$ due to Lemma A.29. In the following, let $\mathbb{N}'$ be a member of the Markov equivalence class of plausible conditional genetic MAGs represented by $\mathbb{Q}'$ and $\mathbb{N}_{k'}$ be the marginal of $\mathbb{N}'$ over $\mathbf{G}_k \cup \{P\}$.

The edges incident to G_2 in $\mathbb{S}$ are oriented into G_2 at line 92 if there is an edge $G_1 \ast\!\!-\!\!\circ\, G_2$ in $\mathbb{S}$ such that G_1 is not adjacent to P in $\mathbb{L}$ and there is some k such that G_1 is not adjacent to P in $\mathbb{P}_k$, and, subsequently, in $\mathbb{N}_{k'}$, and G_2 is not in $\mathbf{G}_k$. Suppose that the edges incident to G_2 in $\mathbb{N}'$ are out of G_2. Then edge $G_2 -\!\!\ast\, P$ exists

in $\mathbb{N}'$ due to Lemma A.44 and because G_1 and G_2 are adjacent in $\mathbb{N}'$ (as $\mathbb{S}$ has the same skeleton as $\mathbb{Q}'$), $[G_1, G_2, P]$ is an inducing path between G_1 and P with respect to $\mathbf{G} \setminus \mathbf{G}_k$ and $\emptyset$ in $\mathbb{N}'$. Owing to Theorem 2.4, G_1 and P are adjacent in $\mathbb{N}'_k$, which is contradiction. Therefore, the edges incident to G_2 in $\mathbb{N}'$, and, subsequently, in $\mathbb{Q}'$, are into G_2.

The edges incident to potential neighbour G_3 of P in $\mathbb{S}$ are oriented into G_3 at line 100 if there is an unshielded triple $[G_1, G_3, G_2]$ such that G_1 and G_2 are not adjacent to P in $\mathbb{S}$. Suppose that the edges incident to G_3 are out of P in $\mathbb{Q}'$. Owing to Lemma A.36, G_1 and G_2 are adjacent in $\mathbb{Q}'$. This is a contradiction, as $\mathbb{Q}'$ has the same skeleton as $\mathbb{S}$. Therefore, the edges incident to G_3 are into P in $\mathbb{Q}'$.

Thus, some $\mathbb{Q}$ is considered with the same skeleton as $\mathbb{Q}'$ and the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes at line 106. □

Proof of Theorem 4.2. Let $\mathbb{Q}'$ be a maximally-informative plausible conditional genetic PAG over $\mathbf{O}$ that is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$. Owing to Lemma A.74, some $\mathbb{Q}$ is considered with the same skeleton as $\mathbb{Q}'$ and the orientations at P, the orientations at a subset of the unshielded genotypes adjacent to P, and the orientations at the genotypes not adjacent to P in $\mathbb{Q}'$, and no orientations at the remaining genotypes at line 106 of Algorithm 15. Since both orientations for the rest unshielded genotypes adjacent to P are considered in Part 6 of the algorithm, $\mathbb{Q} = \mathbb{Q}'$ holds at line 112 at some iteration of the loop. Let $\mathbb{N}$ be a member of the Markov equivalence class of plausible conditional genetic MAGs represented by $\mathbb{Q}$ and $\mathbb{N}_k$ be the marginal of $\mathbb{N}$ over $\mathbf{G}_k \cup \{P\}$. Since $\mathbb{Q}$ is consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$, $\mathbb{N}_k$ and $\mathbb{M}_k$ have the same skeleton due to Theorem 3.3. Owing to Theorem 2.4 and Lemma A.63, if G_1 and P are adjacent in $\mathbb{M}_k$, then either G_1 and G_2 are adjacent in $\mathbb{N}$ or edge $G_1 *\!\!-\!\!* G_2$ such that $G_2 \in \mathbf{G} \setminus \mathbf{G}_k$ exists in $\mathbb{N}$. Therefore, either G_1 and G_2 are adjacent in $\mathbb{N}$ or edge $G_1 *\!\!-\!\!* G_2$ such that $G_2 \in \mathbf{G} \setminus \mathbf{G}_k$ exists in $\mathbb{Q}$. Thus, the condition at line 126 succeeds and $\mathbb{Q}$ is added to $\mathbf{Q}$ at line 127. □

Theorem 4.3

The proof of Theorem 4.3 follows from the following lemma.

Lemma A.75. *Let $\mathbb{N}$ be a potential plausible conditional genetic causal MAG that is obtained from $\mathbb{L}$ at line 25 of Algorithm 15 by the following procedure:*

1. *Remove a subset of the $G \leftarrow\!\!* P$ and $G \circ\!\!* P$ edges such that there is no k such that G and the neighbours and potential neighbours of P in $\mathbb{L}$ that are on the same chromosome as G are all in $\mathbf{G}_k$.*

2. *For each G adjacent to P such that the endpoints at G are unoriented in $\mathbb{L}$, orient the edges incident to G either all out of all into G.*

3. *For each G not adjacent to P such that the endpoints at G are unoriented in $\mathbb{L}$, orient the edges incident to G into G.*

4. *Remove $G_1 *\!\!* G_2$ edges such that G_1 and G_2 are not both adjacent to P in $\mathbb{L}$ and not on the same chromosome.*

5. *Remove $G_1 *\!\!- G_2$ edges such that G_1 is not adjacent to P in $\mathbb{L}$ and there is some k such that G_1 is not adjacent to P in $\mathbb{P}_k$ and G_2 is not in $\mathbf{G}_k$.*

Then $\mathbb{N}$ is a plausible conditional genetic causal MAG consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$.

Proof. It will be first proved that $\mathbb{N}$ satisfies the conditions of Theorem 3.8.

- *Condition (1):* Edges incident to P in $\mathbb{N}$ were oriented out of P at line 10 of Algorithm 15.

- *Condition (2):* Edges incident to each G adjacent to P in $\mathbb{N}$ were all oriented either all out of G or all into G at line 14, 18, or in Step (2) of the procedure above.

- *Condition (3):* Edges incident to each G not adjacent to P in $\mathbb{N}$ were oriented into G at line 24 or in Step (3).

- *Condition (4):* Owing to Lemma A.24, genotypes adjacent to P are adjacent in $\mathbb{M}$. Lemma A.60 implies that $\mathbb{L}$ at line 25 contained a superset of the links in $\mathbb{M}$. Therefore, genotypes adjacent to P were adjacent in $\mathbb{L}$ at line 25. Since no edge between genotypes adjacent to P is removed by the procedure above, G_1 and G_2 are adjacent in $\mathbb{N}$.

- *Condition (5):* If there is a triple $[G_1, G_3, G_2]$ such that G_1 and G_2 are not adjacent to P, G_3 is adjacent to P, and the edge between G_3 and P is out of G_3 in $\mathbb{N}$, then there is some k such that G_1 and P are not adjacent in $\mathbb{P}_k$ because otherwise G_1 and P would not become adjacent in $\mathbb{L}$ at line 6. Owing to Step (5), G_3 is in $\mathbf{G}_k$. G_1 and G_3 are adjacent in $\mathbb{P}_k$, and subsequently in $\mathbb{M}_k$, because otherwise they would not become adjacent in $\mathbb{L}$ at line 6. Lemma A.67 therefore implies that G_1 and G_3 are adjacent in $\mathbb{M}$. Furthermore, G_1 and P are not adjacent in $\mathbb{M}$ due to Theorem 2.4. Similarly, G_2 and G_3 are adjacent in $\mathbb{M}$ and G_2 and P are not adjacent in $\mathbb{M}$. Owing to Lemma A.69, edge $G -\!\!* P$ exists in $\mathbb{M}_k$, and due to Lemma A.66, in $\mathbb{M}$. Owing to Lemma A.36, G_1 and G_2 are adjacent in $\mathbb{M}$. Lemma A.60 then says that G_1 and G_2 were adjacent in $\mathbb{L}$ at line 25. G_1 and G_2 are on the same chromosome due to Lemma A.38. Therefore, the edge between G_1 and G_2 was not removed in Step (4). Since G_2 is not adjacent to P in $\mathbb{N}$ and $\mathbb{N}$ was shown to satisfy Condition (3) of Theorem 3.8, the edge between G_1 and G_2 was into G_2. Therefore, the edge was not removed in Step (5). Thus, G_1 and G_2 are adjacent in $\mathbb{N}$.

