Modeling causality for pairs of phenotypes
in system genetics

Elias Chaibub Neto
Aimee T. Broman
Mark P Keller
Alan D Attie
Bin Zhang
Jun Zhu
Brian S Yandell

1 Sage Bionetworks, Seattle, Washington, 98109.
2 Department of Biochemistry, University of Wisconsin - Madison, Madison, Wisconsin, 53706.
3 Department of Statistics, University of Wisconsin - Madison, Madison, Wisconsin, 53706.
4 Department of Horticulture, University of Wisconsin - Madison, Madison, Wisconsin, 53706.
Running title: Causality for Pairs of Phenotypes

Key words:

1. Causality.
3. Hypothesis tests.
4. Systems genetics.
5. Quantitative trait loci.

Corresponding author:

Brian S. Yandell
Department of Statistics
University of Wisconsin - Madison
1300 University Avenue
Madison, Wisconsin 53706
byandell@wisc.edu
Abstract

Current efforts in systems genetics have focused on the development of statistical approaches that aim to disentangle causal relationships among molecular phenotypes in segregating populations. Reverse engineering of transcriptional networks plays a key role in the understanding of gene regulation. However, transcriptional regulation is only one possible mechanism, as methylation, phosphorylation, direct protein-protein interaction, transcription factor binding, etc., can also contribute to gene regulation. These additional modes of regulation can be interpreted as unobserved variables in the transcriptional gene network, and can potentially impact its reconstruction accuracy. We develop tests of causal direction for a pair of phenotypes that may be embedded in a more complicated but unobserved network by extending Vuong’s selection tests for misspecified models. Our tests provide a significance level, which is unavailable for the widely used AIC and BIC criteria. We evaluate the performance of our tests against the AIC, BIC and a recently published causality inference test in simulation studies. We compare the precision of causal calls using biologically validated causal relationships extracted from a database of 247 knockout experiments in yeast. Our model selection tests are more precise, showing greatly reduced false positive rates compared to the alternative approaches. In practice, this is a useful feature since follow up studies tend to be time consuming and expensive and, hence, it is important for the experimentalist to have causal predictions with low false positive rates.

Introduction

A key objective of biomedical research is to unravel the biochemical mechanisms underlying complex disease traits. Integration of genetic information with genomic, proteomic and metabolomic data has been used to infer causal relationships among phenotypes

Two key assumptions for causal inference in systems genetics are genetic variation preceding phenotypic variation and Mendelian randomization of alleles in unlinked loci. These conditions together, which provide temporal order and eliminate confounding of other factors, justify causal claims between QTLs and phenotypes. Causal inference among phenotypes is justified by conditional independence relations under Markov properties (Li et al. 2006, Chaibub Neto et al. 2010).

Given a pair of phenotypes, \( Y_1 \) and \( Y_2 \), that co-map to the same quantitative trait locus, \( Q \), our objective is to learn which of the four distinct models, \( M_1, M_2, M_3 \) and \( M_4 \), depicted in Figure 1, is the best representation for the true relation between \( Y_1 \) and \( Y_2 \). Models \( M_1, M_2, M_3 \) and \( M_4 \) represent, respectively, the causal, reactive, independence and full models as collapsed versions of more complex regulatory networks. For instance, when the data are transcriptional and one gene is upstream of other genes, the regulation of the upstream gene may affect those downstream, even when the regulation takes place via post-transcriptional mechanisms, and, hence, is mediated by unobserved variables. Transcriptional networks should be interpreted as collapsed versions of more
complicated networks, where the presence of an arrow from a QTL to a phenotype or from one phenotype to another simply means that there is a directional influence of one node on another (that is, there is at least one path in the network where the node in the tail of the arrow is upstream of the node in the head). Figure S1 in the Supplement shows a few examples of networks and their collapsed versions. Our goal in this paper is to infer the causal direction between two nodes, and the term “causal” should be interpreted as causal direction, meaning either direct or indirect causal relations.

In this paper, we propose novel causal model selection hypothesis tests, and compare their performance to the AIC and BIC model selection criteria and to a causality inference test (CIT) proposed by Millstein et al (2009). AIC (Akaike 1974) and BIC (Schwarz 1978) are widely used penalized likelihood criteria to perform model selection among models of different sizes. Over-parameterized models tend to over-fit the data and, when comparing models with different dimension, it is necessary to counter-balance model fit and model parsimony by adding a penalty term that depends on the number of parameters. CIT is an intersection-union test, in which a number of equivalence and conditional F tests are conservatively combined in a single test. P-values are computed for models $M_1$ and $M_2$ in Figure 1, but not for the $M_3$ or $M_4$ models, and the decision rule for model calling goes as follows: (1) call $M_1$ if the $M_1$ p-value is less than a significance threshold $\alpha$ and the $M_2$ p-value is greater than $\alpha$; (2) call $M_2$ if it is the other way around; (3) call $M_i$ if both p-values are greater than $\alpha$; and (4) make a “no call” if both p-values are less than $\alpha$. The $M_i$ call actually means that the model is not $M_1$ or $M_2$ and could correspond to an $M_3$ or $M_4$ model. Note that the CIT makes a “no call” when both $M_1$ and $M_2$ models are simultaneously significant.

