Causal Model Selection Hypothesis Tests in Systems Genetics

Elias Chaibub Neto[1]

Aimee T. Broman[2]

Mark P Keller[2]

Alan D Attie[2]

Bin Zhang[1]

Jun Zhu[1]

Brian S Yandell[3,4]

[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: Causal Model Selection Tests

Key words:

1. Causality.
2. Model selection.
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. Model selection criterions, such as the AIC and BIC, have been widely used for this purpose, in spite of being unable to quantify the uncertainty associated with the model selection call. In this paper we propose three novel hypothesis tests to perform model selection among models representing distinct causal relationships. We focus on models composed of pairs of phenotypes and use their common QTL to determine which phenotype has a causal effect on the other, or whether the phenotypes are not causally related, and are only statistically associated. Our hypothesis tests are fully analytical and avoid the use of computationally expensive permutation or re-sampling strategies. They adapt and extend Vuong's model selection test to the comparison of four possibly misspecified models, handling the full range of possible causal relationships among a pair of phenotypes. We evaluate the performance of our tests against the AIC, BIC and a recently published causality inference test in simulation studies. Furthermore, we compare the precision of the causal predictions made by the methods using biologically validated causal relationships extracted from a database of 247 knockout experiments in yeast. Overall, our model selection tests tend to be conservative but also more precise than alternative approaches. In practice, this is a useful feature since most biologists can only investigate a few genes from a rank-ordered list of candidates, and shorter and more accurate lists are often desired.

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 (Schadt et al. 2005; Li et al. 2006; Kulp and Jagalur 2006; Chen et al. 2006; Zhu et al. 2004, 2007, 2008; Aten et al. 2008; Liu et al. 2008; Chaibub Neto et al. 2008, 2009; Winrow et al. 2009; Millstein et al. 2009). Current approaches for causal inference in systems genetics can be classified into whole network scoring methods (Li et al. 2006; Zhu et al. 2004, 2007, 2008; Liu et al. 2008; Chaibub Neto et al. 2008, 2010; Winrow et al. 2009; Hageman et al. 2010) or pairwise methods, which focus on the inference of causal relationships among pairs of phenotypes (Schadt et al. 2005; Li et al. 2006; Kulp and Jagalur 2006; Chen et al. 2006; Aten et al. 2008; Millstein et al. 2009; Li et al. 2010; Duarte and Zeng 2011). In this paper we develop a pairwise approach for causal inference among pairs of phenotypes.

Given a pair of phenotypes, Y_1 and Y_2, that co-map to a 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. Note that the models in Figure 1 can represent collapsed versions of more complex networks. A directed edge from the QTL to a phenotype or from one phenotype to the other simply means that there exists at least one path in the network where the node in the tail of the arrow is upstream to the node in the head. Hence, the term "causal" should be interpreted as either direct or indirect causal relations. Figure S1 in the Supplement shows a few examples of networks and their collapsed versions.

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). 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 criterium, 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 test the null hypothesis that both models are equally close to the true data generating process according to the Kullback-Leibler distance (Kullback 1958). 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.

One drawback of Vuong's test is that it tends to be conservative and low powered (Clarke 2007). In order to circumvent this problem Clarke (2007) proposed a non-parametric version of Vuong's test that achieves an increase in power at the expense of higher miss-calling error rates. While Voung's null hypothesis tests whether the average (penalized) log-likelihood ratio of two models is zero, Clarke's null hypothesis tests whether the median (penalized) log-likelihood ratio is zero.

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. An interesting property of the CMST tests, inherited from Vuong's test, is their ability to perform model selection among misspecified models. That is, the true data generating process need not be one of the models under consideration.

As the simulations and real data analysis presented in the next sections show, the CMST tests tend to be more precise, but also more conservative than alternative approaches. Rather than a weakness, we see this property as a desired feature of the CMST tests. Most biologists are interested in identifying a rank-ordered list of candidates for further study with a low false positive rate. In many situations, only a few candidates can be actively investigated in detail. A long list of putative causal traits is not useful if most prove to be false positives; high power to detect causal relations alone is not enough. A low-powered method that conservatively identifies candidates with high confidence and few errors can be more appealing (a similar point is made by 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.

As with most methods for causal inference in the context of segregating populations, our approach relies on the fact that, in general, genetic variation precedes phenotypic variation, and on the fact that Mendelian randomization of alleles in unlinked loci provides a mechanism to eliminate the effects of confounding. Both conditions need to be met in order to justify causal claims between QTLs and phenotypes. Causal inference among phenotypes, on the other hand, is justified by conditional independence relations under Markov properties (Li et al. 2006, Chaibub Neto et al. 2010).

Methods

In this Section we present our parametric, non-parametric and joint parametric causal

model selection tests. As a pre-requisite to understand our CMST tests, we briefly present Vuong's and Clarke's model selection tests, before we derive our own. Because the CMST test statistics are based on AIC and BIC scores and we compare our tests to the CIT approach in the simulations a real data analysis, we first briefly describe the AIC, BIC and CIT approaches in the next subsection.

AIC, BIC and CIT: a brief review

The AIC (Akaike 1974) and BIC (Schwarz 1978) are widely used penalized likelihood criteria to perform model selection among models with different number of parameters. 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 proportional to the number of parameters. The AIC penalty is proportional to the number of parameters, whereas the BIC penalty accounts for the sample size and number of parameters.

The CIT (Millstein et al., 2009) corresponds to 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 α and the M_2 p-value is greater than α; (2) call M_2 call if it is the other way around; (3) call M_i if both p-values are greater than α; and (4) make a "no call" if both p-values are less than α. 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. As we will see in the next subsections the CMST do not suffer from this incoherent behavior.

Vuong's model selection test

Vuong's test (Vuong 1989) derives from the Kullback-Leibler (1959) Information Criterion (KLIC) that measures the closeness of a probability model to the true distribution generating the data. Formally, let $\{f(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ represent a parametric family of conditional models. Then

$$\begin{aligned} KLIC(h^0; f) &= E^0\left[\log h^0(\mathbf{y} \mid \mathbf{x})\right] - E^0\left[\log f(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_*)\right] \\ &= \int_{\mathbf{x}} \left[\int_{\mathbf{y}} h^0(\mathbf{y} \mid \mathbf{x}) \, \log \frac{h^0(\mathbf{y} \mid \mathbf{x})}{f(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_*)} \, d\mathbf{y}\right] h^0(\mathbf{x}) \, d\mathbf{x}, \end{aligned} \tag{1}$$

where E^0 represents the expectation with respect to the true joint distribution $h^0(\mathbf{y}, \mathbf{x}) = h^0(\mathbf{y} \mid \mathbf{x})h^0(\mathbf{x})$, and $\boldsymbol{\theta}_*$ 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 $f_1(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{1*})$, denoted f_1 for short, is regarded as a better approximation to the true model $h^0(\mathbf{y} \mid \mathbf{x})$, than the alternative model $f_2(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{2*})$ 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 later criterion and the null and alternative hypotheses are defined as

$$H_0: \quad E^0\left[LR_{12}\right] = 0, \qquad H_1: \quad E^0\left[LR_{12}\right] > 0, \qquad H_2: \quad E^0\left[LR_{12}\right] < 0, \tag{2}$$

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 Vuong (1989) showed that 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 $E^0\left[LR_{12}\right]$ and $Var^0\left[LR_{12}\right] = \sigma_{12.12}$, where $\hat{f}_{1,i} = f_1(\mathbf{y}_i \mid \mathbf{x}_i\,; \hat{\boldsymbol{\theta}}_1)$ and $\hat{\boldsymbol{\theta}}_1$ is the maximum likelihood estimate of $\boldsymbol{\theta}_1$. Let $L\hat{R}_{12} = \sum_{i=1}^{n} L\hat{R}_{12,i}$, then, under H_0,

$$n^{-1/2} L\hat{R}_{12}/\sqrt{\hat{\sigma}_{12.12}} \quad \rightarrow^d \quad N(0,1)\ . \tag{3}$$

Under H_1 this test statistic converges almost surely to ∞, whereas, under H_2, it converges to $-\infty$.

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 $D_{12} = k_{f_1} - k_{f_2}$ or $D_{12} = (k_{f_1} - k_{f_2})(\log n)/2$ for the AIC and BIC penalties, respectively, and k_{f_1} and k_{f_2} represent the number of parameters of f_1 and f_2, respectively. Because the penalty term divided by $n^{1/2}$ converges to zero, $n^{-1/2} L\hat{R}^*_{12}/\sqrt{\hat{\sigma}_{12.12}}$ has the same asymptotic properties as $n^{-1/2} L\hat{R}_{12}/\sqrt{\hat{\sigma}_{12.12}}$ and we can use the adjusted log likelihood ratio for the model selection test (Vuong 1989). Because in our applications we consider models with different dimensions, we adopt

$$Z_{12} = n^{-1/2} L\hat{R}^*_{12}/\sqrt{\hat{\sigma}_{12.12}} \tag{4}$$

as a test statistic in this paper.

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. Note that since $Z_{12} = -Z_{21}$, we have that $p_{21} = 1 - \Phi(z_{21}) = \Phi(z_{12})$, so that $p_{12} + p_{21} = 1$. As we will see later, this interesting property of the Vuong's test ensures that the p-values of the intersection-union tests that we develop next, cannot be simultaneously significant.

Figure S2 in the Supplement illustrates how Vuong's test trades a decrease in detection of false positives by a reduction in statistical power to detect true positives. Finally, on

a technical note, we point out that the hypothesis test presented in this sub-section corresponds to Vuong's test for strictly non-nested models. Nonetheless, as it is clear from Figure 1 some of our model comparisons involve nested models (for instance, model M_1 in Figure 1 is nested on M_4^a). In the Supplement we justify why we can still use the above test for the comparison of nested models when our test statistics are based on penalized log-likelihood scores.

Clarke's model selection paired sign test

The model selection paired sign test, proposed by Clarke (2007), represents a non-parametric alternative to Vuong's test. Instead of testing the null hypothesis that the mean log-likelihood ratio is 0, it tests the null hypothesis that the median of the individual log-likelihood ratios is equal to zero.

The test statistic adopted by Clarke's test, T_{12}, is a version of the 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

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

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.

Since our models can have different dimensions, we actually adopt

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

as a test statistic for the sign test, where D_{12} represents the AIC or BIC penalty.

Causal Model Selection Tests (CMST)

In our applications we consider four models M_1, M_2, M_3 and M_4. In this section we 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 in a different context. Here, we present 3 distinct versions of the CMST: (1) the parametric; (2) the non-parametric; and (3) the joint-parametric CMST tests. The first two versions overlook the dependency among the test statistics, although we revisit this point in the derivation of the multivariate version. As above, we implement the tests with penalized log-likelihoods, although the results are stated in terms of log-likelihoods for the sake of lighter notation.

Starting with the parametric version, we test the null H_0: model M_1 is not 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,

$$H_0 : \left\{E^0\left[LR_{12}\right] = 0\right\} \cup \left\{E^0\left[LR_{13}\right] = 0\right\} \cup \left\{E^0\left[LR_{14}\right] = 0\right\}, \tag{7}$$

against

$$H_1 : \left\{E^0\left[LR_{12}\right] > 0\right\} \cap \left\{E^0\left[LR_{13}\right] > 0\right\} \cap \left\{E^0\left[LR_{14}\right] > 0\right\}. \tag{8}$$

The rejection region for this test is given by $\min\{z_{12}, z_{13}, z_{14}\} > c_\alpha$, where c_α represents the critical value derived from a standard normal distribution. The intersection-union p-value is given by $p_1 = \max\{p_{12}, p_{13}, p_{14}\}$. Intersection-union p-values for the other comparisons are similar. Note that for a fixed level α, if $p_1 \leq \alpha$, then $\min\{p_2, p_3, p_4\} \geq 1-\alpha$. Therefore, the proposed CMST test ensures the detection of at most one significant model p-value at a time, in contrast to the CIT approach.

The non-parametric CMST test correspond to an intersection-union of Clarke's tests, exactly analogous to the one developed above for the parametric version. Because in practice $p_{12} + p_{21} \approx 1$ for Clarke's test, it follows that the non-parametric CMST test does not allow the detection of more than one significant model p-values at the same time, as well.

Simple application of separate Vuong tests, overlooks the dependency among the test statistics. Therefore, we consider a multivariate extension. We develop next the joint parametric CMST test for model M_1. Under the same general regularity conditions of Vuong (1989), the sample covariance of $L\hat{R}_{12,i}$ and $L\hat{R}_{13,i}$, $\hat{\sigma}_{12.13}$, converges almost surely to $Cov^0\left[LR_{12}, LR_{13}\right] = \sigma_{12.13}$ (and similarly for all other covariance terms). Therefore, the sample covariance matrix, $\hat{\mathbf{\Sigma}}_1$, converges almost surely to $\mathbf{\Sigma}_1$. It follows from the multivariate central limit and Slutsky's theorems that when

$$\begin{pmatrix} E^0\big[LR_{12}\big] \\ E^0\big[LR_{13}\big] \\ E^0\big[LR_{14}\big] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{9}$$

we have that

$$\mathbf{Z}_1 \;=\; \mathrm{diag}(\hat{\mathbf{\Sigma}}_1)^{-\frac{1}{2}}\, \mathbf{L\hat{R}}_1/\sqrt{n} \;\longrightarrow^d\; N_3(\mathbf{0}\,,\, \boldsymbol{\rho}_1)\;, \tag{10}$$

where $\mathbf{L\hat{R}}_1 = (L\hat{R}_{12}\,,\, L\hat{R}_{13}\,,\, L\hat{R}_{14})^T$ and $\boldsymbol{\rho}_1 = \mathrm{diag}(\mathbf{\Sigma}_1)^{-\frac{1}{2}}\, \mathbf{\Sigma}_1\, \mathrm{diag}(\mathbf{\Sigma}_1)^{-\frac{1}{2}}$ is the correlation matrix

$$\boldsymbol{\rho}_1 = \begin{pmatrix} 1 & \rho_{12.13} & \rho_{12.14} \\ \rho_{12.13} & 1 & \rho_{13.14} \\ \rho_{12.14} & \rho_{13.14} & 1 \end{pmatrix}\;. \tag{11}$$

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

$$H_0 : \min\left\{E^0\left[LR_{12}\right], E^0\left[LR_{13}\right], E^0\left[LR_{14}\right]\right\} \leq 0 \ . \tag{12}$$

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

$$H_1 : \min\left\{E^0\left[LR_{12}\right], E^0\left[LR_{13}\right], E^0\left[LR_{14}\right]\right\} > 0 \ , \tag{13}$$

using the statistic $W_1 = \min\{\mathbf{Z}_1\}$, with p-value

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

Note that by adopting the W_1 test statistic, the joint parametric CMST test follows the same spirit of an intersection union test. However, it does so by accounting for the dependency among the test statistics. The derivation of the tests for models M_2, M_3 and M_4 is analogous to the one just presented. Table 1 depicts the joint CMST tests for all models.

The CMST tests are implemented in the `R/qtlhot` package available at CRAN.

Simulation studies

We conducted two simulation studies. In the first, denoted the "pilot study", we focus on performance comparison of different methods with data generated from simple yet diverse causal models. The goal is to understand the behavior of our methods in simple settings. In the second, denoted the "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.

In the pilot simulation study we generated data from Models A to E in Figure 2. A detailed description of the simulation experiment and QTL mapping analysis is given in the Supplement. We evaluated the method's statistical performance using statistical power, miss-calling error rate and precision. These quantities were computed as,

$$\text{Power} = \frac{\text{TP}}{\text{N}}, \qquad \text{Miss-calling error} = \frac{\text{FP}}{\text{N}}, \qquad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

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 was 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 for now on), as well as, no multiple testing correction. We simulate data from the models in Figure 5 which emulate QTL hotspot patterns (i.e., genomic regions to which hundreds or thousands of traits co-map). A frequent goal in QTL hotspots studies is to determine a master regulator, i.e., a transcript that regulates the transcription of the traits mapping to the hotspot. A promising strategy towards this end is to test the *cis* trais (i.e., transcripts physically located close to the QTL hotspot) against all other co-mapping traits. Our simulations are aimed to evaluate the performance of the CMST tests in this setting. Details on the simulation experiment design and QTL mapping are provided in the Supplement.

