Comparison between instrumental variable and mediation-based methods for reconstructing causal gene networks in yeast

Adriaan-Alexander Ludl and Tom Michoel*

November 18, 2020

Computational Biology Unit, Department of Informatics, University of Bergen, PO Box 7803, 5020 Bergen, Norway

* Corresponding author, email: tom.michoel@uib.no

Abstract

Causal gene networks model the flow of information within a cell. Reconstructing causal networks from omics data is challenging because correlation does not imply causation. When genomics and transcriptomics data from a segregating population are combined, genomic variants can be used to orient the direction of causality between gene expression traits. Instrumental variable methods use a local expression quantitative trait locus (eQTL) as a randomized instrument for a gene's expression level, and assign target genes based on distal eQTL associations. Mediation-based methods additionally require that distal eQTL associations are mediated by the source gene. A detailed comparison between these methods has not yet been conducted, due to the lack of a standardized implementation of different methods, the limited sample size of most multi-omics datasets, and the absence of ground-truth networks for most organisms. Here we used Findr, a software package providing uniform implementations of instrumental variable, mediation, and coexpression-based methods, a recent dataset of 1,012 segregants from a cross between two budding yeast strains, and the YEASTRACT database of known transcriptional interactions to compare causal gene network inference methods. We found that causal inference methods result in a significant overlap with the ground-truth, whereas coexpression did not perform better than random. A subsampling analysis revealed that the performance of mediation saturates at large sample sizes, due to a loss of sensitivity when residual correlations become significant. Instrumental variable methods on the other hand contain false positive predictions, due to genomic linkage between eQTL instruments. Instrumental variable and mediation-based methods also have complementary roles for identifying causal genes underlying transcriptional hotspots. Instrumental variable methods correctly predicted *STB5* targets for a hotspot centred on the transcription factor *STB5*, whereas mediation failed due to Stb5p auto-regulating its own expression. Mediation suggests a new candidate gene, *DNM1*, for a hotspot on Chr XII, whereas instrumental variable methods could not distinguish between multiple genes located within the hotspot. In conclusion, causal inference from genomics and transcriptomics data is a powerful approach for reconstructing causal gene networks, which could be further improved by the development of methods to control for residual correlations in mediation analyses and genomic linkage and pleiotropic effects from transcriptional hotspots in instrumental variable analyses.

1 Introduction

Causal gene networks model the flow of information from genotype to phenotype within a cell or whole organism [1–4]. Reconstructing causal networks from omics data is challenging because correlation does not imply causation. However, when genomics and transcriptomics data from a large number of individuals in a segregating population are combined, genomic variants can be used to orient the direction of causality between gene expression traits. This is based on the fact that alleles are randomly segregated during meiosis and genotypes remain fixed during an individual's lifetime, such that genomic variants act as causal anchors from which all arrows are directed outward [1, 2, 5]. Moreover, local and distal expression quantitative trait locus (eQTL) associations have biologically distinct interpretations, because genomic variation at regulatory DNA elements leads to altered transcription of nearby genes by *cis*-acting, epigenetic mechanisms, whereas distal associations must be intermediated by *trans*-acting factors [6, 7]

These principles are combined in different ways in two classes of causal inference methods that use genomic variants as causal anchors: *instrumental variable* (known as *Mendelian randomization* in genetic epidemiology) or *mediation*-based [8]. *Mediation* infers the direction of causality between two traits that are statistically associated to the same genomic variant by testing whether the association between the variant and one of the traits is mediated by the other trait, in which case there must be a causal relation from the mediating trait to the other one [9, 10]. Mediation does not require that one of the traits has a "preferential" relation to the genomic variant (as in *cis* or *trans*). However, mediation fails in the presence of high measurement noise or hidden confounders, such as common upstream factors coregulating both traits, where it rejects true interactions (i.e. reports false negatives) [11].

Instrumental variable or *Mendelian randomization* methods assume that the genomic variant acts as a randomized "instrument" for one of the traits, similar to the random assignment of individuals to treatment groups in randomized controlled trials, such that a statistical association between the variant and the second trait is evidence for a causal relation from the first to the second trait. The random group assignment, in genetics the random segregation of alleles, ensures that causal effects can be detected even in the presence of confounding. However, instrumental variable methods fail if there are pathways from the variant to the second trait other than through the first trait (pleiotropic effects) [12–14].

A detailed comparison between these two approaches requires a standardized implementation where pre-processing (e.g. data normalization) and post-processing (e.g. multiple testing correction) are handled uniformly. Previously, we developed *Findr*, a computationally efficient software package implementing six likelihood ratio tests that can be combined in multiple ways to reconstruct instrumental variable as well as mediation-based causal gene networks [11]. *Findr* expresses the result of each test as a posterior probability (one minus the local false discovery rate), allowing tests to be combined by the usual rules of probability theory [10]. This results in causal network inference methods that are representative for the broader field. For instance, the implementation of the mediation-based method in *Findr* is identical to the method of Chen *et al.* [10], which had its roots in the "likelihood-based causal model selection" (LCMS) procedure of Schadt *et al.* [9]. The Causal Inference Test (CIT) software [15, 16] is another implementation of an LCMS-based mediation method, which combines statistical tests using omnibus p-values and FDR estimates. We found previously that it results in similar inferences as the mediation-based method implemented in *Findr* [11]. Instrumental variable methods on the other hand are based on genetic associations, for which *Findr* uses categorical regression of gene expression profiles on genotype values, similar to for instance the ANOVA option in Matrix-eQTL [17].

Using simulated data from the DREAM5 Systems Genetics challenge [18, 19], we found previously that instrumental variable methods generally outperformed mediation-based methods in terms of area under the precision-recall curve, and that the performance of mediation-based methods *decreased* with increasing sample size, due to increased statistical significance of confounding effects [11]. However, at that time, no real-world dataset with sufficient sample size as well as an accurate ground-truth network of causal interactions was available to test these predictions in a real biological system.

Fortuitously, a dataset has now become available of genomic variation and gene expression data in more than 1,000 segregants from a cross between two strains of budding yeast, a popular eukaryotic model organism [20]. By learning networks from these data, and comparing against the wealth of transcriptional regulatory interactions and other functional validation data available for budding yeast [21], a thorough benchmarking of methods for reconstructing causal gene networks has become possible.

2 Methods

2.1 Selecting strongest *cis*-eQTLs

Using the data on expression quantitative trait loci (eQTLs) from [20], we selected the strongest *cis* acting eQTLs for 2884 genes. The eQTLs were ranked in descending order according to the absolute value of the correlation coefficient between scaled expression levels and marker genotype (r, obtained from [20, Source data 4]), and for each gene the highest ranked eQTL was retained. Among the selected eQTLs 2044 occured once, 337 eQTLs were strongest for two genes, 44 were strongest for three genes, 6 were strongest for four genes, 2 were strongest for five genes.

2.2 Network inference methods

We used the inference methods implemented in *Findr* [11]. The source code is available at `https://github.com/lingfeiwang/findr`. The test P_0 only uses gene expression data. For the other tests (P, P_2, P_3, P_5), we used the genotype and gene expression data from [20] (see section 2.7 below for details) with *cis*-eQTLs as causal anchors for the inference tests. Composite tests are obtained by element-wise multiplication of the matrices containing the results of individual tests.

2.3 Performance measures

The Precision-Recall curves and area under the curve (AUPR) for interactions predicted by a given test were computed using the scikit-learn package [22] and three ground-truth matrices (see Data section 2.7). Recall is equivalent to the true positive rate (TPR), i.e. the number of true positive predictions as a fraction of all known positive interactions in the network. Precision or positive predictive value is 1 – FDR where FDR is the global false discovery rate.

