Exploiting Population Samples to Enhance Genome-Wide Association Studies of Disease

Shachar Kaufman and Saharon Rosset[1]
School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel 6997801

ABSTRACT It is widely acknowledged that genome-wide association studies (GWAS) of complex human disease fail to explain a large portion of heritability, primarily due to lack of statistical power—a problem that is exacerbated when seeking detection of interactions of multiple genomic loci. An untapped source of information that is already widely available, and that is expected to grow in coming years, is population samples. Such samples contain genetic marker data for additional individuals, but not their relevant phenotypes. In this article we develop a highly efficient testing framework based on a constrained maximum-likelihood estimate in a case–control–population setting. We leverage the available population data and optional modeling assumptions, such as Hardy–Weinberg equilibrium (HWE) in the population and linkage equilibrium (LE) between distal loci, to substantially improve power of association and interaction tests. We demonstrate, via simulation and application to actual GWAS data sets, that our approach is substantially more powerful and robust than standard testing approaches that ignore or make naive use of the population sample. We report several novel and credible pairwise interactions, in bipolar disorder, coronary artery disease, Crohn's disease, and rheumatoid arthritis.

GENOME-WIDE association studies (GWAS) have implicated thousands of single-nucleotide polymorphisms (SNPs) in the human genome as associated with hundreds of phenotypes (Johnson and O'Donnell 2009). However, as many researchers have pointed out (Manolio *et al.* 2009; Eichler *et al.* 2010), the results from GWAS fail to explain the observed heritability of many phenotypes, including complex human diseases, whose genetic architectures remain largely unknown. One often-cited reason for this problem is that the high multiple-testing burden requires an exceedingly stringent statistical significance level. Furthermore, while most studies have employed univariate (locus-by-locus) testing approaches, complex diseases are likely to be affected by interactions between loci (Eichler *et al.* 2010). Such interactions arise when there is a dependence of genotypic effects of one locus on genotypes at other loci (Cordell 2009).

In the case of interactions, due to the overwhelming number of locus subsets, the multiple-testing problem becomes a serious computational and statistical challenge. Even when limiting exploration to pairwise SNP–SNP interactions, a modest study including 300,000 usable loci requires testing $\sim$45 billion SNP pairs, and associations must have P-values $< \sim 10^{-12}$ (the 0.05 Bonferroni-corrected significance level) to be declared statistically significant genome-wide. Recently, several authors have suggested sophisticated approximate and exhaustive methods for detecting pairwise interactions (Brinza *et al.* 2010; Liu *et al.* 2011; Prabhu and Pe'er 2012). These methods constitute a major step in dealing with the computational issue of carrying out the large number of tests, but their application to actual studies has led to surprisingly few replicable discoveries (for example, one pair in Liu *et al.* 2011 and another in Prabhu and Pe'er 2012). Many of the other reported pairwise discoveries fail after careful scrutiny (see, for example, the discussion in Liu *et al.* 2011 and the *Discussion* in the present article). Given findings in other organisms (Shao *et al.* 2008; Bloom *et al.* 2013), and the biological plausibility of the existence of interactions, the likely explanation for the limited GWAS results is that modest interaction effects comprising common SNPs do exist; however, due to the aforementioned statistical challenge, the tests employed are not powerful enough to detect them given available sample sizes.

Improving power of tests used in GWAS is therefore an extremely important research question today, especially

doi: 10.1534/genetics.114.162511
Manuscript received February 6, 2014; accepted for publication February 27, 2014; published Early Online March 10, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162511/-/DC1.
[1]Corresponding author: School of Mathematical Sciences, Tel Aviv University, Box 39040, Tel Aviv, Israel 6997801. E-mail: saharon@post.tau.ac.il

when interactions are considered. One solution is to increase sample sizes by collecting more case–control data. Indeed, this is an ongoing trend, and ever-larger studies are in preparation by various worldwide consortia. Often, however, independent samples from the studied population but with unknown case–control status are already available. Sources include, for example, (i) population studies (Waye 2005; Siva 2008), (ii) quantitative trait studies (Yang *et al.* 2010), and (iii) binary trait studies that employ a case–population design [like the Wellcome Trust Case Control Consortium (WTCCC) study, discussed in more detail below]. The coming years are expected to make such *population samples* available in increasingly large numbers, given recently published government plans in multiple countries to sequence hundreds of thousands of genomes (http://news.sciencemag.org/scienceinsider/2012/12/uk-unveils-plan-to-sequence-whol.html).

A key challenge is to make optimal use of these population samples to increase power of GWAS in general and interaction detection efforts in particular. The common existing practice is to use only the case–control data and ignore the population data completely. Some studies employ population data as reference panels from which linkage disequilibrium structure is inferred. This can improve power by reducing the effective multiple-comparisons burden, as well as by the finer localization of detected associations through imputation. These methods are different from the one we develop and could be used as well. In another approach, which is closer to the one studied here, population samples are treated as additional controls with the intention of increasing sample size and thus power (Burton *et al.* 2007; Lippert *et al.* 2013). Arguably, for rare diseases the latter approach is satisfactory (because controls are very similar to the population). However, since this leads to mislabeling of any cases in the population sample as controls, for common disease this becomes a problem, possibly causing more harm than good. We suggest a more appropriate approach here that models the joint likelihood of the case–control and population samples, in the expectation that correctly using all available data would result in significantly improved power. This expectation is verified in our simulations and is further reflected in application to real GWAS data for seven different phenotypes studied by the WTCCC.

An important feature of our approach is its ability to incorporate assumptions about the studied population. The idea is that by exploiting additional properties of the data, one can reduce the parameter space describing associations of interest (Song and Nicolae 2009). This improves power to detect associations that comply with the assumptions, at the often acceptable cost of reduced power to detect implausible associations. Two popular and broadly applicable population assumptions that constrain the parameter space are Hardy–Weinberg equilibrium (HWE) (the independence of maternal and paternal allele values that make up each genotype locus) and linkage equilibrium (LE) (the independence of the two loci that make up a pair) (Yang *et al.* 1999; Chatterjee and Carroll 2005; Zhao *et al.* 2006). Imperfect as they may be (see *Discussion*), these assumptions are expected to hold for the vast majority of associations considered in a typical GWAS.

Thus, the main contribution of our present work is the proposal of a method for use in GWAS that combines (a) the correct handling of case–population or case–control–population designs and (b) inclusion of population assumptions, to maximize efficiency of association testing. We show in simulation results that our direct constrained maximum-likelihood (CMLE) approach is substantially more powerful than ignoring extraneous population data in a case–control study or using them naively to extend controls. We then apply this method to the seven phenotypes from the WTCCC study, which employs a case–population design, with the cases of unrelated diseases potentially serving as a source for additional population samples. Our pairwise analysis of these data reveals multiple new associations in several diseases. Examination of all significant findings shows that a large number of them include SNPs in genomic regions not implicated by the original WTCCC study. Interestingly, these regions include several loci that were identified and replicated in later studies employing standard analysis methods but larger sample sizes—illustrating and validating the improved power of our method. We report several novel and credible associations, including pairwise interactions in bipolar disorder, coronary artery disease, Crohn's disease, and rheumatoid arthritis.

An implementation of all tests described herein for pairwise and univariate GWAS is available in the *R* statistical computing software package *CCpop*. As our theory and results demonstrate, our new approach can lead to substantial gains in power (up to sevenfold in standard models, see *Results*). We believe that practically every case–control GWAS can identify relevant population samples to use as additional data and that many of these studies can further justify population assumptions when analyzing the data (any doubt about such assumptions can be eliminated by using more general tests during replication). Thus our recommendation is that previous studies be revisited with our approach and that new studies use our tests rather than the traditional ones.

Methods

Association testing with case–control–population data

Here we describe our new approach that correctly handles case–control–population designs within a maximum-likelihood framework. First, the method is presented in the context of a single SNP. Second, we extend it to the case of two or more SNPs. Finally, we show how to incorporate population assumptions into the model.

The univariate case: Let x be a diploid SNP locus with genotypes in {0, 1, 2}, and let y be a binary disease phenotype coded 0 for controls (which are truly unaffected) and 1 for cases. Let $\overrightarrow{n} = \{n_{i,j}\}_{i=0,j=0}^{1\ 2}$ denote the observed number of samples in a standard retrospective case–control sample $\mathcal{C}$ with $y = i$ and $x = j$. Let $n_{i,} = \sum_j n_{i,j}$ denote the number of individuals with phenotype i and $n_{,j} = \sum_i n_{i,j}$ denote the number of

individuals with genotype j. Suppose there exists, in addition, an independent population sample $\mathcal{P}$ with the observed genotype counts at x denoted $\vec{m} = \{m_j\}_{j=0}^{2}$. Let $\vec{p} = \{p_{i|j}\}_{i=0,j=0}^{1\ 2}$ be the probability of disease status i given the genotype j. For $i = 1$, this is referred to as disease *penetrance*. Denote $\vec{\pi} = \{\pi_j\}_{j=0}^{2}$ the marginal distribution of x in the population and K the prevalence of the disease, which is assumed known.

The log-likelihood of the pooled data $(\mathcal{C}, \mathcal{P})$ may be written as

$$\begin{aligned} l(\vec{p}, \vec{\pi}; \vec{n}, \vec{m}) &= \sum_{i,j} n_{i,j} \log\left(p_{i|j}\frac{\pi_j}{K}\right) + \sum_j m_j \log \pi_j \\ &\propto \sum_{i,j} n_{i,j} \log p_{i|j} + \sum_j (n_{\cdot,j} + m_j) \log \pi_j. \end{aligned} \tag{1}$$

Without further assumptions on the natural parameter space $\{\vec{p}, \vec{\pi}\}$, this space is of dimension 5, since $\vec{\pi}$ must sum to 1, and $p_{1|j} = 1 - p_{0|j}$. The known prevalence K imposes a constraint on this space:

$$\sum_j p_{1|j}\pi_j = K. \tag{2}$$

Solving (1) for the maximum-likelihood estimator (MLE) under the alternative hypothesis (of association between y and x) may be approached directly as a nonlinear optimization problem with a nonlinear equality constraint (2) and "box" inequality constraints that keep all parameters in [0, 1]. Under the null hypothesis (of no association), an additional constraint is imposed that the penetrance be flat $p_{1|j} \equiv c$. Trivially, due to (2),

$$c = K, \quad \pi_j = \frac{n_{\cdot,j} + m_j}{n + m}. \tag{3}$$

This estimation approach works well, but convergence under the alternative may not be fast enough to manage the large number of tests that must be performed during GWAS. Rewriting the problem in standard convex form (Boyd and Vandenberghe 2004) and using modern solvers lead to faster model fitting. Define $q_j = p_{1|j}\pi_j$; solving (1) subject to (2) now takes the following convex optimization problem form:

$$\begin{aligned} \underset{\vec{q}, \vec{\pi}}{\text{minimize}} \ & -\sum_j \left(n_{0,j} \log(\pi_j - q_j) + n_{1,j} \log q_j + m_j \log \pi_j\right) \\ \text{subject to} \ & \sum_j q_j = K, \\ & \sum_j \pi_j = 1. \end{aligned} \tag{4}$$

Notably, problem (4) is also subject to the implicit domain constraints $q_j \geq 0$, $\pi_j \geq 0$, and $q_j \leq \pi_j$, but all constraints are linear. It is possible to solve on the order of tens of thousands of such small problems per second on a modern personal computer, using a variety of off-the-shelf open-source or commercial solvers. For example, one can substitute the equality constraint into the objective function, add a logarithmic barrier for the box constraints, and solve with Newton's method or the BFGS algorithm (Ruszczynski 2011).

Problem (4) facilitates association testing within the template of the generalized likelihood-ratio test (GLRT), contrasting the maximum likelihood under the alternative hypothesis with that under the null. The GLRT statistic under the null is asymptotically distributed as a centered chi-square random variable with 3 d.f. As described below, this procedure readily incorporates standard population-level assumptions that can be used to decrease the degrees of freedom of the test.

The multivariate case: Case–control–population analysis can be similarly applied for detecting associations that involve multiple SNPs working together. The methods we develop here are equally applicable to interactions of an arbitrary number of SNPs. For exposition purposes, however, we describe our methods and provide results in the context of pairwise interactions, where power is desperately needed but detection is still a feasible task.

Similarly to the univariate case, let $\vec{x} = (x_1, x_2)$, with $x_1 \in \{0, 1, 2\}$ and $x_2 \in \{0, 1, 2\}$, be a pair of diploid SNPs. Let $n_{i,j,k}$ denote the observed counts in $\mathcal{C}$ with $y = i$, $x_1 = j$, $x_2 = k$, and let $n_i = \sum_{j,k} n_{i,j,k}$ and $n_{j,k} = \sum_i n_{i,j,k}$ be the disease status and pairwise genotypic counts, respectively. The pairwise genotype counts in $\mathcal{P}$ are denoted $m_{j,k}$. Let $p_{i|j,k}$ be the probability of disease status i given the pair of genotypes and $\pi_{j,k}$ be the bivariate marginal distribution of the pair.

Testing for pairwise genetic association with disease is typically formulated using a logistic parameterization of the penetrance,

$$p_{1|j,k} = \frac{1}{1 + \exp(-\xi_{j,k})},$$

where $\xi_{j,k}$ may take different forms, depending on the specific formulation (see *Discussion*). The approach we focus on here tests for *association while allowing for interaction* (Cordell 2009). In this approach one is interested in the overall significance of the model for $\vec{p}$ that includes effects at both loci, compared to a null model that includes neither,

$$\begin{aligned} \xi_{j,k}^{\text{flat}} &= \mu \\ \xi_{j,k}^{\text{full}} &= \mu + \alpha_1 \mathbb{I}_{j=1} + \alpha_2 \mathbb{I}_{j=2} + \beta_1 \mathbb{I}_{k=1} + \beta_2 \mathbb{I}_{k=2} \\ &\quad + \gamma_{1,1} \mathbb{I}_{j=1,k=1} + \gamma_{1,2} \mathbb{I}_{j=1,k=2} + \gamma_{2,1} \mathbb{I}_{j=2,k=1} \\ &\quad + \gamma_{2,2} \mathbb{I}_{j=2,k=2}, \end{aligned} \tag{5}$$

with $\mathbb{I}$ being the indicator function.

By writing the log-likelihood of the pooled data and repeating the steps taken in the univariate case, we obtain the MLE as the standard convex form,

$$\begin{aligned}\underset{\vec{q},\vec{\pi}}{\text{minimize}} \; & -\sum_{j,k}\left(n_{0,j,k}\log\left(\pi_{j,k}-q_{j,k}\right)+n_{1,j,k}\log q_{j,k}\right.\\ & \left.+m_{j,k}\log\pi_{j,k}\right)\\ \text{subject to} \; & \sum_{j,k} q_{j,k}=K,\\ & \sum_{j,k}\pi_{j,k}=1,\end{aligned} \tag{6}$$

with $q_{j,k} = p_{1|j,k}\pi_{j,k}$. Now the full parameter space is of dimension 17 (again, population assumptions will reduce this dimension). The same optimizers used for solving problem (4) can be used for problem (6) to fit thousands of pairwise models per second. While on a single personal computer this may still be too computationally demanding for a genome-wide exhaustive pairwise search with its billions of candidates, we show in *Results* that this approach can be combined with a filter such as the fast methods mentioned in the Introduction. An appropriate filter rapidly generates a ranking of all pairs according to a score that is roughly related to the likelihood-ratio statistic of interest. Subsequently, millions of top-ranking candidate pairs can be processed with our approach, and the whole analysis takes <1 hr in the case of the WTCCC data.

The flat model, $\xi^{\text{flat}}_{j,k}$, is again trivially solved by

$$\mu = K, \quad \pi_{j,k} = \frac{n_{j,k}+m_{j,k}}{n+m}.$$

Incorporating distributional assumptions

The nonparametric tests described above (where the specifications of $\vec{p}$ and $\vec{\pi}$ are saturated with respect to the included SNPs) are consistent against all forms of dependence and genotypic distributions. However, when additional assumptions can be made about the nature of plausible effects and distributions, the space of underlying estimated parameters is reduced, and testing can become more efficient (Song and Nicolae 2009; Zheng *et al.* 2012). A considerable amount of attention has been given in the literature to developing such testing approaches. Some of the most widely used assumptions for simplifying $\vec{p}$ are allelic, dominant, recessive, and additive SNP effects (Sasieni 1997; Freidlin *et al.* 2009), which can be tested with the aforementioned nonparametric approaches similarly to standard ones, *i.e.*, by using transformed genotypes or treating them as continuous (Zheng *et al.* 2012).

As for population assumptions that constrain the parameter space for π, two popular and broadly applicable ones are HWE (the independence of maternal and paternal allele values that make up each SNP) and LE (the independence of the two loci that make up a pair) (Yang *et al.* 1999; Chatterjee and Carroll 2005; Zhao *et al.* 2006). Although such assumptions have their limitations (Albert *et al.* 2001; Mukherjee and Chatterjee 2008), HWE can be expected to hold for homogeneous populations, and LE is plausible in the case of distal locus pairs, *i.e.*, almost all pairs considered in a typical GWAS (Chen and Chatterjee 2007; Song and Nicolae 2009). Consider, for example, the pairwise case. Under LE, $\pi_{j,k} = \pi^{(1)}_j \pi^{(2)}_k$, where $\pi^{(1)}$ and $\pi^{(2)}$ are the marginal distributions of x_1 and x_2 genotypes in the population, respectively. This assumption reduces the dimension of the parameter space (and the degrees of freedom of the GLRT) by 4. Further, under HWE at x_1, suppose the minor allele frequency is f_1; then we have that $\pi^{(1)}_0 = (1-f_1)^2$, $\pi^{(1)}_1 = (1-f_1)f_1$, $\pi^{(1)}_2 = f_1^2$—eliminating an additional parameter. Assume HWE at x_2 has the same effect (with the minor allele frequency f_2 parameterizing $\pi^{(2)}$). Thus problem (6), for example, simplifies to

$$\begin{aligned}\underset{\vec{q},f_1,f_2}{\text{minimize}} \; & -\sum_{j,k}\left(n_{0,j,k}\log\left(\pi_j(f_1)\pi_k(f_2)-q_{j,k}\right)\right.\\ & \left.+n_{1,j,k}\log q_{j,k}+m_{j,k}\log\left(\pi_j(f_1)\pi_k(f_2)\right)\right)\\ \text{subject to} \; & \sum_{j,k} q_{j,k}=K.\end{aligned} \tag{7}$$

Other methods and design of simulation study

Assuming known population parameters: Since population data contain only information about population parameters, it is interesting to compare in our simulation study the performance of the GLRT based on problems (4), (6), and (7) to an unrealistic test, where the genotype frequencies are known (or, under HWE, and under LE for the pairwise case, where the minor allele frequencies are known). This represents a theoretical upper bound on performance when infinitely many population samples are available. In the univariate case, for example, one has to maximize with respect to $\vec{p}$ the log-likelihood:

$$l(\vec{n};\vec{p}) = \sum_{i,j} n_{i,j}\log\left(p_{i|j}\frac{\pi_j}{K}\right) \propto \sum_{i,j} n_{i,j}\log\left(p_{i|j}\right). \tag{8}$$

This is already a concave optimization problem with the affine equality constraint

$$\sum_j p_{1|j}\pi_j = K,$$

and box constraints, and as such can be handled efficiently. Treatment of the multivariate case is similar.

The existing naive approach: A simple way of testing for association, while exploiting an extraneous population sample, is to include the samples $\mathcal{P}$ as additional controls and perform standard association testing on the resulting extended case–control sample, denoted $\mathcal{C}^+$. This approach is taken, for example, in the univariate "expanded reference group analysis" of the WTCCC study (Burton *et al.* 2007) (although there are some additional issues to consider in that context, as we discuss in *Results*). While this approach

Table 1 Summary of testing approaches

Name	Description
Case–control G	The $2 \times d$ GLRT of the phenotype y and a genotype (or a pair of genotypes) treated as a $d = 3$ (or $d = 9$) category variable.
Chen and Chatterjee G	The univariate case–control Wald test assuming HWE among controls (Chen and Chatterjee 2007).
Case-only G	The pairwise 3×3 GLRT examining the dependence between genotypes x_1 and x_2 among cases ($y = 1$).
Three-way G	The pairwise $2 \times 3 \times 3$ GLRT, testing for three-way independence of genotype x_1, genotype x_2, and the phenotype.
Known P_x	The (unrealistic) GLRT based on Equation 8, where marginal/pairwise SNP distribution and the prevalence are known.
CMLE	The GLRT based on Equation 4 (for a single SNP) or Equation 6 (in the pairwise case), assuming known prevalence.
CMLE HWE LE	Same as CMLE above, but also assuming HWE and LE.