Our causal model selection tests (CMSTs) adapt and extend Vuong’s (1989) and Clarke’s (2007) tests to the comparison of four models. Vuong’s model selection test
is a formal parametric hypothesis test devised to quantify the uncertainty associated with a model selection criterion, comparing two models based on their (penalized) likelihood scores. It uses the (penalized) log-likelihood ratio scaled by its standard error as a test statistic, and tests the null hypothesis that both models are equally close to the true data generating process. While the (penalized) log-likelihood scores can only determine whether, for example, model A fits the data better than model B, Vuong’s test goes one step further and attaches a p-value to the scaled contrast of (penalized) log-likelihood scores. In this way it can interrogate whether the better fit of model A compared to model B is statistically significant or not. Vuong’s test tends to be conservative and low powered. Clarke (2007) proposed a non-parametric version that achieves an increase in power at the expense of higher miss-calling error rates by using the median rather than the mean of (penalized) log-likelihood ratio.

We propose 3 distinct versions of causal model selection tests: (1) the parametric CMST test, that corresponds to an intersection-union test of six separate Vuong’s tests; (2) the non-parametric CMST test, constructed as an intersection-union tests of six Clarke’s tests; and (3) the joint-parametric CMST test, that mimics an intersection-union test, and is derived from the joint distribution of Vuong’s test statistics. These CMST tests inherit from Vuong’s test the property that none of the models being compared need be correct. That is, the true model may describe a more complicated network, including unobserved factors. Our approach simply selects the wrong model that is closest to the (unknown) true model.

**Methods**

**Vuong’s model selection test**

The Kullback-Leibler (1959) Information Criterion (KLIC) measures the closeness of
a probability model to the true distribution of data. Sawa (1978) showed that the KLIC orders approximate models by comparing the expected value of the log likelihood under the true model. Vuong (1989) used this result to develop an empirical test of two models based on the sample mean of the log likelihood ratio scaled by its sample standard error.

Formally, \( \{f(y \mid x; \theta) : \theta \in \Theta\} \) represents a parametric family of conditional models and

\[
KLIC(h^0; f) = E^0 \left[ \log h^0(y \mid x) \right] - E^0 \left[ \log f(y \mid x; \theta_*) \right] \\
= \int_x \left[ \int_y h^0(y \mid x) \log \frac{h^0(y \mid x)}{f(y \mid x; \theta_*)} \, dy \right] h^0(x) \, dx,
\]

where \( E^0 \) represents the expectation with respect to the true joint distribution \( h^0(y, x) = h^0(y \mid x) h^0(x) \), and \( \theta_* \) is the parameter value that minimizes the KLIC distance from \( f \) to the true model (Sawa 1978). Note that \( f \) need not belong to the same parametric family as \( h^0 \).

A model \( f_1(y \mid x; \theta_{1*}) \), denoted \( f_1 \) for short, is regarded as a better approximation to the true model \( h^0(y \mid x) \), than the alternative model \( f_2(y \mid x; \theta_{2*}) \) if and only if \( KLIC(h^0; f_1) < KLIC(h^0; f_2) \), or alternatively, \( E^0[\log f_1] > E^0[\log f_2] \) (Sawa 1978). Vuong’s model selection test is based on the latter criterion and the null and alternative hypotheses are defined as

\[
H_0 : E^0[LR_{12}] = 0, \quad H_1 : E^0[LR_{12}] > 0, \quad H_2 : E^0[LR_{12}] < 0,
\]

where \( LR_{12} = \log f_1 - \log f_2 \). The null hypothesis is \( f_1 \) and \( f_2 \) are equally close to the true distribution. The alternative hypothesis \( H_1 \) means that \( f_1 \) is better than \( f_2 \) and conversely for the alternative \( H_2 \).

The quantity \( E^0[LR_{12}] \) is unknown, but under fairly general conditions the sample
mean and variance of $L\hat{R}_{12,i} = \log \hat{f}_{1,i} - \log \hat{f}_{2,i}$ converge almost surely to $E^0[LR_{12}]$ and $Var^0[LR_{12}] = \sigma_{12,12}$, where $\hat{f}_{1,i} = f_1(y_i | x_i; \hat{\theta}_1)$ and $\hat{\theta}_1$ is the maximum likelihood estimate of $\theta_1$ (Vuong 1989). Let $L\hat{R}_{12} = \sum_{i=1}^n L\hat{R}_{12,i}$, then, under $H_0$,

$$n^{-1/2}L\hat{R}_{12}/\sqrt{\hat{\sigma}_{12,12}} \to^d N(0,1). \tag{3}$$

Under $H_1$ this test statistic converges almost surely to $\infty$, whereas, under $H_2$, it converges to $-\infty$ (Vuong 1989).

Vuong’s test is based on the unadjusted log likelihood ratio statistic. However, competing models may have different dimensions, requiring a complexity penalty. The penalized log-likelihood ratio is given by $L\hat{R}^*_{12} = L\hat{R}_{12} - D_{12}$, where the penalty $D_{12}$ is the difference of the number of parameters between models 1 and 2 (for AIC), or this value rescaled by $(\log n)/2$ (for BIC). Because the penalty term is of smaller size than $n^{1/2}$, the adjusted log likelihood ratio accounting for different model dimensions

$$Z_{12} = n^{-1/2}L\hat{R}^*_{12}/\sqrt{\hat{\sigma}_{12,12}} \tag{4}$$

has the same asymptotic properties as $n^{-1/2}L\hat{R}_{12}/\sqrt{\hat{\sigma}_{12,12}}$ (Vuong 1989).