AUPR-ratio or fold-change is the AUPR divided by the theoretical value for random predictions on a given ground truth. The latter is obtained as the precision for random predictions given by $prec_{\text{random}} = N_E/(N_R * N_T)$ where N_R is the number of regulating genes, N_T is the number of target genes, N_E is the number of edges, i.e. the number of ones in the ground-truth adjacency matrix.

2.4 Subsampling

We performed subsampling on the segregants to evaluate the change in performance of our inference methods on various sample sizes. Four subsamples of randomly selected segregants were drawn for the following sizes: 10, 100, 200, 400, 600, 800 and 1,000. The inference methods were run on each sample. We report the average AUPR and its statistical standard deviation over the four subsamples in Fig. 4.

2.5 Genotypes covariance and target counts

We computed the covariance matrix of the genotypes at the retained eQTLs for all 1,012 segregants (Fig. 5). The rows and columns of the matrix were reordered according to the genome position of the eQTL, the ordering algorithm is described below in 2.7.

Findr posterior probability matrices were thresholded to obtain discrete networks with an expected FDR target value as described previously [10, 11]: because for each interaction the local false discovery rate is given by $\mathrm{fdr} = 1 - p$, where p is the posterior probability value obtained by the test, the expected FDR of a network consisting of all interactions with $p \geq p_{\mathrm{th}}$ is the average of the local fdr of the retained interactions. We determined p_{th} as the threshold that gave the greatest expected FDR below the target value (5 or 10 %). We counted the number of targets for each source gene whose $p > p_{\mathrm{th}}$.

2.6 Software and Data availability

The inferred regulatory relationships for the thresholds reported in Tab. 2 for the causal tests (P_2P_3, P_2, P_2P_5, P) and scripts to reproduce the analysis are provided in the repository `https://github.com/michoel-lab/FindrCausalNetworkInferenceOnYeast`. Running all Findr inference tests on the data from [20] takes about 10 to 15 seconds on a typical desktop computer.

2.7 Data

We used gene expression data for 5,720 genes and genotypes for a panel of 1,012 segregants from crosses of one laboratory strain (BY) and a wine strain (RM) from [20]. Batch and optical density (OD) effects, as given by the covariates provided in [20], were removed from the expression data using categorical regression, as implemented in the statsmodels python package [23]. The paper also provides data on expression quantitative trait loci (eQTLs) that was used to select the strongest *cis*-eQTLs, as well as a file with annotations to the 102 hotspots that they identified.

For validation we used networks of known transcriptional regulatory interactions in yeast (S. cerevisiae) from YEASTRACT [21]. Regulation matrices were obtained from `http://www.yeastract.com/formregmatrix.php`. We retrieved the full ground-truth matrices containing all reported interactions of the following types from the YEASTRACT website: *DNA binding evidence* was used as the "Binding", *expression evidence* including TFs acting as activators and those acting as inhibitors was used as the "Expression", *DNA binding and expression evidence* was used as the "Binding & Expression". Self regulation was removed from all ground truths. The numbers of regulators, targets and interactions for these three ground-truth networks are shown in Tab. 1.

Ground-Truth Network	N_R	N_T	N_E	N_{sE}
Binding	90	5,151	19,099	28
Binding&Expression	80	3,394	5,680	24
Expression	113	5,369	92,646	77

Table 1: **Properties of the YEASTRACT ground-truth networks.** N_R is the number of regulating genes, N_T is the number of target genes, N_E is the number of edges excluding self-edges, N_{sE} is the number of self-edges. Data was retrieved from YEASTRACT [21].

Annotations of the yeast genome were used to map gene names to their identifiers and order them according to the position of their causal anchor (eQTL) along the full genome, first by chromosome and then by position along the chromosome. The sorting algorithm places mitochondrial genes first (when present) and orders the chromosomes according to the numerical value of the roman numerals. We used the gff3 file (`Saccharomyces_cerevisiae.R64-1-1.83.gff3.gz`) from the Ensembl database (release 83, December 2015), [24], which is the version used by [20]. The file is accessible at `ftp://ftp.ensembl.org/pub/release-83/gff3/saccharomyces_cerevisiae/`.

3 Results

3.1 Findr reconstructs instrumental variable and mediation-based causal gene networks in yeast

We used the software *Findr* [11] to reconstruct causal and non-causal gene networks in yeast from a dataset of genomic variation and expression data for 5,720 genes in 1,012 segregants from a cross between two strains of budding yeast [20]. 2,884 genes had an associated genomic causal anchor, here defined as the variant most strongly associated to the gene and present in the list of genome-wide significant eQTLs whose confidence interval (of variable size) overlaps with an interval covering the gene, 1,000 bp upstream and 200 bp downstream of the gene position [20]. *Findr* implements six likelihood ratio (LLR) tests between triplets (E,A,B), where A and B are genes, and E is the causal anchor for A. For each test i, *Findr* outputs the posterior probability P_i of the selected hypothesis being true (Fig. 1A). These posterior probabilities can then be combined to obtain the posterior probabilities of various compound hypotheses being true. Here we considered four causal tests and one non-causal test to reconstruct directed gene networks:

- *Mediation.* Mediation-based approaches infer a causal interaction $A \rightarrow B$ if gene B is statistically associated to the causal anchor E, and the association is abolished after conditioning on gene A [9,10,16]. In *Findr* this is accomplished by the compound hypothesis that test 2 and 3 are both true, i.e. by the posterior probability $P_2 \times P_3$. Mediation can distinguish true positive (TP) from true negative (TN) causal interactions in the absence of hidden confounders, but will report a false negative (FN) if a real causal interaction is confounded by a hidden factor (Fig. 1B, row 1), due to a collider effect [10, 11].

- *Instrumental variables without pleiotropy.* Instrumental variable approaches assume that the causal anchor E acts as a randomized instrument for gene A, and, in their simplest form, infer