Results

In this section we evaluate and compare the performance of our AIC- and BIC-based

causal model selection tests against the AIC and BIC model selection criterions and the CIT. We first present the results of our two simulation studies and then turn to the analysis of real data from yeast genetical genomics experiment. For the real data illustration we use a compendium of 247 knockout experiments (Hughes et al. 2000, Zhu et at. 2008) in yeast in order to evaluate the "biological" precision of causal predictions made by the methods under study.

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 pulled together. Panels 3(a-c) represent the simulations using 112 subjects, and panels 3(d-f) the 1,000 subjects simulation results. (The choice 112 was motivated by the sample size in our real data example.) The x-axis represents the significance levels used for computing the results. Note that 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 smallest AIC (BIC) 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 S4-S8 in the Supplement show the simulation results data for each one of Models A to E, when sample size is 112. Figures S9-S13 show the same results for sample

size 1,000. Inspection of these figures clearly suggests that 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 S4, S9, S7, S12). 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 S5, S10, S8, S13). Model C is particularly challenging (Figures S6 and S11), showing the highest false positive detection rates among all models.

We point out, however, that, in genetical genomics experiments we often restrict our attention to the analysis of *cis*-genes against *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 methods performance, 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.

Nonetheless, in the analysis of *trans* against *trans* genes there is no a priori reason to discard the relationship depicted in Model C, and an extra load of false positives should be expected. We point out, however, that 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 S6 and S11.

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 F. The top left panel shows that, except for the BIC-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 an accentuated decrease 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, nonetheless, seen to be too conservative even for the non-parametric tests.

Figure 7 presents the results for the simulations based on model G. Overall we see the same patterns as for model F, only more clear cut. For instance, application of BH correction seems quite deleterious for the joint and parametric CMST tests, where no significant calls were detected. For the non-parametric tests, on the other hand, BH correction seems to produce a slightly decrease in power (specially for the larger p-value cutoffs), with noticeable decrease in FDR (specially for the AIC based tests). The BY correction seems to be too conservative in this case again.

Because we fit almost three million hypothesis tests in the present simulation study, we did not include the CIT tests in this investigation restricting our attention to the

computationally more efficient CMST tests.

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 `R/qtlyeast` package at GITHUB.

To use this information, we: (i) determined which of the 247 ko-genes also showed

a significant QTL 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 8 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. In terms of the number of true positives, the CIT, BIC and AIC outperformed the CMST approaches, with the AIC-based CMST methods tending to be less powered than the BIC-based ones. However, the CIT, BIC and AIC also inferred the highest numbers of false positives (panel 8b), and showed low prediction precisions (panel 8c). From panel 8c we see that the CMST tests dominated the AIC, BIC and CIT meth-

ods, showing higher precision rates across all target significance levels. Among the CMST approaches, the joint parametric CMST tended to show the highest precisions, followed by the non-parametric and parametric CMST tests.

The results presented on Figure 8 were computed using all 135 ko-genes. However, in light of our simulation results, that suggest that the analysis of *cis* against *trans* genes is usually easier than the analysis of *trans* against *trans* genes, we investigated the results restricting ourselves to ko-genes with significant *cis* QTLs. Only 28 out of the 135 ko-genes were *cis* traits, but, nonetheless, were responsible for 2,947 out the total 31,975 "ko-gene/putative target gene" relationships. Figure 9 presents the results restricted to the *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 show better controlled 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. Even though an ideal method should show 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 ranked list of promising genes that causally affect a phenotype of interest, and it is time consuming and expensive to conduct validation experiments, a biologist might be more inclined to use a less powered and more precise method. However, if many genes can be easily validated, then the biologist might find the larger lists generated by more powered and less precise methods more appealing.

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 a hard time to pick up the correct causal ordering of 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 we need to be cautions. The problem is that the true causal relationships might take place outside the transcriptional regulation level. For instance, the true causal regulations 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. Nonetheless, our CMST tests include model M_4

(see Figure 1) and can, in principle, account for these latent variables.

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 of: (i) the incorporation of prior information via Bayesian reasoning; (ii) the adjustment for experimental factors, such as sex and age, that might induce correlations not explained the the causal relations; and (iii) the consideration of 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 two accounts. First, they tend to be conservative and precise, forfeiting to make calls in situations where alternative approaches might deliver a flood of false positive calls. Second, the CMST tests can adjust for experimental factors by modeling them as additive and interactive covariates (this feature is already implemented in our code, although we didn't need it for the yeast data analysis). Furthermore, because our tests can be applied to non-nested models of different dimensions, they can be readily extended to incorporate a larger number of models. For the parametric and non-parametric versions it simply means implementing intersection-union tests on a larger number of Vuong's tests. For the joint-parametric test we just need to handle a higher dimensional null distribution. Finally, even though we do not attempt it here, the incorporation of prior information via Bayesian reasoning represents an exciting direction for future work.

In theory, 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 Benjamini-Hochberg procedure does not hold. Nonetheless, we still evaluated the performance of this method, in addition to the Benjamini-Yekutieli (BY) procedure, that relaxes the independent test statistics assumption, in our simulations.

Our results suggest that multiple testing correction should not be performed for the joint and the parametric CMST tests, as they achieve low FDR levels without any correction and show severe reduction in statistical power with the application of BH and BY control procedures. 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 S17, S18, S19, and S20 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

1. Akaike H., 1974 A new look at the statistical model identification. IEEE Transactions on Automatic Control **19:** 716723.

2. Aten J. E., T. F. Fuller, A. J. Lusis, S. Horvath, 2008 Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Systems Biology **2**: 34.

3. Benjamini Y., Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society (Series B) **57:** 289-300.

4. Brem R., L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression trait in yeast. PNAS **102:** 1572-1577.

5. Broman K., H. Wu, S. Sen, G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19**: 889-890.

6. Clarke K. A., 2007 A simple distribution-free test for nonnested model selection. Political Analysis **15**: 347-363.

7. Chaibub Neto E., C. Ferrara, A. D. Attie, B. S. Yandell, 2008 Inferring causal phenotype networks from segregating populations. Genetics **179**: 1089-1100.

8. Chaibub Neto E., M. P. Keller, A. D. Attie, B. S. Yandell, 2009 Causal graphical models in system genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. Annals of Applied Statistics **4**: 320-339.

9. Chen L. S., F. Emmert-Streib, J. D. Storey, 2007 Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biology **8**: R219.

10. Churchill G. A., R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138**: 963-971.

11. Duarte C. W., Z. B. Zeng, 2011 High-confidence discovery of genetic network regulators in expression quantitative trait loci data. Genetics **187**: 955-964.

12. Dupuis J. and D. Siegmund, 1999 Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics **151:** 373-386.

13. Ghazalpour A., S. Doss, B. Zhang, S. Wang, C. Plaisier, et al., 2006 Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight. PLoS Genetics **2**(8): e130.

14. Haley C., S. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**: 315-324.

15. Hughes T. R., M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, et al, 2000 Functional discovery via a compendium of expression profiles. Cell **102:** 109-116.

16. Kulp D. C., M. Jagalur, 2006 Causal inference of regulator-target pairs by gene mapping of expression phenotypes. BMC Genomics **7**: 125.

17. Kullback S., 1959 Information theory and statistics. John Wiley and Sons. New York.

18. Lander E. S., D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185-199.

19. Li R., S. W. Tsaih, K. Shockley, I. M. Stylianou, J. Wergedal, B. Paigen, G. A. Churchill, 2006 Structural model analysis of multiple quantitative traits. PLoS Genetics **2**: e114.

20. Li Y., B. M. Tesson, G. A. Churchill, R. C. Jansen, 2010 Critical preconditions for causal inference in genome-wide association studies. Trends in Genetics **26:** 493-498.

21. Liu B., A. de la Fuente, I. Hoeschele, 2008 Gene network inference via structural equation modeling in genetical genomics experiments. Genetics **178**: 1763-1776.

22. Manichaikul A., J. Dupuis, S. Sen, and K. W. Broman, 2006 Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. Genetics **174:** 481-489.

23. Manichaikul A., J. Y. Moon, S. Sen, B. S. Yandell, K. W. Broman, 2009 A model selection approach for the identification of quantitative trait loci in experimental crosses. Genetics **181**: 1077-1086.

24. Margolin A., A. Califano, 2007 Theory and limitations of genetic network inference from microarray data. Annals of the New York Academy of Sciences **1115**: 51-72.

25. Millstein J., B. Zhang, J. Zhu, E. E. Schadt, 2009 Disentangling molecular relationships with a causal inference test. BMC Genetics **10**: 23 doi:10.1186/1471-2156-10-23.

26. Rockman M. V., 2008 Reverse engineering the genotype-phenotype map with natural genetic variation. Nature **456**: 738-744.

27. Sawa T., 1978 Information criteria for discriminating among alternative regression models. Econometrica **46**: 1273-1291.

28. Schadt E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, et al., 2005 An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics **37**: 710-717.

29. Schwarz G. E., 1978 Estimating the dimension of a model. Annals of Statistics **6:** 461464.

30. Storey J, Tibshirani R (2003) Statistical significance for genomewide studies. PNAS **100**: 9440-9445.

31. Sun W., T. Yu, K. C. Li, 2007 Detection of eQTL modules mediated by activity levels of transcription factors. Bioinformatics **23**: 2290-2297.

32. Vuong Q. H., 1989 Likelihood ratio tests for model selection and non-nested hypothesis. Econometrica **57**: 307-333.

33. Winrow C. J., D. L. Williams, A. Kasarskis, J. Millstein, A. D. Laposky, et al., 2009 Uncovering the genetic landscape for multiple sleep-wake traits. PLoS ONE **4**: e5161.

34. Zhu J., M. C. Wiener, C. Zhang, A. Fridman, E. Minch, et al., 2007 Increasing the Power to Detect Causal Associations by Combining Genotypic and Expression Data in Segregating Populations. PLoS Computational Biology **3**(4): e69. doi:10.1371/journal.pcbi.0030069

35. Zhu J., B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, E. E. Schadt, 2008 Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature Genetics **40**: 854-861.

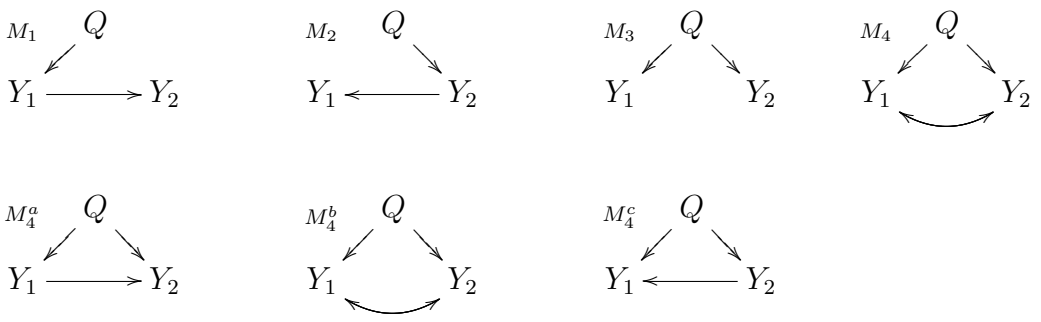

Figure 1: Pairwise causal models fitted by the CMST tests. Y_1 and Y_2 represent phenotypes that co-map to a same QTL, Q. Models M_1, M_2, M_3 and M_4 represent, respectively, the causal, reactive, independent and full model. In model M_1 the phenotype Y_1 has a causal effect on Y_2. In M_2, the phenotype 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, actually corresponds to three distribution equivalent models M_4^a, M_4^b, and M_4^c which cannot be distinguished using the data because their maximized likelihood scores are identical. Model M_4^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_4^a and M_4^c correspond to causal-pleiotropic and reactive-pleiotropic relations, respectively.

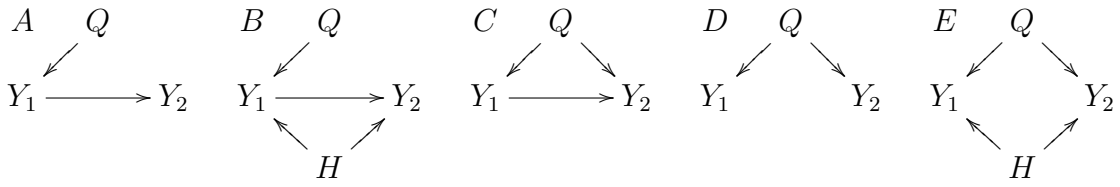

Figure 2: Models used in the simulation study. Y_1 and Y_2 represent phenotypes that co-map to a same QTL, Q. Model A represents a causal effect of Y_1 on Y_2. Model B represents the same, with the additional complication that part of the correlation between Y_1 and Y_2 is due to a hidden-variable H. 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.

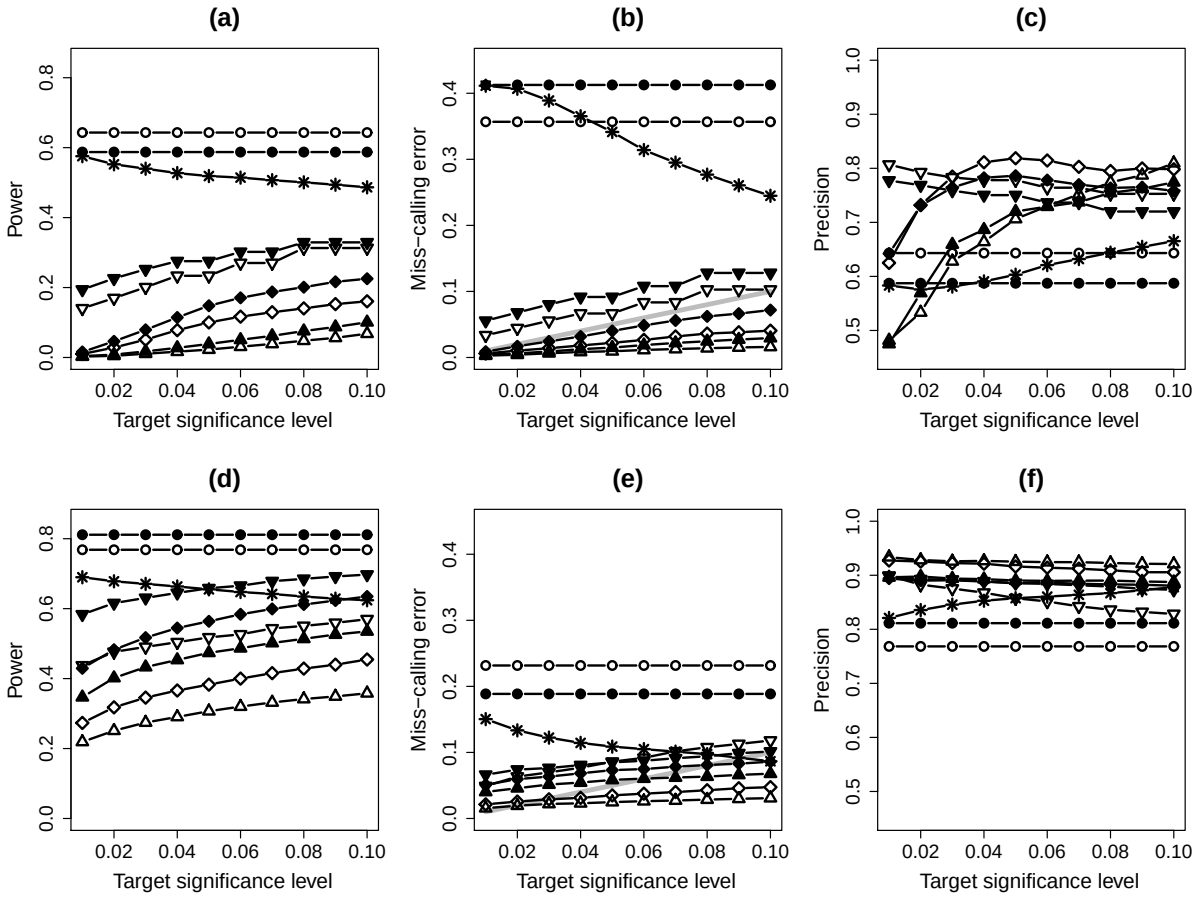

Figure 3: Overall power, miss-calling error rate and precision across the simulated models depicted in Figure 2. Panels a-c represent the simulations based on sample size 112, whereas panels d-f present the results for sample size 1,000. Asterisk represents the CIT. Empty and filled symbols represent, respectively, AIC- and BIC-based methods. Diamonds: parametric CMST. Point-down triangles: non-parametric CMST. Point-up triangles: joint-parametric CMST. Circles: AIC and BIC.

