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.

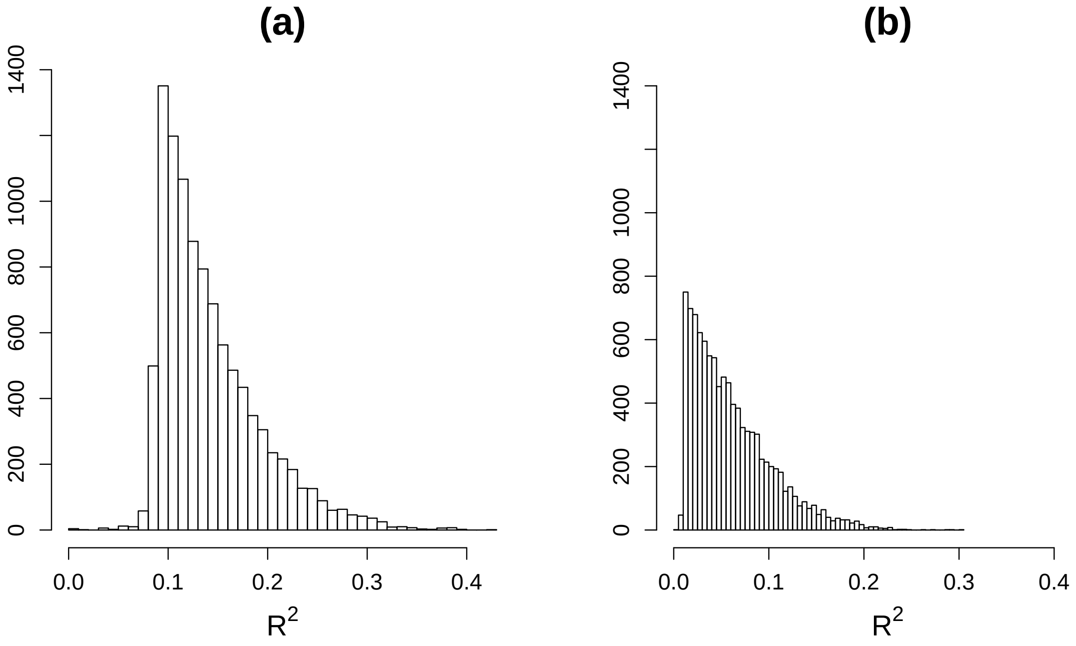

Figure SI.2 Overall distribution of the R^2 statistics across all simulated models in Figure 2. Panels a and b present the R^2 statistics for sample sizes 112 and 1,000, respectively.

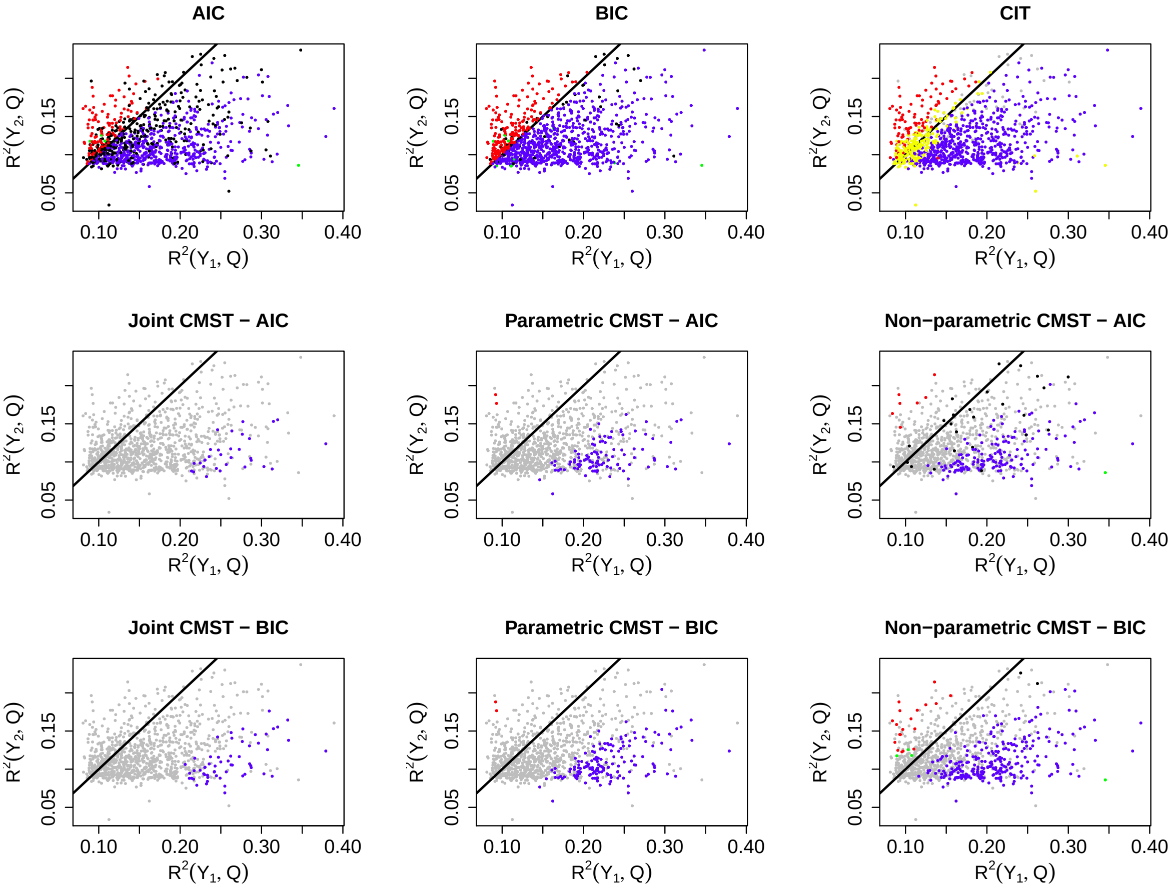

Figure S3 Simulation results for Model A in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this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