The p-value of Vuong’s test is given by $p_{12} = P(Z_{12} \geq z_{12}) = 1 - \Phi(z_{12})$, where $\Phi()$ represents the cumulative density function of a standard normal variable (Vuong 1989). Note that since $Z_{12} = -Z_{21}$; $p_{21} = 1 - \Phi(z_{21}) = \Phi(z_{12})$, so that $p_{12} + p_{21} = 1$. This property of the Vuong’s test ensures that the p-values of the intersection-union tests cannot be simultaneously significant.

Figure S2 in the Supplement illustrates how Vuong’s test trades a reduction in false positives against a reduction in statistical power. In our applications we need to account for both nested and non-nested models. While the presented test corresponds to Vuong’s
test for strictly non-nested models, it is also valid for nested models when we adopt penalized likelihood scores (see the Supplement for further details).

**Clarke’s model selection paired sign test**

The model selection paired sign test (Clarke 2007) is a non-parametric alternative to Vuong’s test, testing the null hypothesis that the median log-likelihood ratio is 0. Clarke’s test statistic, $T_{12}$, is a sign test on $L\hat{R}_{12,i}$. Under the null hypothesis that the median log-likelihood ratio is zero, $T_{12}$ has a binomial distribution, and the p-value for comparing models 1 and 2 is

$$p_{12} = P(T_{12} \geq t_{12}) = \sum_{k=t_{12}}^{n} C_{k}^{n} 2^{-n}, \quad (5)$$

with $C_{k}^{n} = n!/(k!(n-k)!$. The p-values for $T_{12}$ and $T_{21}$ do not add to 1 since the statistics are discrete, $p_{12} + p_{21} = 1 + C_{t_{12}}^{n} 2^{-n}$. Nonetheless, the $C_{t_{12}}^{n} 2^{-n}$ term decreases to 0 as $n$ increases, and, in practice, $p_{12} + p_{21} \approx 1$ even for moderate sample sizes. We adjust this test using the AIC or BIC penalty $D_{12}$,

$$T_{12} = \sum_{i=1}^{n} \mathbb{1} \left\{ L\hat{R}_{12,i} - n^{-1} D_{12} > 0 \right\}, \quad (6)$$

to account for the varying dimensionality of the models.

**Causal Model Selection Tests (CMST)**

The four models $M_1$, $M_2$, $M_3$ and $M_4$ (Figure 1) are used to derive intersection-union tests based on the application of six separate Vuong (or Clarke) tests comparing, namely, $f_1 \times f_2$, $f_1 \times f_3$, $f_1 \times f_4$, $f_2 \times f_3$, $f_2 \times f_4$ and $f_3 \times f_4$. Sun et al. (2007) implicitly used intersection-unions of Vuong’s tests to select among three non-nested models. Here, we present 3 distinct versions of the CMST: (1) parametric, (2) non-parametric and (3)
joint-parametric CMST tests. We implement the tests with penalized log-likelihoods, but state results for log-likelihoods.

Here we focus on model $M_1$ and p-value $p_1$, with analogous results and notation for the other three models. Starting with the parametric version, we test the null $H_0$: model $M_1$ is no closer to the true model than $M_2$, $M_3$ or $M_4$, against the alternative $H_1$: $M_1$ is closer to the true model than $M_2$, $M_3$ and $M_4$. More explicitly, we test,

$$H_0 : \{ E_0[LR_{12}] = 0 \} \cup \{ E_0[LR_{13}] = 0 \} \cup \{ E_0[LR_{14}] = 0 \} ; \quad (7)$$

against

$$H_1 : \{ E_0[LR_{12}] > 0 \} \cap \{ E_0[LR_{13}] > 0 \} \cap \{ E_0[LR_{14}] > 0 \} . \quad (8)$$

The rejection region for this test is $\min\{z_{12}, z_{13}, z_{14}\} > c_{\alpha}$, where $c_{\alpha}$ is the $\alpha$-critical value of the standard normal. The intersection-union p-value is $p_1 = \max\{p_{12}, p_{13}, p_{14}\}$. For any $\alpha$, if $p_1 \leq \alpha$, then $\min\{p_2, p_3, p_4\} \geq 1 - \alpha$. Therefore, the proposed CMST test has at most one significant model p-value at a time, in contrast to the CIT approach.

The non-parametric CMST test corresponds to an intersection-union of Clarke’s tests, exactly analogous to the parametric version. Because in practice $p_{12} + p_{21} \approx 1$ for Clarke’s test, the non-parametric CMST test also does not allow the detection of more than one significant model p-value.

Simple application of separate Vuong tests overlooks the dependency among the test statistics. A multivariate extension, the joint parametric CMST test, can be developed to address this caveat. For model $M_1$, and under the same general regularity conditions of Vuong (1989), the sample covariance of $\hat{LR}_{12,i}$ and $\hat{LR}_{13,i}$, $\hat{\sigma}_{12,13}$, converges almost surely to $Cov^0[LR_{12}, LR_{13}] = \sigma_{12,13}$ (and similarly for all other covariance terms). Therefore, the sample covariance matrix, $\hat{\Sigma}_1$, converges almost surely to $\Sigma_1$. From the multivariate
central limit and Slutsky’s theorems (Shao 2003), if

\[
\begin{pmatrix}
  E^0[LR_{12}] \\
  E^0[LR_{13}] \\
  E^0[LR_{14}]
\end{pmatrix}
= \begin{pmatrix}
  0 \\
  0 \\
  0
\end{pmatrix}
\] (9)

