

1 **Efficient and accurate causal inference with hidden con-** 2 **founders from genome-transcriptome variation data**

4 Lingfei Wang and Tom Michoel*

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

7

8 **Abstract**

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

17 **Author summary**

18 Understanding how genetic variation between individuals determines variation in observable traits
19 or disease risk is one of the core aims of genetics. It is known that genetic variation often affects
20 gene regulatory DNA elements and directly causes variation in expression of nearby genes. This
21 effect in turn cascades down to other genes via the complex pathways and gene interaction net-
22 works that ultimately govern how cells operate in an ever changing environment. In theory, when
23 genetic variation and gene expression levels are measured simultaneously in a large number of
24 individuals, the causal effects of genes on each other can be inferred using statistical models
25 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

26 software Findr which, unlike existing methods, takes into account the complex but unknown net-
27 work context when predicting causality between specific gene pairs. Findr's predictions have a
28 significantly higher overlap with known gene networks compared to existing methods, using both
29 simulated and real data. Findr is also nearly a million times faster, and hence the only software in
30 its class that can handle modern datasets where the expression levels of ten-thousands of genes
31 are simultaneously measured in hundreds to thousands of individuals.

32 **1 Introduction**

33 Genetic variation in non-coding genomic regions, including at loci associated with complex traits
34 and diseases identified by genome-wide association studies, predominantly plays a gene-regulato-
35 ry role¹. Whole genome and transcriptome analysis of natural populations has therefore be-
36 come a common practice to understand how genetic variation leads to variation in phenotypes².
37 The number and size of studies mapping genome and transcriptome variation has surged in
38 recent years due to the advent of high-throughput sequencing technologies, and ever more ex-
39 pensive catalogues of expression-associated DNA variants, termed expression quantitative trait
40 loci (eQTLs), are being mapped in humans, model organisms, crops and other species^{1,3-5}.
41 Unravelling the causal hierarchies between DNA variants and their associated genes and pheno-
42 types is now the key challenge to enable the discovery of novel molecular mechanisms, disease
43 biomarkers or candidate drug targets from this type of data^{6,7}.

44 It is believed that genetic variation can be used to infer the causal directions of regulation between
45 coexpressed genes, based on the principle that genetic variation causes variation in nearby gene
46 expression and acts as a causal anchor for identifying downstream genes^{8,9}. Although numerous
47 statistical models have been proposed for causal inference with genotype and gene expression
48 data from matching samples¹⁰⁻¹⁵, no software implementation in the public domain is efficient
49 enough to handle the volume of contemporary datasets, hindering any attempts to evaluate their
50 performances. Moreover, existing statistical models rely on a conditional independence test which
51 assumes that no hidden confounding factors affect the coexpression of causally related gene
52 pairs. However gene regulatory networks are known to exhibit redundancy¹⁶ and are organized
53 into higher order network motifs¹⁷, suggesting that confounding of causal relations by known or
54 unknown common upstream regulators is the rule rather than the exception. Moreover, it is also

55 known that the conditional independence test is susceptible to variations in relative measurement
56 errors between genes^{8,9}, an inherent feature of both microarray and RNA-seq based expression
57 data¹⁸.

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

63 2 Results

64 2.1 Findr incorporates existing and novel causal inference tests

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

79 **2.2 Traditional causal inference fails in the presence of hidden confounders and** 80 **weak regulations**

81 Findr's computational speed allowed us to systematically evaluate traditional causal inference
82 methods for the first time. We obtained five datasets with 999 samples simulated from synthetic
83 gene regulatory networks of 1,000 genes with known genetic architecture from the DREAM5
84 Systems Genetics Challenge (**Section 4.1**), and subsampled each dataset to observe how per-
85 formance depends on sample size (**Section 4.7**). The correlation test (P_0) does not incorporate
86 genotype information and was used as a benchmark for performance evaluations in terms of
87 areas under the receiver operating characteristic (AUROC) and precision-recall (AUPR) curves
88 (**Section 4.7**). The traditional method¹¹ combines the secondary (P_2) and independence (P_3)
89 tests sequentially (**Table 1, Section 4.6**), and was evaluated by comparing P_2 and P_2P_3 sep-
90 arately against the correlation test. Both the secondary test alone and the traditional causal
91 inference test combination were found to *underperform* the correlation test (**Figure 1A,B**). More-
92 over, the inclusion of the conditional independence test *worsened* inference accuracy, more
93 so with increasing sample size (**Figure 1A,B**) and increasing number of regulations per gene
94 (**Supplementary Material S2.3**). Similar performance drops were also observed for the Causal
95 Inference Test (CIT)^{13,15} software, which also is based on the conditional independence test
96 (**Figure S3**).

97 We believe that the failure of traditional causal inference is due to an elevated false negative rate
98 (FNR) coming from two sources. First, the secondary test is less powerful in identifying weak
99 interactions than the correlation test. In a true regulation $E \rightarrow A \rightarrow B$, the secondary linkage
100 ($E \rightarrow B$) is the result of two direct linkages chained together, and is harder to detect than either of
101 them. The secondary test hence picks up fewer true regulations, and consequently has a higher
102 FNR. Second, the conditional independence test is counter-productive in the presence of hidden
103 confounders (i.e. common upstream regulators). In such cases, even if $E \rightarrow A \rightarrow B$ is genuine,
104 the conditional independence test will find E and B to be still correlated after conditioning on
105 A due to a collider effect (**Figure S5**)¹⁹. Hence the conditional independence test only reports
106 positive on $E \rightarrow A \rightarrow B$ relations without confounder, further raising the FNR. This is supported
107 by the observation of worsening performance with increasing sample size (where confounding
108 effects become more distinguishable) and increasing number of regulations per gene (which leads

109 to more confounding).

110 To further support this claim, we examined the inference precision among the top predictions
111 from the traditional test, separately for gene pairs directly unconfounded or confounded by at
112 least one gene (**Section 4.7**). Compared to unconfounded gene pairs, confounded ones resulted
113 in significantly more false positives among the top predictions (**Figure 1C**). Furthermore, the
114 vast majority of real interactions fell outside the top 1% of predictions (i.e. had small posterior
115 probability) [92% (651/706) for confounded and 86% (609/709) for unconfounded interactions,
116 **Figure 1C**]. Together, these results again showed the failure of the traditional test on confounded
117 interactions and its high false negative rate overall.

118 **2.3 Findr accounts for weak secondary linkage, allows for hidden confounders,** 119 **and outperforms existing methods on simulated data**

120 To overcome the limitations of traditional causal inference methods, Findr incorporates two addi-
121 tional tests (**Table 1** and **Section 4.3**). The relevance test (P_4) verifies that B is not independent
122 from A and E simultaneously and is more sensitive for picking up weak secondary linkages than
123 the secondary linkage test. The controlled test (P_5) ensures that the correlation between A and
124 B cannot be fully explained by E , i.e. excludes pleiotropy. The same subsampling analysis re-
125 vealed that P_4 performed best in terms of AUROC, and AUPR with small sample sizes, whilst
126 the combination P_2P_5 achieved highest AUPR for larger sample sizes (**Figure 1A,B**). Most impor-
127 tantly, both tests consistently outperformed the correlation test (P_0), particularly for AUPR. This
128 demonstrates conclusively in a comparative setting that the inclusion of genotype data indeed can
129 improve regulatory network inference. These observations are consistent across all five DREAM
130 datasets (**Figure S2**).