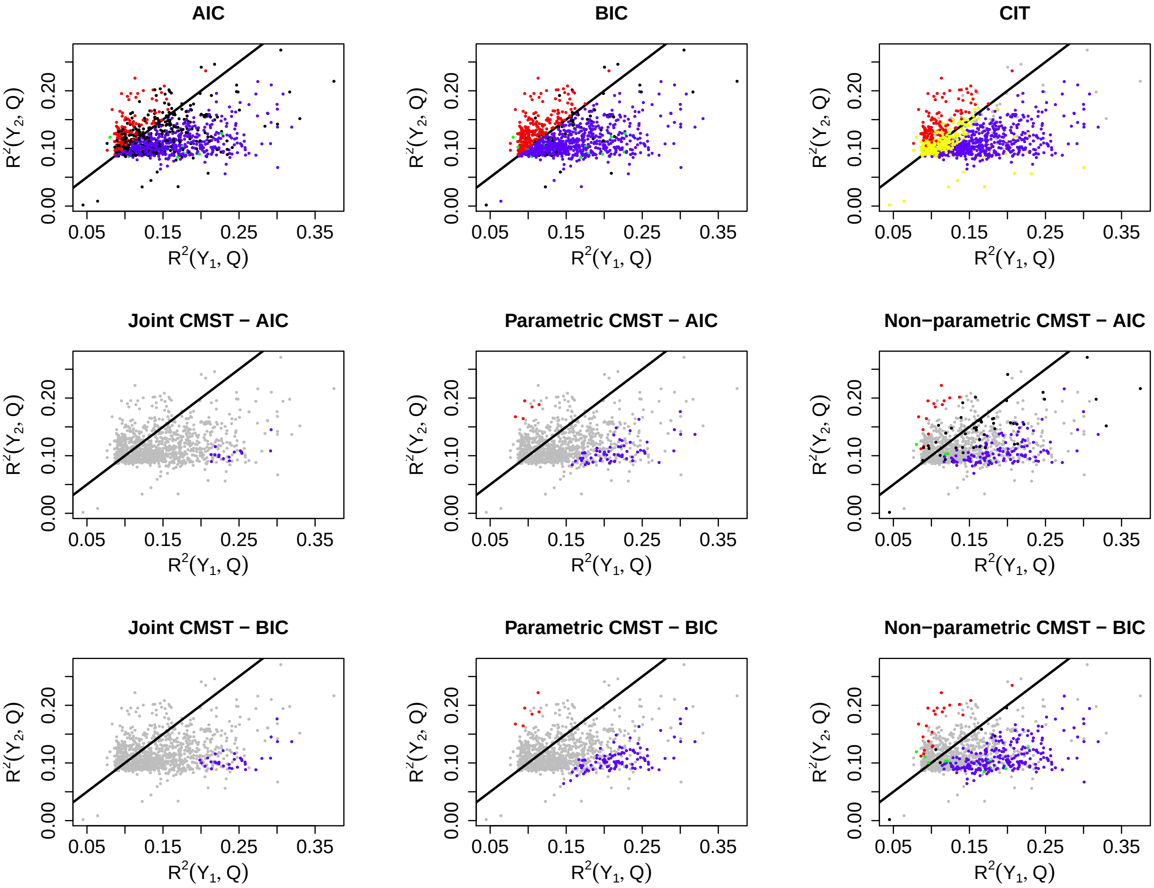

Figure S4 Simulation results for Model B in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this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

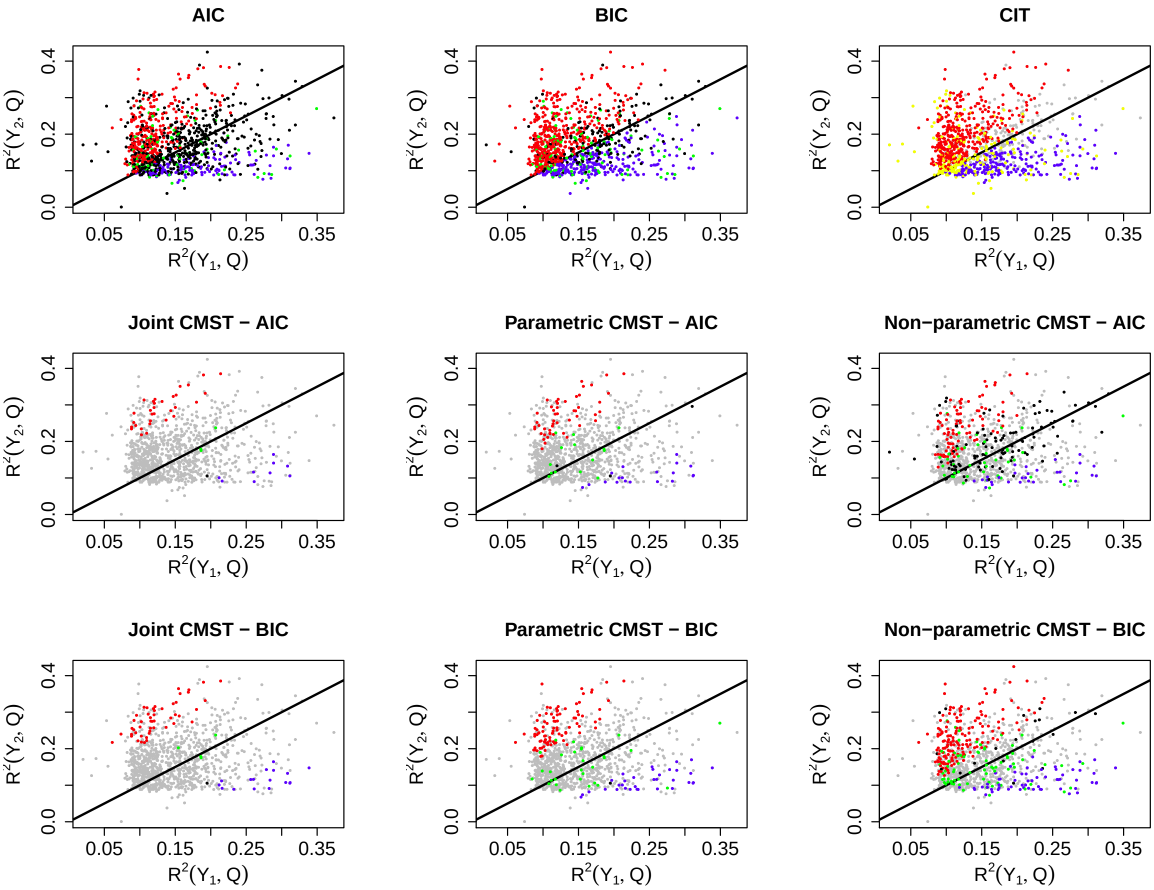

Figure S5 Simulation results for Model C in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the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