then \( Z_1 = \text{diag}(\hat{\Sigma}_1)^{-\frac{1}{2}} L\hat{R}_1 / \sqrt{n} \rightarrow^d N_3(0, \rho_1) \), where \( L\hat{R}_1 = (L\hat{R}_{12}, L\hat{R}_{13}, L\hat{R}_{14})^T \) and \( \rho_1 = \text{diag}(\Sigma_1)^{-\frac{1}{2}} \Sigma_1 \text{diag}(\Sigma_1)^{-\frac{1}{2}} \) is the correlation matrix

\[
\rho_1 = \begin{pmatrix}
  1 & \rho_{12,13} & \rho_{12,14} \\
  \rho_{12,13} & 1 & \rho_{13,14} \\
  \rho_{12,14} & \rho_{13,14} & 1
\end{pmatrix}.
\] (10)

The condition in 9 is the worst case of the more general null hypothesis that \( M_1 \) is not better than at least one of \( M_2, M_3, M_4 \), or

\[
H_0 : \min \{ E^0[LR_{12}], E^0[LR_{13}], E^0[LR_{14}] \} \leq 0. \] (11)

We test this against the alternative that \( M_1 \) is better than all three, or

\[
H_1 : \min \{ E^0[LR_{12}], E^0[LR_{13}], E^0[LR_{14}] \} > 0., \] (12)

using the statistic \( W_1 = \min \{ Z_1 \} \), with p-value

\[
P(W_1 \geq w_1) = P(\min \{ Z_{12}, Z_{13}, Z_{14} \} \geq w_1) = P(Z_{12} \geq w_1, Z_{13} \geq w_1, Z_{14} \geq w_1). \] (13)

The joint parametric CMST test with \( W_1 \) follows the spirit of an intersection union
test while accounting for dependency among test statistics. Table 1 depicts the joint CMST tests for all models.

The CMST tests are implemented in the \texttt{R/qtlhot} package available at CRAN. Although not explicitly stated in the notation, the pairwise models can easily account for additive and interactive covariates, and our code already implements this feature. When using this package please cite this paper.

**Simulation studies**

We conducted two simulation studies. In the first “pilot study”, we focus on performance comparison of the AIC, BIC, CIT and CMST methods with data generated from simple causal models. The goal is to understand the behavior of our methods in diverse settings. In the second “large scale study”, we perform a simulation experiment, with data generated from causal models emulating QTL hotspot patterns. The goal is to understand the impact of multiple testing on the performance of our causality tests.

The pilot simulation study has data generated from Models A to E in Figure 2. We conducted 10 simulation studies, generating data from the five models described above under sample sizes 112 (the size of our real data example) and 1,000. For each model, we simulated 1,000 backcrosses. We chose simulation parameters to ensure that 99% of the $R^2$ coefficients between phenotypes and QTL ranged between 0.08 and 0.32 for the simulations based on sample size of 112 subjects and between 0.01 to 0.20 for the simulations based on 1,000 subjects (see Supplement for details). We evaluated the method’s performance using statistical power, miss-calling error rate and precision. These quantities were computed as,

$$\text{Power} = \frac{TP}{N}, \quad \text{Miss-calling error} = \frac{FP}{N}, \quad \text{Precision} = \frac{TP}{TP + FP},$$
where \( N \) is the total number of tests, and TP (true positives) and FP (false positives) are defined according to Table 2, which depicts possible calls against simulated models, and tabulates whether a specific call correctly represents the causal relationship between the phenotypes in the model from which the data were generated.

In the large scale simulation study we investigate the empirical FDR (1 minus the precision) and power levels achieved by the CMST tests using the Benjamini and Hochberg (1995) and the Benjamini and Yekutieli (2001) FDR control procedures (denoted, respectively, by BH and BY), as well as no multiple testing correction. We simulate data from the models in Figure 5, which emulate eQTL hotspot patterns, i.e., genomic regions to which hundreds or thousands of traits co-map (West et al. 2007). In each simulation we generated 1,000 distinct backcrosses with phenotypic data on 5,001 traits on 112 individuals. We simulated unequally spaced markers for model \( F \), but equally spaced markers for \( G \), with \( Q_1 \) and \( Q \) set 1cM apart. Because we fit almost three million hypothesis tests in this simulation study, we did not include the CIT tests in this investigation, restricting our attention to the computationally more efficient CMST tests. The details for our choice of simulation parameters and QTL mapping are presented in the Supplement. A frequent goal in eQTL hotspots studies is to determine a master regulator, i.e., a transcript that regulates the transcription of the other traits mapping to the hotspot. A promising strategy towards this end is to test the cis traits (i.e., transcripts physically located close to the QTL hotspot) against all other co-mapping traits. Our simulations evaluate the performance of the CMST tests in this setting.

**Results**

**Pilot simulation study results**

Figure 3 depicts the power, miss-calling error rate and precision of each of the methods
based on the simulation results of all five models in Figure 2. The results of the AIC and BIC approaches are constant across all significance levels since these approaches do not provide a measure of statistical significance. For those methods, we simply fit the models to the data and select the model with the best (smallest) score.

