PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

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

Lingfei Wang, Tom Michoel*

Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian, United Kingdom

* Tom.Michoel@roslin.ed.ac.uk

OPEN ACCESS

Citation: Wang L, Michoel T (2017) Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. PLoS Comput Biol 13(8): e1005703. https://doi.org/10.1371/journal.pcbi.1005703

Editor: Jennifer Listgarten, Microsoft Research, UNITED STATES

Received: February 2, 2017

Accepted: July 26, 2017

Published: August 18, 2017

Data Availability Statement: The following data were downloaded from public repositories; data processing normalization steps, if any, are detailed in the Methods section. DREAM5 Systems Genetics challenges. Available from: https://www.synapse.org/#!Synapse:syn2820440/wiki/. Geuvadis genotype data. Available from: ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/genotypes/. Geuvadis gene expression data. Available from: ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/GD462.GeneQuantRPKM.50FN.samplename.resk10.txt.gz.

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 higly accurate causal inference between gene expression traits using cis-regulatory DNA variations as causal anchors, which improves current methods by taking into consideration 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 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.

Geuvadis miRNA expression data. Available from: ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-2/analysis_results/GD452.MirnaQuantCount.1.2N.50FN.samplename.resk10.txt. Geuvadis best eQTL data for mRNA. Available from: ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/EUR373.gene.cis.FDR5.best.rs137.txt.gz. Geuvadis best eQTL data for miRNA. Available from: ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-2/analysis_results/EUR363.mi.cis.FDR5.best.rs137.txt.gz. Groundtruth microRNA target genes. Available from: https://downloads.sourceforge.net/project/mirlab/groundtruth_Strong.csv. siRNA silencing and DNA binding data of transcription factors in lymphoblastoid cell lines. Available from: https://doi.org/10.1371/journal.pgen.1004226.s011. ENCODE filtered proximal transcription factor - target gene network. Available from: http://encodenets.gersteinlab.org/enets2.Proximal_filtered.txt.

Funding: This research was supported by grants from the Biotechnology and Biological Sciences Research Council (BBSRC, http://www.bbsrc.ac.uk, grant numbers: BB/J004235/1, BB/M020053/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Genetic variation in non-coding genomic regions, including at loci associated with complex traits and diseases identified by genome-wide association studies (GWAS), predominantly plays a gene-regulatory role [1]. Whole genome and transcriptome analysis of natural populations has therefore become a common practice to understand how genetic variation leads to variation in phenotypes [2]. 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 [1, 3–5]. 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 [6, 7].

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 [8, 9]. Although numerous statistical models have been proposed for causal inference with genotype and gene expression data from matching samples [10–15], 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 [16] and are organized into higher order network motifs [17], 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 [8, 9, 18], an inherent feature of both microarray and RNA-seq based expression data [19].

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.

Results

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 (Fig 1, Materials and methods). 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, $0 \leq P_i \leq 1$, Materials and methods). For this purpose, Findr utilizes newly derived analytical formulae for the null distributions of the likelihood ratios of the implemented tests (Materials and methods, S1 Fig). 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 (Materials and methods). Each of these tests verifies whether the data arose from a specific subset of (E, A, B) relations (Fig 1) among the full hypothesis space of all their possible interactions, and results in a

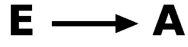
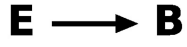
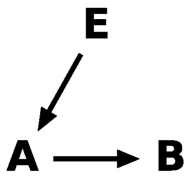
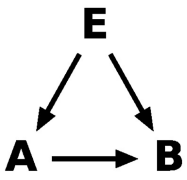
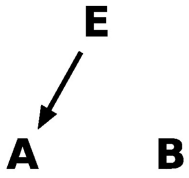
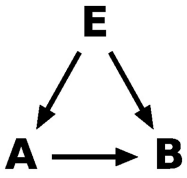
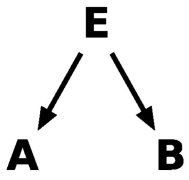
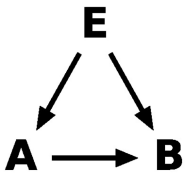
Test ID	Test name	Null (hypothesis)	Alternative (hypothesis)	Selected hypothesis
0	Correlation			Alternative
1	Primary (Linkage)			Alternative
2	Secondary (Linkage)			Alternative
3	(Conditional) Independence			Null
4	Relevance			Alternative
5	Controlled			Alternative

Fig 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.

https://doi.org/10.1371/journal.pcbi.1005703.g001

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.

The traditional causal inference test 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, and subsampled each dataset to observe how performance depends on sample size (Materials and methods). 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 (Materials and methods). The traditional method [11] combines the secondary (P_2) and independence (P_3) tests sequentially (Fig 1, Materials and methods), and was evaluated by comparing P_2 and P_2 P_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 (Fig 2A and 2B). Moreover, the inclusion of the conditional independence test *worsened* inference accuracy, more so with increasing sample size (Fig 2A and 2B) and increasing number of regulations per gene (S1 Text, S2 Fig). Similar performance drops were also observed for the Causal Inference Test (CIT) [13, 15] software, which also is based on the conditional independence test (S3 Fig).

We believe that the failure of the traditional causal inference test 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 (S4 Fig) [20]. Hence the conditional independence test only reports positive on $E \rightarrow A \rightarrow B$ relations without any 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 more confounding).

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 (Materials and methods). Compared to unconfounded gene pairs, confounded ones resulted in significantly more false positives among the top predictions (Fig 2C). 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, Fig 2C]. Together, these results again showed the failure of the traditional test on confounded interactions and its high false negative rate overall.

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 (Fig 1 and Materials and methods). The relevance test (P_4) verifies that B is not