131 We combined the advantages of P_4 and P_2P_5 by averaging them in a composite test (P) (**Section**
132 **4.6**), which outperformed P_4 and P_2P_5 at all sample sizes (**Figure 1** and **Figure S2**) and hence
133 was appointed as Findr's new test for causal inference. Findr's new test (P) obtained consistently
134 higher levels of local precision (i.e. one minus local FDR) on confounded and unconfounded gene
135 pairs compared to Findr's traditional causal inference test (P_T) (**Figure 1C,D**), and outperformed
136 the traditional test (P_T), correlation test (P_0), CIT, and every participating method of the DREAM5
137 Systems Genetics Challenge (**Section 4.1**) in terms of AUROC and AUPR on all 15 datasets

138 (Figure 1E,F, Table S1, Figure S4).

139 Specifically, Findr's new test was able to address the inflated FNR of the traditional method due
140 to confounded interactions. It performed almost equally well on confounded and unconfounded
141 gene pairs and, compared to the traditional test, significantly fewer real interactions fell outside
142 the top 1% of predictions (55% vs. 92% for confounded and 45% vs. 86% for unconfounded
143 interactions, Figure 1D, Figure S6).

144 2.4 The conditional independence test incurs false negatives for unconfounded 145 regulations due to measurement error

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

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

160 We verified our hypothesis with a simple simulation (Section 4.8). In a typical scenario with 300
161 samples from a monoallelic species, minor allele frequency 0.1, and a third of the total variance
162 of B coming from $A^{(t)}$, the conditional independence test reported false negatives (likelihood
163 ratio p-value $\ll 1$, i.e. rejecting the null hypothesis of conditional independence, cf. Table 1) as
164 long as measurement error contributed more than half of A 's total unexplained variance (Figure
165 2B). False negatives occurred at even weaker measurement errors, when the sample sizes were
166 larger or when stronger $A \rightarrow B$ regulations were assumed (Figure S7).

167 This observation goes beyond the well-known problems that arise from a large measurement
168 error in all variables, which acts like a hidden confounder⁹, or from a much larger measurement
169 error in A than B , which can result in B becoming a better measurement of $A^{(t)}$ than A itself⁸. In
170 this simulation, the false negatives persisted even if $E \rightarrow A$ was observationally much stronger
171 than $E \rightarrow B$, such as when A 's measurement error was only 10% ($\sigma_{A1}^2 = 0.1$) compared to up to
172 67% for B (**Figure 2B**). This suggested a unique and mostly neglected source of false negatives
173 that would not affect other tests. Indeed, the secondary, relevance, and controlled tests were
174 much less sensitive to measurement errors (**Figure 2A,C,D**).

175 **2.5 Findr outperforms traditional causal inference and machine learning methods** 176 **on microRNA target prediction**

177 In order to evaluate Findr on a real dataset, we performed causal inference on miRNA and mRNA
178 sequencing data in lymphoblastoid cell lines from 360 European individuals in the Geuvadis
179 study³ (**Section 4.1**). We first tested 55 miRNAs with reported significant cis-eQTLs against
180 23,722 genes. Since miRNA target predictions from sequence complementarity alone result in
181 high numbers of false positives, prediction methods based on correlating miRNA and gene ex-
182 pression profiles are of great interest²¹. Although miRNA target prediction using causal inference
183 from genotype and gene expression data has been considered²², it remains unknown whether
184 the inclusion of genotype data improves existing expression-based methods. To compare Findr
185 against the state-of-the-art for expression-based miRNA target prediction, we used miRLAB, an
186 integrated database of experimentally confirmed human miRNA target genes with a uniform inter-
187 face to predict targets using twelve methods, including linear and non-linear, pairwise correlation
188 and multivariate regression methods²³. We were able to infer miRNA targets with 11/12 miRLAB
189 methods, and also applied the GENIE3 random forest regression method²⁴, CIT, and the three
190 tests in Findr: the new (P) and traditional (P_T) causal inference tests and the correlation test (P_0)
191 (**Supplementary Material S2.4**). Findr's new test achieved highest AUROC and AUPR among
192 the 16 methods attempted. In particular, Findr's new test significantly outperformed the traditional
193 test and CIT, the two other genotype-assisted methods, while also being over 500,000 times faster
194 than CIT (**Figure 3, Table S2, Figure S8**). Findr's correlation test outperformed all other methods
195 not using genotype information, including correlation, regression, and random forest methods,
196 and was 500 to 100,000 times faster (**Figure 3, Table S2, Figure S8**). This further illustrates the

197 power of the Bayesian gene-specific background estimation method implemented in all Findr's
198 tests (**Section 4.5**).

199 **2.6 Findr predicts transcription factor targets with more accurate FDR estimates**

200 We considered 3,172 genes with significant cis-eQTLs in the Geuvadis data³ (**Section 4.1**) and
201 inferred regulatory interactions to the 23,722 target genes using Findr's traditional (P_T), new (P)
202 and correlation (P_0) tests, and CIT. Groundtruths of experimentally confirmed causal gene inter-
203 actions in human, and mammalian systems more generally, are of limited availability and mainly
204 concern transcription or transcription-associated DNA-binding factors (TFs). Here we focused on
205 a set of 25 TFs in the set of eQTL-genes for which either differential expression data following
206 siRNA silencing (6 TFs) or TF-binding data inferred from ChIP-sequencing and/or DNase foot-
207 printing (20 TFs) in a lymphoblastoid cell line (GM12878) was available²⁵ (**Section 4.1**). AUPRs
208 and AUROCs did not exhibit substantial differences, other than modest improvement over random
209 predictions (**Figure S9**). To test for enrichment of true positives among the top-ranked predictions,
210 which would be missed by global evaluation measures such as AUPR or AUROC, we took ad-
211 vantage of the fact that Findr's probabilities are empirical local precision estimates for each test
212 (**Section 4.5**), and assessed how estimated local precisions of new, traditional, and correlation
213 tests reflected the actual precision. Findr's new test correctly reflected the precision values at
214 various threshold levels, and was able to identify true regulations at high precision control levels
215 (**Figure 4**). However, the traditional test significantly underestimated precision due to its elevated
216 FNR. This lead to a lack of predictions at high precision thresholds but enrichment of true regula-
217 tions at low thresholds, essentially nullifying the statistical meaning of its output probability P_T . On
218 the other hand, the correlation test significantly overestimated precisions because it is unable to
219 distinguish causal, reversed causal or confounded interactions, which raises its FDR. The same
220 results were observed when alternative groundtruth ChIP-sequencing networks were considered
221 (**Figure S9, Figure S10**).

222 3 Discussion

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

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

241 Although the newly proposed test avoids the elevated FNR from the conditional independence
242 test, it is not without its own limitations. Unlike the conditional independence test, the relevance
243 and controlled tests (**Table 1**) are symmetric between the two genes considered. Therefore the
244 direction of causality in the new test arises predominantly from using a different eQTL when testing
245 the reverse interaction, potentially leading to a higher FDR as a minor trade-off. About 10% of
246 cis-regulatory eQTLs are linked (as *cis*-eQTLs) to the expression of more than one gene²⁶. In
247 these cases, it appears that the shared *cis*-eQTL regulates the genes independently²⁶, which
248 in Findr is accounted for by the 'controlled' test (**Table 1**). When causality between genes and
249 phenotypes or among phenotypes is tested, sharing or linkage of (e)QTLs can be more common.
250 Resolving causality in these cases may require the use of Findr's conservative, traditional causal
251 inference test, in conjunction with the new test.