Overall, the AIC, BIC and CIT showed high power, high miss-calling error rates and low precision. The CMST methods, on the other hand, showed lower power, lower miss-calling error rates and higher precision. The non-parametric CMST tended to be more powerful but less precise than the other CMST approaches. As expected, for sample size 1,000, all methods showed an increase in power and precision and decrease in miss-calling error rate.

Figures S5-S9 in the Supplement show the simulation results data for each one of Models A to E, when sample size is 112. Figures S10-S14 show the same results for sample size 1,000. Some of the simulated models were intrinsically more challenging than others. For instance, in the absence of latent variables the causal and independence relations can often be correctly inferred by all methods (see the results for Models A and D in Figures S5, S10, S8, S13). However, the presence of hidden-variables in Models B and E tend to complicate matters. Nonetheless, although the AIC, BIC and CIT methods tend to detect false positives at high rates in these complicated situations, the CMST tests tend to forfeit making calls and tend to detect fewer false positives (see Figures S6, S11, S9, S14). Model C is particularly challenging (Figures S7 and S12), showing the highest false positive detection rates among all models.

In genetical genomics experiments we often restrict our attention to the analysis of \textit{cis}-genes against \textit{trans}-genes. In this special case it is reasonable to expect the pleiotropic causal relationship depicted in Model C to be much less frequent than the relationships shown in Models A, B, D and E, so that the performance statistics shown in Figure 3,
might be negatively impacted to an unnecessary degree by the simulation results from Model C.

In order to investigate the performance of methods in the cis against trans case, we present in Figure 4 the simulation results based on Models A, B, D and E only. Comparison of Figures 3 and 4 show an overall improvement in power, decrease in miss-calling rates and increase in precision.

In the analysis of trans against trans genes there is no a priori reason to discard the relationship depicted in Model C, and more false positives should be expected. The CMST approaches, specially the joint parametric and parametric CMST methods, tend to detect a much smaller number of false positives than the AIC, BIC and CIT approaches, as shown in Figures S7 and S12.

**Large scale simulation study results**

With the possible exception of the non-parametric version, the previous simulation study suggests that the CMST tests can be quite conservative. Therefore, it is reasonable to ask whether multiple testing correction is really necessary in order to achieve reasonable false discovery rates (FDR).

Figure 6 presents the observed FDR and power using uncorrected, BH corrected and BY corrected p-values for the simulations based on model G. The top left panel shows that, except for the AIC-based non-parametric CMST, the observed FDRs were considerably lower than the p-value cutoff, suggesting that multiple testing adjustment is not necessary for the CMST tests. Furthermore, comparison of the bottom panels shows that the BH and BY adjustments leads to a reduction in power (specially for the BY adjustment) for the joint and parametric tests at the expense of small drop in FDR levels (that were already low without any correction). For the non-parametric tests, on the other
hand, BH corrections leads to bigger drops in FDR (specially for the AIC based test), and smaller drops in power. The BY correction appears too conservative even for the non-parametric tests. The results for model $F$ are similar (Figure S15).

**Yeast data analysis and biologically validated predictions**

We analyzed a budding yeast genetical genomics data-set derived from a cross of a standard laboratory strain, and a wild isolate from a California vineyard (Brem and Kruglyak 2005). The data consists of expression measurements on 5,740 transcripts measured on 112 segregant strains with dense genotype data on 2,956 markers. Processing of the expression measurements raw data was done as described in Brem and Kruglyak (2005), with an additional step of converting the processed measurements to normal scores. We performed QTL analysis using Haley-Knott regression (Haley and Knott 1992) with the R/qtl software (Broman et al. 2003). We used Haldane’s map function, genotype error rate of 0.0001, and set the maximum distance between positions at which genotype probabilities were calculated to 2cM. We adopted a permutation LOD threshold (Churchill and Doerge 1994) of 3.48, controlling the genome wide error rate of falsely detecting a QTL at a significance level of 5%.

In order to evaluate the precision of the causal predictions made by the methods we used validated causal relationships extracted from a data-base of 247 knock-out experiments in yeast (Hughes et al. 2000, Zhu et al. 2008). In each of these experiments, one gene was knocked-out, and the expression levels of the remainder genes in control and knocked-out strains were interrogated for differential expression. The set of differentially expressed genes form the knock-out signature (ko-signature) of the knocked-out gene (ko-gene), and show direct evidence of a causal effect of the ko-gene on the ko-signature genes. The yeast data cross and knocked-out data analyzed in this section is available in
the \texttt{R/qtlyeast} package at GITHUB (https://github.com/byandell/qtlyeast).

To use this information, we: (i) determined which of the 247 ko-genes also showed a significant eQTL in our data-set; (ii) for each one of the ko-genes showing significant linkages, we determined which other genes in our data-set also co-mapped to the same QTL of the ko-gene, generating, in this way, a list of putative targets of the ko-gene; (iii) for each of the ko-gene/putative targets list, we applied all methods using the ko-gene as the $Y_1$ phenotype, the putative target genes as the $Y_2$ phenotypes and the ko-gene QTL as the causal anchor; (iv) for the AIC- and BIC-based non-parametric CMST tests we adjusted the p-values according to the Benjamini and Hochberg FDR control procedure; and (v) for each method we determined the “validated precision”, computed as the ratio of true positives by the sum of true and false positives, where a true positive is defined as an inferred causal relationship where the target gene belongs to the ko-signature of the ko-gene, and a false positive is given by an inferred causal relation where the target gene does not belong to the ko-signature.

