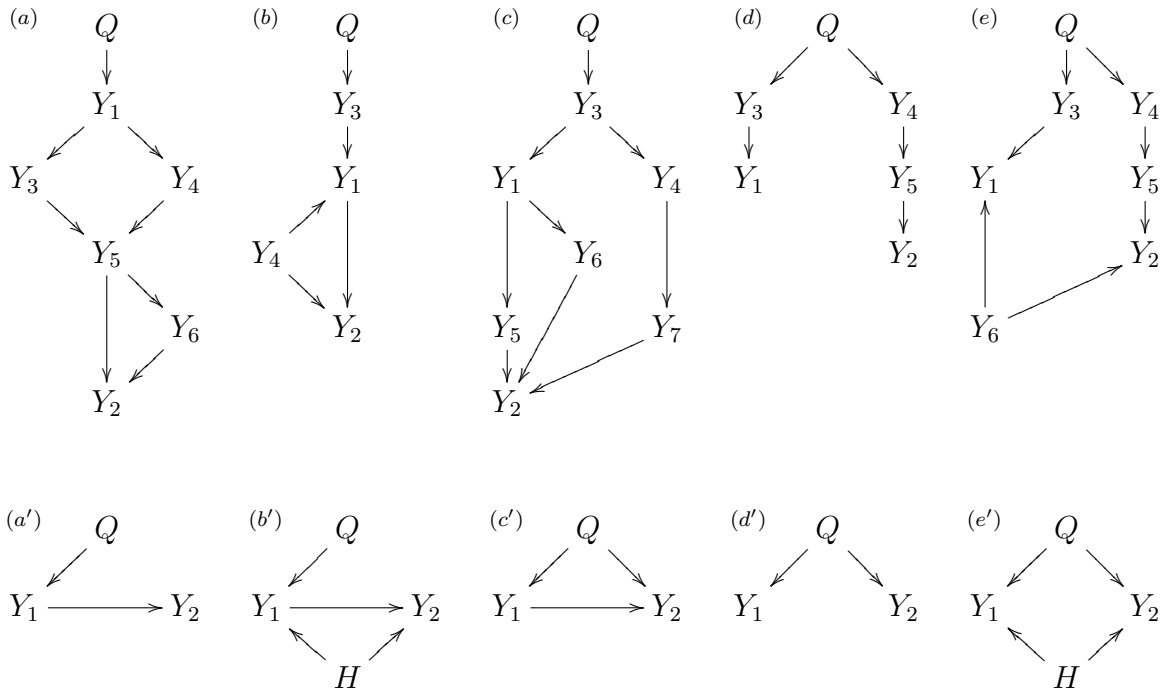

Figure S 1 Network models and their collapsed versions. The collapsed networks (bottom panels) represent simplified versions of the true networks (top panels), where nodes other than Q, Y_1 and Y_2 are ignored, even though they still represent the correct causal flow among these three nodes in the true 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'.

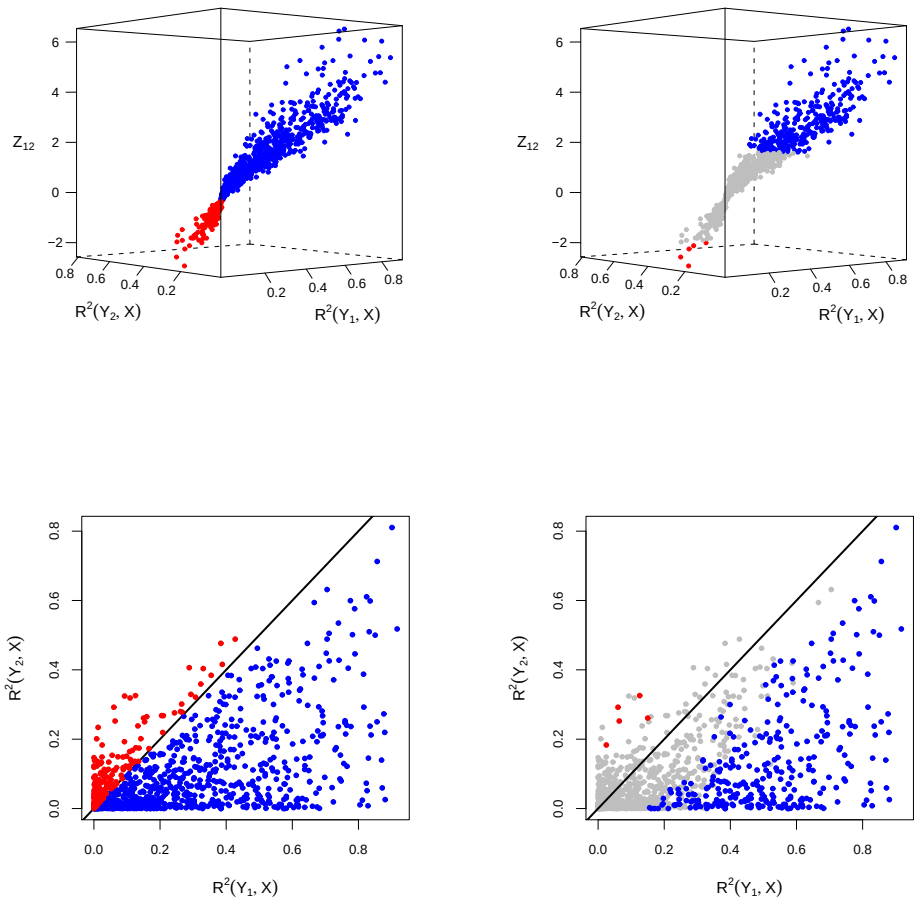

Figure S2 Model selection via log-likelihood ratio versus Vuong's test.

Figure S2 illustrates how Vuong's test works. We generated 1,000 data-sets from the model $X \to Y_1 \to Y_2$ and applied Vuong's test to the comparison of models $M_1 : X \to Y_1 \to Y_2$ against $M_2 : X \to Y_2 \to Y_1$. 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". Note that because model M_1 corresponds to the true model, we have that the a M_1 call is always correct, whereas

a M_2 call is always incorrect in this example. Therefore, 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 $X \rightarrow Y_1 \rightarrow Y_2$, 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 $Z_{12} \geq 1.64$ for 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 an increase in precision for a reduction in statistical power to detect true positives.

File S1

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.

File S2

Simulation studies

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

File S3

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 SI.2a, and the axis scales on Figures S3-S7) and between 0.01 to 0.20 for the simulations based on 1,000 subjects (see Figure SI.2b, and the axis scales on Figures S8-S12).

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.

Often times the phenotypes map to nearby but not precisely the same QTL, and we need 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

$$\log_{10}\left\{\frac{f(y_1, y_2 \mid q)}{f(y_1, y_2)}\right\} = \log_{10}\left\{\frac{f(y_1 \mid q)}{f(y_1)}\right\} + \log_{10}\left\{\frac{f(y_2 \mid y_1, q)}{f(y_2 \mid y_1)}\right\}, \tag{1}$$

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 SI1 illustrates our approach. When both phenotypes co-map to more than one QTL we select the QTL with the highest joint mapping peak.

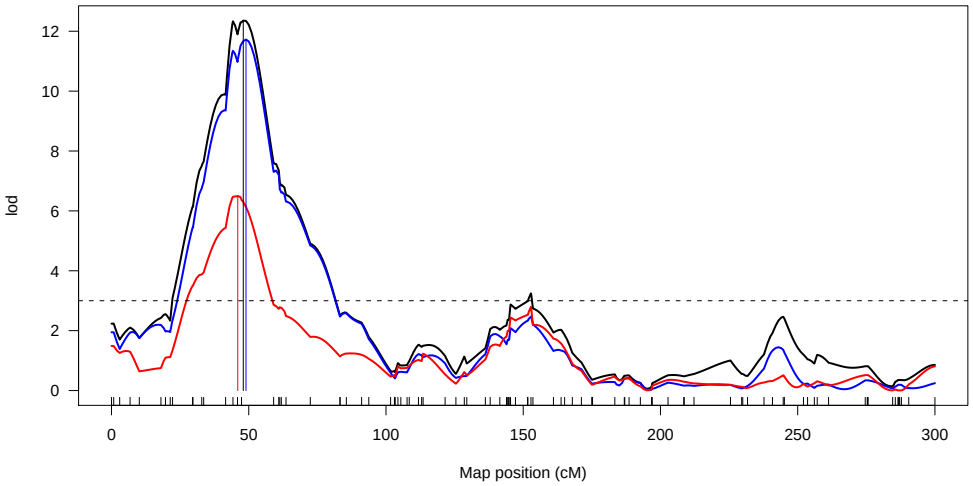

Figure SI.1 We simulated data from a model $Q \to Y_1 \to Y_2$, with a QTL, Q, at 50cM. The blue and red curves show the (unconditional) LOD profiles of phenotypes Y_1 and Y_2, respectively. The black curve depicts the joint LOD curve, and the peak QTL position λ is given by the black vertical line. 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.

