Learning causal biological networks with generalized Mendelian randomization

Md. Bahadur Badsha[1*] and Audrey Qiuyan Fu[1*]

[1] Department of Statistical Science, Institute for Bioinformatics and Evolutionary Studies, Center for Modeling Complex Interactions, University of Idaho, Moscow, ID, USA.
*Corresponding author (mdbadsha@uidaho.edu; audreyf@uidaho.edu)

Learning causality from biological data remains a challenge. We present MRPC, a novel machine learning algorithm that employs generalized Mendelian randomization and learns a causal biological network with directed edges. Our method has several desirable statistical features: it controls the false discovery rate, and performs robust inference. Using MRPC, we distinguished direct and indirect targets among multiple genes associated with eQTLs, and constructed a network for frequently altered cancer genes.

Whereas experiments (e.g., temporal transcription or protein expression assays, gene knockouts or knockdowns) have been conducted to understand the causal relationships among genes[1, 2], or between an expression quantitative trait loci (eQTL) and its direct and indirect target genes[3], it remains a challenge to learn causality directly from genomic data. Correlation (or association) is often used as a proxy of a potentially causal relationship, but similar levels of correlation can arise from different causal mechanisms (Models 1-4 in **Fig. 1a**). For example, between two genes with correlated expression levels, it is plausible that one gene regulates the other gene (Models 1 and 2 in **Fig. 1a**); it is also plausible that they do not regulate each other directly, but both are regulated by a common genetic variant (Model 3 in **Fig. 1a**).

Correlation between the expression, or any molecular phenotype, of two genes is symmetrical. However, if a genetic variant (e.g., a SNP) is significantly associated with the expression of one of the two genes, then we may assign a directed edge from the variant to the gene, as it is reasonable to assume that the genotype causes changes in the phenotype (expression), not the other way around. This additional, directed edge breaks the symmetry between the two genes, and makes it possible to infer the causal direction (e.g., compare Model 1 and Model 2 in **Fig. 1a**). This is the rationale behind generalized Mendelian Randomization (gMR). Like standard MR, gMR assumes that the alleles of a genetic variant are randomly assigned to individuals in a population, similar to a perturbation experiment performed by Nature[4]. Standard MR has been widely used in epidemiology studies, where genetic variants are used as instrumental variables to facilitate the estimate of causal effect between an exposure and a disease phenotype[4]. It received increasing attention in genetics in recent years[5-10]. Here we generalize standard MR[4], which typically refers to the topology depicted as Model 1 in **Fig. 1a**, and allow genetic variants to have a variety of relationships with the phenotypes[8] (e.g., Models 2-4 in **Fig. 1a**).

However, application of MR in genomics has not been efficient: existing methods generally work with a small number of nodes, may require temporal information, or are not easily generalizable to large graphs. In machine learning, on the other hand, a class of algorithms, such as those based on the classic PC algorithm[11-15], have been developed to efficiently learn causal graphs for a large number of nodes. These algorithms typically consist of two main steps (**Fig.**

1b): i) *inferring the graph skeleton* through a series of statistical independence tests. The graph skeleton is the same as the final graph except that the edges are undirected; and ii) *determining the direction of the edges* in the skeleton. Variants of the original PC algorithm have been developed to reduce the impact of the ordering of the nodes on inference (e.g., the R package pcalg[14, 15]), or to reduce the number of statistical tests needed for inferring the skeleton (e.g., the R package bnlearn[12, 13]).

Here we incorporate gMR into PC algorithms and aim to learn a causal graph where the nodes are genetic variants and molecular phenotypes (such as gene expression), and where the edges between nodes are undirected or directed, with the direction indicating causality. Crucially, by combining gMR with machine learning, our method is efficient and accurate. gMR can be thought of as a way of introducing useful constraints in graph learning and effectively reducing the search space of possible topologies. Our method builds on five basic topologies that describe the causal relationship in a triplet of one genetic variant and two molecular phenotypes (**Fig. 1a**). One of these basic topologies is the null model, where the two molecular phenotypes are marginally and conditionally independent (Model 0 in **Fig. 1a**). All the non-null topologies can produce similar correlation between the two phenotypes (Models 1-4 in **Fig. 1a**).

Our method, namely MRPC, is a novel causal graph inference method for genomic data (**Fig. 1b**; **Supplementary Fig. 1)**. This method analyzes a data matrix with each row being an individual, and each column a genetic variant or a molecular phenotype. Our method also consists of the two main steps as described above. The first step of learning the graph skeleton is similar to that of other PC algorithms, but with an online control of the false discovery rate (FDR), which is explained in detail below. We incorporated gMR in the second step of edge orientation (**Fig. 1b**), which involves three scenarios: i) MRPC first identify edges involving the genetic variants and orient these edges to point to the molecular phenotype; ii) MRPC then looks for three nodes with a potential V-structure (e.g., Model 2 in **Fig. 1a**, or among three molecular phenotypes, $T_1 \rightarrow T_2 \leftarrow T_3$). MRPC conducts additional conditional independence tests if no such test has been performed in the first step; and iii) among the remaining edges, MRPC iteratively finds node triplets with only one undirected edge. It examines the results from the independence tests from the first step to identify which of the five basic topologies is consistent with the test results for this triplet. In MRPC we use Fisher's z transformation for Pearson's correlation in all the marginal test and for the partial correlation in all the conditional test, consistent with the default test in pcalg[14]. However, other parametric or nonparametric tests for marginal and conditional independence tests may be performed in place of Fisher's z transformation test.