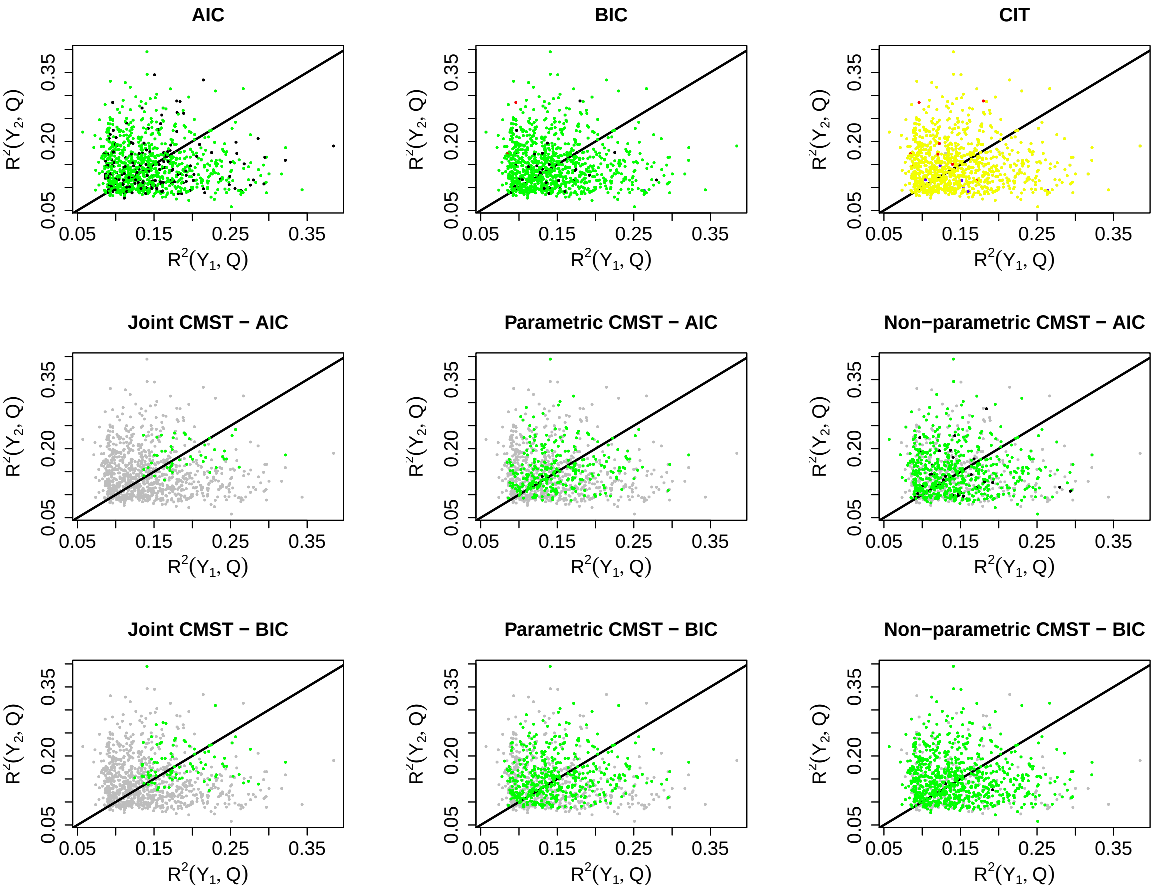

Figure S6 Simulation results for Model D in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the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

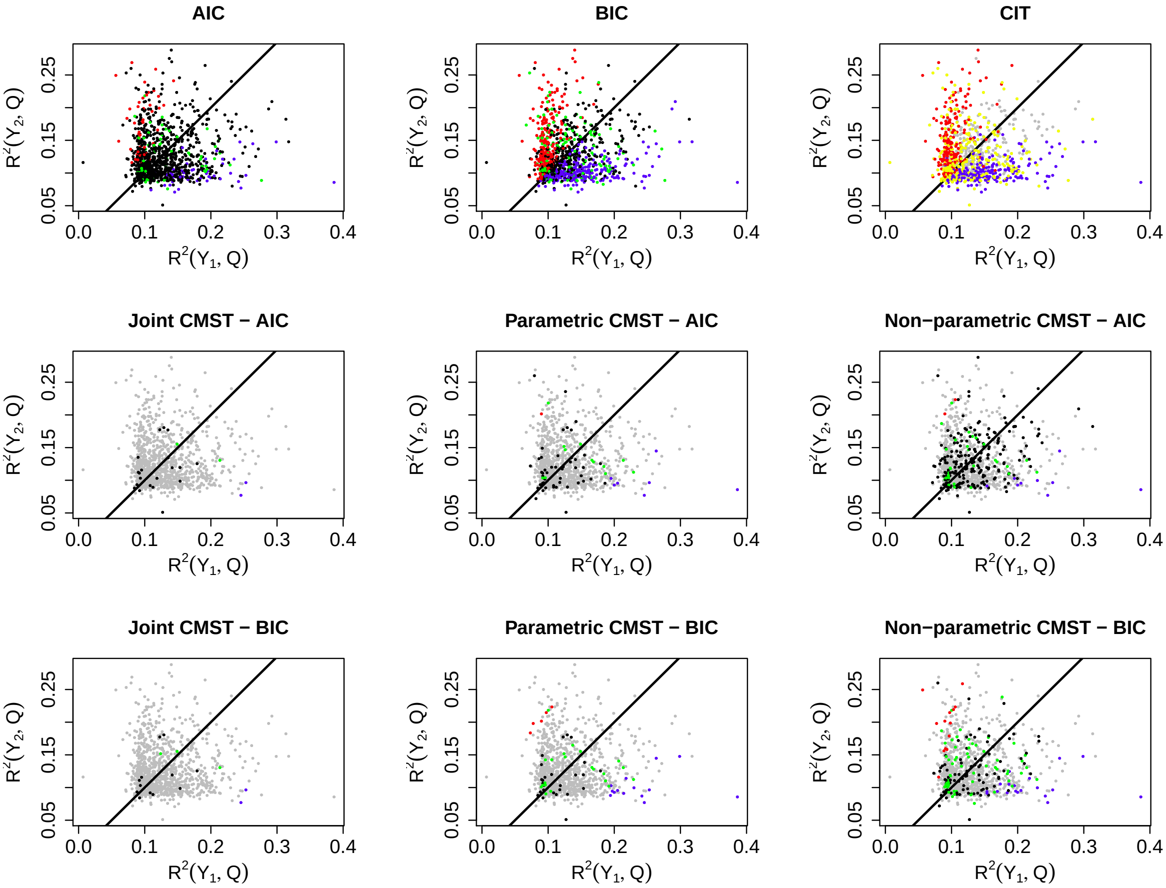

Figure S7 Simulation results for Model E in Figure 2 and sample size 112. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the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