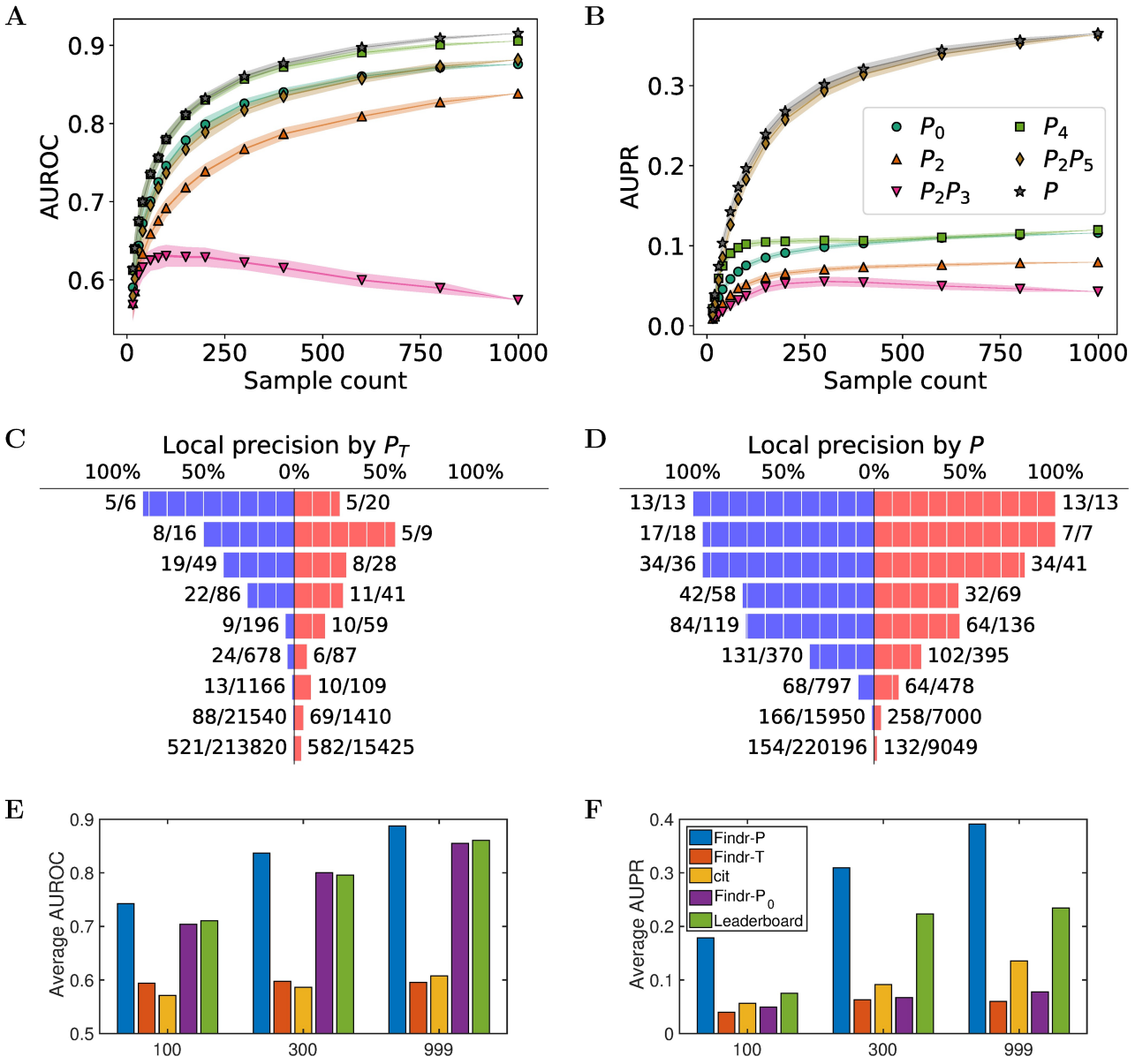


Fig 2. Findr achieves best prediction accuracy on the DREAM5 systems genetics challenge. (**A**, **B**) The mean AUROC (**A**) and AUPR (**B**) on subsampled data are shown for traditional (P_2, P_2 P_3) and newly proposed (P_4, P_2 P_5, P) causal inference tests against the baseline correlation test (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 Fig 1; P is the new composite test (Materials and methods). This figure is for dataset 4 of the DREAM challenge. For results on other datasets of the same challenge, see S2 Fig. (**C**, **D**) Local precision of top predictions (bars top to bottom: 0% to 0.01%, 0.01% to 0.02%, 0.02% to 0.05%, 0.05% to 0.1%, 0.1% to 0.2%, 0.2% to 0.5%, 0.5% to 1%, 1% to 10%, and 10% to 100% top predictions) for the traditional (**C**) and novel (**D**) tests for dataset 4 of the DREAM challenge. Gene pairs unconfounded (left, blue) and confounded by a third gene (right, red) are visualized separately. Each full brick corresponds to 10% in precision. 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 S5E and S5F Fig. (**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_T), and correlation (Findr-P_0) tests, for CIT and for the best scores on the DREAM challenge leaderboad. For individual results on all 15 datasets, see S1 Table.

https://doi.org/10.1371/journal.pcbi.1005703.g002

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 (Fig 2A and 2B). 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 (S2 Fig).

We combined the advantages of P_4 and P_2 P_5 by averaging them in a composite test (P) (Materials and methods), which outperformed P_4 and P_2 P_5 at all sample sizes (Fig 2 and S2 Fig) 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) (Fig 2C and 2D, S5 Fig), and outperformed the traditional causal inference test (P_T), correlation test (P_0), CIT, and every participating method of the DREAM5 Systems Genetics Challenge (Materials and methods) in terms of AUROC and AUPR on all 15 datasets (Fig 2E and 2F, S1 Table, S6 Fig).

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, Fig 2D, S5 Fig).

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

The traditional causal inference method based on the conditional indepedence 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 (Fig 2C and 2D, S5 Fig).

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 Fig 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" [21].

We verified our hypothesis with a simple simulation (Materials and methods). 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 (likeilihood ratio p-value $\ll 1$, i.e. rejecting the null hypothesis of conditional indepencence, cf. Fig 1) as long as measurement error contributed more than half of A's total unexplained variance (Fig 3B). False negatives occurred at even weaker measurement errors, when the sample sizes were larger or when stronger $A \rightarrow B$ regulations were assumed (S7 Fig).