Existing PC algorithms (such as those implemented in pcalg and bnlearn) control the type I error rate for each individual statistical test, but not the family-wise error rate (FWER) or the FDR, as controlling both the FWER and FDR requires the knowledge of the total number of tests, which is not known in advance in graph learning. Lack of correction for multiple comparison often leads to too many edges being retained in the inferred graph, especially when the graph is large (see our simulation results below). We implemented in MRPC the LOND method for controlling the FDR in an online matter[16] (see Methods). The LOND method estimates the expected FDR conditioned on the number of tests performed so far and the number of rejections for these tests.

Furthermore, genomic data may contain outliers[17], which can greatly distort the inferred graph (see our simulation results below). Like pcalg, MRPC uses the correlation matrix, rather than the individual-feature matrix, as the input. We implemented in MRPC a method for calculating the robust correlation matrix[17] (see Methods) to alleviate the impact of outliers.

We assessed the performance of MRPC through synthetic data. We simulated data using linear models for the five basic topologies, three common topologies in biology (such as star, multi-parent, and layered graphs), as well as a complex topology with over 20 nodes (**Fig. 2a**). We varied the sample size, as well as the signal strength through the coefficients in the linear models (see Methods), learned the graph from the synthetic data, and calculated the adjusted Structural Hamming Distance (aSHD) between the inferred graph and the truth (see Methods). For each topology, we compared the mean aSHDs (with standard deviation) over 1000 data sets simulated with different combinations of signal strength and sample size across three methods: MRPC, the pc method (implemented in pcalg) and the mmhc method (implemented in bnlearn) (**Fig. 2b**; **Supplementary Table 1)**. Smaller aSHD corresponds to higher accuracy. MRPC performs better than pc and mmhc in most cases, and recovers the true graph particularly well at moderate or stronger signal with a medium or larger sample size. For the complex topology, MRPC performs consistently better than pc and mmhc. In the presence of outliers, MRPC also outperforms pc and mmhc (**Supplementary Fig. 2)**. In terms of runtime, MRPC is essentially the same as pc, both of which a bit slower than mmhc. Incidentally, mmhc does not handle missing values, which are common in genomic data.

We next applied MRPC to two causal inference problems that are common in biology. First, we are interested in identifying true targets when a single SNP is statistically associated with the expression of multiple genes. Multiple genes are potential targets often because these genes are physically close to one another on the genome, and the eQTL analysis usually examines the association between one SNP-gene pair at a time, ignoring the dependence structure among the genes. Indeed, among the eQTLs identified from the GEUVADIS data[18] (i.e., gene expression measured in lymphoblastoid cell lines of a subset of individuals genotyped in the 1000 Genomes Project), 62 eQTLs discovered under the most stringent criteria have more than one associated gene (see Methods). We applied MRPC to each of these eQTLs and their associated genes in the 373 Europeans, and identified 11 types of topologies (**Supplementary Table 2; Supplementary Fig. 3;** also see comparison with mmhc and pc in **Supplementary Fig. 4)**. Three of these 11 types are Models 1, 3 and 4 shown in **Fig. 1a**. Seven other topologies are identified for eight eQTLs each with three associated genes (**Supplementary Table 2)**. Although the multiple associated genes of the same eQTL are physically near one another, our method managed to tease apart the different dependence (or regulatory relationships) among these genes. For example, the SNP rs479844 (chr11:65,784,486), one of the 62 eQTLs, turns out to be significant in at least three genome-wide association studies (GWASs) for atopic march and more specifically, atopic dermatitis (p values ranging from 10^{-10} to 10^{-18})[19-22]. This SNP has been linked with two genes, AP5B1 (chr11:65,775,893-65,780,802) and OVOL1 (chr11:65,787,022-65,797,219), in these GWASs, but it is unclear which is the real target. Our MRPC infers Model 1 for the triplet: rs479844→OVOL1→AP5B1, which suggests that OVOL1 is more likely to be the direct target, and AP5B1 the indirect one. Meanwhile, for HLA-DQA1 (chr6:32,637,403-32,654,846) and HLA-DQB1 (chr6:32,659,464-32,666,689), both genes are associated with the SNP rs9274660 and located in the major histocompatibility (MHC) region of high linkage

disequilibrium. As expected, MRPC infers an undirected edge between the two genes, as the information on the two genes is highly symmetric in the genotype and gene expression data. By contrast, mmhc and pc often mis-specify edges or their directions (**Supplementary Fig. 4**). We focused on the European sample in this analysis, as the sample size of the Africans is small (89). However, we managed to replicate part of the topologies for the few eQTLs discovered in both populations (**Supplementary Note**).

Secondly, we applied MRPC to the genomic data of breast cancer patients from the TCGA cohort[23], with the aim of learning the causal gene regulatory network in breast cancer. Copy number variation usually has a much stronger effect on gene expression than SNPs do in breast cancer. We therefore used the copy number variation as the genotype, and gene expression as the molecular phenotype. Similar to an earlier investigation[24], we extracted 85 frequently altered genes (e.g., BRCA1, BRCA2, TP53, etc.) in breast cancer and their copy number variation data. We calculated the Pearson correlation matrix (**Fig. 3a**), and applied MRPC at the FDR of 5% (**Fig. 3b**), and subsequently at 0.01, 0.10 and 0.15. The inferred graphs appeared reasonably stable: each graph contains around 200 edges; when the FDR changes by 0.05, the number of edges inferred differently tends to be around 10, which is roughly 1/20 of all the edges (**Fig. 3c**; **Supplementary Figs. 5-7**); this is consistent with the change of 0.05 in FDR, as this rate implies that on average roughly 5% of all the edges are likely to be false positives and therefore would not be consistently inferred at another FDR. In other words, most of the edges are inferred reliably across different FDR levels.