In boldface type are the novel tests that are introduced in this article. "Known P_x" is unrealistic and merely represents a theoretical upper bound on the benefit of exploiting population samples. The remaining tests represent existing approaches.

mislabels as controls any cases that exist in $\mathcal{P}$, leading to a potential loss of power, this loss will be small for a rare disease. Also, the mislabeling of true cases in $\mathcal{P}$ acts as another (independent) source of noise, meaning that tests that are valid on $\mathcal{C}$ are also valid when applied to $\mathcal{C}^+$ and do not suffer increased type I error. Finally, the simplest approach ignores the extraneous population data altogether, keeping to $\mathcal{C}$ alone.

To represent existing testing methods, we consider four popular approaches. The first approach is standard case–control analysis, which is represented by the logistic GLRT contrasting the saturated and flat models of Equation 5. This test is also known as a "G test" of independence in the two-dimensional contingency table formed by $y \times \vec{x}$. For univariate testing this table is 2×3, while for pairwise testing it is 2×9. This table may be compiled either from $\mathcal{C}$ or from $\mathcal{C}^+$. This test does not exploit the HWE and LE assumptions, nor does it make use of the known prevalence of disease. It is worth noting that the commonly used Pearson chi-square test is an asymptotic approximation to the G test. We refrain from using the chi-square test because it generally gives less accurate P-values at the far tails that are of interest in pairwise testing. The second approach is that of Chen and Chatterjee (2007). It applies to the univariate case only and assumes HWE among controls (and therefore, for rare disease, approximately HWE in the population).

The two remaining approaches are relevant only for the pairwise testing scenario. In the case-only 3×3 GLRT, one uses the contingency table for x_1 and x_2 in cases only. Thus using $\mathcal{C}^+$ (which has the same cases, but more controls) for compiling the table provides no benefit over using $\mathcal{C}$. This test assumes LE among the controls population (or in the general population, under a rare disease assumption) and ignores *any* marginal association signal (Piegorsch *et al.* 1994; Song and Nicolae 2009). While this test is a popular approach to harnessing the LE assumption, by definition, it cannot benefit from including extraneous population samples as additional controls. We therefore consider the related "three-way independence" $2 \times 3 \times 3$ G test, which uses the full three-dimensional case–control contingency table $y \times x_1 \times x_2$. This test assumes LE in the population under the null of no association, but not under the alternative, thus wasting degrees of freedom on modeling a dependence between the SNPs. If the disease is rare, then the degrees of freedom invested in modeling the dependence between the SNPs among controls go to waste as well, as this can capture only spurious dependence signals. Contrary to the case-only test, the three-way independence test is sensitive to marginal effects.

Design of the simulation study: The association testing methods described above are summarized in Table 1. The methods differ in the assumptions they make and are further classified as (i) ignoring $\mathcal{P}$ and using $\mathcal{C}$ only as done in most GWAS to date, (ii) extending $\mathcal{C}$ by including $\mathcal{P}$ as if they were additional controls (denoted $\mathcal{C}^+$), or (iii) correctly analyzing all samples [referred to as ($\mathcal{C}$, $\mathcal{P}$) analysis]. The "known P_x" and case-only G tests are applied only to $\mathcal{C}$ (since the former cannot benefit from population samples, and the latter ignores any control data), and the case–control and three-way G tests and the method of Chen and Chatterjee (2007) are applicable either to $\mathcal{C}$ or to $\mathcal{C}^+$. The CMLE test can be applied to $\mathcal{C}$ and $\mathcal{C}^+$ and is the only method that can be applied to ($\mathcal{C}$, $\mathcal{P}$).

Several problem parameters affect testing performance, notably samples sizes,

$$n_0 = \sum_{i=0,j} n_{i,j}, \quad n_1 = \sum_{i=1,j} n_{i,j}, \quad m = \sum_j m_j$$

(here for the univariate testing scenario), and the true values of K and π, the latter possibly including deviation from LE and HWE.

As the basis for simulated alternatives (*i.e.*, nonnull setups) in the univariate case we use the six partitions of the three genotypes in two risk levels: an "at risk" set of genotype values G_1 *vs.* a "protected" set G_0. These models include the recessive, dominant, and heterozygous partitions (*e.g.*, recessive minor-at-risk: $G_0 = \{0, 1\}$, $G_1 = \{2\}$). An "effect size" parameter is defined, which governs $p_{1|j\in G_1} - p_{1|j\in G_0}$,

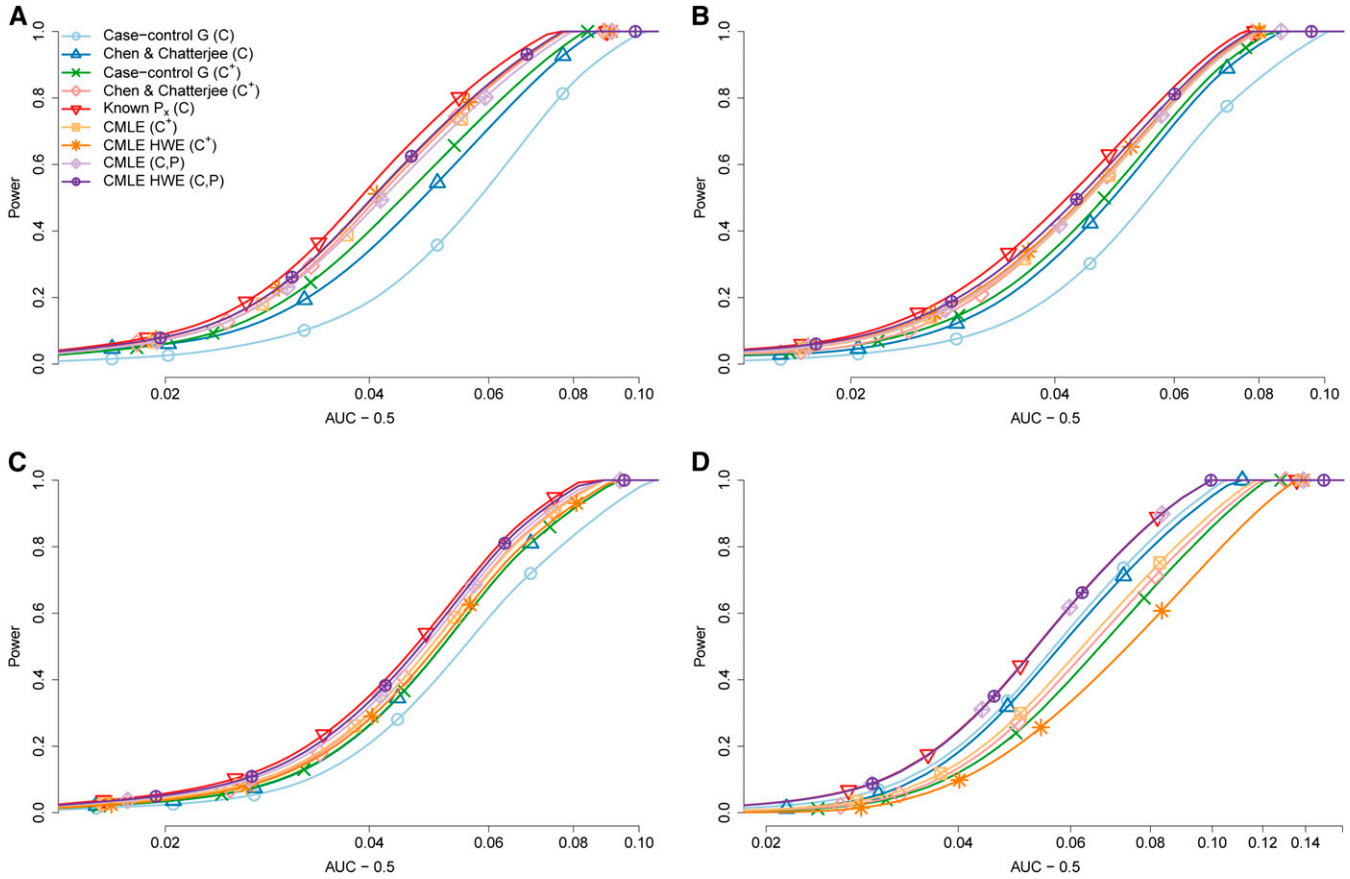


Figure 1 Univariate power simulation. Power is averaged across univariate models with two risk levels. AUC is area under ROC curve, a measure of detection difficulty (see main text). A–D show results for prevalences (K) 0.05, 0.1, 0.2, and 0.4, respectively.

the difference in risk levels between the two groups. Risk levels themselves for G_1 and G_0 are determined from the remaining parameters.

The minor allele frequency (MAF) is taken uniformly from [0.05, 0.5]. This mimics the prior distribution typical to data genotyped using SNP array technology (Waye 2005). We repeat the simulations with prevalence of disease $K \in \{0.05, 0.1, 0.2, 0.4\}$ spanning the common disease spectrum, *e.g.*, bipolar disorder to hypertension to obesity. Sample sizes are fixed at a balanced $n_0 = n_1 = 1000$, and we assume a modest $m = 5000$ independent population samples are also available. Nominal test levels are set according to the 0.05 Bonferroni correction for the number of loci in the study, M, assuming throughout $M = 300{,}000$.

In the pairwise GWAS scenario, we focus on studying association while allowing for interaction, as in Equation 5. Because in this case we are not testing directly for a "pure" interaction, marginal associations ("main effects") alone can lead to a successful discovery. A pairwise GWAS is typically preceded, however, by a univariate study (see, for example, the classification in Liu *et al.* 2011 to marginal, conditional, and pairwise testing), and this makes pairs with marginally detectable SNPs less interesting. We thus perform a marginal association test at both loci and remove simulated data sets from consideration if either univariate test leads to a successful detection. In other words, what is measured is the power of the tests, conditional on not having a marginally detectable signal.

As in the univariate case, we use as alternative hypothesis setups all the possible partitions of the 3×3 pairwise SNP–SNP covariate space into two risk levels. These models are the same as the fully penetrant models of Li and Reich (2000), which have been used in multiple GWAS simulations (Evans *et al.* 2006; Song and Nicolae 2009; Wan *et al.* 2010), except that we admit partially penetrant models by controlling the risk levels. The MAF changes independently for each SNP, and we use the same sample sizes as in the univariate simulation described above. The nominal test level in the pairwise case is adjusted for all $M(M-1)/2$ locus pairs.

Results

Simulation study

An extensive simulation was performed to compare the tests from Table 1 on the basis of statistical power to detect univariate associations and pairwise interactions under various

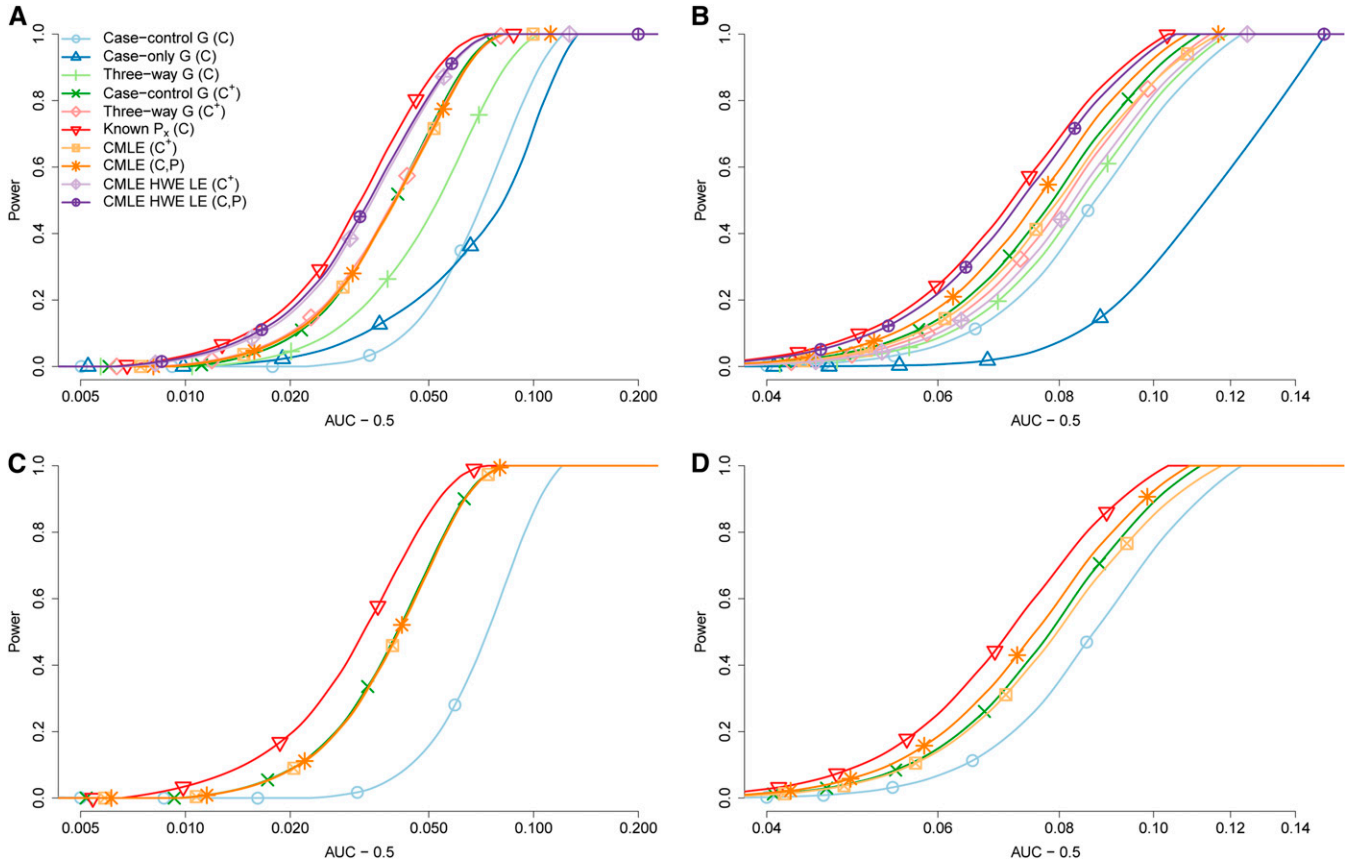


Figure 2 Pairwise power simulation. Power is averaged across all materially different models as described in Li and Reich (2000) with our extensions (see main text). The horizontal axis is defined as in Figure 1. (A and B) Under LE. (C and D) Under LD with $\rho^2 = 0.01$. (A and C) Disease prevalence 0.05. (B and D) Disease prevalence 0.2.

disease models. Additional simulations compared performance under deviation from the modeling assumptions.

Power analysis: We first examined the performance of the methods in the single-SNP scenario. Power is averaged over all models described in *Methods* and presented in Figure 1. The horizontal axis is chosen to objectively reflect problem difficulty—the true area under the curve (AUC) is the probability of the penetrance of a randomly chosen case being higher than that of a randomly chosen control (Jostins and Barrett 2011). It is immediately evident that, as expected, the unrealistic approach that uses an oracle for the marginal SNP distribution is uniformly most powerful among the methods compared. The least powerful approach in all scenarios except $K = 0.4$ is the one most often used in actual studies—the case–control analysis that ignores any population samples. Even with a modest population sample, naive use of these data can greatly improve detection power when K is small.

For higher prevalence of disease, however, the effect of mislabeling cases in the population sample as if they were controls becomes more evident, until for $K = 0.4$ this makes all methods using this approach substantially inferior to ignoring the population sample altogether. Exploiting HWE in controls pays off for low K. This is also expected, since when the disease is relatively rare, the HWE that holds in the population also holds approximately among controls. For $K = 0.4$ this is no longer true, and it is better to perform the standard case–control test.

The inclusion of four variants of our proposed CMLE test, combined with the selection of existing tests, allows us to examine the individual benefits obtained by the two ingredients of our approach, namely, correct handling of population samples and use of the HWE assumption. For low prevalence of disease, there is little to gain from the correct $(\mathcal{C}, \mathcal{P})$ analysis, but HWE (and use of the known K) provides a consistent modest edge. Comparing in this case CMLE HWE to the approach of Chen and Chatterjee (2007), again we see how there is little difference between assuming HWE in controls and in the population. These differences, however, grow with K, and the effect of each ingredient becomes apparent. Finally, we note that the power of the most appropriate test, CMLE HWE $(\mathcal{C}, \mathcal{P})$, is the highest of all the realistic tests in all cases and quickly approaches the theoretical upper bound of known marginals, which means the potential benefits from the exogenous data are fully realized by our approach.

Next we considered the pairwise GWAS scenario, where we were first interested in the power gains under the HWE

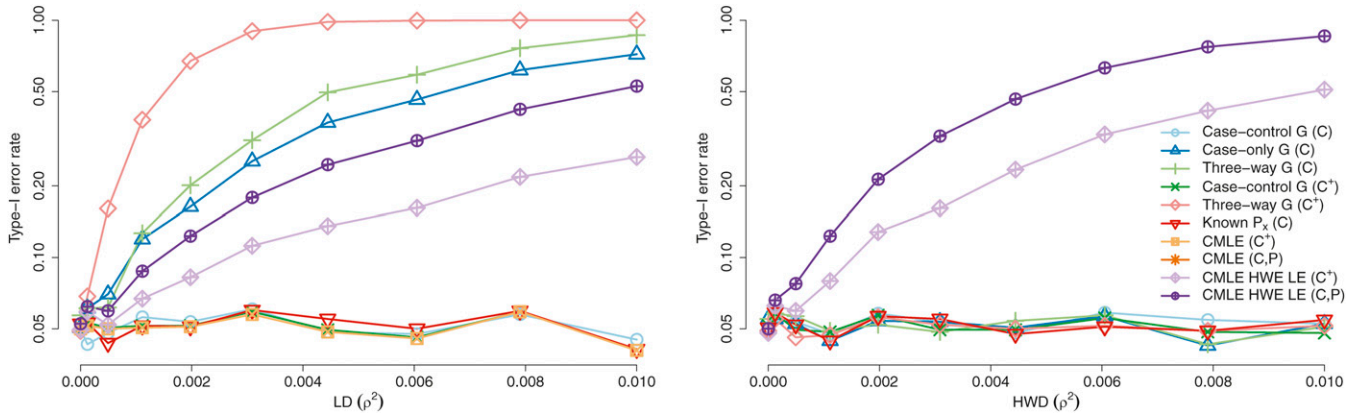


Figure 3 Type I error under deviation from assumptions. Left, LE; right, HWE. ρ is the allelic Pearson correlation coefficient (for LD, between any allele in x_1 and any allele in x_2; for HWD, between the two alleles of the same SNP, with both SNPs simulated to have the same degree of HWD for simplicity).

and LE assumptions. Figure 2, A and B, summarizes the results across all considered models. Supporting Information, File S1, Figure S1, Figure S2, Figure S3, Figure S4 includes detailed results for each Li and Reich (2000)-like model class. Given that few interactions have been detected in all-pairs studies (exhaustive or otherwise) conducted so far, interest lies mainly in those setups where current methods have low power. Incorporating available population samples or known parameters clearly offers substantial power gains for interaction detection under many of the possible setups that fall in the region of interest. Not surprisingly, the best approach is again the unrealistic one, where the pairwise distribution is known (or in this case of HWE and LE, the MAFs at both loci are known). Here too, the performance of this approach defines an upper bound for the realistic approaches, and the important conclusion stemming from these results is that the suggested CMLE test approaches this bound. Notably, the gains from performing $\mathcal{C}^+$ or ($\mathcal{C}$, $\mathcal{P}$) analysis beyond standard case–control $\mathcal{C}$ or case-only analysis are great, reaching sevenfold power gain, averaged over all models (when power increases from ~0.1 of G ($\mathcal{C}$) to ~0.7 of CMLE HWE LE ($\mathcal{C}$, $\mathcal{P}$) in Figure 2A). The differences between ($\mathcal{C}$, $\mathcal{P}$) and $\mathcal{C}^+$ are not as great, but become more substantial for higher prevalence values.

We repeated the simulation under a modest amount of linkage disequilibrium (LD) ($\rho^2 = 0.01$), applying only methods that do not rely on the LE assumption. The results were similar, but naturally some of the power gains in the CMLE approach were attenuated, to the point it is clearly superior to standard $\mathcal{C}^+$ analysis only for prevalence >0.1 (see Figure 2, C and D).

Sensitivity analysis: We compared the type I error of the tests to quantify the effects of deviation from HWE and LE assumptions. Figure 3 shows type I error rates for the pairwise scenario with $n_0 = n_1 = 1000$, $m = 5000$, and the MAF at both loci held at 0.3. In Figure 3, left, there is HWE, and the degree of LD between the loci, measured by Pearson's product-moment correlation coefficient squared, is varied. In Figure 3, right, LE holds but there is a varying degree of Hardy–Weinberg disequilibrium (HWD) (measured again by the squared correlation, here between the two alleles at the same locus). In all tests, the nominal type I error rate is set to 0.05.