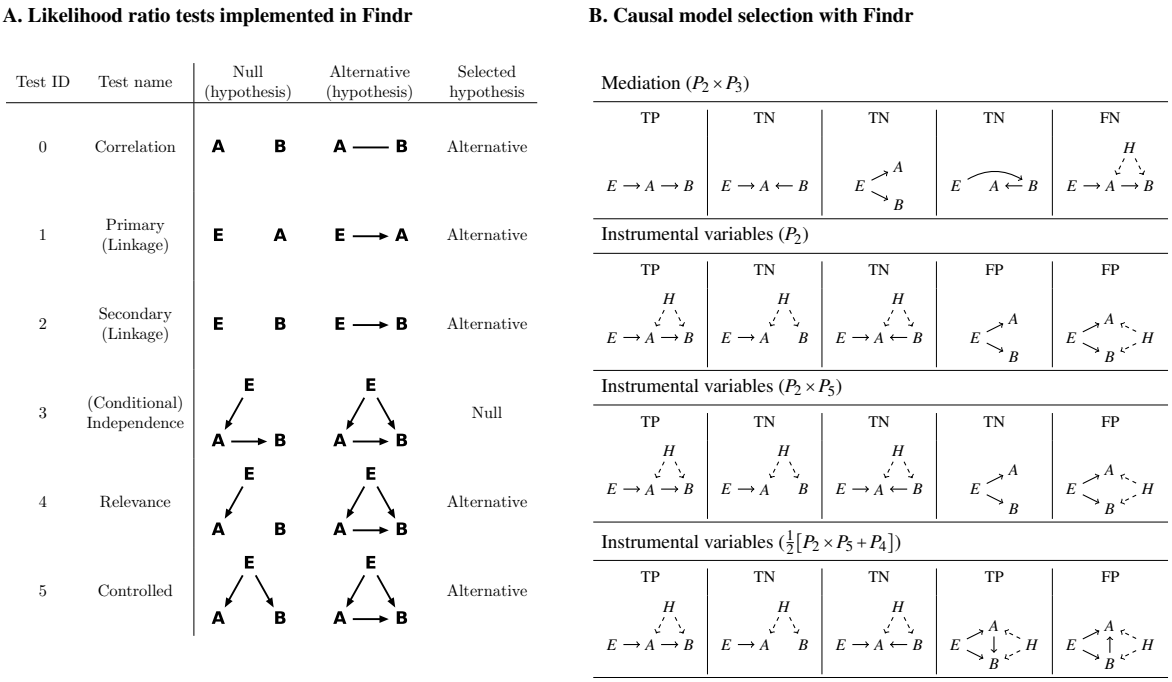

Figure 1: **A. Likelihood ratio (LLR) tests implemented in Findr.** E is a causal anchor of gene A. Arrows in a hypothesis indicate directed regulatory relations. Genes A and B each follow a normal distribution, whose mean depends additively on its regulator(s), as determined in the corresponding hypothesis. The dependency is categorical on genotypes and linear on gene expression levels. The undirected line represents a multi-variate normal distribution between the relevant variables. In each test, either the null or the alternative hypothesis is selected, as shown. Figure ©2017 Wang, Michoel, reproduced by permission from [11] under Creative Commons Attribution License. **B. Causal model selection with Findr.** By combining the posterior probabilities P_i of the selected hypothesis for test i being true, Findr determines whether coexpressed genes A and B are connected by a causal $A \to B$ relation. **(Row 1)** In the absence of hidden confounders (H), mediation-based causal inference, combining Findr tests 2 and 3, correctly identifies true positive (TP; correlation due to causal $A \to B$ relation) and true negative (TN; correlation without causal $A \to B$ relation) models. However, it reports a false negative (FN) if the causal relation is affected by a hidden confounder. **(Row 2)** If the causal anchor is "exclusive" to gene A, then the instrumental variable method based on Findr test 2 correctly identifies TP and TN models, even in the presence of hidden confounding. However, it reports a false positive (FP) if the association between E and B is due to other paths than through A (pleiotropy). **(Row 3)** An instrumental variable method that combines Findr tests 2 and 5 correctly identifies a true negative if the correlation between A and B is entirely due to a pleitotropic effect of E, but will still report a false positive if there is an additional effect from a hidden confounder. **(Row 4)** An instrumental variable method based on the compound hypothesis that test 4 is true, or test 2 and test 5 are true, reports a TP for causal relations where $E \to A \to B$ is not the only path from E to B, with or without confounding, but will report a FP if the true causal relation is $B \to A$ (or absent).

a causal interaction $A \to B$ if gene B is statistically associated to the causal anchor E, i.e. by the posterior probability P_2 that test 2 is true. Instrumental variables can distinguish true positive from true negative causal interactions even in the presence of hidden confounders, but will report a false positive (FP) if there are other pathways than through A that cause a statistical association between E and B (pleiotropy) (Fig. 1B, row 2).

- *Instrumental variables with perfect pleiotropy.* To address the problem of pleiotropy, we can additionally require that genes A and B are not independent after conditioning on E, accomplished

by the compound hypothesis that test 2 and 5 are both true, i.e. by the posterior probability $P_2 \times P_5$. This correctly identifies a true negative if E explains all of the correlation between A and B, but will still result in a false positive if there is a hidden confounder (Fig. 1B, row 3).

- *Instrumental variables with partial pleiotropy.* To overcome the problem of FP predictions in the "confounded pleiotropy" situation, we introduced test 4 in *Findr*, which tests whether gene B is not independent of E and A simultaneously, and found empirically that the combination $P = \frac{1}{2}(P_2P_5 + P_4)$ performs better than $P_2 \times P_5$ alone [11]. In particular, it identifies a TP for causal $A \to B$ relations even in the presence of alternative $E \to B$ paths and hidden confounding, at the expense of FP predictions when the relation is reversed or absent (Fig. 1B, row 4).

- *Coexpression.* As a basic reference, we reconstructed a gene network based on coexpression alone, using *Findr* test 0. Note that the posterior probability P_0 is not symmetric ($P_0(A \to B) \neq P_0(B \to A)$), because it is estimated from the observed distribution of LLR test statistics for each A separately [11].

To illustrate the differences between coexpression, instrumental variable, and mediation-based gene networks, we considered the sub-networks inferred between the 2,884 genes that had a causal anchor (i.e. the sub-network where the probability of an edge can be estimated for *both* edge directions). As expected, the coexpression network (P_0) is largely symmetric (Fig. 2, left), whereas the causal instrumental variable (P_2, Fig. 2, center) and mediation-based (P_2P_3, Fig. 2, right) networks show a clear asymmetric structure with some genes having a very large number of high-confidence targets. These genes correspond to transcriptional hotspots, regions of the genome with a large, genome-wide effect on gene expression [20]. The overall structure of the causal networks appears consistent with the general considerations above. The overall signal (strength of posterior probabilities) is weaker in the mediation-based network, consistent with an increased false negative rate (Fig. 2, right). On the other hand, the instrumental variable network appears to have a genomic structure, where nearby genes are mutually connected and have a similar target profile (Fig. 2, middle). This could be due to genomic linkage between causal anchors: if two genes A and A' share the same or highly correlated instruments E and E', then their predicted target sets would also be very similar, and probably include a large proportion of false positive predictions for either gene.

3.2 Causal gene networks overlap significantly with known transcriptional regulatory networks

We assessed the performance of networks predicted by Findr on three ground-truth networks of transcriptional regulatory interactions in yeast, where targets of a transcription factor (TF) are defined by TF-DNA binding interactions ("Binding" network), differential expression upon TF perturbation ("Expression" network), or the intersection of them ("Binding & Expression" network) (see Methods and Table 1). The precision-recall curves for the four causal inference methods showed the characteristic peak of high precision at low recall indicative of an enrichment of true positives among the predictions with highest posterior probabilities, and confirmed by increased area under the precision-recall curve (AUPR) compared to random predictions (Figure 3). This was markedly the case for the Binding & Expression ground-truth, with AUPR more than 1.3 times higher than random. This is consistent with the notion that genes that are bound by a TF as well as differentially expressed upon TF perturbation are more likely to be real TF targets, that is, that the Binding & Expression ground-truth is of higher quality than the others. Differences between causal inference methods were

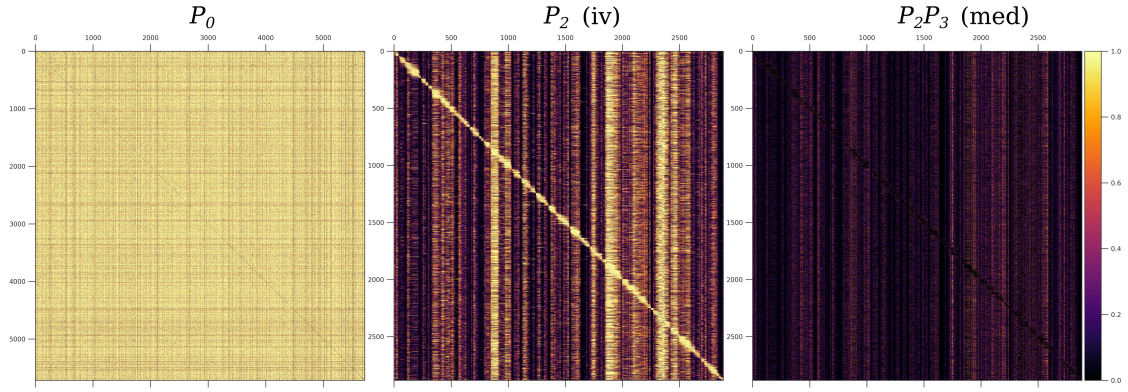

Figure 2: **Matrices of predicted gene interactions.** These square matrices represent the interactions between 2884 genes with causal anchors (eQTLs), posterior probability values are color coded. Vertical bands correspond to hotspots. **Left**: The correlation based test P_0. **Center**: The instrumental variable test P_2. **Right**: The mediation test P_2P_3. The genes are ordered according to the position of their causal anchor in the full yeast genome. See Sup. Fig. S1 for the instrumental variable tests P_2P_5 and P.