File S4

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 SI.3 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 SI.4 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 generally advisable (Benjamini and Hochberg 1995). 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.

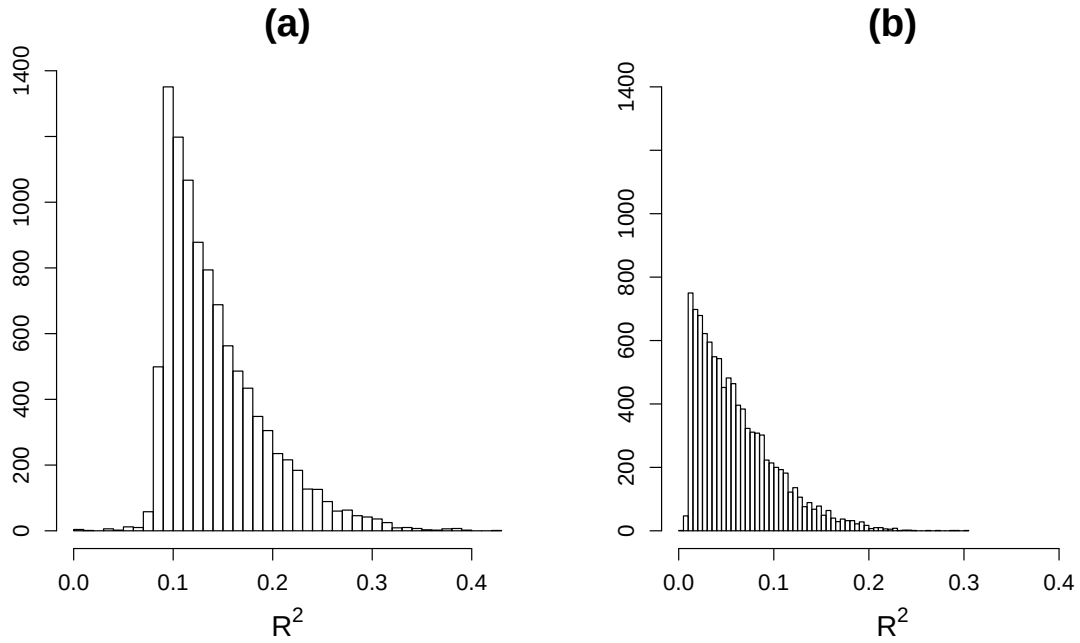

Figure SI.2 Overall distribution of the R^2 statistics across all simulated models in Figure 2. Panels a and b present the R^2 statistics for sample sizes 112 and 1,000, respectively.

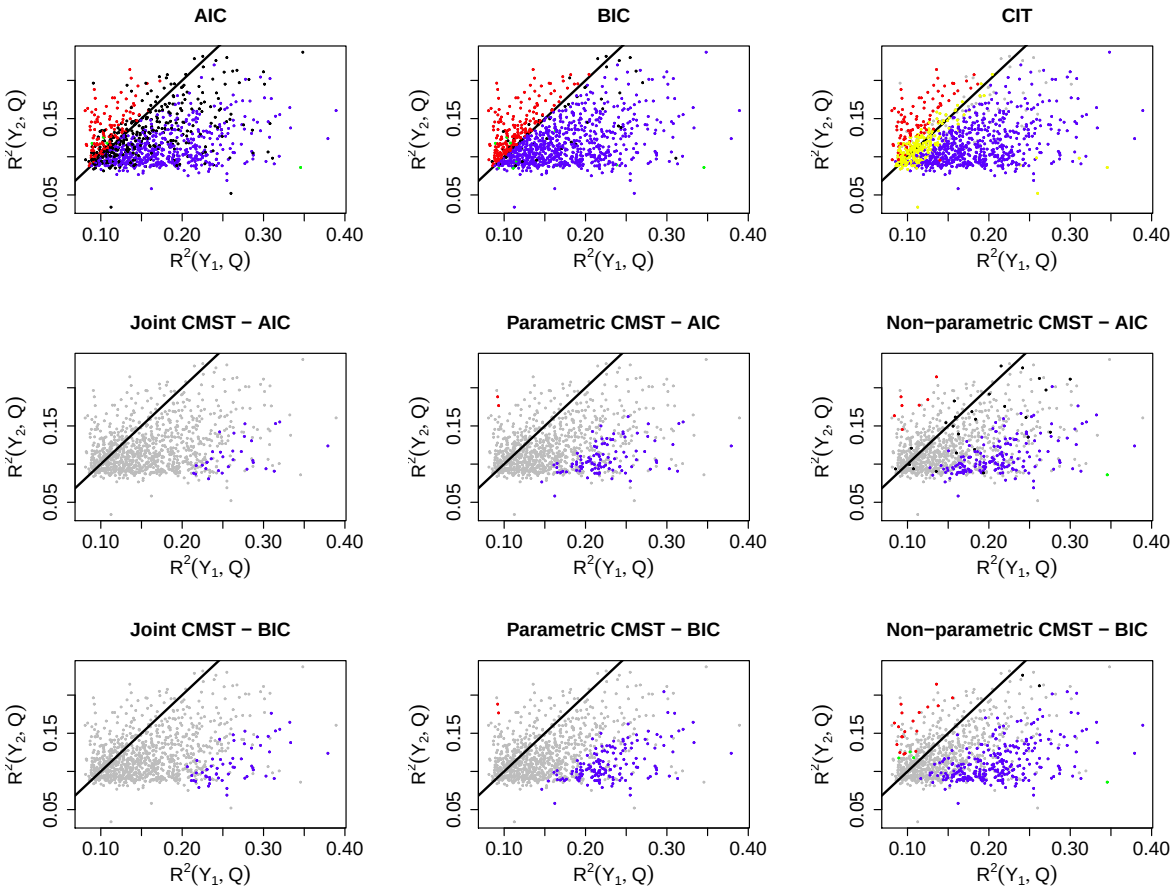

Figure S3 Simulation results for Model A in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this 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.