- *Condition (6):* Owing to Step (4), there are no G_1 and G_2 such that G_1 and P are not adjacent, G_1 and G_2 are adjacent, and G_1 and G_2 are on the same chromosome in $\mathbb{N}$.

Theorem 3.8 therefore implies that $\mathbb{N}$ is a plausible conditional genetic causal MAG. It will be now proved that $\mathbb{N}$ satisfies the conditions of Lemma A.65 with every $\mathbb{M}_k$ ($1 \leq k \leq n$).

- *Condition (1):* For each edge $G -\!\!* P$ such that G is unshielded in $\mathbb{M}_k$, edge $G -\!\!* P$ exists in $\mathbb{L}$ at line 25 due to Lemma A.68 and was not removed from

$\mathbb{N}$ in Step (1), (4), or (5).

- *Condition (3):* For each edge between G_1 and G_2 in $\mathbb{M}_k$ such that G_1 and P are not adjacent in $\mathbb{M}_k$, Lemma A.67 says that G_1 and G_2 are adjacent in $\mathbb{M}$. Lemma A.60 therefore implies that G_1 and G_2 were adjacent in $\mathbb{L}$ at line 25. Furthermore, since G_1 and P are not adjacent in $\mathbb{M}_k$, and therefore in $\mathbb{P}_k$, G_1 and P did not become adjacent in $\mathbb{L}$ at line 6. Owing to Lemma A.60, G_1 and P are not adjacent in $\mathbb{M}$. Lemma A.38 says that G_1 and G_2 are on the same chromosome. Therefore, the edge between G_1 and G_2 in $\mathbb{N}$ was not removed in Step (4). Suppose that the edge is removed in Step (5). Then the edge between G_1 and G_2 in $\mathbb{N}$ was out of G_2 and there is some k' such that G_1 is not adjacent to P in $\mathbb{P}_{k'}$ and G_2 is not in $\mathbf{G}_{k'}$. The edges incident to G_2 in $\mathbb{N}$ were oriented out of G_2 at line 14 or in Step (2) of the procedure above. Therefore, G_2 and P are adjacent in $\mathbb{N}$. G_2 and P are adjacent in $\mathbb{M}_k$ because otherwise G_2 and P would not become adjacent in $\mathbb{L}$ at line 6. The edge between G_2 and P in $\mathbb{M}_k$ is out of G_2 due to Lemma A.69. Owing to Lemma A.66, edge $G_2 -\!\!* P$ exists in $\mathbb{M}$. The edge between G_1 and G_2 in $\mathbb{M}$ is out of G_2 due to Lemma A.28. Owing to Lemma A.63, there is an inducing path between G_1 and P with respect to $\mathbf{G} \setminus \mathbf{G}_{k'}$ and $\emptyset$ in $\mathbb{M}$. Due to Theorem 2.4, G_1 and P are adjacent in $\mathbb{M}_{k'}$, and, subsequently, in $\mathbb{P}_{k'}$, which is a contradiction. Therefore, G_1 and G_2 are adjacent in $\mathbb{N}$.

- *Condition (2):* For each edge $G_1 *\!\!-\!\!* P$ such that either G_1 is shielded or the edge is into G_1 in $\mathbb{M}_k$, if G_1 and P are not adjacent in $\mathbb{N}$, there are two cases:

    1. *G_1 and P are not adjacent in $\mathbb{L}$ at line 6:* There is some k' such that G_1 and P are not adjacent in $\mathbb{P}_{k'}$, and, subsequently, in $\mathbb{M}_{k'}$. Therefore, G_1 and P are strictly m-separated by some set $\mathbf{Z}$ in $\mathbb{M}_{k'}$. Lemma A.45 says that edge $G_1 *\!\!- G_2$ exists in $\mathbb{M}_{k'}$ for each G_2 in $\mathbf{Z}$. There is some G_2 in $\mathbf{Z}$ that is not in $\mathbf{G}_k$ because otherwise G_1 and P would be m-separated in $\mathbb{M}_k$. Since $\mathbb{N}$ was shown to satisfy Condition (3) of Lemma A.65, G_1 and G_2 are adjacent in $\mathbb{N}$. Owing to Lemma A.44, $G_2 -\!\!* P$ exists in

$\mathbb{M}_{k'}$. Since $\mathbb{N}$ was shown to satisfy Condition (1) of Lemma A.65, edge $G_2 -\!\!* P$ exists in $\mathbb{N}$. Therefore, the edge between G_1 and G_2 in $\mathbb{N}$ is out of G_2 due to Lemma A.28.

2. *G_1 and P are adjacent in $\mathbb{L}$ at line 6 and removed in Step (1) of the procedure above:* There was some neighbour or potential neighbour G_2 of P in $\mathbb{L}$ at line 25 on the same chromosome as G_1 that is not in $\mathbf{G}_k$. G_1 and G_2 were adjacent in $\mathbb{L}$ at line 7 due to Lemma A.53. Since G_1 and G_2 are on the same chromosome, the edge between them was not removed in Step (4). There is no k such that G_1 is not adjacent to P in $\mathbb{P}_k$, because otherwise G_1 and P would not become adjacent in $\mathbb{L}$ at line 6. Therefore, the edge between G_1 and G_2 was not removed in Step (5).

- *Condition (4):* For each G_1 not adjacent to P in $\mathbb{M}_k$, G_1 and P did not become adjacent in $\mathbb{L}$ at line 6. Therefore, G_1 and P are not adjacent in $\mathbb{N}$. Owing to Step (5), there is no edge $G_1 *\!\!- G_2$ such that $G_2 \in \mathbf{G} \setminus \mathbf{G}_k$ in $\mathbb{N}$.

- *Condition (5):* For each G_1 and G_2 not adjacent in $\mathbb{M}_k$, G_1 and G_2 are not adjacent in $\mathbb{P}_k$. Therefore, G_1 and G_2 do not become adjacent in $\mathbb{L}$ at line 6. Thus, G_1 and G_2 are not adjacent in $\mathbb{N}$.

- *Condition (6):* If G is an unshielded genotype adjacent to P and the edges incident to G are into G in $\mathbb{M}_k$, Lemma A.69 says that edge $G -\!\!* P$ does not exist in $\mathbb{L}$ at line 25, and, therefore, in $\mathbb{N}$.

Owing to Lemma A.65, $\mathbb{N}$ is a plausible conditional genetic causal MAG consistent with $\mathbb{M}_1, \ldots, \mathbb{M}_n$.

