Efficient and accurate causal inference with hidden founders from genome-transcriptome variation data

Lingfei Wang and Tom Michoel

Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK

Abstract

Mapping gene expression as a quantitative trait using whole genome-sequencing and transcriptome analysis allows to discover the functional consequences of genetic variation. We developed a novel method and ultra-fast software Findr for highly accurate causal inference between gene expression traits using cis-regulatory DNA variations as causal anchors, which improves current methods by taking into account hidden confounders and weak regulations. Findr outperformed existing methods on the DREAM5 Systems Genetics challenge and on the prediction of microRNA and transcription factor targets in human lymphoblastoid cells, while being nearly a million times faster. Findr is publicly available at https://github.com/lingfeiwang/findr.

Author summary

Understanding how genetic variation between individuals determines variation in observable traits or disease risk is one of the core aims of genetics. It is known that genetic variation often affects gene regulatory DNA elements and directly causes variation in expression of nearby genes. This effect in turn cascades down to other genes via the complex pathways and gene interaction networks that ultimately govern how cells operate in an ever changing environment. In theory, when genetic variation and gene expression levels are measured simultaneously in a large number of individuals, the causal effects of genes on each other can be inferred using statistical models similar to those used in randomized controlled trials. We developed a novel method and ultra-fast

*Corresponding author. Email: Tom.Michoel@roslin.ed.ac.uk
software Findr which, unlike existing methods, takes into account the complex but unknown network context when predicting causality between specific gene pairs. Findr’s predictions have a significantly higher overlap with known gene networks compared to existing methods, using both simulated and real data. Findr is also nearly a million times faster, and hence the only software in its class that can handle modern datasets where the expression levels of ten-thousands of genes are simultaneously measured in hundreds to thousands of individuals.

1 Introduction

Genetic variation in non-coding genomic regions, including at loci associated with complex traits and diseases identified by genome-wide association studies, predominantly plays a gene-regulatory role. Whole genome and transcriptome analysis of natural populations has therefore become a common practice to understand how genetic variation leads to variation in phenotypes. The number and size of studies mapping genome and transcriptome variation has surged in recent years due to the advent of high-throughput sequencing technologies, and ever more expansive catalogues of expression-associated DNA variants, termed expression quantitative trait loci (eQTLs), are being mapped in humans, model organisms, crops and other species. Unravelling the causal hierarchies between DNA variants and their associated genes and phenotypes is now the key challenge to enable the discovery of novel molecular mechanisms, disease biomarkers or candidate drug targets from this type of data.

It is believed that genetic variation can be used to infer the causal directions of regulation between coexpressed genes, based on the principle that genetic variation causes variation in nearby gene expression and acts as a causal anchor for identifying downstream genes. Although numerous statistical models have been proposed for causal inference with genotype and gene expression data from matching samples, no software implementation in the public domain is efficient enough to handle the volume of contemporary datasets, hindering any attempts to evaluate their performances. Moreover, existing statistical models rely on a conditional independence test which assumes that no hidden confounding factors affect the coexpression of causally related gene pairs. However gene regulatory networks are known to exhibit redundancy and are organized into higher order network motifs, suggesting that confounding of causal relations by known or unknown common upstream regulators is the rule rather than the exception. Moreover, it is also
known that the conditional independence test is susceptible to variations in relative measurement errors between genes\textsuperscript{8,9}, an inherent feature of both microarray and RNA-seq based expression data\textsuperscript{18}.

To investigate and address these issues, we developed Findr (Fast Inference of Networks from Directed Regulations), an ultra-fast software package that incorporates existing and novel statistical causal inference tests. The novel tests were designed to take into account the presence of unknown confounding effects, and were evaluated systematically against multiple existing methods using both simulated and real data.

2 Results

2.1 Findr incorporates existing and novel causal inference tests

Findr performs six likelihood ratio tests involving pairs of genes (or exons or transcripts) $A$, $B$, and an eQTL $E$ of $A$ (Table 1, Section 4.3). Findr then calculates Bayesian posterior probabilities of the hypothesis of interest being true based on the observed likelihood ratio test statistics (denoted $P_i$, $i = 0$ to $5$, $0 \leq P_i \leq 1$, Section 4.5). For this purpose, Findr utilizes newly derived analytical formulae for the null distributions of the likelihood ratios of the implemented tests (Section 4.4, Figure S1). This, together with efficient programming, resulted in a dramatic speedup compared to the standard computationally expensive approach of generating random permutations. The six posterior probabilities are then combined into the traditional causal inference test, our new causal inference test, and separately a correlation test that does not incorporate genotype information (Section 4.6). Each of these tests verifies whether the data arose from a specific subset of $(E, A, B)$ relations (Table 1) among the full hypothesis space of all their possible interactions, and results in a probability of a causal interaction $A \rightarrow B$ being true, which can be used to rank predictions according to significance or to reconstruct directed networks of gene regulations by keeping all interactions exceeding a probability threshold.
2.2 Traditional causal inference fails in the presence of hidden confounders and weak regulations

Findr’s computational speed allowed us to systematically evaluate traditional causal inference methods for the first time. We obtained five datasets with 999 samples simulated from synthetic gene regulatory networks of 1,000 genes with known genetic architecture from the DREAM5 Systems Genetics Challenge (Section 4.1), and subsampled each dataset to observe how performance depends on sample size (Section 4.7). The correlation test \( P_0 \) does not incorporate genotype information and was used as a benchmark for performance evaluations in terms of areas under the receiver operating characteristic (AUROC) and precision-recall (AUPR) curves (Section 4.7). The traditional method \(^1^1\) combines the secondary \( P_2 \) and independence \( P_3 \) tests sequentially (Table 1, Section 4.6), and was evaluated by comparing \( P_2 \) and \( P_2P_3 \) separately against the correlation test. Both the secondary test alone and the traditional causal inference test combination were found to underperform the correlation test (Figure 1A,B). Moreover, the inclusion of the conditional independence test worsened inference accuracy, more so with increasing sample size (Figure 1A,B) and increasing number of regulations per gene (Supplementary Material S2.3). Similar performance drops were also observed for the Causal Inference Test (CIT)\(^1^3\) software, which also is based on the conditional independence test (Figure S3).

We believe that the failure of traditional causal inference is due to an elevated false negative rate (FNR) coming from two sources. First, the secondary test is less powerful in identifying weak interactions than the correlation test. In a true regulation \( E \rightarrow A \rightarrow B \), the secondary linkage \( (E \rightarrow B) \) is the result of two direct linkages chained together, and is harder to detect than either of them. The secondary test hence picks up fewer true regulations, and consequently has a higher FNR. Second, the conditional independence test is counter-productive in the presence of hidden confounders (i.e. common upstream regulators). In such cases, even if \( E \rightarrow A \rightarrow B \) is genuine, the conditional independence test will find \( E \) and \( B \) to be still correlated after conditioning on \( A \) due to a collider effect (Figure S5 \(^1^9\)). Hence the conditional independence test only reports positive on \( E \rightarrow A \rightarrow B \) relations without confounder, further raising the FNR. This is supported by the observation of worsening performance with increasing sample size (where confounding effects become more distinguishable) and increasing number of regulations per gene (which leads
To further support this claim, we examined the inference precision among the top predictions from the traditional test, separately for gene pairs directly unconfounded or confounded by at least one gene (Section 4.7). Compared to unconfounded gene pairs, confounded ones resulted in significantly more false positives among the top predictions (Figure 1C). Furthermore, the vast majority of real interactions fell outside the top 1% of predictions (i.e. had small posterior probability) [92% (651/706) for confounded and 86% (609/709) for unconfounded interactions, Figure 1C]. Together, these results again showed the failure of the traditional test on confounded interactions and its high false negative rate overall.