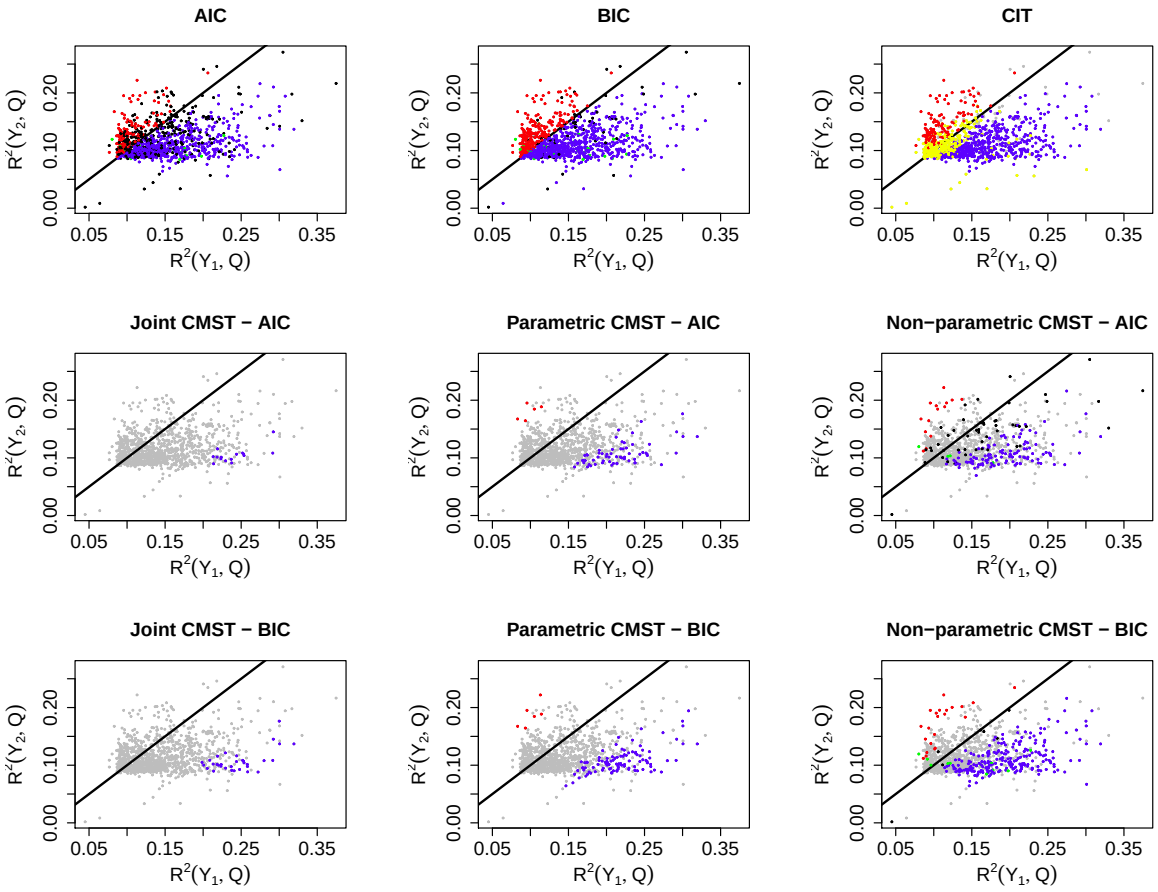

Figure S4 Simulation results for Model B in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this 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.

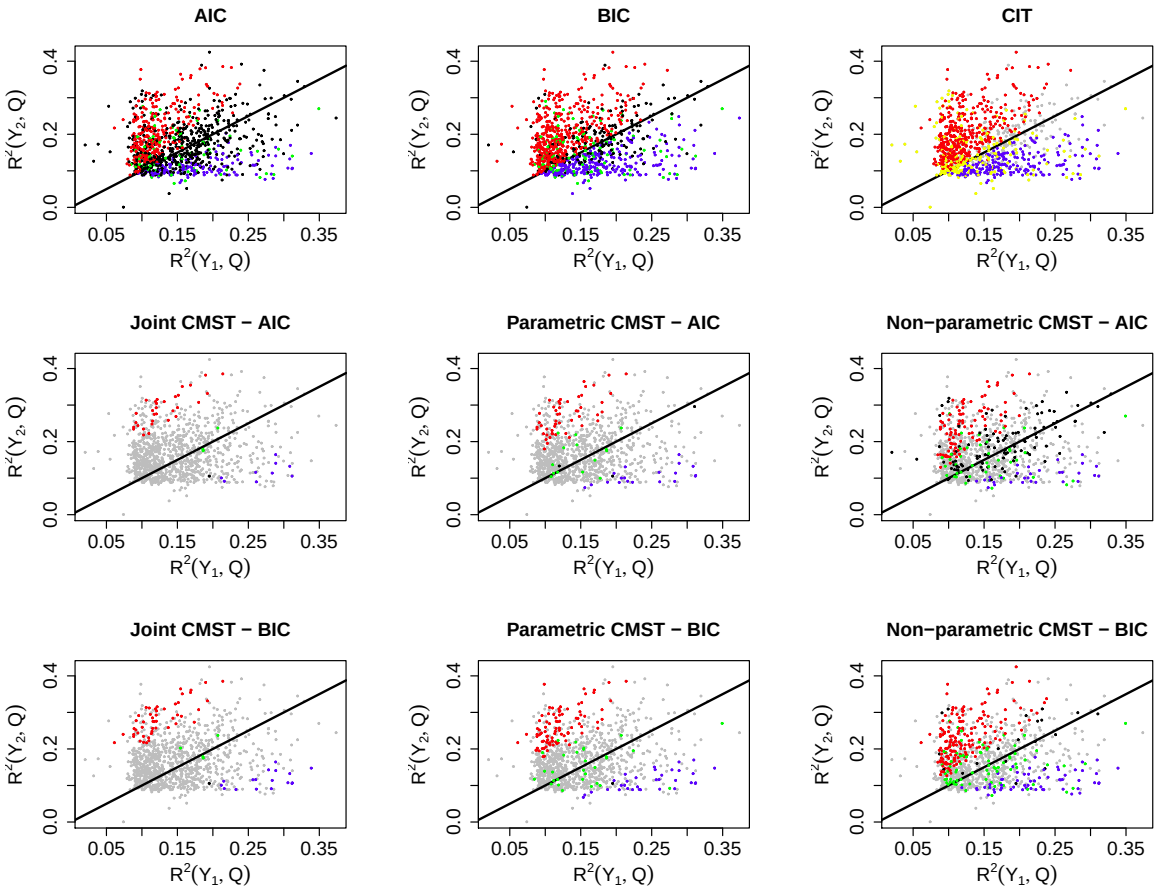

Figure S5 Simulation results for Model C in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the 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.

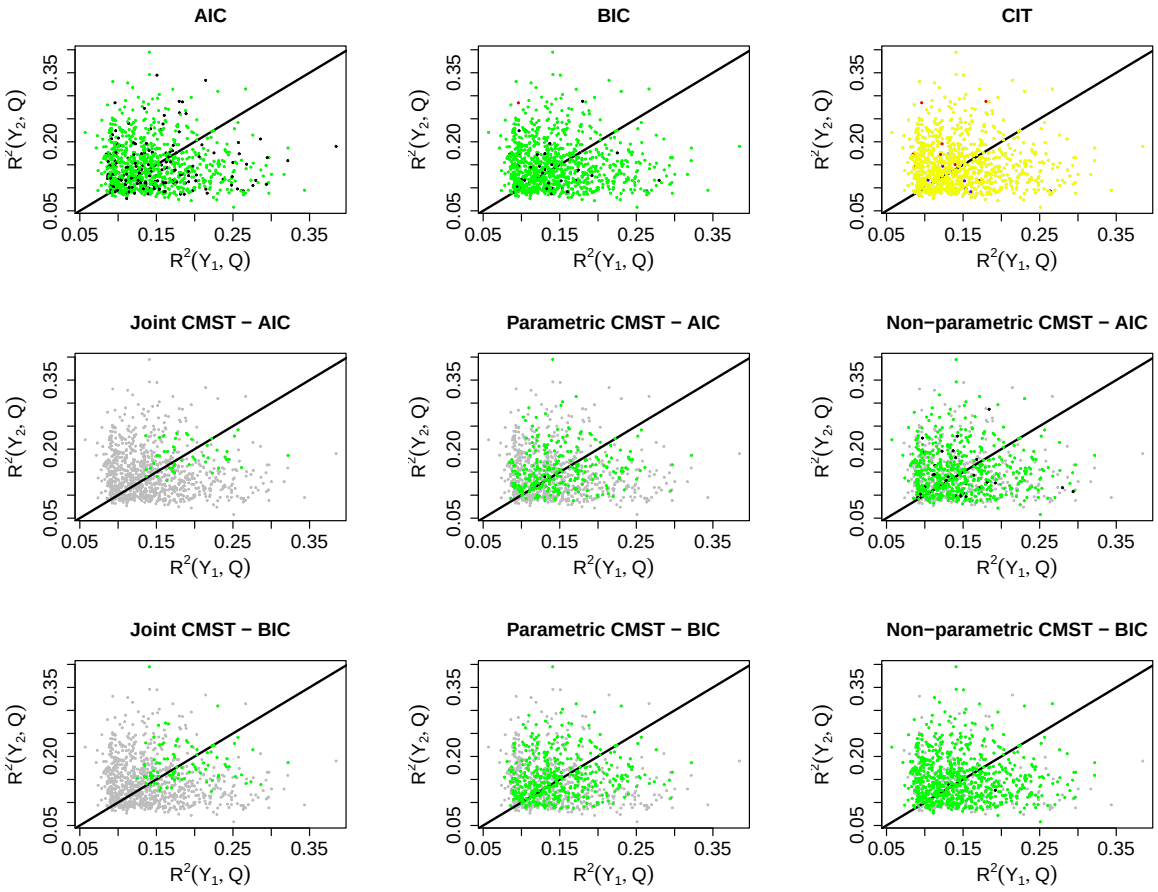

Figure S6 Simulation results for Model D in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the 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.