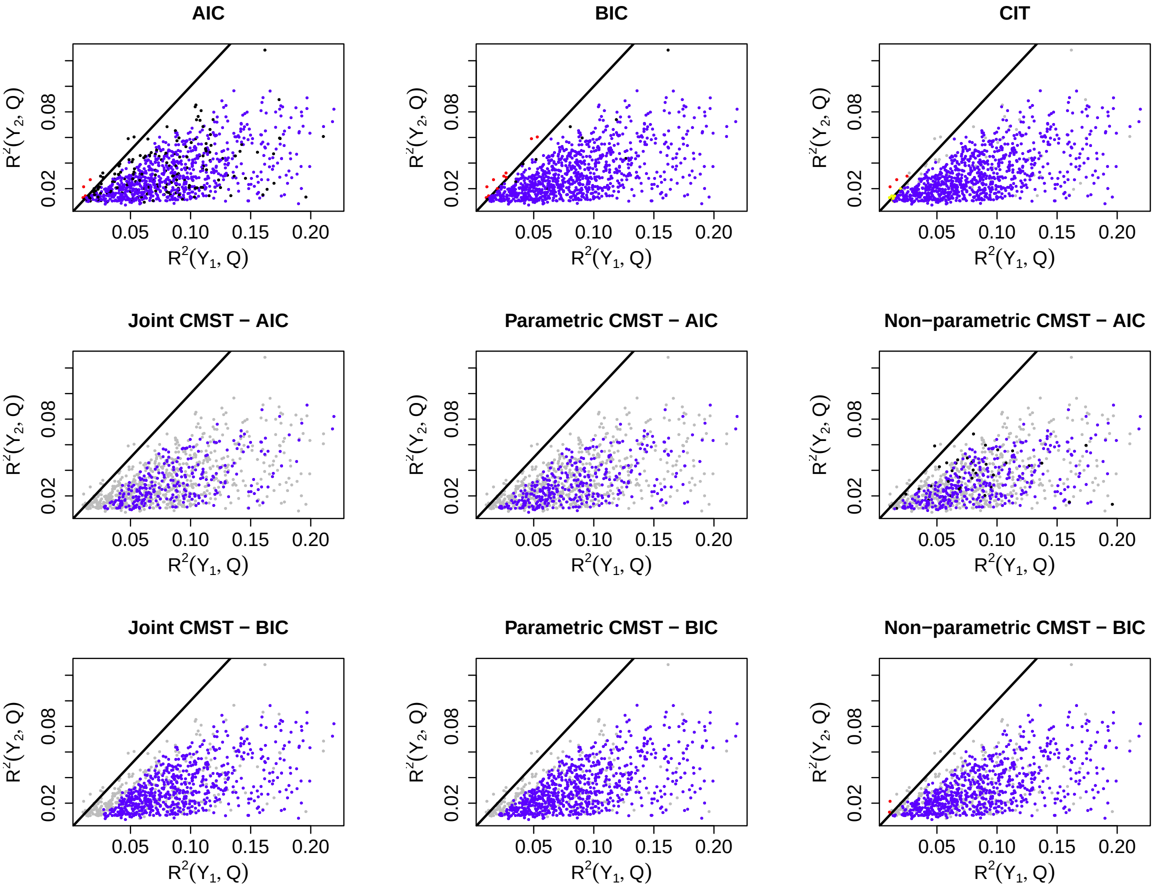

Figure S8 Simulation results for Model A in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the “no calls”. Results were computed using significance level 0.05. For this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

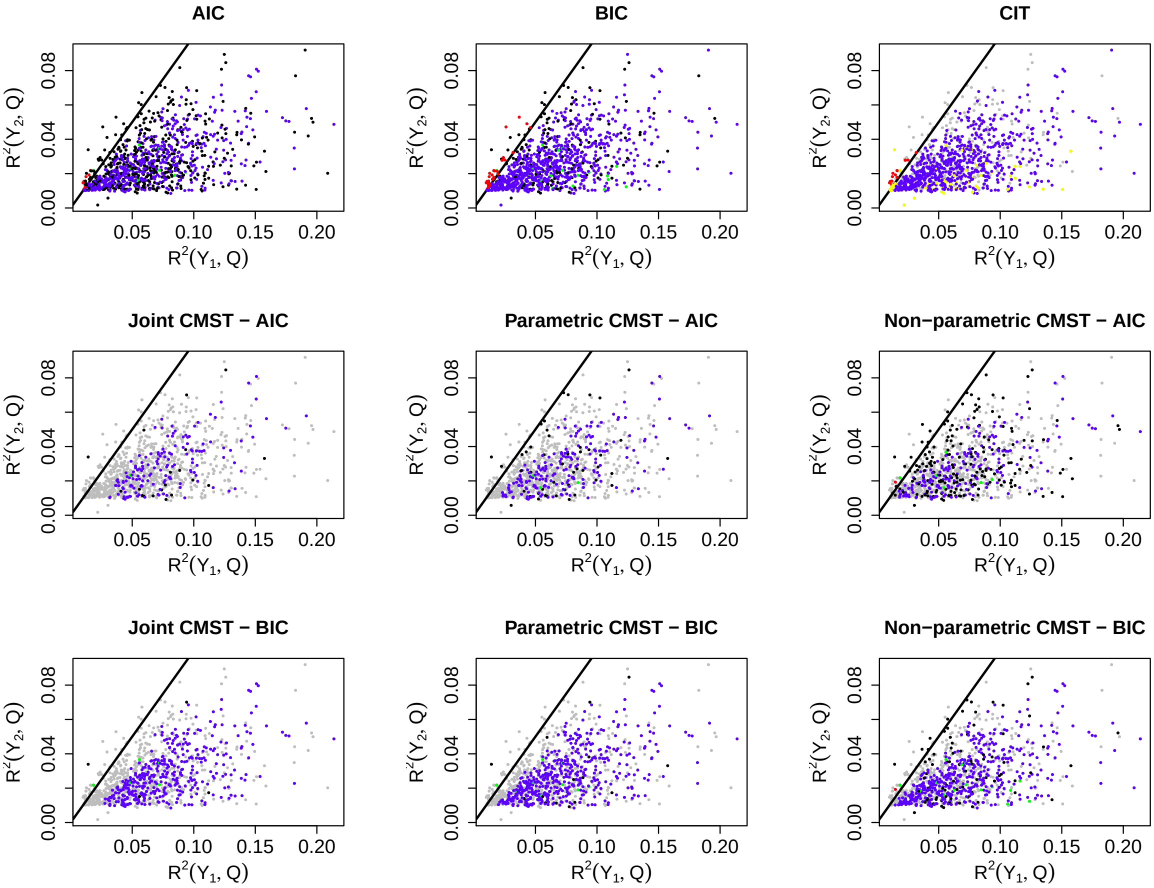

Figure S9 Simulation results for Model B in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For this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