252 In this paper we have addressed the challenge of pairwise causal inference, but to reconstruct
253 the actual pathways and networks that affect a phenotypic trait, two important limitations have
254 to be considered. First, linear pathways propagate causality, and may thus appear as densely
255 connected sets of triangles in pairwise causal networks. Secondly, most genes are regulated by
256 multiple upstream factors, and hence some true edges may only have a small posterior probability
257 unless they are considered in an appropriate multivariate context. The most straightforward way
258 to address these issues would be to model the real directed interaction network as a Bayesian
259 network with sparsity constraints. A major advantage of Findr is that it outputs probability val-
260 ues which can be directly incorporated as prior edge probabilities in existing network inference
261 softwares.

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

269 **4 Methods**

270 **4.1 Datasets**

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

- 272 1. Simulated genotype and transcriptome data of synthetic gene regulatory networks from the
273 DREAM5 Systems Genetics challenge A (DREAM for short)²⁷, generated by the SysGen-
274 SIM software²⁸. DREAM provides 15 sub-datasets, obtained by simulating 100, 300, and
275 999 samples of 5 different networks each, containing 1000 genes in every sub-dataset
276 but more regulations for sub-datasets with higher numbering. In every sub-dataset, each
277 gene has exactly one matching genotype variable. 25% of the genotype variables are cis-
278 expression Quantitative Trait Loci (eQTL), defined in DREAM as: their variation changes
279 the expression level of the corresponding gene directly. The other 75% are trans-eQTLs,

280 defined as: their variation affects the expression levels of only the *downstream targets* of
281 the corresponding gene, but not the gene itself. Because the identities of cis-eQTLs are un-
282 known, we calculated the P-values of genotype-gene expression associations with kruX²⁹,
283 and kept all genes with a P-value less than 1/750 to filter out genes without cis-eQTL. For
284 the subsampling analysis (detailed in **Section 4.7**), we restricted the evaluation to the pre-
285 diction of target genes from these cis-genes only, in line with the assumption that Findr
286 and other causal inference methods require as input a list of genes whose expression is
287 significantly associated with at least one cis-eQTL. For the full comparison of Findr to the
288 DREAM leaderboard results, we predicted target genes for all genes, regardless of whether
289 they had a cis-eQTL.

290 2. Genotype and transcriptome sequencing data on 465 human lymphoblastoid cell line sam-
291 ples from the Geuvadis project³ consisting of the following data products:

- 292 • Genotype data (ArrayExpress accession E-GEUV-1)³⁰.
- 293 • Gene quantification data for 23722 genes from nonredundant unique samples and
294 after quality control and normalization (ArrayExpress accession E-GEUV-1)³¹.
- 295 • Quantification data of miRNA, with the same standard as gene quantification data
296 (ArrayExpress accession E-GEUV-2)³².
- 297 • Best eQTLs of mRNAs and miRNAs (ArrayExpress accessions E-GEUV-1 and E-
298 GEUV-2)^{33,34}.

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

302 3. For validation of predicted miRNA-gene interactions, we extracted the “strong” ground-truth
303 table from miRLAB^{23,35}, which contains experimentally confirmed miRNA-gene regulations
304 from the following databases: TarBase³⁶, miRecords³⁷, miRWalk³⁸, and miRTarBase³⁹.
305 The intersection of the Geuvadis and ground-truth table contains 20 miRNAs and 1054
306 genes with 1217 confirmed regulations, which are considered for prediction validation. In-
307 teractions that are present in the ground-truth table are regarded as true while others as
308 false.

309 4. For verification of predicted gene-gene interactions, we obtained differential expression data
310 following siRNA silencing of 59 transcription-associated factors (TFs) and DNA-binding data
311 of 201 TFs for 8872 genes in a reference lymphoblastoid cell line (GM12878) from²⁵. Six
312 siRNA-targeted TFs, 20 DNA-binding TFs, and 6,790 target genes without missing differ-
313 ential expression data intersected with the set of 3172 eQTL-genes and 23722 target genes
314 in Geuvadis and were considered for validation. We reproduced the pipeline of²⁵ with the
315 criteria for true targets as having a False Discovery Rate (FDR) < 0.05 from R package
316 *qvalue* for differential expression in siRNA silencing, or having at least 2 TF-binding peaks
317 within 10kb of their transcription start site. We also obtained the filtered proximal TF-target
318 network from⁴⁰, which had 14 TFs and 7,000 target genes in common with the Geuvadis
319 data.

320 4.2 General inference algorithm

321 Consider a set of observations sampled from a mixture distribution of a null and an alternative
322 hypothesis. For instance in gene regulation, every observation can correspond to expression
323 levels of a pair of genes which are sampled from a bivariate normal distribution with zero (null
324 hypothesis) or non-zero (alternative hypothesis) correlation coefficient. In Findr, we predict the
325 probability that any sample follows the alternative hypothesis with the following algorithm (based
326 on and modified from¹¹):

- 327 1. For robustness against outliers, we convert every continuous variable into standard nor-
328 mally distributed $N(0, 1)$ values using a rank-based inverse normal transformation across
329 all samples. We name this step as *supernormalization*.
- 330 2. We propose a null and an alternative hypothesis for every likelihood ratio test (LRT) of inter-
331 est where, by definition, the null hypothesis space is a subset of the alternative hypothesis.
332 Model parameters are replaced with their maximum likelihood estimators (MLEs) to obtain
333 the log likelihood ratio (LLR) between the alternative and null hypotheses (**Section 4.3**).
- 334 3. We derive the analytical expression for the probability density function (PDF) of the LLR
335 when samples follow the null hypothesis (**Section 4.4**).

336 4. We convert LLRs into posterior probabilities of the hypothesis of interest with the empirical
337 estimation of local FDR (**Section 4.5**).

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

339 4.3 Likelihood ratio tests

340 Consider correlated genes A , B , and a third variable E upstream of A and B , such as a significant
341 eQTL of A . The eQTLs can be obtained either *de novo* using eQTL identification tools such as
342 matrix-eQTL⁴¹ or kruX²⁹, or from published analyses. Throughout this article, we assume that E
343 is a significant eQTL of A , whereas extension to other data types is straightforward. We use A_i
344 and B_i for the expression levels of gene A and B respectively, which are assumed to have gone
345 through the supernormalization in **Section 4.2**, and optionally the genotypes of the best eQTL
346 of A as E_i , where $i = 1, \dots, n$ across samples. Genotypes are assumed to have a total of n_a
347 alleles, so $E_i \in \{0, \dots, n_a\}$. We define the null and alternative hypotheses for a total of six tests,
348 as shown in **Table 1**. LLRs of every test are calculated separately as follows:

349 0. **Correlation test:** Define the null hypothesis as A and B are independent, and the alterna-
350 tive hypothesis as they are correlated:

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

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

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

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

354 Maximum likelihood estimators (MLE) for the model parameters ρ , σ_{A0} , and σ_{B0} are

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

355 and the LLR is simply

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

356 In the absence of genotype information, we use nonzero correlation between A and B as
 357 the indicator for $A \rightarrow B$ regulation, giving the posterior probability

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

358 false negative

359 **1. Primary (linkage) test:** Verify that E regulates A from $\mathcal{H}_{\text{alt}}^{(1)} \equiv E \rightarrow A$ and $\mathcal{H}_{\text{null}}^{(1)} \equiv E \not\rightarrow A$.
 360 For $\mathcal{H}_{\text{alt}}^{(1)}$, we model $E \rightarrow A$ as A follows a normal distribution whose mean is determined
 361 by E categorically, i.e.