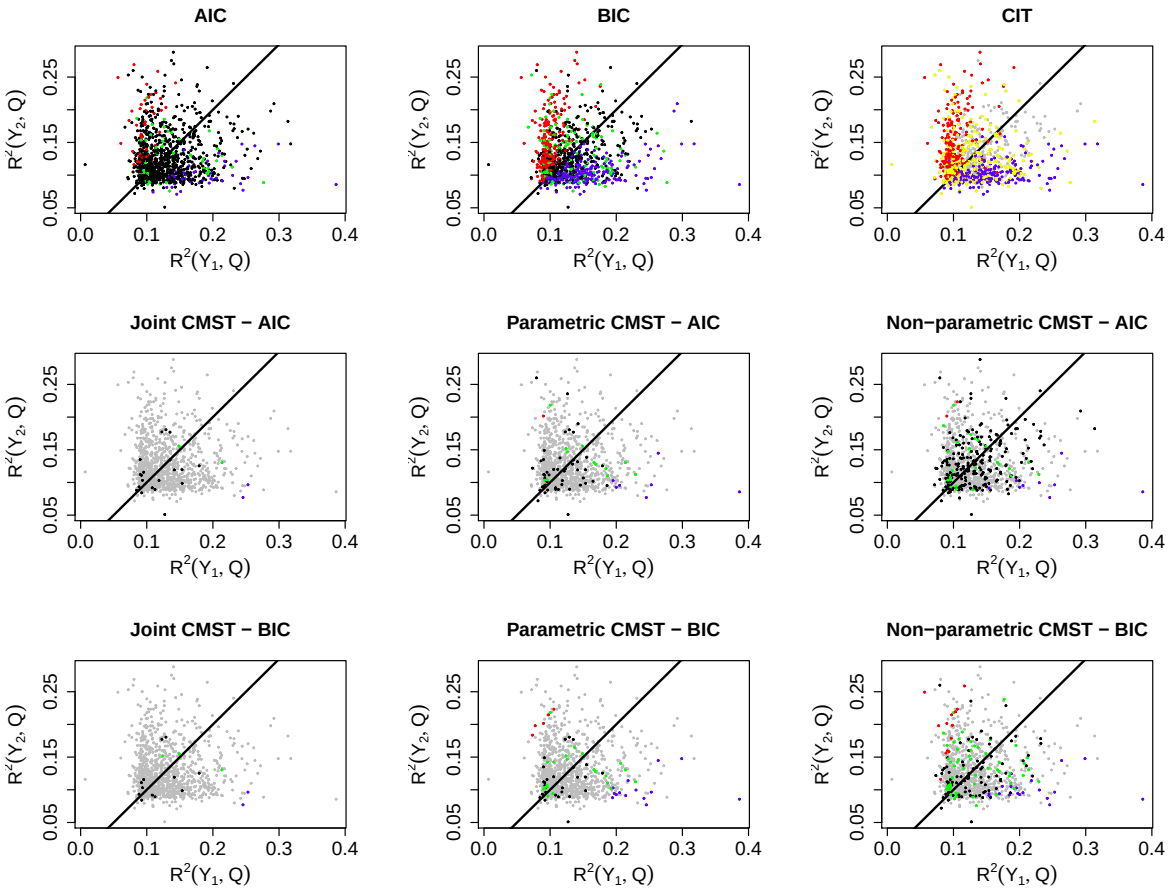

Figure S7 Simulation results for Model E in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the 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.

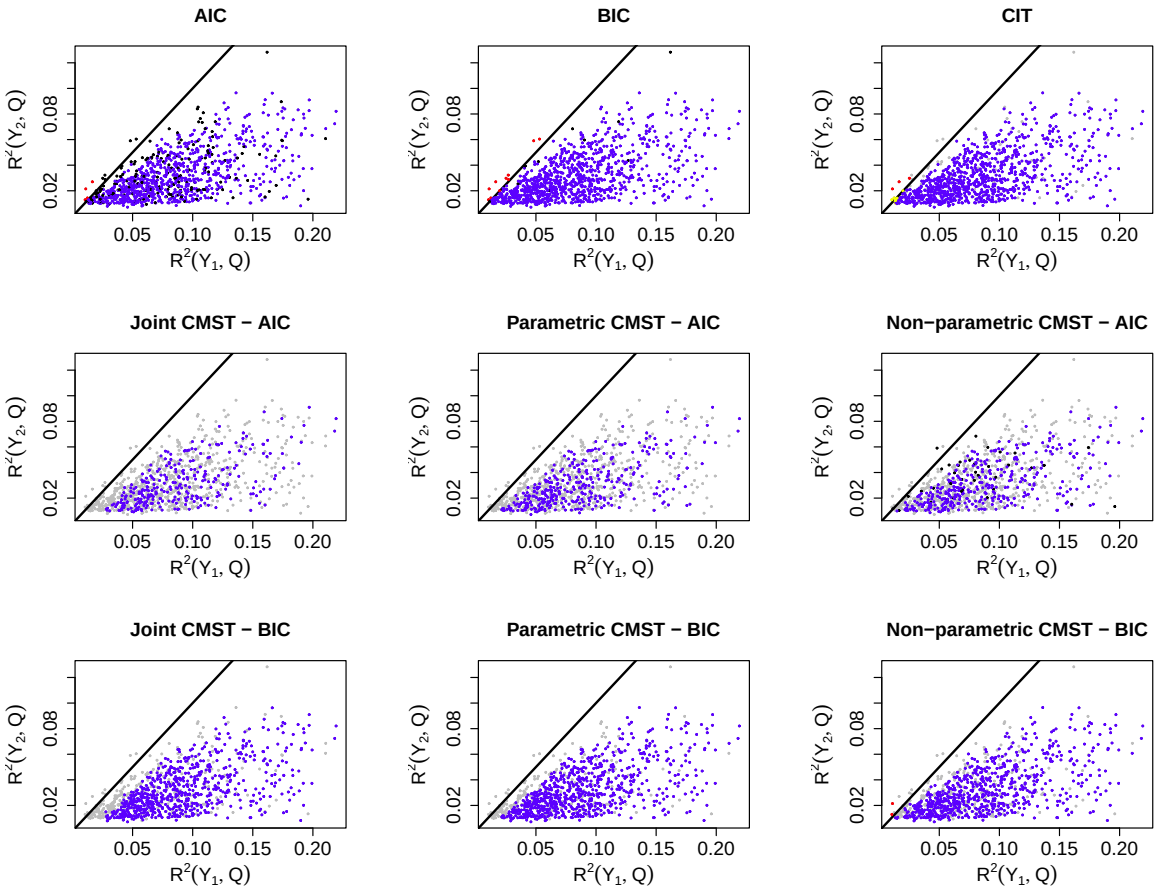

Figure S8 Simulation results for Model A in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this 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.