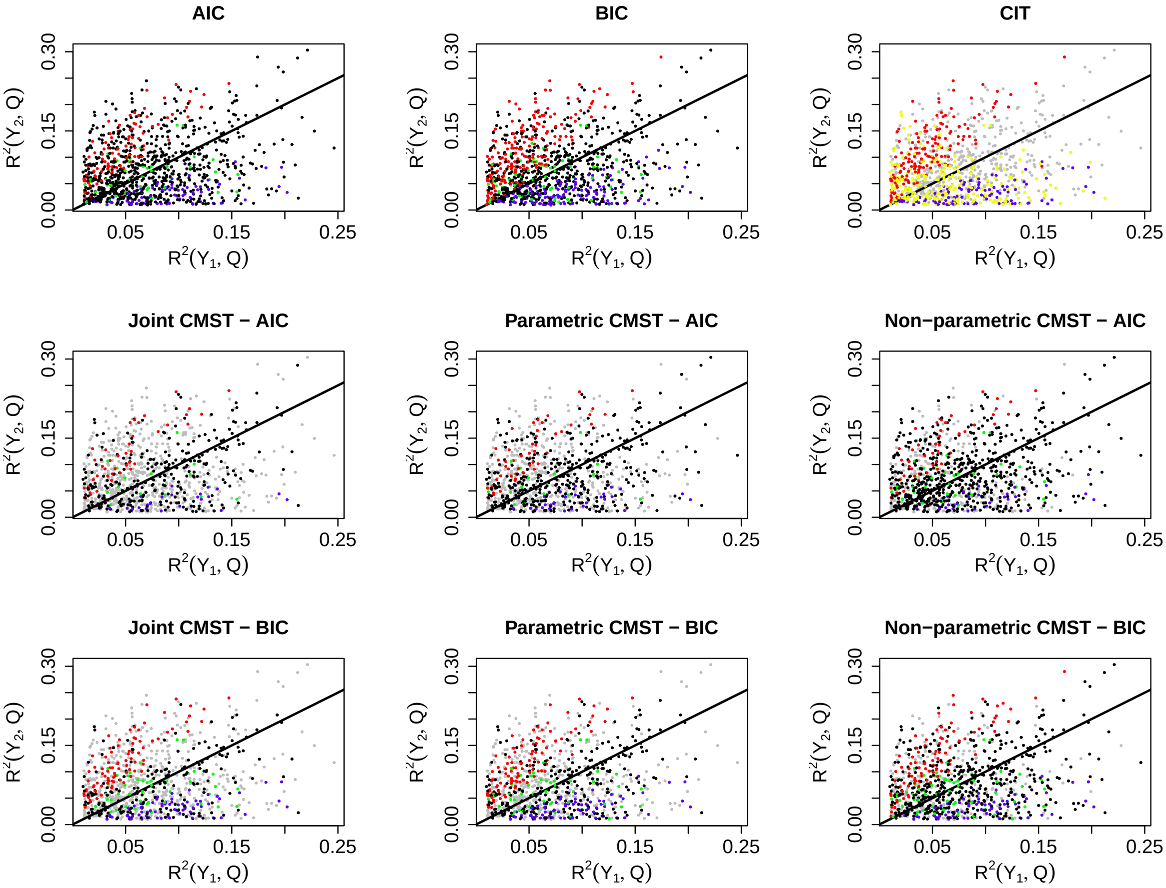

Figure S10 Simulation results for Model C in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the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

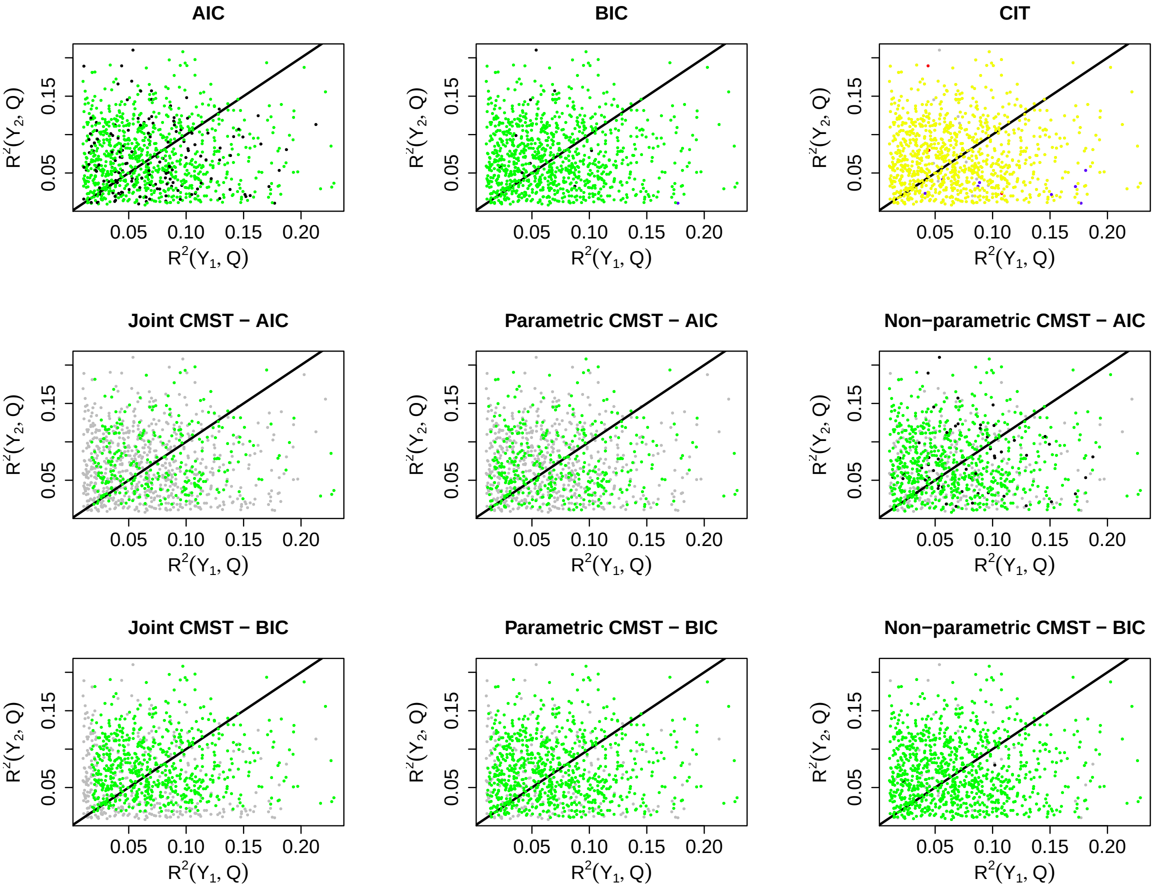

Figure S11 Simulation results for Model D in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the “no calls”. Results were computed using significance level 0.05. For the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