□

Proof of Theorem 4.3. The set $\mathbf{fNE}(P)$, $\mathbf{cCH}(P)$, and $\mathbf{cPNE}(P)$ constructed in Algorithm 16 is the set of neighbours, children, and potential neighbours, respectively, of P at line 25 of Algorithm 15. Owing to Lemmas A.56 and A.71, $\mathbf{fNE}(P)$ is a superset of the set of fixed neighbours of P and the set of removable neighbours of P is empty. Lemmas A.75 and A.28 therefore imply that $\mathbf{fNE}(P)$ is the set of fixed

neighbours of P. Owing to Lemmas A.56 and A.72, the set $\mathbf{fCH}(P)$ constructed in Algorithm 16 is a superset of the set of fixed children of P. Lemmas A.75 and A.28 therefore imply that $\mathbf{fCH}(P)$ is the set of fixed children of P. Finally, Lemmas A.56, A.75, and A.28 imply that the set $\mathbf{fPNE}(P)$, $\mathbf{rCH}(P)$, and $\mathbf{rPNE}(P)$ constructed in Algorithm 16 is the set of fixed potential neighbours, removable children, and removable potential neighbours, respectively, of P. □

Theorem 4.4

Proof of Theorem 4.4. The hypothesis of absence of a consistent link between P and G is equivalent to the union of the hypotheses of d-separation of P and G given each set in a union of collections that are separation-sufficient for G and P in $\mathbb{M}_i$ for each $1 \leq i \leq n$ such that $G \in \mathbf{G}_i$. From the proof of Theorem 3.13, the conditioning sets of the tests of conditional independence of G and P performed by the algorithm on D_i are separation-sufficient for G and P in $\mathbb{M}_i$ for each $1 \leq i \leq n$ such that $G \in \mathbf{G}_i$. Theorem 2.3 and Condition (1) therefore imply that the p-value corresponding to the hypothesis of absence of a consistent link between P and G is upper-bounded by the maximal among the p-values corresponding to the tests of conditional independence of G and P performed by the algorithm (note that the p-value of the same test is the same across datasets as Algorithm 7 is used to perform the tests). Thus, an appropriate FDR-controlling procedure applied to the upper bounds of the consistent-link-absence p-values corresponding to the genotypes in $\widehat{\mathbf{cAD}}(P)$ controls the FDR of the genotypes below q. □

Appendix B

Exome-array association study

B.1 Introduction

Exome arrays are an inexpensive means to genotype exonic SNPs identified in exome-sequencing studies [Grove et al., 2013]. The Prion Unit used exome arrays with UK, German, and US sCJD samples, and I created four datasets (described in Table B.1) using the generated data and external control data. Association analysis was then applied to each of the datasets. The fourth analysis is a mega-analysis.

#	Dataset	n	m
1	UK sCJD cases vs. GERAD controls	1444	86789
2	German sCJD cases vs. German controls	3476	106462
3	US sCJD cases vs. Coriell controls	1654	93493
4	UK, German, and US cases vs. GERAD, German, and Coriell controls	6536	111910

Table B.1: Datasets in the exome-array association study. *GERAD* refers to the GERAD consortium, while *Coriell* refers to the *Coriell Institute for Medical Research*. *n* and *m* denote the number of samples and variants, respectively.

B.2 Method

An exome-array association study involves the same quality control steps before and after association analysis as a GWAS [Weale, 2010, Anderson et al., 2010]. However, the accuracy of the genotype clusters provided by Illumina, the manufacturer of the chips used, is decreased for rare SNPs compared to common ones [Ritchie et al., 2011], and most SNPs in exome arrays are rare. The *Cohorts for Heart and*

Aging Research in Genomic Epidemiology (CHARGE) consortium overcame this problem by generating new clusters for Illumina's HumanExome v1.0 chip from a much larger sample than Illumina used [Grove et al., 2013]. These clusters could not be used in this study because chips other than HumanExome v1.0 were also used. Re-clustering using the Prion Unit's datasets was also not a option, since the sample sizes are relatively small. Therefore, the *zCall* rare-variant caller [Goldstein et al., 2012] was used to post-process the calls made using Illumina's clusters, in order to improve the accuracy of the calls. This approach was also followed by the *Genetic and Environmental Risk in Alzheimer's Disease (GERAD)* consortium, which supplied controls used in this study.

A further complicating factor in this study was the fact that the Prion Unit's original datasets were generated on different chip models and different versions of the same model (see Table B.2 for the list of original datasets). In order to eliminate false positives due to chip differences, strict quality control was performed before merging the datasets (see below).

#	Cohort	Chip
1	UK sCJD cases	HumanExome v1.1
2		HumanOmniExpressExome v1.0
3	German sCJD cases	HumanExome v1.1
4		HumanOmniExpressExome v1.0
5	US sCJD cases	HumanOmniExpressExome v1.0
6	GERAD controls	HumanExome v1.0
7	German controls	HumanExome v1.0
8	Coriell controls	NeuroX (HumanExome v.1.1 plus custom content)

Table B.2: Original datasets used in the exome-array association study. All chips were manufactured by Illumina. *GERAD* refers to the GERAD consortium, while *Coriell* refers to the *Coriell Institute for Medical Research*. Among the datasets, only the one by GERAD was already processed.

For each dataset except the GERAD one (because it was already processed using zCall), initial quality control, pre-zCall quality control, zCall processing, and post-zCall quality control were performed (see Figure B.1 for the description of the quality control steps), based on the standard operating procedure used by GERAD (personal communication).