In the graph inferred at the FDR of 5%, one gene (MAML2) has three targets (NFIB, MET, and PIK3R1), followed by nine genes (ATM, CANT1, ELK4, ERCC4, IL6ST, KMT2C, KMT2E, MAP3K1, and MET) with two targets, 31 genes with one target, and 44 without targets (**Fig. 3b**). For better visualization of the result, we then applied WGCNA[25] to the inferred graphs. We experimented with several module sizes, and in the end divided the graph into modules with at least seven nodes (including four genes) per module, such that all the visibly large clusters of nodes were represented (**Fig. 3b**; **Supplementary Fig. 8**; **Supplementary Table 3**). Genes have higher connectivity within the module than with other modules, although most modules have edges connecting one another, consistent with the notion that multiple biological pathways are involved, and possibly interacting in cancer[24] (**Fig. 3b**). We ran gene ontology (GO) enrichment analysis[26] on the genes in each module (excluding the grey nodes, which are not allocated to any module). Except for the green module, which contains only four genes, each module is significantly associated with distinct biological processes or PANTHER pathways[27], suggesting that the causal network we learned has a structure consistent with the underlying biological functions (**Fig. 3b**; **Supplementary Tables 4, 5**).

Additionally, our results also illustrate that causal inference distinguishes ‘direct’ from ‘indirect’ correlation. For example, following hierarchical clustering, the correlation heatmap indicates that NF1, ERCC4, and TRIP11 have higher correlation with one another and are therefore clustered together (**Fig. 3a**). However, no edge connects NF1 and ERCC4 at any of the FDRs we examined (**Fig. 3b**; **Supplementary Figs. 5-7**). A closer look at the conditional independence tests showed that the relatively strong correlation between the two genes is explained away by TRIP11, BRCA1 and the copy number variation of NF1 (p value: 3e-6; the significance level after accounting for the FDR: 5e-7). In other words, the correlation between NF1 and ERCC4 is

indirect: it is induced by their association with three other nodes. Indeed, there is no interaction between NF1 and ERCC4 in the literature to the best of our knowledge. Instead, NF1 has been shown to interact with the KMT2 family[28], also shown in our inferred network (Fig. 3b), whereas the DNA repair gene ERCC4 is recently shown to be involved in the translesion DNA synthesis together with TRIP11 and other genes[29], consistent with the edge between these two genes in our inferred network (Fig. 3b).

In summary, our MRPC method extends standard MR and takes advantage of the development of machine learning algorithms for causal graph inference. MPRC integrates genotypes with molecular phenotypes, and can efficiently and accurately learn causal networks. Our method is flexible as it requires only the genotype data (SNPs or other types of variants) and the molecular phenotype measurements (gene expression, or other functional data, such as exon expression, RNA editing, DNA methylation, etc.), and can be applied to a variety of causal inference problems. By incorporating generalized MR, MRPC makes it possible to examine how genotypes induce changes in phenotypes. Our method is nonparametric in that no explicit distributions are assumed for the underlying graph. Our method resolves several issues present in popular implementations of causal graph inference algorithms, which have led to large biases in inferred graphs. Like most causal graph learning methods, a key assumption behind MRPC is that there are no hidden nodes that are connected to the observed nodes in the graph. Whereas this assumption may not hold in biology, we can take additional measures to alleviate the impact of hidden nodes. For example, genes are often grouped in clusters that tend to have higher correlation within the cluster. Our method can be applied to genes within a gene cluster to build the detailed causal network. As the next step, we are working on extensions of MRPC to deal with hidden variables[8, 14]. The current version of MRPC has already demonstrated its power in tackling several biological problems on causality and in effectively using existing large amounts of genomic data.

1. Segal, E. et al. *Nat Genet* **34**, 166-176 (2003).
2. Housden, B.E. et al. *PLoS Genet* **9**, e1003162 (2013).
3. Cheung, V.G. & Spielman, R.S. *Nat Rev Genet* **10**, 595-604 (2009).
4. Davey Smith, G. & Hemani, G. *Hum Mol Genet* **23**, R89-98 (2014).
5. Zhu, Z. et al. *Nat Genet* **48**, 481-487 (2016).
6. Hill, S.M. et al. *Nat Methods* **13**, 310-318 (2016).
7. Zhang, B. et al. *Cell* **153**, 707-720 (2013).
8. Chaibub Neto, E. et al. *Genetics* **193**, 1003-1013 (2013).
9. Gutierrez-Arcelus, M. et al. *Elife* **2**, e00523 (2013).
10. Oren, Y., Nachshon, A., Frishberg, A., Wilentzik, R. & Gat-Viks, I. *Elife* **4** (2015).
11. Spirtes, P., Glymour, C. & Scheines, R. *The MIT Press* (2000).
12. Tsamardinos, I., Brown, L.E. & Aliferis, C.F. *Machine Learning* **65**, 31-78 (2006).
13. Scutari, M. *J. Stat. Softw.* **35**, 22 (2010).
14. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H. & Bühlmann, P. *J. Stat. Softw.* **47**, 26 (2012).
15. Colombo, D. & Maathuis, M.H. *J. Mach. Learn. Res.* **15**, 3741-3782 (2014).
16. Javanmard, A. & Montanari, A. *arXiv:1502.06197v2 [stat.ME]* (March 5, 2015).
17. Badsha, M.B., Mollah, M.N., Jahan, N. & Kurata, H. *J Biosci Bioeng* **116**, 397-407 (2013).