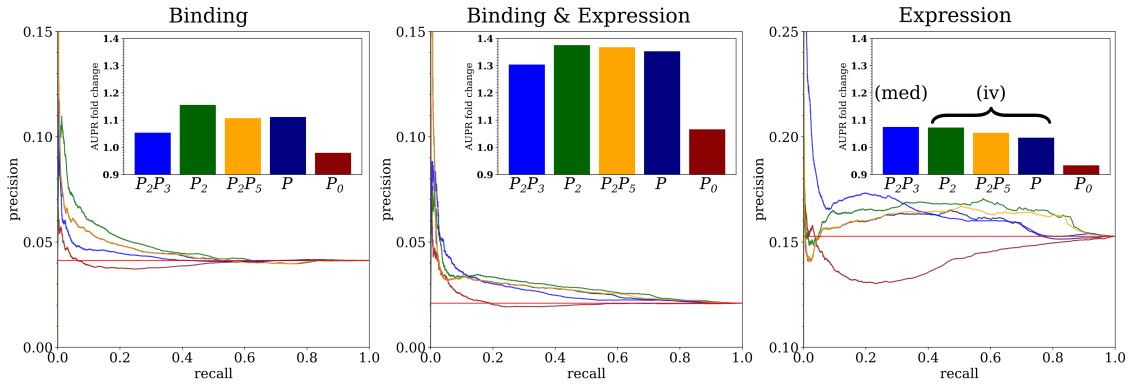

Figure 3: **Performance of causal inference on YEASTRACT ground truths.** Precision-recall curves for four causal inference methods (P_2P_3, P_2, P_2P_5, P) and one coexpression method (P_0) are shown for the Binding (left), Binding and Expression (center) and Expression (right) ground-truth networks. The insets show the area under the precision-recall curves (AUPR) as the the "fold change" relative to the baseline performance for random predictions. The horizontal red line shows the baseline performance for random predictions and is used as reference for AUPR fold change (insets). The network inference methods are described in Section 3.1.

modest, with instrumental variable methods (P_2, P_2P_5, P) showing somewhat better performance than the mediation-based method (P_2P_3) on the Binding and Binding & Expression ground-truths, and vice versa on the Expression ground-truth (Figure 3). In contrast to the causal inference methods, the coexpression-based method (P_0) did not show any improvement over random predictions. This is not surprising. An unbiased evaluation of 35 diverse methods for network inference from expression alone did not find any improvement over random predictions on a comparable ground-truth network for yeast [25].

3.3 The performance of mediation saturates at large sample size

The availability of more than 1,000 segregants in the genotype and gene expression dataset allowed us to evaluate the performance of network inference across sample sizes by random subsampling of the data. The clearest pattern was again observed for the Binding & Expression ground-truth, consisting of the most reliable known transcriptional regulatory interactions, where the three instrumental variable methods (P_2, P_2P_5, P) showed a monotonous increase in AUPR with increasing sample size (Fig. 4). The mediation based method (P_2P_3) initially showed a similar performance as the instrumental variable methods, but saturated above 400 samples when accounting for statistical error and dropped from having the highest to the lowest average performance of all causal inference methods. The same pattern is also observed on the Binding ground-truth, albeit in a less pronounced way, presumably due to lower AUPR values relative to random predictions for all methods.

These results are consistent with previous work on simulated data, where we observed a decrease in performance with increasing sample size for mediation-based methods [11]. There we showed that hidden confounders and measurement noise can lead to a residual correlation between the causal anchor E and a target gene B after adjusting for the regulatory gene A (cf. Fig. 1). At sufficiently large sample size, this residual correlation becomes significantly different from zero and thereby leads to a false negative prediction.

Sample size showed little effect on the coexpression method P_0 for sample sizes larger than 400 for all ground truths. This is consistent with the notion that estimates of correlations will stabilize around their true values at smaller sample sizes than estimates of causal effects.

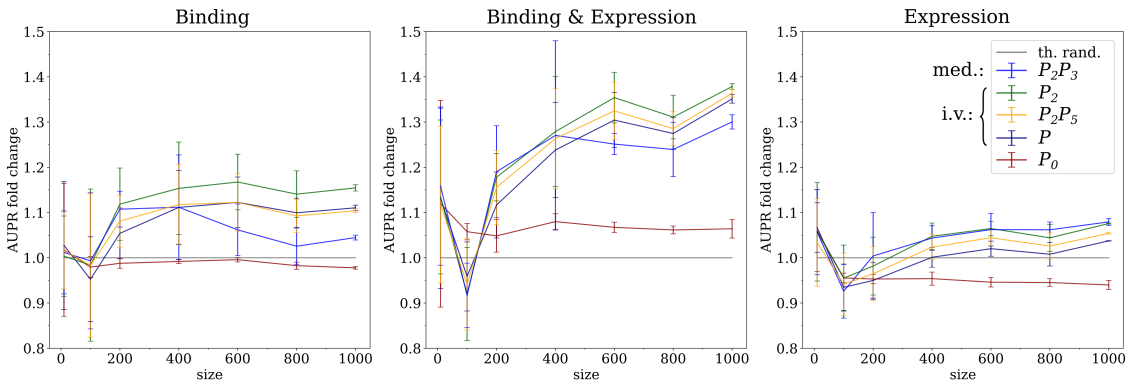

Figure 4: **Performance of causal inference across sample sizes.** AUPR fold change values for four causal inference methods (P_2P_3, P_2, P_2P_5, P) and one coexpression method (P_0) (see Section 3.1) at various sample sizes for the Binding (left), Binding & Expression (center) and Expression (right) ground-truth networks. Four samples were randomly drawn from the expression data and evaluated with each test. Error bars represent the standard deviation across the four subsets. The horizontal grey line indicates the level of random predictions. The fold change is relative to the baseline performance for random predictions.

3.4 Instrumental variable methods are affected by genomic linkage blocks

Next, we assessed the extent to which instrumental variable methods are affected by genomic linkage between causal anchors which would lead to false positive predictions due to (real or apparent) pleiotropic effects (cf. Fig. 1). For instance for the P_2 method, if two genes in the same genomic

neighbourhood have causal anchors with strongly correlated genotype values, the method would predict them to have similar sets of target genes.

To perform the analysis, we first truncated the posterior probability values in order to obtain discrete, directed networks. Thresholds were determined to obtain networks with an expected FDR ≤ 5% (for the instrumental variable methods) or ≤ 10% (for the mediation-based method) (see Methods section). The larger FDR value for mediation was chosen to counterbalance its increased false negative rate, and resulted in posterior probability thresholds that were comparable between all methods (Table 2).

Test	p_{th}	FDR	N_R	N_T	N_E	ρ
P_2P_3	0.8175	0.09953	1,808	5,628	144,091	0.014
P_2	0.825	0.04974	2,884	5,720	2,319,854	0.141
P_2P_5	0.8375	0.04994	2,884	5,719	1,740,251	0.106
P	0.8575	0.04982	2,884	5,720	2,428,039	0.147

Table 2: **Properties of thresholded predicted networks.** We report the thresholds (p_{th}) used to select significant interactions for the four causal inference methods, the corresponding global False Discovery Rate (FDR), as well as descriptors for the resulting networks: N_R is the number of regulating genes, N_T is the number of target genes, N_E is the number of edges, and ρ is the filling ratio of the adjacency matrix, i.e. the ratio of non-zero and zero values in the thresholded matrices.