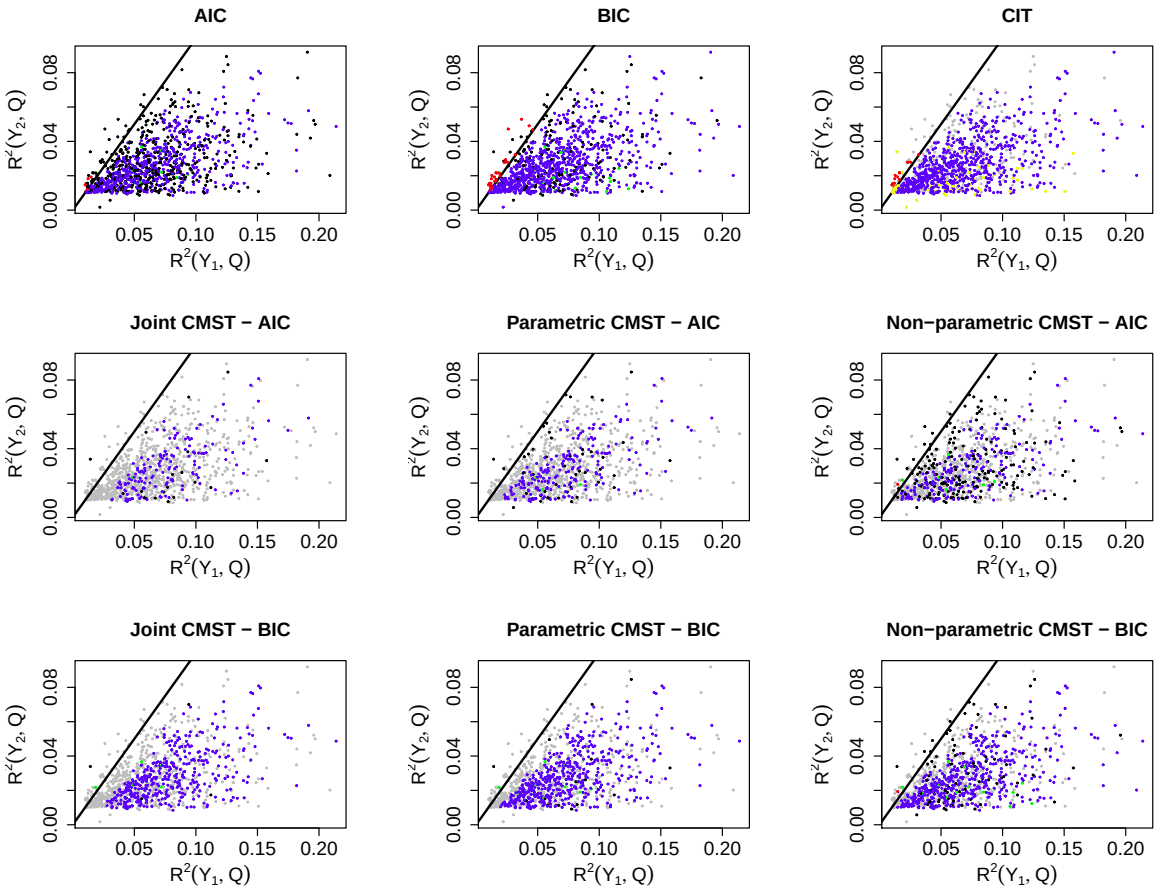

Figure S9 Simulation results for Model B in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this 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.

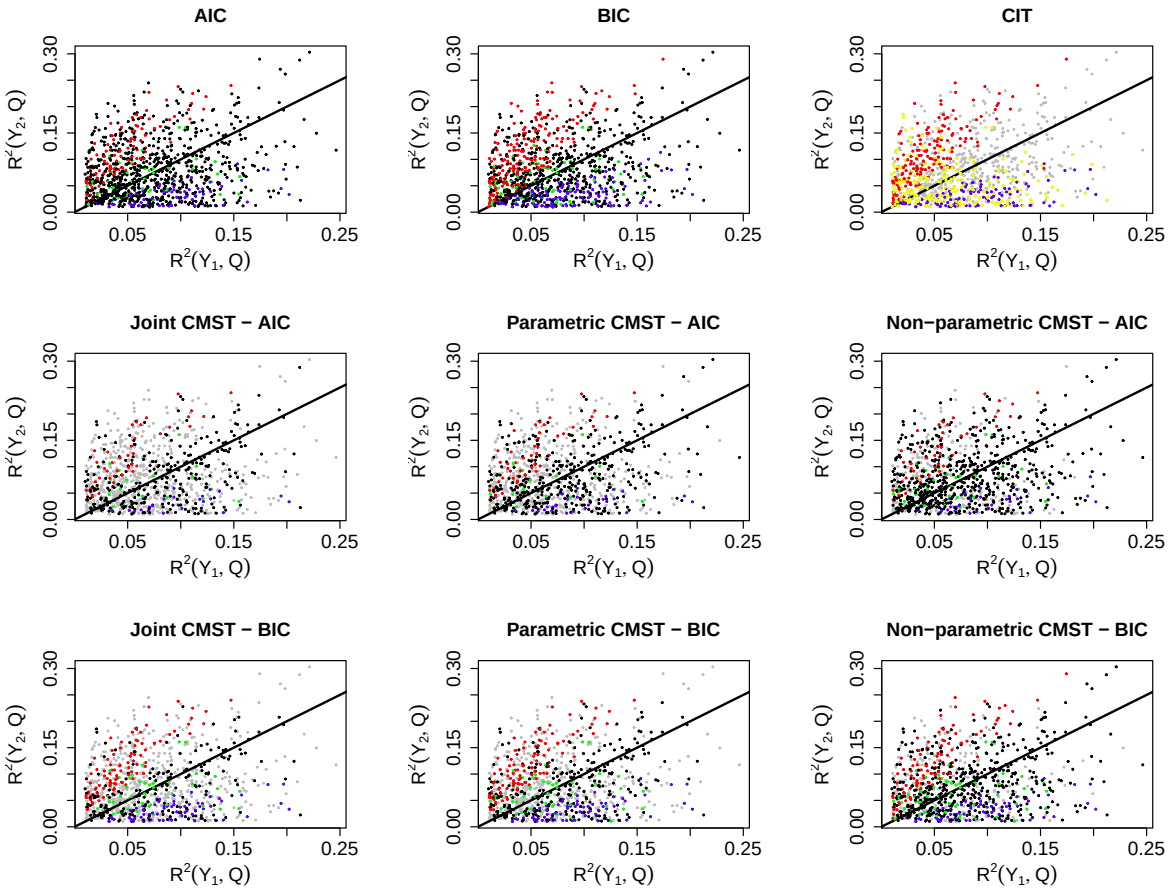

Figure S10 Simulation results for Model C in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the 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.

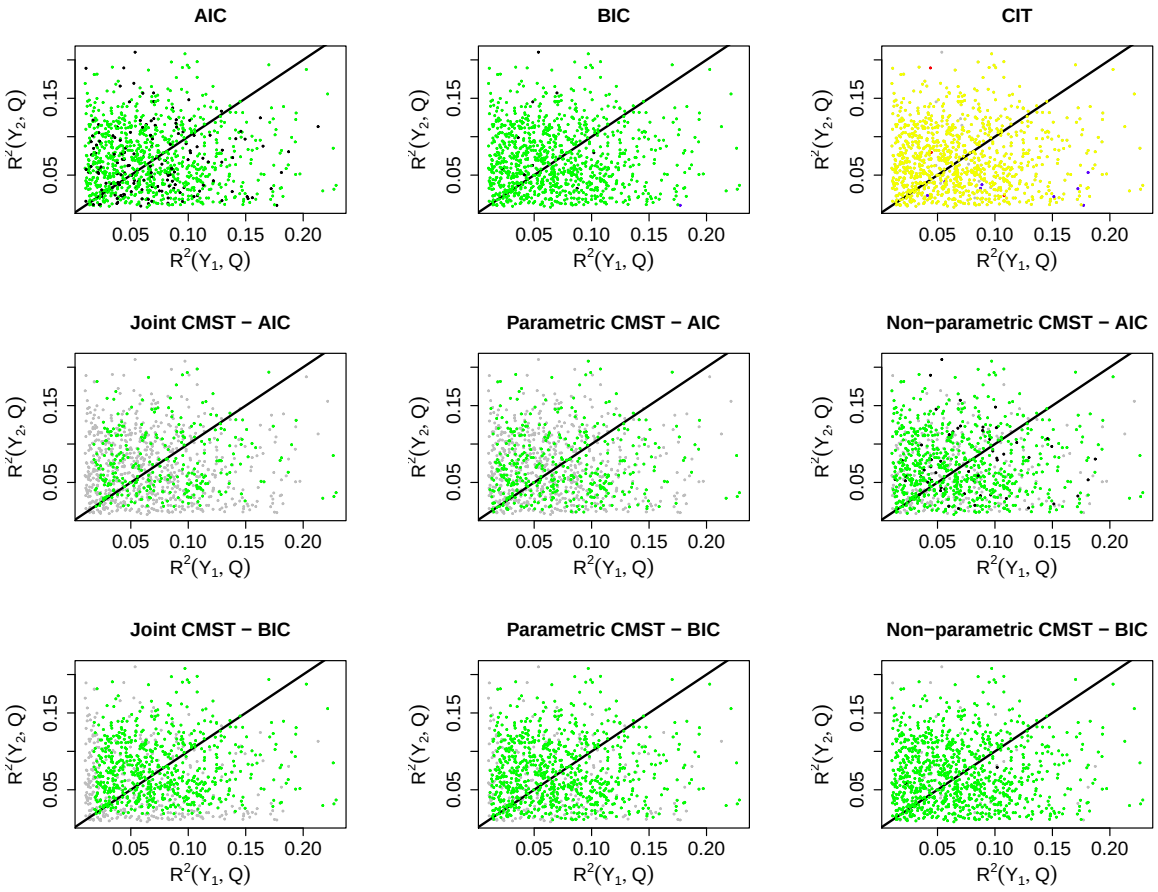

Figure S11 Simulation results for Model D in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the 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.