In total 135 of the ko-genes showed a significant QTL, generating 135 putative target lists. A gene belonged to the putative target list of a ko-gene when its 1.5 LOD support interval (Lander and Botstein 1989; Dupuis and Siegmund 1999; Manichaikul et al. 2006) contained the location of the ko-gene QTL. The number of genes in each of the putative target lists varied from list to list, but in total we tested 31,975 “ko-gene/putative target gene” relationships.

Figure 7 presents the number of inferred true positives, number of inferred false positives and the prediction precision across varying target significance levels for each one of the methods. The CIT, BIC and AIC had a higher number of true positives than the CMST approaches, with the AIC-based CMST methods having less power than the BIC-based CMST methods. However, the CIT, BIC and AIC also inferred the highest
numbers of false positives (panel 7b), and showed low prediction precisions (panel 7c). From panel 7c we see that the CMST tests show substantially higher precision rates across all target significance levels compared to the AIC, BIC and CIT methods. Among the CMST approaches, the joint parametric CMST tended to show the highest precision, followed by the non-parametric and parametric CMST tests.

The results presented on Figure 7 were computed using all 135 ko-genes. However, in light of our simulation results, that suggest that the analysis of \textit{cis} against \textit{trans} genes is usually easier than the analysis of \textit{trans} against \textit{trans} genes, we investigated the results restricting ourselves to ko-genes with significant \textit{cis} QTLs. Only 28 out of the 135 ko-genes were \textit{cis} traits, but, nonetheless, were responsible for 2,947 out the total 31,975 “ko-gene/putative target gene” relationships. Figure 8 presents the results restricted to the \textit{cis} ko-genes. All methods show improvement in precision, corroborating our simulation results. Once again, the CMST tests showed higher precision than the CIT, AIC and BIC.

**Discussion**

In this paper, we proposed three novel hypothesis tests that adapt and extend Vuong’s and Clarke’s model selection tests, to the comparison of four models, spanning the full range of possible causal relationships among a pair of phenotypes. Our CMST tests scale well to large genome wide analyzes because they are fully analytical and avoid computationally expensive permutation or re-sampling strategies.

Another useful property of the CMST tests, inherited from Vuong’s test, is their ability to perform model selection among misspecified models. That is, the correct model need not be one of the models under consideration. Accounting for the misspecification of the models is key. In general, any two phenotypes of interest are embedded in a complex
network and are affected by many other phenotypes not considered in the grossly simplified (and thus misspecified) pairwise models.

Overall, our simulations and real data analysis show that the CMST tests are better at controlling miss-calling error rates and tend to outperform the AIC, BIC and CIT methods in terms of statistical precision. However, they do so at the expense of a decrease in statistical power. While an ideal method would have high precision and power, in practice there is always a trade-off between these quantities. Whether a more powerful and less precise, or a less powerful and more precise method is more adequate, depends on the biologist’s research goals and resources. For instance, if the goal is to generate a rank-ordered list of promising candidates genes that might causally affect a phenotype of interest, and the biologist can easily validate several genes, a larger list generated by more powered and less precise methods might be more appealing. However, in general, follow up studies tend to be time consuming and expensive, and only a few candidates can be studied in detail. A long list of putative causal traits is not useful if most are false positives. High power to detect causal relations alone is not enough. A more precise method that conservatively identifies candidates with high confidence can be more appealing (see also Chen et al. 2006).

Further, the exploratory goal is often to identify causal agents without attempting to reconstruct entire pathways. Therefore, much information about the larger networks in which the tested pairs of traits reside is unknown and generally unknowable, and contributes to the large unexplained variation that in turn results in low power. Our method accurately reflects this difficulty to detect causal relationships in the presence of noisy high throughput data and poorly understood networks.

Interestingly, our data analysis and simulations also suggest that the analysis of cis against trans gene pairs is less prone to detect false positives than the analysis of trans
against trans gene pairs. Our simulations suggest that model selection approaches have difficulty ordering the phenotypes when the QTL effect reaches the truly reactive gene by two or more distinct paths, only one of which is mediated by the truly causal gene (see Figure S1c in the Supplement, for an example).

When we test causal relationships among gene expression phenotypes, the true relationships might not be a direct result of transcriptional regulation. For instance, the true causal regulation might be due to methylation, phosphorylation, direct protein-protein interaction, transcription factor binding, etc. Margolin and Califano (2007) have pointed out the limitations of causal inference at the transcriptional level, where molecular phenotypes at other layers of regulation might represent latent variables. Model $M_4$ (see Figure 1) can account for these latent variables and can test this scenario explicitly.

Furthermore, as pointed out by Li et al. (2010), causal inference depends on the detection of subtle patterns in the correlation between traits. Hence, it can be challenging even when the true causal relations take place at the transcriptional level. The authors point out that reliable causal inference in genome-wide linkage and association studies require large sample sizes and would benefit from: (i) incorporating prior information via Bayesian reasoning; (ii) adjusting for experimental factors, such as sex and age, that might induce correlations not explained the the causal relations; and (iii) considering a richer set of models than the four models accounted in this paper.

The CMST tests represent a step in the direction of reliable causal inference in three accounts. First, they tend to be precise, declining to make calls in situations where alternative approaches usually deliver a flood of false positive calls. Second, the CMST tests can adjust for experimental factors by modeling them as additive and interactive covariates. Third, the CMST tests can be applied to non-nested models of different dimensions, and can be readily extended to incorporate a larger number of models by
implementing intersection-union tests on a larger number of Vuong’s tests. For the joint-parametric test a higher dimensional null distribution is required.