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

362 From the total likelihood $p(A \mid E) = \prod_{i=1}^n p(A_i \mid E_i)$, we find MLEs for model parameters
 363 $\mu_j, j = 0, 1, \dots, n_a$, and σ_A , as

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

364 where n_j is the sample count by genotype category,

$$n_j \equiv \sum_{i=1}^n \delta_{E_i j}.$$

365 The Kronecker delta function is defined as $\delta_{xy} = 1$ for $x = y$, and 0 otherwise. When
 366 summing over all genotype values ($j = 0, \dots, n_a$), we only pick those that exist ($n_j >$
 367 0) throughout this article. Since the null hypothesis is simply that A_i is sampled from a
 368 genotype-independent normal distribution, with MLEs of mean zero and standard deviation
 369 one due to the supernormalization (**Section 4.2**), the LLR for test 1 becomes

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

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

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

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

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

374 and the LLR as

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

375 $\mathcal{H}_{\text{alt}}^{(2)}$ is chosen to verify that E regulates B .

376 **3. (Conditional) independence test:** Verify that E and B are independent when conditioning
 377 on A . This can be achieved by comparing $\mathcal{H}_{\text{alt}}^{(3)} \equiv B \leftarrow E \rightarrow A \wedge (A \text{ correlates with } B)$
 378 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
 379 regulates B only through A :

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

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

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

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

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

382 Substituting parameters $\mu_j, \nu_j, \sigma_A, \sigma_B, \rho$ of $\mathcal{H}_{\text{alt}}^{(3)}$ and $\mu_j, \rho, \sigma_A, \sigma_B$ of $\mathcal{H}_{\text{null}}^{(3)}$ with their MLEs,

383 we obtain the LLR:

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

384 where

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

385 and $\hat{\rho}$ is defined in Eq 2.

386 **4. Relevance test:** Since the indirect regulation $E \rightarrow B$ tends to be weaker than any of its
 387 direct regulation components ($E \rightarrow A$ or $A \rightarrow B$), we propose to test $E \rightarrow A \rightarrow B$ with
 388 indirect regulation $E \rightarrow B$ as well as the direct regulation $A \rightarrow B$ for stronger distinguishing
 389 power on weak regulations. We define $\mathcal{H}_{\text{alt}}^{(4)} \equiv E \rightarrow A \wedge E \rightarrow B \leftarrow A$ and $\mathcal{H}_{\text{null}}^{(4)} \equiv E \rightarrow$
 390 $A \quad B$. This simply verifies that B is not independent from both A and E simultaneously.
 391 In the alternative hypothesis, B is regulated by E and A , which is modeled as a normal
 392 distribution whose mean is additively determined by E categorically and A linearly, i.e.

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

393 We can hence solve its LLR as

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

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

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

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

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

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

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

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

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

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

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

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

$$X_1 \sim N(0, 1), \quad X_2 \sim N(0, 1), \quad X_3 \sim \chi^2(n-2). \quad (12)$$

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

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

412 in which

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

413 follows the Beta distribution.

414 We define distribution $\mathcal{D}(k_1, k_2)$ as the distribution of a random variable $Z = -\frac{1}{2} \ln(1 - Y)$
 415 for $Y \sim \text{Beta}(k_1/2, k_2/2)$, i.e.

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

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

$$p(z | k_1, k_2) = \frac{2}{B(k_1/2, k_2/2)} (1 - e^{-2z})^{(k_1/2-1)} e^{-k_2z}, \quad (15)$$

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

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

419 1. **Primary test:** $\mathcal{H}_{\text{null}}^{(1)} = E \rightarrow A$ indicates no regulation from E to A . Therefore, similarly
 420 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
 421 zero mean and unit variance.

422 The expression of $\text{LLR}^{(1)}$ then becomes:

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

423 where

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

424 For now, assume all possible genotypes are present, i.e. $n_j > 0$ for $j = 0, \dots, n_a$. Trans-

425 form $\{\tilde{A}_i\}$ by defining

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

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

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

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

$$\begin{aligned} Y_1 &\equiv \sum_{j=0}^{n_a} \sqrt{\frac{n_j}{n}} X_j \sim N(0, 1), \\ Y_2 &\equiv \left(\sum_{j=0}^{n_a} X_j^2 \right) - Y_1^2 \sim \chi^2(n_a), \\ Y_3 &\equiv X_{n_a+1} \sim \chi^2(n - n_a - 1). \end{aligned}$$

428 Some calculation would reveal

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

429 i.e.

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

430 To account for genotypes that do not show up in the samples, define $n_v \equiv \sum_{j \in \{j|n_j > 0\}} 1$
431 as the number of different genotype values across all samples. Then

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

432 2. **Secondary test:** Since the null hypotheses and LLRs of primary and secondary tests are
433 identical, $\text{LLR}^{(2)}$ follows the same null distribution as Eq 19.

434 **3. Independence test:** The independence test verifies if E and B are uncorrelated when
 435 conditioning on A , with $\mathcal{H}_{\text{null}}^{(3)} = E \rightarrow A \rightarrow B$. For this purpose, we keep E and A intact
 436 while randomizing \tilde{B}_i according to B 's correlation with A :

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

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

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

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

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

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

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

442 and then normalize \tilde{B}_i into B_i according to Eq 8. Some calculation reveals the null distri-
 443 bution as

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

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

446 **4.5 Bayesian inference of posterior probabilities**

447 After obtaining the PDFs for the LLRs from real data and the null hypotheses, we can convert LLR
 448 values into posterior probabilities $P(\mathcal{H}_{\text{alt}} | \text{LLR})$. We use a similar technique as in¹¹, which itself
 449 was based on a more general framework to estimate local FDRs in genome-wide studies⁴². This
 450 framework assumes that the real distribution of a certain test statistic forms a mixture distribution

451 of null and alternative hypotheses. After estimating the null distribution, either analytically or by
 452 simulation, it can be compared against the real distribution to determine the proportion of null
 453 hypotheses, and consequently the posterior probability that the alternative hypothesis is true at
 454 any value of the statistic.

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

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

459 The priors $P(\mathcal{H}_{\text{null}})$ and $P(\mathcal{H}_{\text{alt}})$ sum to unity and correspond to the proportions of null and
 460 alternative hypotheses in the mixture distribution. For any test $i = 0, \dots, 5$, Bayes' theorem then
 461 yields its posterior probability as

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

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

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

464 After obtaining the LLR distribution of the null hypothesis [$p(\text{LLR} | \mathcal{H}_{\text{null}})$], we can determine its
 465 proportion [$P(\mathcal{H}_{\text{null}})$] by aligning $p(\text{LLR} | \mathcal{H}_{\text{null}})$ with the real distribution $p(\text{LLR})$ at the LLR \rightarrow
 466 0^+ side. This provides all the prerequisites to perform Bayesian inference and obtain any P_i from
 467 Eq 23.