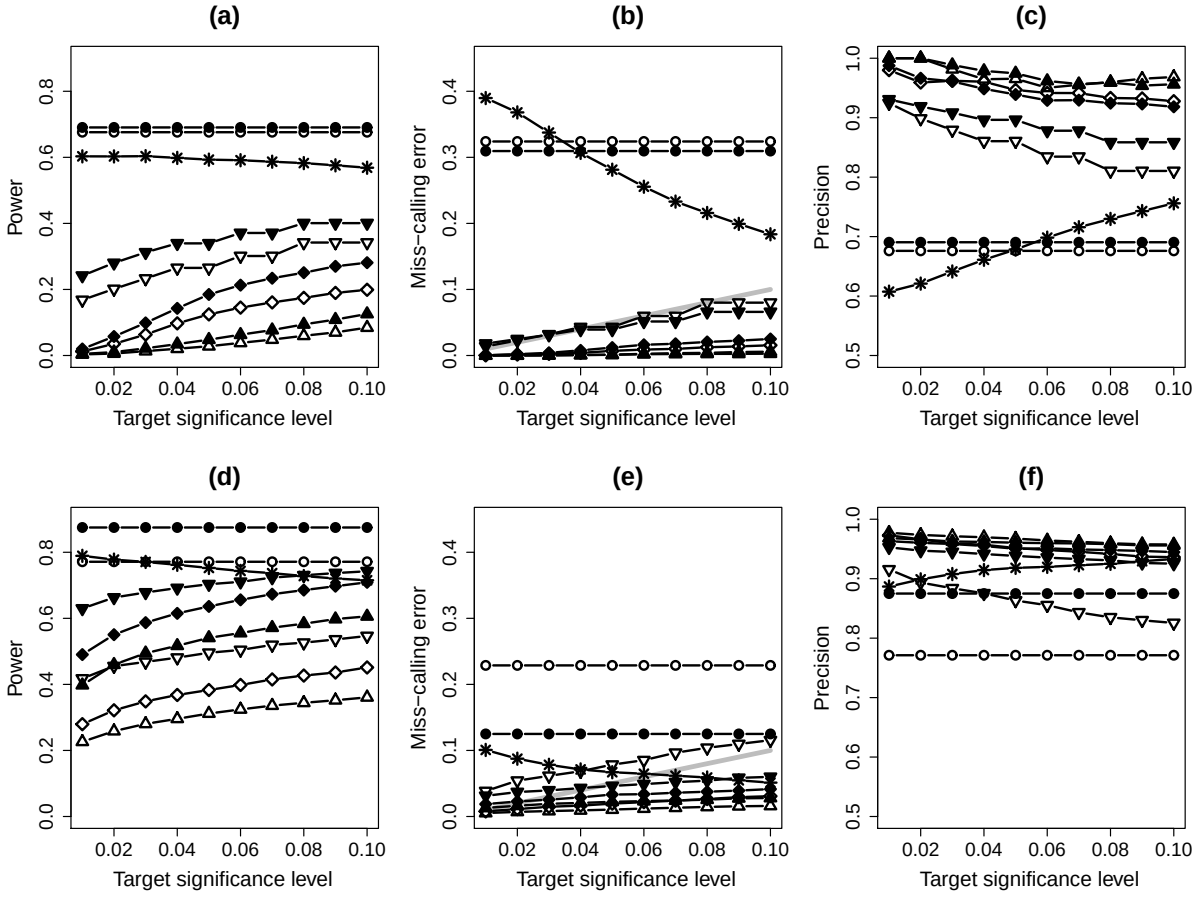

Figure 4: Overall power, miss-calling error rate and precision restricted to the *cis* versus *trans* cases. The results were computed using only the simulated models A, B, D and E in Figure 2, since the pleiotropic causal relationship depicted in Model C is expected to be much less frequent than the 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. Asterisk represents the CIT. Empty and filled symbols represent, respectively, AIC- and BIC-based methods. Diamonds: parametric CMST. Point-down triangles: non-parametric CMST. Point-up triangles: joint-parametric CMST. Circles: AIC and BIC.

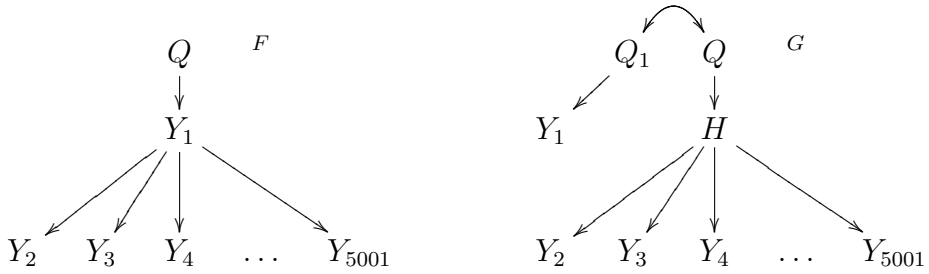

Figure 5: Models generating hotspot patterns. Y_1 represents a *cis* expression trait. Y_k, $k = 2, \ldots, 5001$ represent expression traits mapping in *trans* to the hotspot QTL Q. H represents an unobserved expression trait. Model F generates a hotspot pattern derived from the causal effect of the 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.

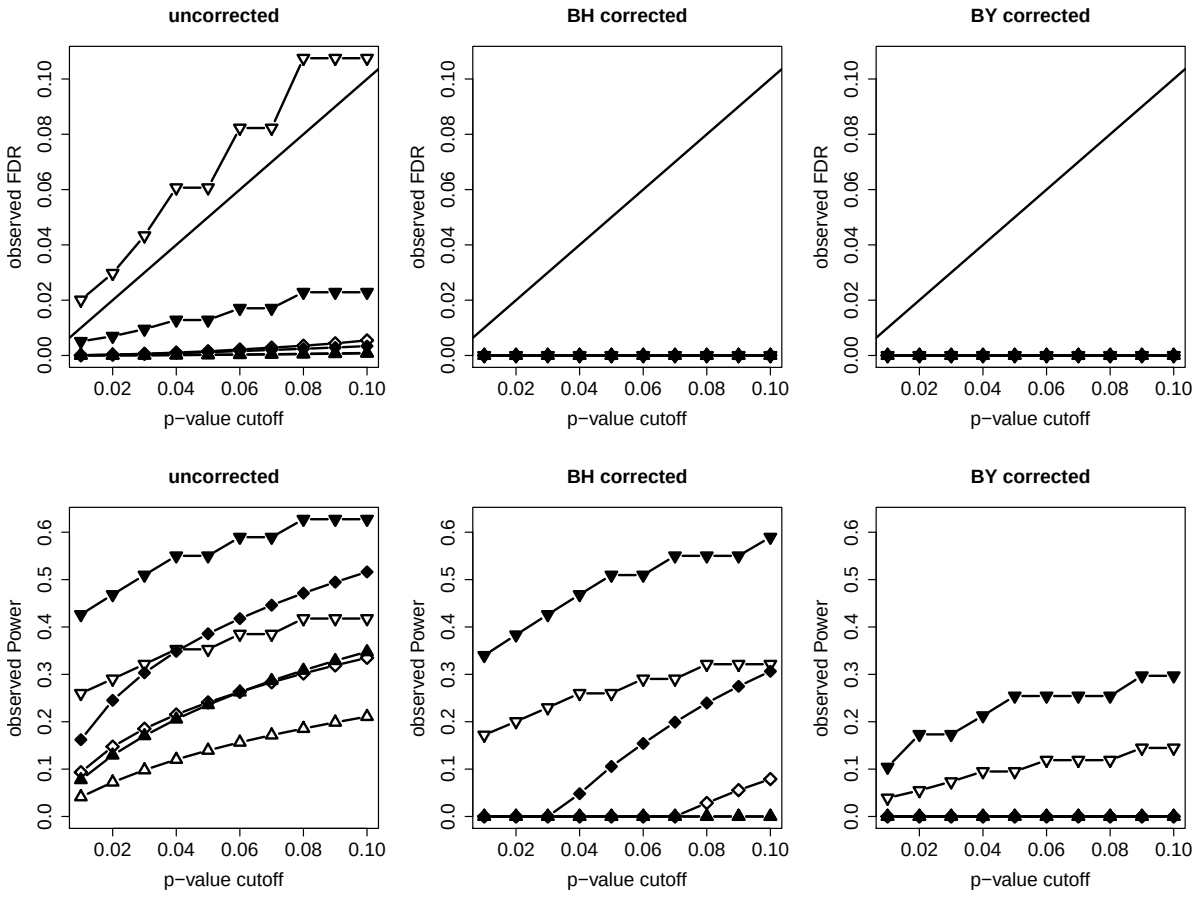

Figure 6: Observed FDR and power for the simulations based on model F. Empty and filled symbols represent, respectively, AIC- and BIC-based methods. Diamonds: parametric CMST. Point-down triangles: non-parametric CMST. Point-up triangles: joint-parametric CMST. Circles: AIC and BIC.

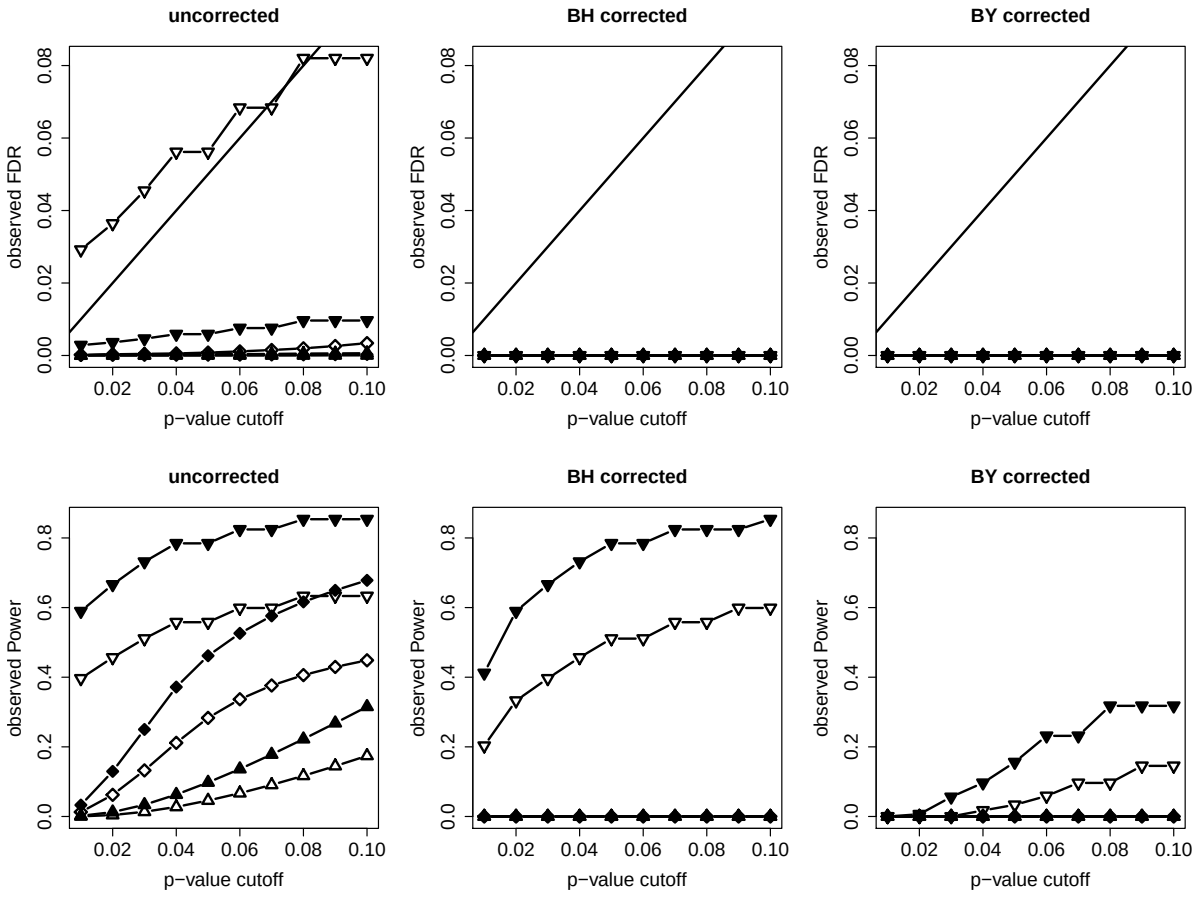

Figure 7: Observed FDR and power for the simulations based on model G. Empty and filled symbols represent, respectively, AIC- and BIC-based methods. Diamonds: parametric CMST. Point-down triangles: non-parametric CMST. Point-up triangles: joint-parametric CMST. Circles: AIC and BIC.

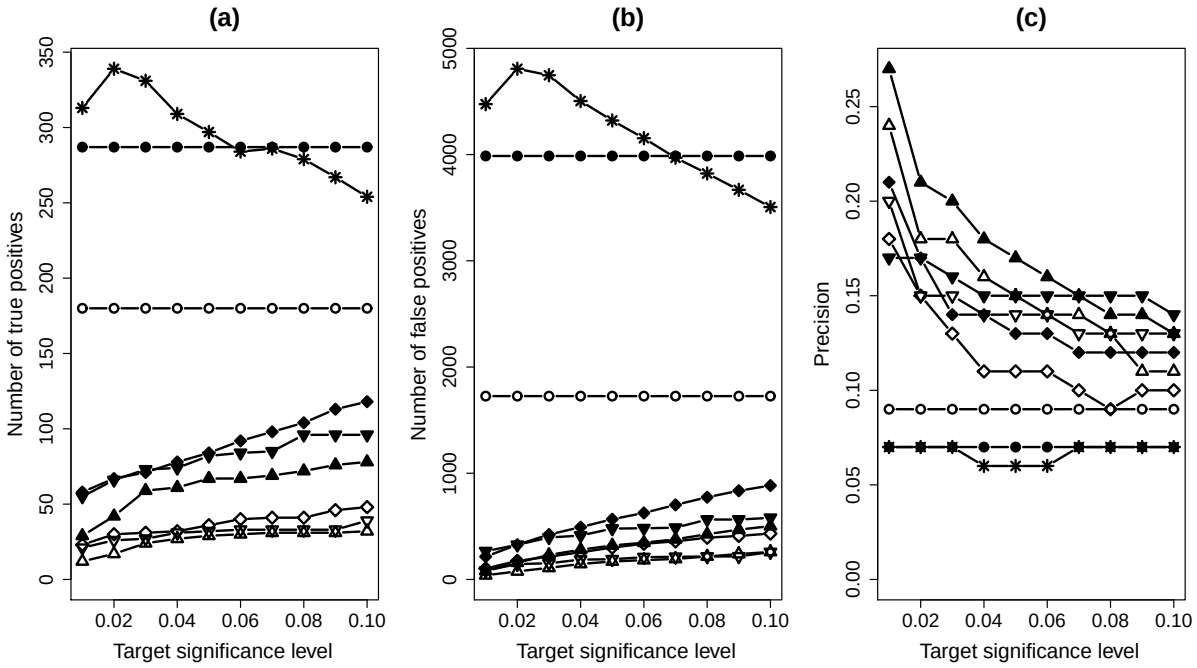

Figure 8: Overall number of true positives, number of false positives and precision across all 135 ko-gene/putative target lists. Asterisk represents the CIT. Empty and filled symbols represent, respectively, AIC- and BIC-based methods. Diamonds: parametric CMST. Point-down triangles: non-parametric CMST. Point-up triangles: joint-parametric CMST. Circles: AIC and BIC.