Despite the similar posterior probability thresholds, the instrumental variable networks are around ten times more densely connected than those obtained by the mediation-based method (Table 2); a difference that cannot be explained by the lower sensitivity of the latter alone. We show that in the instrumental variable networks, high interaction counts occur in blocks that roughly follow the structure of the causal-anchor genotype covariance, whereas they occur more in spikes in the mediation network (Fig. 5). This becomes apparent when plotting the number of targets for each regulatory gene (i.e. each gene with a significant *cis*-eQTL) *versus* its position on the genome. In instrumental variable methods, the genomic causal anchor is used as a "proxy" for the regulatory gene. Hence, if the causal anchor genotypes of two genes within the same locus are correlated due to genomic linkage, then their target sets will unavoidably be similar as well, resulting in the pattern observed in Fig. 5. In mediation-based methods, the expression profile of the regulatory gene is used as the mediator (in test 3, cf. Fig. 1A), and hence a target set will be specific to a regulator, even when its causal anchor is correlated or shared with other genes.

3.5 Causal network inference suggests causal genes for transcriptional hotspots

Regions of the genome that are statistically associated with variation in expression of a high number of genes (the peaks in Fig. 5B) are called transcriptional "hotspots", and finding the causal genes underlying a hotspot is an important problem in quantitative genetics [26]. Albert *et al.* [20] identified 102 hotspot loci using their data, and developed a fine-mapping strategy to narrow the confidence intervals for the hotspot locations. Overlaying the hotspot markers (median bootstrap hotspot locations [20]) with the P_2 target counts in the 5% FDR network (Fig. 5B) showed good consistency, as expected; 37 of those hotspot markers were in our list of causal anchors (i.e. strongest local eQTL for at least one gene). Albert *et al.* [20] defined candidate causal hotspot genes as the genes located within the fine-mapped hotspot regions, and for 26 hotspots they obtained three or fewer candidate genes. Here

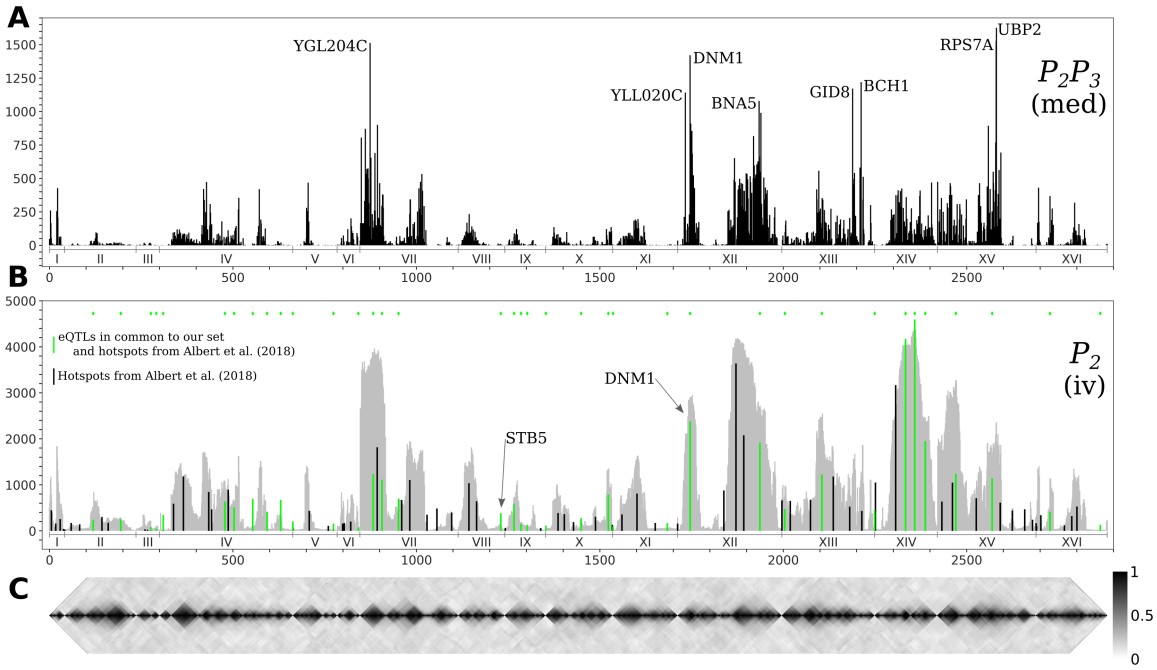

Figure 5: **Hotspots and genotype covariance. A.** The counts of significant interactions for the mediation-based method P_2P_3 at FDR below 10%, with annotations for eight regulatory genes with more than 1,000 targets. **B.** The counts of significant interactions for the instrumental variable method P_2 at FDR below 5% (in grey), and the number of non-zero effects for 102 hotspot markers from [20] (in black); the subset of these hotspots that are also a causal anchor (i.e. the strongest local eQTL for at least one gene) for the *Findr* analysis are marked in green and are also indicated by diamonds at the top of the panel. Interaction count plots for the other instrumental variable methods are in Supp. Fig. S3. **C.** The diagonal of the genotype covariance matrix for the 2884 causal anchor eQTLs. Genes are ordered along the horizontal axis according to the position of their causal anchor in the yeast genome.

we illustrate how causal gene network inference can contribute to the identification of causal hotspot genes using two representative examples *STB5* and *DNM1*.

STB5 is a transcription activator of multidrug resistance genes [27], and the only gene located in one hotspot region on chromosome VIII. The hotspot marker, chrVIII:459310_C/G, is located 11 bp upstream from *STB5*, and is the causal anchor for *STB5* and for no other genes (Fig. 5B and Fig. 6 Left). The instrumental variable method P_2 predicted 131 targets at FDR below 5% for *STB5*, which are strongly enriched for *STB5* targets in the Binding (hypergeometric p-value $2.3 \cdot 10^{-12}$) and Binding & Expression (hypergeometric p-value $1.9 \cdot 10^{-10}$) ground-truth networks. This suggests that when a hotpspot can be confidently mapped to a single gene, instrumental variable methods predict biologically plausible target sets confirming the candidate causal hotspot gene. In contrast, the mediation-based method P_2P_3 predicted only nine *STB5* targets at FDR below 10%, with no enrichment in the ground-truth networks. A possible mechanism that could explain the loss of sensitivity of mediation in this case was already suggested by Albert *et al.* [20]: *STB5* does not show allele-specific expression, but carries protein-altering variants between the two yeast strains that were crossed, suggesting that the causal variants in this hotspot act by directly altering Stb5p protein activity; moreover Stb5p is predicted to target its own promoter in YEASTRACT. Taken together, this leads to a model where transcription of *STB5* is a noisy measurement of Stb5p level, that does not block the path from the

protein-altering variants to *STB5* target genes via Stb5p protein level (Supp. Fig. S4). Hence conditioning on *STB5* transcription in P_2P_3 does not remove the association between these variants and the target genes completely, resulting in false negative predictions by a process similar to the measurement noise model studied in [11].

DNM1 is a gene located near a hotspot region on chromosome XII, and is among the genes with highest target count in the mediation-based network (Fig. 5A, Fig. 6 Right). The hotspot marker is also the causal anchor of *DNM1*, which is located 11,363 bp downstream of this marker and outside the hotspot region mapped by Albert *et al.* Comparison with the target counts in the instrumental variable network, which closely follow the genotype covariance pattern, shows that *DNM1* is the gene in this region that retains the most targets by far in the mediation network (P_2, 2910 targets; P_2P_3, 1421 targets; Fig. 6 Right). This is particularly true when compared with two of the four candidate causal genes of Albert *et al.* that also have a local eQTL within the hotspot region, *YLL007C* (also known as *LMO1*) (P_2, 2846 targets, P_2P_3, 139 targets) and *MMM1* (P_2, 3610 targets; P_2P_3, 8 targets). Based on the high specificity of the P_2P_3 test, we conjecture that *DNM1* is a more likely causal gene for this hotspot than *LMO1* or *MMM1*. Functional analysis in this case does not help to distinguish between these candidates, because Dnm1p and Mmm1p are both essential proteins for the maintenance of mitochondrial morphology [28], and Lmo1p is a signaling protein involved in mitophagy [29]. However, deletion of *DNM1* and *MMM1* results in distinct mitochondrial phenotypes [28], and hence this prediction is experimentally testable in principle.