2.3 Findr accounts for weak secondary linkage, allows for hidden confounders, and outperforms existing methods on simulated data

To overcome the limitations of traditional causal inference methods, Findr incorporates two additional tests (Table 1 and Section 4.3). The relevance test \( P_4 \) verifies that \( B \) is not independent from \( A \) and \( E \) simultaneously and is more sensitive for picking up weak secondary linkages than the secondary linkage test. The controlled test \( P_5 \) ensures that the correlation between \( A \) and \( B \) cannot be fully explained by \( E \), i.e. excludes pleiotropy. The same subsampling analysis revealed that \( P_4 \) performed best in terms of AUROC, and AUPR with small sample sizes, whilst the combination \( P_2 P_5 \) achieved highest AUPR for larger sample sizes (Figure 1A,B). Most importantly, both tests consistently outperformed the correlation test \( P_0 \), particularly for AUPR. This demonstrates conclusively in a comparative setting that the inclusion of genotype data indeed can improve regulatory network inference. These observations are consistent across all five DREAM datasets (Figure S2).

We combined the advantages of \( P_4 \) and \( P_2 P_5 \) by averaging them in a composite test \( P \) (Section 4.6), which outperformed \( P_4 \) and \( P_2 P_5 \) at all sample sizes (Figure 1 and Figure S2) and hence was appointed as Findr’s new test for causal inference. Findr’s new test \( P \) obtained consistently higher levels of local precision (i.e. one minus local FDR) on confounded and unconfounded gene pairs compared to Findr’s traditional causal inference test \( P_T \) (Figure 1C,D), and outperformed the traditional test \( P_T \), correlation test \( P_0 \), CIT, and every participating method of the DREAM5 Systems Genetics Challenge (Section 4.1) in terms of AUROC and AUPR on all 15 datasets.
Specifically, Findr's new test was able to address the inflated FNR of the traditional method due to confounded interactions. It performed almost equally well on confounded and unconfounded gene pairs and, compared to the traditional test, significantly fewer real interactions fell outside the top 1% of predictions (55% vs. 92% for confounded and 45% vs. 86% for unconfounded interactions, Figure 1D, Figure S6).

2.4 The conditional independence test incurs false negatives for unconfounded regulations due to measurement error

The traditional causal inference method based on the conditional independence test results in false negatives for confounded interactions, whose effect was shown significant for the simulated DREAM datasets. However, the traditional test surprisingly reported more confounded gene pairs than the new test in its top predictions (albeit with lower precision), and correspondingly fewer unconfounded gene pairs (Figure 1C, D, Figure S6).

We hypothesized that this inconsistency originated from yet another source of false negatives, where measurement error can confuse the conditional independence test. Measurement error in an upstream variable (called A in Table 1) does not affect the expression levels of its downstream targets, and hence a more realistic model for gene regulation is $E \rightarrow A^{(t)} \rightarrow B$ with $A^{(t)} \rightarrow A$, where the measured quantities are $E$, $A$, and $B$, but the true value for $A$, noted $A^{(t)}$, remains unknown. When the measurement error (in $A^{(t)} \rightarrow A$) is significant, conditioning on $A$ instead of $A^{(t)}$ cannot remove all the correlation between $E$ and $B$ and would therefore report false negatives for unconfounded interactions as well. This effect has been previously studied, for example in epidemiology as the “spurious appearance of odds-ratio heterogeneity”[20].

We verified our hypothesis with a simple simulation (Section 4.8). In a typical scenario with 300 samples from a monoallelic species, minor allele frequency 0.1, and a third of the total variance of $B$ coming from $A^{(t)}$, the conditional independence test reported false negatives (likelihood ratio p-value $\ll 1$, i.e. rejecting the null hypothesis of conditional independence, cf. Table 1) as long as measurement error contributed more than half of $A$’s total unexplained variance (Figure 2B). False negatives occurred at even weaker measurement errors, when the sample sizes were larger or when stronger $A \rightarrow B$ regulations were assumed (Figure S7).
This observation goes beyond the well-known problems that arise from a large measurement error in all variables, which acts like a hidden confounder, or from a much larger measurement error in A than B, which can result in B becoming a better measurement of $A^{(1)}$ than $A$ itself. In this simulation, the false negatives persisted even if $E \rightarrow A$ was observationally much stronger than $E \rightarrow B$, such as when $A$'s measurement error was only 10% ($\sigma_{A1}^2 = 0.1$) compared to up to 67% for $B$ (Figure 2B). This suggested a unique and mostly neglected source of false negatives that would not affect other tests. Indeed, the secondary, relevance, and controlled tests were much less sensitive to measurement errors (Figure 2A,C,D).

2.5 Findr outperforms traditional causal inference and machine learning methods on microRNA target prediction

In order to evaluate Findr on a real dataset, we performed causal inference on miRNA and mRNA sequencing data in lymphoblastoid cell lines from 360 European individuals in the Geuvadis study (Section 4.1). We first tested 55 miRNAs with reported significant cis-eQTLs against 23,722 genes. Since miRNA target predictions from sequence complimentarity alone result in high numbers of false positives, prediction methods based on correlating miRNA and gene expression profiles are of great interest. Although miRNA target prediction using causal inference from genotype and gene expression data has been considered, it remains unknown whether the inclusion of genotype data improves existing expression-based methods. To compare Findr against the state-of-the-art for expression-based miRNA target prediction, we used miRLAB, an integrated database of experimentally confirmed human miRNA target genes with a uniform interface to predict targets using twelve methods, including linear and non-linear, pairwise correlation and multivariate regression methods. We were able to infer miRNA targets with 11/12 miRLAB methods, and also applied the GENIE3 random forest regression method, CIT, and the three tests in Findr: the new ($P$) and traditional ($P_T$) causal inference tests and the correlation test ($P_0$) (Supplementary Material S2.4). Findr’s new test achieved highest AUROC and AUPR among the 16 methods attempted. In particular, Findr’s new test significantly outperformed the traditional test and CIT, the two other genotype-assisted methods, while also being over 500,000 times faster than CIT (Figure 3, Table S2, Figure S8). Findr’s correlation test outperformed all other methods not using genotype information, including correlation, regression, and random forest methods, and was 500 to 100,000 times faster (Figure 3, Table S2, Figure S8). This further illustrates the
power of the Bayesian gene-specific background estimation method implemented in all Findr’s
tests (Section 4.5).