As long as modeling assumptions are met, all tests maintain the nominal level as expected. In accordance with similar past observations (Albert *et al.* 2001; Mukherjee and Chatterjee 2008), methods that rely on LE, namely, the case-only, three-way independence, and CMLE LE tests, are sensitive to this assumption and quickly break down under LD. Of these methods, CMLE is the most robust to departure from LE. The sensitivity to deviation from HWE is high as well. This calls for discretion when applying such tests and for use of the general CMLE in the case of doubt regarding the validity of assumptions. As mentioned in the Introduction and as elaborated in the *Discussion*, the HWE and/or LE assumptions will often be applicable for random sampling in homogeneous populations and for distal pairs; however, it is good practice to perform replication analysis with tests that do not make these assumptions to minimize the chance of a false discovery.

Application to the WTCCC study

The proposed CMLE tests were used to analyze the data sets from the WTCCC study (Burton *et al.* 2007). The WTCCC data contain independent case cohorts for seven common diseases: bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), Crohn's disease (CD), rheumatoid arthritis (RA), type I diabetes (T1D), and type II diabetes (T2D). Each case cohort is on the order of 2000 individuals, and an additional shared cohort is available with ~3000 population controls (*i.e.*, controls were not screened to exclude cases of the seven diseases). Each individual was genotyped at ~500,000 SNP loci, which after standard quality control (QC) reduces to ~350,000 loci. For pairwise testing we follow the recommendations from Liu *et al.* (2011) and further filter out SNPs with MAF < 0.1, leading to ~300,000 loci.

In the main analysis undertaken by the WTCCC (Burton *et al.* 2007), each case cohort is analyzed separately against the shared controls, when the latter are treated as pure controls rather than population controls. This simplifies the work and is a reasonable approximation since the prevalences of the (sometimes nonstandard) phenotype definitions considered by WTCCC are not high [the prevalences for the seven phenotypes often quoted for the relevant UK population vary between 0.001 and 0.15, but the most common phenotypes, such as HT, are actually defined in the WTCCC study as *extreme cases* of HT, leading to lower prevalence (Burton *et al.* 2007)]. This approach is essentially a $\mathcal{C}^+$ analysis with zero pure controls. In the expanded reference group analysis (ERGA), to increase power, Burton *et al.* (2007) extend the controls set for any one disease, using the cases from remaining diseases that are thought to be unrelated (autoimmune diseases CD, RA, and T1D are related and so are the cardiovascular phenotypes CAD and HT). Because there is no guarantee that a case in an unrelated data set is not also a case with respect to the current phenotype, this too is a $\mathcal{C}^+$ analysis, using additional population samples. Both analyses in Burton *et al.* (2007) are univariate, but several other authors have since performed interaction testing on these data (*e.g.*, Emily *et al.* 2009; Liu *et al.* 2011; Prabhu and Pe'er 2012; Lippert *et al.* 2013).

The pairwise problem is less extensively studied, more challenging, and according to the simulation results above holds the greatest potential for benefiting from our methodology. We therefore performed a pairwise analysis where the shared controls and extending case cohorts were taken as *population* samples. This is more appropriate almost everywhere in the genome, except for loci associated with the unrelated phenotypes—which we excluded. In addition, we imposed the prevalence numbers obtained from UK health organizations (Burton *et al.* 2007; Allender *et al.* 2012).

To speed up the processing of the seven data sets, pairs were first filtered using the software package PIAM (Liu *et al.* 2011), such that SNPs with significant association signals at the univariate Bonferroni level were removed. Then pairwise testing was performed over all remaining SNPs that show some marginal signal ($P < 0.1$), using an optimized implementation of the unconstrained G test applied to $\mathcal{C}^+$ (see *Methods*). The 100,000 pairs with the smallest P-value from this analysis that are also well separated (at least 5 Mbp away) were further analyzed by our CMLE HWE LE approach. The entire process was completed in ~30 min per data set on a modern personal computer (using an exhaustive pairwise search filter is feasible as well and adds several hours of running time per data set).

Overall, across six of the seven diseases, 736 pairs were detected that are significant at the pairwise Bonferroni level ($\approx 10^{-12}$, varying slightly between data sets) and contain distal SNPs. After filtering pairs with SNPs that are marginally significant at the univariate Bonferroni level according to the univariate CMLE HWE test (which as we have seen above has more power than the $\mathcal{C}^+$ G test used in the PIAM filter), 9 pairs in four diseases survive. Seven of these pairs, in three diseases, were not implicated by the original WTCCC study (defined as both SNPs being at least 1 Mb away from any WTCCC significant association). These results are summarized in Table 2. Table 3 briefly highlights some of the most promising pairwise findings overall, and Figure 4 and Figure 5 give the underlying contingency tables and estimated odds ratios for a few of these top pairs. The detailed listing of all results is provided in File S1, and a short discussion for each disease is given next.

Table 2 Significant pairwise findings in WTCCC

Disease	BD	CAD	CD	HT	RA	T1D	T2D
Significant pairs	0	134	3	576	12	8	3
With weak marginals	0	2	3	0	2	2	0
Implicating novel loci	0	2	3	0	2	0	0

Pairs with weak marginals are defined as pairs where both SNPs have univariate CMLE *P*-values larger than the relevant univariate Bonferroni significance level. Novel loci are defined as pairs where both SNPs are >1 Mb away from any association reported in the original WTCCC study.

BD: Although no significant pairs were detected at the (somewhat arbitrary and probably quite conservative) Bonferroni level (1.09×10^{-12}), it is interesting to note that the pair ranked first (rs9865654, rs17600642), with a pairwise P-value of 1.13×10^{-12}, has a leg on chromosome 10 (rs17600642), which was not detected in the ERGA, but has been noted more recently (Jiang and Zhang 2011). Because of our more stringent filtering, our data do not include rs10925490 (MAF $\approx$ 0.06) and thus the pair identified by Prabhu and Pe'er (2012).

CAD: The immediate regions of the SNPs composing the topmost pair (rs5007171, rs2329902) have not been reported in past GWAS results [according to the GaP and dbGaP catalogs (Mailman *et al.* 2007)], but because rs5007171 has a strong marginal signal (P-value of 7×10^{-6}), it is possible that this result is due to a marginal association here. Many of the following ranked pairs also include rs5007171 and are thus suspect as well. On the other hand, a noteworthy significant pair with weak marginals is (rs16905928, rs3781575) which repeats with small SNP location shifts. This pair shows a mostly recessive–recessive effect; see Figure 4.

CD: The highest ranked pair (rs962087, rs7028357) appears to be a dominant–dominant association (see Figure 4), where neither SNP has a known association with any disease. The pair ranked second (rs7554511, rs11945978) involves SNPs not implicated in the WTCCC study, but a proximal locus of rs7554511 (rs11584383) was found later to be strongly associated with CD in a larger study (P-value of 10^{-11}) (Barrett *et al.* 2008), in which the WTCCC data were combined with additional sources. This pair is notably also detected as a significant pairwise association in Liu *et al.* (2011), where it is the only pair that is found to be a strictly

Table 3 Promising pairwise associations in WTCCC

Phenotype	SNP_1		SNP_2		P_1	P_2	P_{CO}	P_{CMLE}
BD	rs9865654	3p25.2	rs17600642	10q22.1	7.64×10^{-4}	1.04×10^{-6}	2.66×10^{-6}	1.13×10^{-12}
CAD	rs16905928	10p12.31	rs3781575	11p13	6.15×10^{-3}	1.30×10^{-5}	5.12×10^{-9}	$\mathbf{1.83 \times 10^{-13}}$
CD	rs962087	5p14.1	rs7028357	9p24.1	3.83×10^{-4}	3.24×10^{-5}	2.11×10^{-8}	$\mathbf{1.28 \times 10^{-13}}$
CD	rs7554511	1q32.1	rs11945978	4p12	1.10×10^{-6}	4.93×10^{-6}	1.01×10^{-4}	$\mathbf{4.17 \times 10^{-13}}$
RA	rs1605705	3p26.1	rs6831911	4q34.3	3.58×10^{-5}	5.28×10^{-7}	1.73×10^{-10}	$\mathbf{2.78 \times 10^{-18}}$
RA	rs894848	1p13.2	rs6427122	1q24.2	1.18×10^{-4}	2.04×10^{-5}	2.57×10^{-7}	$\mathbf{3.18 \times 10^{-13}}$

We selected pairs that achieved or came close to achieving genome-wide significance at the 0.05 pairwise Bonferroni level, that have weak marginal signals, and that appear to be novel associations when compared to the discoveries of the original WTCCC study (Burton *et al.* 2007). P_1 and P_2 are *P*-values for the univariate CMLE HWE test for SNP_1 and SNP_2, respectively; P_{CO} is the *P*-value for the (pairwise) case-only test; and P_{CMLE} is the *P*-value for the pairwise CMLE HWE LE test. Strictly significant *P*-values are in boldface type.

significant association. Liu *et al.* (2011) thoroughly analyzed this interaction and have in fact validated it in an independent data set (IBDGC-non-Jewish, which is also a part of the data used in Barrett *et al.* 2008). Both rs7554511 and rs11945978 have already been suggested as marginal associations in Barrett *et al.* (2009) (which studied ulcerative colitis, a closely related condition) and Duerr *et al.* (2006) (studying CD), respectively. Pairs composed of respectively neighboring SNPs have small *P*-values as well and corroborate the association signal (although none are strictly significant).

HT: While the original (univariate) WTCCC study did not produce any significant findings, hundreds of alleged associations are detected in our pairwise analysis. Examining these associations more closely, it seems that all of these are in fact due to a few marginal associations that are just below the univariate Bonferroni threshold. One such association is represented by rs17509005, which repeats in many pairs. It is probably more accurately captured by rs11047543, which has been implicated more recently in Pfeufer *et al.* (2010) with a *P*-value of $10^{-12.5}$, or by rs17287293, described in Eijgelsheim *et al.* (2010) with a *P*-value of $10^{-9.7}$. This marginal association comes up only as a moderate signal (10^{-5}), using standard univariate testing of the WTCCC ERGA, but comes up stronger in our results (10^{-6}).

RA: The first pair here is (rs1605705, rs6831911), showing a dominant–dominant deleterious effect (Figure 5, top); both SNPs are not mentioned in the WTCCC study results. rs1605705 is noted as a moderate association in dbGaP. The second pair (rs894848, rs6427122) has the dominant–dominant pattern again (see Figure 5, bottom). These SNPs were not found by the WTCCC analysis as well, but more recently rs864537, which is <1 Mb from rs6427122, has been implicated in RA (*P*-value of $10^{-10.7}$) (Stahl *et al.* 2010). Another strong marginal proximal association is rs840016 (Zhernakova *et al.* 2011).

T1D: Significant pairs in this data set can be attributed to marginal signals, mostly in rs1377748. The pair (rs2077749, rs10849946) is proximal to a region found on chromosome 12 in the original study; however, careful examination indicates that this is a separate association discovered more recently in two studies applying different methods to the WTCCC data (rs1265564 in Huang *et al.* 2012 with a *P*-value of 10^{-16} and rs3184504 in Plagnol *et al.* 2011 with a *P*-value of 10^{-37}).

T2D: Of the three significant pairs, both SNPs making up the first pair are also implicated in the WTCCC ERGA. The remaining two pairs are due to a marginal effect in rs962087.

Discussion

Existing tests used in GWAS are often not powerful enough to detect modest associations given the available sample sizes. Specifically, the search for effects that are statistically significant only when considering multiple loci jointly is an important ongoing effort that has so far produced very few replicable results. We introduce a powerful new approach for univariate and multivariate testing that is based on constrained maximum-likelihood estimation. Our method makes efficient use of extraneous population samples in the context of case–control, case–population, and case–control–population studies of complex human disease. The underlying modeling approach has the important capability to exploit population assumptions, such as HWE and LE of constituent SNPs.

The resulting association testing approach is quite flexible and shown through extensive simulation to be valid and powerful for a wide range of problem settings. The successful application of our approach to the WTCCC study data demonstrates this as well and shows in particular that an efficient pairwise analysis can uncover novel associations that, when approached by marginal testing, are visible only in larger samples (if at all). These results have an immediate implication for disease association studies in general—most studied populations have samples that are easily accessible to researchers and that can now be utilized to substantially improve power of standard testing approaches. The increasing number and richness of genetic data sources such as those mentioned in the Introduction will make the suggested approach only more relevant.

A critical examination of past results, especially from pairwise analyses, raises many suspicions regarding reported discoveries: there is strong reliance on sparse data due to low

Controls	rs3781575 = 0	1	2
rs16905928 = 0	6970	2617	285
1	1611	615	61
2	104	33	5
Cases			
rs16905928 = 0	1117	392	56
1	215	95	17
2	17	4	13

Controls	rs7028357 = 0	1	2
rs962087 = 0	4868	2249	254
1	1875	1004	109
2	180	64	5
Cases			
rs962087 = 0	749	356	48
1	277	257	22
2	20	15	4

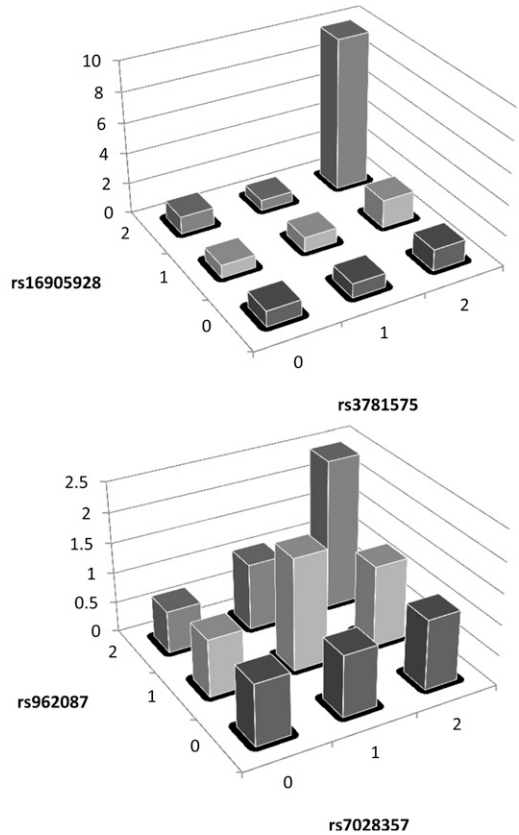


Figure 4 WTCCC promising pairs (part 1). Shown are contingency tables and odds ratios estimated using CMLE HWE LE for promising pairwise associations in CAD (top, OR truncated at 10 for presentation) and CD (bottom) data; see text for discussion.

MAFs (Lippert *et al.* 2013) or LD (Brinza *et al.* 2010; Wan *et al.* 2010; Lippert *et al.* 2013) or on SNPs showing traces of a range of genotyping errors (Cordell 2009; Brinza *et al.* 2010; Liu *et al.* 2011), as well as a host of other problems. Although there are certainly other explanations (Marchini *et al.* 2005), this could help account for the low replicability of such association results. We reiterate the conclusions of Liu *et al.* (2011), who have identified similar problems in other studies. They suggested that quality control for pairwise analysis be stricter than for univariate GWAS, and we have implemented such control in this study. Another issue that contributes to the difficulty of replicating pairwise associations is due to the unconstrained search space employed by most studies. The present article is part of a larger body of work that attempts to improve the situation by restricting the search for multilocus associations to more biologically plausible ones (Chatterjee and Carroll 2005; Wang and Sheffield 2005; Chen and Chatterjee 2007; Song and Nicolae 2009; Han *et al.* 2012; Luss *et al.* 2012). Various combinations of the constraints considered by others and those described here are reasonable, and while it is straightforward to apply them to data, asymptotic theory to support fast testing is not always available.

Many of the limitations discussed in Liu *et al.* (2011) are relevant to our approach as well. It is worth noting a few limitations specific to the additional assumptions taken here. One issue is that joining samples from different sources often requires stratification adjustment, which we expect to address in future work. Another issue noted by various authors is the high sensitivity of tests based on HWE and LE to departure from these assumptions (as is also evident in our simulations), and ways of mitigating this sensitivity have been suggested (Albert *et al.* 2001; Mukherjee and Chatterjee 2008). Indeed, a prudent step following discoveries arising from our analysis is to confirm them on independent data, using an unconstrained test. On the other hand, additional simulations (results not shown) imply that the sensitivity of the suggested approach to misspecification of disease prevalence is low, which is important when accurate estimates are not available, for example in populations of developing countries.

While we have focused on univariate associations and on pairwise associations allowing for interaction, let us stress two points. First, the methods described are equally applicable to SNP subsets of arbitrary order. Second, the literature is rich with other formulations for the genetic interaction problem, and many of them plug naturally into our approach as well. For example, our framework also allows conditional association and pure interaction tests that (similarly to the case-only test) attempt to identify interactions that cannot be due to combinations of marginal effects alone (Cordell 2009; Liu *et al.* 2011). This task is significantly more challenging

Controls	rs6831911 = 0	1	2
rs1605705 = 0	5472	1793	153
1	2154	713	60
2	192	64	7
Cases			
rs1605705 = 0	887	299	20
1	336	241	16
2	39	21	1

Controls	rs6427122 = 0	1	2
rs894848 = 0	4587	3328	602
1	1046	771	151
2	69	42	12
Cases			
rs894848 = 0	744	583	91
1	156	224	28
2	8	23	3

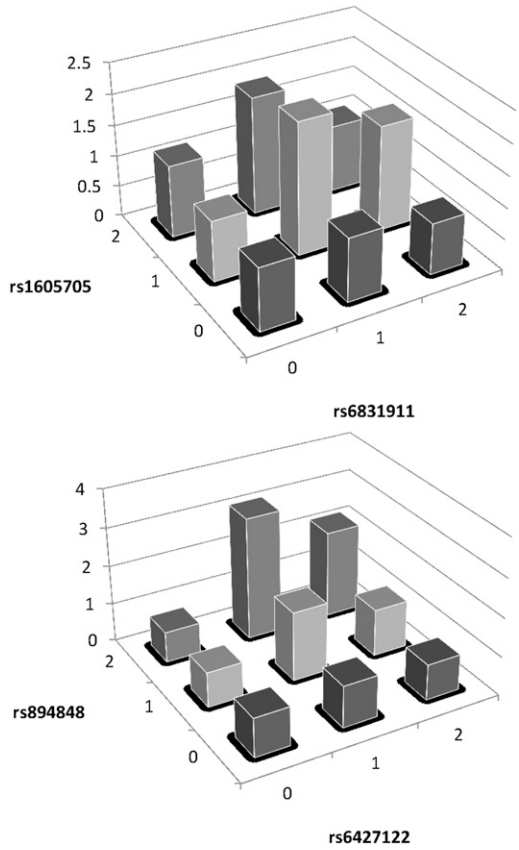


Figure 5 WTCCC promising pairs (part 2). Shown are contingency tables and odds ratios estimated using CMLE HWE LE for promising pairwise associations in RA data; see text for discussion.

(that is, the optimization can be more difficult, and tests may have less power), but we expect our methods when combined with an appropriate filtering method such as that of Prabhu and Pe'er (2012) to be more powerful than existing alternatives in this context as well.

Finally, it would be interesting to examine the benefit of using the methods suggested here in case–control studies of late-onset diseases such as Alzheimer's, because taking controls in a study like this as known to be unaffected is susceptible to the same issues as using population controls. Within our framework the "controls" of such studies can be assumed to be a random population sample and handled properly.

Acknowledgments

The authors thank Yang Liu for supplying additional components for the PIAM software tool used for filtering the WTCCC data, David Golan, an anonymous reviewer, and the associate editor for helpful suggestions. This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv university and by Israeli Science Foundation grant 1487/12. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for that project was provided by the Wellcome Trust under award 076113.

Literature Cited

Albert, P., D. Ratnasinghe, J. Tangrea, and S. Wacholder, 2001 Limitations of the case-only design for identifying gene-environment interactions. Am. J. Epidemiol. 154: 687–693.

Allender, S., V. Peto, P. Scarborough, A. Boxer, and M. Rayner, 2012 *Coronary Heart Disease Statistics*. British Heart Foundation and Stroke Association. London, UK.

Barrett, J., S. Hansoul, D. Nicolae, J. Cho, R. Duerr *et al.*, 2008 Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. 40: 955–962.

Barrett, J., J. Lee, C. Lees, N. Prescott, C. Anderson *et al.*, 2009 Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the hnf4a region. Nat. Genet. 41: 1330–1334.

Bloom, J., I. Ehrenreich, W. Loo, T. Lite, and L. Kruglyak, 2013 Finding the sources of missing heritability in a yeast cross. Nature 494: 234–237.

Boyd, S., and L. Vandenberghe, 2004 *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Brinza, D., M. Schultz, G. Tesler, and V. Bafna, 2010 Rapid detection of gene–gene interactions in genome-wide association studies. Bioinformatics 26: 2856–2862.

Burton, P., D. Clayton, L. Cardon, N. Craddock, P. Deloukas *et al.*, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.

Chatterjee, N., and R. J. Carroll, 2005 Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92: 399–418.