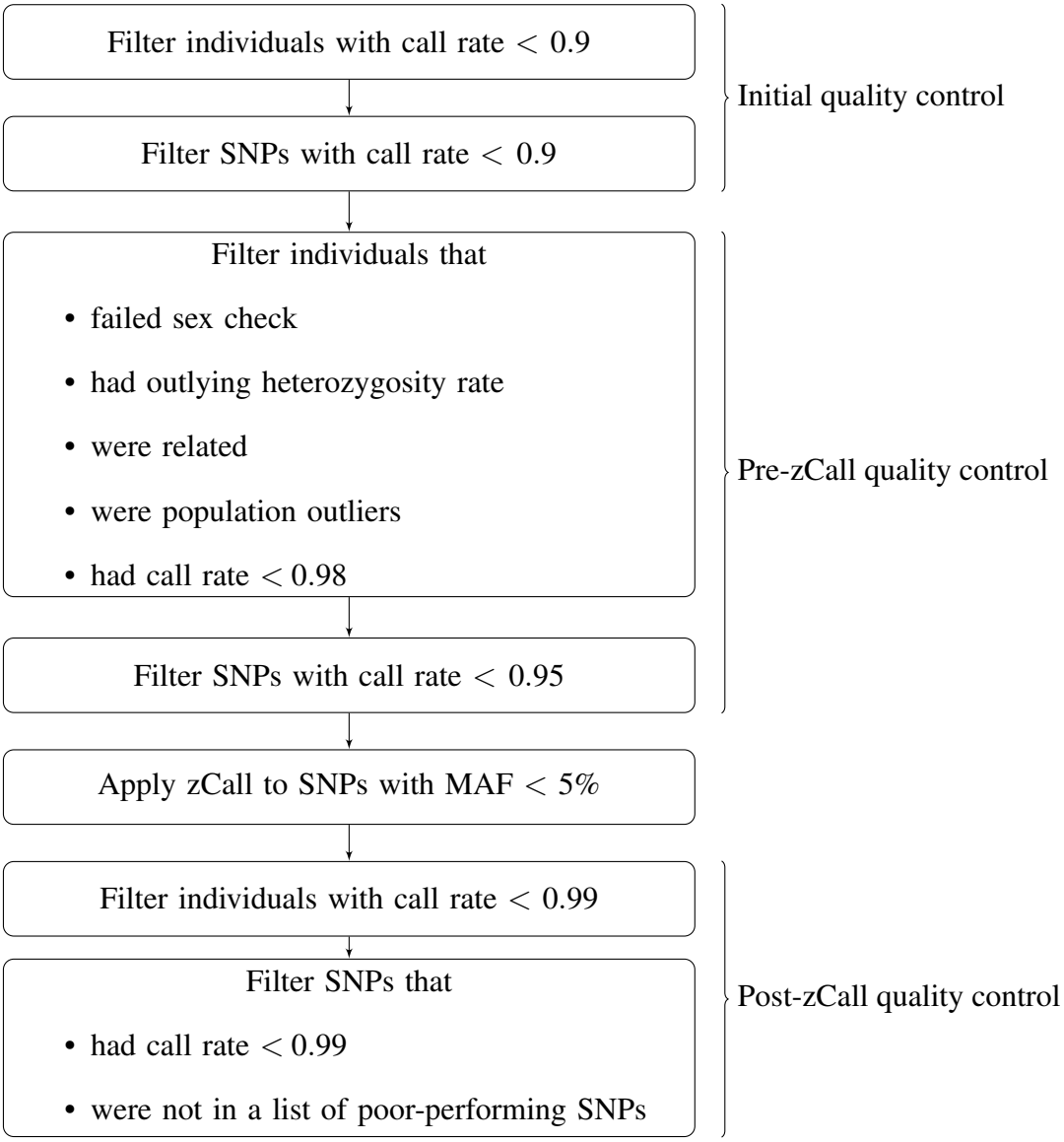


Figure B.1: Quality control of the original exome-array datasets (see Table B.2) except the GERAD one (since it was already processed). Heterozygosity rate was considered outlying if it was > 3 standard deviations away from the mean. Individuals were considered related if they had IBD (identity by descent) > 0.185 and population outliers if their first or the second *multi-dimensional scaling (MDS)* component was > 3 standard deviations away from the mean. MAF refers to the MAF in the original Illumina sample that was used to generate the genotype clusters. A list of poorly-performing SNPs on the Illumina HumanExome v1.0 chip was created by CHARGE [Grove et al., 2013]. SNPs in that list were also dropped from the GERAD dataset.

For each association study, merging of the original datasets, individual quality control, SNP quality control, association analysis, and post-association quality control were performed (see Figure B.2 for the description of the quality control steps).

The observations were taken to be the gametes of the individuals and association analysis was performed using Fisher's exact test.

Post-association quality control involved comparing the MAF in the controls with the MAF in the *European American* population on the *Exome Variant Server*[1] or the *European* population in The 1000 Genomes Project and inspecting the original clusters of the hits.

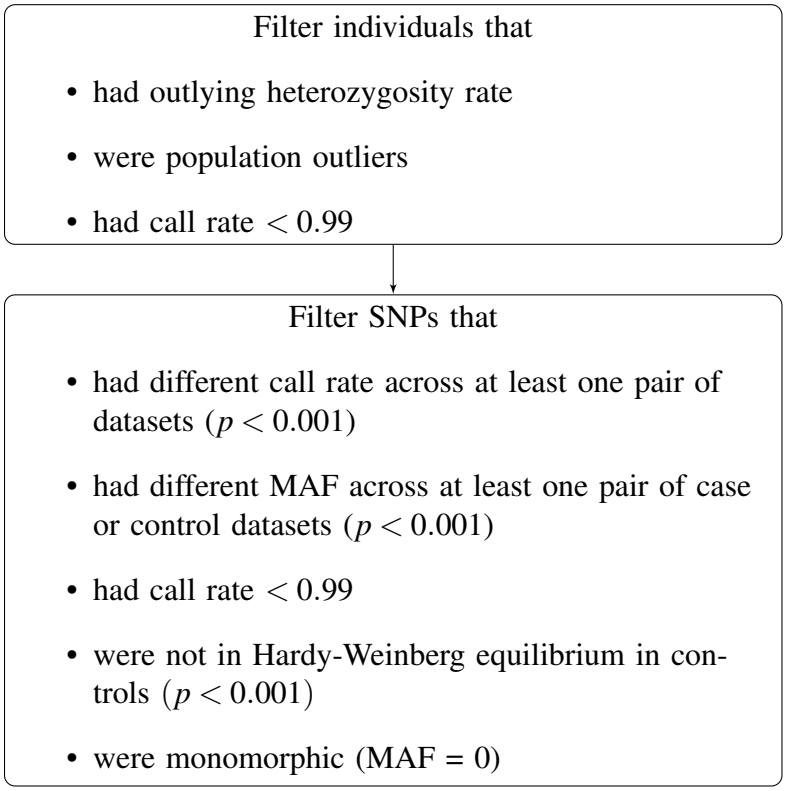


Figure B.2: Quality control of to the exome-array datasets (see Table B.1). MAF stands for minor allele frequency. The second SNP quality control step was performed for control datasets only in Dataset 4, in order to allow for differences in the reporting of sCJD cases across different countries. The p-values were calculated using asymptotic χ^2 tests.

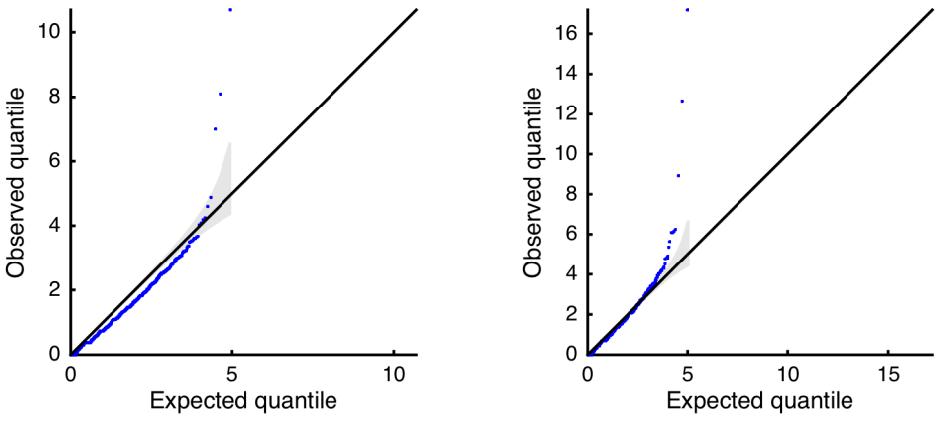
B.3 Results

Figure B.3 shows the *quantile-quantile* plots for each association analysis. Since most SNPs are not expected to be associated with disease, most p-values should fall on the main diagonal. However, if an exact test is used, as in this case, the p-values

[1] http://evs.gs.washington.edu/EVS/

should fall on a line below the main diagonal [see Note 6 in Weale, 2010]. This can be most clearly seen in Figure B.3a.

Table B.3 lists the SNPs that achieved an association p-value $\leq 10^{-6}$ in one of the datasets in Table B.1. For each hit, if the MAF in controls matches the MAF in EA, or the MAF in EUR if the former is not available, the hit was deemed admissible. The hit was also deemed admissible if the MAF is not available in EA or EUR. For each admissible hit, the clusters in the original datasets were inspected (see *Supplementary Information II – Exome array hit clusters*). The clusters for rs144218313, rs200542656, rs114501427, and rs199759206 were deemed problematic. The clusters for rs145985036 and rs116589141 were well-formed but the SNPs were not detected in the cases when Sanger sequencing was performed by a colleague. In conclusion, there were no discoveries in this study.



(a) UK **(b)** DE

(c) US **(d)** Mega-analysis

Figure B.3: Quantile-quantile plots for each exome-array association analysis. The gray areas are the 95% concentration bands [see Weale, 2010].