2.6 Findr predicts transcription factor targets with more accurate FDR estimates

We considered 3,172 genes with significant cis-eQTLs in the Geuvadis data (Section 4.1) and inferred regulatory interactions to the 23,722 target genes using Findr’s traditional ($P_T$), new ($P$) and correlation ($P_0$) tests, and CIT. Groundtruths of experimentally confirmed causal gene interactions in human, and mammalian systems more generally, are of limited availability and mainly concern transcription or transcription-associated DNA-binding factors (TFs). Here we focused on a set of 25 TFs in the set of eQTL-genes for which either differential expression data following siRNA silencing (6 TFs) or TF-binding data inferred from ChIP-sequencing and/or DNase foot-printing (20 TFs) in a lymphoblastoid cell line (GM12878) was available (Section 4.1). AUPRs and AUROCs did not exhibit substantial differences, other than modest improvement over random predictions (Figure S9). To test for enrichment of true positives among the top-ranked predictions, which would be missed by global evaluation measures such as AUPR or AUROC, we took advantage of the fact that Findr’s probabilities are empirical local precision estimates for each test (Section 4.5), and assessed how estimated local precisions of new, traditional, and correlation tests reflected the actual precision. Findr’s new test correctly reflected the precision values at various threshold levels, and was able to identify true regulations at high precision control levels (Figure 4). However, the traditional test significantly underestimated precision due to its elevated FNR. This lead to a lack of predictions at high precision thresholds but enrichment of true regulations at low thresholds, essentially nullifying the statistical meaning of its output probability $P_T$. On the other hand, the correlation test significantly overestimated precisions because it is unable to distinguish causal, reversed causal or confounded interactions, which raises its FDR. The same results were observed when alternative groundtruth ChIP-sequencing networks were considered (Figure S9, Figure S10).
3 Discussion

We developed a highly efficient, scalable software package Findr (Fast Inference of Networks from Directed Regulations) implementing novel and existing causal inference tests. Application of Findr on real and simulated genome and transcriptome variation data showed that our novel tests, which account for weak secondary linkage and hidden confounders at the potential cost of an increased number of false positives, resulted in a significantly improved performance to predict known gene regulatory interactions compared to existing methods, particularly traditional methods based on conditional independence tests, which had highly elevated false negative rates.

Causal inference using eQTLs as causal anchors relies on crucial assumptions which have been discussed in-depth elsewhere[8,9]. Firstly, it is assumed that genetic variation is always causal for variation in gene expression, or quantitative traits more generally, and is independent of any observed or hidden confounding factors. Although this assumption is valid for randomly sampled individuals, caution is required when this is not the case (e.g. case-control studies). Secondly, measurement error is assumed to be independent and comparable across variables. Correlated measurement error acts like a confounding variable, whereas a much larger measurement error in the source variable $A$ than the target variable $B$ may lead to an inversion of the inferred causal direction. The conditional independence test in particular relies on the unrealistic assumptions that hidden confounders and measurement errors are absent, the violation of which incurs false negatives and a failure to correctly predict causal relations, as shown throughout this paper.

Although the newly proposed test avoids the elevated FNR from the conditional independence test, it is not without its own limitations. Unlike the conditional independence test, the relevance and controlled tests (Table 1) are symmetric between the two genes considered. Therefore the direction of causality in the new test arises predominantly from using a different eQTL when testing the reverse interaction, potentially leading to a higher FDR as a minor trade-off. About 10% of cis-regulatory eQTLs are linked (as cis-eQTLs) to the expression of more than one gene[25]. In these cases, it appears that the shared cis-eQTL regulates the genes independently[25], which in Findr is accounted for by the ‘controlled’ test (Table 1). When causality between genes and phenotypes or among phenotypes is tested, sharing or linkage of (e)QTLs can be more common. Resolving causality in these cases may require the use of Findr’s conservative, traditional causal inference test, in conjunction with the new test.
In this paper we have addressed the challenge of pairwise causal inference, but to reconstruct the actual pathways and networks that affect a phenotypic trait, two important limitations have to be considered. First, linear pathways propagate causality, and may thus appear as densely connected sets of triangles in pairwise causal networks. Secondly, most genes are regulated by multiple upstream factors, and hence some true edges may only have a small posterior probability unless they are considered in an appropriate multivariate context. The most straightforward way to address these issues would be to model the real directed interaction network as a Bayesian network with sparsity constraints. A major advantage of Findr is that it outputs probability values which can be directly incorporated as prior edge probabilities in existing network inference softwares.

In conclusion, Findr is a highly efficient and accurate open source software tool for causal inference from large-scale genome-transcriptome variation data. Its nonparametric nature ensures robust performances across datasets without parameter tuning, with easily interpretable output in the form of accurate precision and FDR estimates. Findr is able to predict causal interactions in the context of complex regulatory networks where unknown upstream regulators confound traditional conditional independence tests, and more generically in any context with discrete or continuous causal anchors.

4 Methods

4.1 Datasets

We used the following datasets/databases for evaluating causal inference methods:

1. Simulated genotype and transcriptome data of synthetic gene regulatory networks from the DREAM5 Systems Genetics challenge A (DREAM for short), generated by the SysGen-SIM software. DREAM provides 15 sub-datasets, obtained by simulating 100, 300, and 999 samples of 5 different networks each, containing 1000 genes in every sub-dataset but more regulations for sub-datasets with higher numbering. In every sub-dataset, each gene has exactly one matching genotype variable. 25% of the genotype variables are cis-expression Quantitative Trait Loci (eQTL), defined in DREAM as: their variation changes the expression level of the corresponding gene directly. The other 75% are trans-eQTLs,
defined as: their variation affects the expression levels of only the downstream targets of the corresponding gene, but not the gene itself. Because the identities of cis-eQTLs are unknown, we calculated the P-values of genotype-gene expression associations with kruX and kept all genes with a P-value less than $1/750$ to filter out genes without cis-eQTL. For the subsampling analysis (detailed in Section 4.7), we restricted the evaluation to the prediction of target genes from these cis-genes only, in line with the assumption that Findr and other causal inference methods require as input a list of genes whose expression is significantly associated with at least one cis-eQTL. For the full comparison of Findr to the DREAM leaderboard results, we predicted target genes for all genes, regardless of whether they had a cis-eQTL.

2. Genotype and transcriptome sequencing data on 465 human lymphoblastoid cell line samples from the Geuvadis project consisting of the following data products:

- Genotype data (ArrayExpress accession E-GEUV-1)
- Gene quantification data for 23722 genes from nonredundant unique samples and after quality control and normalization (ArrayExpress accession E-GEUV-1)
- Quantification data of miRNA, with the same standard as gene quantification data (ArrayExpress accession E-GEUV-2)
- Best eQTLs of mRNAs and miRNAs (ArrayExpress accessions E-GEUV-1 and E-GEUV-2)

We restricted our analysis to 360 European samples which are shared by gene and miRNA quantifications. Excluding invalid eQTLs from the Geuvadis analysis, such as single-valued genotypes, 55 miRNA-eQTL pairs and 3172 gene-eQTL pairs were retained.

3. For validation of predicted miRNA-gene interactions, we extracted the “strong” ground-truth table from miRLAB, which contains experimentally confirmed miRNA-gene regulations from the following databases: TarBase, miRecords, miRWalk, and miRTarBase. The intersection of the Geuvadis and ground-truth table contains 20 miRNAs and 1054 genes with 1217 confirmed regulations, which are considered for prediction validation. Interactions that are present in the ground-truth table are regarded as true while others as false.
4. For verification of predicted gene-gene interactions, we obtained differential expression data following siRNA silencing of 59 transcription-associated factors (TFs) and DNA-binding data of 201 TFs for 8872 genes in a reference lymphoblastoid cell line (GM12878) from\cite{25}. Six siRNA-targeted TFs, 20 DNA-binding TFs, and 6,790 target genes without missing differential expression data intersected with the set of 3172 eQTL-genes and 23722 target genes in Geuvadis and were considered for validation. We reproduced the pipeline of\cite{25} with the criteria for true targets as having a False Discovery Rate (FDR) < 0.05 from R package \textit{qvalue} for differential expression in siRNA silencing, or having at least 2 TF-binding peaks within 10kb of their transcription start site. We also obtained the filtered proximal TF-target network from\cite{40}, which had 14 TFs and 7,000 target genes in common with the Geuvadis data.