Chen, J., and N. Chatterjee, 2007 Exploiting Hardy-Weinberg equilibrium for efficient screening of single SNP associations from case-control studies. Hum. Hered. 63: 196–204.

Cordell, H., 2009 Detecting gene–gene interactions that underlie human diseases. Nat. Rev. Genet. 10: 392–404.

Couzin-Frankel, J., 2012 U.K. unveils plan to sequence whole genomes of 100,000 patients. ScienceInsider: 10. Available at: http://news.sciencemag.org/scienceinsider/2012/12/uk-unveils-planto-sequence-whol.html.

Duerr, R., K. Taylor, S. Brant, J. Rioux, M. Silverberg *et al.*, 2006 A genome-wide association study identifies il23r as an inflammatory bowel disease gene. Sci. Signal. 314: 1461.

Eichler, E., J. Flint, G. Gibson, A. Kong, S. Leal *et al.*, 2010 Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11: 446–450.

Eijgelsheim, M., C. Newton-Cheh, N. Sotoodehnia, P. de Bakker, M. Müller *et al.*, 2010 Genome-wide association analysis identifies multiple loci related to resting heart rate. Hum. Mol. Genet. 19: 3885–3894.

Emily, M., T. Mailund, J. Hein, L. Schauser, and M. Schierup, 2009 Using biological networks to search for interacting loci in genome-wide association studies. Eur. J. Hum. Genet. 17: 1231–1240.

Evans, D., J. Marchini, A. Morris, and L. Cardon, 2006 Two-stage two-locus models in genome-wide association. PLoS Genet. 2: e157.

Freidlin, B., G. Zheng, Z. Li, and J. Gastwirth, 2009 Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum. Hered. 53: 146–152.

Han, S., P. Rosenberg, and N. Chatterjee, 2012 Testing for gene–environment and gene–gene interactions under monotonicity constraints. J. Am. Stat. Assoc. 107: 1441–1452.

Huang, J., D. Ellinghaus, A. Franke, B. Howie, and Y. Li, 2012 1000 genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 data. Eur. J. Hum. Genet. 20: 801–805.

Jiang, Y., and H. Zhang, 2011 Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. Genet. Epidemiol. 35: 125–132.

Johnson, A., and C. O'Donnell, 2009 An open access database of genome-wide association results. BMC Med. Genet. 10: 6.

Jostins, L., and J. Barrett, 2011 Genetic risk prediction in complex disease. Hum. Mol. Genet. 20: R182–R188.

Li, W., and J. Reich, 2000 A complete enumeration and classification of two-locus disease models. Hum. Hered. 50: 334–349.

Lippert, C., J. Listgarten, R. Davidson, S. Baxter, H. Poong *et al.*, 2013 An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. Sci. Rep. 3: 1099.

Liu, Y., H. Xu, S. Chen, X. Chen, Z. Zhang *et al.*, 2011 Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. PLoS Genet. 7: e1001338.

Luss, R., S. Rosset, and M. Shahar, 2012 Efficient regularized isotonic regression with application to gene–gene interaction search. Ann. Appl. Stat. 6: 253–283.

Mailman, M., M. Feolo, Y. Jin, M. Kimura, K. Tryka *et al.*, 2007 The NCBI dbgap database of genotypes and phenotypes. Nat. Genet. 39: 1181–1186.

Manolio, T., F. Collins, N. Cox, D. Goldstein, L. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461: 747–753.

Marchini, J., P. Donnelly, and L. Cardon, 2005 Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat. Genet. 37: 413–417.

Mukherjee, B., and N. Chatterjee, 2008 Exploiting gene-environment independence for analysis of case–control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64: 685–694.

Pfeufer, A., C. van Noord, K. Marciante, D. Arking, M. Larson *et al.*, 2010 Genome-wide association study of pr interval. Nat. Genet. 42: 153–159.

Piegorsch, W., C. Weinberg, and J. Taylor, 1994 Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat. Med. 13: 153–162.

Plagnol, V., J. Howson, D. Smyth, N. Walker, J. Hafler *et al.*, 2011 Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. PLoS Genet. 7: e1002216.

Prabhu, S., and I. Pe'er, 2012 Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. Genome Res. 22: 2230–2240.

Ruszczynski, A., 2011 *Nonlinear Optimization*, Vol. 13. Princeton University Press, Princeton, NJ.

Sasieni, P., 1997 From genotypes to genes: doubling the sample size. Biometrics 53: 1253–1261.

Shao, H., L. C. Burrage, D. Sinasac, A. Hill, S. Ernest *et al.*, 2008 Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. Proc. Natl. Acad. Sci. USA 105: 19910–19914.

Siva, N., 2008 1000 genomes project. Nat. Biotechnol. 26: 256.

Song, M., and D. Nicolae, 2009 Restricted parameter space models for testing gene-gene interaction. Genet. Epidemiol. 33: 386–393.

Stahl, E., S. Raychaudhuri, E. Remmers, G. Xie, S. Eyre *et al.*, 2010 Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42: 508–514.

Wan, X., C. Yang, Q. Yang, H. Xue, X. Fan *et al.*, 2010 Boost: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am. J. Hum. Genet. 87: 325.

Wang, K., and V. Sheffield, 2005 A constrained-likelihood approach to marker-trait association studies. Am. J. Hum. Genet. 77: 768–780.

Waye, M., 2005 A haplotype map of the human genome. Nature 437: 1299–1320.

Yang, J., B. Benyamin, B. McEvoy, S. Gordon, A. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–569.

Yang, Q., M. Khoury, F. Sun, and W. Flanders, 1999 Case-only design to measure gene-gene interaction. Epidemiology 10: 167–170.

Zhao, J., L. Jin, and M. Xiong, 2006 Test for interaction between two unlinked loci. Am. J. Hum. Genet. 79: 831–845.

Zheng, G., Y. Yang, X. Zhu, and R. Elston, 2012 *Analysis of Genetic Association Studies*. Springer-Verlag, New York.

Zhernakova, A., E. Stahl, G. Trynka, S. Raychaudhuri, E. Festen *et al.*, 2011 Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet. 7: e1002004.

Communicating editor: J. Wall

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162511/-/DC1

Exploiting Population Samples to Enhance Genome-Wide Association Studies of Disease

Shachar Kaufman and Saharon Rosset

DOI: 10.1534/genetics.114.162511

Exploiting Population Samples to Enhance Genome-Wide Association Studies of Disease – Supplementary Material

Shachar Kaufman and Saharon Rosset

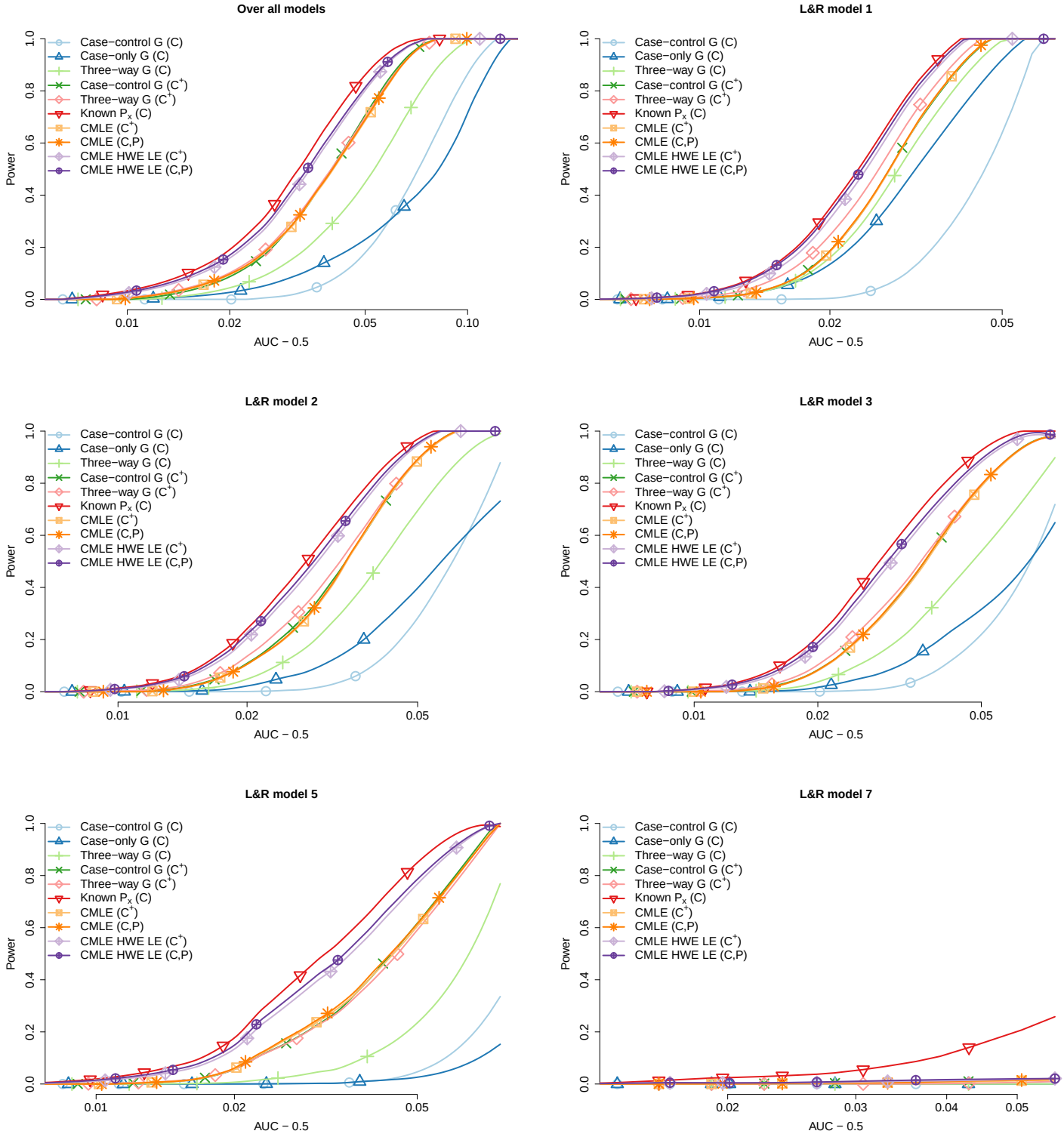
March 8, 2014

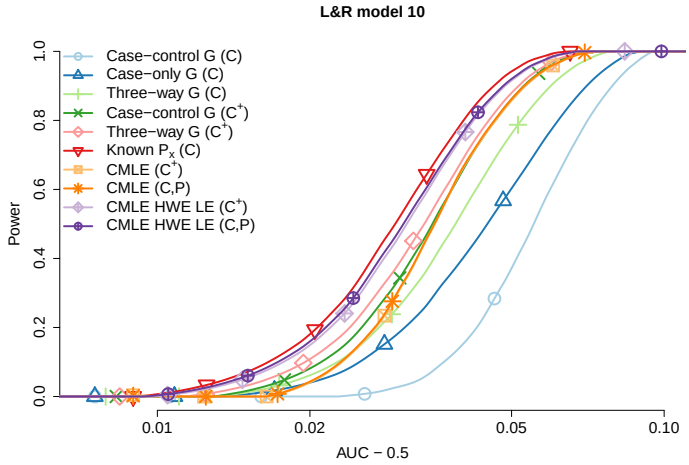
List of Figures

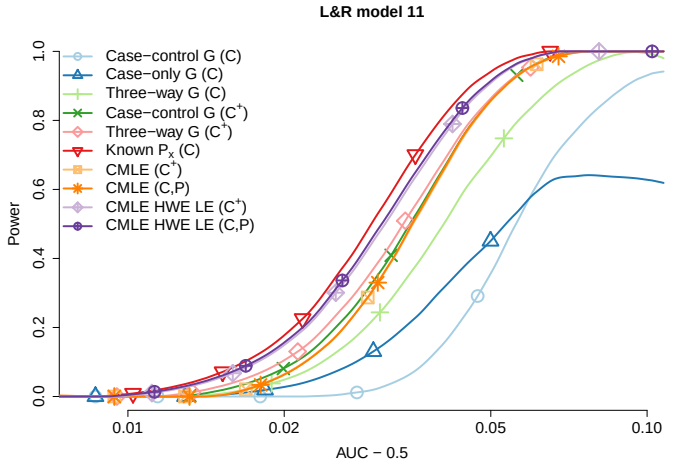
Figures S1 through S4 give detailed results for the pairwise power simulations described in the main text. The 50 two-SNP, two-risk-level base the 50 fully penetrant models of Li and Reich [2000]. These models represent equivalence classes in the space of all 512 possible models, that are not degenerate and cannot be received from one another using a set of fundamental operations such as 1's complement and order reversal. These include disease models considered by earlier literature [Neuman et al., 2005], such as: recessive-recessive (model 1), dominant-dominant (model 27), recessive-dominant (model 3). Model 15 is known a a modifying effect model, model 11 can be seen as a threshold model, model 27 is a XOR (exclusive or) model. Models 7 and 56 are degenerate pairwise associations which are actually univariate.

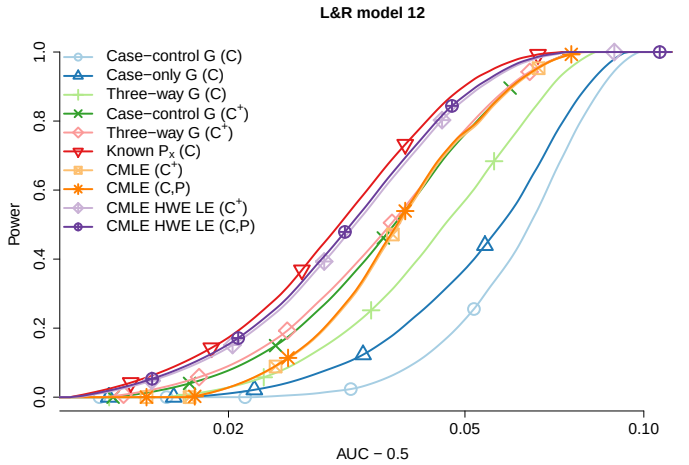
Note: As described in the main text, a wide range of simulation setups was implemented for all settings and models, but we then eliminated all setups where either marginal effects was detectable or where no model had any power. Thus the plots are based on a varying number of "relevant" setups that were left for each base model. We eliminated any plot where less than 20 relevant setups remained.

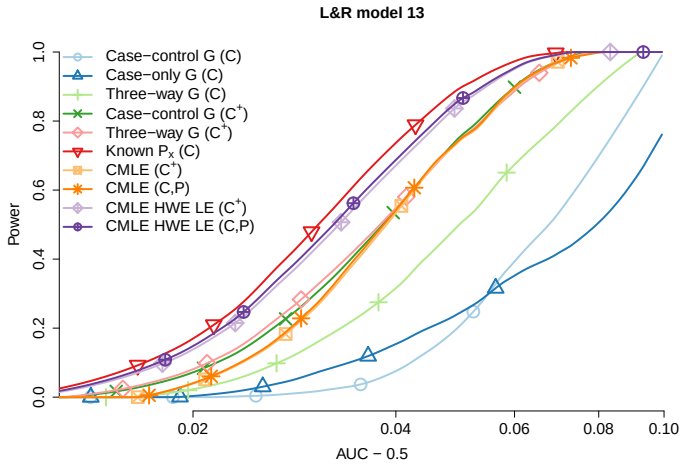
Figure S1: Power simulation results under LE for 0.05 prevalence.