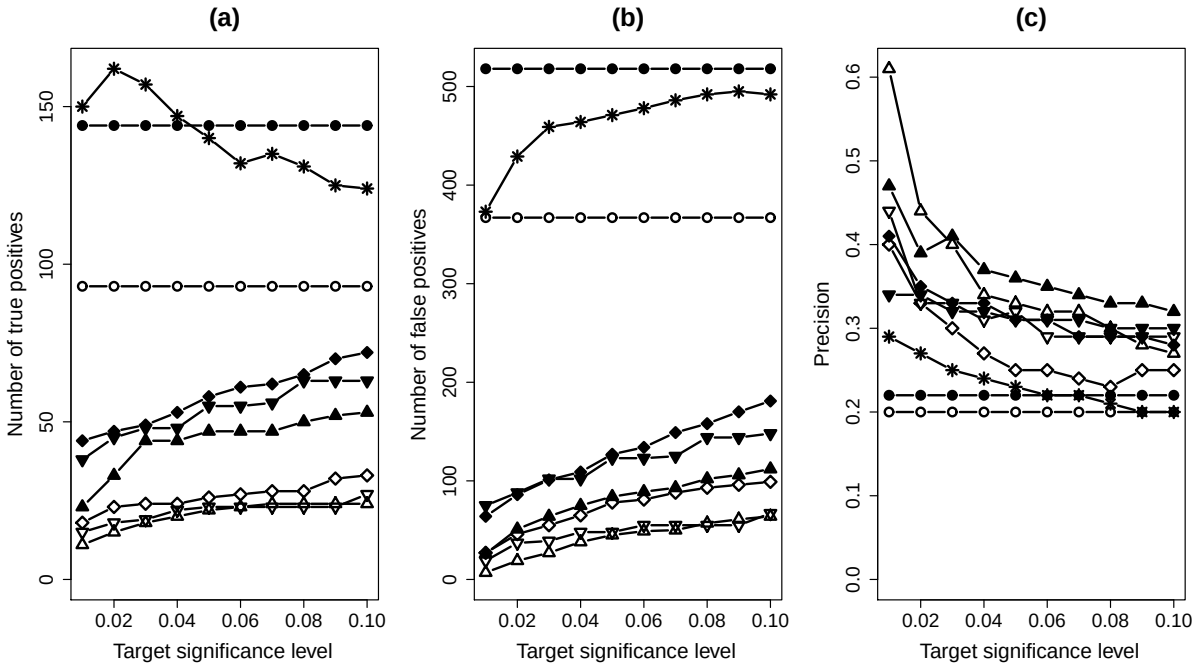

Figure 9: Overall number of true positives, number of false positives and precision restricted to 28 *cis* ko-gene/putative target lists. Asterisk represents the CIT. Empty and filled symbols represent, respectively, AIC- and BIC-based methods. Diamonds: parametric CMST. Point-down triangles: non-parametric CMST. Point-up triangles: joint-parametric CMST. Circles: AIC and BIC.

H_0	Null distribution	P-value
$H_0^{M_1}$	$\mathbf{Z}_1 = (Z_{12}, Z_{13}, Z_{14})^T \sim N_3(\mathbf{0}, \hat{\boldsymbol{\rho}}_1)$	$P(Z_{12} \geq w_1,\, Z_{13} \geq w_1,\, Z_{14} \geq w_1)$
$H_0^{M_2}$	$\mathbf{Z}_2 = (Z_{21}, Z_{23}, Z_{24})^T \sim N_3(\mathbf{0}, \hat{\boldsymbol{\rho}}_2)$	$P(Z_{21} \geq w_2,\, Z_{23} \geq w_2,\, Z_{24} \geq w_2)$
$H_0^{M_3}$	$\mathbf{Z}_3 = (Z_{31}, Z_{32}, Z_{34})^T \sim N_3(\mathbf{0}, \hat{\boldsymbol{\rho}}_3)$	$P(Z_{31} \geq w_3,\, Z_{32} \geq w_3,\, Z_{34} \geq w_3)$
$H_0^{M_4}$	$\mathbf{Z}_4 = (Z_{41}, Z_{42}, Z_{43})^T \sim N_3(\mathbf{0}, \hat{\boldsymbol{\rho}}_4)$	$P(Z_{41} \geq w_4,\, Z_{42} \geq w_4,\, Z_{43} \geq w_4)$

Table 1: Model selection tests for models M_1, M_2, M_3 and M_4. Here $w_k = \min\{\mathbf{z}_k\}$ for $k = 1, \ldots, 4$, and $\boldsymbol{\rho}_k$ is defined in analogy with $\boldsymbol{\rho}_1$ in equation 11.

CMST	Model A	Model B	Model C	Model D	Model E
M_1	TP	TP	FP	FP	FP
M_2	FP	FP	FP	FP	FP
M_3	FP	FP	FP	TP	FP
M_4	FP	FP	TP	FP	TP
CIT	Model A	Model B	Model C	Model D	Model E
M_1	TP	TP	FP	FP	FP
M_2	FP	FP	FP	FP	FP
M_k	FP	FP	TP	TP	TP

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 is generated from the model on column j. For instance, when data is 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 is 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.

Supplement

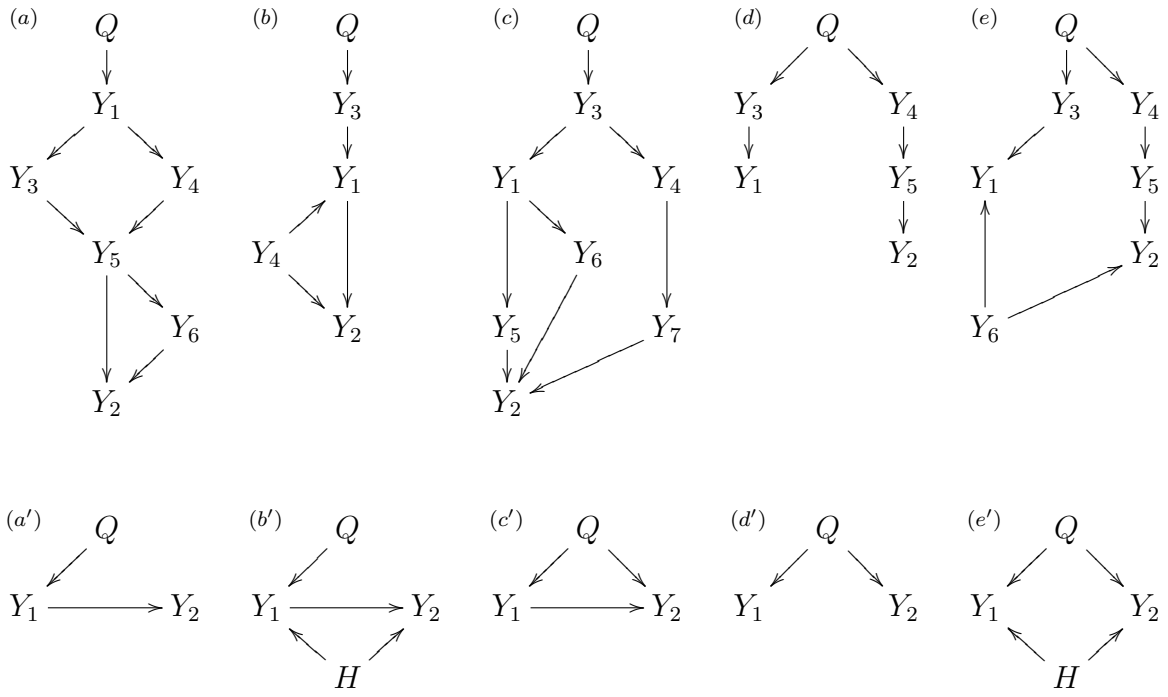

Figure S 1: Network models and their collapsed versions. The collapse networks (bottom panels) represent simplified versions of the true networks (top panels), where nodes other than Q, Y_1 and Y_2 are ignored, even though they still represent correct the causal flow among these three nodes in the true network. Consider, for example, network c and its collapsed version c'. The path $Q \to Y_3 \to Y_1$ in c is collapsed to $Q \to Y_1$ in c'. The paths $Y_1 \to Y_5 \to Y_2$ and $Y_1 \to Y_6 \to Y_2$ in c are collapsed to $Y_1 \to Y_2$ in c'. The path $Q \to Y_3 \to Y_4 \to Y_7 \to Y_2$ in c is collapsed to $Q \to Y_2$ in c'.

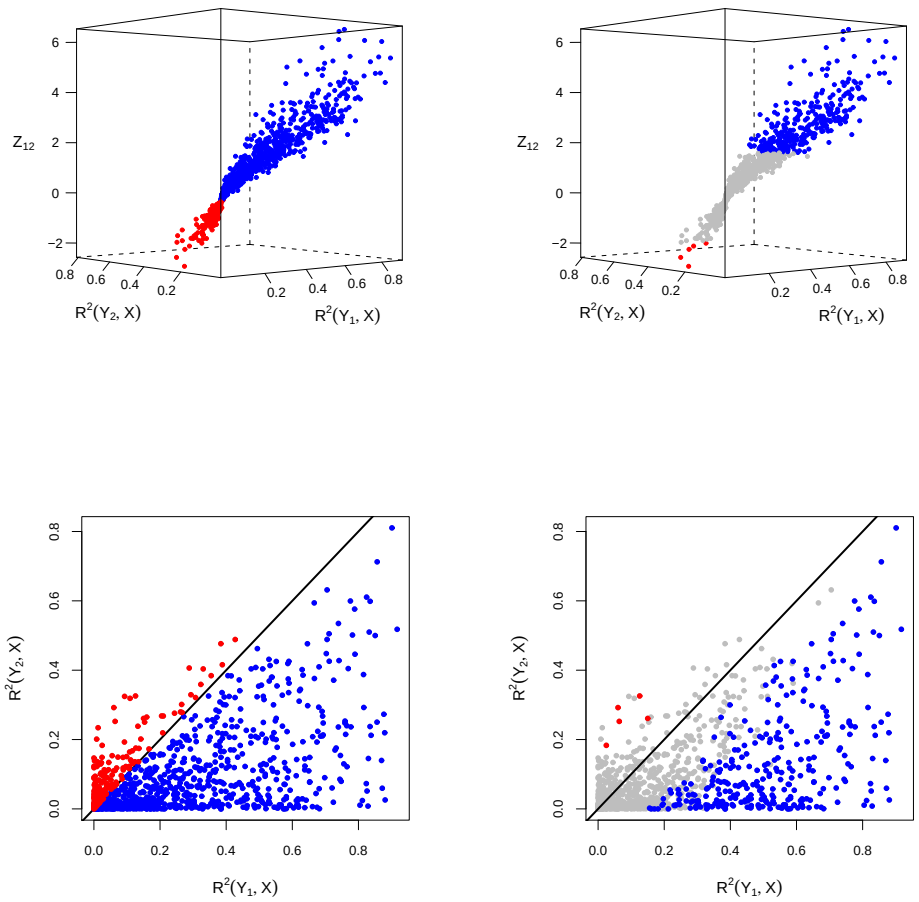

Figure S 2: Model selection via log-likelihood ratio versus Vuong's test.

Figure S2 illustrates how Vuong's test works. We generated 1,000 data-sets from the model $X \rightarrow Y_1 \rightarrow Y_2$ and applied Vuong's test to the comparison of models $M_1 : X \rightarrow Y_1 \rightarrow Y_2$ against $M_2 : X \rightarrow Y_2 \rightarrow Y_1$. The top panels present 3D scatter plots of the test statistics Z_{12} against the R^2 values of the regression of Y_1 on X, $R^2(Y_1, X)$, and the R^2 values of the regression of Y_2 on X, $R^2(Y_2, X)$. The data points are color coded as blue, red and grey, representing, respectively, M_1, M_2 and "no calls". Blue and red points represent, respectively, correct and incorrect calls. The bottom panels follow the

same color coding and show the projections of the 3D scatter plots into the $R^2(Y_1, X)$ by $R^2(Y_2, X)$ plane.

The left panels of Figure S2 show the model selection results based on the log-likelihood ratio (LR) criterium, where positive $L\hat{R}_{12}$ values support M_1 and negative $L\hat{R}_{12}$ values support M_2 (note that we actually use the Z_{12} test statistics, instead of $L\hat{R}_{12}$ statistics, but the results are equivalent). Because we generate the data from model M_1, it will usually be the case that X explains a greater proportion of the variability of Y_1 than of Y_2. In other words, $R^2(Y_1, X)$ will tend to be higher than $R^2(Y_2, X)$. However, some of the data-sets show the opposite trend due to random noise on the data. The bottom left panel shows that the log-likelihood criterium tends to make incorrect calls when $R^2(Y_1, X) < R^2(Y_2, X)$.

The right panels of Figure S2 show the model selection results derived from Vuong's test. Now we see that most of the incorrect calls made by the log-likelihood criterium (red points) are not significant (grey points) according to Vuong's test, that requires that $Z_{12} \leq -1.64$ or $Z_{12} \geq 1.64$ in order to achieve statistical significance at a 5% level. The drawback is the reduction in power to detect the correct calls, since not only red dots are replaced by grey dots, but many of the blue dots are turned into grey, as well. These figures illustrate how Vuong's test trade a decrease in detection of false positives by a reduction in statistical power to detect true positives.

A technical note on Vuong's test

Vuong (1989) fully characterized the asymptotic distribution of the log-likelihood ratio statistic under the most general conditions. He showed that the form of the asymptotic distribution of the log-likelihood ratio depends on whether the models are observationally identical or not. Two models are observationally identical if their probability densities are the same, when evaluated at the respective pseudo-true parameter values, i.e., $f_1(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{1*}) = f_2(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{2*})$ for almost all $(\mathbf{y}, \mathbf{x})$, where the pseudo-true parameter values, $\boldsymbol{\theta}_{k*}$, corresponds to the parameter value that minimizes the Kullback-Leibler distance from the true model (Sawa 1978).

Explicitly, Vuong showed (Theorem 3.3 on page 313) that under very general conditions:

1. If $f_1(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{1*}) = f_2(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{2*})$, then $2\,LR_{12}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ converges in distribution to a weighted sum of chi-square distributions.

2. If $f_1(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{1*}) \neq f_2(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{2*})$, then

$$\frac{1}{\sqrt{n}} \left(LR_{12}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) - E^0 \left[\log \frac{f_1(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{1*})}{f_2(\mathbf{y} \mid \mathbf{x}\,;\, \boldsymbol{\theta}_{2*})} \right] \right) \rightarrow^d N(0, \sigma_{12.12})$$

Because of this interesting asymptotic behavior Vuong had to proposed 3 distinct model selection tests: one for strictly non-nested models, that are always not observationally identical; another for overlapping models that might or might not be observationally identical; and a third for nested models, that are always observationally identical. (Nested models are always observationally identical because the nested model cannot be better than the full model and both models are equally close to the true model if and only if they are the same.)

In our applications, models M_1, M_2 and M_3 are not nested on each other, but are nested on models M_4^a, M_4^c and M_4^b, respectively (Figure 1 in the main text). Hence, our model selection tests consider pairs of models that are either non-nested or nested. In the Methods section we presented Vuong's test for not observationally identical models, that is suitable for the comparison of strictly non-nested models ($M_1 \times M_2$, $M_1 \times M_3$ and $M_2 \times M_3$).

We point out, however, that even though we perform model selection tests between nested models ($M_1 \times M_4$, $M_2 \times M_4$ and $M_3 \times M_4$) we don't need to use Vuong's test for nested models because our test statistics are based on penalized log-likelihoods instead of log-likelihoods, and our penalized models are not observationally identical for nested models too. In other words, even though $f_1(\mathbf{y} \mid \mathbf{x}\,; \boldsymbol{\theta}_{1*}) = f_4(\mathbf{y} \mid \mathbf{x}\,; \boldsymbol{\theta}_{4*})$ when model 1 is nested in model 4, we have that $f_1(\mathbf{y} \mid \mathbf{x}\,; \boldsymbol{\theta}_{1*}) - p_1 \neq f_4(\mathbf{y} \mid \mathbf{x}\,; \boldsymbol{\theta}_{4*}) - p_4$ since the penalty p_1 is smaller than p_4. Therefore, we can simply use Vuong's test for not observationally identical models in this case too.