4 Discussion

4.1 Causal inference from genomics and transcriptomics data infers truly directed gene networks

Reconstructing transcriptional regulatory networks from transcriptomics data has been a major research focus in computational biology during the last 20 years. Existing methods span the entire range of correlation, mutual information, regression, Bayesian networks, random forest, and other machine learning methods, as well as meta-methods combining multiple of these aproaches [19]. Yet, performance on eukaryotic gene expression data has been disappointing, with overlap between predicted and known networks generally not better than random predictions [19]. To some extent, this is due to the lack of reliable ground-truth data. For instance, there is little overlap between the two most common high-throughput experimental techniques for measuring regulatory interactions, mapping TF-DNA binding sites using ChIP-sequencing and measuring differential expression after TF deletion or overexpression [30], see also Table 1. Exceptions to this rule are the transcriptional networks controlling early development in multi-cellular organisms, which are mapped in exquisite detail in some model organisms [31]. When conventional network inference methods are applied to developmental transcriptome data, good performance is in fact observed [32]. Nevertheless the problem of reconstructing *signalling* transcriptional networks from observational expression data remains, and a key missing ingredient in existing approaches is the *directionality* of the edges. Without additional information, any association inferred from transcriptomics data alone is essentially symmetric.

Causal inference is designed to reconstruct truly directed networks, by integrating genomics and transcriptomics data based on general principles of quantitative trait locus analysis [1]. The publication of a dataset of more than 1,000 yeast segregants has allowed us to demonstrate that causal inference indeed results in directed networks with strong, non-random overlap with networks of known transcrip-

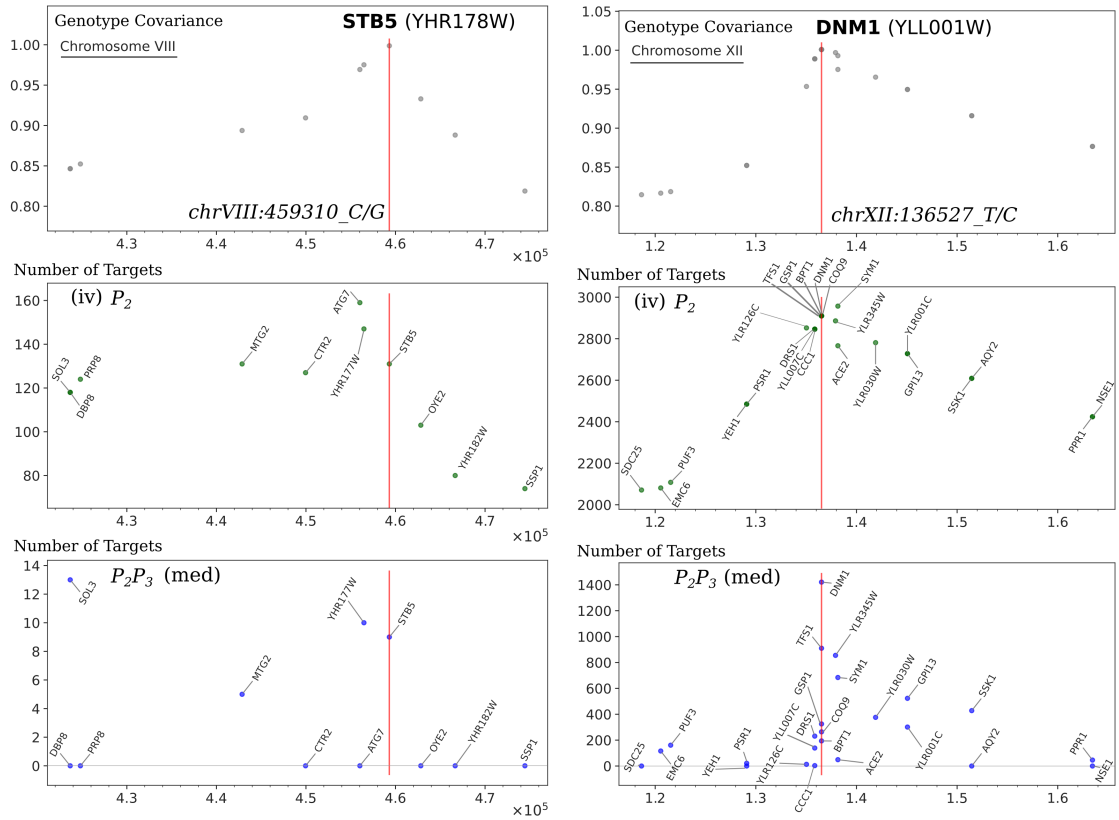

Figure 6: **Details of predicted targets in the vicinity of two genes.** We show the local structure at two genes: *STB5* with eQTL chrVIII:459310_C/G **(left)** and *DNM1* with eQTL chrXII:136527_T/C **(right)**. The **top row** shows genotype covariance in the vicinity of the eQTL (red line) for the gene, in the region where the covariance is greater than 0.8. The **middle row** shows number of targets predicted by P_2 (instrumental variables) at FDR below 5%. The **bottom row** shows number of targets predicted by P_2P_3 (mediation) at FDR below 10%. The horizontal axis gives the position along the chromsome of the eQTL corresponding to each gene. Genes are annotated with their short name where available. Note that data points overlap in genotype covariance and in P_2 for some genes because they share the same eQTL and that P_2P_3 gives no targets on certain genes.

tional interactions. Moreover, the overlap was highest for the most reliable ground-truth that combined two sources of experimental evidence (DNA binding and response to perturbation). Although 1,012 samples for an organism with around 6,000 genes may seem a large number, our analysis also shows that there is no sign yet that performance is saturating as a function of sample size. Causal inference indeed requires larger sample sizes than coexpression-based methods, because it relies on more subtle patterns in the data, something that was already apparent in early considerations of causal inference in this context [5].

Although the integration of genomics and transcriptomics data addresses the key shortcoming pertaining to lack of directionality in network inference when using transcriptomics data alone, important limitations remain. Apart from those already discussed at length in this paper—low sensitivity due to hidden confounders for mediation-based methods, and increased false positive rate due to genomic linkage for instrumental variable methods—another fundamental problem remains: transcriptional regulation is, for the most part, carried out by proteins. Hence, causal interactions inferred from ge-

nomics and transcriptomics data are by definition indirect. If an intermediate, unmeasured protein C (e.g., the protein product of A) lies on the path from gene A to gene B, that is, A also mediates associations between C and local eQTLs for A, then this does not affect the causal inference for the interaction $A \rightarrow B$ (Supp. Fig. S2). However, variation in transcription level of a transcription factor (or other regulatory protein) does not always translate to equal variation in protein level, and *vice versa*. For instance, Albert *et al.* [20] found several protein-altering variants in candidate causal genes mapped to hotspot regions that did not have any local eQTLs. In such cases, our methods would wrongly assign the *trans*-associated target genes to a gene with local eQTL (if one exists), and miss the non-varying (at transcription level) causal gene. This limitation can only be addressed by integrating another layer of information—proteomics data, which is not yet available in comparable sample sizes.

The methods implemented in *Findr* and analyzed in this paper are broadly representative of the current state-of-the-art for causal inference from genomics and transcriptomics data. Nevertheless, some ideas have been proposed recently that we did not evaluate here. For instance, in addition to the statistical tests implemented in *Findr*, Badsha and Fu [33] propose to also use a causal anchor for the target gene B to obtain evidence for a causal interaction $A \rightarrow B$. However, including this test requires limiting the analysis to interactions where both source and target genes have a significant eQTL. Yang *et al.* [34] on the other hand propose to address the hidden confounder problem in mediation by adjusting for selected surrogate variables (e.g. principal components). However, such variables are necessarily composed of combinations of genes and it is challenging to ensure that they only represent common parents and no common children of an $A \rightarrow B$ interaction (which would introduce false positives if conditioned upon, see Supp. Fig. S2). It will be of interest to include these developments in future comparisons.