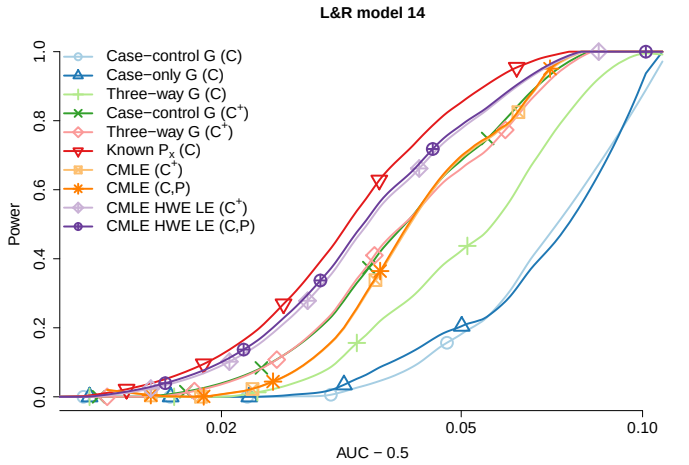


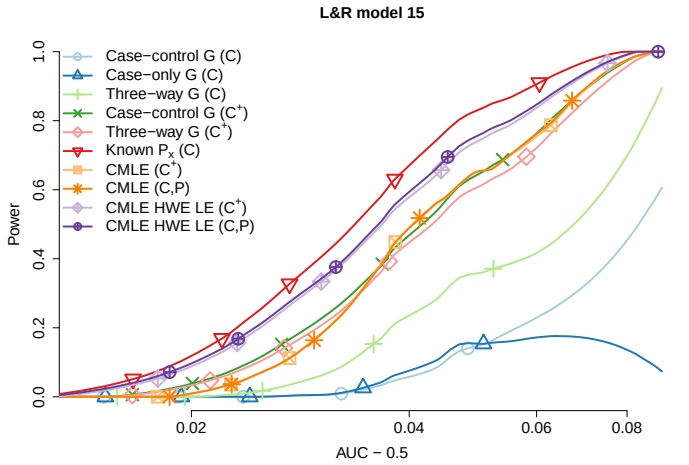


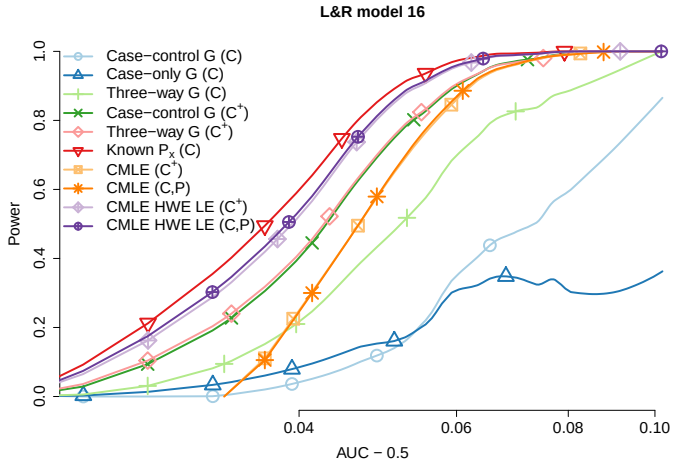


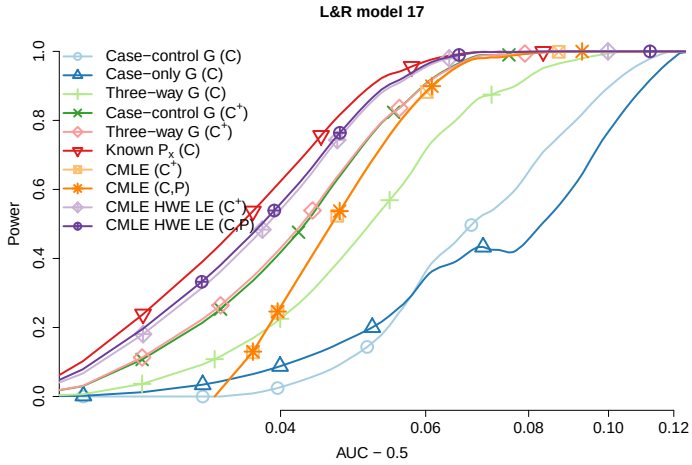


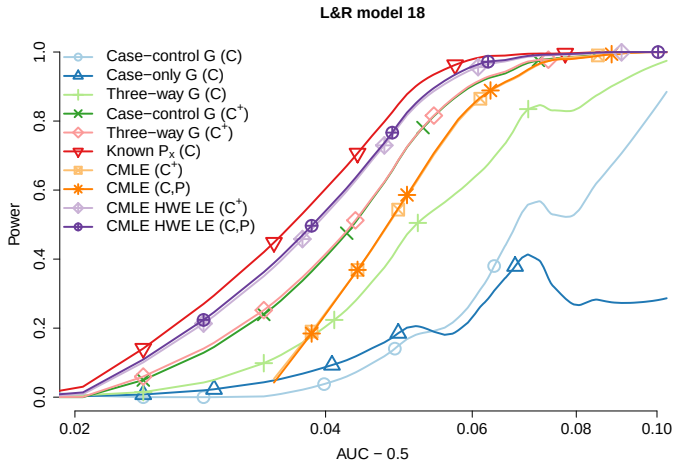


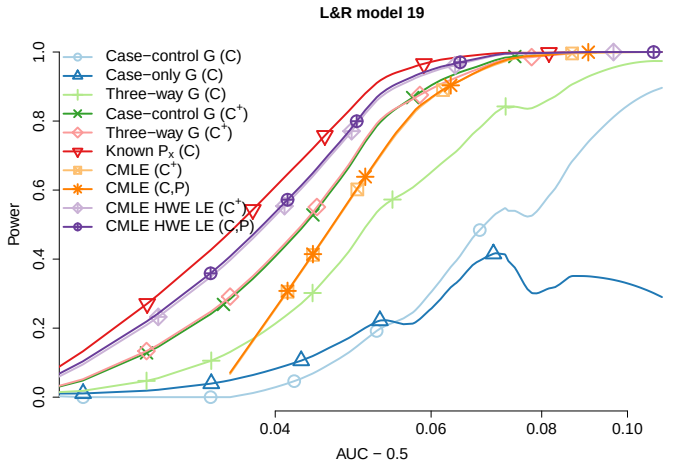


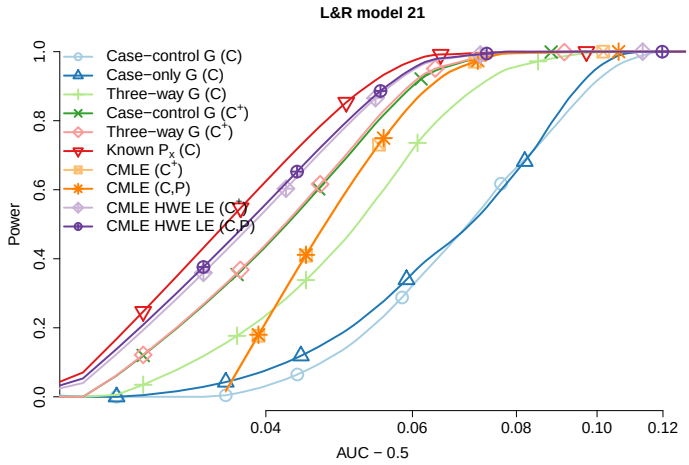


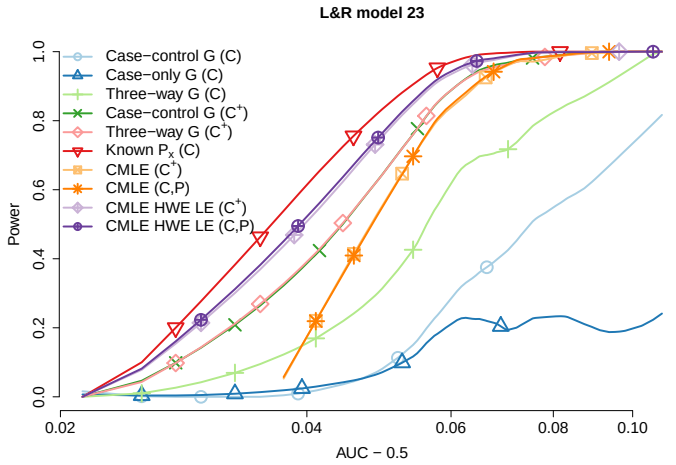


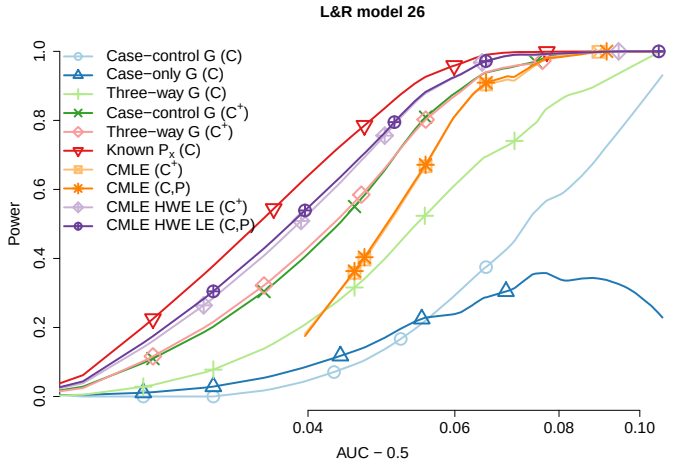


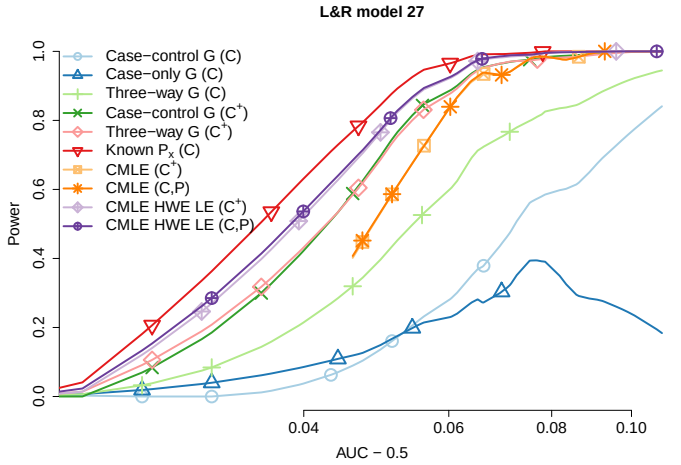


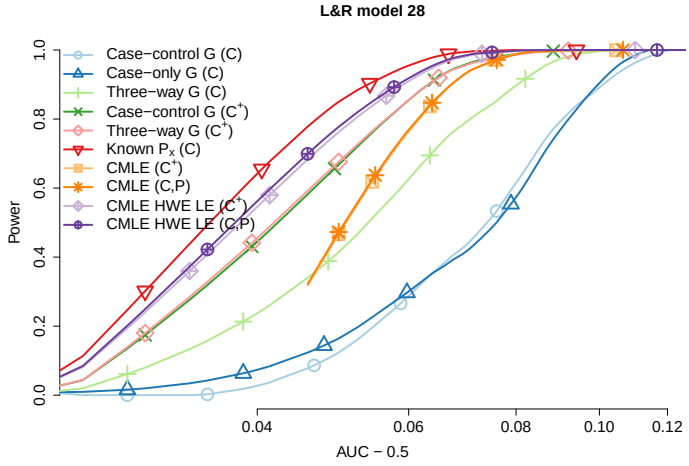


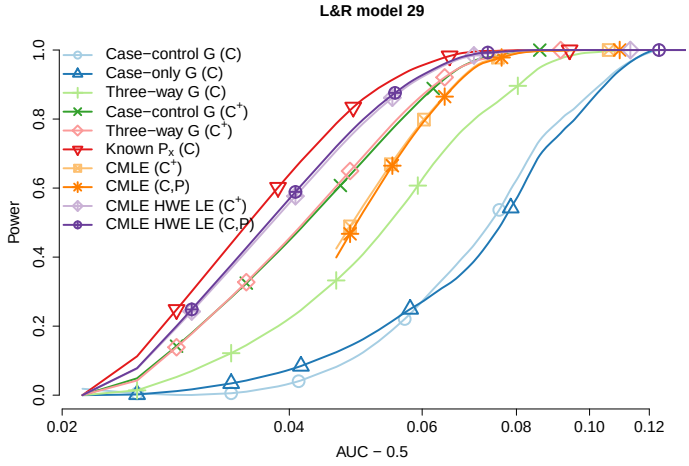


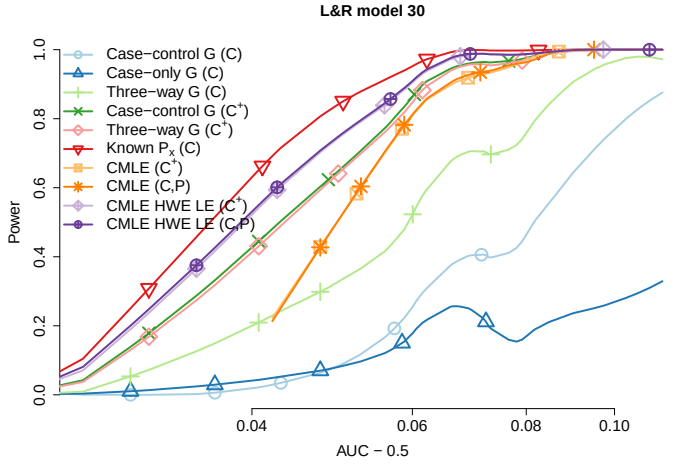


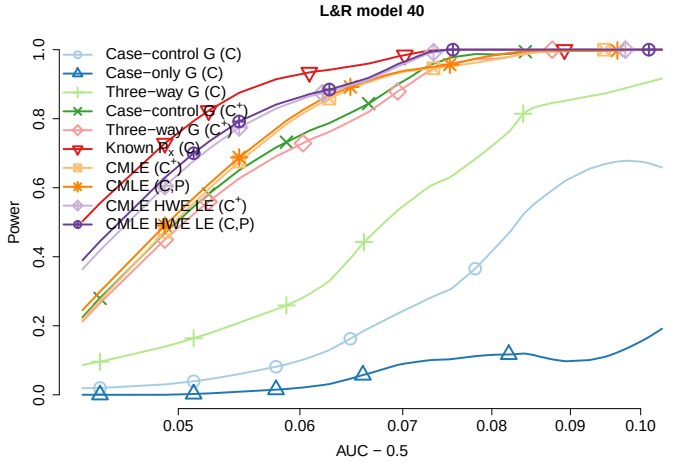


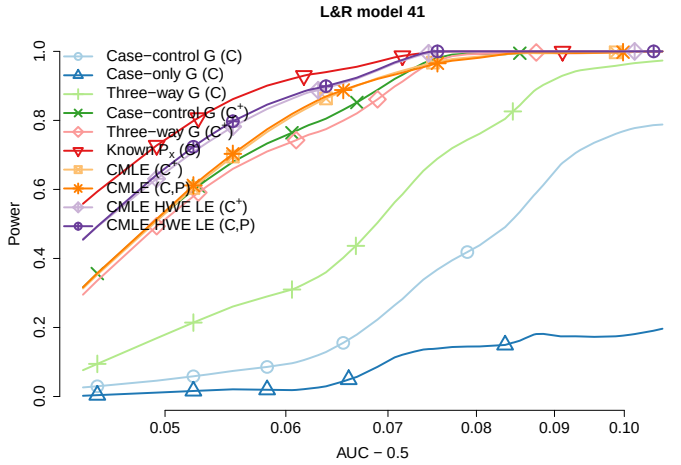


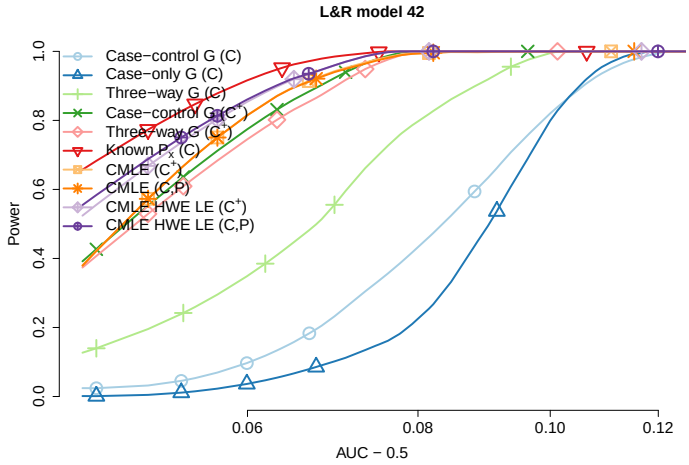


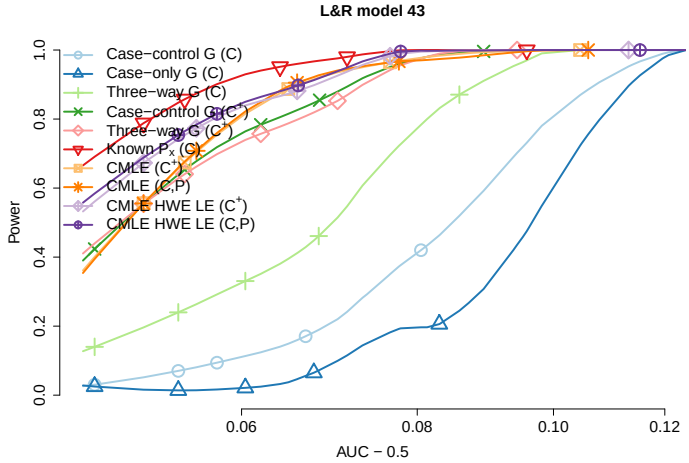


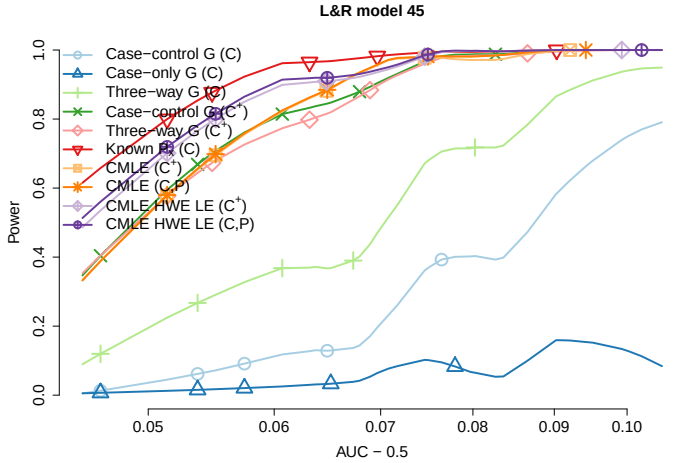


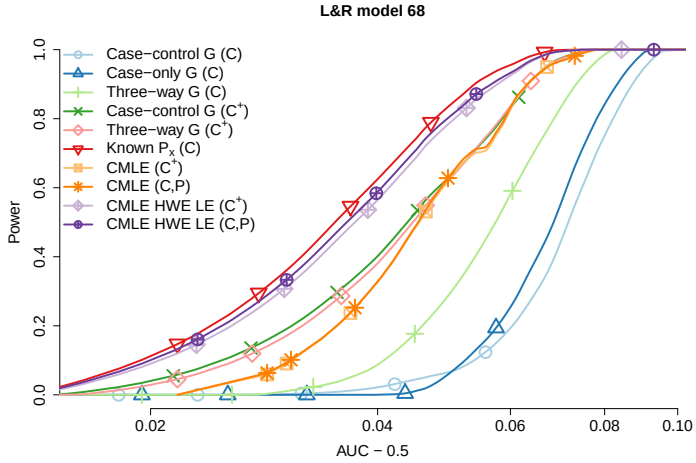