### 4.2 General inference algorithm

Consider a set of observations sampled from a mixture distribution of a null and an alternative hypothesis. For instance in gene regulation, every observation can correspond to expression levels of a pair of genes which are sampled from a bivariate normal distribution with zero (null hypothesis) or non-zero (alternative hypothesis) correlation coefficient. In Findr, we predict the probability that any sample follows the alternative hypothesis with the following algorithm (based on and modified from\cite{11}):

1. For robustness against outliers, we convert every continuous variable into standard normally distributed $N(0, 1)$ values using a rank-based inverse normal transformation across all samples. We name this step as \textit{supernormalization}.

2. We propose a null and an alternative hypothesis for every likelihood ratio test (LRT) of interest where, by definition, the null hypothesis space is a subset of the alternative hypothesis. Model parameters are replaced with their maximum likelihood estimators (MLEs) to obtain the log likelihood ratio (LLR) between the alternative and null hypotheses (\textit{Section 4.3}).

3. We derive the analytical expression for the probability density function (PDF) of the LLR when samples follow the null hypothesis (\textit{Section 4.4}).
4. We convert LLRs into posterior probabilities of the hypothesis of interest with the empirical estimation of local FDR (Section 4.5).

Implementational details can be found in Findr’s source code.

4.3 Likelihood ratio tests

Consider correlated genes $A$, $B$, and a third variable $E$ upstream of $A$ and $B$, such as a significant eQTL of $A$. The eQTLs can be obtained either de novo using eQTL identification tools such as matrix-eQTL\(^\text{41}\) or kruX\(^\text{29}\), or from published analyses. Throughout this article, we assume that $E$ is a significant eQTL of $A$, whereas extension to other data types is straightforward. We use $A_i$ and $B_i$ for the expression levels of gene $A$ and $B$ respectively, which are assumed to have gone through the supernormalization in Section 4.2 and optionally the genotypes of the best eQTL of $A$ as $E_i$, where $i = 1, \ldots, n$ across samples. Genotypes are assumed to have a total of $n_a$ alleles, so $E_i \in \{0, \ldots, n_a\}$. We define the null and alternative hypotheses for a total of six tests, as shown in Table 1. LLRs of every test are calculated separately as follows:

0. **Correlation test:** Define the null hypothesis as $A$ and $B$ are independent, and the alternative hypothesis as they are correlated:

\[ \mathcal{H}^{(0)}_{\text{null}} = A \perp B, \quad \mathcal{H}^{(0)}_{\text{alt}} = A \rightarrow B. \] (1)

The superscript $(0)$ is the numbering of the test. Both hypotheses are modeled with gene expression levels following bivariate normal distributions, as

\[
\begin{pmatrix}
    A_i \\
    B_i
\end{pmatrix}
\sim
\mathcal{N}
\left(
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
\sigma^2_A & \rho \sigma_A \sigma_B \\
\rho \sigma_A \sigma_B & \sigma^2_B
\end{pmatrix}
\right),
\]

for $i = 1, \ldots, n$. The null hypothesis corresponds to $\rho = 0$.

Maximum likelihood estimators (MLE) for the model parameters $\rho$, $\sigma_A$, and $\sigma_B$ are

\[ \hat{\rho} = \frac{1}{n} \sum_{i=1}^{n} A_i B_i, \quad \hat{\sigma}_A = \hat{\sigma}_B = 1, \]

(2)
and the LLR is simply
\[ \text{LLR}^{(0)} = -\frac{n}{2} \ln(1 - \rho^2). \] (3)

In the absence of genotype information, we use nonzero correlation between \( A \) and \( B \) as the indicator for \( A \to B \) regulation, giving the posterior probability
\[ P(A \to B) = P(H_{\text{alt}}^{(0)} | \text{LLR}^{(0)}). \]

false negative

1. **Primary (linkage) test**: Verify that \( E \) regulates \( A \) from \( H_{\text{alt}}^{(1)} \equiv E \to A \) and \( H_{\text{null}}^{(1)} \equiv E \to A \).

For \( H_{\text{alt}}^{(1)} \), we model \( E \to A \) as \( A \) follows a normal distribution whose mean is determined by \( E \) categorically, i.e.
\[ A_i \mid E_i \sim N(\mu_{E_i}, \sigma_A^2). \] (4)

From the total likelihood \( p(A \mid E) = \prod_{i=1}^{n} p(A_i \mid E_i) \), we find MLEs for model parameters \( \mu_j, j = 0, 1, \ldots, n_a \), and \( \sigma_A \), as
\[ \hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n} A_i \delta_{E_i,j}, \quad \hat{\sigma}_A^2 = 1 - \sum_{j=0}^{n_a} \frac{n_j}{n} \hat{\mu}_j^2, \]

where \( n_j \) is the sample count by genotype category,
\[ n_j \equiv \sum_{i=1}^{n} \delta_{E_i,j}. \]

The Kronecker delta function is defined as \( \delta_{xy} = 1 \) for \( x = y \), and 0 otherwise. When summing over all genotype values \( (j = 0, \ldots, n_a) \), we only pick those that exist \( (n_j > 0) \) throughout this article. Since the null hypothesis is simply that \( A_i \) is sampled from a genotype-independent normal distribution, with MLEs of mean zero and standard deviation one due to the supernormalization (Section 4.2), the LLR for test 1 becomes
\[ \text{LLR}^{(1)} = -\frac{n}{2} \ln \hat{\sigma}_A^2. \] (5)
By favoring a large LLR\(^{(1)}\), we select \(H\)\(_{alt}^{(1)}\) and verify that \(E\) regulates \(A\), with

\[ P(E \rightarrow A) = P(H\)\(_{alt}^{(1)} \mid \text{LLR}^{(1)}) . \]

2. **Secondary (linkage) test:** The secondary test is identical with the primary test, except it verifies that \(E\) regulates \(B\). Hence repeat the primary test on \(E\) and \(B\) and obtain the MLEs:

\[ \hat{\nu}_j = \frac{1}{n_j} \sum_{i=1}^n B_i \delta_{E_i j}, \quad \hat{\sigma}_B^2 = 1 - \sum_{j=0}^{n_a} \frac{n_j}{n} \hat{\nu}_j^2 , \]

and the LLR as

\[ \text{LLR}^{(2)} = -\frac{n}{2} \ln \hat{\sigma}_B^2 . \]

\(H\)\(_{alt}^{(2)}\) is chosen to verify that \(E\) regulates \(B\).