On a technical note, we point out that Vuong's Theorem 3.3 still holds when we replace the log-likelihood ratio by the penalized log-likelihood ratio. The demonstration mimics Vuong's original proof presented on page 327. We just need to replace the log-likelihoods by penalized log-likelihoods in the Taylor expansion of the log-likelihoods around the maximum likelihood estimates.

Simulation studies

Here we provide further details on the simulation studies presented in the main text.

Pilot simulation study

We conducted a total of 10 simulation studies, generating data from the five models described in Figure 2 in the main text using sample sizes 112 and 1,000 (the choice 112 was motivated by the sample size in our real data example). For each model, we simulated 1,000 backcrosses composed with 3 chromosomes of length 100cM containing 101 unequally spaced markers per chromosome. For each one of the simulated backcrosses, the additive and dominance genetic effects were sampled, respectively, from the $U[-0.75, 0.75]$ and $U[0, 0.75]$ distributions, where $U[a, b]$ represents the uniform distribution on the interval $[a, b]$. Residual error rates were sampled from $U[0.5, 1.5]$, and the phenotype to phenotype regression coefficients in Figures 2 A, B and C were sampled from $U[-1, 1]$. The hidden-variable to phenotype regression coefficients on Figures 2 B and E were sampled from $U[-1, 1]$ and $U[0.5, 1]$, respectively. This choice of parameters ensured that approximately 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 (see Figure S3a, and the axis scales on Figures S4-S8) and between 0.01 to 0.20 for the simulations based on 1,000 subjects (see Figure S3b, and the axis scales on Figures S9-S13).

The backcross simulations and the QTL mapping analyses were performed using the R/qtl software (Broman et al. 2003). We performed Haley-Knott regression (Haley and Knott 1992) and adopted 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 used a permutation LOD threshold (Churchill and Doerge 1994) of 2.24 for the QTL mapping analysis, aiming to control the genome wide error rate of falsely detecting

a QTL at a 5% rate. The selection of the co-mapping QTL, to be used as a causal anchor in the tests, was performed as described in the next section.

Large scale simulation study

We performed two separate simulation studies generating data from the models in Figure 5 in the main text. In model F, Y_1 plays the role of a master regulator *cis* trait, and all other traits map in *trans* to QTL hotspot QTL Q because of the causal effect of Y_1. In model G, Y_1 plays the role of a *cis* trait mapping to a QTL closely linked to Q, and, therefore, causally independent of the *trans* traits in the hotspot.

In each simulation study we generated 1,000 distinct backcrosses with genetic data composed of 3 chromosomes of length 100cM containing 101 markers per chromosome, and 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. The additive and dominance genetic effects of Q on Y_1 were sampled, respectively, from the $U[0.5, 1]$ and $U[0, 0.5]$ distributions. Residual error rates were sampled from $U[0.5, 1.5]$, and the coefficients of the regressions of Y_k on Y_1 were sampled from $U[0.5, 1]$. Figure S15 shows the overall R^2 distributions. QTL mapping was performed as in the pilot study, but here we used the QTL for trait Y_1 as a causal anchor.

For each simulated data set we tested Y_1 against all other phenotypes Y_k, $k = 2, \ldots, 5001$, that share the QTL with Y_1, so that the number of hypothesis tests varied from simulation to simulation. Figure S16 shows the distribution of the number of tests per simulation study. In total we performed 1,656,261 tests for the simulations with model F, and 1,286,243 tests for the simulations with model G.

The empirical FDR (that corresponds to one minus the precision) was computed as the ratio of the number of FPs by the sum of the number of FPs and TPs across all

tests. The empirical power was computed as before. For model F, a FP is defined as any statistically significant M_2, M_3, or M_4 call, and a TP is given by a significant M_1 call. For model G, on the other hand, a FP corresponds to any statistically significant M_1, M_2, or M_4 call, and a TP is given by a significant M_3 call. For the evaluations without multiple testing correction, a call M_k was statistically significant if the respective p-value, p_k, was smaller than a fixed significance level α.

Multiple testing correction procedures based on the control of family wise error rates tend to be very conservative, and are not advisable. Here, we investigate the performances of the Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) FDR control procedures (denoted, respectively, by BH and BY for now on). The BH and BY adjusted p-values were computed based on the p-values across all simulations pooled together, separately by model call (e.g., for the model F simulations, we pool together all 1,656,261 M_1 p-values and apply the BH adjusted for this set of p-values, and similarly for the M_2, M_3 and M_4 p-values), and then compute the FDR and power empirical estimates using the adjusted p-values.

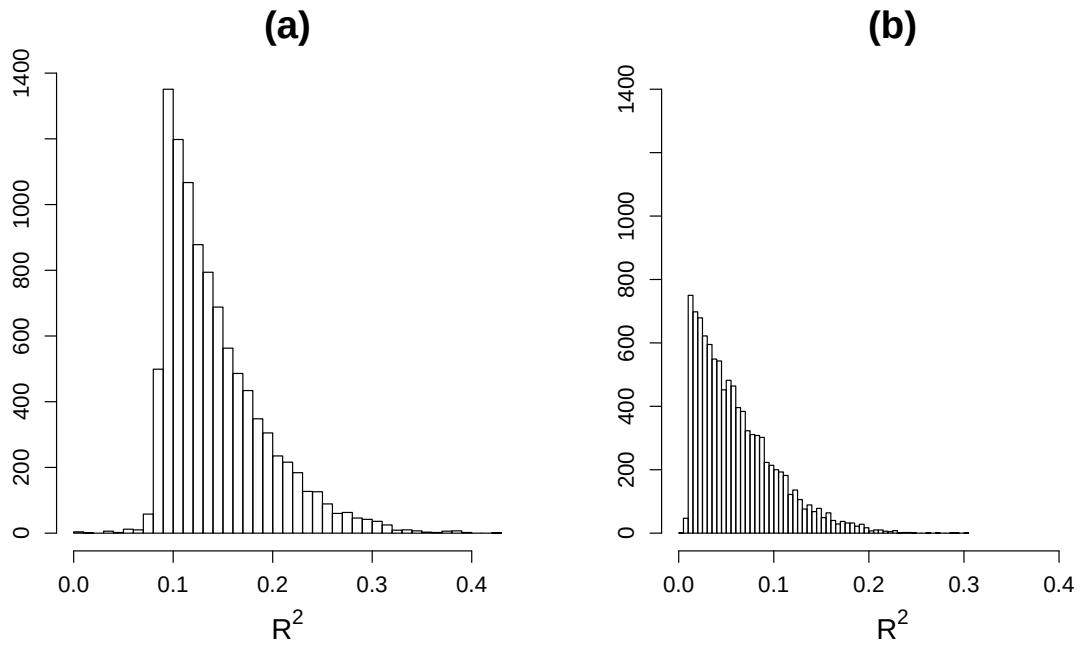

Figure S 3: Overall distribution of the R^2 statistics across all simulated models in Figure 2. Panels a and b present the R^2 statistics for sample sizes 112 and 1,000, respectively.

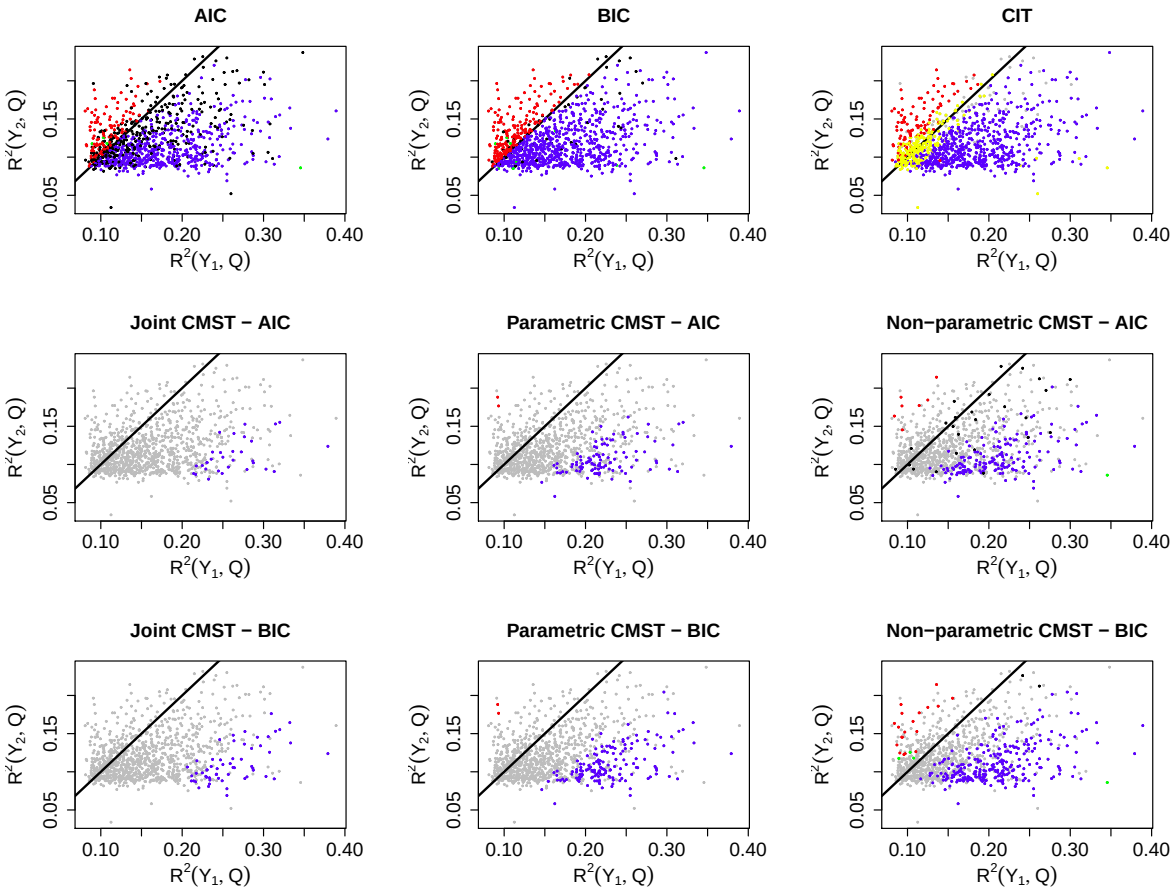

Figure S 4: Simulation results for Model A in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this model, blue dots represent true positives. Red, green and black dots represent false positives for the AIC, BIC and CMST methods. Red and yellow dots represent false positives for the CIT.

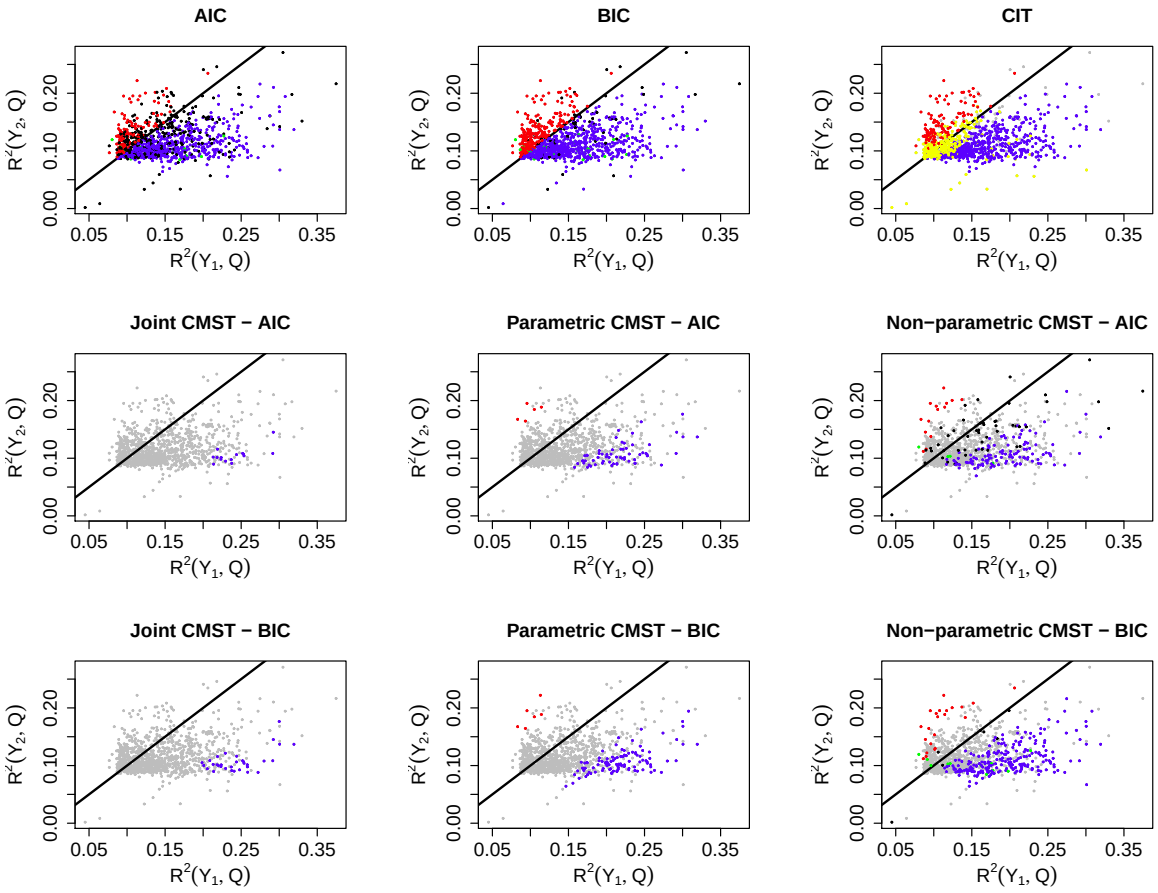

Figure S 5: Simulation results for Model B in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this model, blue dots represent true positives. Red, green and black dots represent false positives for the AIC, BIC and CMST methods. Red and yellow dots represent false positives for the CIT.

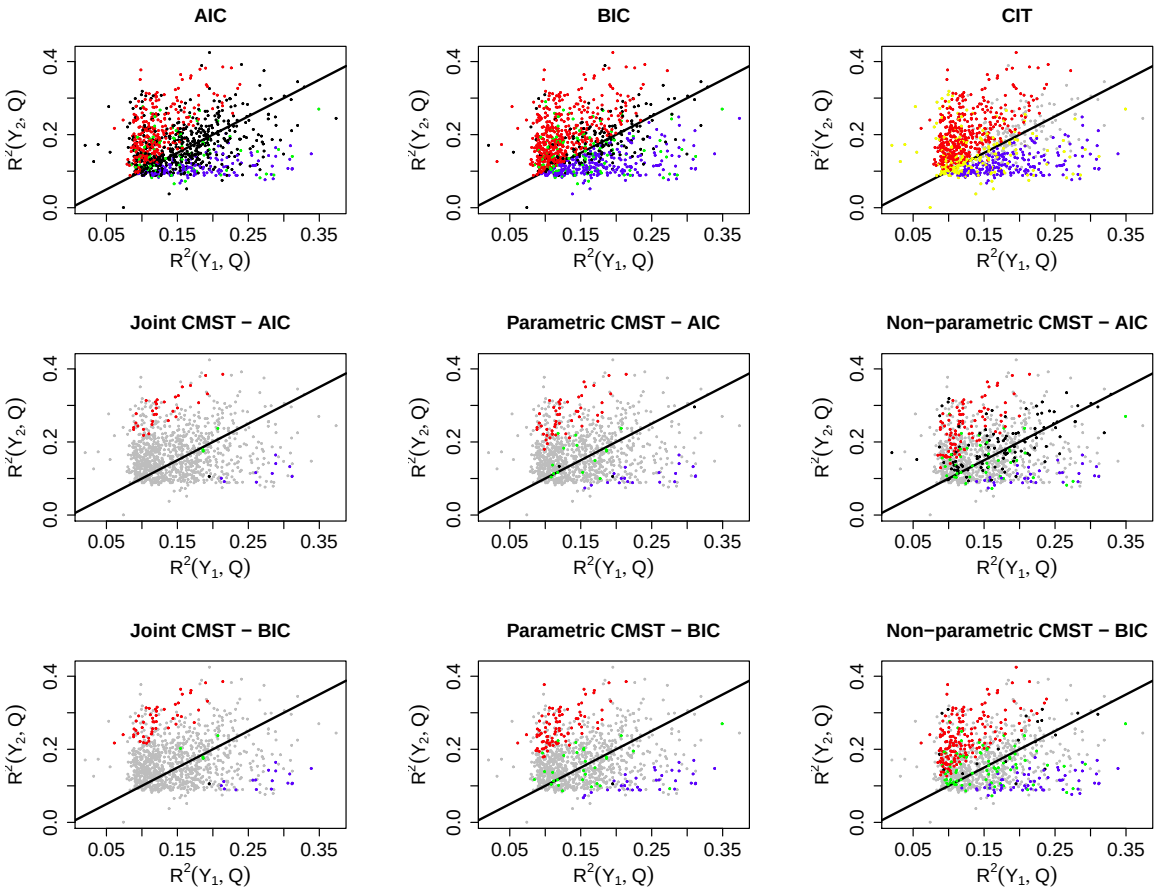

Figure S 6: Simulation results for Model C in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the AIC, BIC and CMST methods, black dots represent true positives, and blue, red and green dots represent false positives. For the CIT test, yellow dots represent true positives and blue and red dots show false positives.