D#	RS#	C#	Position	MAF (cases)	MAF (controls)	P-value	OR	OR 95% CI	MAF (EA)	MAF (EUR)	Admissible?
1	rs1058065	2	27587724	0.015	0.000	$1.02 \cdot 10^{-7}$	N/A	[N/A, N/A]	0.015		No
1	rs144218313	4	983394	0.018	0.000	$8.06 \cdot 10^{-9}$	N/A	[N/A, N/A]	0.000		Yes
1	rs142631461	17	27576202	0.023	0.000	$2.05 \cdot 10^{-11}$	N/A	[N/A, N/A]	0.027		No
2	rs1058065	2	27587724	0.016	0.001	$2.14 \cdot 10^{-13}$	29.86	[8.95, 99.59]	0.015		No
2	exm-rs13088462	3	51071713	0.076	0.043	$8.10 \cdot 10^{-7}$	1.84	[1.46, 2.33]	N/A	0.004	Yes
2	rs6647	14	94847415	0.235	0.177	$6.83 \cdot 10^{-7}$	1.44	[1.25, 1.65]	0.212		No
2	rs4778138	15	28335820	0.124	0.178	$8.05 \cdot 10^{-7}$	0.66	[0.55, 0.78]	N/A	0.002	Yes
2	rs142631461	17	27576202	0.017	0.000	$6.59 \cdot 10^{-18}$	N/A	[N/A, N/A]	0.027		No
2	rs200542656	19	46998601	0.009	0.000	$1.21 \cdot 10^{-9}$	N/A	[N/A, N/A]	0.000		Yes
2	exm1507050	19	55494076	0.011	0.001	$5.77 \cdot 10^{-7}$	8.85	[3.63, 21.56]	0.002		No
3	N/A	1	12837153	0.014	0.000	$1.57 \cdot 10^{-7}$	N/A	[N/A, N/A]	N/A	N/A	Yes
3	rs114501427	2	139326516	0.021	0.000	$2.85 \cdot 10^{-11}$	N/A	[N/A, N/A]	0.000		Yes
3	rs1058065	2	27587724	0.000	0.015	$2.94 \cdot 10^{-8}$	0.00	[0.00, N/A]	0.015		No
3	rs144218313	4	983394	0.038	0.000	$3.95 \cdot 10^{-20}$	N/A	[N/A, N/A]	0.000		Yes
3	rs77363096	5	41160373	0.014	0.000	$1.57 \cdot 10^{-7}$	N/A	[N/A, N/A]	0.000		No
3	rs201768800	9	139411768	0.015	0.000	$3.74 \cdot 10^{-8}$	N/A	[N/A, N/A]	0.000		No
3	rs199759206	9	139734293	0.013	0.000	$3.08 \cdot 10^{-7}$	N/A	[N/A, N/A]	0.000		Yes
3	rs145985036	11	119045520	0.023	0.000	$3.28 \cdot 10^{-12}$	N/A	[N/A, N/A]	0.000		Yes
3	rs151199705	14	105858969	0.028	0.000	$8.32 \cdot 10^{-15}$	N/A	[N/A, N/A]	0.000		No
3	rs142631461	17	27576202	0.000	0.027	$2.64 \cdot 10^{-14}$	0.00	[0.00, N/A]	0.027		No
3	rs115310908	22	37465121	0.023	0.001	$6.21 \cdot 10^{-11}$	39.39	[5.40, 287.50]	0.000		No
3	rs41307393	23	88009170	0.016	0.000	$5.13 \cdot 10^{-7}$	N/A	[N/A, N/A]	0.014		No
4	N/A	1	12837153	0.006	0.001	$6.56 \cdot 10^{-9}$	10.69	[4.10, 27.85]	N/A	N/A	Yes
4	rs34517537	6	25769359	0.003	0.000	$5.12 \cdot 10^{-7}$	N/A	[N/A, N/A]	0.000		Yes
4	rs145985036	11	119045520	0.009	0.000	$1.16 \cdot 10^{-18}$	N/A	[N/A, N/A]	0.000		Yes
4	rs138169747	11	74800380	0.003	0.000	$5.12 \cdot 10^{-7}$	N/A	[N/A, N/A]	0.000		No
4	rs116589141	12	111336788	0.003	0.000	$5.02 \cdot 10^{-7}$	N/A	[N/A, N/A]	0.000		Yes
4	rs4934	14	95080803	0.460	0.507	$4.43 \cdot 10^{-7}$	0.83	[0.77, 0.89]	0.477		No

Table B.3: SNPs with association p-value $\leq 10^{-6}$ in the datasets of Table B.1. *RS#* is the RS number of the variant in dbSNP, with *N/A* denoting that the variant is not in dbSNP. *C#* is the chromosome number. *OR* is the allelic OR, with *N/A* denoting an undefined OR. *OR 95% CI* is the corresponding 95% confidence interval. *MAF (cases)*, *MAF (controls)*, *MAF (EA)*, and *MAF (EUR)*, denotes the MAF in cases, controls, the *European American (EA)* population on the Exome Variant Server, and the *European (EUR)* population in The 1000 Genomes Project, respectively. An *N/A* MAF (EA) or MAF (EUR) denotes that the MAF of the variant in EA and EUR, respectively, was looked up but was not available. A blank MAF (EUR) denotes that the MAF of the variant in EUR was not looked up, because the MAF in EA was available.

Bibliography

Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544, 2006.

Alexander V Alekseyenko, Nikita I Lytkin, Jizhou Ai, Bo Ding, Leonid Padyukov, Constantin F Aliferis, and Alexander Statnikov. Causal graph-based analysis of genome-wide association data in rheumatoid arthritis. *Biology direct*, 6(1):1, 2011.

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research, Special Topic on Causality*, 11:171–234, 2010a.

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification. part ii: Analysis and extensions. *Journal of Machine Learning Research, Special Topic on Causality*, 11:235–284, 2010b.

Michael Alpers and L Rail. Kuru and creutzfeldt-jakob disease: clinical and aetiological aspects. *Proceedings of the Australian Association of Neurologists*, 8: 7–15, 1970.

Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, An-

drew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010.

Bruno Aranda, Hagen Blankenburg, Samuel Kerrien, Fiona SL Brinkman, Arnaud Ceol, Emilie Chautard, Jose M Dana, Javier De Las Rivas, Marine Dumousseau, Eugenia Galeota, et al. Psicquic and psiscore: accessing and scoring molecular interactions. *Nature methods*, 8(7):528–529, 2011.

A.P. Armen and I. Tsamardinos. Estimation and control of the false discovery rate of bayesian network skeleton identification. Technical report, Institute of Computer Science, Foundation for Research and Technology - Hellas, January 2014.

Panu Artimo, Manohar Jonnalagedda, Konstantin Arnold, Delphine Baratin, Gabor Csardi, Edouard De Castro, Séverine Duvaud, Volker Flegel, Arnaud Fortier, Elisabeth Gasteiger, et al. Expasy: Sib bioinformatics resource portal. *Nucleic acids research*, page gks400, 2012.

David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5):706–716, 2008.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 289–300, 1995. ISSN 00359246. URL `http://www.jstor.org/stable/2346101`.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001. ISSN 0090-5364.

Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227, 2009.

Sonja Blasche and Manfred Koegl. Analysis of protein–protein interactions using lumier assays. *Virus-Host Interactions: Methods and Protocols*, pages 17–27, 2013.

Giorgos Borboudakis and Ioannis Tsamardinos. Incorporating causal prior knowledge as path-constraints in bayesian networks and maximal ancestral graphs. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1799–1806, 2012.

Giorgos Borboudakis, Sofia Triantafillou, and Ioannis Tsamardinos. Tools and algorithms for causally interpreting directed edges in maximal ancestral graphs. In *Sixth European Workshop on Probabilistic Graphical Models*, 2012.

Daniela Börnigen, Léon-Charles Tranchevent, Francisco Bonachela-Capdevila, Koenraad Devriendt, Bart De Moor, Patrick De Causmaecker, and Yves Moreau. An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28(23): 3081–3088, 2012.