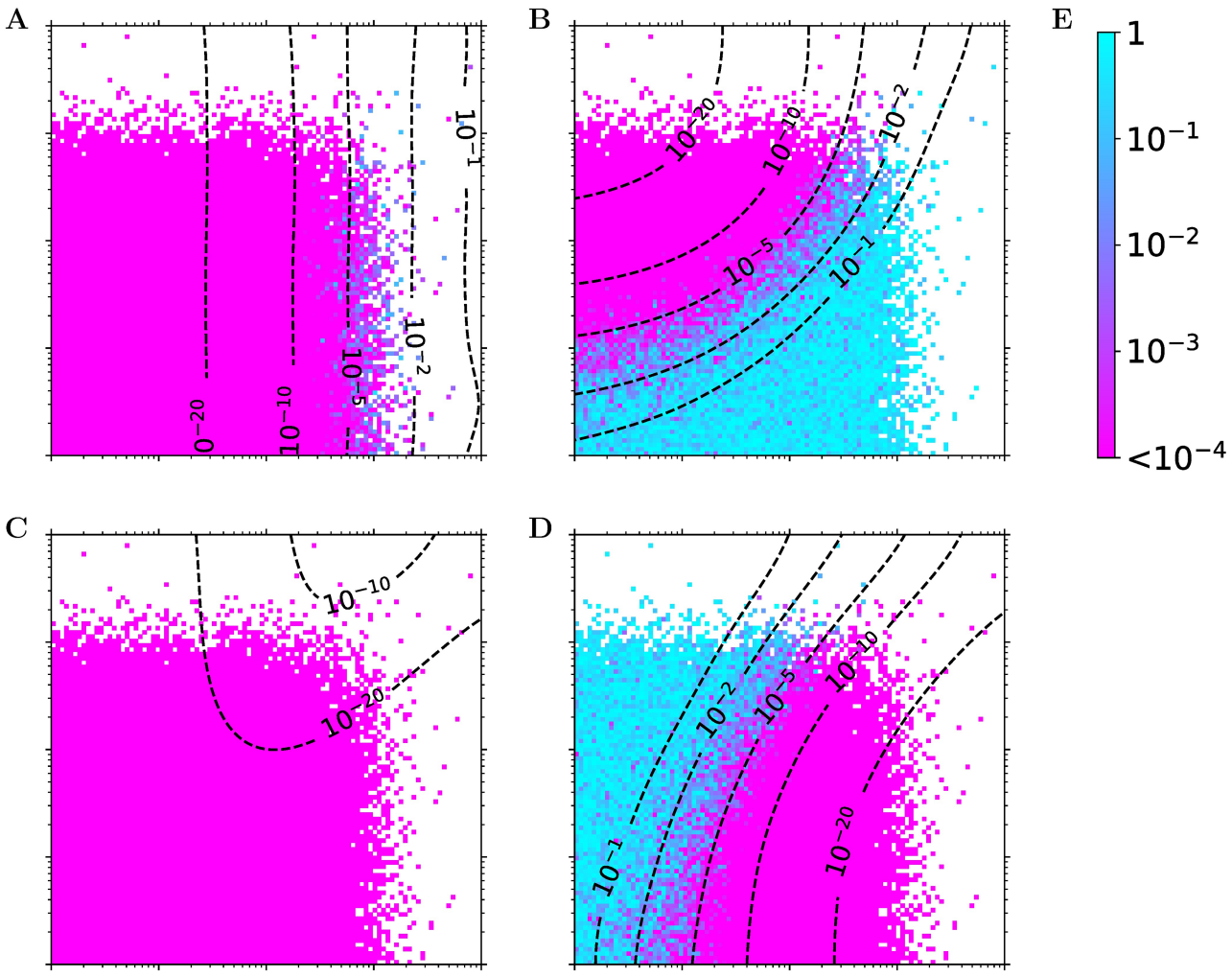


Fig 3. The conditional independence test yields false negatives for unconfounded regulations in the presence of even minor measurement errors. (**A**, **B**, **C**, **D**) Null hypothesis p-values of the secondary linkage (**A**), conditional independence (**B**), relevance (**C**), and controlled (**D**) 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, x axis (σ^2_{A1}) is $A^{(t)}$'s variance from other sources and y axis (σ^2_{A2}) is the variance due to measurement noise. A total of 100 values from 10^{-2} (left, bottom) to 10^2 (right, top) were taken for σ^2_{A1} and σ^2_{A2} each to form the 100 × 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. Fig 1). For details of the simulation and results from other parameter settings, see Materials and methods and S7 Fig respectively. (**E**) Color bar.

https://doi.org/10.1371/journal.pcbi.1005703.g003

This observation goes beyond the well-known problems that arise from a large measurement error in all variables, which acts like a hidden confounder [9], or from a much larger measurement error in A than B, which can result in B becoming a better measurement of $A^{(t)}$ than A itself [8]. 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% ($\sigma^2_{A1} = 0.1$) compared to up to 67% for B (Fig 3B). 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 such measurement errors (Fig 3A, 3C, and 3D).

Findr outperforms the traditional causal inference test 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 [3] (Materials and methods). 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 [22]. Although miRNA target prediction using causal inference from genotype and gene expression data has been considered [23], it remains unknown whether the inclusion of genotype data improves existing expression-based methods. To compare Findr against the state-of-the-art 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 [24]. We were able to infer miRNA targets with 11/12 miRLAB methods, and also applied the GENIE3 random forest regression method [25], CIT, and the three tests in Findr: the new (P) and traditional (P_T) causal inference tests and the correlation test (P_0) (S1 Text). Findr's new test achieved the 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 (Fig 4, S2 Table, S8 Fig). 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 (Fig 4, S2 Table, S8 Fig). This further illustrates the power of the Bayesian gene-specific background estimation method implemented in all Findr's tests (Materials and methods).

Findr predicts transcription factor targets with more accurate FDR estimates

We considered 3,172 genes with significant cis-eQTLs in the Geuvadis data [3] (Materials and methods) 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 footprinting (20 TFs) in a lymphoblastoid cell line (GM12878) was available [26] (Materials and methods). AUPRs and AUROCs did not exhibit substantial differences, other than modest improvement over random predictions (S9 Fig, S3 Table). 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 (Materials and methods), 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 (Fig 5). 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