3. **(Conditional) independence test:** Verify that \(E\) and \(B\) are independent when conditioning on \(A\). This can be achieved by comparing \(H\)\(_{alt}^{(3)} \equiv B \leftarrow E \rightarrow A \wedge (A \text{ correlates with } B)\) against \(H\)\(_{null}^{(3)} \equiv E \rightarrow A \rightarrow B\). LLRs close to zero then prefer \(H\)\(_{null}^{(3)}\), and ensure that \(E\) regulates \(B\) only through \(A\):

\[ P(E \perp B \mid A) = P(H\)\(_{null}^{(3)} \mid \text{LLR}^{(3)}) . \]

For \(H\)\(_{alt}^{(3)}\), the bivariate normal distribution dependent on \(E\) can be represented as

\[
\begin{pmatrix}
A_i \\ B_i
\end{pmatrix} \sim N
\begin{pmatrix}
\begin{pmatrix}
\mu_{E_i} \\ \nu_{E_i}
\end{pmatrix},
\begin{pmatrix}
\sigma_A^2 & \rho \sigma_A \sigma_B \\
\rho \sigma_A \sigma_B & \sigma_B^2
\end{pmatrix}
\end{pmatrix} .
\]

For \(H\)\(_{null}^{(3)}\), the distributions follow Eq 4 as well as

\[ B_i \mid A_i \sim N(\rho A_i, \sigma_B^2) . \]

Substituting parameters \(\mu_j, \nu_j, \sigma_A, \sigma_B, \rho\) of \(H\)\(_{alt}^{(3)}\) and \(\mu_j, \rho, \sigma_A, \sigma_B\) of \(H\)\(_{null}^{(3)}\) with their MLEs,
we obtain the LLR:

\[
\text{LLR}^{(3)} = -\frac{n}{2} \ln \left( \hat{\sigma}_A^2 \hat{\sigma}_B^2 - (\hat{\rho} + \sigma_{AB} - 1)^2 \right) + \frac{n}{2} \ln \hat{\sigma}_A^2 + \frac{n}{2} \ln(1 - \hat{\rho}^2),
\]

(6)

where

\[
\sigma_{AB} \equiv 1 - \sum_{j=0}^{n_a} \frac{n_j}{n} \hat{\mu}_j \hat{\nu}_j,
\]

and \(\hat{\rho}\) is defined in Eq 2.

4. **Relevance test**: Since the indirect regulation \(E \rightarrow B\) tends to be weaker than any of its direct regulation components (\(E \rightarrow A\) or \(A \rightarrow B\)), we propose to test \(E \rightarrow A \rightarrow B\) with indirect regulation \(E \rightarrow B\) as well as the direct regulation \(A \rightarrow B\) for stronger distinguishing power on weak regulations. We define \(\mathcal{H}^{(4)}_{\text{alt}} \equiv E \rightarrow A \land E \rightarrow B \leftarrow A\) and \(\mathcal{H}^{(4)}_{\text{null}} \equiv E \rightarrow A \rightarrow B\). This simply verifies that \(B\) is not independent from both \(A\) and \(E\) simultaneously. In the alternative hypothesis, \(B\) is regulated by \(E\) and \(A\), which is modeled as a normal distribution whose mean is additively determined by \(E\) categorically and \(A\) linearly, i.e.

\[
B_i \mid E_i, A_i \sim N(\nu_{E_i} + \rho A_i, \sigma_B^2).
\]

We can hence solve its LLR as

\[
\text{LLR}^{(4)} = -\frac{n}{2} \ln \left( \hat{\sigma}_A^2 \hat{\sigma}_B^2 - (\hat{\rho} + \sigma_{AB} - 1)^2 \right) + \frac{n}{2} \ln \hat{\sigma}_A^2.
\]

5. **Controlled test**: Based on the positives of the secondary test, we can further distinguish the alternative hypothesis \(\mathcal{H}^{(5)}_{\text{alt}} \equiv B \leftarrow E \rightarrow A \land A \rightarrow B\) from the null \(\mathcal{H}^{(5)}_{\text{null}} \equiv B \leftarrow E \rightarrow A\) to verify that \(E\) does not regulate \(A\) and \(B\) independently. Its LLR can be solved as

\[
\text{LLR}^{(5)} = -\frac{n}{2} \ln \left( \hat{\sigma}_A^2 \hat{\sigma}_B^2 - (\hat{\rho} + \sigma_{AB} - 1)^2 \right) + \frac{n}{2} \ln \hat{\sigma}_A^2 \hat{\sigma}_B^2.
\]
4.4 Null distributions for the log-likelihood ratios

The null distribution of LLR, \( p(\text{LLR} \mid \mathcal{H}_{\text{null}}) \), may be obtained either by simulation or analytically. Simulation, such as random permutations from real data or the generation of random data from statistics of real data, can deal with a much broader range of scenarios in which analytical expressions are unattainable. However, the drawbacks are obvious: simulation can take hundreds of times longer than analytical methods to reach a satisfiable precision. Here we obtained analytical expressions of \( p(\text{LLR} \mid \mathcal{H}_{\text{null}}) \) for all the tests introduced above.

0. Correlation test: \( \mathcal{H}^{(0)}_{\text{null}} = A \sim B \) indicates no correlation between \( A \) and \( B \). Therefore, we can start from
\[
\tilde{B}_i \sim \text{i.i.d } N(0, 1).
\] (7)

In order to simulate the supernormalization step, we normalize \( \tilde{B}_i \) into \( B_i \) with zero mean and unit variance as:
\[
B_i = \frac{\tilde{B}_i - \bar{\tilde{B}}}{\sigma_{\tilde{B}}}, \quad \bar{\tilde{B}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{B}_i, \quad \sigma_{\tilde{B}}^2 = \frac{1}{n} \sum_{i=1}^{n} (\tilde{B}_i - \bar{\tilde{B}})^2.
\] (8)

Transform the random variables \( \{\tilde{B}_i\} \) by defining
\[
X_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_i \tilde{B}_i,
\] (9)
\[
X_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{B}_i,
\] (10)
\[
X_3 = \left( \sum_{i=1}^{n} \tilde{B}_i^2 \right) - X_1^2 - X_2^2.
\] (11)

Since \( \tilde{B}_i \sim \text{i.i.d } N(0, 1) \) (according to Eq 7), we can easily verify that \( X_1, X_2, X_3 \) are independent, and
\[
X_1 \sim N(0, 1), \quad X_2 \sim N(0, 1), \quad X_3 \sim \chi^2(n-2).
\] (12)

Expressing Eq 3 in terms of \( X_1, X_2, X_3 \) gives
\[
\text{LLR}^{(0)} = -\frac{n}{2} \ln(1 - Y),
\] (13)
in which
\[ Y = \frac{X_2^2}{X_1^2 + X_3} \sim \text{Beta}\left(\frac{1}{2}, \frac{n-2}{2}\right) \quad (14) \]
follows the Beta distribution.

We define distribution \( D(k_1, k_2) \) as the distribution of a random variable
\[ Z = -\frac{1}{2} \ln(1 - Y) \]
for \( Y \sim \text{Beta}(k_1/2, k_2/2) \), i.e.
\[ Z = -\frac{1}{2} \ln(1 - Y) \sim D(k_1, k_2). \]
The probability density function (PDF) for \( Z \sim D(k_1, k_2) \) can be derived as: for \( z > 0 \),
\[ p(z \mid k_1, k_2) = \frac{2}{B(k_1/2, k_2/2)} \left(1 - e^{-2z}\right)^{(k_1/2-1)} e^{-k_2 z}, \quad (15) \]
and for \( z \leq 0 \), \( p(z \mid k_1, k_2) = 0 \). Here \( B(a, b) \) is the Beta function. Therefore the null distribution for the correlation test is simply
\[ \text{LLR}^{(0)}/n \sim D(1, n-2). \quad (16) \]

1. **Primary test**: \( H_{null}^{(1)} = E \quad A \) indicates no regulation from \( E \) to \( A \). Therefore, similarly with the correlation test, we start from \( \tilde{A}_i \sim \text{i.i.d } N(0, 1) \) and normalize them to \( A_i \) with zero mean and unit variance.