468 In practice, PDFs are approximated with histograms. This requires proper choices of histogram
 469 bin widths, $P(\mathcal{H}_{\text{null}})$, and techniques to ensure the conversion from LLR to posterior probability
 470 is monotonically increasing and smooth. Implementational details can be found in Findr package
 471 and in **Section S1.1**. Distributions can be estimated either separately for every (E, A) pair or by
 472 pooling across all (E, A) pairs. In practice, we test on the order of 10^3 to 10^4 candidate targets

473 (“ B ”) for every (E, A) such that a separate conversion of LLR values to posterior probabilities is
474 both feasible and recommended, as it accounts for different roles of every gene, especially hub
475 genes, through different rates of alternative hypotheses.

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

479 4.6 Tests to evaluate

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

- 482 • The correlation test is introduced as a benchmark, against which we can compare other
483 methods involving genotype information. Pairwise correlation is a simple measure for the
484 probability of two genes being functionally related either through direct or indirect regulation,
485 or through coregulation by a third factor. Bayesian inference additionally considers different
486 gene roles. Its predicted posterior probability for regulation is P_0 .
- 487 • The traditional causal inference test, as explained in¹¹, suggested that the regulatory rela-
488 tion $E \rightarrow A \rightarrow B$ can be confirmed with the combination of three separate tests: E
489 regulates A , E regulates B , and E only regulates B through A (i.e. E and B become
490 independent when conditioning on A). They correspond to the primary, secondary, and
491 independence tests respectively. The regulatory relation $E \rightarrow A \rightarrow B$ is regarded pos-
492 itive only when all three tests return positive. The three tests filter the initial hypothesis
493 space of all possible relations between E , A , and B , sequentially to $E \rightarrow A$ (primary test),
494 $E \rightarrow A \wedge E \rightarrow B$ (secondary test), and $E \rightarrow A \rightarrow B \wedge (\text{no confounder for } A \text{ and } B)$ (con-
495 ditional independence test). The resulting test is stronger than $E \rightarrow A \rightarrow B$ by disallowing
496 confounders for A and B . So its probability can be broken down as

$$P_T \equiv P_1 P_2 P_3. \quad (24)$$

497 Trigger⁴³ is an R package implementation of the method. However, since Trigger integrates
498 eQTL discovery with causal inference, it is not practical for use on modern datasets. For

499 this reason, we reimplemented this method in Findr, and evaluated it with P_2 and P_2P_3
500 separately, in order to assess the individual effects of secondary and independence tests.
501 As discussed above, we expect a set of significant eQTLs and their associated genes as
502 input, and therefore $P_1 = 1$ is assured and not calculated in this paper or the package Findr.
503 Note that P_T is the estimated local precision, i.e. the probability that tests 2 and 3 are both
504 true. Correspondingly, its local FDR (the probability that one of them is false) is $1 - P_T$.

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

$$P \equiv \frac{1}{2}(P_2P_5 + P_4). \quad (25)$$

507 Specifically, the first term in Eq 25 accounts for hidden confounders. The controlled test re-
508 places the conditional independence test and constrains the hypothesis space more weakly,
509 only requiring the correlation between A and B is not entirely due to pleiotropy. Therefore,
510 P_2P_5 (with $P_1 = 1$) verifies the hypothesis that $B \leftarrow E \rightarrow A \wedge (A \not\perp B \mid E)$, a superset of
511 $E \rightarrow A \rightarrow B$.

512 On the other hand, the relevance test in the second term of Eq 25 addresses weak in-
513 teractions that are undetectable by the secondary test from existing data (P_2 close to 0).
514 This term still grants higher-than-null significance to weak interactions, and verifies that
515 $E \rightarrow A \wedge (E \rightarrow B \vee A \dashv B)$, also a superset of $E \rightarrow A \rightarrow B$. In the extreme undetectable
516 limit where $P_2 = 0$ but $P_4 \neq 0$, the novel test Eq 25 automatically reduces to $P = \frac{1}{2}P_4$,
517 which assumes equal probability of either direction and assigns half of the relevance test
518 probability to $A \rightarrow B$.

519 The composite design of the novel test aims not to miss any genuine regulation whilst dis-
520 tinguishing the full spectrum of possible interactions. When the signal level is too weak for
521 tests 2 and 5, we expect P_4 to still provide distinguishing power better than random predic-
522 tions. When the interaction is strong, P_2P_5 is then able to pick up true targets regardless of
523 the existence of hidden confounders.

524 4.7 Evaluation methods

525 • Evaluation metrics:

526 Given the predicted posterior probabilities for every pair (A, B) from any test, or more gener-
527 ically a score from any inference method, we evaluated the predictions against the direct
528 regulations in the ground-truth tables (**Section 4.1**) with the metrics of Receiver Operating
529 Characteristic (ROC) and Precision-Recall (PR) curves, as well as the Areas Under the
530 ROC (AUROC) and Precision-Recall (AUPR) curves⁴⁴. In particular, AUPR is calculated
531 with the Davis-Goadrich nonlinear interpolation⁴⁵ with R package *PRROC*.

532 • Subsampling:

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

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

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

548 • Local precision of top predictions separately for confounded and unconfounded 549 gene pairs:

550 In order to demonstrate the inferential precision among top predictions for any inference
551 test (here the traditional and novel tests separately), we first ranked all (ordered) gene pairs

552 (A, B) according to the inferred significance for $A \rightarrow B$. All gene pairs were split into
553 groups according to their relative significance ranking (9 groups in **Figure 1C,D**, as top 0%
554 to 0.01%, 0.01% to 0.02%, etc). Each group was divided into two subgroups, based on
555 whether each gene pair shared at least one direct upstream regulator gene (confounded)
556 or not (unconfounded), according to the gold standard. Within each subgroup, the local
557 precision was computed as the number of true directed regulations divided by the total
558 number of gene pairs in the subgroup.

559 **4.8 Simulation studies on causal models with measurement error**

560 We investigated how each statistical test tolerates measurement errors with simulations in a con-
561 trolled setting. We modelled the causal relation $A \rightarrow B$ in a realistic setup as $E \rightarrow A^{(t)} \rightarrow B$
562 with $A^{(t)} \rightarrow A$. E remains as the accurately measured genotype values as the eQTL for the
563 primary target gene A . $A^{(t)}$ is the true expression level of gene A , which is not observable. A
564 is the measured expression level for gene A , containing measurement errors. B is the measured
565 expression level for gene B .

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

$$A_i^{(t)} \sim N(\widetilde{E}_i, \sigma_{A1}^2), \quad (26)$$

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

$$B_i \sim N(\widetilde{A}_i^{(t)}, \sigma_B^2), \quad (28)$$

571 in which σ_{A1} , σ_{A2} , and σ_B are parameters of the model. Note that σ_B^2 is B 's variance from
572 all unknown sources, including expression level variations and measurement errors. The tilde
573 normalizes the variable into zero mean and unit variance, as:

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

574 where \bar{X} and $\text{Var}(X)$ are the mean and variance of $X \equiv \{X_i\}$ respectively.

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

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

586 Acknowledgements

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

588 References

- 589 1. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nature*
590 *Reviews Genetics*. 2015;16:197–212.
- 591 2. Civelek M, Lusk AJ. Systems genetics approaches to understand complex traits. *Nature*
592 *Reviews Genetics*. 2014;15(1):34–48.
- 593 3. Lappalainen T, Sammeth M, Friedlander MR, 't Hoen PAC, Monlong J, Rivas MA, et al.
594 Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*.
595 2013 09;501(7468):506–511. Available from: <http://dx.doi.org/10.1038/nature12531>.
- 596 4. Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-
597 Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*.
598 2015;348(6235):648–660.
- 599 5. Franzén O, Ermel R, Cohain A, Akers N, Di Narzo A, Talukdar H, et al. Cardiometabolic Risk
600 Loci Share Downstream *Cis* and *Trans* Genes Across Tissues and Diseases. *Science*. 2016;.