FDR control for the CMST approaches is a challenging problem as our tests violate the key assumption, made by FDR control procedures, that the distribution of the p-values under the null hypothesis are uniformly distributed (Benjamini and Hochberg 1995, Storey and Tibshirani 2003). Recall that the CMST p-values are computed as the maximum across other p-values, and the maximum of multiple uniform random variables no longer follows a uniform distribution. Additionally, the CMST tests are usually not independent since we often test the same cis-trait against several trans-traits, so that the additional assumption of independent test statistics made by the original Benjamini-Hochberg procedure does not hold. The Benjamini-Yekutieli (BY) procedure, relaxes the independent test statistics assumption, and we explore both these corrections in our simulations. Our results suggest that BH and BY multiple testing correction should not be performed for the joint and the parametric CMST tests, as the FDR levels are lower than the nominal level without any correction and are too conservative with severe reduction in statistical power with the application of BH and BY control. The non-parametric CMST tests, on the other hand, seemed to benefit from BH correction, showing slight decrease in power with concomitant decrease in FDR, in spite of the non-parametric CMST tests being based on discrete test statistics and the BH procedure being developed to handle p-values from continuous statistics. Inspection of the p-value distributions (see Figures S18, S19, S20, and S21 on the Supplement) suggests that the smaller p-values of the non-parametric tests, relative to the other approaches, are the reason for the higher power achieved by the BH corrected non-parametric tests. The BY procedure, on the other hand, tended to be too conservative even for the non-parametric CMST tests.

The CMST approach is currently implemented for inbred line crosses. Extension to
outbred populations involving mixed effects models is yet to be done. Although in this paper we focused on mRNA expression traits, the CMST tests can be applied to any sort of heritable phenotype, including clinical phenotypes and other “omic” molecular phenotypes.

The higher statistical precision and computational efficiency achieved by our fully analytical hypothesis tests will help biologists to perform large scale screening of causal relations, providing a conservative rank-ordered list of promising candidate genes for further investigations.

Acknowledgments

This work was supported by CNPq Brazil (ECN); NCI ICBP grant U54-CA149237 and NIH grant R01MH090948 (ECN); NIDDK grants DK66369, DK58037 and DK06639 (ADA, MPK, ATB, BSY, ECN); NIGMS grants PA02110 and GM069430-01A2 (BSY). We thank Adam Margolin for helpful discussions and comments on this work, and the editor and referees for comments and suggestions that considerably improved this work.

References


Figure 1: Pairwise causal models. $Y_1$ and $Y_2$ represent phenotypes that co-map to a same QTL, $Q$. Models $M_1$, $M_2$, $M_3$ and $M_4$ represent, respectively, the causal, reactive, independent and full model. In model $M_1$ the phenotype $Y_1$ has a causal effect on $Y_2$. In $M_2$, the phenotype $Y_1$ is actually reacting to a causal effect of $Y_2$, hence the name reactive model. In the independence model, $M_3$, there is no causal relationship between $Y_1$ and $Y_2$ and their correlation is solely due to $Q$. The full model, $M_4$, corresponds to three distribution equivalent models $M_{a4}^b$, $M_{b4}^b$, and $M_{c4}^b$ which cannot be distinguished as their maximized likelihood scores are identical. Model $M_{a4}^b$ represents a causal independence relationship where the correlation between $Y_1$ and $Y_2$ is a consequence of latent causal phenotypes, common causal QTLs or of common environmental effects. Models $M_{a4}^a$ and $M_{f4}^f$ correspond to causal-pleiotropic and reactive-pleiotropic relationships, respectively.
Figure 2: Models used in the simulation study. $Y_1$ and $Y_2$ represent phenotypes that co-map to a same QTL, $Q$. Model A represents a causal effect of $Y_1$ on $Y_2$. Model B represents the same, with the additional complication that part of the correlation between $Y_1$ and $Y_2$ is due to a hidden-variable $H$. Model C represents a causal-pleiotropic model, where $Q$ affects both $Y_1$ and $Y_2$ but $Y_1$ also has a causal effect on $Y_2$. Model D shows a purely pleiotropic model, where both $Y_1$ and $Y_2$ are under the control of the same QTL, but one does not causally affect the other. Model E represents the pleiotropic model, where the correlation between $Y_1$ and $Y_2$ is partially explained by a hidden-variable $H$. 
Figure 3: Power (panel a and d), miss-calling error rate (panel b and e) and precision (panel c and f) across the simulated models depicted in Figure 2. The x-axis represents the significance levels used for computing the results. Panels a-c represent the simulations based on sample size 112, whereas panels d-f present the results for sample size 1,000. Dashed and full curves represent, respectively, AIC- and BIC-based methods. Green: parametric CMST. Red: non-parametric CMST. Blue: joint-parametric CMST. Black: AIC and BIC. Orange: CIT. The grey line on panels b and e corresponds to the $\alpha$ levels.
Figure 4: Power (panel a and d), miss-calling error rate (panel b and e) and precision (panel c and f) restricted to the cis versus trans cases. The x-axis represents the significance levels used for computing the results. The results were computed using only the simulated models A, B, D and E in Figure 2, since the pleiotropic causal relationship depicted in Model C is expected to be much less frequent than the others when testing cis versus trans case. Panels a-c represent the simulations based on sample size 112, whereas panels d-f present the results for sample size 1,000. Dashed and full curves represent, respectively, AIC- and BIC-based methods. Green: parametric CMST. Red: non-parametric CMST. Blue: joint-parametric CMST. Black: AIC and BIC. Orange: CIT. The grey line on panels b and e corresponds to the $\alpha$ levels.
Figure 5: Models generating hotspot patterns. \( Y_1 \) represents a cis expression trait. \( Y_k \), \( k = 2, \ldots, 5001 \) represent expression traits mapping in trans to the hotspot QTL \( Q \). \( H \) represents an unobserved expression trait. Model \( F \) generates a hotspot pattern derived from the causal effect of the master regulator, \( Y_1 \), on the transcription of the other traits. Model \( G \) gives rise to a hotspot pattern, due to the causal effect of \( H \) on \( Y_k \), but where the cis-trait \( Y_1 \) maps to \( Q_1 \), a QTL closely linked to the true QTL hotspot \( Q \), and is actually causally independent of the traits mapping in trans to the \( Q \).
Figure 6: Observed FDR and power for the simulations based on model $G$. The x-axis represents the p-value cutoffs used for computing the results. Dashed and full curves represent, respectively, AIC- and BIC-based methods. Green: parametric CMST. Red: non-parametric CMST. Blue: joint-parametric CMST. Black: AIC and BIC. The grey line in the top panels corresponds to the $\alpha$ levels.
Figure 7: Overall number of true positives, number of false positives and precision across all 135 ko-gene/putative target lists. The x-axis represents the significance levels used for computing the results. Dashed and full curves represent, respectively, AIC- and BIC-based methods. Green: parametric CMST. Red: non-parametric CMST. Blue: joint-parametric CMST. Black: AIC and BIC. Orange: CIT.
Figure 8: Overall number of true positives, number of false positives and precision restricted to 28 cis ko-gene/putative target lists. The x-axis represents the significance levels used for computing the results. Dashed and full curves represent, respectively, AIC- and BIC-based methods. Green: parametric CMST. Red: non-parametric CMST. Blue: joint-parametric CMST. Black: AIC and BIC. Orange: CIT.
Table 1: Model selection tests for models $M_1$, $M_2$, $M_3$ and $M_4$. Here $w_k = \min\{z_k\}$ for $k = 1, \ldots, 4$, and $\boldsymbol{\rho}_k$ is defined in analogy with $\boldsymbol{\rho}_1$ in equation 10.