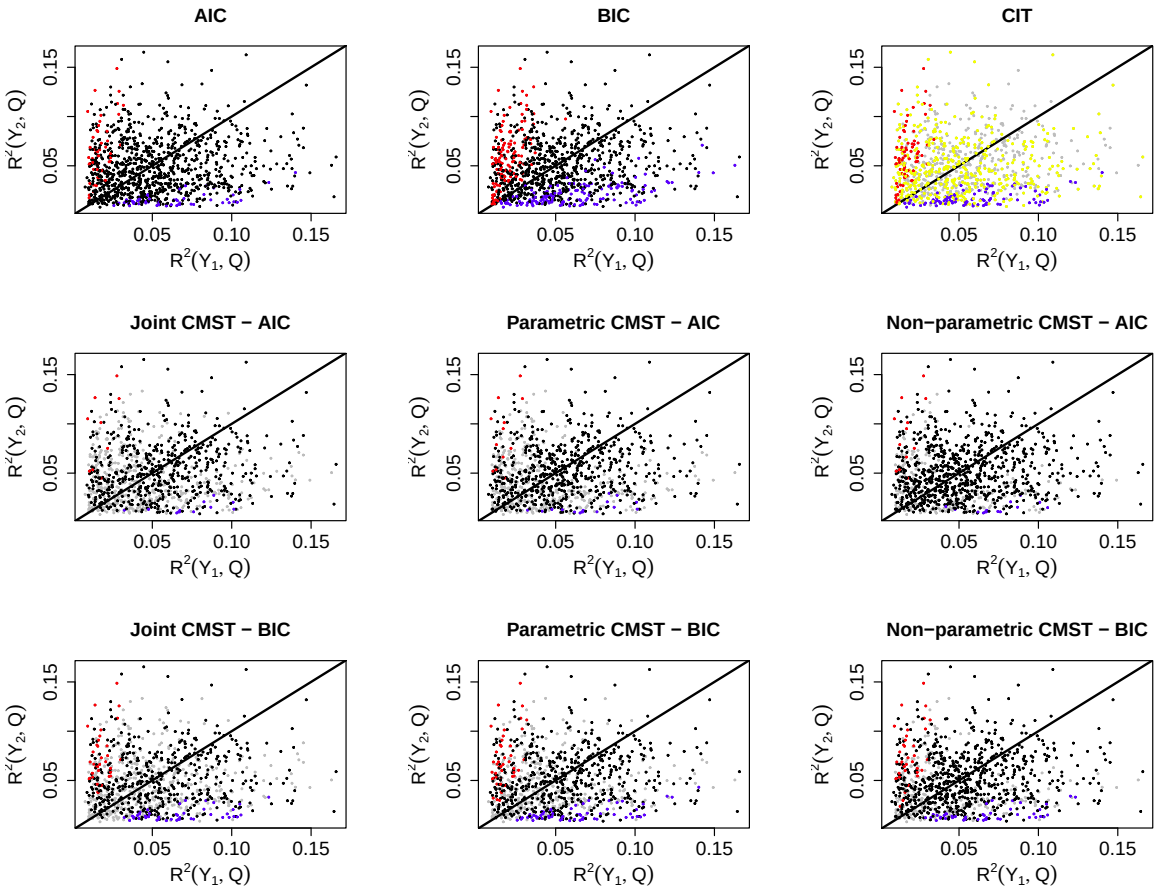

Figure S12 Simulation results for Model E in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the 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.

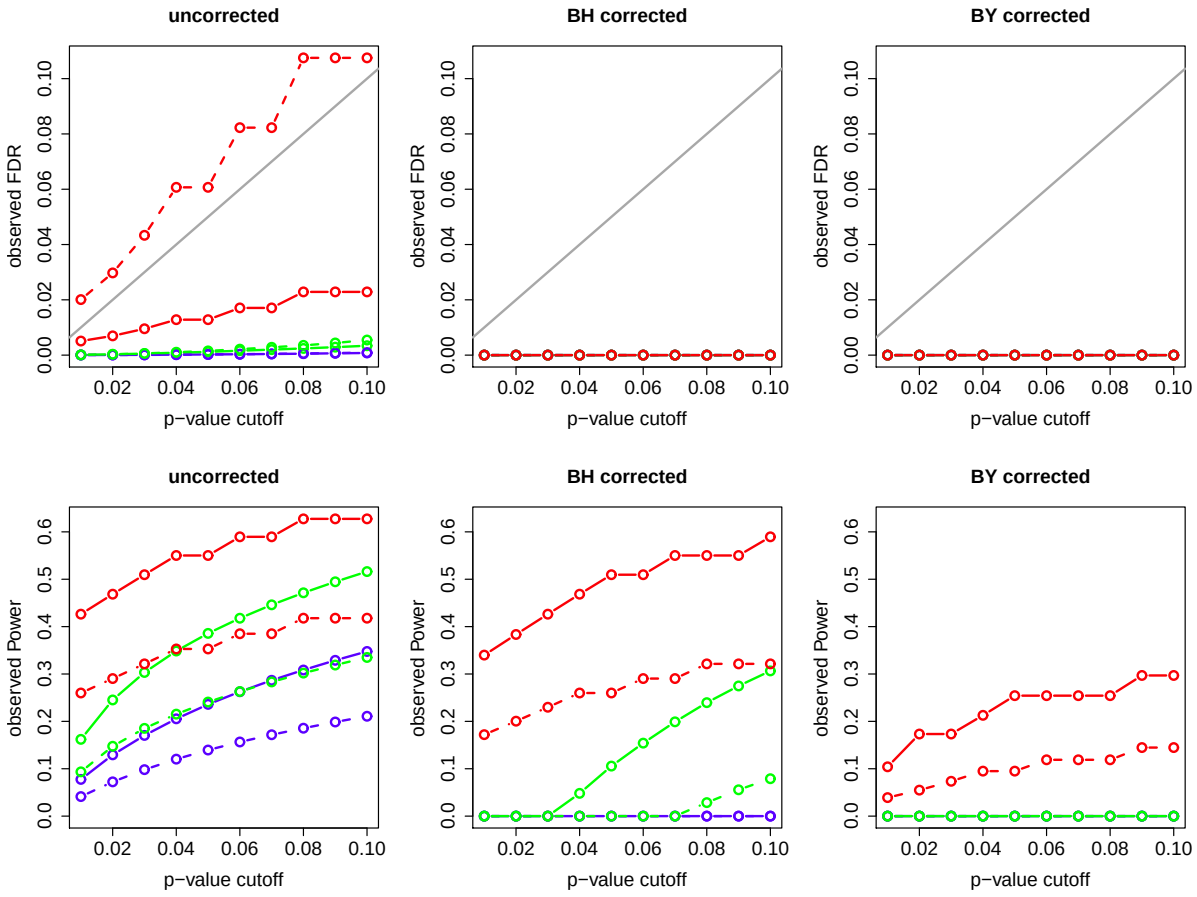

Figure S13 Observed FDR and power for the simulations based on model F. The x-axis represents the p-value cutoffs used for computing the results. Dashed and full curves represent, respectively, AIC- and BIC-based methods. Green: parametric CMST. Red: non-parametric CMST. Blue: joint-parametric CMST. Black: AIC and BIC. The grey line in the top panels corresponds to the α levels.