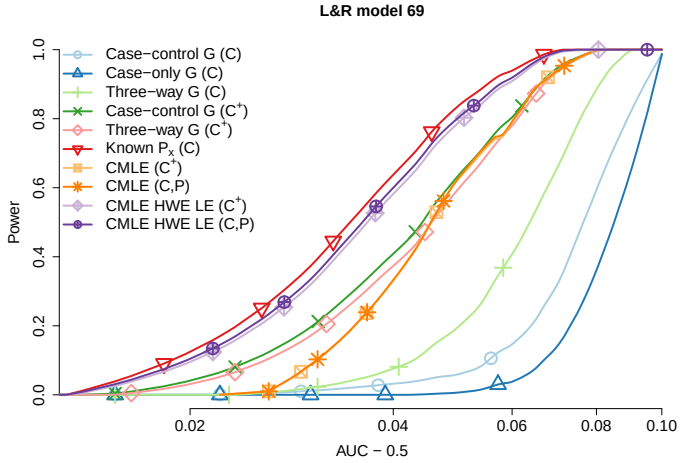


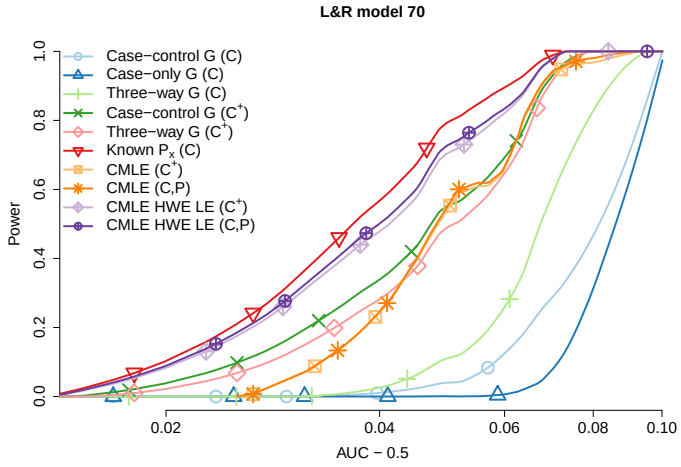


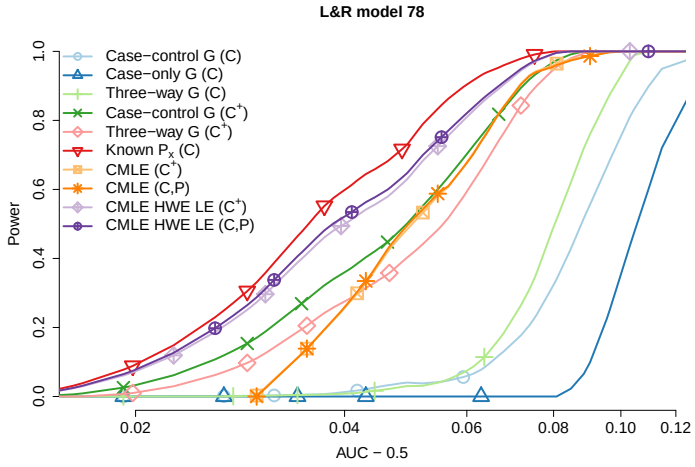


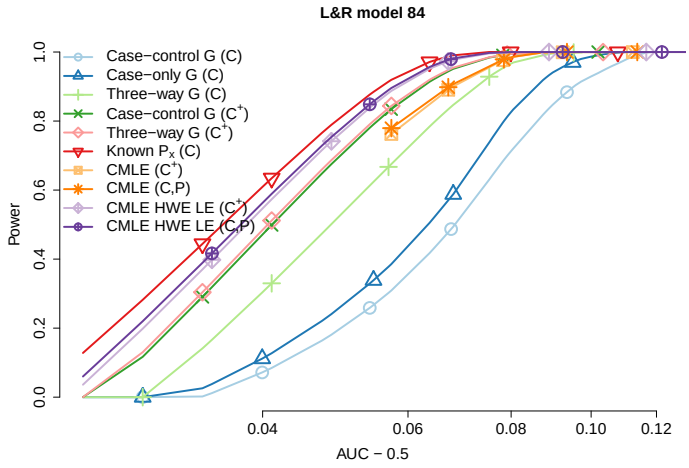


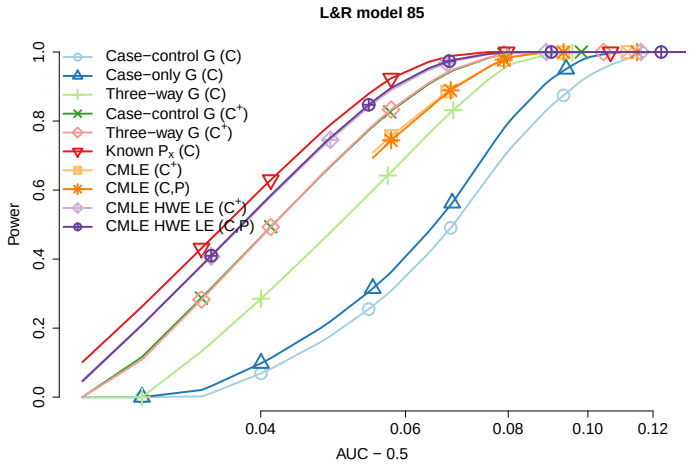


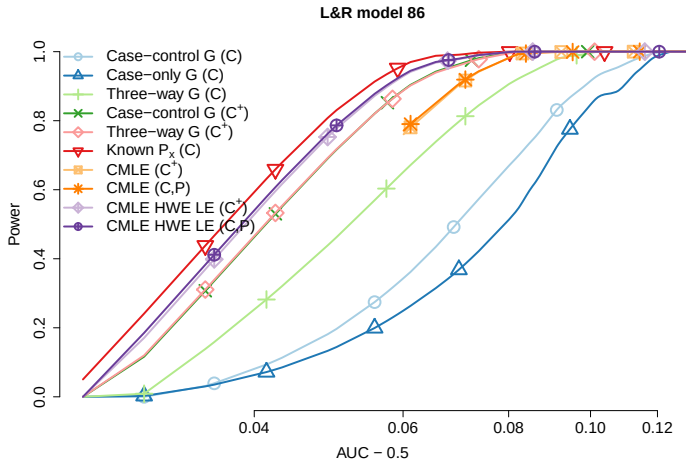


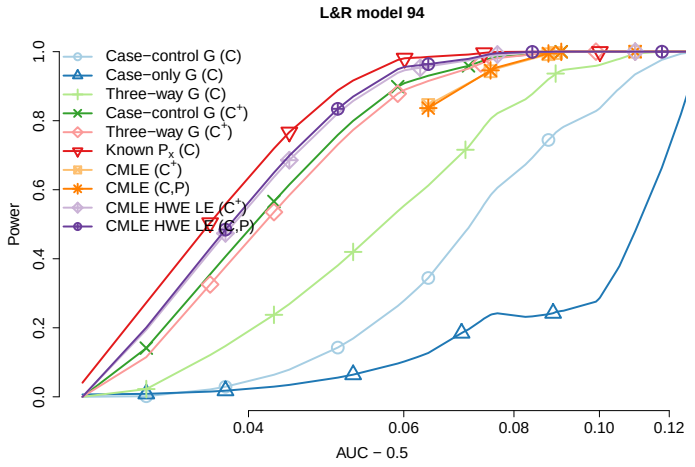


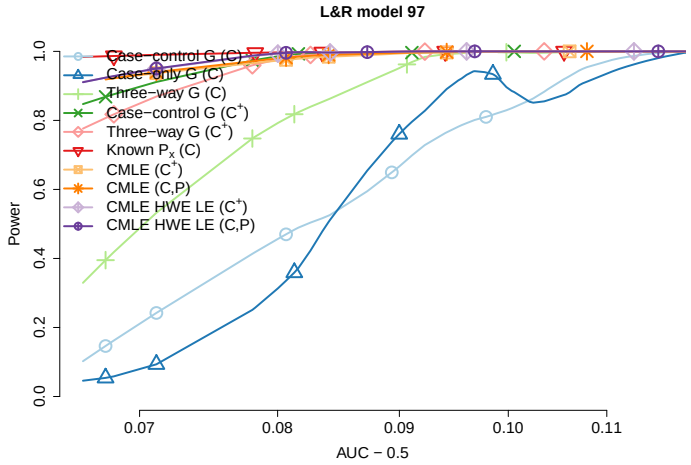


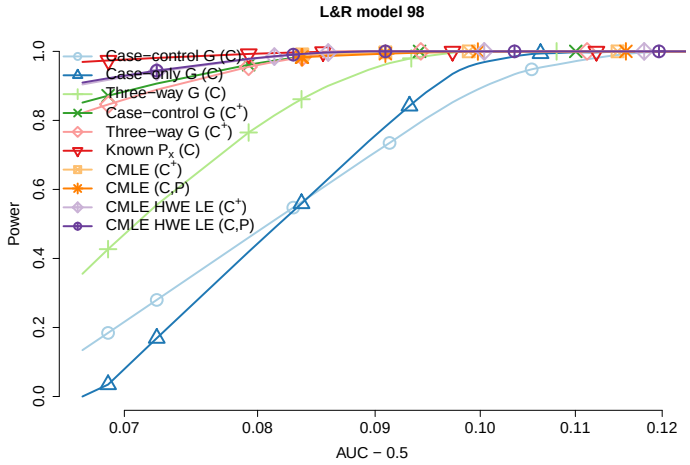


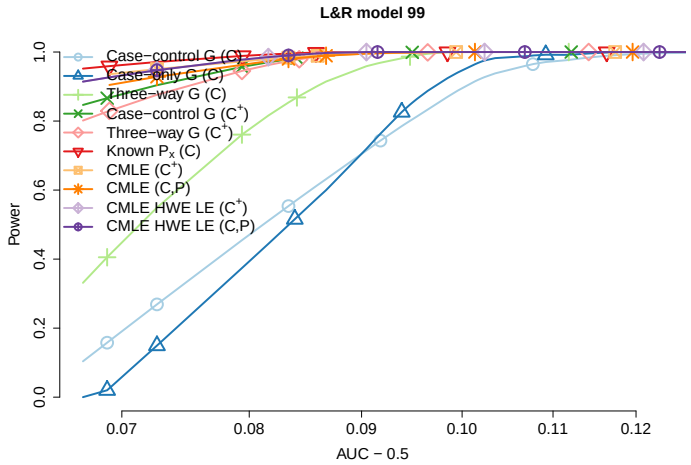


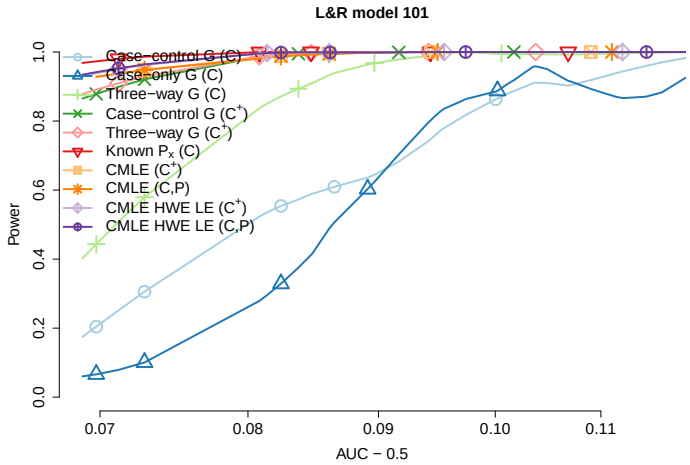


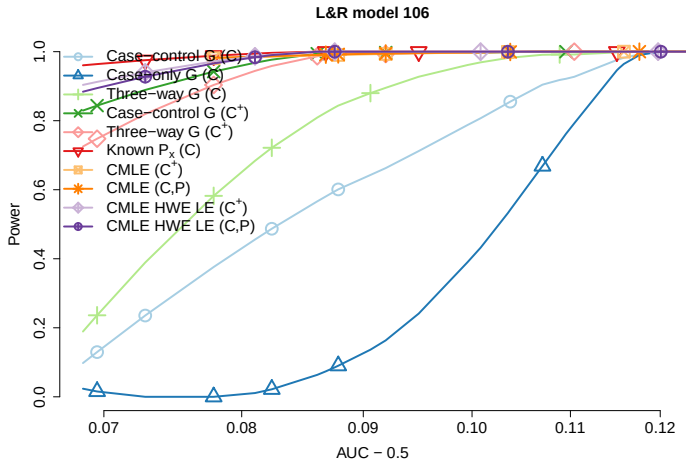












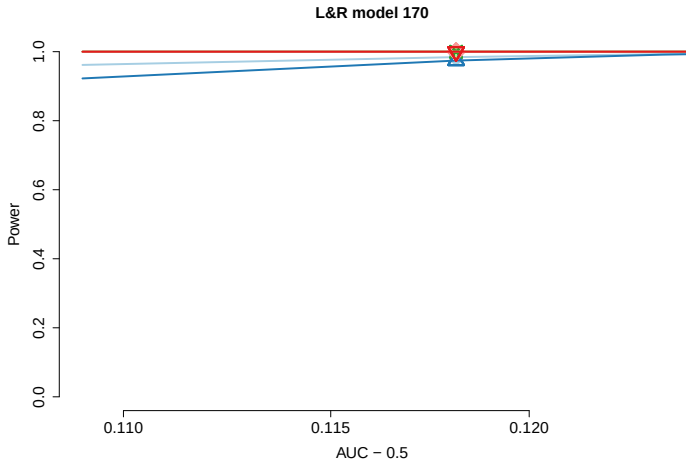
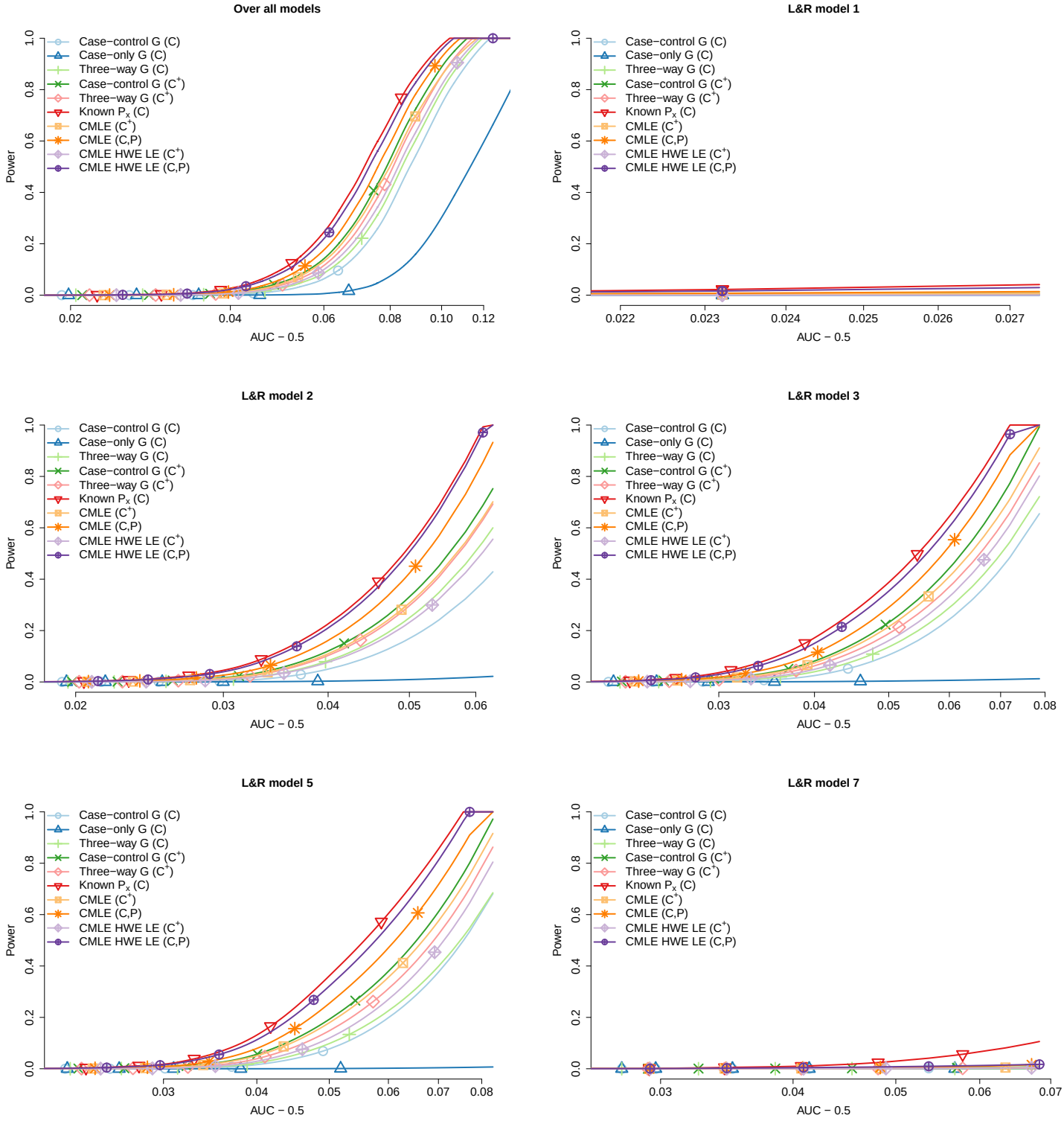
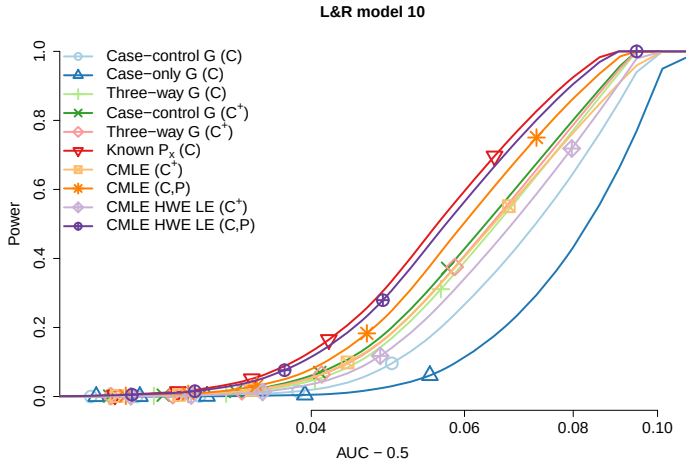
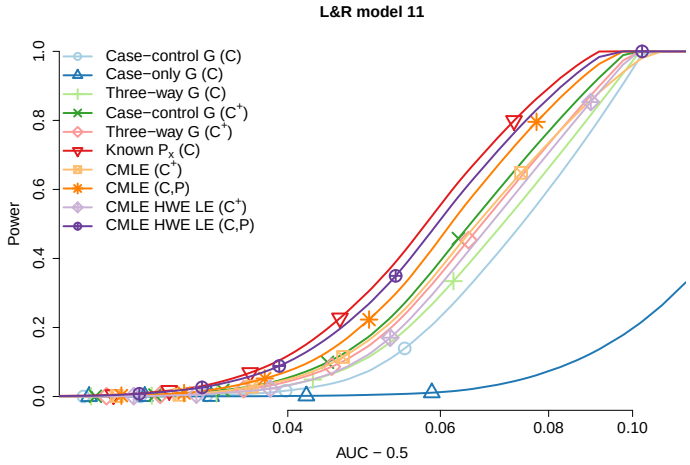
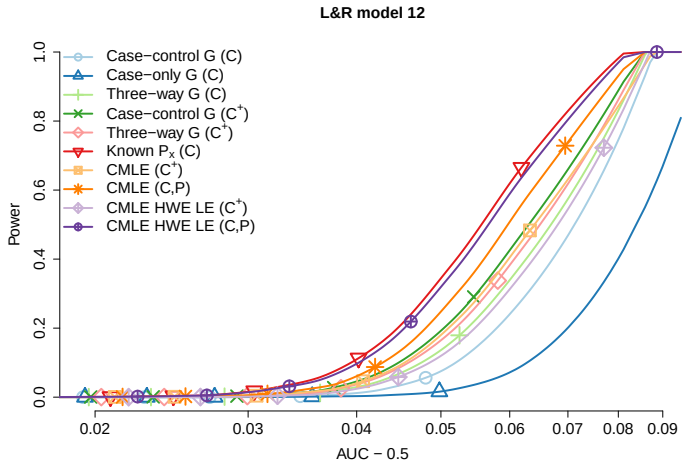


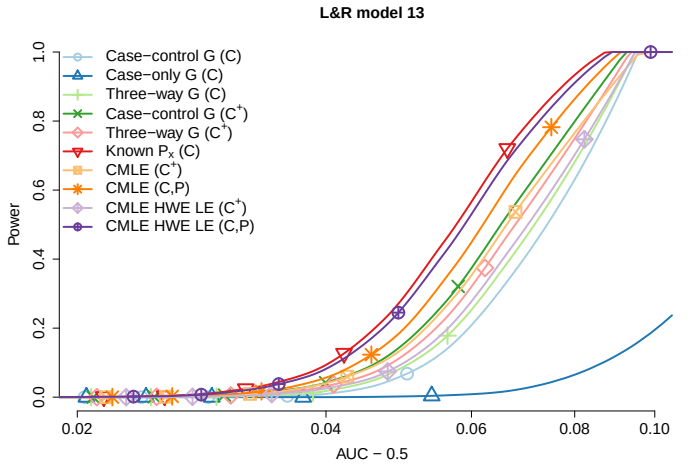
Figure S2: Power simulation results under LE for 0.20 prevalence.