- 601 6. Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature*.
602 2009;461:218–223.
- 603 7. Talukdar H, Foroughi Asl H, Jain R, Ermel R, Ruusalepp A, Franzén O, et al. Cross-tissue
604 regulatory gene networks in coronary artery disease. *Cell Systems*. 2016;2:196–208.
- 605 8. Rockman MV. Reverse engineering the genotype–phenotype map with natural genetic vari-
606 ation. *Nature*. 2008;456(7223):738–744.
- 607 9. Li Y, Tesson BM, Churchill GA, Jansen RC. Critical reasoning on causal inference in genome-
608 wide linkage and association studies. *Trends in Genetics*. 2010;26(12):493–498.
- 609 10. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, et al. An integrative ge-
610 nomics approach to infer causal associations between gene expression and disease. *Nature*
611 *Genetics*. 2005;37(7):710–717.
- 612 11. Chen L, Emmert-Streib F, Storey J. Harnessing naturally randomized transcription to infer
613 regulatory relationships among genes. *Genome Biology*. 2007;8(10):R219. Available from:
614 <http://genomebiology.com/2007/8/10/R219>.
- 615 12. Aten JE, Fuller TF, Lusk AJ, Horvath S. Using genetic markers to orient the edges in quanti-
616 tative trait networks: the NEO software. *BMC Systems Biology*. 2008;2(1):34.
- 617 13. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal
618 inference test. *BMC Genetics*. 2009;10(1):1–15. Available from: [http://dx.doi.org/10.](http://dx.doi.org/10.1186/1471-2156-10-23)
619 [1186/1471-2156-10-23](http://dx.doi.org/10.1186/1471-2156-10-23).
- 620 14. Neto EC, Broman AT, Keller MP, Attie AD, Zhang B, Zhu J, et al. Modeling causality for pairs
621 of phenotypes in system genetics. *Genetics*. 2013;193(3):1003–1013.
- 622 15. Millstein J, Chen GK, Breton CV. *cit*: hypothesis testing software for mediation analysis
623 in genomic applications. *Bioinformatics*. 2016;32(15):2364–2365. Available from: [http:](http://bioinformatics.oxfordjournals.org/content/32/15/2364.abstract)
624 [//bioinformatics.oxfordjournals.org/content/32/15/2364.abstract](http://bioinformatics.oxfordjournals.org/content/32/15/2364.abstract).
- 625 16. Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, Simon I, et al. Backup in gene regu-
626 latory networks explains differences between binding and knockout results. *Mol Syst Biol*.
627 2009;5(1).
- 628 17. Alon U. Network motifs: theory and experimental approaches. *Nat Rev Genet*. 2007;8:450–
629 461.
- 630 18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. *limma* powers differential ex-
631 pression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*.

- 632 2015;43(7):e47.
- 633 19. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating
634 bias due to conditioning on a collider. *International Journal of Epidemiology*. 2010;39(2):417.
635 Available from: [+http://dx.doi.org/10.1093/ije/dyp334](http://dx.doi.org/10.1093/ije/dyp334).
- 636 20. Greenland S. The effect of misclassification in the presence of covari-
637 ates. *American Journal of Epidemiology*. 1980 Oct;112(4):564–569. Avail-
638 able from: [https://academic.oup.com/aje/article/112/4/564/59323/](https://academic.oup.com/aje/article/112/4/564/59323/THE-EFFECT-OF-MISCLASSIFICATION-IN-THE-PRESENCE-OF)
639 [THE-EFFECT-OF-MISCLASSIFICATION-IN-THE-PRESENCE-OF](#).
- 640 21. Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, et al. Using expression profiling
641 data to identify human microRNA targets. *Nat Meth*. 2007 12;4(12):1045–1049. Available
642 from: <http://dx.doi.org/10.1038/nmeth1130>.
- 643 22. Su WL, Kleinhanz RR, Schadt EE. Characterizing the role of miRNAs within gene regula-
644 tory networks using integrative genomics techniques. *Molecular Systems Biology*. 2011;7(1).
645 Available from: <http://msb.embopress.org/content/7/1/490>.
- 646 23. Le TD, Zhang J, Liu L, Liu H, Li J. miRLAB: An R Based Dry Lab for Exploring miRNA-
647 mRNA Regulatory Relationships. *PLoS ONE*. 2015 12;10(12):1–15. Available from: <http://dx.doi.org/10.1371/journal.pone.0145386>.
- 648
- 649 24. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks from Ex-
650 pression Data Using Tree-Based Methods. *PLoS ONE*. 2010 09;5(9):1–10. Available from:
651 <http://dx.doi.org/10.1371/journal.pone.0012776>.
- 652 25. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation
653 in transcription factor binding. *PLoS Genetics*. 2014;10(3):e1004226.
- 654 26. Tong P, Monahan J, Prendergast JG. Shared regulatory sites are abundant in the hu-
655 man genome and shed light on genome evolution and disease pleiotropy. *PLoS genetics*.
656 2017;13(3):e1006673.
- 657 27. DREAM5 Systems Genetics challenges; 2014. Available from: [https://www.synapse.org/](https://www.synapse.org/#!/Synapse:syn2820440/wiki/)
658 [#!/Synapse:syn2820440/wiki/](#).
- 659 28. Pinna A, Soranzo N, Hoeschele I, de la Fuente A. Simulating systems genetics data
660 with SysGenSIM. *Bioinformatics*. 2011;27(17):2459–2462. Available from: [http://](http://bioinformatics.oxfordjournals.org/content/27/17/2459.abstract)
661 [bioinformatics.oxfordjournals.org/content/27/17/2459.abstract](#).
- 662 29. Qi J, Foroughi Asl H, Bjorkegren J, Michoel T. kruX: matrix-based non-parametric eQTL dis-

- 663 covery. BMC Bioinformatics. 2014;15(1):11. Available from: <http://www.biomedcentral.com/1471-2105/15/11>.
- 664
- 665 30. Geuvadis genotype data; 2013. Available from: <ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/genotypes/>.
- 666
- 667 31. Geuvadis gene expression data; 2013. 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.
- 668
- 669
- 670 32. Geuvadis miRNA expression data; 2013. 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.
- 671
- 672
- 673 33. Geuvadis best eQTL data for mRNA; 2013. 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.
- 674
- 675
- 676 34. Geuvadis best eQTL data for miRNA; 2013. 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.
- 677
- 678
- 679 35. miRLAB 'strong' ground-truth data; 2015. Available from: https://downloads.sourceforge.net/project/mirlab/groundtruth_Strong.csv.
- 680
- 681 36. 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. Available from: <http://nar.oxfordjournals.org/content/40/D1/D222.abstract>.
- 682
- 683
- 684
- 685 37. 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. Available from: http://nar.oxfordjournals.org/content/37/suppl_1/D105.abstract.
- 686
- 687
- 688 38. 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. Available from: <http://dx.doi.org/10.1016/j.jbi.2011.05.002>.
- 689
- 690
- 691 39. 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. Available from: <http://nar.oxfordjournals.org/content/42/D1/D78.abstract>.
- 692
- 693

694 org/content/42/D1/D78.abstract.

695 40. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the
696 human regulatory network derived from ENCODE data. *Nature*. 2012;489(7414):91–100.