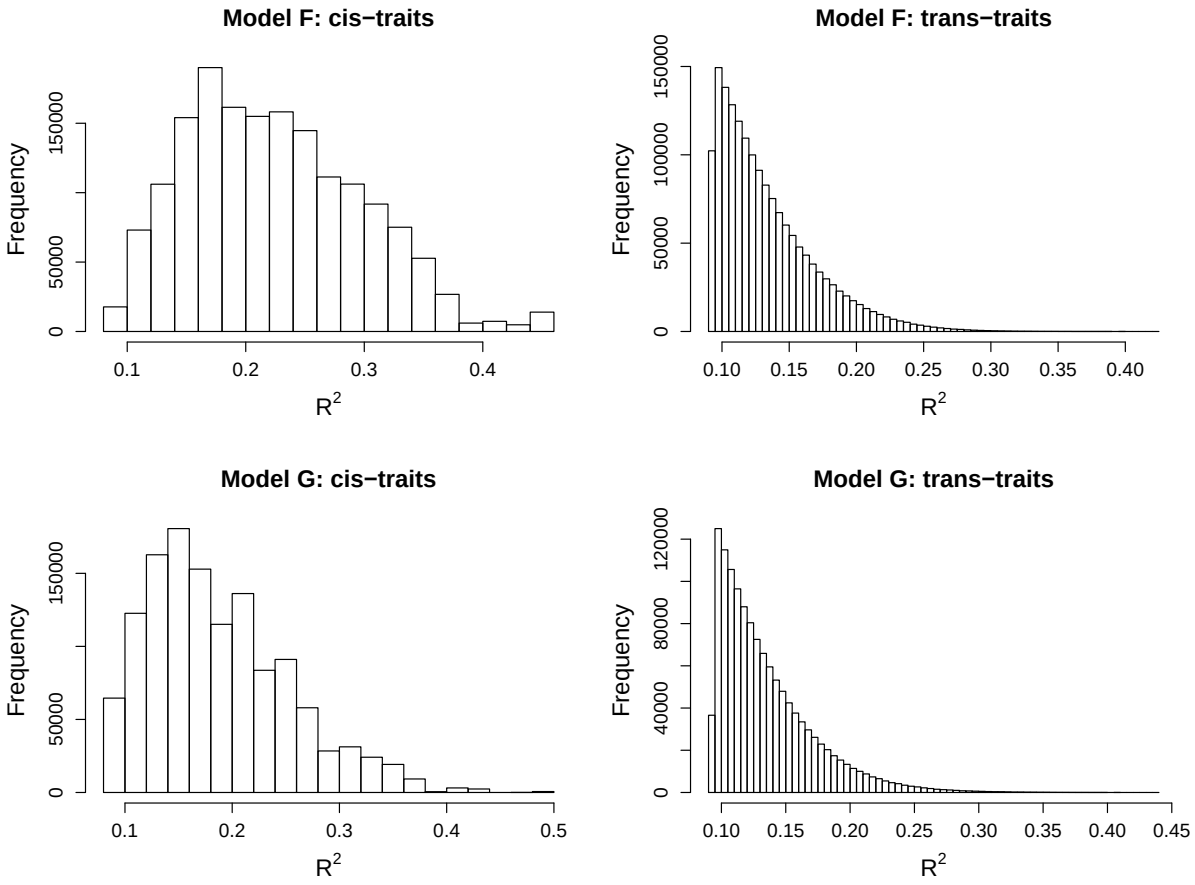

Figure SI.3 Overall R^2 statistics distributions for the large scale simulation study. The left and right panels show the distribution for the *cis*-traits and *trans*-traits, respectively.

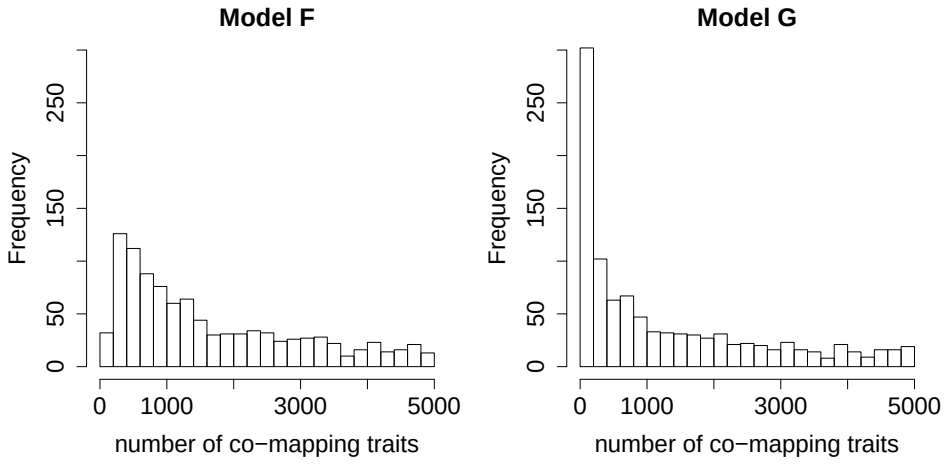

Figure SI.4 For each model F and G we performed 1,000 separate simulations, and tested Y_1 against all other phenotypes Y_k, $k = 2, \ldots, 5001$, that shared the QTL with Y_1, at each simulation. The panels show the distribution of the number of tests, i.e, the number of trans-traits that co-mapped 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.

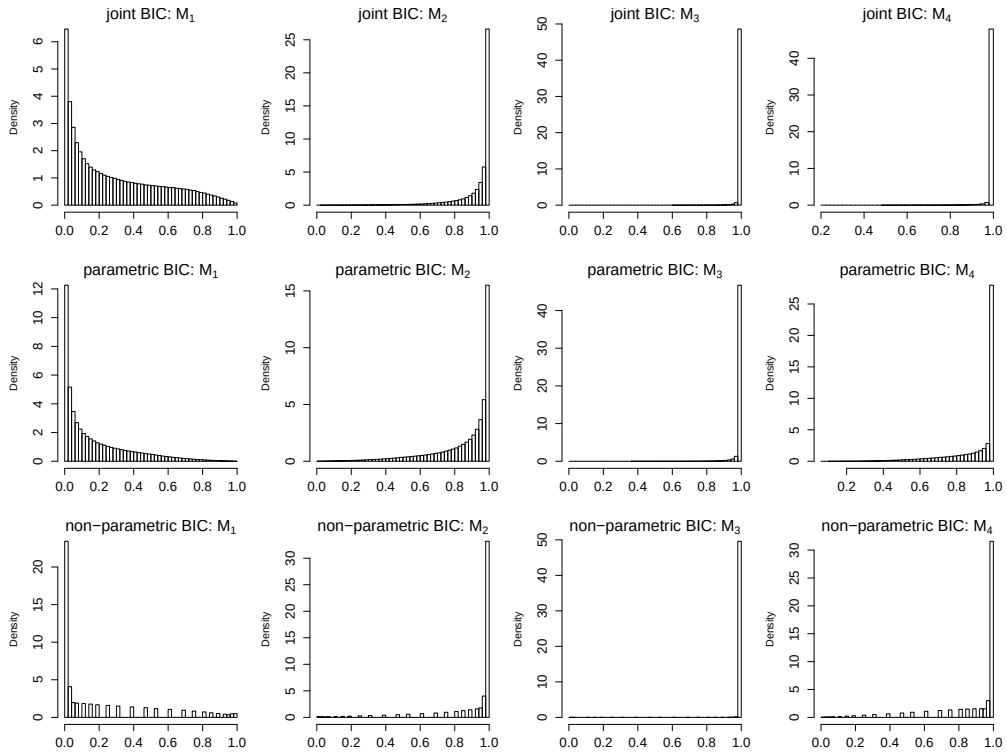

Figure S14 Uncorrected p-value distributions for the BIC-based CMST tests with data simulated from model F in Figure 5. Results based on 1,656,261 tests. For these simulations, the M_1 call is the correct one, hence the skewed distribution towards small p-values at the left panels. The skewness towards 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).