18. Lappalainen, T. et al. *Nature* **501**, 506-511 (2013).
19. Marenholz, I. et al. *Nat Commun* **6**, 8804 (2015).
20. Paternoster, L. et al. *Nat Genet* **47**, 1449-1456 (2015).
21. Paternoster, L. et al. *Nat Genet* **44**, 187-192 (2011).
22. MacArthur, J. et al. *Nucleic Acids Res* **45**, D896-901 (2017).
23. The Cancer Genome Atlas Network. *Nature* **490**, 61-70 (2012).
24. Wang, X., Fu, A.Q., McNerney, M.E. & White, K.P. *Nat Commun* **5**, 4828 (2014).
25. Langfelder, P. & Horvath, S. *BMC Bioinformatics* **9**, 559 (2008).
26. Gene Ontology Consortium. *Nucleic Acids Res* **43**, D1049-1056 (2015).
27. Mi, H., Muruganujan, A. & Thomas, P.D. *Nucleic Acids Res* **41**, D377-386 (2013).
28. Rao, R.C. & Dou, Y. *Nature reviews. Cancer* **15**, 334-346 (2015).
29. Ziv, O. et al. *Nat Commun* **5**, 5437 (2014 Nov 25).

METHODS

Calculation of robust correlation. We implemented the method in Badsha et al.[17] to calculate the robust correlation matrix as the input to the MRPC inference. Specifically, for data that are approximately normal (usually after preprocessing of the data), we calculated iteratively the robust mean vector $\boldsymbol{\mu}$ and the robust covariance matrix $\boldsymbol{v}$ until convergence. At the t+1st iteration,

$$\boldsymbol{\mu}_{t+1} = \frac{\sum_{i=1}^{n}[\varphi_\beta(\boldsymbol{x}_i;\boldsymbol{\mu}_t,\boldsymbol{v}_t)\boldsymbol{x}_i]}{\sum_{i=1}^{n}\varphi_\beta(\boldsymbol{x}_i;\boldsymbol{\mu}_t,\boldsymbol{v}_t)} \quad (1)$$

and

$$\boldsymbol{v}_{t+1} = \frac{\sum_{i=1}^{n}[\varphi_\beta(\boldsymbol{x}_i;\boldsymbol{\mu}_t,\boldsymbol{v}_t)(\boldsymbol{x}_i-\boldsymbol{\mu}_t)(\boldsymbol{x}_i-\boldsymbol{\mu}_t)^T]}{(1+\beta)^{-1}\sum_{i=1}^{n}\varphi_\beta(\boldsymbol{x}_i;\boldsymbol{\mu}_t,\boldsymbol{v}_t)}, \quad (2)$$

where,

$$\varphi_\beta(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{v}) = exp\left(-\frac{\beta}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{v}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \quad (3)$$

In the equations above, $\boldsymbol{x}_i$ is the vector of gene expression in the ith sample, n the sample size, and β the tuning parameter. Equation (3) downweighs the outliers through β, which takes values in [0,1]. Larger β leads to smaller weights on the outliers. When $\beta = 0$, equation (2) is similar to the standard definition of the variance, except that the scalar is 1/n, whereas the unbiased estimator of the variance has a scalar of 1/(n-1). In our applications, we generally use $\beta = 0.005$ if the robust correlation matrix is calculated, consistent with Badsha et al.[17]. When the data matrix contains missing values, we perform imputation using the R package mice[30]. Alternative, one may impute the data using other appropriate methods, and calculate the correlation matrix as the input for MRPC.

Sequential FDR control. We implemented the LOND algorithm that control FDR in an online manner, as we did not know the number of tests beforehand in learning the causal graph. Specifically, consider a sequence of null hypotheses (marginal or conditional independence between two molecular phenotypes) $H(m) = H_1, H_2, H_3, \ldots, H_m$, with corresponding p-values $p(m) = p_1, p_2, p_3, \ldots, p_m$. The LOND algorithm aims to determine a sequence of significance level α_i, such that the decision for the ith test is

$$R_i = \begin{cases} 1, & \text{if } p_i \leq \alpha_i \quad (\text{reject } H_i) \\ 0, & \text{if } p_i > \alpha_i \quad (\text{accept } H_i) \end{cases}.$$

The number of rejections over m tests is then

$$D_{(m)} = \sum_{i=1}^{m} R_i.$$

For the overall FDR to be δ, the significance level α_i is set to be

$$\alpha_i = \delta_i \left[D_{(i-1)} + 1 \right],$$

where the FDR for the *i*th test is

$$\delta_i = \frac{c}{i^a},$$

such that

$$\sum_{i=1}^{\infty} \delta_i = \delta,$$

for integer a>1 and a constant c. We choose a nonnegative sequence δ_i, such that $\sum_{i=1}^{\infty} \delta_i = FDR$. The default value for a is set to 2 in MRPC. At an FDR of 0.05 and $a = 2$, we have

$$\sum_{i=1}^{\infty} \delta_i = \sum_{i=1}^{\infty} \frac{c}{i^2} = c \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{c\pi^2}{6} = 0.05.$$