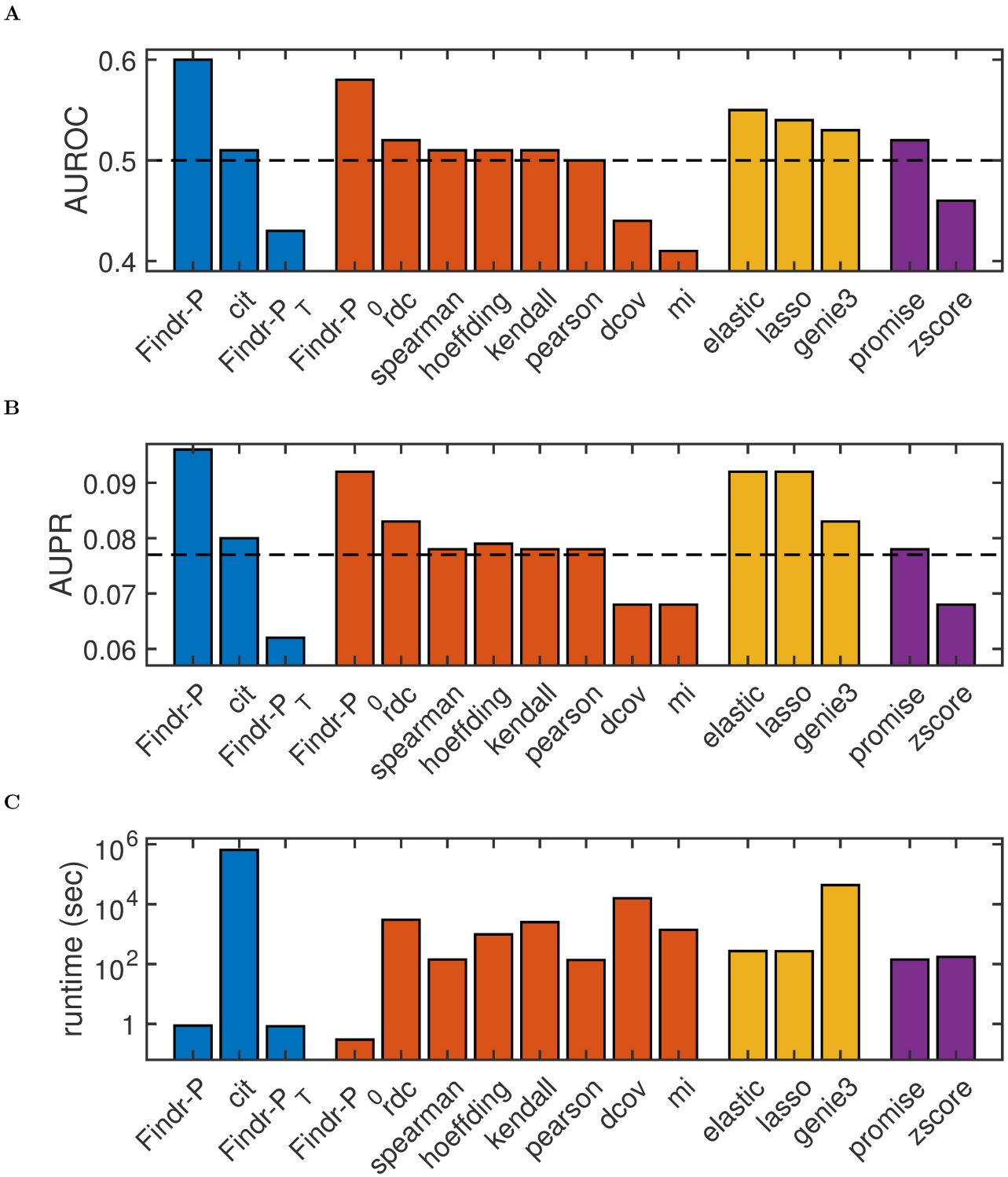


Fig 4. Findr achieves highest accuracy and speed on the prediction of miRNA target genes from the Geuvadis data. Shown are the AUROC (**A**), AUPR (**B**) and runtime (**C**) for 16 miRNA target prediction methods. Methods are colored by type: blue, genotype-assisted causal inference methods; red, pairwise correlation methods; yellow, multivariate regression methods; purple, other methods. Dashed lines are the AUROC and AUPR from random predictions. For method details, see S1 Text.

https://doi.org/10.1371/journal.pcbi.1005703.g004

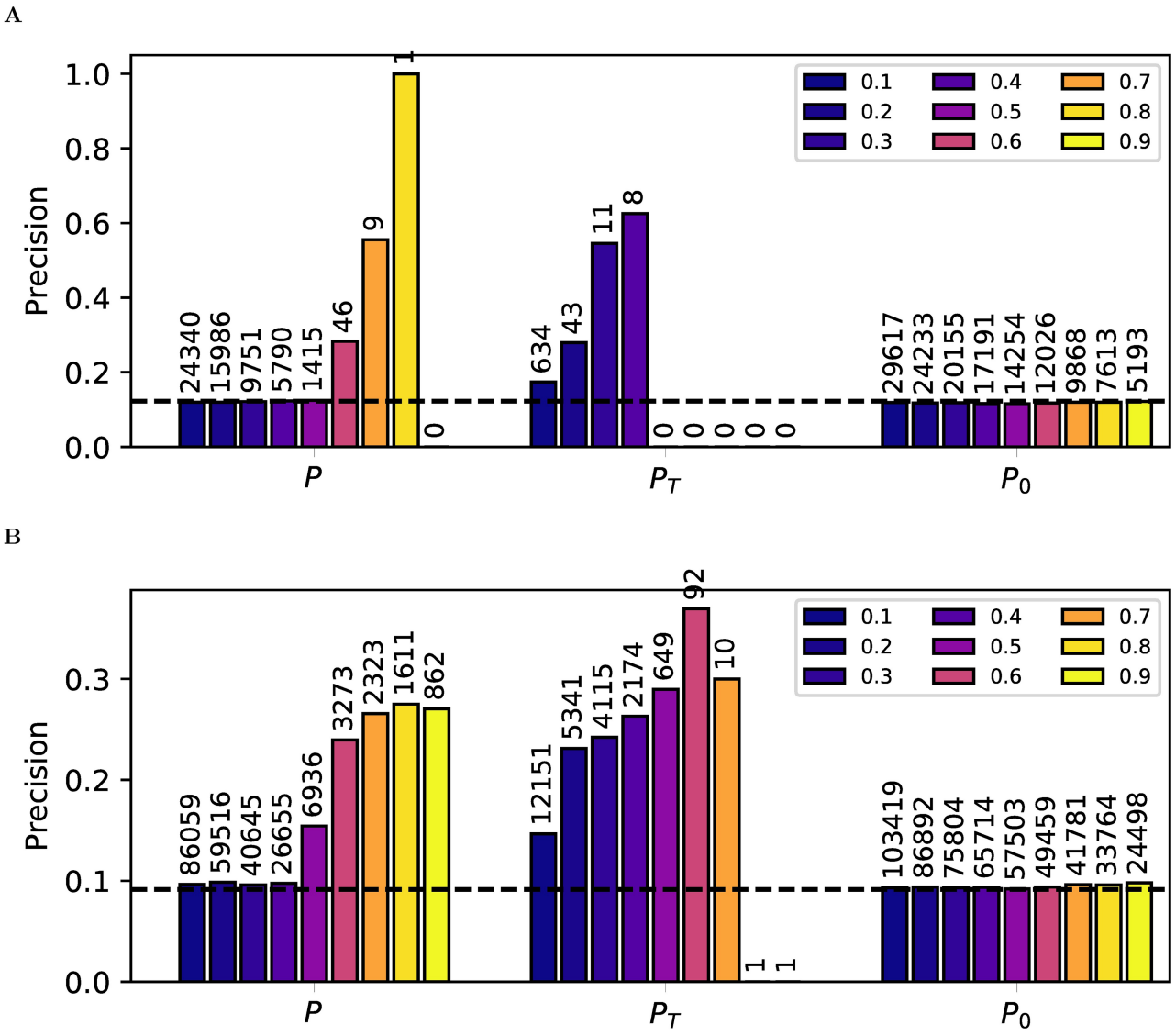


Fig 5. Findr predicts TF targets with more accurate FDR estimates from the Geuvadis data. The precision (i.e. 1-FDR) of TF target predictions is shown at probability cutoffs 0.1 to 0.9 (blue to yellow) with respect to known functional targets from siRNA silencing of 6 TFs (**A**) and known TF-binding 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.

https://doi.org/10.1371/journal.pcbi.1005703.g005

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 (S9 and S10 Figs).

Materials and methods

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 SysGenSIM

software [27]. 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 [28], and kept all genes with a P-value less than 1/750 to filter out genes without cis-eQTL. For the subsampling analysis, 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 [3] 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 [24], which contains experimentally confirmed miRNA-gene regulations from the following databases: TarBase [29], miRecords [30], miRWalk [31], and miRTarBase [32]. 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 [26]. 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 [26] with the criteria for true targets as having a False Discovery Rate (FDR) < 0.05 from R package *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 [33], which had 14 TFs and 7,000 target genes in common with the Geuvadis data.

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 wich 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 [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 *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.
3. We derive the analytical expression for the probablity density function (PDF) of the LLR when samples follow the null hypothesis.
4. We convert LLRs into posterior probabilities of the hypothesis of interest with the empirical estimation of local FDR.