697 41. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformat-*
698 *ics*. 2012;28(10):1353–1358. Available from: [http://bioinformatics.oxfordjournals.](http://bioinformatics.oxfordjournals.org/content/28/10/1353.abstract)
699 [org/content/28/10/1353.abstract](http://bioinformatics.oxfordjournals.org/content/28/10/1353.abstract).

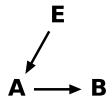
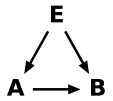
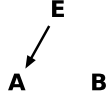
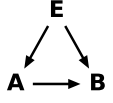
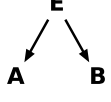
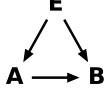
700 42. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of*
701 *the National Academy of Sciences*. 2003;100(16):9440–9445. Available from: [http://www.](http://www.pnas.org/content/100/16/9440.abstract)
702 [pnas.org/content/100/16/9440.abstract](http://www.pnas.org/content/100/16/9440.abstract).

703 43. Chen LS, Sangurdekar DP, Storey JD. trigger: Transcriptional Regulatory Inference from
704 Genetics of Gene Expression; 2007. R package version 1.16.0.

705 44. Stolovitzky G, Prill RJ, Califano A. Lessons from the DREAM2 Challenges. *Annals of the*
706 *New York Academy of Sciences*. 2009;1158(1):159–195.

707 45. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. In: *Pro-*
708 *ceedings of the 23rd International Conference on Machine Learning. ICML '06. New York, NY,*
709 *USA: ACM; 2006. p. 233–240. Available from: [http://doi.acm.org/10.1145/1143844.](http://doi.acm.org/10.1145/1143844.1143874)*
710 [1143874](http://doi.acm.org/10.1145/1143844.1143874).

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.

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

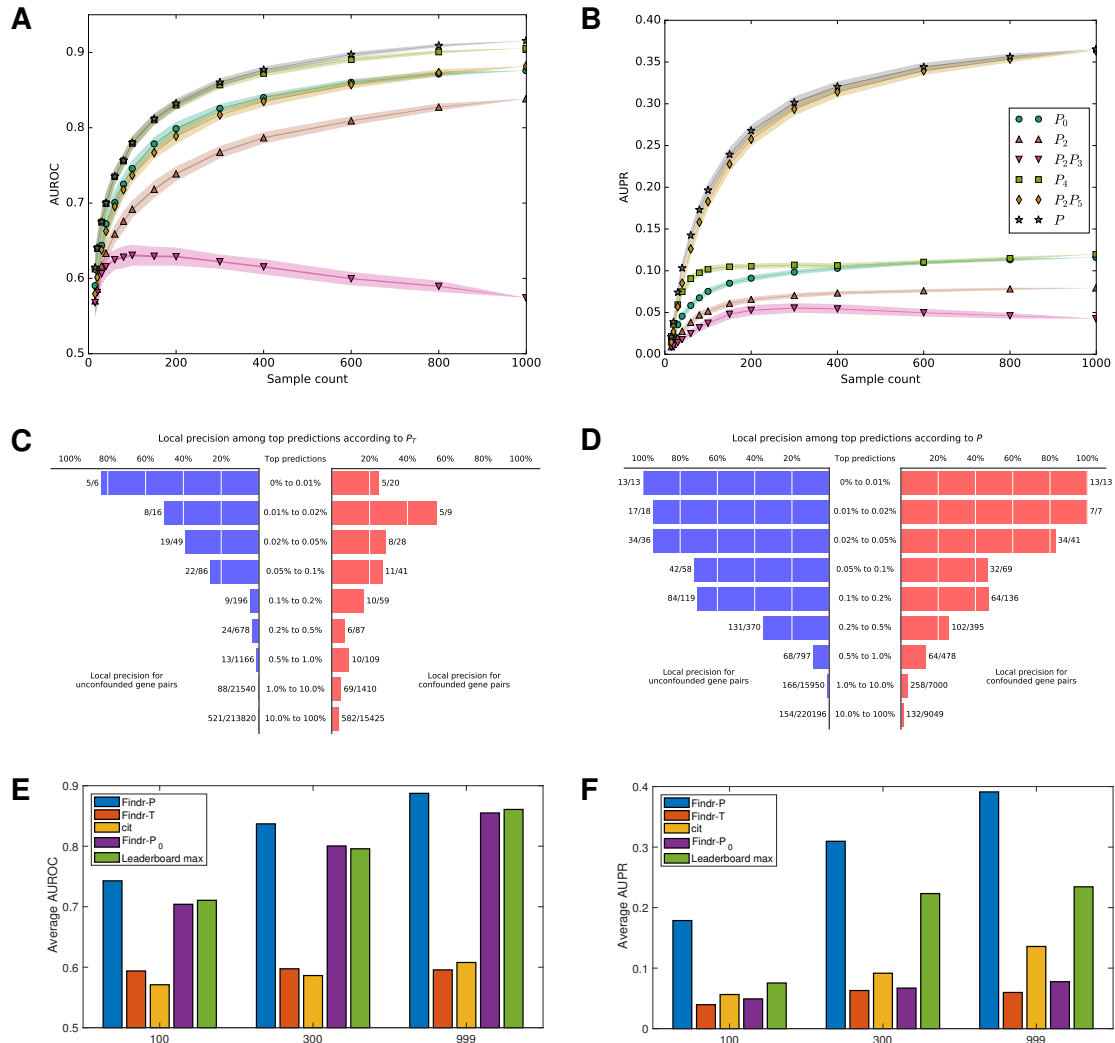


Figure 1: Findr achieves best prediction accuracy on the DREAM5 Systems Genetics Challenge. (A, B) The mean AUROC (A) and AUPR (B) on subsampled data are shown for traditional (P_2 , P_2P_3) and newly proposed (P_4 , P_2P_5 , P) causal inference tests against the baseline correlation test (P_0). Every 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 **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 of top predictions 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_T), 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**.