Then

$$c = \frac{6 * 0.05}{\pi^2} = 0.0304.$$

Values of δ_i and α_i for the first 10 tests are listed in an example given in **Supplementary Table 6.**

Simulation. We generated synthetic data for a variety of graphs, which fall into three categories depending on the complexity (**Fig. 2a**): i) basic topologies of a triplet; ii) topologies common in biological networks: star (i.e., one molecular phenotype has multiple targets); multi-parent (i.e., one molecular phenotype has multiple regulators apart from the genetic variants); and layered; and iii) a complex topology.

In each topology, we generated the data first for the nodes without parents, and then for other nodes. Genetic variants are nodes without parents, and we assume them to be biallelic SNPs with three genotypes 0, 1, and 2. Denote the minor allele frequency by q and assume Hardy-Weinberg equilibrium. Then the genotype of the *i*th variant, V_i , follows a multinomial distribution:

$$\Pr(V_i = 0) = (1-q)^2, \Pr(V_i = 1) = 2q(1-q), \Pr(V_i = 2) = q^2.$$

Denote the *j*th molecular phenotype by T_j and the set of its parent nodes by P, which may be empty, or may include variant nodes or nodes of other molecular phenotypes. We assume that the molecular phenotype T_j follows a normal distribution

$$T_j \sim N(\gamma_0 + \sum_{k \in \mathrm{P}} \gamma_k V_k + \sum_{l \in \mathrm{P}} \gamma_l T_l \,,\, \sigma^2).$$

The variance may be different for different nodes. For simplicity, we use the same value for all the nodes.

We treat undirected edges as bidirected edges and interpret such an edge as an average of the two directions with equal weights. For example, for the undirected edge in Model 4 in **Fig. 1a**, we generate data for $T_1 \rightarrow T_2$:

$$T_1 \sim N\left(\gamma_0 + \gamma_1 V,\ \sigma^2\right);\ T_2 \sim N\left(\gamma_0 + \gamma_1 V + \gamma_2 T_1,\ \sigma^2\right),$$

and separately for $T1 \leftarrow T2$:

$$T_1 \sim N(\gamma_0 + \gamma_1 V + \gamma_2 T_2,\ \sigma^2);\ \ T_2 \sim N(\gamma_0 + \gamma_1 V,\ \sigma^2).$$

We then randomly choose a pair of values with 50:50 probability for each sample.

Calculation of adjusted Structural Hamming Distance (aSHD). The SHD, as is implemented in the R package pcalg[14] and bnlearn[13], counts how many differences exist between two directed graphs. This distance is 1 if an edge exists in one graph but missing in the other, or if the direction of an edge is different in the two graphs. The larger this distance, the more different the two graphs are. We adjusted the SHD to reduce the penalty on the wrong direction of an edge to 0.5. For example, between two graphs $V \rightarrow T_1 \leftarrow T_2$ and $V \rightarrow T_1 \rightarrow T_2$, the SHD is 1 and our aSHD is 0.5. By contrast, between graphs $V \rightarrow T_1 \leftarrow T_2$ and $V \rightarrow T_1$, T_2 (no edge between T_1 and T_2), both the SHD and aSHD are 1. Therefore, our adjustment penalizes the wrong direction less than the wrong inference of the edge.

Analysis of the GEUVADIS data. The GEUVADIS project (http://www.ebi.ac.uk/Tools/geuvadis-das/) performed RNA-seq (gene expression) on 373 Europeans and 89 Africans from the 1000 Genomes Project. The GEUVADIS project combined the gene expression data with the genotype data, and identified eQTLs across the human genome. Among the most stringent set of eQTLs, 62 have more than one target gene. We extracted the genotypes of these eQTLs and the expression of the target genes in the 373 Europeans, and applied MRPC to each eQTL with its target genes.

Analysis of the TCGA breast cancer data. We used the frequently altered genes identified earlier[24] for this analysis. We downloaded the copy number variation, expression and methylation data for these genes from cBioPortal (http://www.cbioportal.org/), which provides processed and normalized data. We also downloaded the clinical data of the breast cancer patients from the TCGA website (https://cancergenome.nih.gov/). Using the clinical data, we selected 566 patients that were ER+, were identified as white, and also had genetic and molecular data, for our analysis.

Software. MRPC is implemented in an R package at https://github.com/audreyqyfu/mrpc.

30. van Buuren, S. & Groothuis-Oudshoorn, K. *J. Stat. Softw.* **45**, 67 (2011).

ACKNOWLEDGMENTS

We thank Jonathan Pritchard, Anand Bhaskar, Towfique Raj and Boxiang Liu for helpful discussions, and Diego Calderon for detailed and constructive comments on an earlier version of the paper. This research is supported by NIH R00HG007368 (to A.Q.F.).