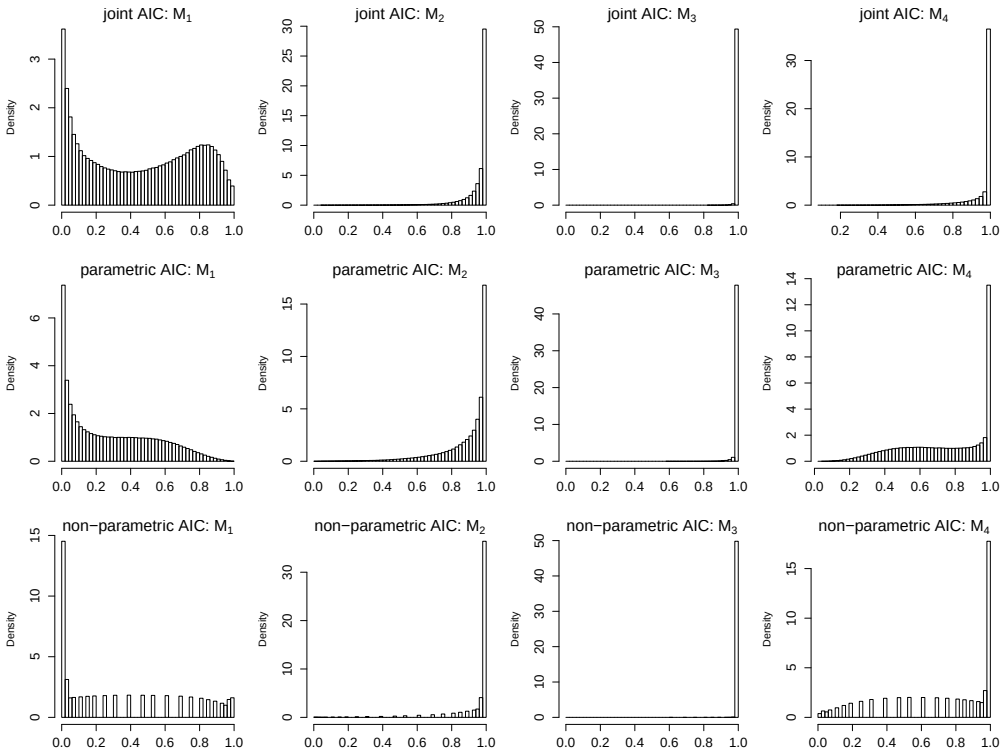

Figure S15 Uncorrected p-value distributions for the AIC-based CMST tests with data simulated from model F in Figure 5. Results based on 1,656,261 tests. For these simulations, the M_1 call is the correct one, hence the skewed distribution towards small p-values at the left panels. The skewness towards 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).

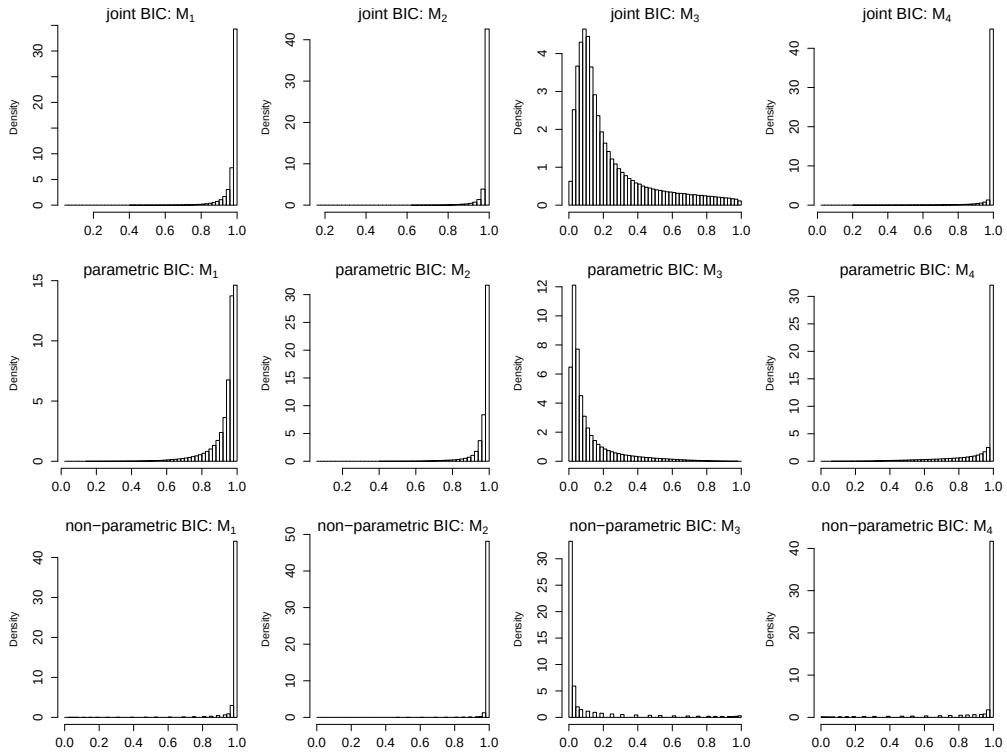

Figure S16 Uncorrected p-value distributions for the BIC-based CMST tests with data simulated from model G in Figure 5. Results based on 1,286,243 tests. For these simulations, the M_3 call is the correct one, hence the skewed distribution towards small p-values at the M_3 panels. The skewness towards 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).

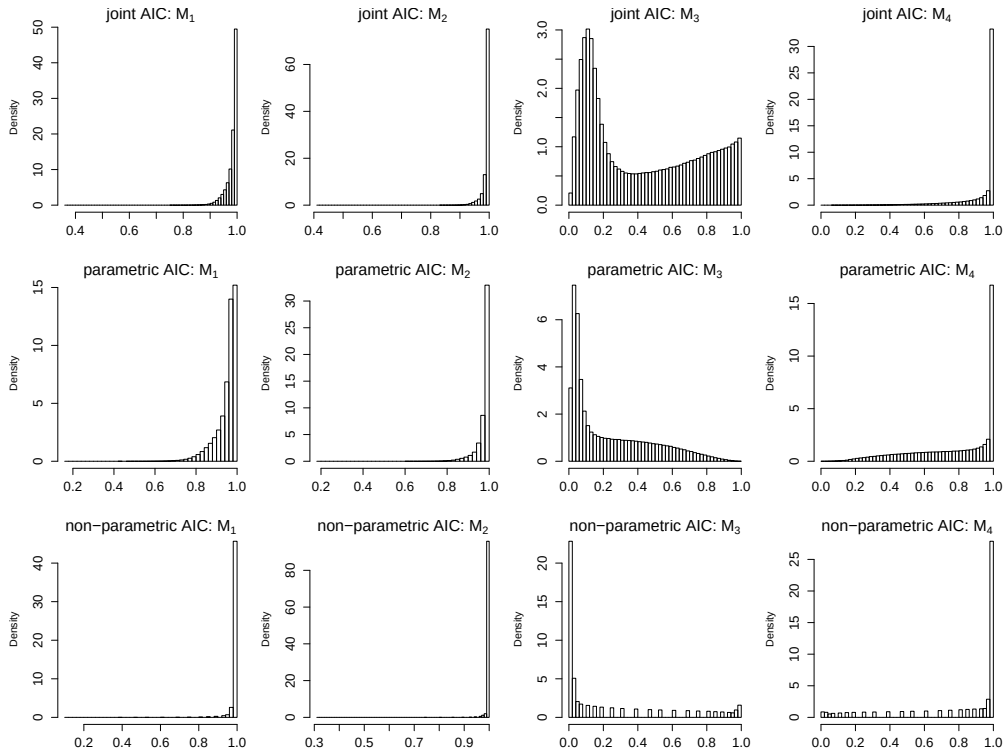

Figure S17 Uncorrected p-value distributions for the AIC-based CMST tests with data simulated from model G in Figure 5. Results based on 1,286,243 tests. For these simulations, the M_3 call is the correct one, hence the skewed distribution towards small p-values at the M_3 panels. The skewness towards 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).