4.2 Biological data matches theoretical predictions

Causal inference is in essence a hypothesis-driven approach: the causal diagrams in Figure 1 encode prior knowledge and assumptions of how genotypes, genes, and unknown confounding factors influence each other. Based on these diagrams, we can make certain predictions about the patterns we expect to find in the data, such as the relative sensitivity and specificity of mediation *versus* instrumental variable methods, the different situations where each method will be successful or not, etc. It is gratifying to see these predictions confirmed using real data, strengthening significantly our previous findings on simulated data [11].

The hypothesis-driven nature of causal inference lies in between the use of biophysical models of gene regulation and the application of unsupervised machine learning methods for reconstructing gene regulatory networks. Biophysical approaches attempt to include quantitative models of TF-DNA interactions into the network inference process [35, 36], but are hampered by a lack of resolution in omics data (due to both noise within a sample, and limited sampling density). Unsupervised machine learning approaches search for non-random patterns within the data, but without prior specification of the type of pattern being sought, and they lack the ability to identify truly directed associations. Supervised or semi-supervised methods could potentially overcome this limitation, but require data of known interactions for training, which is sparse for non-model organisms [37].

The agreement between theoretical predictions and empirical results indicates that we have a correct understanding of how causal gene network inference algorithms work, and how to interpret results in terms of what type of interactions these algorithms do and do not identify, albeit without any reference

to the underlying biophysical mechanisms.

4.3 Practical recommendations and future work

We conclude this paper by sharing practical recommendations for researchers wanting to apply causal inference methods for the integration of genomics and transcriptomics data.

In general, we recommend instrumental variable over mediation-based methods, as their increased sensitivity tends to outweigh the higher specificity of mediation-based methods. The saturation of performance of mediation-based methods with increasing sample sizes is particularly worrying, although for most current datasets the point where performance saturates is probably not yet reached.

We found limited differences between instrumental variable methods. In the absence of any ground-truth data to evaluate results, we would generally recommend to use the P_2P_5 method, because it will remove at least the most obvious cases of pleiotropy from P_2, while having an easier interpretation than the P method.

The main weakness of instrumental variable methods is their susceptibility to false positive predictions due to genomic linkage. This is a particular concern in data from experimental crosses or inbred organisms, where linkage blocks are large. However, also in human data it has been found that around 10% of non-redundant local eQTLs are associated to the expression of multiple nearby genes [38]. As illustrated, mediation-based causal inference and manual analysis of gene function can sometimes be used to resolve linkage of causal anchors.

In conclusion, causal inference from genomics and transcriptomics data is a more powerful approach for reconstructing causal gene networks than using transcriptomics data alone, which could be further improved by the inclusion of additional layers of omics data and the development of methods to control for or find signal in residual correlations among genes in mediation analyses, and to resolve genomic linkage and pleiotropic effects from transcriptional hotspots in instrumental variable analyses.

References

[1] Jansen RC and Nap JP. Genetical genomics: the added value from segregation. *Trends in Genetics* **17**:388–391 (2001).

[2] Rockman MV. Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature* **456**:738–744 (2008).

[3] Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**:218–223 (2009).

[4] Boyle EA, Li YI and Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**:1177–1186 (2017).

[5] Li Y *et al.* Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends in Genetics* **26**:493–498 (2010).

[6] Albert FW and Kruglyak L. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* **16**:197–212 (2015).

[7] Pai AA, Pritchard JK and Gilad Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet* **11**:e1004857 (2015).

[8] Hemani G, Tilling K and Smith GD. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genetics* **13**:e1007081 (2017).

[9] Schadt EE *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**:710–717 (2005).

[10] Chen LS, Emmert-Streib F and Storey JD. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology* **8**:R219 (2007).

[11] Wang L and Michoel T. Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. *PLoS Computational Biology* **13**:e1005703 (2017).

[12] Davey Smith G and Ebrahim S. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**:1–22 (2003).

[13] Lawlor DA *et al.* Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**:1133–1163 (2008).

[14] Davey Smith G and Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* **23**:R89–R98 (2014).

[15] Millstein J *et al.* Disentangling molecular relationships with a causal inference test. *BMC Genetics* **10**:23 (2009).

[16] Millstein J, Chen GK and Breton CV. cit: hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics* **32**:2364–2365 (2016).

[17] Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**:1353–1358 (2012).

[18] Pinna A *et al.* Simulating systems genetics data with SysGenSIM. *Bioinformatics* **27**:2459–2462 (2011).

[19] Marbach D *et al.* Wisdom of crowds for robust gene network inference. *Nature Methods* **9**:796–804 (2012).

[20] Albert FW *et al.* Genetics of trans-regulatory variation in gene expression. *Elife* **7**:e35471 (2018).

[21] Monteiro PT *et al.* YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic Acids Research* **48**:D642–D649 (2020).

[22] Pedregosa F *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**:2825–2830 (2011).

[23] Seabold S and Perktold J. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference* (2010).

[24] Yates AD *et al.* Ensembl 2020. *Nucleic Acids Research* **48**:D682–D688 (2019).

[25] Marbach D *et al.* Predictive regulatory models in drosophila melanogaster by integrative inference of transcriptional networks. *Genome Research* **22**:1334–1349 (2012).

[26] Rockman MV and Kruglyak L. Genetics of global gene expression. *Nature Reviews Genetics* **7**:862–872 (2006).

[27] Kasten M and Stillman D. Identification of the saccharomyces cerevisiae genes STB1–STB5 encoding Sin3p binding proteins. *Molecular and General Genetics MGG* **256**:376–386 (1997).

[28] Otsuga D *et al.* The dynamin-related GTPase, Dnm1p, controls mitochondrial morphology in yeast. *The Journal of Cell Biology* **143**:333–349 (1998).

[29] Schmitz HP *et al.* Identification of Dck1 and Lmo1 as upstream regulators of the small GTPase Rho5 in Saccharomyces cerevisiae. *Molecular Microbiology* **96**:306–324 (2015).

[30] Cusanovich DA *et al.* The functional consequences of variation in transcription factor binding. *PLoS Genetics* **10**:e1004226 (2014).

[31] MacArthur S *et al.* Developmental roles of 21 drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol* **10**:R80 (2009).

[32] Joshi A, Beck Y and Michoel T. Multi-species network inference improves gene regulatory network reconstruction for early embryonic development in Drosophila. *Journal of Computational Biology* **22**:253–265 (2015).

[33] Badsha M and Fu AQ. Learning causal biological networks with the principle of Mendelian randomization. *Frontiers in Genetics* **10**:460 (2019).

[34] Yang F *et al.* Identifying cis-mediators for trans-eqtls across many human tissues using genomic mediation analysis. *Genome research* **27**:1859–1871 (2017).

[35] Bussemaker HJ, Foat BC and Ward LD. Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu Rev Biophys Biomol Struct* **36**:329–347 (2007).

[36] Äijö T and Bonneau R. Biophysically motivated regulatory network inference: progress and prospects. *Human heredity* **81**:62–77 (2016).

[37] Maetschke SR *et al.* Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in bioinformatics* **15**:195–211 (2014).

[38] Tong P, Monahan J and Prendergast JG. Shared regulatory sites are abundant in the human genome and shed light on genome evolution and disease pleiotropy. *PLoS genetics* **13**:e1006673 (2017).

Supplementary Information

Method	p_{th}	FDR
P_2P_3	0.8175	0.09953
P_2	0.825	0.04974
P_2P_5	0.8375	0.04994
P	0.8575	0.04982
P_0	0.86	0.00986

Table S1: **FDR thresholds.** The thresholds (p_{th}) reported here were used to select significant interactions for the methods shown in figure 5 and S3.