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

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 [34] or kruX [28], 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 supernormalization, 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 Fig 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:

$$\mathcal{H}^{(0)}_{\text{null}} = A \quad B, \qquad \mathcal{H}^{(0)}_{\text{alt}} = A \text{ — } B. \tag{1}$$

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

$$\begin{pmatrix} A_i \\ B_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{A0} & \rho\,\sigma_{A0}\sigma_{B0} \\ \rho\,\sigma_{A0}\sigma_{B0} & \sigma^2_{B0} \end{pmatrix}\right),$$

   for $i = 1, \ldots, n$. The null hypothesis corresponds to $\rho = 0$.
   Maximum likelihood estimators (MLE) for the model parameters ρ, σ_{A0}, and σ_{B0} are

$$\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} A_i B_i, \qquad \hat{\sigma}_{A0} = \hat{\sigma}_{B0} = 1, \tag{2}$$

and the LLR is simply

$$\mathrm{LLR}^{(0)} = -\frac{n}{2}\ln(1-\hat{\rho}^2). \tag{3}$$

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

$$P(A \text{ — } B) = P(\mathcal{H}_{\mathrm{alt}}^{(0)} \mid \mathrm{LLR}^{(0)}).$$

1. **Primary (linkage) test**: Verify that E regulates A from $\mathcal{H}_{\mathrm{alt}}^{(1)} \equiv E \to A$ and $\mathcal{H}_{\mathrm{null}}^{(1)} \equiv E \quad A$. For $\mathcal{H}_{\mathrm{alt}}^{(1)}$, we model $E \to A$ as A follows a normal distribution whose mean is determined by E categorically, i.e.

$$A_i \mid E_i \sim N(\mu_{E_i}, \sigma_A^2). \tag{4}$$

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 σ_A, as

$$\hat{\mu}_j = \frac{1}{n_j}\sum_{i=1}^{n} A_i \delta_{E_i j}, \qquad \hat{\sigma}_A^2 = 1 - \sum_{j=0}^{n_a}\frac{n_j}{n}\hat{\mu}_j^2,$$

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 supernormalization, the LLR for test 1 becomes

$$\mathrm{LLR}^{(1)} = -\frac{n}{2}\ln\hat{\sigma}_A^2. \tag{5}$$

By favoring a large $\mathrm{LLR}^{(1)}$, we select $\mathcal{H}_{\mathrm{alt}}^{(1)}$ and verify that E regulates A, with

$$P(E \to A) = P(\mathcal{H}_{\mathrm{alt}}^{(1)} \mid \mathrm{LLR}^{(1)}).$$

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:

$$\hat{v}_j = \frac{1}{n_j}\sum_{i=1}^{n} B_i \delta_{E_i j}, \qquad \hat{\sigma}_B^2 = 1 - \sum_{j=0}^{n_a}\frac{n_j}{n}\hat{v}_j^2,$$

and the LLR as

$$\mathrm{LLR}^{(2)} = -\frac{n}{2}\ln\hat{\sigma}_B^2.$$

$\mathcal{H}_{\mathrm{alt}}^{(2)}$ 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 $\mathcal{H}_{\text{alt}}^{(3)} \equiv B \leftarrow E \rightarrow A \wedge (A \text{ correlates with } B)$ against $\mathcal{H}_{\text{null}}^{(3)} \equiv E \rightarrow A \rightarrow B$. LLRs close to zero then prefer $\mathcal{H}_{\text{null}}^{(3)}$, and ensure that E regulates B only through A:

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

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

$$\begin{pmatrix} A_i \\ B_i \end{pmatrix} \Bigg| \; E_i \sim N\left(\begin{pmatrix} \mu_{E_i} \\ v_{E_i} \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} \right).$$

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

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

Substituting parameters μ_j, v_j, σ_A, σ_B, ρ of $\mathcal{H}_{\text{alt}}^{(3)}$ and μ_j, ρ, σ_A, σ_B of $\mathcal{H}_{\text{null}}^{(3)}$ with their MLEs, we obtain the LLR:

$$\begin{aligned} \text{LLR}^{(3)} \quad = \quad & -\frac{n}{2}\ln\left(\hat{\sigma}_A^2\hat{\sigma}_B^2 - (\hat{\rho} + \sigma_{AB} - 1)^2\right) \\ & + \frac{n}{2}\ln\hat{\sigma}_A^2 + \frac{n}{2}\ln(1 - \hat{\rho}^2), \end{aligned} \tag{6}$$

where

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

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 $\mathcal{H}_{\text{alt}}^{(4)} \equiv E \rightarrow A \wedge E \rightarrow B \leftarrow A$ and $\mathcal{H}_{\text{null}}^{(4)} \equiv E \rightarrow A \quad B$. 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.

$$B_i \mid E_i, A_i \sim N(v_{E_i} + \rho A_i, \sigma_B^2).$$

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 $\mathcal{H}_{\text{alt}}^{(5)} \equiv B \leftarrow E \rightarrow A \wedge A \rightarrow B$ from the null $\mathcal{H}_{\text{null}}^{(5)} \equiv B \leftarrow E \rightarrow A$

to verify that E does not regulate A and B independently. Its LLR can be solved as

$$\mathrm{LLR}^{(5)} = -\frac{n}{2}\ln\left(\hat{\sigma}_A^2\hat{\sigma}_B^2 - (\hat{\rho}+\sigma_{AB}-1)^2\right) + \frac{n}{2}\ln\hat{\sigma}_A^2\hat{\sigma}_B^2.$$

Null distributions for the log-likelihood ratios

The null distribution of LLR, $p(\mathrm{LLR} \mid \mathcal{H}_{\mathrm{null}})$, 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 $p(\mathrm{LLR} \mid \mathcal{H}_{\mathrm{null}})$ for all the tests introduced above.

0. **Correlation test**: $\mathcal{H}_{\mathrm{null}}^{(0)} = A \quad B$ indicates no correlation between A and B. Therefore, we can start from

$$\tilde{B}_i \sim \text{i.i.d } N(0,1). \tag{7}$$

In order to simulate the supernormalization step, we normalize $\tilde{B}_i$ into B_i with zero mean and unit variance as:

$$B_i \equiv \frac{\tilde{B}_i - \bar{\tilde{B}}_i}{\sigma\tilde{B}}, \qquad \bar{\tilde{B}} \equiv \frac{1}{n}\sum_{i=1}^{n}\tilde{B}_i, \qquad \sigma_{\tilde{B}}^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(\tilde{B}_i - \bar{\tilde{B}})^2. \tag{8}$$

Transform the random variables $\{\tilde{B}_i\}$ by defining

$$X_1 \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\tilde{B}_i, \tag{9}$$

$$X_2 \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{B}_i, \tag{10}$$

$$X_3 \equiv \left(\sum_{i=1}^{n}\tilde{B}_i^2\right) - X_1^2 - X_2^2. \tag{11}$$

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

$$X_1 \sim N(0,1), \qquad X_2 \sim N(0,1), \qquad X_3 \sim \chi^2(n-2). \tag{12}$$

Expressing Eq 3 in terms of X_1, X_2, X_3 gives

$$\mathrm{LLR}^{(0)} = -\frac{n}{2}\ln(1-Y), \tag{13}$$

in which

$$Y \equiv \frac{X_1^2}{X_1^2 + X_3} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{n-2}{2}\right) \tag{14}$$

follows the Beta distribution.