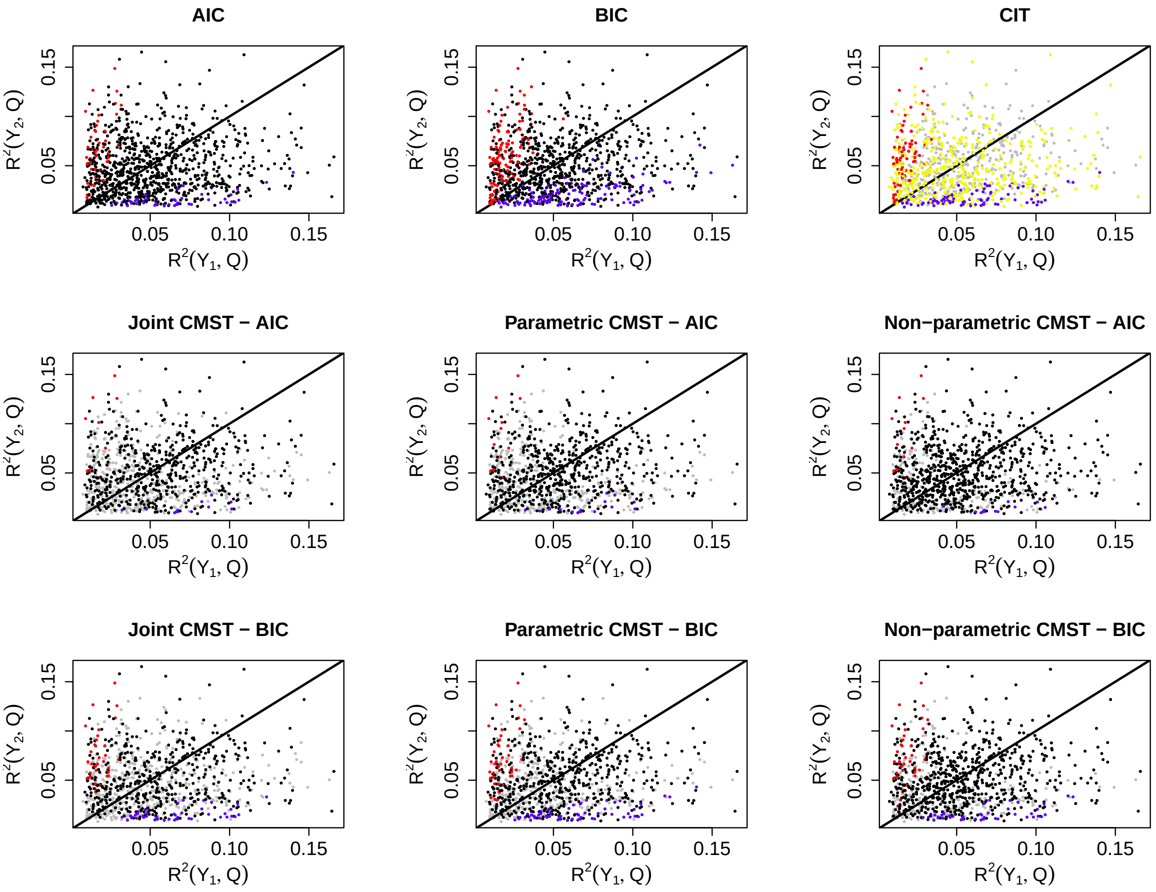

Figure S12 Simulation results for Model E in Figure 2 and sample size 1,000. Blue, red, green and black dots represent, respectively, M_1, M_2, M_3 and M_4 calls. Yellow dots (CIT plot only) represent M_i calls. Grey dots show the "no calls". Results were computed using significance level 0.05. For the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

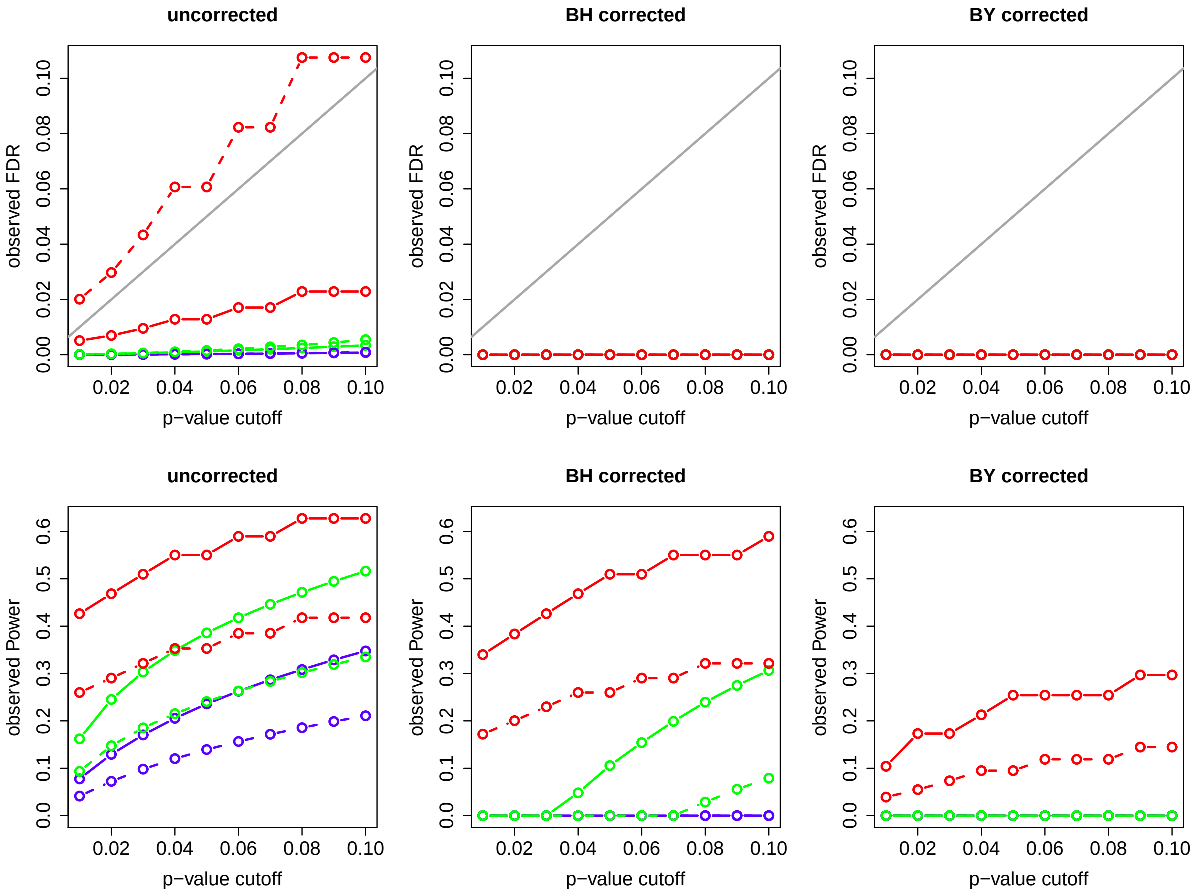

Figure S13 Observed FDR and power for the simulations based on model F. The x-axis represents the p-value cutoffs used for computing the results. Dashed and full curves represent, respectively, AIC- and BIC-based methods. Green: parametric CMST. Red: non-parametric CMST. Blue: joint-parametric CMST.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