AUTHOR CONTRIBUTIONS

A.Q.F. conceived project. M.B.B. and A.Q.F. developed the method. M.B.B. implemented the software. M.B.B. and A.Q.F. performed all the analyses and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

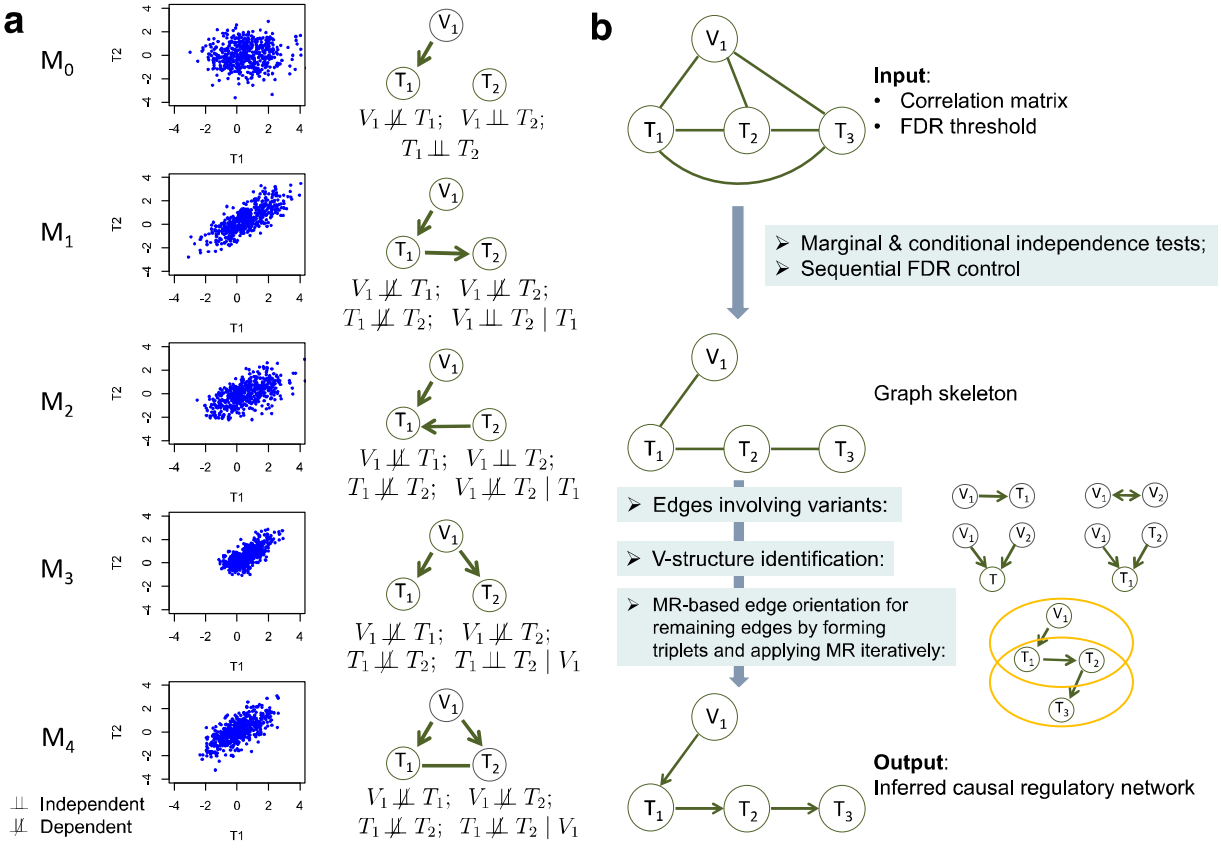
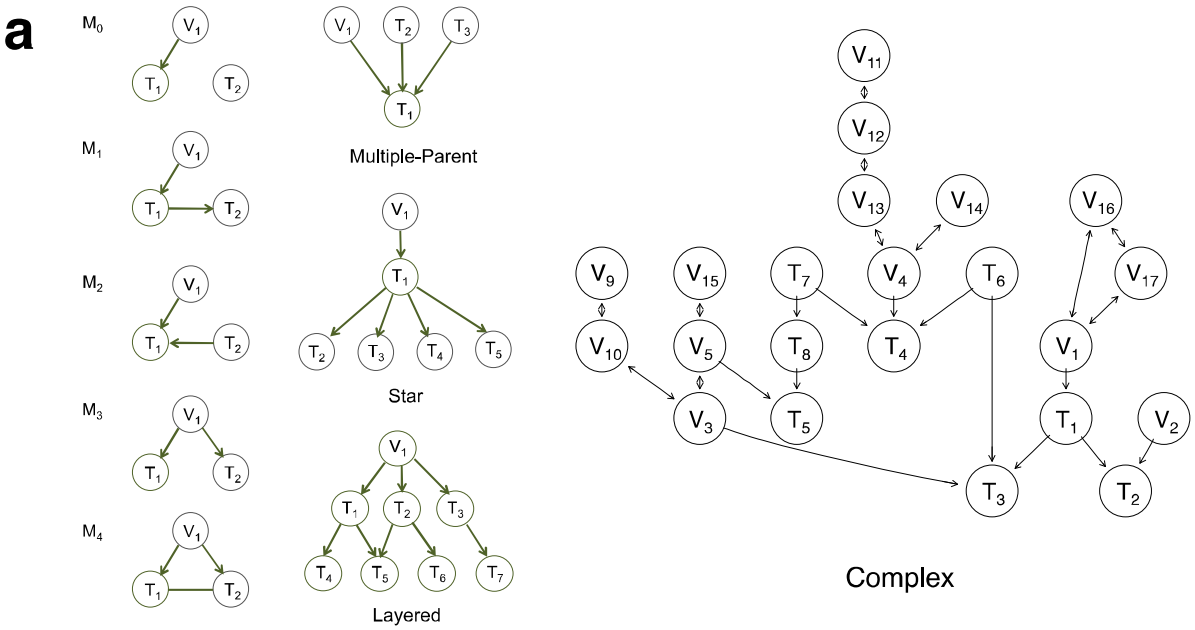