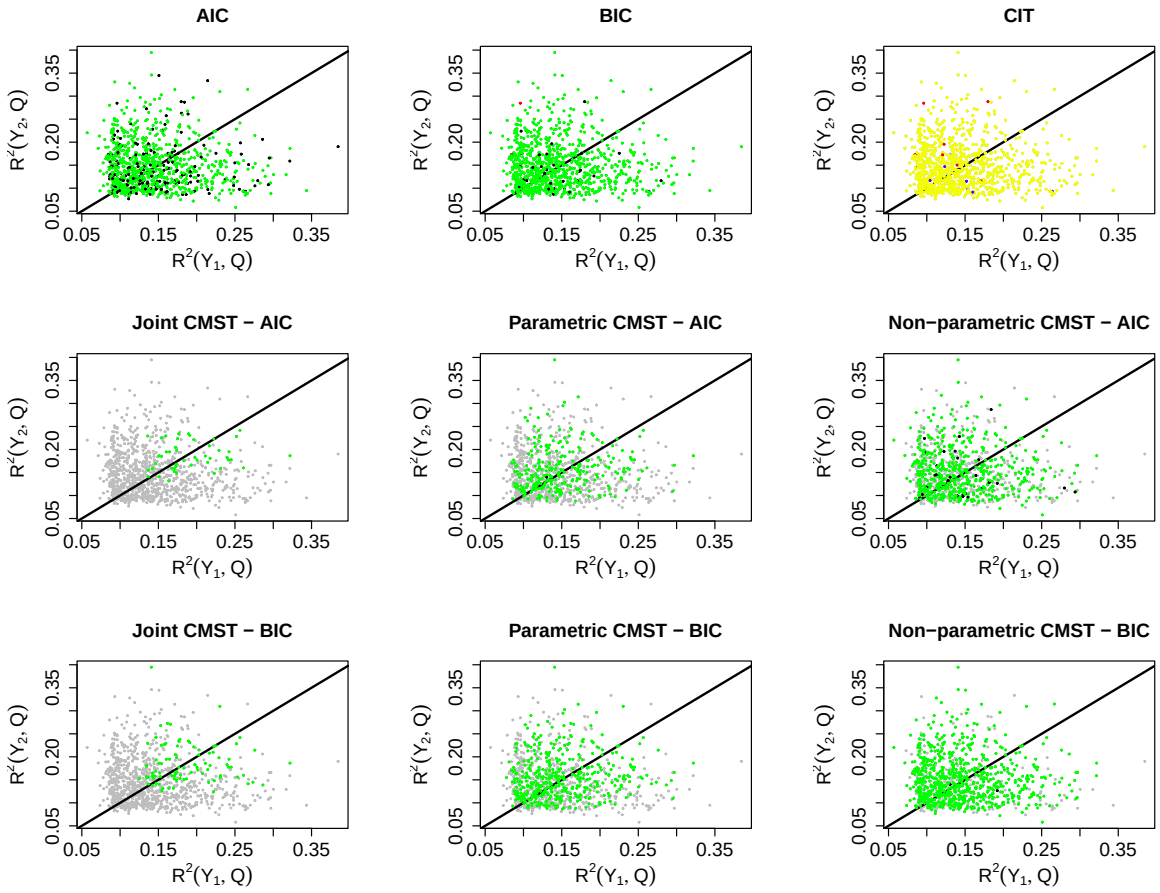

Figure S 7: Simulation results for Model D in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the AIC, BIC and CMST methods, green dots represent true positives, and blue, red and black dots represent false positives. For the CIT test, yellow dots represent true positives and blue and red dots show false positives.

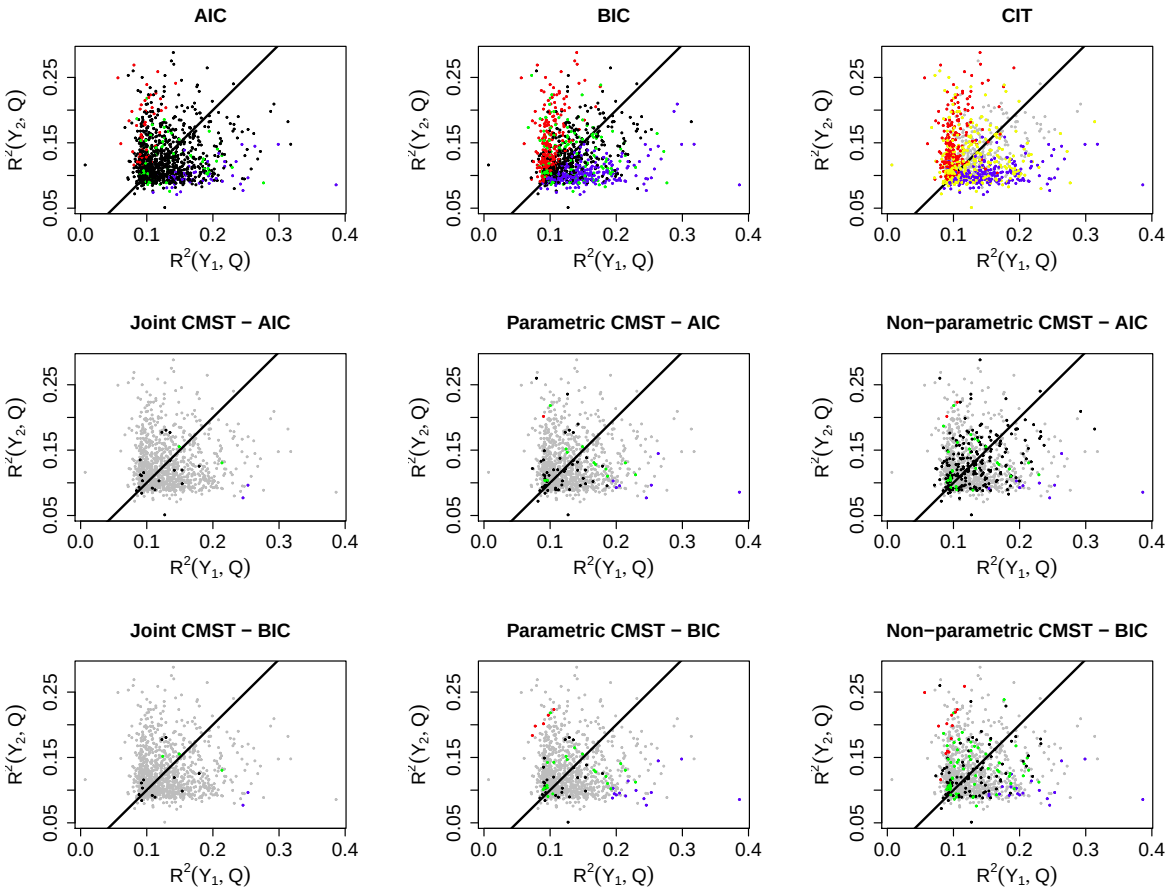

Figure S 8: Simulation results for Model E in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the AIC, BIC and CMST methods, black dots represent true positives, and blue, red and green dots represent false positives. For the CIT test, yellow dots represent true positives and blue and red dots show false positives.

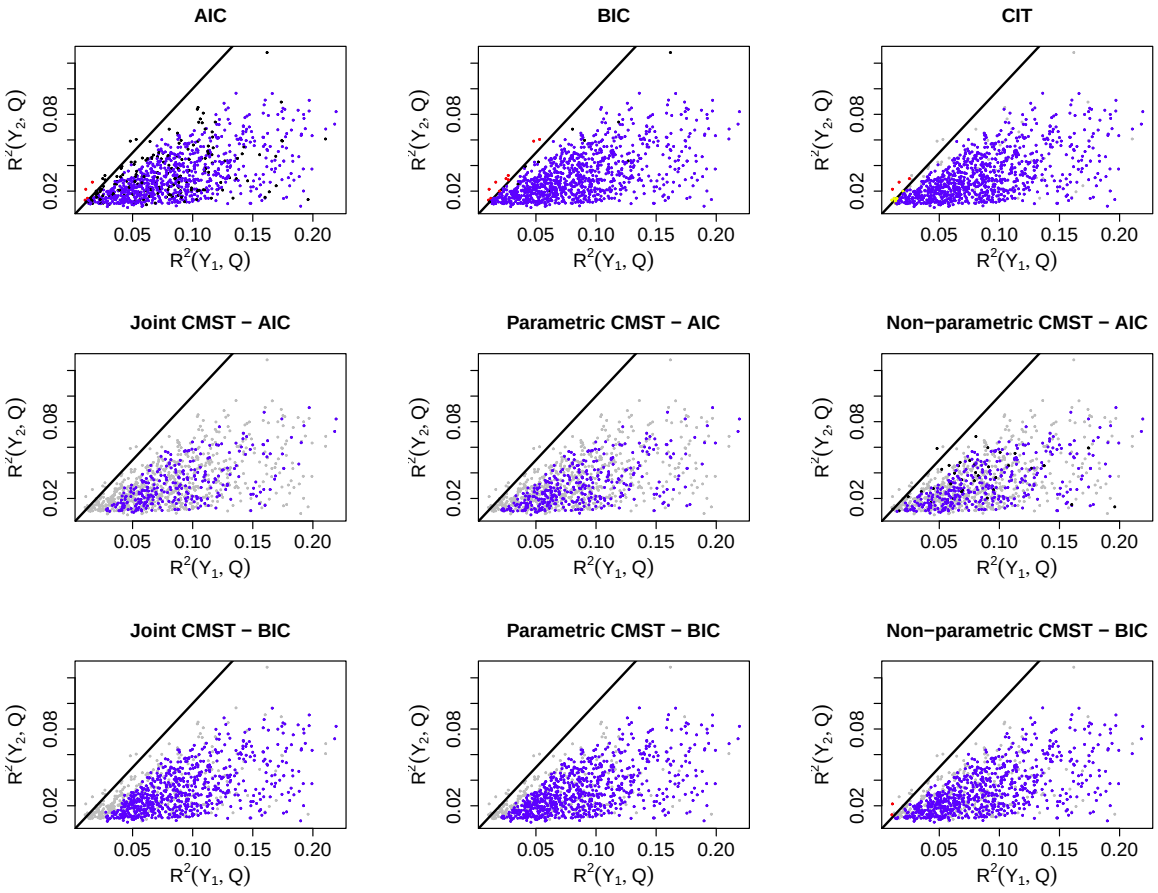

Figure S 9: Simulation results for Model A in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this model, blue dots represent true positives. Red, green and black dots represent false positives for the AIC, BIC and CMST methods. Red and yellow dots represent false positives for the CIT.

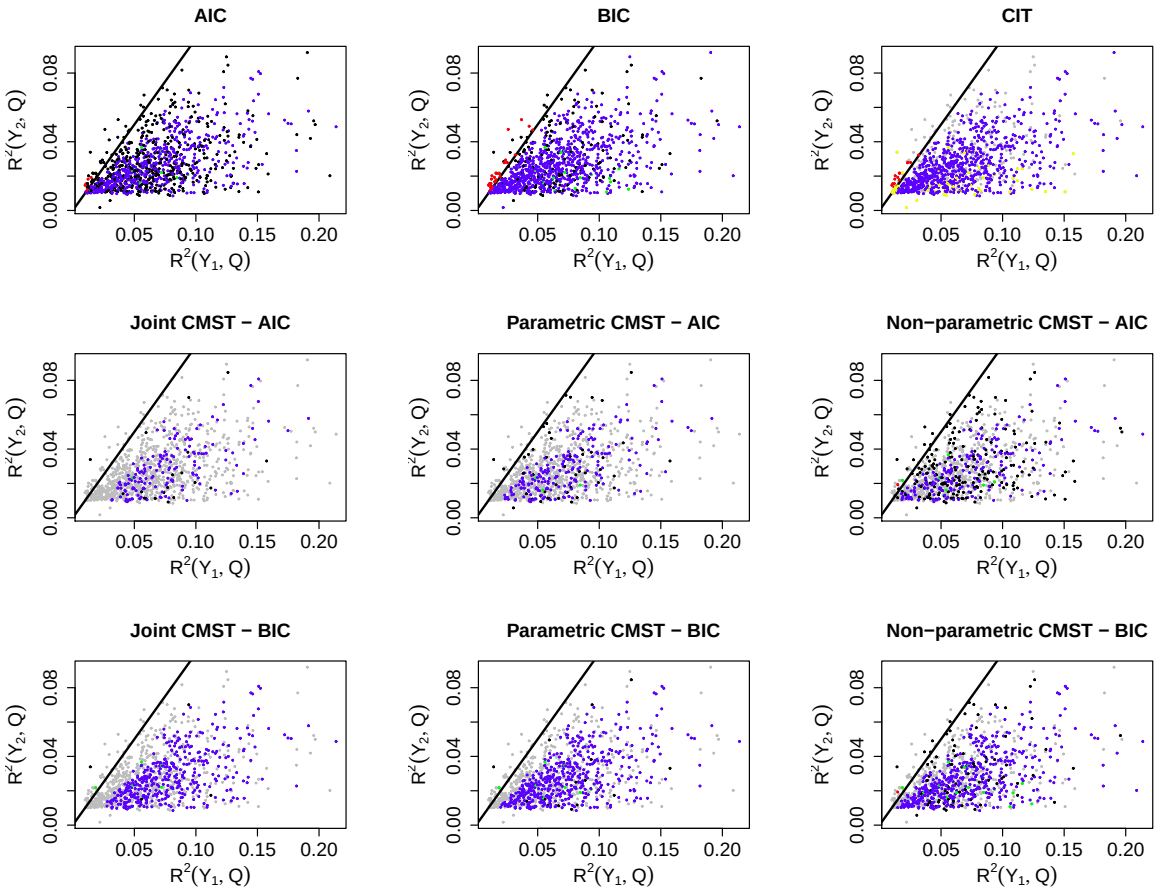

Figure S 10: Simulation results for Model B in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this model, blue dots represent true positives. Red, green and black dots represent false positives for the AIC, BIC and CMST methods. Red and yellow dots represent false positives for the CIT.