Juliane Bremer, Frank Baumann, Cinzia Tiberi, Carsten Wessig, Heike Fischer, Petra Schwarz, Andrew D Steele, Klaus V Toyka, Klaus-Armin Nave, Joachim Weis, et al. Axonal prion protein is required for peripheral myelin maintenance. *Nature neuroscience*, 13(3):310–318, 2010.

P Brown, M Preece, J-P Brandel, T Sato, L McShane, I Zerr, A Fletcher, RG Will, M Pocchiari, NR Cashman, et al. Iatrogenic creutzfeldt–jakob disease at the millennium. *Neurology*, 55(8):1075–1081, 2000.

Paul Brown, Clarence J Gibbs, Pamela Rodgers-Johnson, David M Asher, Michael P Sulima, Alfred Bacote, Lev G Goldfarb, and D Carleton Gajdusek. Human spongiform encephalopathy: the national institutes of health series of

300 cases of experimentally transmitted disease. *Annals of neurology*, 35(5): 513–529, 1994.

Moira E Bruce, R Gl Will, JW Ironside, I McConnell, D Drummond, A Suttie, L McCardle, A Chree, J Hope, C Birkett, et al. Transmissions to mice indicate that ?new variant?cjd is caused by the bse agent. *Nature*, 389(6650):498–501, 1997.

Hansruedi Büeler, Marek Fischer, Yolande Lang, Horst Bluethmann, Hans-Peter Lipp, Stephen J DeArmond, Stanley B Prusiner, Michel Aguet, and Charles Weissmann. Normal development and behaviour of mice lacking the neuronal cell-surface prp protein. *Nature*, 356(6370):577–582, 1992.

Hansruedi Büeler, Adriano Aguzzi, A Sailer, R-A Greiner, P Autenried, M Aguet, and C Weissmann. Mice devoid of prp are resistant to scrapie. *Cell*, 73(7):1339–1347, 1993.

Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145): 661–678, 2007.

William S Bush, Scott M Dudek, and Marylyn D Ritchie. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 368. NIH Public Access, 2009.

Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39 (suppl 1):D685–D690, 2011.

Tao Chen, Nevin L Zhang, Tengfei Liu, Kin Man Poon, and Yi Wang. Model-based

multidimensional clustering of categorical data. *Artificial Intelligence*, 176(1): 2246–2269, 2012.

Geraldine M Clarke, Carl A Anderson, Fredrik H Pettersson, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

David W Colby and Stanley B Prusiner. Prions. *Cold Spring Harbor perspectives in biology*, 3(1):a006833, 2011.

John Collinge. Prion diseases of humans and animals: their causes and molecular basis. *Annual review of neuroscience*, 24(1):519–550, 2001.

John Collinge, Jerome Whitfield, Edward McKintosh, John Beck, Simon Mead, Dafydd J Thomas, and Michael P Alpers. Kuru in the 21st century?an acquired human prion disease with very long incubation periods. *The Lancet*, 367(9528): 2068–2074, 2006.

John Collinge, Jerome Whitfield, Edward McKintosh, Adam Frosh, Simon Mead, Andrew F Hill, Sebastian Brandner, Dafydd Thomas, and Michael P Alpers. A clinical study of kuru patients with long incubation periods at the end of the epidemic in papua new guinea. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1510):3725–3739, 2008.

SJ Collins, Pascual Sanchez-Juan, CL Masters, GM Klug, Cock van Duijn, Anna Poleggi, M Pocchiari, S Almonti, Natividad Cuadrado-Corrales, J de Pedro-Cuesta, et al. Determinants of diagnostic investigation sensitivities across the clinical spectrum of sporadic creutzfeldt–jakob disease. *Brain*, 129(9):2278–2287, 2006.

Gregory Cooper. An overview of the representation and discovery of causal relationships using bayesian networks. *Computation, causation, and discovery*, pages 4–62, 1999.

Gregory M Cooper and Jay Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9): 628–640, 2011.

Heather J Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

Juan D Correa and Elias Bareinboim. Causal effect identification by adjustment under confounding and selection biases. 2017.

Nello Cristianini and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. 2000.

David Danks, Clark Glymour, and Robert E Tillman. Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems*, pages 1665–1672, 2008.

Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.

Carmen Dering, Claudia Hemmelmann, Elizabeth Pugh, and Andreas Ziegler. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genetic epidemiology*, 35(S1):S12–S17, 2011.

Vanessa Didelez, Svend Kreiner, Niels Keiding, et al. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387, 2010.

Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11 (6):446–450, 2010.

Mathieu Emily, Thomas Mailund, Jotun Hein, Leif Schauser, and Mikkel Heide Schierup. Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17(10):1231–1240, 2009.

David Eppstein, Maarten Löffler, and Darren Strash. Listing all maximal cliques in sparse graphs in near-optimal time. In *International Symposium on Algorithms and Computation*, pages 403–414. Springer, 2010.

Evangelos Evangelou and John PA Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389, 2013.

Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 44(D1):D481–D487, 2016.

A.S. Fast. *Learning the Structure of Bayesian Networks with Constraint Satisfaction*. PhD thesis, University of Massachusetts Amherst, 2010.

T. Fawcett. ROC graphs: Notes and practical considerations for data mining researchers. *HP Laboratories technical report*, 2003.

S.E. Fienberg. The analysis of cross-classified categorical data. 1977.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

Nir Friedman. The bayesian structural em algorithm. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 129–138. Morgan Kaufmann Publishers Inc., 1998.

Stephanie D Gan and Kruti R Patel. Enzyme immunoassay and enzyme-linked immunosorbent assay. *Journal of Investigative Dermatology*, 133(9):1–3, 2013.

Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

Jacqueline I Goldstein, Andrew Crenshaw, Jason Carey, George B Grant, Jared Maguire, Menachem Fromer, Colm O?Dushlaine, Jennifer L Moran, Kimberly Chambert, Christine Stevens, et al. zcall: a rare variant caller for array-based genotyping genetics and population analysis. *Bioinformatics*, 28(19):2543–2545, 2012.

Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17 (6):333–351, 2016.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

Megan L Grove, Bing Yu, Barbara J Cochran, Talin Haritunians, Joshua C Bis, Kent D Taylor, Mark Hansen, Ingrid B Borecki, L Adrienne Cupples, Myriam Fornage, et al. Best practices and joint calling of the humanexome beadchip: the charge consortium. *PloS one*, 8(7):e68095, 2013.

Jemila S Hamid, Pingzhao Hu, Nicole M Roslin, Vicki Ling, Celia MT Greenwood, and Joseph Beyene. Data integration in genetics and genomics: methods and challenges. *Human genomics and proteomics*, 1(1), 2009.

Bing Han, Meeyoung Park, and Xue-wen Chen. A markov blanket-based method for detecting causal snps in gwas. *BMC bioinformatics*, 11(3):1, 2010.

Bing Han, Xue-wen Chen, and Zohreh Talebizadeh. Fepi-mb: identifying snps-disease association using a markov blanket-based approach. *BMC bioinformatics*, 12(12):1, 2011.

Traver Hart, H Kiyomi Komori, Sarah LaMere, Katie Podshivalova, and Daniel R Salomon. Finding the active genes in deep rna-seq gene expression studies. *BMC genomics*, 14(1):778, 2013.

Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, et al. The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463, 2013.

David G Hendrickson, Daniel J Hogan, Daniel Herschlag, James E Ferrell, and Patrick O Brown. Systematic identification of mrnas recruited to argonaute 2 by specific microrna and corresponding changes in transcript abundance. *PloS one*, 3(5):e2126, 2008.

Andrew F Hill and John Collinge. Subclinical prion infection. *Trends in microbiology*, 11(12):578–584, 2003.