Figure 1: Five basic topologies under generalized Mendelian randomization and the MRPC algorithm. **(a)** Each topology involves three nodes: a genetic variant (V_1), and two molecular phenotypes (T_1 and T_2). Directed edges indicate direction of causality, and undirected edges indicate that the direction is undetermined (or equivalently, both directions are equally likely). For each topology (or model), a scatterplot between the two phenotypes is generated using simulated data, the topology is shown, and the marginal and conditional dependence relationships are given. M_0 is the null model where T_1 and T_2 are marginally independent, and therefore the scatterplot does not show correlation. All the other models show scatterplots with similar levels of correlation. Our MRPC can distinguish the non-null models despite similar correlation. **(b)** The MRPC algorithm takes as input the correlation matrix and FDR threshold. It starts with a fully connected graph, first learns a graph skeleton, whose edges are present in the final graph but are undirected, and then orients these edges. MRPC implements sequential FDR control for skeleton inference and generalized MR for edge orientation.



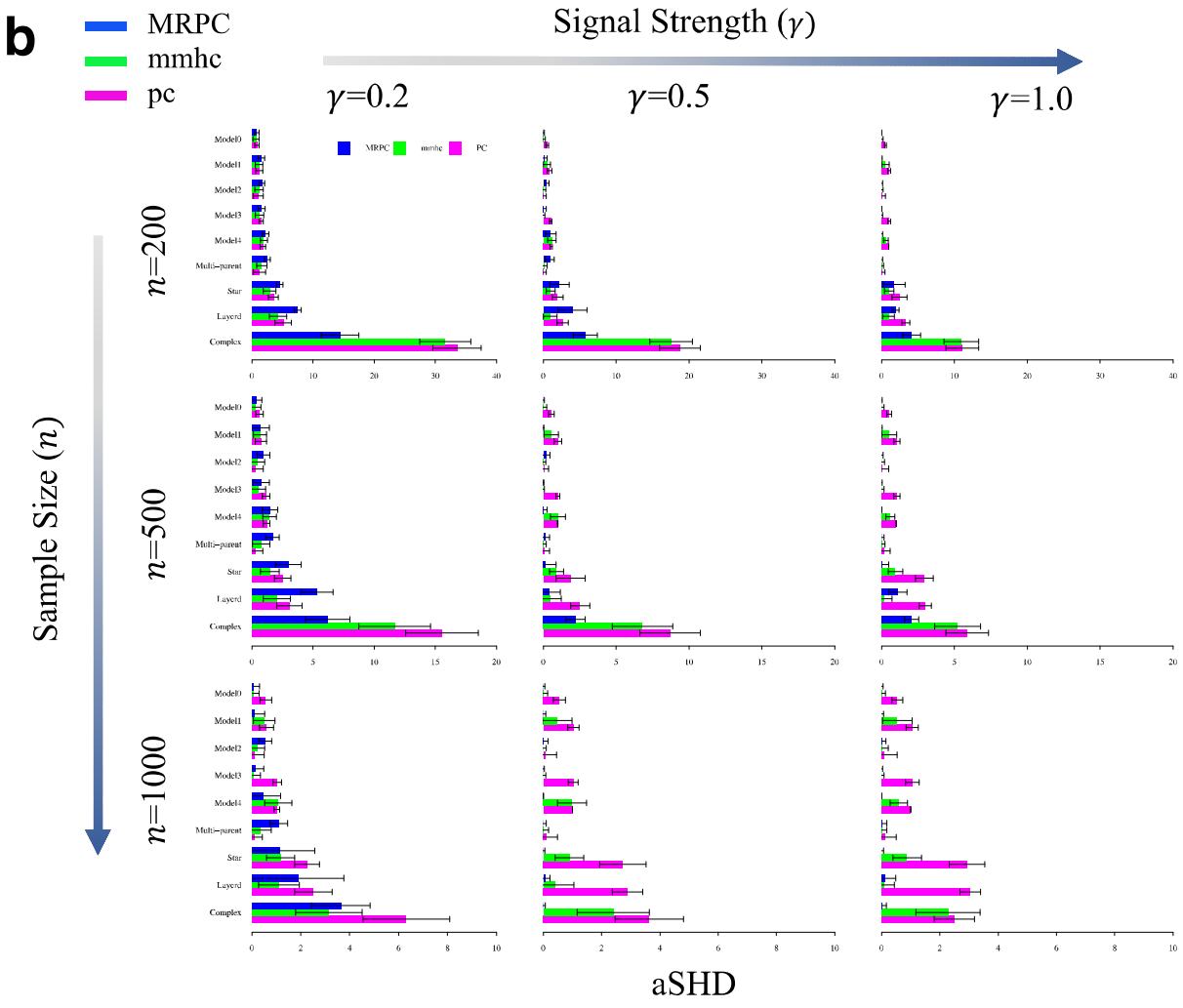


Figure 2: Performance of MRPC on synthetic data. **(a)** Topologies used to generate synthetic data (see Methods for details of simulation). **(b)** Mean adjusted Structural Hamming Distances (aSHDs) for MRPC, mmhc from bnlearn and pc from pcalg on synthetic data. For each of the topologies in (a), 1000 datasets were generated for different signal strengths (γ, which is the coefficient of parent nodes in the linear model; see Methods for simulation details) and different sample sizes (n). Each of the three methods was applied and the aSHD was calculated for the inferred graph relative to the truth. The mean distance over 1000 datasets is represented by the horizontal bar, and the standard deviation the black segment on the bar. Smaller aSHD indicates higher accuracy.