The expression of \( \text{LLR}^{(1)} \) then becomes:
\[ \text{LLR}^{(1)} = -\frac{n}{2} \ln \left(1 - \sum_{j=0}^{n_a} \frac{n_j}{n} \left(\tilde{\mu}_j - \bar{\tilde{A}}\right)^2 / \sigma^2_{\tilde{A}}\right), \]
where
\[ \tilde{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n} \tilde{A}_i \delta_{E,i,j}. \]

For now, assume all possible genotypes are present, i.e. \( n_j > 0 \) for \( j = 0, \ldots, n_a \). Trans-
form \{\tilde{A}_i\} by defining
\[
X_j \equiv \sqrt{n_j \tilde{\mu}_j}, \quad \text{for } j = 0, \ldots, n_a,
\]
\[
X_{n_a+1} \equiv \left( \sum_{i=1}^{n} \tilde{A}_i^2 \right) - \left( \sum_{j=0}^{n_a} X_j^2 \right).
\] (17)

Then we can similarly verify that \{X_i\} are pairwise independent, and
\[
X_i \sim \mathcal{N}(0, 1), \text{ for } i = 0, \ldots, n_a,
\]
\[
X_{n_a+1} \sim \chi^2(n - n_a - 1).
\] (18)

Again transform \{X_i\} by defining independent random variables
\[
Y_1 \equiv \sum_{j=0}^{n_a} \sqrt{\frac{n_j}{n}} X_j \sim \mathcal{N}(0, 1),
\]
\[
Y_2 \equiv \left( \sum_{j=0}^{n_a} X_j^2 \right) - Y_1^2 \sim \chi^2(n_a),
\]
\[
Y_3 \equiv X_{n_a+1} \sim \chi^2(n - n_a - 1).
\]

Some calculation would reveal
\[
\text{LLR}^{(1)} = -\frac{n}{2} \ln \left( 1 - \frac{Y_2}{Y_2 + Y_3} \right),
\]
i.e.
\[
\text{LLR}^{(1)}/n \sim \mathcal{D}(n_a, n - n_a - 1).
\]

To account for genotypes that do not show up in the samples, define \(n_v \equiv \sum_{j \in \{j | n_j > 0\}} 1\) as the number of different genotype values across all samples. Then
\[
\text{LLR}^{(1)}/n \sim \mathcal{D}(n_v - 1, n - n_a).
\] (19)

2. Secondary test: Since the null hypotheses and LLRs of primary and secondary tests are identical, \text{LLR}^{(2)} follows the same null distribution as Eq.19.
3. Independence test: The independence test verifies if $E$ and $B$ are uncorrelated when conditioning on $A$, with $\mathcal{H}^{(3)}_{\text{null}} = E \rightarrow A \rightarrow B$. For this purpose, we keep $E$ and $A$ intact while randomizing $\tilde{B}_i$ according to $B$’s correlation with $A$:

$$\tilde{B}_i \equiv \hat{\rho} A_i + \sqrt{1 - \hat{\rho}^2} X_i, \quad X_i \sim \text{i.i.d } N(0, 1).$$

Then $\tilde{B}_i$ is normalized to $B_i$ according to Eq 8. The null distribution of $\text{LLR}^{(3)}$ can be obtained with similar but more complex computations from Eq 6 as

$$\text{LLR}^{(3)} / n \sim \mathcal{D}(n_v - 1, n - n_v - 1). \quad (20)$$

4. Relevance test: The null distribution of $\text{LLR}^{(4)}$ can be obtained similarly by randomizing $B_i$ according to Eq 7 and Eq 8 as

$$\text{LLR}^{(4)} / n \sim \mathcal{D}(n_v, n - n_v - 1).$$

5. Controlled test: To compute the null distribution for the controlled test, we start from

$$\tilde{B}_i = \hat{\nu}_E + \hat{\sigma}_B X_i, \quad X_i \sim N(0, 1), \quad (21)$$

and then normalize $\tilde{B}_i$ into $B_i$ according to Eq 8. Some calculation reveals the null distribution as

$$\text{LLR}^{(5)} / n \sim \mathcal{D}(1, n - n_v - 1).$$

We verified our analytical method of deriving null distributions by comparing the analytical null distribution v.s. null distribution from permutation for the relevance test in Section S2.2.

4.5 Bayesian inference of posterior probabilities

After obtaining the PDFs for the LLRs from real data and the null hypotheses, we can convert LLR values into posterior probabilities $P(\mathcal{H}_{\text{alt}} \mid \text{LLR})$. We use a similar technique as in 11, which itself was based on a more general framework to estimate local FDRs in genome-wide studies 12. This framework assumes that the real distribution of a certain test statistic forms a mixture distribution
of null and alternative hypotheses. After estimating the null distribution, either analytically or by simulation, it can be compared against the real distribution to determine the proportion of null hypotheses, and consequently the posterior probability that the alternative hypothesis is true at any value of the statistic.

To be precise, consider an arbitrary likelihood ratio test. The fundamental assumption is that in the limit $\text{LLR} \to 0^+$, all test cases come from the null hypothesis ($H_{null}$), whilst as $\text{LLR}$ increases, the proportion of alternative hypotheses ($H_{alt}$) also grows. The mixture distribution of real LLR values is assumed to have a PDF as

$$p(\text{LLR}) = P(H_{null})p(\text{LLR} \mid H_{null}) + P(H_{alt})p(\text{LLR} \mid H_{alt}).$$

The priors $P(H_{null})$ and $P(H_{alt})$ sum to unity and correspond to the proportions of null and alternative hypotheses in the mixture distribution. For any test $i = 0, \ldots, 5$, Bayes’ theorem then yields its posterior probability as

$$P(H^{(i)}_{alt} \mid \text{LLR}^{(i)}) = \frac{p(\text{LLR}^{(i)} \mid H^{(i)}_{alt})}{p(\text{LLR}^{(i)})} P(H^{(i)}_{alt}).$$  \hspace{1cm}(22)

Based on this, we can define the posterior probabilities of the selected hypotheses according to Table 1, i.e. the alternative for tests $0, 1, 2, 4, 5$ and the null for test 3 as

$$P_i \equiv \begin{cases} P(H^{(i)}_{alt} \mid \text{LLR}^{(i)}), & i = 0, 1, 2, 4, 5, \\ P(H^{(i)}_{null} \mid \text{LLR}^{(i)}), & i = 3. \end{cases}$$  \hspace{1cm}(23)

After obtaining the LLR distribution of the null hypothesis $[p(\text{LLR} \mid H_{null})]$, we can determine its proportion $[P(H_{null})]$ by aligning $p(\text{LLR} \mid H_{null})$ with the real distribution $p(\text{LLR})$ at the LLR $\to 0^+$ side. This provides all the prerequisites to perform Bayesian inference and obtain any $P_i$ from Eq 23.