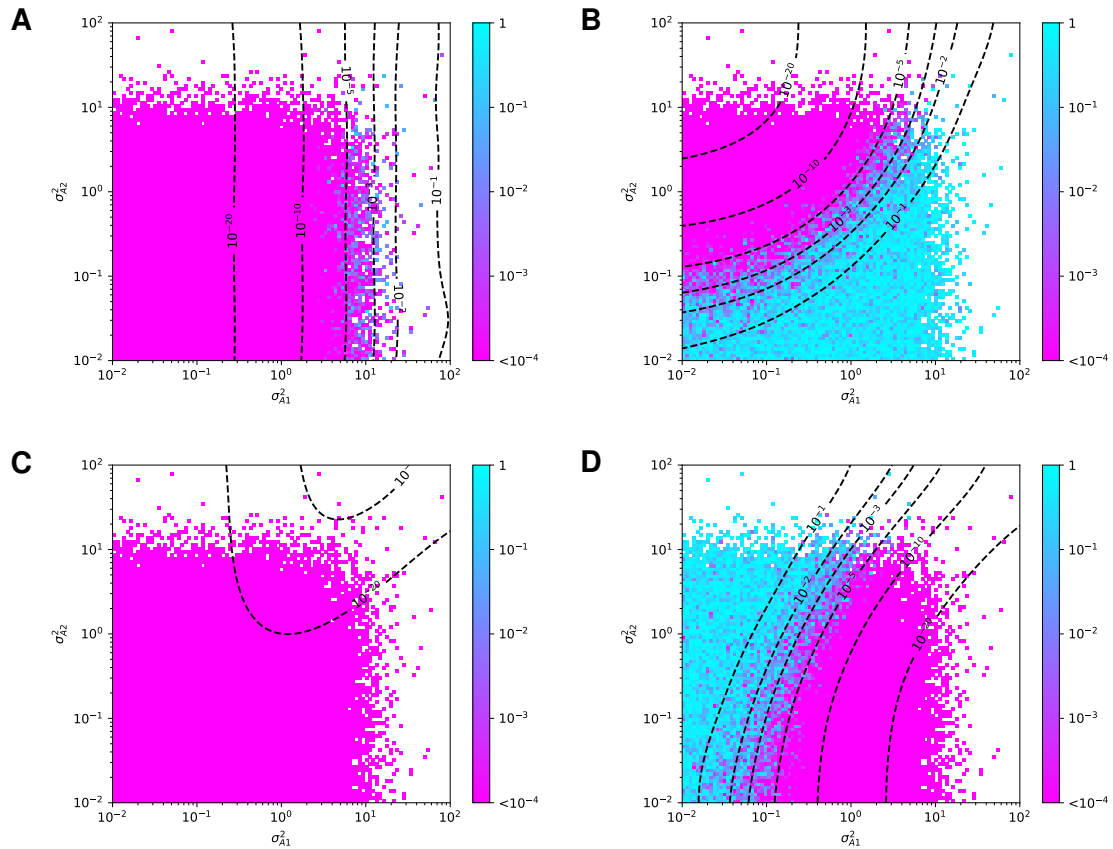


Figure 2: The conditional independence test yields false negatives for unconfounded regulations in the presence of even minor measurement errors. Null hypothesis p-values of the secondary linkage (A), conditional independence (B), relevance (C), and controlled (D) tests are shown on simulated data from 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, σ_{A1}^2 is $A^{(t)}$'s variance from other sources and σ_{A2}^2 is the variance due to measurement noise. A total of 100 values from 10^{-2} to 10^2 were taken for σ_{A1}^2 and σ_{A2}^2 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 smoothed 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.

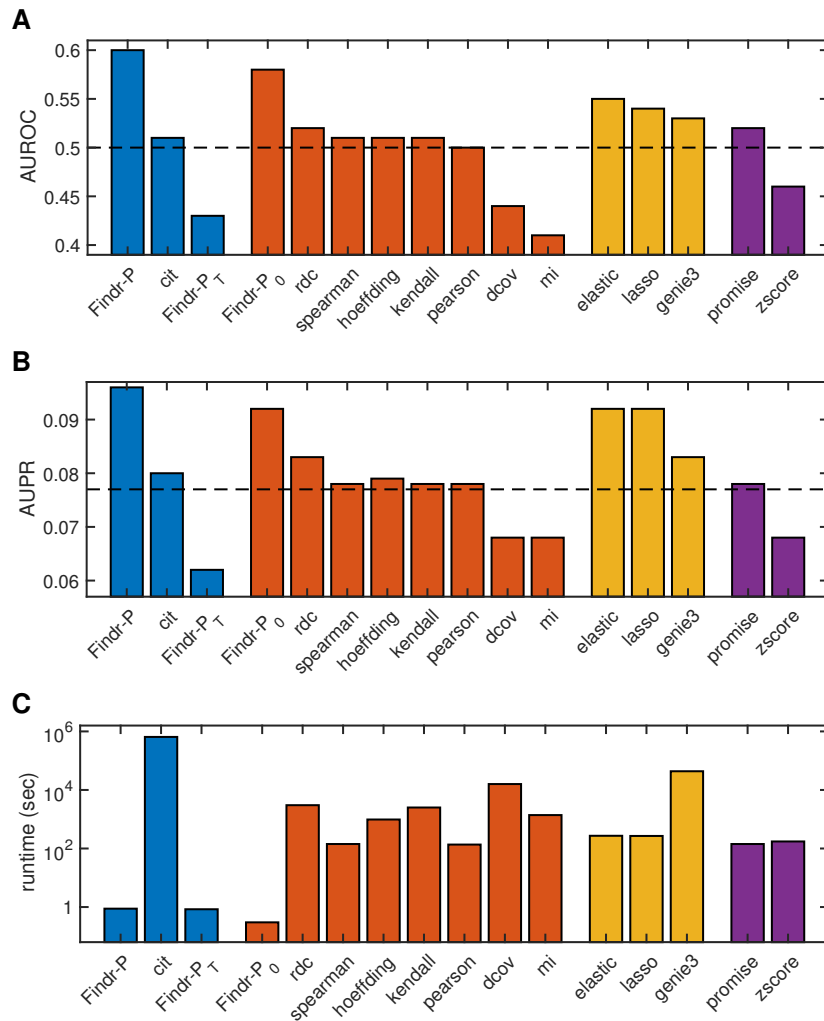


Figure 3: Findr achieves highest accuracy and speed on the prediction of miRNA target genes from the Geuvadis data. Shown are the AUROC (A), AUPR (B) and runtime (C) for 16 miRNA target prediction methods. Methods are colored by type: blue, genotype-assisted causal inference methods; red, pairwise correlation 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**.

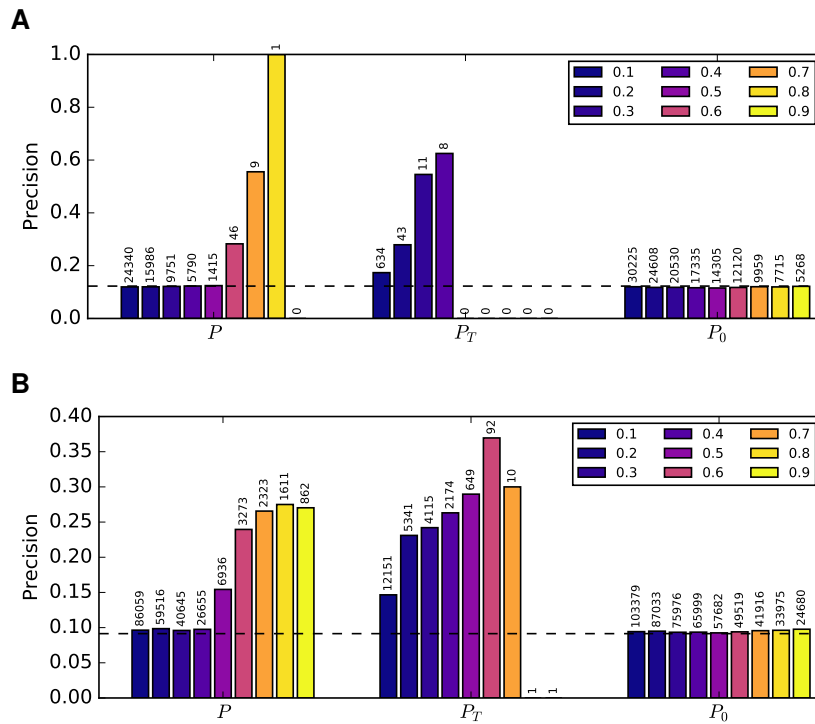


Figure 4: **Findr predicts TF targets with more accurate FDR estimates from the Geuvadis data.** The precision (i.e. 1-FDR) of TF target predictions is shown at probability cutoffs 0.1 to 0.9 (blue to yellow) with respect to known functional targets from siRNA silencing of 6 TFs (**A**) and known TF-binding 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.