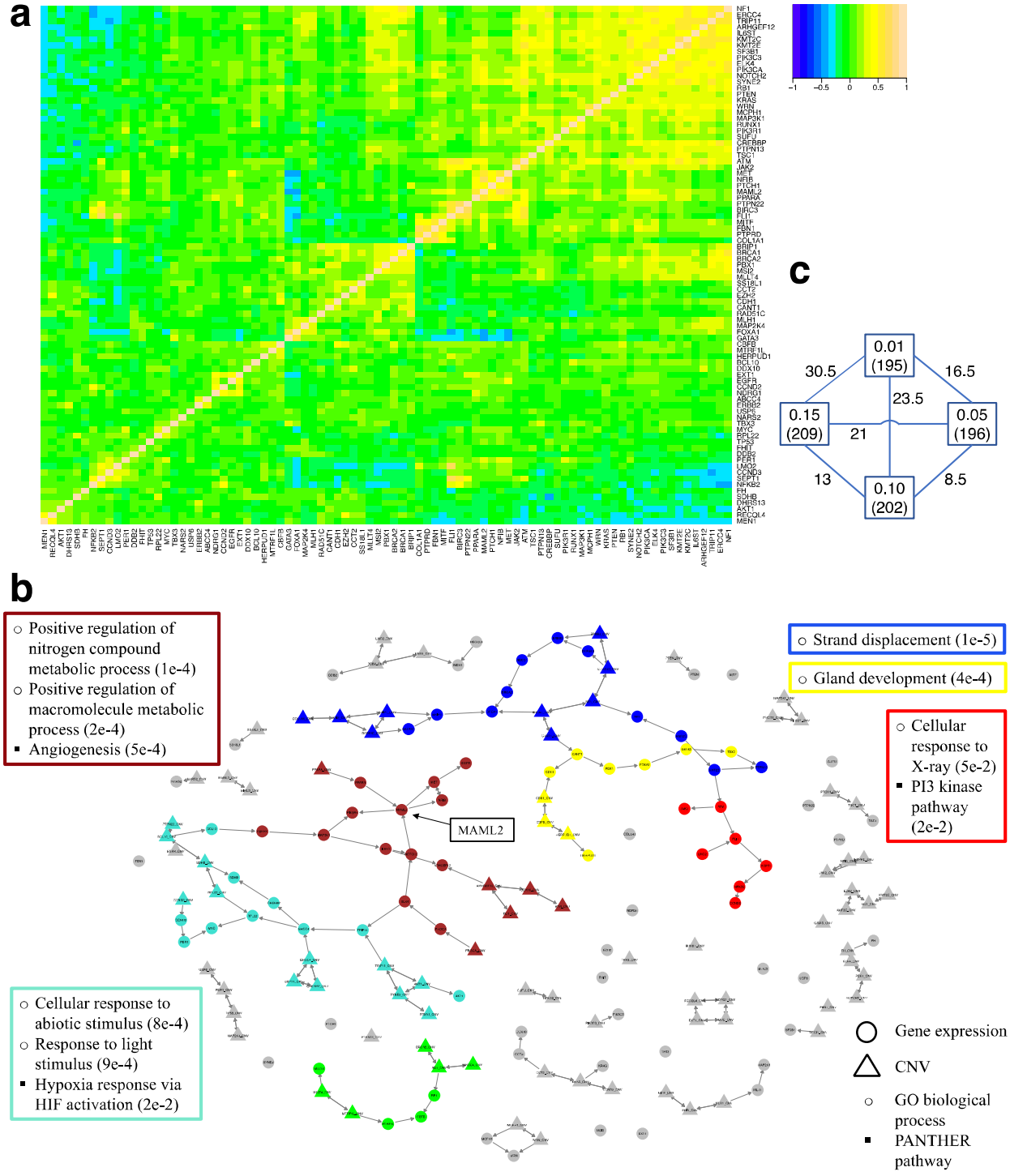


Figure 3: MRPC learns a causal regulatory network for frequently altered cancer genes using the TCGA breast cancer data. **(a)** Pearson correlation heatmap for the 85 genes with hierarchical clustering in rows and columns. **(b)** The causal network inferred at FDR of 5% by MRPC. Modules were identified by WGCNA, such that each non-grey module contains at least seven nodes and four genes. Grey nodes were not assigned to any module. For each module, the box with the corresponding color contains the top GO biological processes and PANTHER

pathways (if exist) enriched for the module, with p values in parentheses (complete results in **Supplementary Tables 4,5**). **(c)** The aSHD between networks inferred by MRPC at different values of FDR. The square indicates the FDR with the total number of edges in parentheses, and the numbers on the lines are the aSHD. These numbers demonstrate the stability of the MRPC inference; see main text for detail.