In practice, PDFs are approximated with histograms. This requires proper choices of histogram bin widths, $P(H_{null})$, and techniques to ensure the conversion from LLR to posterior probability is monotonically increasing and smooth. Implementational details can be found in Findr package and in Section S1.1. Distributions can be estimated either separately for every $(E, A)$ pair or by pooling across all $(E, A)$ pairs. In practice, we test on the order of $10^3$ to $10^4$ candidate targets.
("B") for every \((E, A)\) such that a separate conversion of LLR values to posterior probabilities is both feasible and recommended, as it accounts for different roles of every gene, especially hub genes, through different rates of alternative hypotheses.

Lastly, in a typical application of Findr, inputs of \((E, A)\) pairs will have been pre-determined as the set of significant eQTL-gene pairs from a genome-wide eQTL association analysis. In such cases, we may naturally assume \(P_1 = 1\) for all considered pairs, and skip the primary test.

4.6 Tests to evaluate

Based on the six tests in Section 4.3, we use the following tests and test combinations for the inference of genetic regulations, and evaluate them in the results.

- The correlation test is introduced as a benchmark, against which we can compare other methods involving genotype information. Pairwise correlation is a simple measure for the probability of two genes being functionally related either through direct or indirect regulation, or through coregulation by a third factor. Bayesian inference additionally considers different gene roles. Its predicted posterior probability for regulation is \(P_0\).

- The traditional causal inference test, as explained in \([11]\), suggested that the regulatory relation \(E \rightarrow A \rightarrow B\) can be confirmed with the combination of three separate tests: \(E\) regulates \(A\), \(E\) regulates \(B\), and \(E\) only regulates \(B\) through \(A\) (i.e. \(E\) and \(B\) become independent when conditioning on \(A\)). They correspond to the primary, secondary, and independence tests respectively. The regulatory relation \(E \rightarrow A \rightarrow B\) is regarded positive only when all three tests return positive. The three tests filter the initial hypothesis space of all possible relations between \(E, A,\) and \(B,\) sequentially to \(E \rightarrow A\) (primary test), \(E \rightarrow A \land E \rightarrow B\) (secondary test), and \(E \rightarrow A \rightarrow B \land (\text{no confounder for } A \text{ and } B)\) (conditional independence test). The resulting test is stronger than \(E \rightarrow A \rightarrow B\) by disallowing confounders for \(A\) and \(B\). So its probability can be broken down as

\[
P_T \equiv P_1P_2P_3. \tag{24}\]

Trigger\([13]\) is an R package implementation of the method. However, since Trigger integrates eQTL discovery with causal inference, it is not practical for use on modern datasets. For
this reason, we reimplemented this method in Findr, and evaluated it with $P_2$ and $P_2P_3$
separately, in order to assess the individual effects of secondary and independence tests.

As discussed above, we expect a set of significant eQTLs and their associated genes as
input, and therefore $P_1 = 1$ is assured and not calculated in this paper or the package Findr.

Note that $P_T$ is the estimated local precision, i.e. the probability that tests 2 and 3 are both
true. Correspondingly, its local FDR (the probability that one of them is false) is $1 - P_T$.

- The novel test, aimed specifically at addressing the failures of the traditional causal infer-
ence test, combines the tests differently:

$$P \equiv \frac{1}{2}(P_2P_5 + P_4).$$

Specifically, the first term in Eq 25 accounts for hidden confounders. The controlled test re-
places the conditional independence test and constrains the hypothesis space more weakly,
only requiring the correlation between $A$ and $B$ is not entirely due to pleiotropy. Therefore,
$P_2P_5$ (with $P_1 = 1$) verifies the hypothesis that $B \leftarrow E \rightarrow A \land (A \not\perp B \mid E)$, a superset of
$E \rightarrow A \rightarrow B$.

On the other hand, the relevance test in the second term of Eq 25 addresses weak in-
teractions that are undetectable by the secondary test from existing data ($P_2$ close to 0).

This term still grants higher-than-null significance to weak interactions, and verifies that
$E \rightarrow A \land (E \rightarrow B \lor A \rightarrow B)$, also a superset of $E \rightarrow A \rightarrow B$. In the extreme undetectable
limit where $P_2 = 0$ but $P_4 \neq 0$, the novel test Eq 25 automatically reduces to $P = \frac{1}{2}P_4$,
which assumes equal probability of either direction and assigns half of the relevance test
probability to $A \rightarrow B$.

The composite design of the novel test aims not to miss any genuine regulation whilst dis-
tinguishing the full spectrum of possible interactions. When the signal level is too weak for
tests 2 and 5, we expect $P_1$ to still provide distinguishing power better than random predic-
tions. When the interaction is strong, $P_2P_5$ is then able to pick up true targets regardless of
the existence of hidden confounders.
4.7 Evaluation methods

- **Evaluation metrics:**

  Given the predicted posterior probabilities for every pair \((A, B)\) from any test, or more generically a score from any inference method, we evaluated the predictions against the direct regulations in the ground-truth tables (Section 4.1) with the metrics of Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, as well as the Areas Under the ROC (AUROC) and Precision-Recall (AUPR) curves. In particular, AUPR is calculated with the Davis-Goadrich nonlinear interpolation with R package `PRROC`.

- **Subsampling:**

  In order to assess the effect of sample size on the performances of inference methods, we performed subsampling evaluations. This is made practically possible by the DREAM datasets which contain 999 samples with sufficient variance, as well as the computational efficiency from Findr which makes subsampling computationally feasible. With a given dataset and ground-truth table, the total number of samples \(n\), and the number of samples of our actual interest \(N < n\), we performed subsampling by repeating following steps \(k\) times:

  1. Randomly select \(N\) samples out of the total \(n\) samples without replacement.
  2. Infer regulations only based on the selected samples.
  3. Compute and record the evaluation metrics of interest (e.g. AUROC and AUPR) with the inference results and ground-truths.

    Evaluation metrics are recorded in every loop, and their means, standard deviations, and standard errors over the \(k\) runs, are calculated. The mean indicates how the inference method performs on the metric in average, while the standard deviation reflects how every individual subsampling deviates from the average performance.

- **Local precision of top predictions separately for confounded and unconfounded gene pairs:**

  In order to demonstrate the inferential precision among top predictions for any inference test (here the traditional and novel tests separately), we first ranked all (ordered) gene pairs
(A, B) according to the inferred significance for A → B. All gene pairs were split into
groups according to their relative significance ranking (9 groups in Figure 1C,D, as top 0%
to 0.01%, 0.01% to 0.02%, etc). Each group was divided into two subgroups, based on
whether each gene pair shared at least one direct upstream regulator gene (confounded)
or not (unconfounded), according to the gold standard. Within each subgroup, the local
precision was computed as the number of true directed regulations divided by the total
number of gene pairs in the subgroup.

4.8 Simulation studies on causal models with measurement error

We investigated how each statistical test tolerates measurement errors with simulations in a con-
trolled setting. We modelled the causal relation A → B in a realistic setup as E → A(t) → B
with A(t) → A. E remains as the accurately measured genotype values as the eQTL for the
primary target gene A. A(t) is the true expression level of gene A, which is not observable. A is
the measured expression level for gene A, containing measurement errors. B is the measured
expression level for gene B.