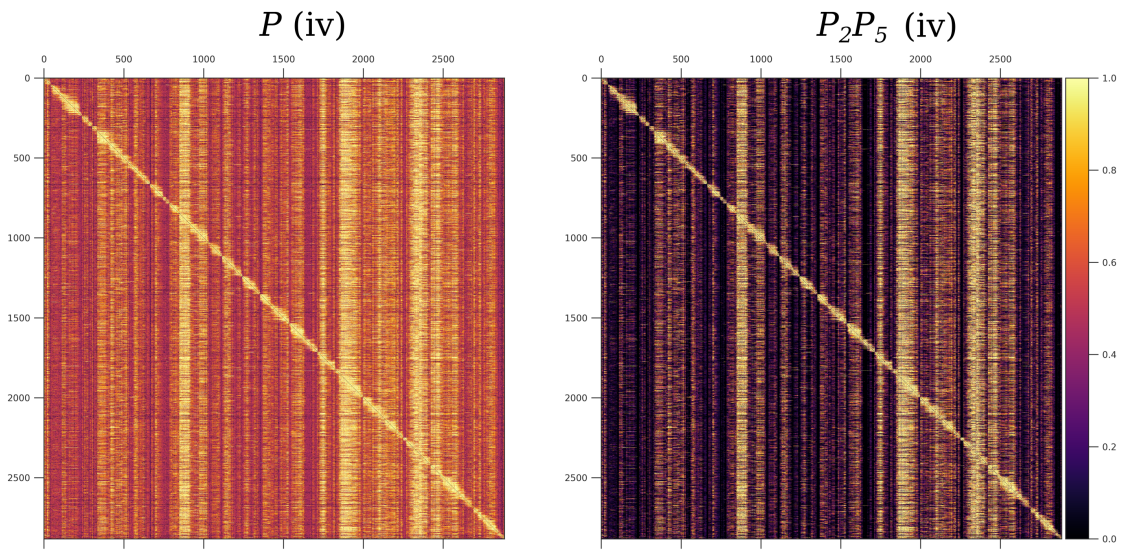

Figure S1: **Matrices of predicted gene interactions.** These square matrices represent the interactions between 2884 genes with causal anchors (eQTLs), posterior probability values are color coded. Vertical bands correspond to hotspots. **Left**: The instrumental variable test with partial pleiotropy P. **Right**: The instrumental variable test with perfect pleiotropy P_2P_5. The genes are ordered according to the position of their causal anchor in the full yeast genome. Definitions of the tests are given in the Methods section. This figure complements Fig. 2 in the main text.

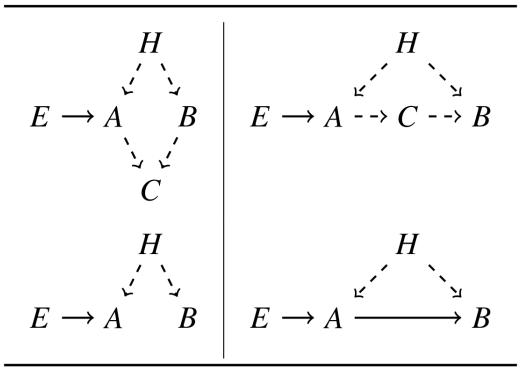

Figure S2: **Comparison of causal models.** These graphical models illustrate possible interactions between genes. The hidden confounder H would be a common parent of A and B as shown also in Fig. 1[B]. The upper row shows a possible ground truth, the lower row shows how this scenario would be classified by *Findr*. **Left**: A common child C of A and B does not affect the predictions of *Findr* regardless of the presence of a common parent H. **Right**: An intermediary node C could mediate the interaction between A and B, regardless of a common parent H.

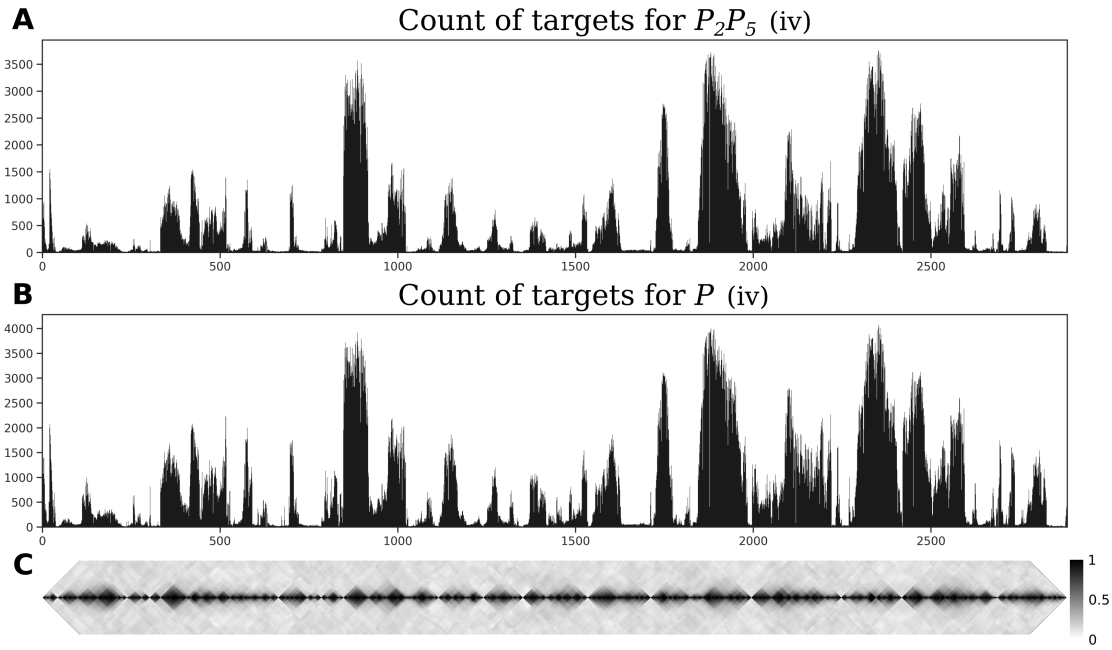

Figure S3: **Hotspots and genotype covariance.** **A** and **B** show the counts of significant interactions for two inference methods. Genes are ordered along the horizontal axis according to the position of their causal anchor in the full yeast genome. **A**: instrumental variables with perfect pleiotropy (P_2P_5) at FDR 5%. **B**: instrumental variables with partial pleiotropy (P) at FDR 5%. The thresholds used are reported in Tab. S1. **C**: The diagonal of the genotype covariance matrix for the 2884 eQTLs.

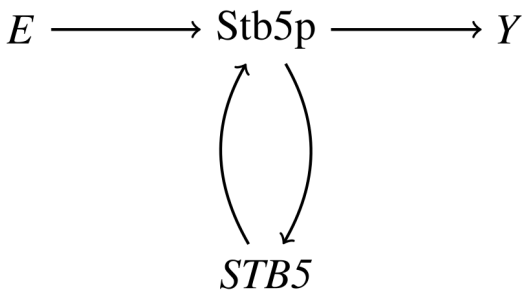

Figure S4: **Hypothetical model for the *STB5* hotspot.** Stb5p protein level is determined by *STB5* transcription level and the genotype of one or more protein-altering variants E, and in turn affects *STB5* transcription level by an auto-regulatory loop. Expression of *STB5* target genes Y is determined by *STB5* transcription only through Stb5p level. Even in the absence of any hidden confounders, *STB5* transcription does not block the path between E and Y, and unless the correlation between *STB5* transcription and Stb5p level is perfect (no biological or experiment noise), conditioning on *STB5* transcription level will not remove the statistical association between E and Y. This model is consistent with the observed lack of allele-specific expression of *STB5* [20], and with the fact that the instrumental variable method P_2 correctly identifies target genes with Stb5p binding sites, but the mediation-based method P_2P_3 does not.