We define distribution $\mathcal{D}(k_1, k_2)$ 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.

$$Z = -\frac{1}{2}\ln(1-Y) \sim \mathcal{D}(k_1, k_2).$$

The probability density function (PDF) for $Z \sim \mathcal{D}(k_1, k_2)$ can be derived as: for $z > 0$,

$$p(z \mid k_1, k_2) = \frac{2}{B(k_1/2, k_2/2)}(1-e^{-2z})^{(k_1/2-1)}e^{-k_2 z}, \tag{15}$$

and for $z \leq 0$, $p(z|k_1, k_2) = 0$. Here $B(a, b)$ is the Beta function. Therefore the null distribution for the correlation test is simply

$$\text{LLR}^{(0)}/n \sim \mathcal{D}(1, n-2). \tag{16}$$

1. **Primary test**: $\mathcal{H}^{(1)}_{\text{null}} = E \quad A$ 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:

$$\text{LLR}^{(1)} = -\frac{n}{2}\ln\left(1 - \sum_{j=0}^{n_a}\frac{n_j}{n}\frac{(\hat{\tilde{\mu}}_j - \bar{\tilde{A}})^2}{\sigma^2_{\tilde{A}}}\right),$$

where

$$\hat{\tilde{\mu}}_j \equiv \frac{1}{n_j}\sum_{i=1}^{n}\tilde{A}_i\delta_{E_i j}.$$

For now, assume all possible genotypes are present, i.e. $n_j > 0$ for $j = 0, \ldots, n_a$. Transform $\{\tilde{A}_i\}$ by defining

$$\begin{aligned} X_j &\equiv \sqrt{n_j}\,\hat{\tilde{\mu}}_j, && \text{for } j = 0, \ldots, n_a, \\ X_{n_a+1} &\equiv \left(\sum_{i=1}^{n}\tilde{A}_i^2\right) - \left(\sum_{j=0}^{n_a}X_j^2\right). \end{aligned} \tag{17}$$

Then we can similarly verify that $\{X_i\}$ are pairwise independent, and

$$\begin{aligned} X_i &\sim N(0,1), \text{ for } i = 0, \ldots, n_a, \\ X_{n_a+1} &\sim \chi^2(n - n_a - 1). \end{aligned} \tag{18}$$

Again transform $\{X_i\}$ by defining independent random variables

$$Y_1 \equiv \sum_{j=0}^{n_a}\sqrt{\frac{n_j}{n}}X_j \sim N(0,1),$$

$$Y_2 \equiv \left(\sum_{j=0}^{n_a}X_j^2\right) - Y_1^2 \sim \chi^2(n_a),$$

$$Y_3 \equiv X_{n_a+1} \sim \chi^2(n - n_a - 1).$$

Some calculation would reveal

$$\mathrm{LLR}^{(1)} = -\frac{n}{2}\ln\left(1-\frac{Y_2}{Y_2+Y_3}\right),$$

i.e.

$$\mathrm{LLR}^{(1)}/n \sim \mathcal{D}(n_a, n-n_a-1).$$

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

$$\mathrm{LLR}^{(1)}/n \sim \mathcal{D}(n_v-1, n-n_v). \tag{19}$$

2. **Secondary test**: Since the null hypotheses and LLRs of primary and secondary tests are identical, $\mathrm{LLR}^{(2)}$ 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 $\mathcal{H}^{(3)}_{\text{null}} = E \to A \to B$. For this purpose, we keep E and A intact while randomizing $\tilde{B}_i$ according to B's correlation with A:

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

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

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

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

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

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

$$\tilde{B}_i = \hat{v}_{E_i} + \hat{\sigma}_B X_i, \qquad X_i \sim N(0,1), \tag{21}$$

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

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

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.

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 [35]. 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 LLR $\to 0^+$, all test cases come from the null hypothesis ($\mathcal{H}_{\text{null}}$), whilst as LLR increases, the proportion of alternative hypotheses ($\mathcal{H}_{\text{alt}}$) also grows. The mixture distribution of real LLR values is assumed to have a PDF as

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

The priors $P(\mathcal{H}_{\text{null}})$ and $P(\mathcal{H}_{\text{alt}})$ 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

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

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

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

After obtaining the LLR distribution of the null hypothesis [$p(\text{LLR} \mid \mathcal{H}_{\text{null}})$], we can determine its proportion [$P(\mathcal{H}_{\text{null}})$] by aligning $p(\text{LLR} \mid \mathcal{H}_{\text{null}})$ 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, $P(\mathcal{H}_{\text{null}})$, 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 S1 Text. 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 associaton analysis. In such cases, we may naturally assume $P_1 = 1$ for all considered pairs, and skip the primary test.

Tests to evaluate

Based on the six tests in Fig 1, we use the following tests and test combinations for the inference of genetic regulations, and evalute 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), $E \rightarrow A \wedge E \rightarrow B$ (secondary test), and $E \rightarrow A \rightarrow B \wedge$ (no confounder for A 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

$$P_T \equiv P_1 P_2 P_3. \tag{24}$$

  Trigger [36] 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 $P_2\,P_3$ 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 inference test, combines the tests differently:

$$P \equiv \frac{1}{2}(P_2 P_5 + P_4). \tag{25}$$

  Specifically, the first term in Eq 25 accounts for hidden confounders. The controlled test replaces 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, $P_2\,P_5$ (with $P_1 = 1$) verifies the hypothesis that $B \leftarrow E \rightarrow A \wedge (A \perp B|E)$, a superset of $E \rightarrow A \rightarrow B$.
  On the other hand, the relevance test in the second term of Eq 25 addresses weak interactions 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 $E \rightarrow A \wedge (E \rightarrow B \vee A \text{ — } B)$, 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 distinguishing the full spectrum of possible interactions. When the signal level is too weak for tests 2 and 5, we expect P_4 to still provide distinguishing power better than random predictions. When the interaction is strong, $P_2\,P_5$ is then able to pick up true targets regardless of the existence of hidden confounders.

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 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 [37]. In particular, AUPR is calculated with the Davis-Goadrich nonlinear interpolation [38] 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 $A \to B$. All gene pairs were split into groups according to their relative significance ranking (9 groups in Fig 2C and 2D, 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.

Simulation studies on causal models with measurement error

We investigated how each statistical test tolerates measurement errors with simulations in a controlled setting. We modelled the causal relation $A \to B$ in a realistic setup as $E \to A^{(t)} \to B$ with $A^{(t)} \to A$. 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 $E \to A^{(t)}$, $A^{(t)} \to A$, and $A^{(t)} \to B$) correponded to a normal distribution whose mean was linearly dependent on the regulator variable. In particular, for sample i:

$$A_i^{(t)} \sim N(\widetilde{E_i}, \sigma_{A1}^2), \tag{26}$$

$$A_i \sim N(A_i^{(t)}, \sigma_{A2}^2), \tag{27}$$

$$B_i \sim N(\widetilde{A_i^{(t)}}, \sigma_B^2), \tag{28}$$

in which σ_{A1}, σ_{A2}, and σ_B are parameters of the model. Note that σ_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:

$$\widetilde{X}_i \equiv \frac{X_i - \bar{X}}{\sqrt{\mathrm{Var}(X)}}, \tag{29}$$

where $\bar{X}$ and $Var(X)$ 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, σ_{A1}, σ_{A2}, and σ_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 σ_B. For each configuration, we varied σ_{A1} and σ_{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.