For simplicity, we only considered monoallelic species. Therefore the genotype E in each sample
followed the Bernoulli distribution, parameterized by the predetermined minor allele frequency.
Each regulatory relation (of E → A(t), A(t) → A, and A(t) → B) corresponded to a normal
distribution whose mean was linearly dependent on the regulator variable. In particular, for sample
i:

\[
A(i) \sim N(\tilde{E}, \sigma_{A1}^2),
\]

\[
A_i \sim N(A(i), \sigma_{A2}^2),
\]

\[
B_i \sim N(A(i), \sigma_B^2),
\]

in which \(\sigma_{A1}^2\), \(\sigma_{A2}^2\), and \(\sigma_B^2\) are parameters of the model. Note that \(\sigma_B^2\) is B’s variance from
all unknown sources, including expression level variations and measurement errors. The tilde
normalizes the variable into zero mean and unit variance, as:

\[
\tilde{X}_i \equiv \frac{X_i - \bar{X}}{\sqrt{\text{Var}(X)}},
\]
where $\bar{X}$ and $\text{Var}(X)$ are the mean and variance of $X \equiv \{X_i\}$ respectively.

Given the five parameters of the model (the number of samples, the minor allele frequency, $\sigma_{A1}$, $\sigma_{A2}$, and $\sigma_B$), we could simulate the observed data for $E$, $A$, and $B$, which were then fed into Findr for tests 2-5 and their p-values of the respective null hypotheses. Supernormalization step was replaced with normalization which merely shifted and scaled variables into zero mean and unit variance.

We then chose different configurations on the number of samples, the minor allele frequency, and $\sigma_B$. For each configuration, we varied $\sigma_{A1}$ and $\sigma_{A2}$ in a wide range to obtain a 2-dimensional heatmap plot for the p-value of each test, thereby exploring how each test was affected by measurement errors of different strengths. Only tiles with a significant $E \rightarrow A$ eQTL relation were retained. The same initial random seed was employed for different configurations to allow for replicability.

**Acknowledgements**

This work was supported by the BBSRC (grant numbers BB/J004235/1 and BB/M020053/1).

**References**


29. Qi J, Foroughi Asl H, Bjørkengren J, Michoel T. kruX: matrix-based non-parametric eQTL dis-


GeneQuantRPKM.50FN.samplename.resk10.txt.gz.


MirnaQuantCount.1.2N.50FN.samplename.resk10.txt.


Table 1: Six likelihood ratio tests are performed to test the regulation $A \rightarrow B$, numbered, named, and defined as shown. $E$ is the best eQTL of $A$. 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 discrete regulators (genotypes) and linear on continuous regulators (gene expression levels). The undirected line represents a multi-variate normal distribution between the relevant variables. In order to identify $A \rightarrow B$ regulation, we select either the null or the alternative hypothesis depending on the test, as shown.

<table>
<thead>
<tr>
<th>Test ID</th>
<th>Test name</th>
<th>Null (hypothesis)</th>
<th>Alternative (hypothesis)</th>
<th>Selected hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Correlation</td>
<td>$A \quad B$</td>
<td>$A \quad B$</td>
<td>Alternative</td>
</tr>
<tr>
<td>1</td>
<td>Primary (Linkage)</td>
<td>$E \quad A$</td>
<td>$E \quad A$</td>
<td>Alternative</td>
</tr>
<tr>
<td>2</td>
<td>Secondary (Linkage)</td>
<td>$E \quad B$</td>
<td>$E \quad B$</td>
<td>Alternative</td>
</tr>
<tr>
<td>3</td>
<td>(Conditional) Independence</td>
<td>$E \quad B$</td>
<td>$E \quad B$</td>
<td>Null</td>
</tr>
<tr>
<td>4</td>
<td>Relevance</td>
<td>$A \quad B$</td>
<td>$A \quad B$</td>
<td>Alternative</td>
</tr>
<tr>
<td>5</td>
<td>Controlled</td>
<td>$A \quad B$</td>
<td>$A \quad B$</td>
<td>Alternative</td>
</tr>
</tbody>
</table>
Figure 1: Findr achieves best prediction accuracy on the DREAM5 Systems Genetics Challenge. (A, B) The mean AUROC (A) and AUPR (B) on subsampled data are shown for traditional ($P_2, P_{2P_3}$) and newly proposed ($P_1, P_{P_{2P_3}}, P$) causal inference tests against the baseline correlation test ($P_0$). Every marker corresponds to the average AUROC or AUPR at specific sample sizes. Random subsampling at every sample size was performed 100 times. Half widths of the lines and shades are the standard errors and standard deviations respectively. $P_i$ corresponds to test $i$ numbered in Table 1. $P$ is the new composite test (Section 4.6). This figure is for dataset 4 of the DREAM challenge. For results on other datasets of the same challenge, see Figure S2. (C, D) Local precision among top predictions according to $P_i$ for the traditional (C) and novel (D) tests for dataset 4 of the DREAM challenge. Numbers next to each bar ($x/y$) indicate the number of true regulations ($x$) and the total number of gene pairs ($y$) within the respective range of prediction scores. For results on other datasets, see Figure S6. (E, F) The average AUROC (E) and AUPR (F) over 5 DREAM datasets with respectively 100, 300 and 999 samples are shown for Findr’s new (Findr-$P$), traditional (Findr-$P_{2P}$), and correlation (Findr-$P_0$) tests, for CIT and for the best scores on the DREAM challenge leaderboard. For individual results on all 15 datasets, see Table S1.
Figure 2: The conditional independence test yields false negatives for unconfounded regulations in the presence of even minor measurement errors. Null hypothesis p-values of the secondary linkage (A), conditional independence (B), relevance (C), and controlled (D) tests are shown on simulated data from the ground truth model $E \rightarrow A(t) \rightarrow B$ with $A(t) \rightarrow A$. $A(t)$’s variance coming from $E$ is set to one, $\sigma^2_{A1}$ is $A(t)$’s variance from other sources and $\sigma^2_{A2}$ is the variance due to measurement noise. A total of 100 values from $10^{-2}$ to $10^2$ were taken for $\sigma^2_{A1}$ and $\sigma^2_{A2}$ to form the $100 \times 100$ tiles. Tiles that did not produce a significant eQTL relation $E \rightarrow A$ with p-value $\leq 10^{-6}$ were discarded. Contour lines are for the log-average of smoothened tile values. Note that for the conditional independence test (B), the true model corresponds to the null hypothesis, i.e. small (purple) p-values correspond to false negatives, whereas for the other tests the true model corresponds to the alternative hypothesis, i.e. small (purple) p-values correspond to true positives (cf. Table 1). For details of the simulation and results from other parameter settings, see Section 4.8 and Figure S7 respectively.
Figure 3: Findr achieves highest accuracy and speed on the prediction of miRNA target genes from the Geuvadis data. Shown are the AUROC (A), AUPR (B) and runtime (C) for 16 miRNA target prediction methods. Methods are colored by type: blue, genotype-assisted causal inference methods; red, pairwise correlation methods; yellow, multivariate regression methods; purple, other methods. Dashed lines are the AUROC and AUPR from random predictions. For method details, see Supplementary Material S2.4.
Figure 4: **Findr predicts TF targets with more accurate FDR estimates from the Geuvadis data.** The precision (i.e. 1-FDR) of TF target predictions is shown at probability cutoffs 0.1 to 0.9 (blue to yellow) with respect to known functional targets from siRNA silencing of 6 TFs (A) and known TF-binding targets of 20 TFs (B). The number above each bar indicates the number of predictions at the corresponding threshold. Dashed lines are precisions from random predictions.