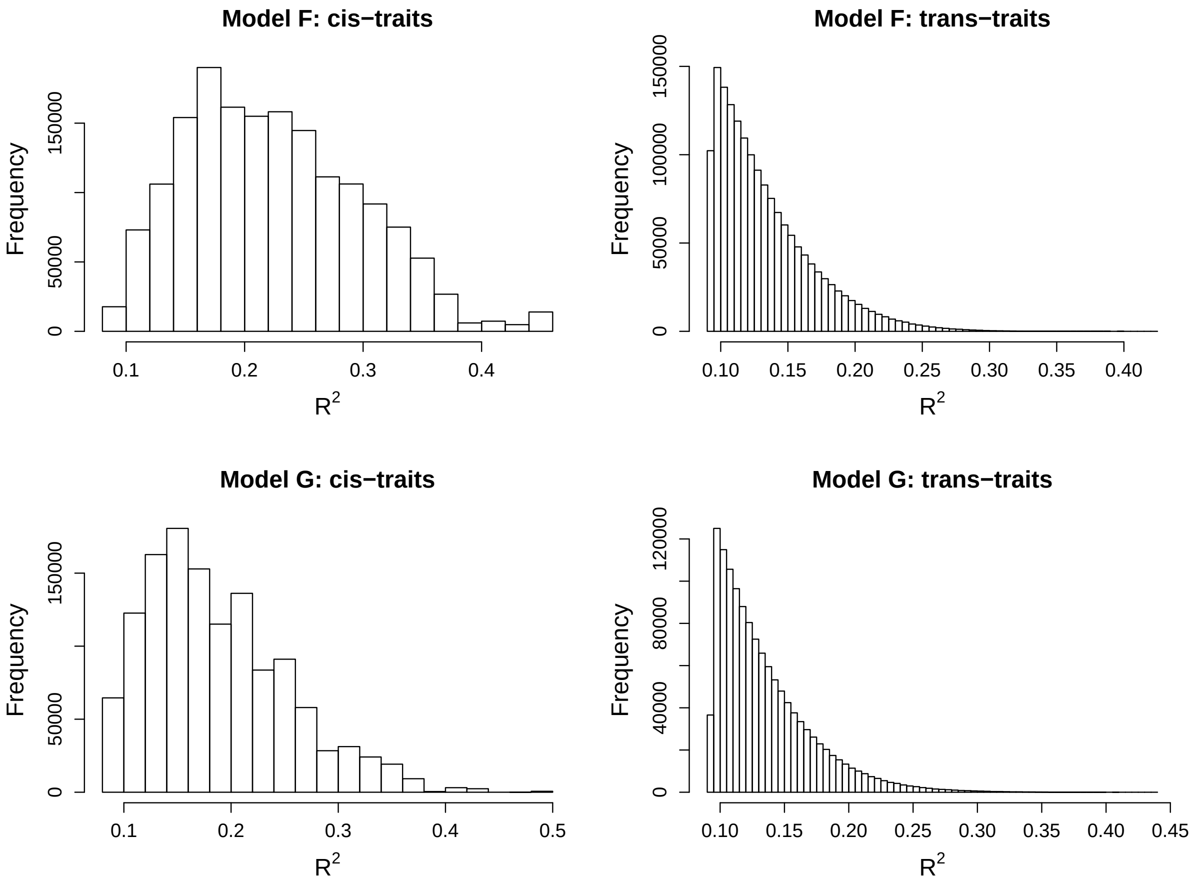

Figure SI.3 Overall R^2 statistics distributions for the large scale simulation study. The left and right panels show the distribution for the *cis*-traits and *trans*-traits, respectively.

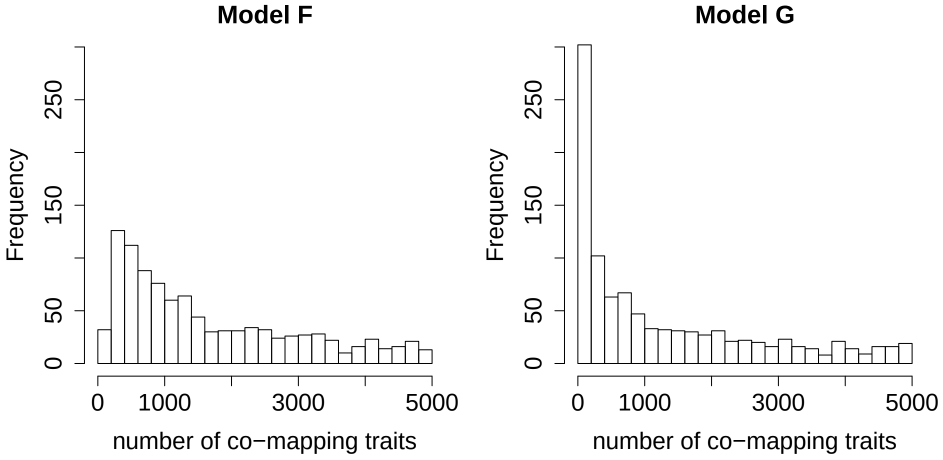

Figure SI.4 For each model F and G we performed 1,000 separate simulations, and tested Y_1 against all other phenotypes Y_k, $k = 2, \ldots, 5001$, that shared the QTL with Y_1, at each simulation. The panels show the distribution of the number of tests, i.e, the number of trans-traits that co-mapped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

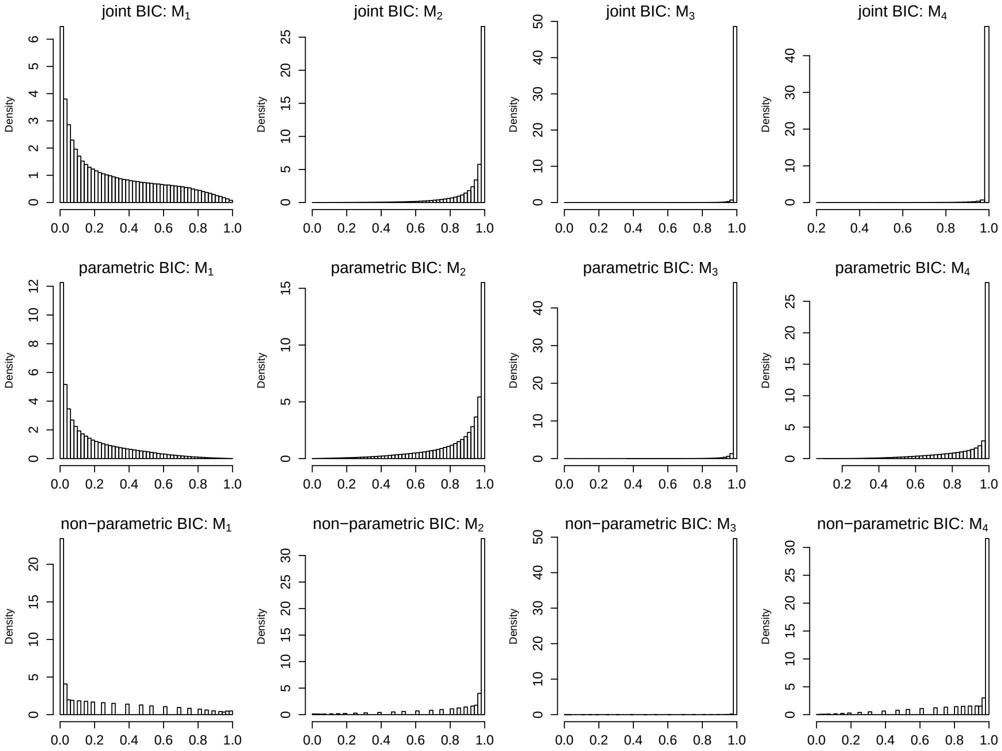

Figure S14 Uncorrected p-value distributions for the BIC-based CMST tests with data simulated from model F in Figure 5. Results based on 1,656,261 tests. For these simulations, the M_1 call is the correct one, hence the skewed distribution towards small p-values at the left panels. The skewness towards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