Conclusion

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 (Fig 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 [39]. In these cases, it appears that the shared *cis*-eQTL regulates the genes independently [39], which in Findr is accounted for by the 'controlled' test (Fig 1). When

causality between genes and phenotypes or among phenotypes is tested, sharing or linkage of (e)QTLs will be more common. Resolving causality in these cases will likely require the use of Findr's conservative, traditional causal inference test in conjunction with the new test, and/or the combination of association signals from multiple (e)QTLs [40]. Lastly, Findr currently operates on individual-level genotype and (molecular or phenotypic) trait data only, and is thus not directly extendable to emerging Mendelian randomization methods that use summary data from independent eQTL and GWAS studies to attempt to infer causality between genes and phenotypic traits [40].

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.

Supporting information

S1 Text. Supplementary text.
(PDF)

S1 Fig. LLR distributions of the relevance test for hsa-miR-200b-3p on 23722 potential targets of Geuvadis dataset. Real, analytical null, and permuted null distributions are demonstrated in the figure, together with the curve of inferred posterior probability of alternative hypothesis. Permutations were randomly conducted on all potential target genes for 100 times. The alignment between analytical and permuted null distributions and the consistent incremental trend of posterior probability verify our method in deriving analytical null distributions.
(PDF)

S2 Fig. The mean AUROC and AUPR on subsampled data are shown for causal inference with traditional and new tests, together with the baseline correlation test. Every marker corresponds to the average AUROC or AUPR at specific sample sizes. At every sample size we performed 100 subsampling. Half widths of the lines and shades are the standard errors and standard deviations respectively, of AUROC or AUPR. Figures from top to bottom correspond to datasets 1, 2, 3, 5. For dataset 4, see Fig 2.
(PDF)

S3 Fig. The AUROC and AUPR of CIT are shown for all 15 datasets of DREAM challenge. Every marker corresponds to the AUROC or AUPR of one dataset. CIT is an R package that includes the conditional independence test, along with tests 2 and 5, while also comparing

$E \rightarrow A \rightarrow B$ against $E \rightarrow B \rightarrow A$. The subsampling analysis on CIT was not feasible due to its low speed.
(PDF)

S4 Fig. The conditional independence test fails in the presence of hidden confounders. When A and B are both regulated by a hidden confounder C, which is independent of E (left), A becomes a collider and conditioning on A would introduce inter-dependency between E and C, which maintains $E \rightarrow B$ regulation (right).
(PDF)

S5 Fig. Local precision of top predictions for the traditional (left) and novel (right) tests for datasets (top to bottom) 1, 2, 3, and 5 of the DREAM challenge.
(PDF)

S6 Fig. Estimated and real precision-recall curves for dataset 4 of the DREAM challenge. The real precision was computed according to the groundtruth, whilst the estimated precision was obtained from the estimated FDR from the respective inference method (precision = 1 − FDR). Only genes with cis-eQTLs were considered as primary targets in prediction and validation. Both the novel (**A**, **B**) and the traditional (**C**, **D**) tests were evaluated. In **A**, **C** the original groundtruth table was used to validate predictions, whereas in **B**, **D** an extended groundtruth was used that also included indirect regulations at any level based on the original groundtruth.
(PDF)

S7 Fig. Null hypothesis p-values of the conditional independence test on simulated data from the ground truth model $E \rightarrow A^{(t)} \rightarrow B$ with $A^{(t)} \rightarrow A$ under parameter settings other than Fig 3. (**A**, **B**) 100 (**A**) or 999 (**B**) samples. (**C**, **D**) Minor allele frequency is 0.05 (**C**) or 0.3 (**D**). (**E**, **F**) Regarding B's variance from $A^{(t)} \rightarrow B$ as unit variance, B's variance from other sources including measurement errrors is 0.2 (**E**) or 20 (**F**). Unmentioned parameters remain the same as in Fig 3.
(PDF)

S8 Fig. ROC (top) and PR (bottom) curves of miRNA target predictions were compared for Findr's traditional, new, and correlation tests, GENIE3, CIT, and 11 methods in miRLAB, based on Geuvadis data. The solid black lines correspond to expected performances from random predictions. A higher curve indicates better prediction performance.
(PDF)

S9 Fig. Three methods of causal inference were evaluated and compared against the baseline correlation test method (P_0): Findr's new test (P), traditional causal inference test in Findr (P_T), and CIT (C). AUROC and AUPR metrics are measured for three inference tasks. MiRNA compares miRNA target predictions based on Geuvadis miRNA and mRNA expression levels against groundtruths from miRLAB. SiRNA and TF-binding compares gene-gene interaction predictions based on Geuvadis gene expression levels against groundtruths from siRNA silencing and TF-binding measurements respectively. ENCODE compares the same gene-gene interaction predictions against TF-binding networks derived from ENCODE data. Dashed lines indicate expected performances from random predictions.
(PDF)

S10 Fig. Inference precision at estimated precision cutoffs 0.1 to 0.9 with respect to groundtruth network derived from TF binding of 14 TFs from ENCODE data. The number

above each bar indicates the number of positive predictions at the corresponding threshold. The dashed line is precision from random predictions.
(PDF)

S1 Table. Predictions from Findr's new (P), traditional (P_T), and correlation (P_0) tests, and CIT were compared against DREAM challenge leaders on AUROC and AUPR for all 15 DREAM datasets. All cis- and trans-genes are included. DREAM challenge constrained the maximum number of submitted regulations by 100,000, which were also applied in our evaluation. Findr's new test consistently obtained higher AUROC and AUPR than all other methods, including the leaders of DREAM challenge.
(PDF)

S2 Table. AUROCs and AUPRs of miRNA target predictions were compared for Findr's traditional, new, and correlation tests, GENIE3, CIT, and 11 methods in miRLAB, based on Geuvadis data. Higher AUROC and AUPR values signify stronger predictive power. Program running times have units in seconds (s), minutes (m), hours (h), or days (d). Findr outperformed other methods in statistical power and speed, with or without genotype information.
(PDF)

S3 Table. AUROCs and AUPRs of gene target predictions were compared for a selected subset of methods in S2 Table, also based on Geuvadis data. The three gold standards could not agree on method.
(PDF)

Author Contributions

Conceptualization: Tom Michoel.

Data curation: Lingfei Wang.

Formal analysis: Lingfei Wang, Tom Michoel.

Funding acquisition: Tom Michoel.

Investigation: Lingfei Wang, Tom Michoel.

Methodology: Lingfei Wang, Tom Michoel.

Project administration: Tom Michoel.

Software: Lingfei Wang.

Supervision: Tom Michoel.

Validation: Lingfei Wang, Tom Michoel.

Writing – original draft: Lingfei Wang, Tom Michoel.

Writing – review & editing: Lingfei Wang, Tom Michoel.