Andrew F Hill, Melanie Desbruslais, Susan Joiner, Katie CL Sidle, Ian Gowland, John Collinge, Lawrence J Doey, and Peter Lantos. The same prion strain causes vcjd and bse. *Nature*, 389(6650):448–450, 1997.

Andrew F Hill, Susan Joiner, Jackie Linehan, Melanie Desbruslais, Peter L Lantos, and John Collinge. Species-barrier-independent prion replication in apparently resistant species. *Proceedings of the National Academy of Sciences*, 97(18): 10248–10253, 2000.

C. Huang and A. Darwiche. Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning*, 15(3):225–263, 1996.

Rachael P Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J Martin, and Claire O'Donovan. The goa database: gene ontology annotation updates for 2015. *Nucleic acids research*, 43(D1): D1057–D1063, 2015.

Xia Jiang and Richard E Neapolitan. Mining pure, strict epistatic interactions from high-dimensional datasets: Ameliorating the curse of dimensionality. *PloS one*, 7(10):e46771, 2012.

Pall F Jonsson, Tamara Cavanna, Daniel Zicha, and Paul A Bates. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC bioinformatics*, 7(1):1, 2006.

Diego Kaski, Simon Mead, Harpreet Hyare, Sarah Cooper, Ravi Jampana, James Overell, Richard Knight, John Collinge, and Peter Rudge. Variant cjd in an individual heterozygous for prnp codon 129. *The Lancet*, 374(9707):2128, 2009.

Xiayi Ke, Sarah Hunt, William Tapper, Robert Lawrence, George Stavrides, Jilur Ghori, Pamela Whittaker, Andrew Collins, Andrew P Morris, David Bentley, et al. The impact of snp density on fine-scale patterns of linkage disequilibrium. *Human Molecular Genetics*, 13(6):577–588, 2004.

Samuel Kerrien, Sandra Orchard, Luisa Montecchi-Palazzi, Bruno Aranda, Antony F Quinn, Nisha Vinod, Gary D Bader, Ioannis Xenarios, Jérôme Wojcik, David Sherman, et al. Broadening the horizon–level 2.5 of the hupo-psi format for molecular interactions. *BMC biology*, 5(1):44, 2007.

Samantha Kleinberg and George Hripcsak. A review of causal inference for biomedical informatics. *Journal of biomedical informatics*, 44(6):1102–1112, 2011.

Samantha Kleinberg and Bud Mishra. The temporal logic of causal structures. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 303–312. AUAI Press, 2009.

Ana Kozomara and Sam Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73, 2014.

Vasileios Lapatas, Michalis Stefanidakis, Rafael C Jimenez, Allegra Via, and Maria Victoria Schneider. Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki*, 22(1):1, 2015.

Guillaume Launay, R Salza, D Multedo, Nicolas Thierry-Mieg, and Sylvie Ricard-Blum. Matrixdb, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. *Nucleic acids research*, page gku1091, 2014.

Nicole A Lazar, Beatriz Luna, John A Sweeney, and William F Eddy. Combining brains: a survey of methods for statistical pooling of information. *Neuroimage*, 16(2):538–550, 2002.

Pierre Legrain and Luc Selig. Genome-wide protein interaction maps using two-hybrid systems. *FEBS letters*, 480(1):32–36, 2000.

J. Lemeire, S. Meganck, F. Cartella, and T. Liu. Conservative independence-based causal structure learning in absence of adjacency faithfulness. *International Journal of Approximate Reasoning*, 2012.

Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM, 2002.

J. Li and Z. J. Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *J. Mach. Learn. Res.*, 10:475–514, 2009. ISSN 1532-4435.

Anthony M Liekens, Jeroen De Knijf, Walter Daelemans, Bart Goethals, Peter De Rijk, Jurgen Del-Favero, et al. Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol*, 12(6):R57, 2011.

Sarah E Lloyd, Obia N Onwuazor, Jonathan A Beck, Gary Mallinson, Martin Farrall, Paul Targonski, John Collinge, and Elizabeth MC Fisher. Identification of

multiple quantitative trait loci linked to prion disease incubation period in mice. *Proceedings of the National Academy of Sciences*, 98(11):6279–6283, 2001.

Ana Lukic and Simon Mead. Genome wide association studies and prion disease. *Prion*, 5(3):154–160, 2011.

Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

Ian Martin, Valina L Dawson, and Ted M Dawson. Recent advances in the genetics of parkinson?s disease. *Annual review of genomics and human genetics*, 12:301, 2011.

David Maxwell Chickering and David Heckerman. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Machine learning*, 29(2):181–212, 1997.

Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369, 2008.

Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

Brett A McKinney, David M Reif, Marylyn D Ritchie, and Jason H Moore. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88, 2006.

Simon Mead. Prion disease genetics. *European Journal of Human Genetics*, 14(3): 273–281, 2006.

Simon Mead, Michael PH Stumpf, Jerome Whitfield, Jonathan A Beck, Mark Poulter, Tracy Campbell, James B Uphill, David Goldstein, Michael Alpers, Elizabeth MC Fisher, et al. Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. *Science*, 300(5619):640–643, 2003.

Simon Mead, James Uphill, John Beck, Mark Poulter, Tracy Campbell, Jessica Lowe, Gary Adamson, Holger Hummerich, Norman Klopp, Ina-Maria Rückert, et al. Genome-wide association study in multiple human prion diseases suggests genetic risk factors additional to prnp. *Human molecular genetics*, 21(8):1897–1906, 2012.

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the eleventh international conference on uncertainty in artificial intelligence*, 1995.

Luisa Montecchi-Palazzi, Ron Beavis, Pierre-Alain Binz, Robert J Chalkley, John Cottrell, David Creasy, Jim Shofstahl, Sean L Seymour, and John S Garavelli. The psi-mod community standard for representation of protein modification data. *Nature biotechnology*, 26(8):864–866, 2008.

Yves Moreau and Léon-Charles Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13 (8):523–536, 2012.

Julie A Morris and Martin J Gardner. Statistics in medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British medical journal (Clinical research ed.)*, 296(6632):1313, 1988.

R.E. Neapolitan. *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ, 2004. ISBN 0130125342.

Clément Niel, Christine Sinoquet, Christian Dina, and Ghislain Rocheleau. A survey about methods dedicated to epistasis detection. *Frontiers in genetics*, 6, 2015.

Noa Novershtern, Aviv Regev, and Nir Friedman. Physical module networks: an integrative approach for reconstructing transcription regulation. *Bioinformatics*, 27(13):i177–i185, 2011.

Sandra Orchard, Samuel Kerrien, Sara Abbani, Bruno Aranda, Jignesh Bhate, Shelby Bidwell, Alan Bridge, Leonardo Briganti, Fiona SL Brinkman, Gianni Cesareni, et al. Protein interaction data curation: the international molecular exchange (imex) consortium. *Nature methods*, 9(4):345–350, 2012.

Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. The mintact project?intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, page gkt1115, 2013.

Mark S Palmer, Aidan J Dryden, J Trevor Hughes, and John Collinge. Homozygous prion protein genotype predisposes to sporadic creutzfeldt–jakob disease. 1991.

Kristine A Pattin and Jason H Moore. Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Human genetics*, 124(1):19–29, 2008.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):1–62, 2010.

Eric M Phizicky and Stanley Fields. Protein-protein interactions: methods for detection and analysis. *Microbiological reviews*, 59(1):94–123, 1995.

Stanley B Prusiner et al. Novel proteinaceous infectious particles cause scrapie. *Science*, 216(4542):136–144, 1982.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *Annals of Statistics*, pages 962–1030, 2002.

Marylyn D Ritchie. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Annals of human genetics*, 75(1):172–182, 2011.

Marylyn D Ritchie and William S Bush. Genome simulation: Approaches for synthesizing in silico datasets for human genomics. *Advances in genetics*, 72:1, 2010.

Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.

Matthew E Ritchie, Ruijie Liu, Benilton S Carvalho, and Rafael A Irizarry. Comparing genotyping algorithms for illumina's infinium whole-genome snp beadchips. *BMC bioinformatics*, 12(1):1, 2011.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

J.G. Ruby, C.H. Jan, and D.P. Bartel. Intronic microRNA precursors that bypass Drosha processing. *NATURE-LONDON-*, 448(7149):83, 2007.

Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(suppl 1):D449–D451, 2004.

Joseph Sambrook and David W Russell. Detection of protein-protein interactions using the gst fusion protein pulldown technique. *CSH protocols*, 2006(1), 2006.

Martin H Schaefer, Tiago JS Lopes, Nancy Mah, Jason E Shoemaker, Yukiko Matsuoka, Jean-Fred Fontaine, Caroline Louis-Jeune, Amie J Eisfeld, Gabriele Neumann, Carol Perez-Iratxeta, et al. Adding protein context to the human protein-

protein interaction network to reveal meaningful interactions. *PLoS Comput Biol*, 9(1):e1002860, 2013.

Mike Schmidt, Elizabeth R Hauser, Eden R Martin, Silke Schmidt, et al. Extension of the simla package for generating pedigrees with complex inheritance patterns: environmental covariates, gene-gene and gene-environment interaction. *Statistical applications in genetics and molecular biology*, 4(1):1133, 2005.

Pak C Sham and Shaun M Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346, 2014.

Show-Ling Shyng, John E Heuser, and David A Harris. A glycolipid-anchored prion protein is endocytosed via clathrin-coated pits. *The Journal of cell biology*, 125(6):1239–1250, 1994.

Nives Škunca, Adrian Altenhoff, and Christophe Dessimoz. Quality of computationally inferred gene ontology annotations. *PLoS Comput Biol*, 8(5):e1002533, 2012.

Peter G Smith and Ray Bradley. Bovine spongiform encephalopathy (bse) and its epidemiology. *British Medical Bulletin*, 66(1):185–198, 2003.

Michael D Spencer, Richard SG Knight, and Robert G Will. First hundred cases of variant creutzfeldt-jakob disease: retrospective case note review of early psychiatric and neurological features. *Bmj*, 324(7352):1479–1482, 2002.

P Spirtes, C Meek, and T Richardson. An algorithm for causal inference in the presence of latent variables and selection bias in computation, causation and discovery, 1999, 1999.

P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, prediction, and search*. 2000. ISBN 0262194406.

Peter Spirtes and Thomas Richardson. A polynomial time algorithm for determining dag equivalence in the presence of latent variables and selection bias. In *Proceed-*

ings of the 6th International Workshop on Artificial Intelligence and Statistics, pages 489–500, 1996.

Alexander Statnikov, Nikita I Lytkin, Jan Lemeire, and Constantin F Aliferis. Algorithms for discovery of multiple markov boundaries. *Journal of Machine Learning Research*, 14(Feb):499–566, 2013.

Harald Steck and Volker Tresp. Bayesian belief networks for data mining. In *Proceedings of the 2. Workshop on Data Mining und Data Warehousing als Grundlage moderner entscheidungsunterstützender Systeme*, pages 145–154. Citeseer, 1999.

J.D. Storey and R. Tibshirani. *Estimating the positive false discovery rate under dependence, with applications to DNA microarrays*. 2001.

The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056, 2015.

The UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014.

Daniel W Thomson, Cameron P Bracken, and Gregory J Goodall. Experimental strategies for microrna target identification. *Nucleic acids research*, 39(16):6845–6853, 2011.

Robert E Tillman. Structure learning with independent non-identically distributed data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1041–1048. ACM, 2009.

Robert E Tillman and Peter Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *International Conference on Artificial Intelligence and Statistics*, pages 3–15, 2011.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *J Machine Learn Res*, 16:2147–2205, 2015.

Sofia Triantafilou, Ioannis Tsamardinos, and Ioannis G Tollis. Learning causal structure from overlapping variable sets. In *International Conference on Artificial Intelligence and Statistics*, pages 860–867, 2010.

I. Tsamardinos and G. Borboudakis. Permutation Testing Improves Bayesian Network Learning. *Machine Learning and Knowledge Discovery in Databases*, pages 322–337, 2010.

I. Tsamardinos and L. E. Brown. Bounding the false discovery rate in local bayesian network learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, volume 2, pages 1100–1105. AAAI Press, 2008.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning Journal*, 65: 31–78, 2006.

Ioannis Tsamardinos and Constantin F Aliferis. Towards principled feature selection: relevancy, filters and wrappers. In *AISTATS*, 2003.

Ioannis Tsamardinos, Sofia Triantafillou, and Vincenzo Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. *The Journal of Machine Learning Research*, 98888:1097–1157, 2012.

Leslie G Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

Daniel V Veres, Dávid M Gyurkó, Benedek Thaler, Kristóf Z Szalay, Dávid Fazekas, Tamás Korcsmáros, and Peter Csermely. Comppi: a cellular compartment-specific database for protein–protein interaction network analysis. *Nucleic acids research*, page gku1007, 2014.

Tom S Verma and Judea Pearl. Causal networks: Semantics and expressiveness. 1990.

Haris G Vikis and Kun-Liang Guan. Glutathione-s-transferase-fusion based assays for studying protein-protein interactions. *Protein-protein interactions: methods and applications*, pages 175–186, 2004.

Ioannis S Vlachos, Maria D Paraskevopoulou, Dimitra Karagkouni, Georgios Georgakilas, Thanasis Vergoulis, Ilias Kanellos, Ioannis-Laertis Anastasopoulos, Sofia Maniou, Konstantina Karathanou, Despina Kalfakakou, et al. Diana-tarbase v7. 0: indexing more than half a million experimentally supported mirna: mrna interactions. *Nucleic acids research*, 43(D1):D153–D159, 2015.

Jonathan DF Wadsworth and John Collinge. Update on human prion disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1772(6):598–609, 2007.

Kai Wang, Mingyao Li, and Hakon Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854, 2010a.

Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010b.

Michael E Weale. Quality control for genome-wide association studies. *Genetic Variation: Methods and Protocols*, pages 341–372, 2010.

Irene Weibrecht, Karl-Johan Leuchowius, Carl-Magnus Clausson, Tim Conze, Malin Jarvius, W Mathias Howell, Masood Kamali-Moghaddam, and Ola Söderberg. Proximity ligation assays: a recent addition to the proteomics toolbox. *Expert review of proteomics*, 2014.

H-E Wichmann, C Gieger, T Illig, MONICA/KORA study group, et al. Kora-

gen-resource for population genetics, controls and a broad spectrum of disease phenotypes. *Das Gesundheitswesen*, 67(S 01):26–30, 2005.

Robert G Will, JW Ironside, M Zeidler, K Estibeiro, SN Cousens, PG Smith, A Alperovitch, S Poser, M Pocchiari, and A Hofman. A new variant of creutzfeldt-jakob disease in the uk. *The Lancet*, 347(9006):921–925, 1996.

Ilka Wittig and Hermann Schägger. Native electrophoretic techniques to identify protein–protein interactions. *Proteomics*, 9(23):5214–5223, 2009.

Jian-Hua Yang, Jun-Hao Li, Shan Jiang, Hui Zhou, and Liang-Hu Qu. Chipbase: a database for decoding the transcriptional regulation of long non-coding rna and microrna genes from chip-seq data. *Nucleic acids research*, 41(D1):D177–D187, 2013.

Andrew Yates, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, et al. Ensembl 2016. *Nucleic acids research*, 44(D1):D710–D716, 2016.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172 (16):1873–1896, 2008.