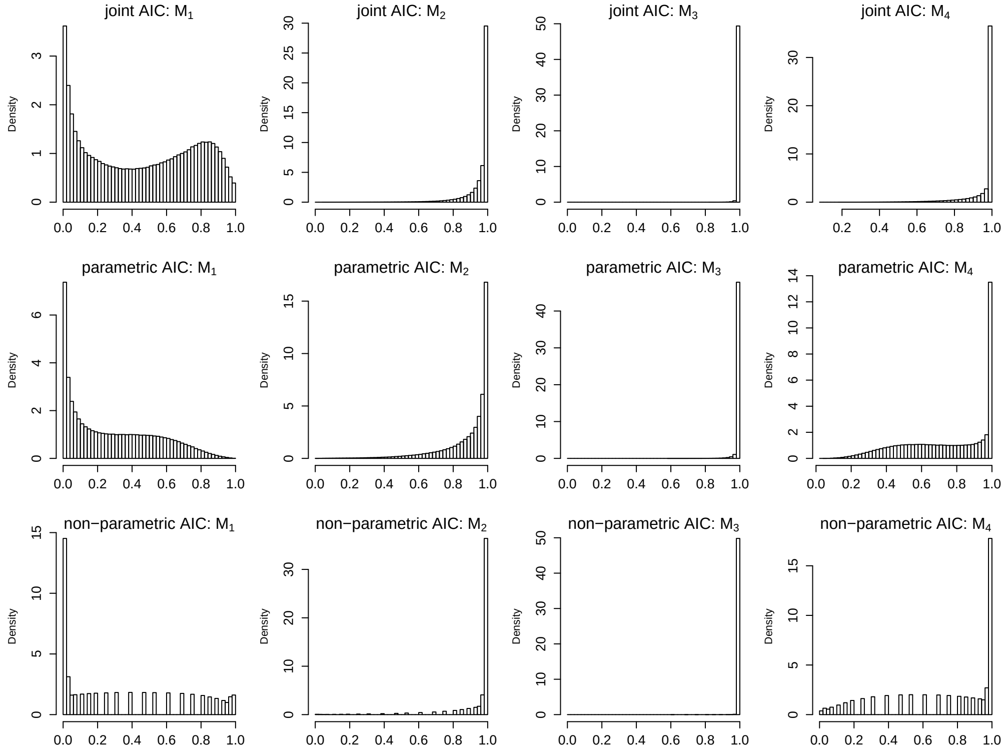

Figure S15 Uncorrected p-value distributions for the AIC-based CMST tests with data simulated from model F in Figure 5. Results based on 1,656,261 tests. For these simulations, the M_1 call is the correct one, hence the skewed distribution towards small p-values at the left panels. The skewness towards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

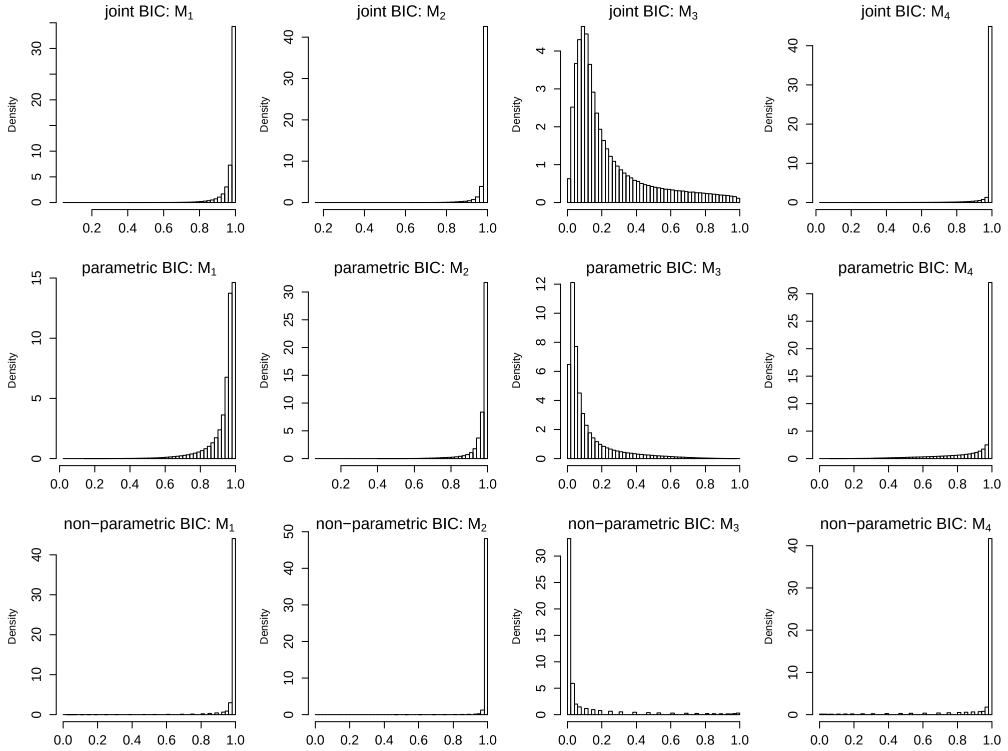

Figure S16 Uncorrected p-value distributions for the BIC-based CMST tests with data simulated from model G in Figure 5. Results based on 1,286,243 tests. For these simulations, the M_3 call is the correct one, hence the skewed distribution towards small p-values at the M_3 panels. The skewness towards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

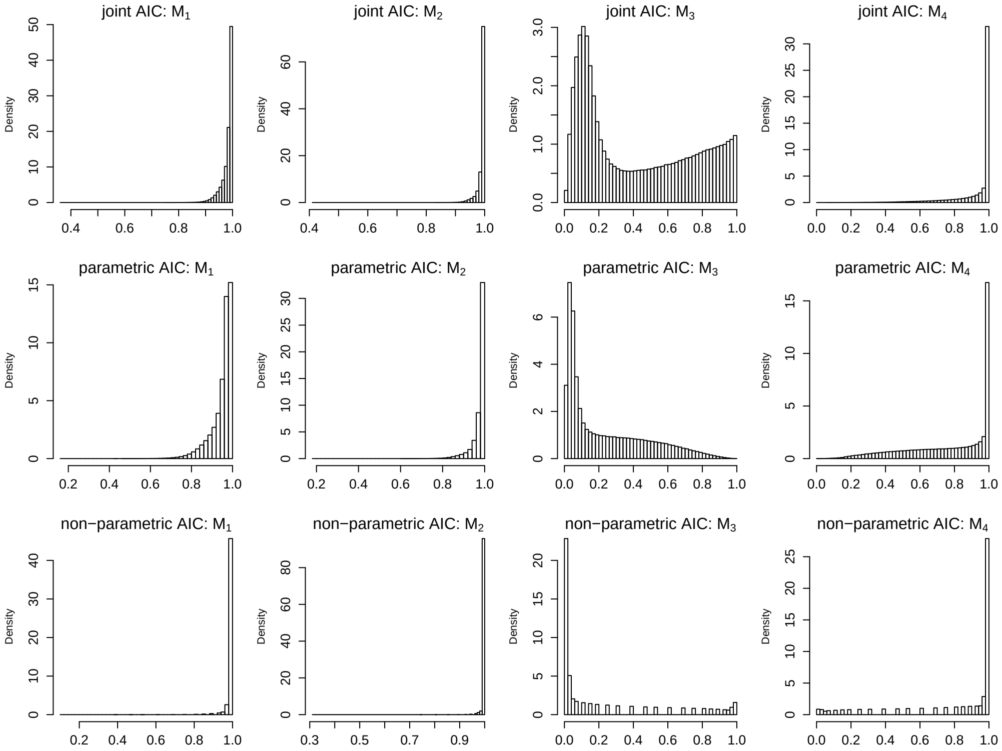

Figure S17 Uncorrected p-value distributions for the AIC-based CMST tests with data simulated from model G in Figure 5. Results based on 1,286,243 tests. For these simulations, the M_3 call is the correct one, hence the skewed distribution towards small p-values at the M_3 panels. The skewness towards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