References

1. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics. 2015; 16:197–212. https://doi.org/10.1038/nrg3891 PMID: 25707927
2. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. Nature Reviews Genetics. 2014; 15(1):34–48. https://doi.org/10.1038/nrg3575 PMID: 24296534

3. Lappalainen T, Sammeth M, Friedlander MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501(7468):506–511. https://doi.org/10.1038/nature12531 PMID: 24037378
4. Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015; 348(6235): 648–660. https://doi.org/10.1126/science.1262110
5. Franzén O, Ermel R, Cohain A, Akers N, Di Narzo A, Talukdar H, et al. Cardiometabolic Risk Loci Share Downstream *Cis* and *Trans* Genes Across Tissues and Diseases. Science. 2016;.
6. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature. 2009; 461:218–223. https://doi.org/10.1038/nature08454 PMID: 19741703
7. Talukdar H, Foroughi Asl H, Jain R, Ermel R, Ruusalepp A, Franzén O, et al. Cross-tissue regulatory gene networks in coronary artery disease. Cell Systems. 2016; 2:196–208. https://doi.org/10.1016/j.cels.2016.02.002 PMID: 27135365
8. Rockman MV. Reverse engineering the genotype—phenotype map with natural genetic variation. Nature. 2008; 456(7223):738–744. https://doi.org/10.1038/nature07633 PMID: 19079051
9. Li Y, Tesson BM, Churchill GA, Jansen RC. Critical reasoning on causal inference in genome-wide linkage and association studies. Trends in Genetics. 2010; 26(12):493–498. https://doi.org/10.1016/j.tig.2010.09.002 PMID: 20951462
10. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics. 2005; 37(7):710–717. https://doi.org/10.1038/ng1589 PMID: 15965475
11. Chen L, Emmert-Streib F, Storey J. Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biology. 2007; 8(10):R219. https://doi.org/10.1186/gb-2007-8-10-r219 PMID: 17931418
12. Aten JE, Fuller TF, Lusis AJ, Horvath S. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Systems Biology. 2008; 2(1):34. https://doi.org/10.1186/1752-0509-2-34 PMID: 18412962
13. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. BMC Genetics. 2009; 10(1):1–15. https://doi.org/10.1186/1471-2156-10-23
14. Neto EC, Broman AT, Keller MP, Attie AD, Zhang B, Zhu J, et al. Modeling causality for pairs of phenotypes in system genetics. Genetics. 2013; 193(3):1003–1013. https://doi.org/10.1534/genetics.112.147124 PMID: 23288936
15. Millstein J, Chen GK, Breton CV. cit: hypothesis testing software for mediation analysis in genomic applications. Bioinformatics. 2016; 32(15):2364–2365. https://doi.org/10.1093/bioinformatics/btw135 PMID: 27153715
16. Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, Simon I, et al. Backup in gene regulatory networks explains differences between binding and knockout results. Mol Syst Biol. 2009; 5(1). https://doi.org/10.1038/msb.2009.33 PMID: 19536199
17. Alon U. Network motifs: theory and experimental approaches. Nat Rev Genet. 2007; 8:450–461. https://doi.org/10.1038/nrg2102 PMID: 17510665
18. Hemani G, Tilling K, Smith GD. Orienting The Causal Relationship Between Imprecisely Measured Traits Using Genetic Instruments. bioRxiv. 2017; p. 117101.
19. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015; 43(7):e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792
20. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. International Journal of Epidemiology. 2010; 39(2):417. https://doi.org/10.1093/ije/dyp334 PMID: 19926667
21. Greenland S. The effect of misclassification in the presence of covariates. American Journal of Epidemiology. 1980; 112(4):564–569. https://doi.org/10.1093/oxfordjournals.aje.a113025 PMID: 7424903
22. Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, et al. Using expression profiling data to identify human microRNA targets. Nat Meth. 2007; 4(12):1045–1049. https://doi.org/10.1038/nmeth1130
23. Su WL, Kleinhanz RR, Schadt EE. Characterizing the role of miRNAs within gene regulatory networks using integrative genomics techniques. Molecular Systems Biology. 2011; 7(1).
24. Le TD, Zhang J, Liu L, Liu H, Li J. miRLAB: An R Based Dry Lab for Exploring miRNA-mRNA Regulatory Relationships. PLoS ONE. 2015; 10(12):1–15. https://doi.org/10.1371/journal.pone.0145386

25. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE. 2010; 5(9):1–10. https://doi.org/10.1371/journal.pone.0012776

26. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation in transcription factor binding. PLoS Genetics. 2014; 10(3):e1004226. https://doi.org/10.1371/journal.pgen.1004226 PMID: 24603674

27. Pinna A, Soranzo N, Hoeschele I, de la Fuente A. Simulating systems genetics data with SysGenSIM. Bioinformatics. 2011; 27(17):2459–2462. https://doi.org/10.1093/bioinformatics/btr407 PMID: 21737438

28. Qi J, Foroughi Asl H, Bjorkegren J, Michoel T. kruX: matrix-based non-parametric eQTL discovery. BMC Bioinformatics. 2014; 15(1):11. https://doi.org/10.1186/1471-2105-15-11 PMID: 24423115

29. Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, et al. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Research. 2012; 40(D1):D222–D229. https://doi.org/10.1093/nar/gkr1161 PMID: 22135297

30. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA—target interactions. Nucleic Acids Research. 2009; 37(suppl 1):D105–D110. https://doi.org/10.1093/nar/gkn851 PMID: 18996891

31. Dweep H, Sticht C, Pandey P, Gretz N. miRWalk—Database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes. Journal of Biomedical Informatics. 2011; 44(5):839–847. https://doi.org/10.1016/j.jbi.2011.05.002 PMID: 21605702

32. Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Research. 2014; 42(D1):D78–D85. https://doi.org/10.1093/nar/gkt1266 PMID: 24304892

33. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012; 489(7414):91–100. https://doi.org/10.1038/nature11245 PMID: 22955619

34. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012; 28(10):1353–1358. https://doi.org/10.1093/bioinformatics/bts163 PMID: 22492648

35. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 2003; 100(16):9440–9445. https://doi.org/10.1073/pnas.1530509100

36. Chen LS, Sangurdekar DP, Storey JD. trigger: Transcriptional Regulatory Inference from Genetics of Gene ExpRession; 2007.

37. Stolovitzky G, Prill RJ, Califano A. Lessons from the DREAM2 Challenges. Annals of the New York Academy of Sciences. 2009; 1158(1):159–195. https://doi.org/10.1111/j.1749-6632.2009.04497.x PMID: 19348640

38. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. In: Proceedings of the 23rd International Conference on Machine Learning. ICML'06. New York, NY, USA: ACM; 2006. p. 233–240. Available from: http://doi.acm.org/10.1145/1143844.1143874.

39. Tong P, Monahan J, Prendergast JG. Shared regulatory sites are abundant in the human genome and shed light on genome evolution and disease pleiotropy. PLoS genetics. 2017; 13(3):e1006673. https://doi.org/10.1371/journal.pgen.1006673 PMID: 28282383

40. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics. 2016; 48(5):481–490. https://doi.org/10.1038/ng.3538 PMID: 27019110