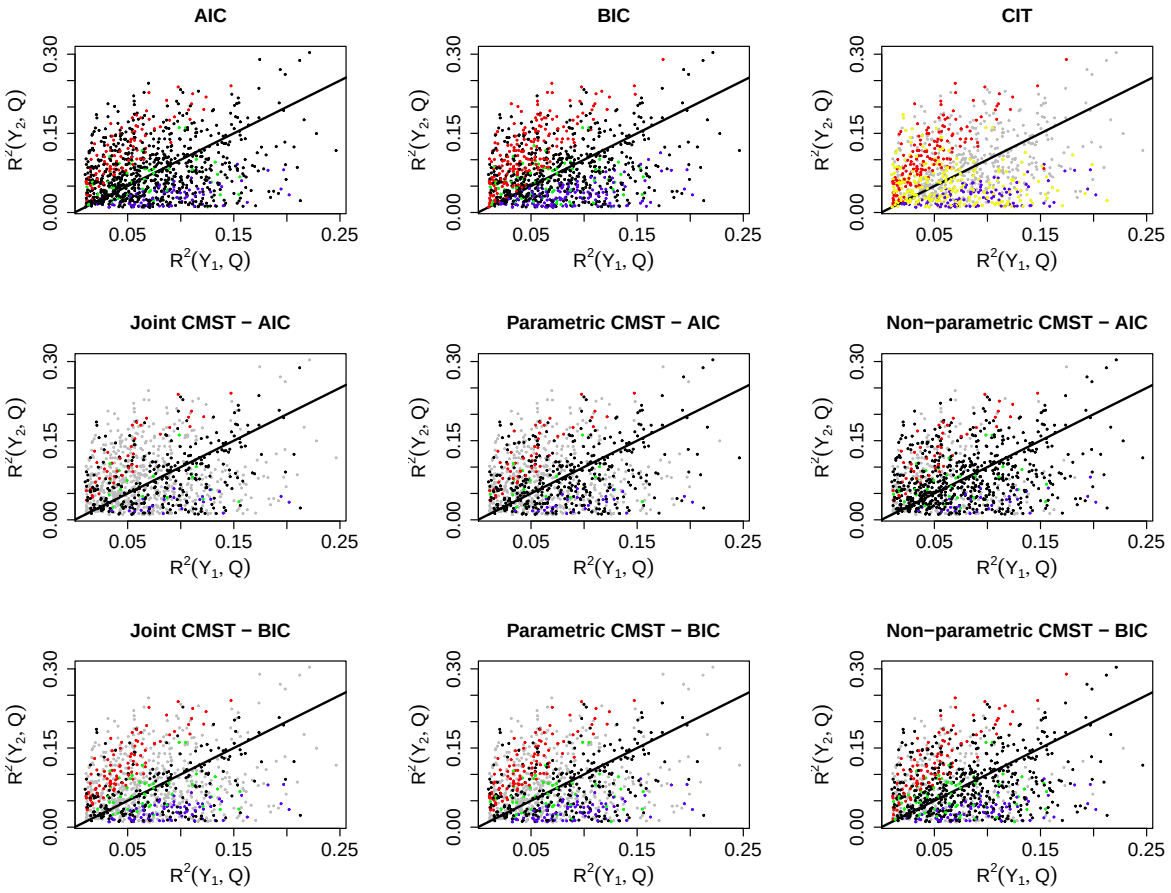

Figure S 11: Simulation results for Model C in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the AIC, BIC and CMST methods, black dots represent true positives, and blue, red and green dots represent false positives. For the CIT test, yellow dots represent true positives and blue and red dots show false positives.

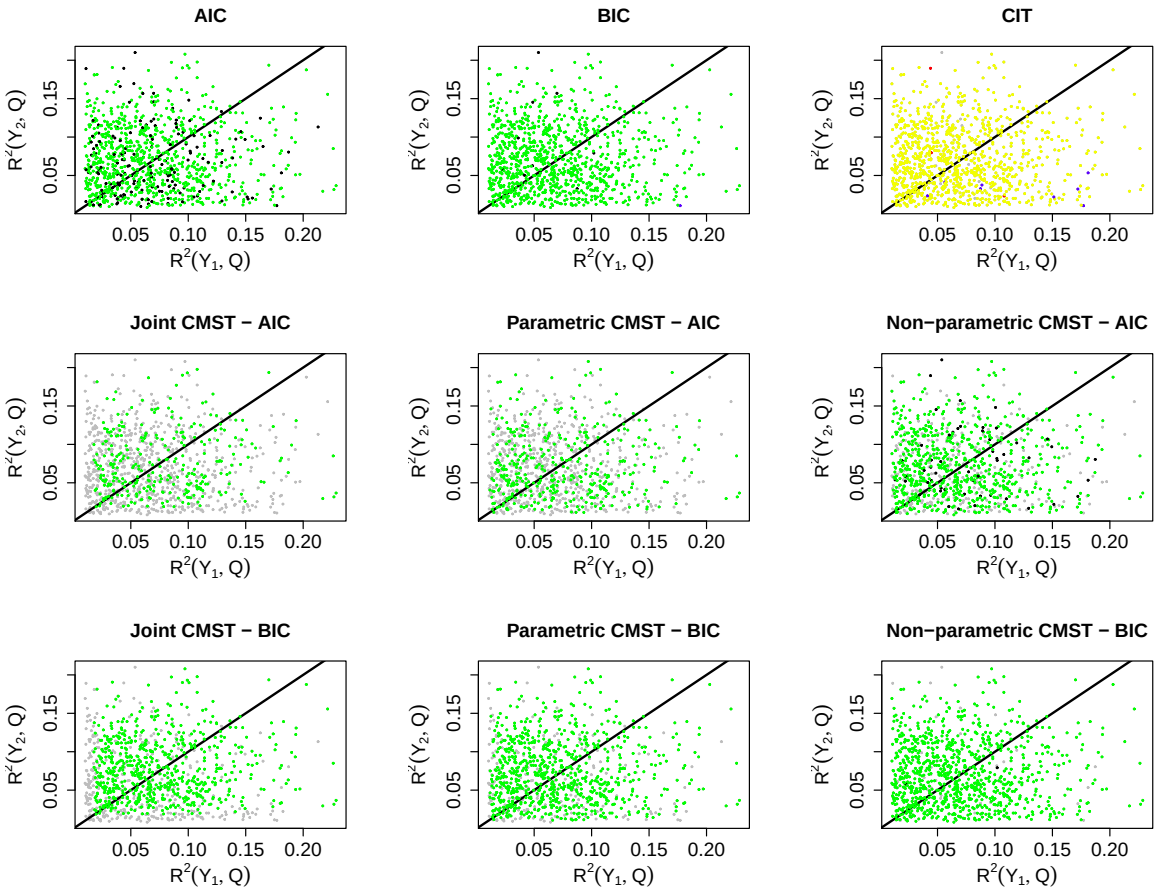

Figure S 12: Simulation results for Model D in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the AIC, BIC and CMST methods, green dots represent true positives, and blue, red and black dots represent false positives. For the CIT test, yellow dots represent true positives and blue and red dots show false positives.

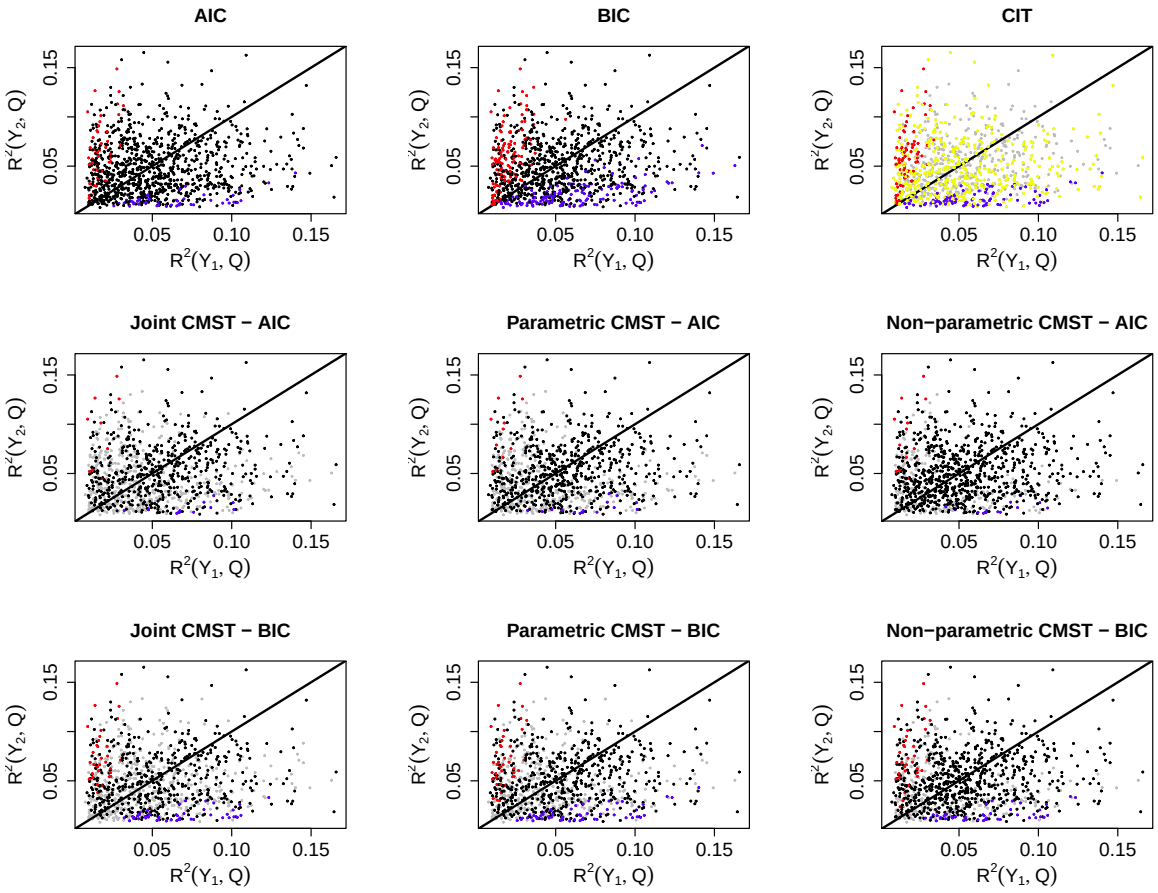

Figure S 13: Simulation results for Model E in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the AIC, BIC and CMST methods, black dots represent true positives, and blue, red and green dots represent false positives. For the CIT test, yellow dots represent true positives and blue and red dots show false positives.

Co-mapping QTL selection

Often times the phenotypes map to nearby but not precisely the same QTL, and one needs to decide which QTL to use as the causal anchor. When testing expression traits against clinical traits, Millstein et al. (2009) and Schadt et al. (2005) suggest using the clinical trait QTL as the anchor.

We adopt a different approach. When the phenotypes map to distinct regions that are less than 2cM apart we determine the QTL position using both phenotypes, jointly, as follows. For each pair of phenotypes (Y_1,Y_2) we perform unconditional mapping analysis for Y_1 and Y_2 and conditional mapping analysis for Y_2 given Y_1. Let LOD_1 represent a LOD score for the mapping analysis of Y_1, and $LOD_{2|1}$ for the mapping analysis of Y_2 given Y_1. Since

$$\log_{10}\left\{\frac{f(y_1, y_2 \mid q)}{f(y_1, y_2)}\right\} = \log_{10}\left\{\frac{f(y_1 \mid q)}{f(y_1)}\right\} + \log_{10}\left\{\frac{f(y_2 \mid y_1, q)}{f(y_2 \mid y_1)}\right\}, \tag{15}$$

we compute the joint LOD score of (Y_1,Y_2) as $LOD_{1,2} = LOD_1 + LOD_{2|1}$ (or equivalently as $LOD_{1,2} = LOD_2 + LOD_{1|2}$). We determine the peak QTL position, λ, using the $LOD_{1,2}$ scores profile and assign the QTL to Y_1 and Y_2 if LOD_1 and LOD_2 are greater than the mapping threshold at the λ position. Figure S14 illustrates our approach. When both phenotypes co-map to more than one QTL we select the QTL with the highest joint mapping peak.

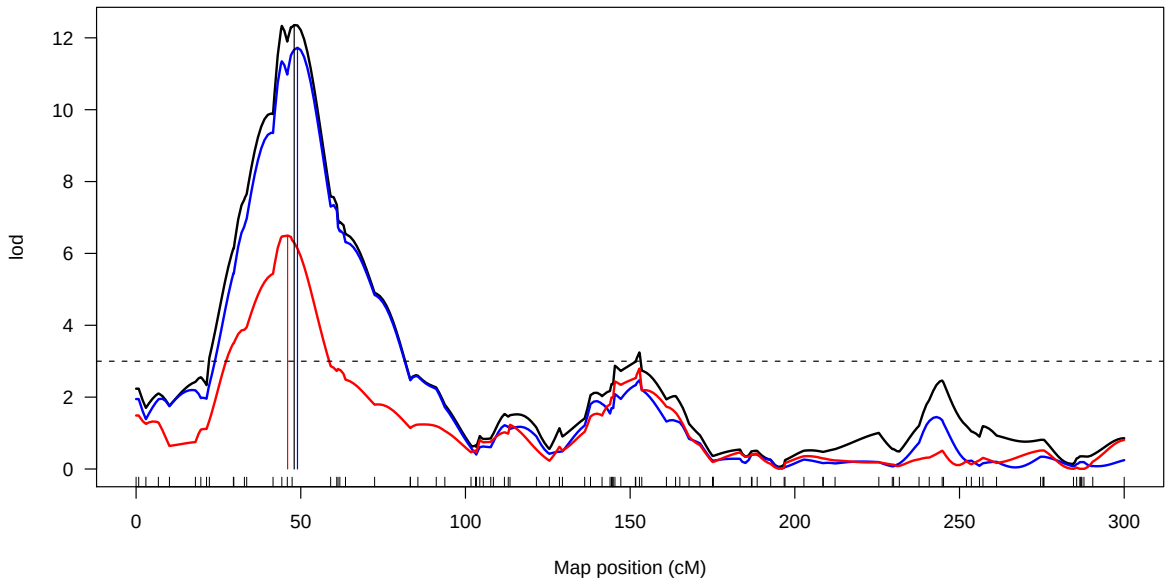

Figure S 14: We simulated data from a model $Q \rightarrow Y_1 \rightarrow Y_2$, with a QTL, Q, at 50cM. The blue and red curves show the (unconditional) LOD profiles of phenotypes Y_1 and Y_2, respectively. The black curve depicts the joint LOD curve, and the peak QTL position λ is given by the black vertical line. Instead of having to perform an arbitrary choice between the QTLs given by the red and blue vertical lines we use the QTL given by the black line. The dashed line shows the QTL mapping threshold.