<table>
<thead>
<tr>
<th>$H_0$</th>
<th>Null distribution</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0^{M_1}$</td>
<td>$Z_1 = (Z_{12}, Z_{13}, Z_{14})^T \sim N_3(\mathbf{0}, \hat{\boldsymbol{\rho}}_1)$</td>
<td>$p_1 = P(Z_{12} \geq w_1, Z_{13} \geq w_1, Z_{14} \geq w_1)$</td>
</tr>
<tr>
<td>$H_0^{M_2}$</td>
<td>$Z_2 = (Z_{21}, Z_{23}, Z_{24})^T \sim N_3(\mathbf{0}, \hat{\boldsymbol{\rho}}_2)$</td>
<td>$p_2 = P(Z_{21} \geq w_2, Z_{23} \geq w_2, Z_{24} \geq w_2)$</td>
</tr>
<tr>
<td>$H_0^{M_3}$</td>
<td>$Z_3 = (Z_{31}, Z_{32}, Z_{34})^T \sim N_3(\mathbf{0}, \hat{\boldsymbol{\rho}}_3)$</td>
<td>$p_3 = P(Z_{31} \geq w_3, Z_{32} \geq w_3, Z_{34} \geq w_3)$</td>
</tr>
<tr>
<td>$H_0^{M_4}$</td>
<td>$Z_4 = (Z_{41}, Z_{42}, Z_{43})^T \sim N_3(\mathbf{0}, \hat{\boldsymbol{\rho}}_4)$</td>
<td>$p_4 = P(Z_{41} \geq w_4, Z_{42} \geq w_4, Z_{43} \geq w_4)$</td>
</tr>
</tbody>
</table>
Table 2: True and false positives table. Each entry $i, j$ represents whether the call on row $i$ is a true positive (TP) or as false positive (FP), when the data are generated from the model on column $j$. For instance, when data are generated from Models A or B, a $M_1$ call represents a true positive, whereas a $M_2$, $M_3$ or $M_4$ call represents a false positive for the AIC, BIC and CMSTs approaches (for the CIT a $M_2$ or $M_i$ call represents false positive). Note that a $M_4$ call is considered a true positive for Model C (in addition to Model E) because it corresponds to Model $M_4^a$ on Figure 1 and, hence, is distribution equivalent to Model $M_4$. Please note too that because the CIT only provides p-values for the $M_1$ and $M_2$ calls, but not for the $M_3$ and $M_4$ calls, and its output is either $M_1$, $M_2$ or $M_i$, we classify a $M_i$ call as a true positive for Models C, D and E. Observe that by doing so we are actually giving an unfair advantage for the CIT approach, since when the data are generated from, say, Model E, the CIT only needs to discard models $M_1$ and $M_2$ as non-significant in order to detect a “true positive”. The AIC, BIC and CMST approaches, on the other hand, need to discard models $M_1$, $M_2$ and $M_3$ as non-significant and accept model $M_4$ as significant.