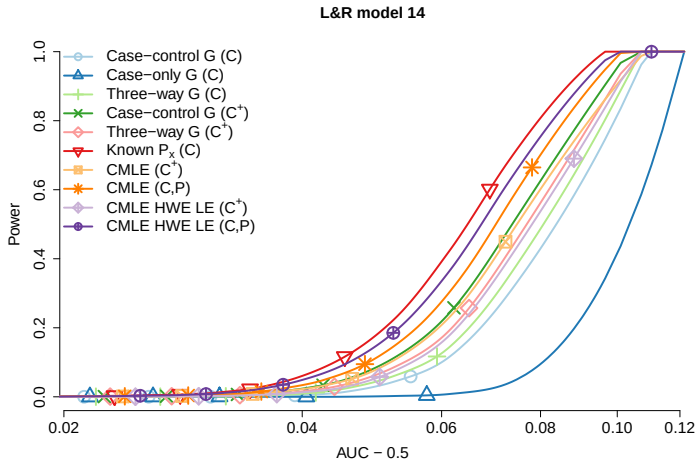


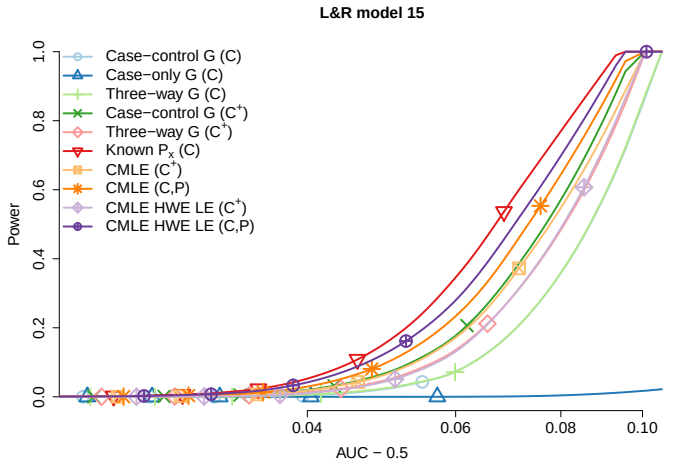


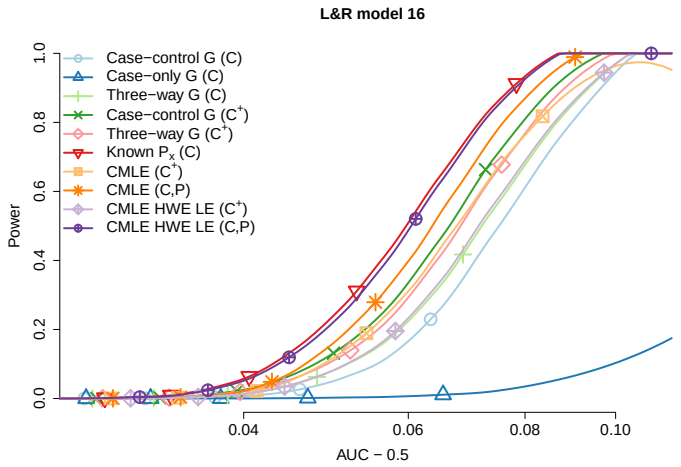


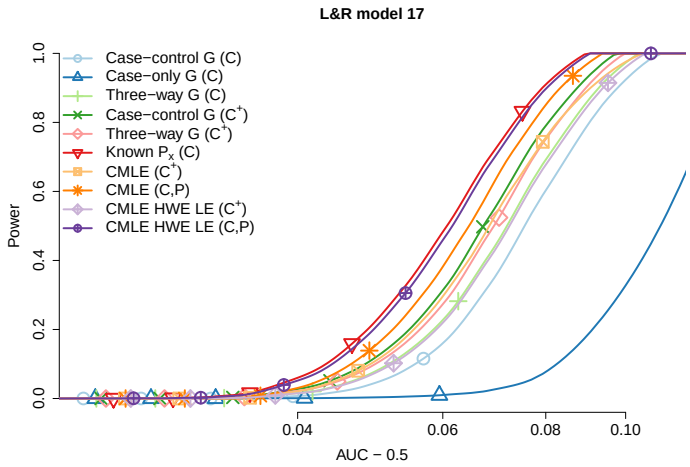


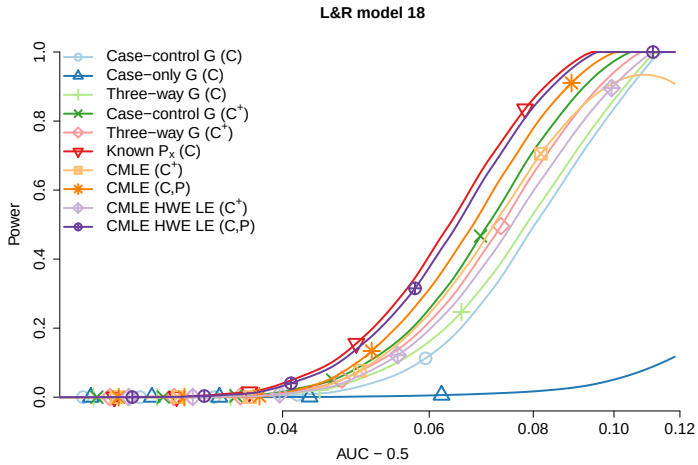


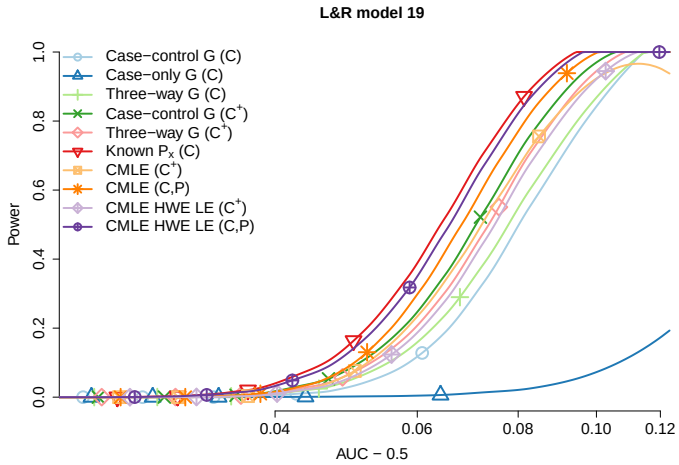


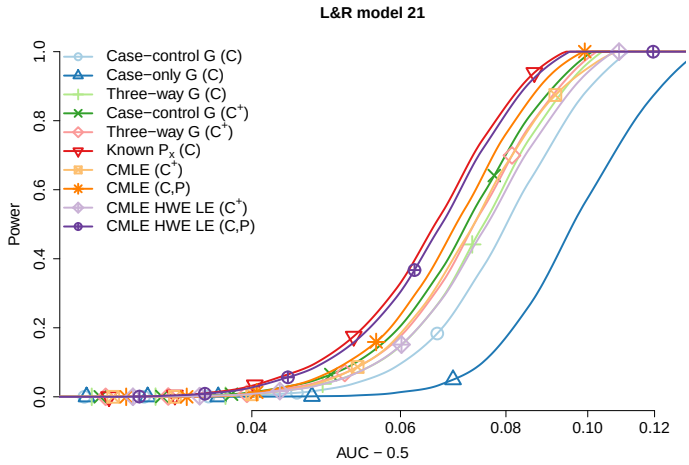


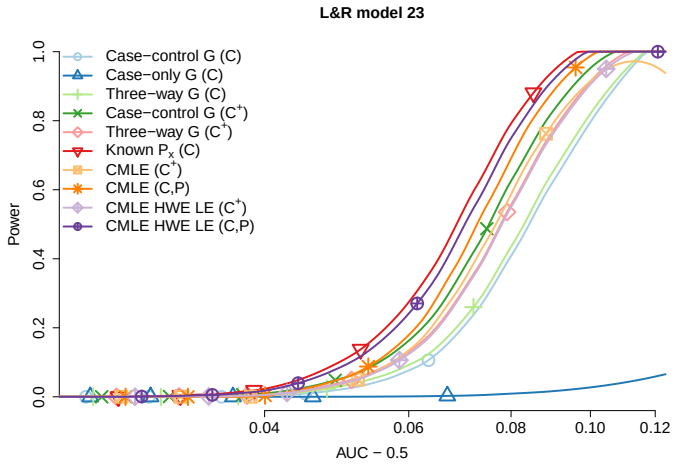


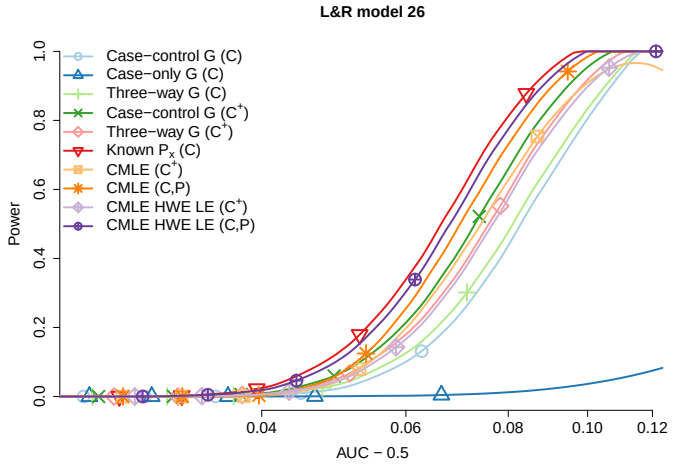


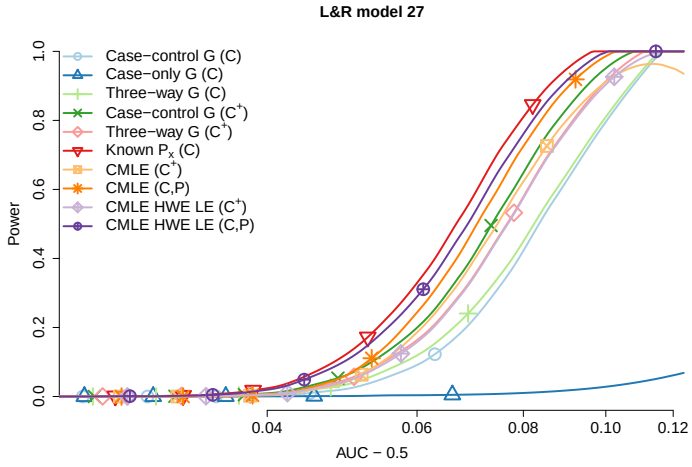


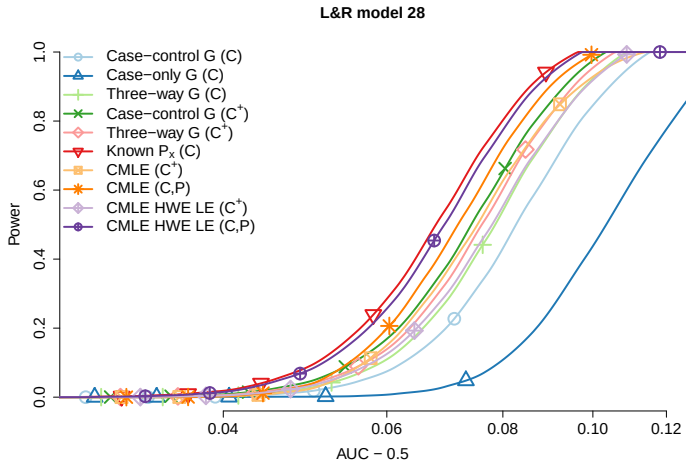


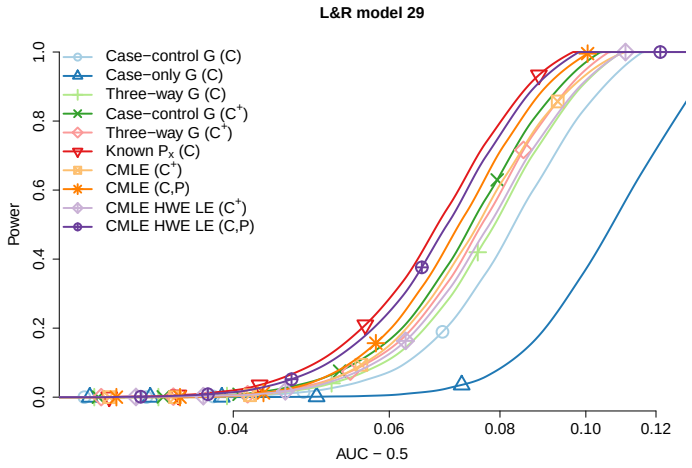


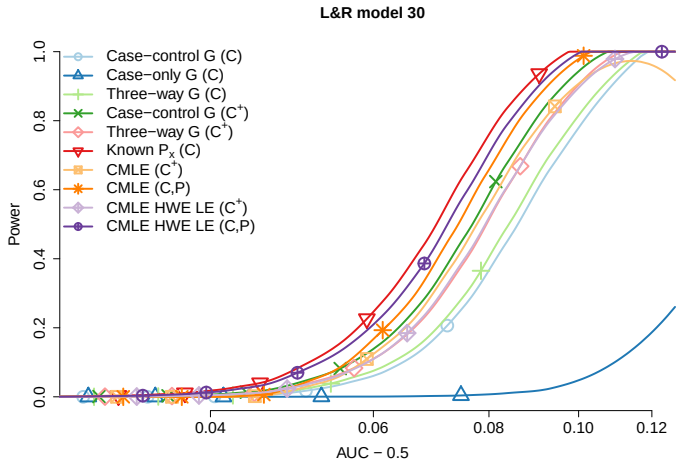


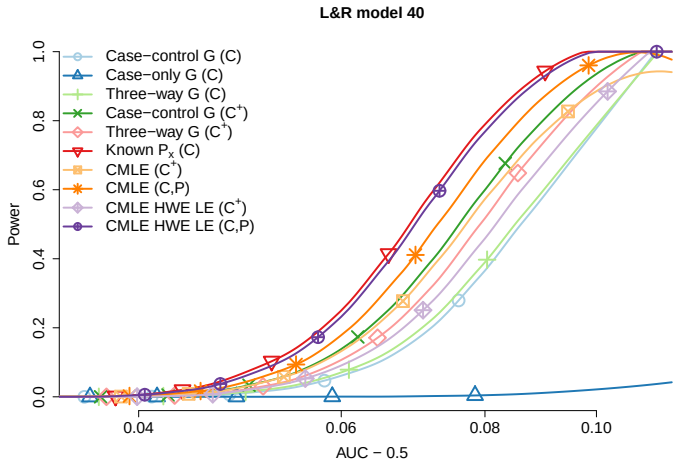


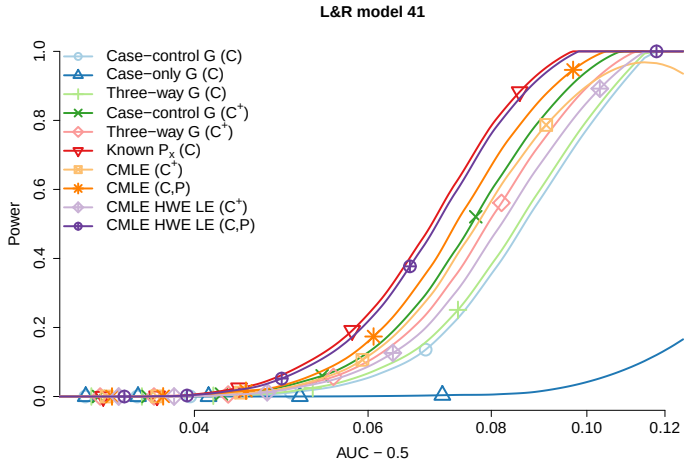


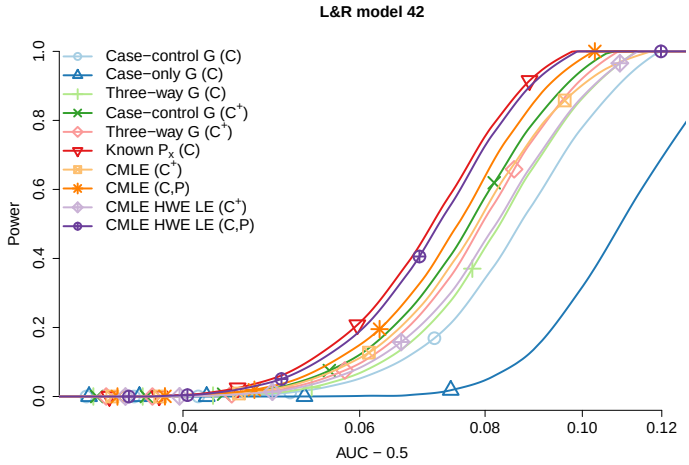


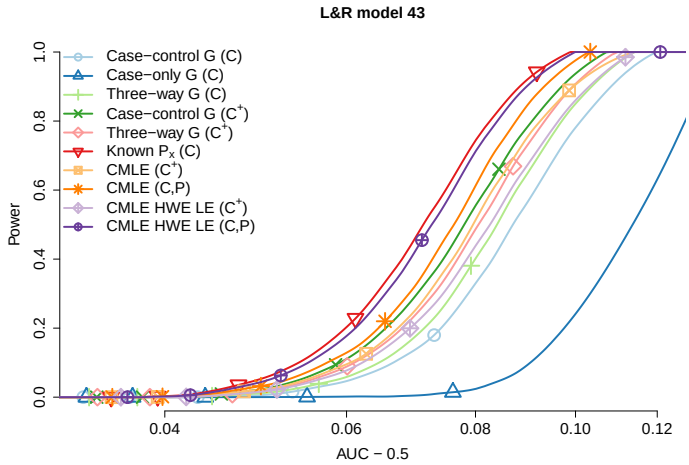


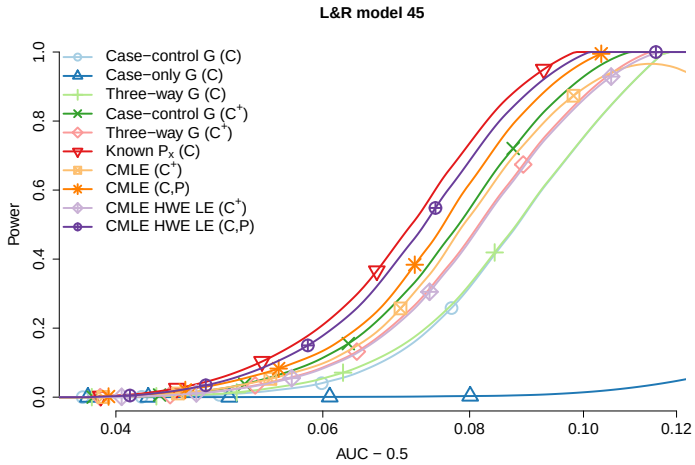


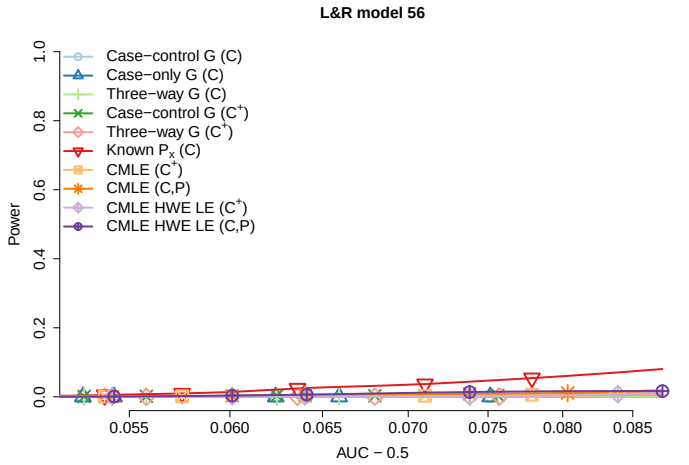


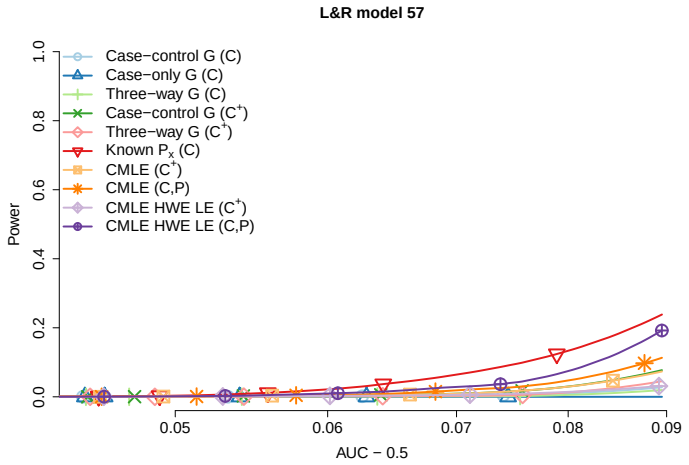


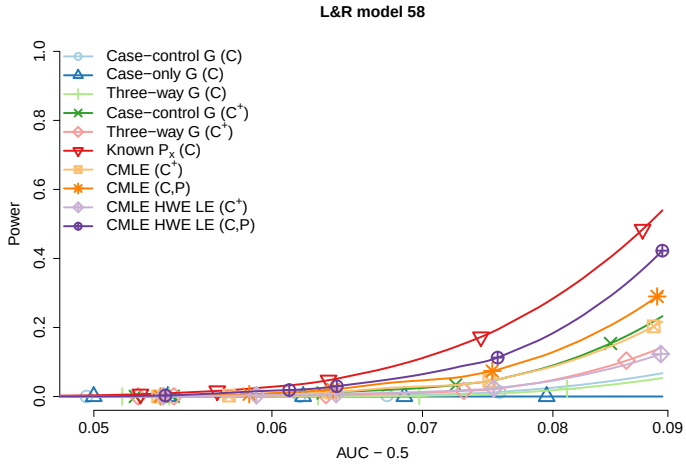


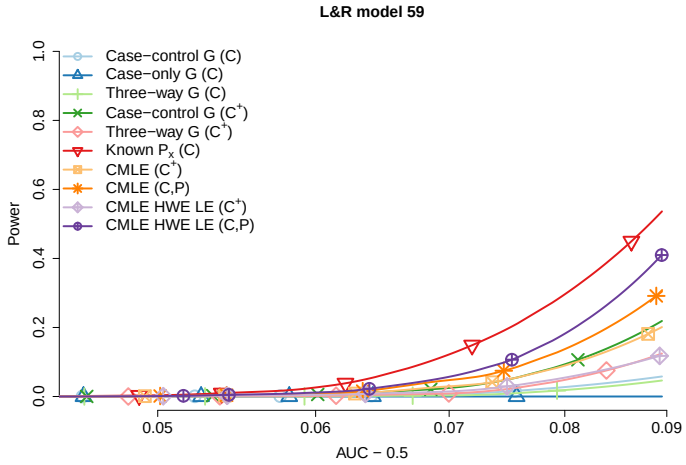