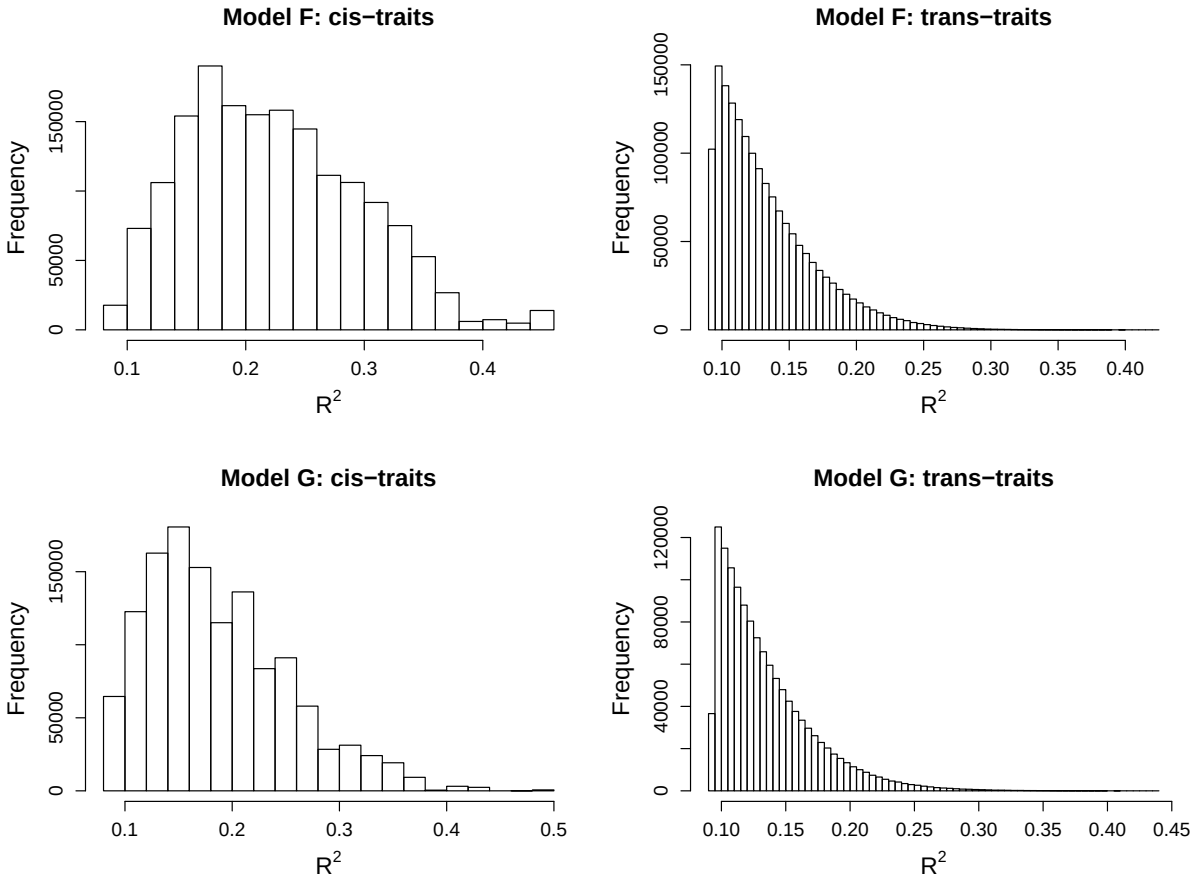

Figure S 15: Overall R^2 statistics distributions for the large scale simulation study. The left and right panels show the distribution for the *cis*-traits and *trans*-traits, respectively.

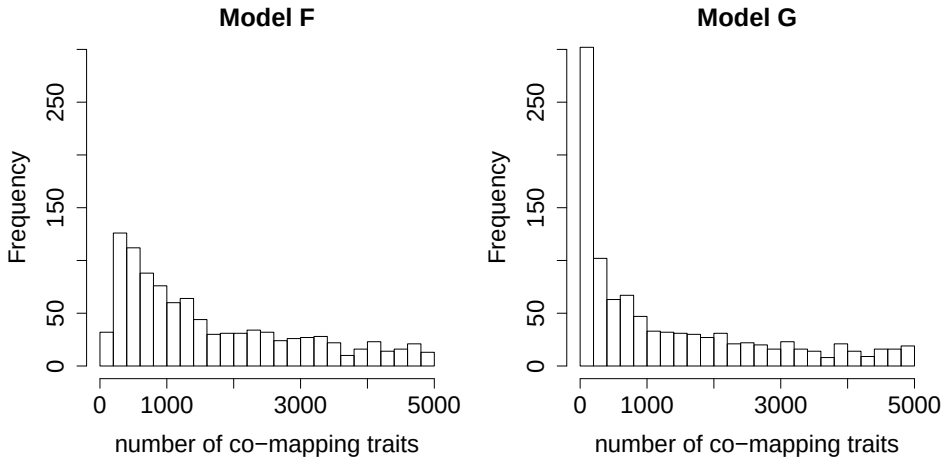

Figure S 16: For each model F and G we performed 1,000 separate simulations, and tested Y_1 against all other phenotypes Y_k, $k = 2, \ldots, 5001$, that shared the QTL with Y_1, at each simulation. The panels show the distribution of the number of tests, i.e, the number of trans-traits that co-mapped to Y_1, per simulation study. In total, we performed 1,656,261 tests across the 1,000 simulations with model F, and 1,286,243 tests across the simulations with model G.

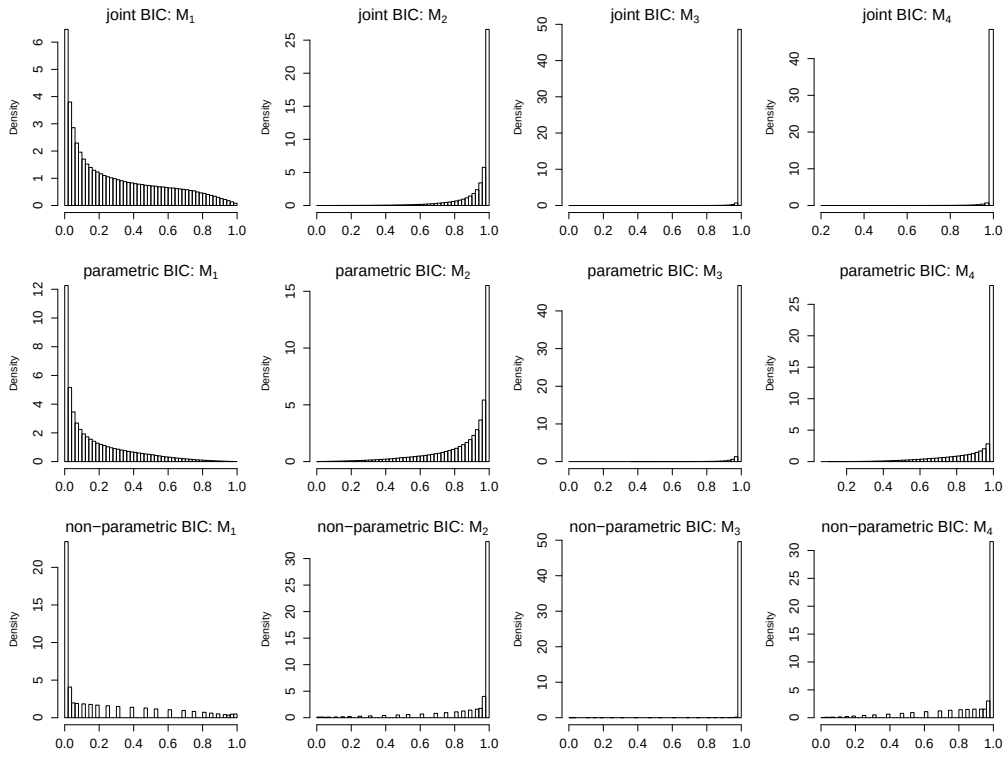

Figure S 17: Uncorrected p-value distributions for the BIC-based CMST tests with data simulated from model F in Figure 5. Results based on 1,656,261 tests. For these simulations, the M_1 call is the correct one, hence the skewed distribution towards small p-values at the left panels. The skewness towards larger p-values for the M_2, M_3, and M_4 calls follows from the fact that whenever a p-value for one model is smaller than α, then the p-values for the other three models are greater than $1-\alpha$. Note the larger frequency of small M_1 p-values for the non-parametric CMST test (bottom left panel - the discrete nature of the histogram is a consequence of the test statistic being discrete for the non-parametric test).

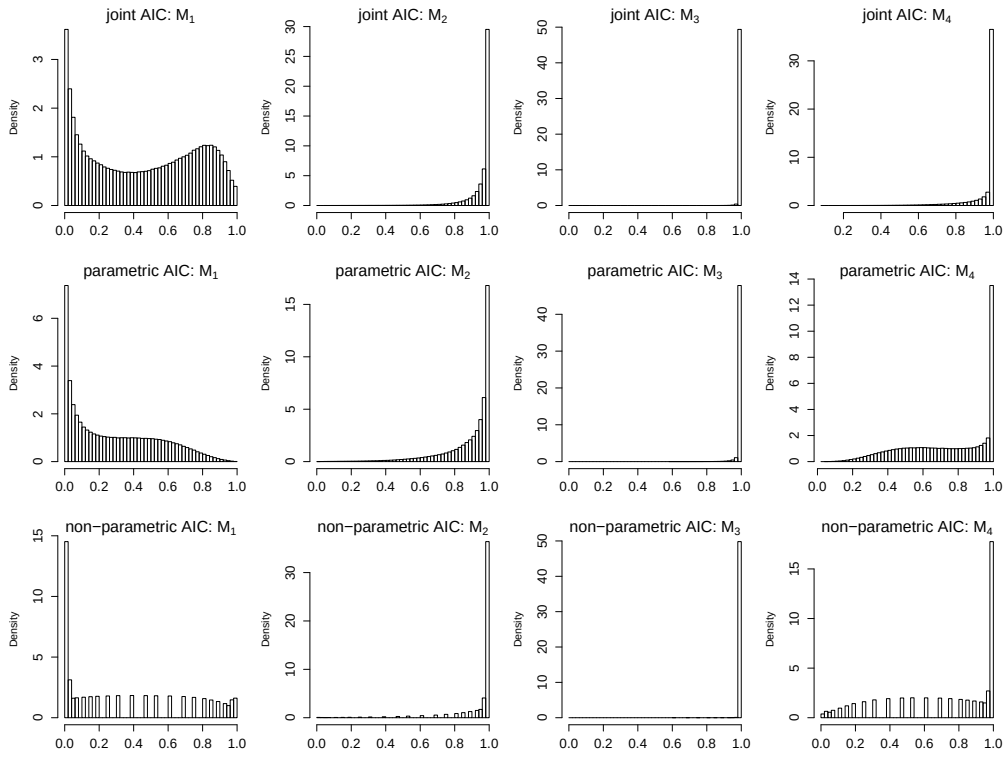

Figure S 18: Uncorrected p-value distributions for the AIC-based CMST tests with data simulated from model F in Figure 5. Results based on 1,656,261 tests. For these simulations, the M_1 call is the correct one, hence the skewed distribution towards small p-values at the left panels. The skewness towards larger p-values for the M_2, M_3, and M_4 calls follows from the fact that whenever a p-value for one model is smaller than α, then the p-values for the other three models are greater than $1-\alpha$. Note the larger frequency of small M_1 p-values for the non-parametric CMST test (bottom left panel - the discrete nature of the histogram is a consequence of the test statistic being discrete for the non-parametric test).

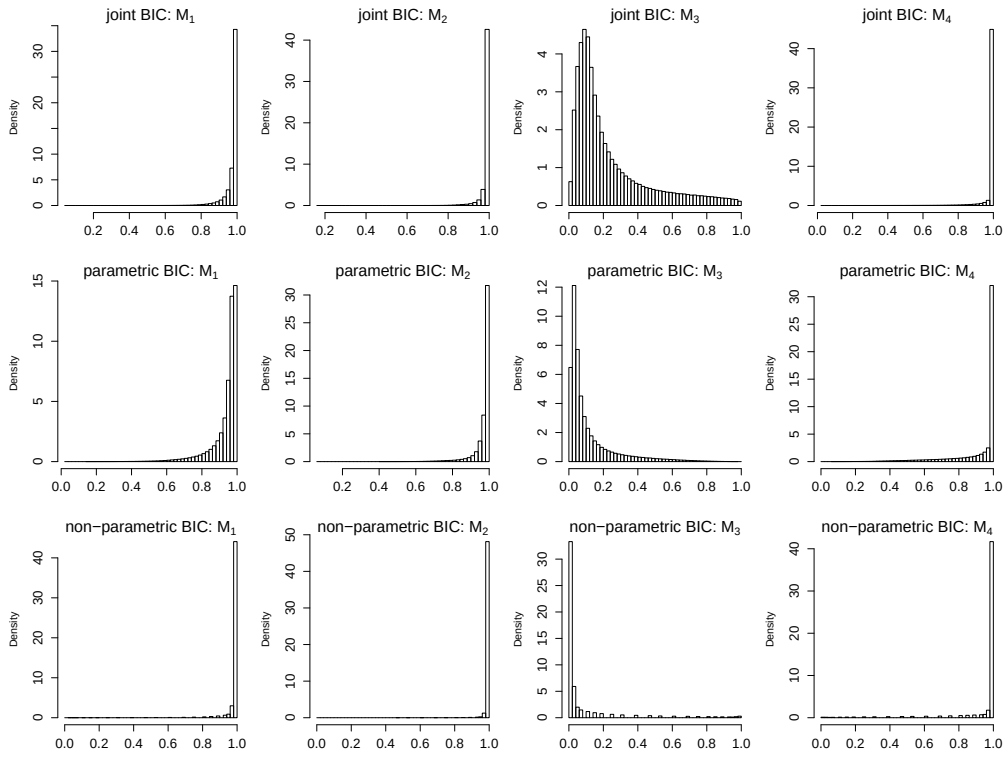

Figure S 19: Uncorrected p-value distributions for the BIC-based CMST tests with data simulated from model G in Figure 5. Results based on 1,286,243 tests. For these simulations, the M_3 call is the correct one, hence the skewed distribution towards small p-values at the M_3 panels. The skewness towards larger p-values for the M_1, M_2, and M_4 calls follows from the fact that whenever a p-value for one model is smaller than α, then the p-values for the other three models are greater than $1-\alpha$. Note the larger frequency of small M_3 p-values for the non-parametric CMST test (bottom left panel - the discrete nature of the histogram is a consequence of the test statistic being discrete for the non-parametric test).

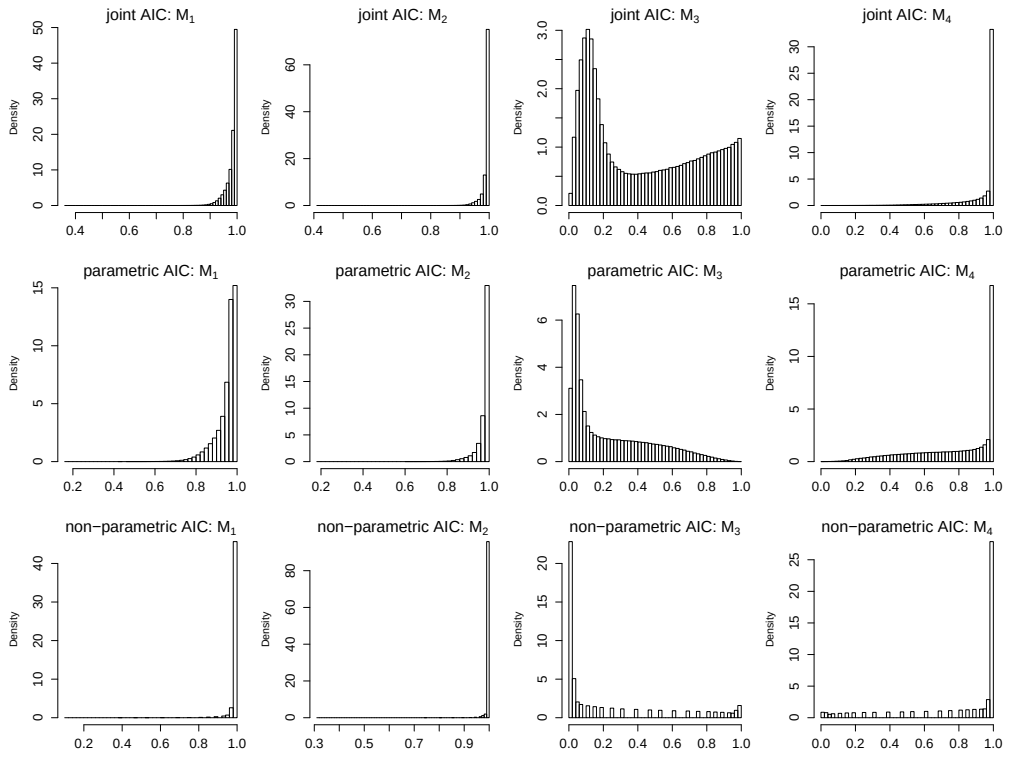

Figure S 20: Uncorrected p-value distributions for the AIC-based CMST tests with data simulated from model G in Figure 5. Results based on 1,286,243 tests. For these simulations, the M_3 call is the correct one, hence the skewed distribution towards small p-values at the M_3 panels. The skewness towards larger p-values for the M_1, M_2, and M_4 calls follows from the fact that whenever a p-value for one model is smaller than α, then the p-values for the other three models are greater than $1-\alpha$. Note the larger frequency of small M_3 p-values for the non-parametric CMST test (bottom left panel - the discrete nature of the histogram is a consequence of the test statistic being discrete for the non-parametric test).