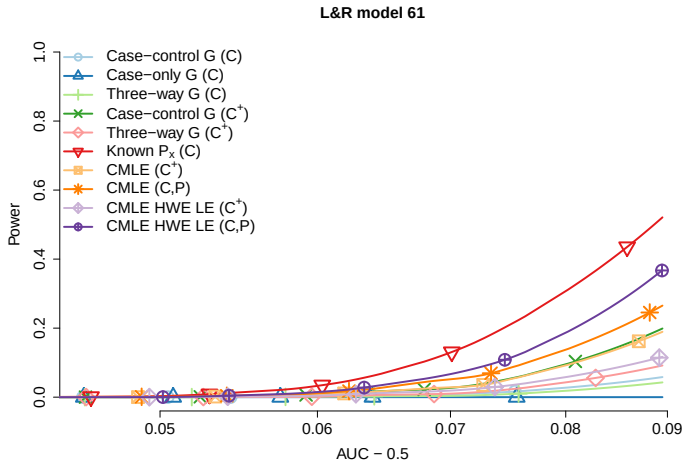


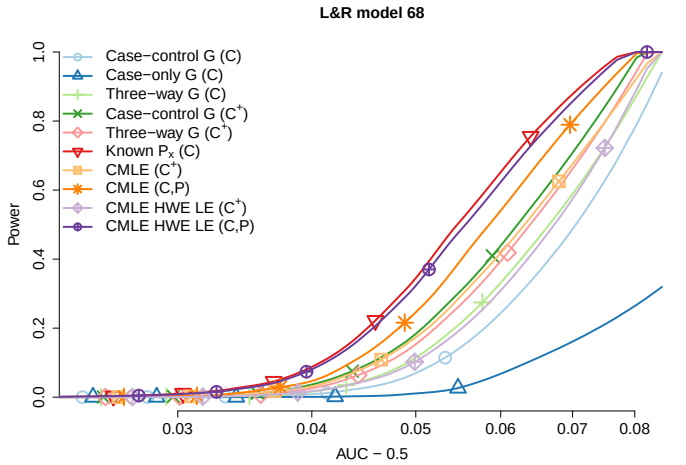


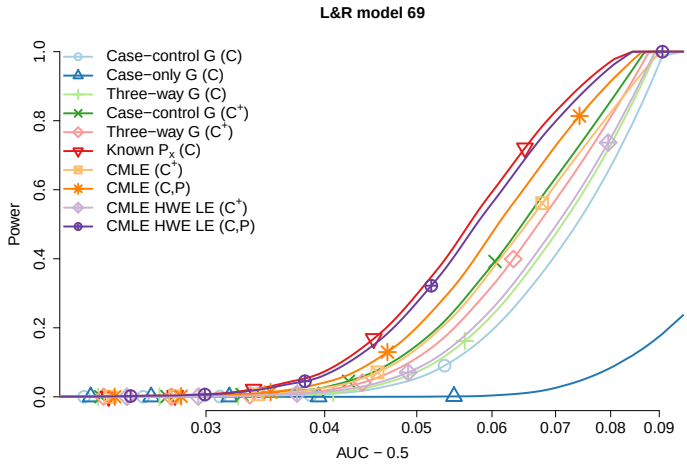


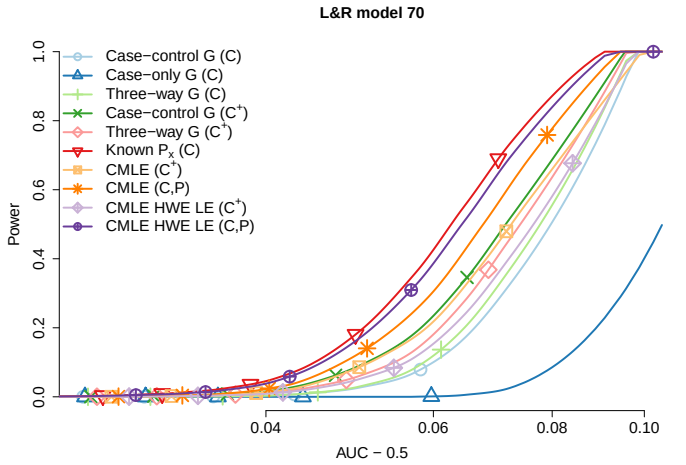


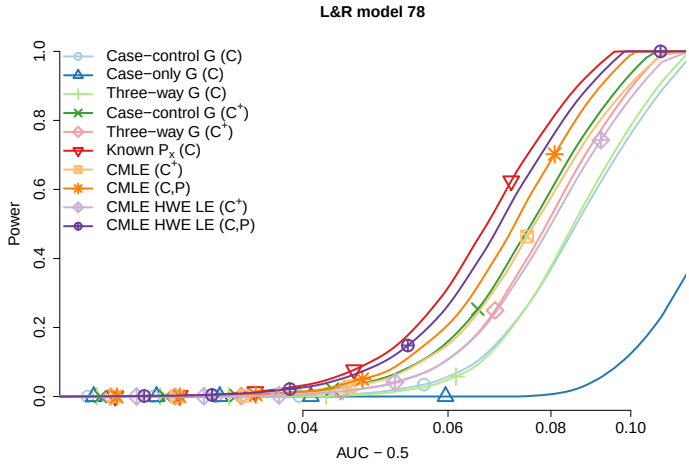


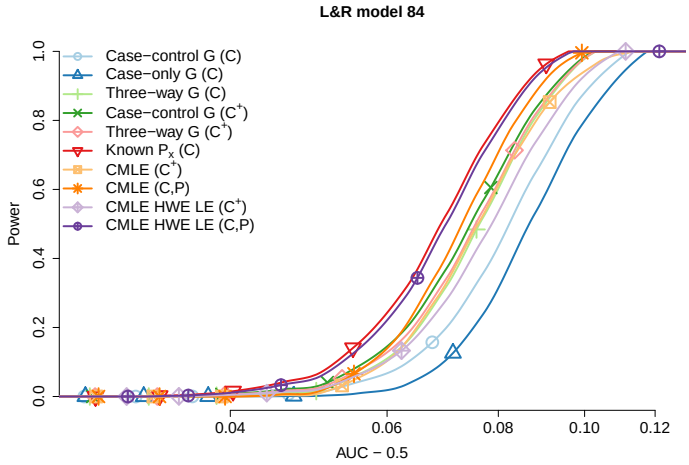


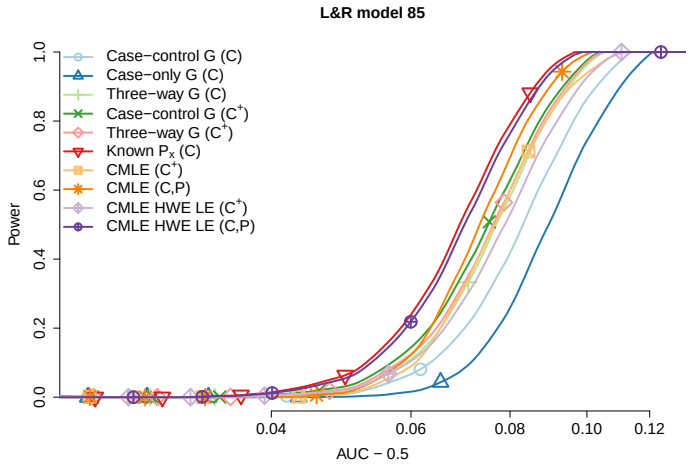


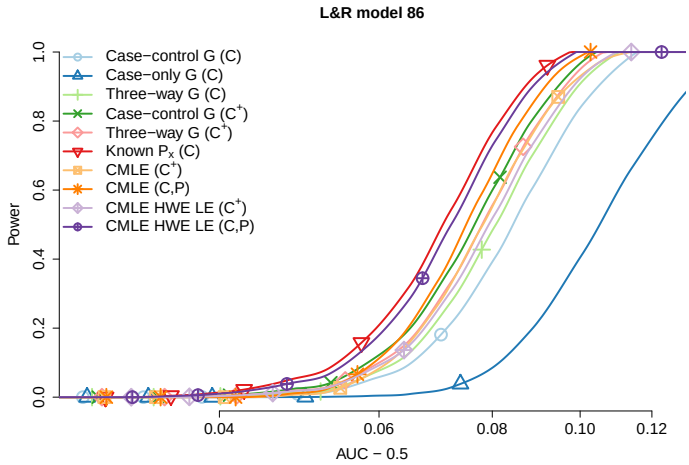


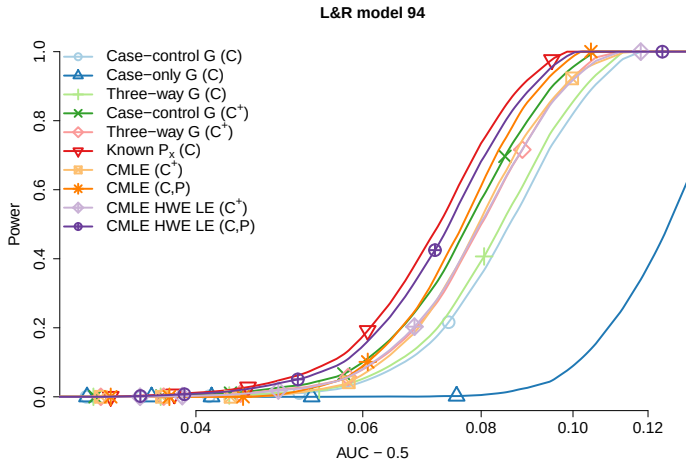


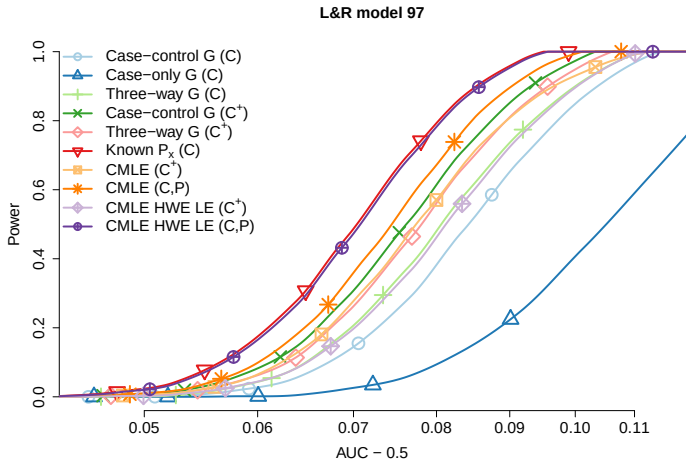


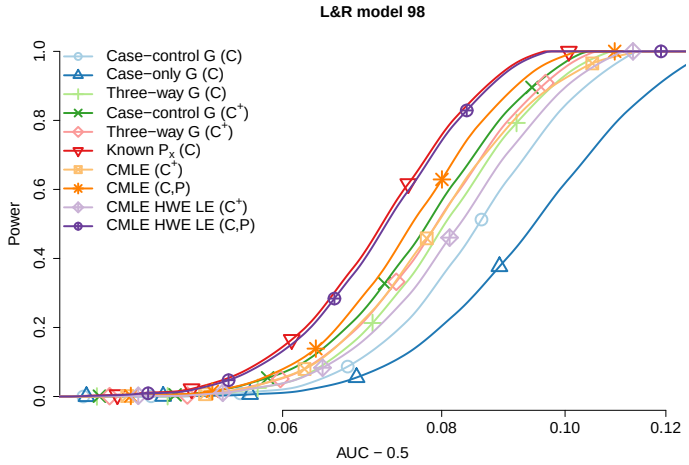


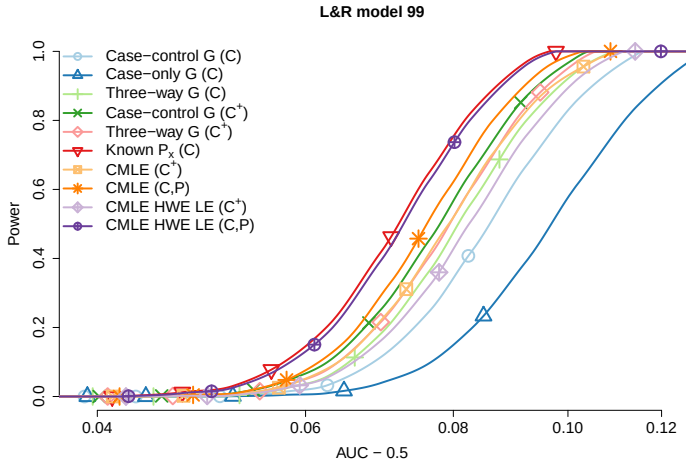


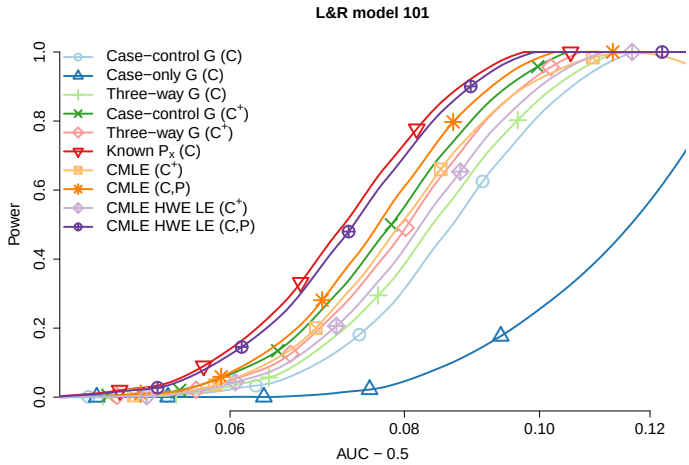


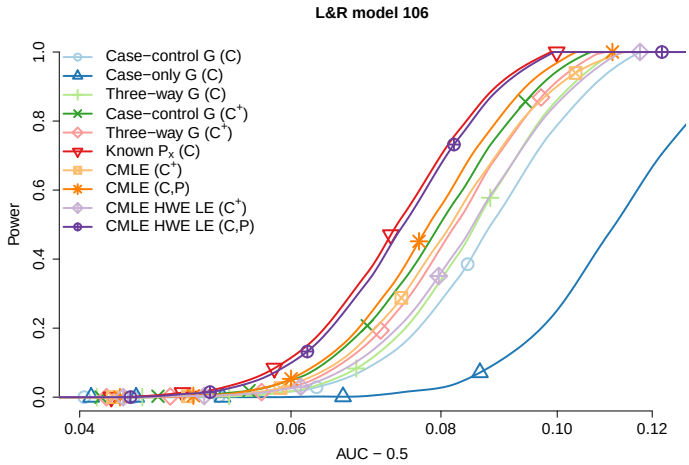


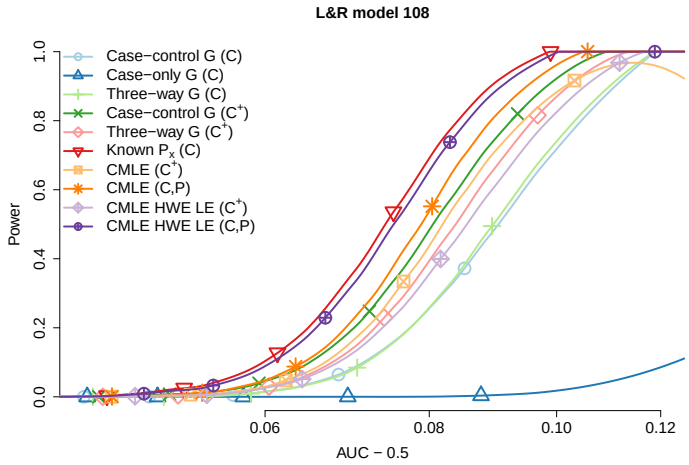


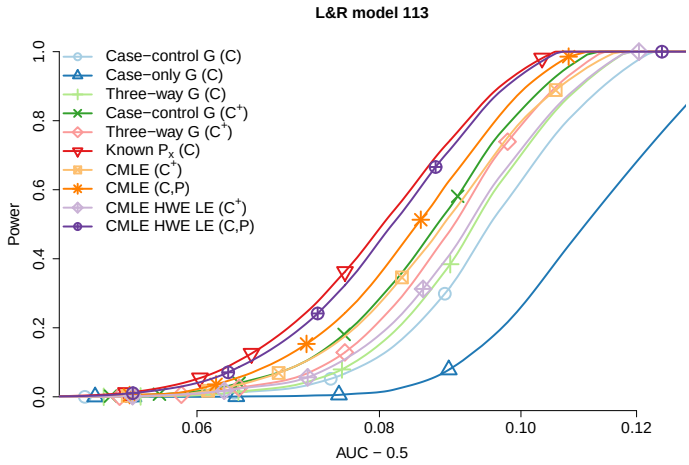


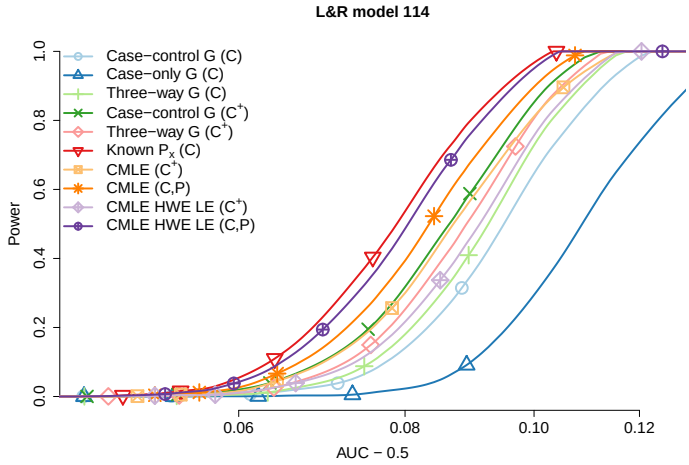


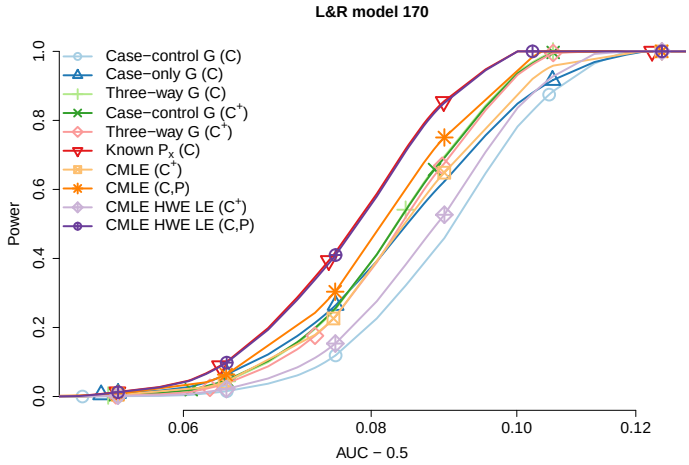












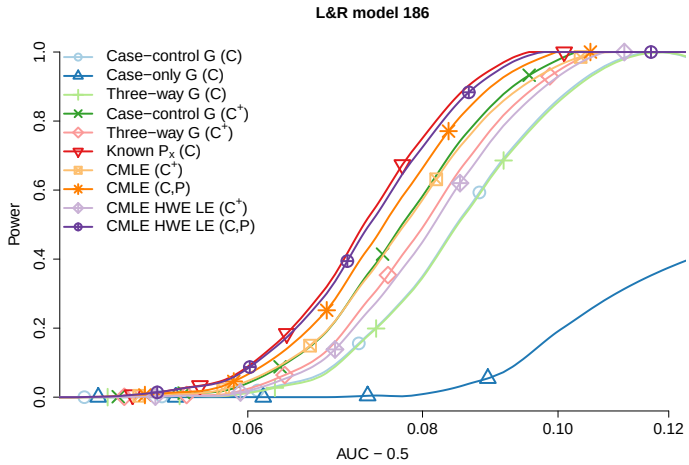
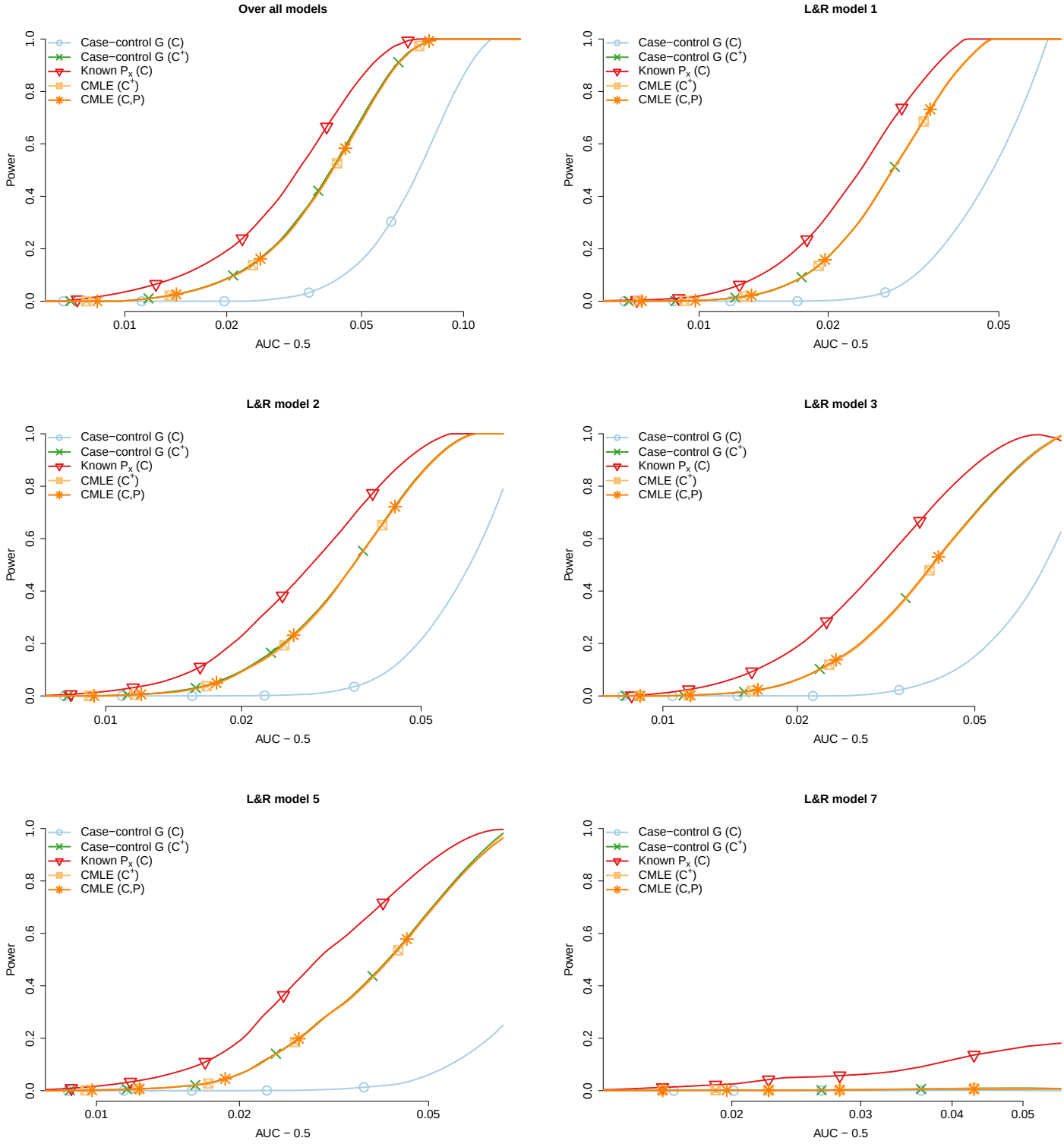
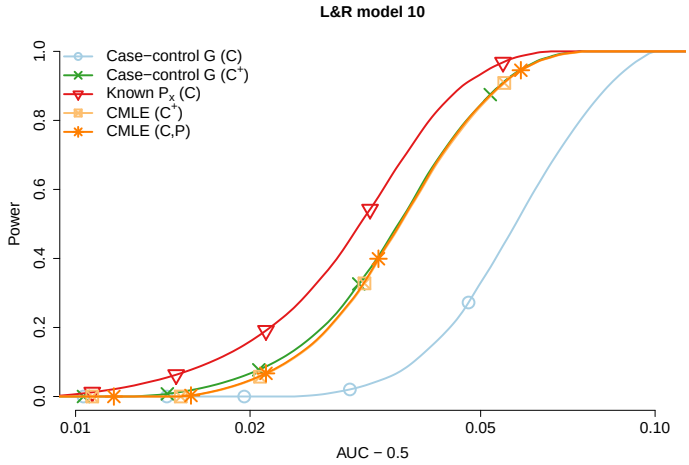
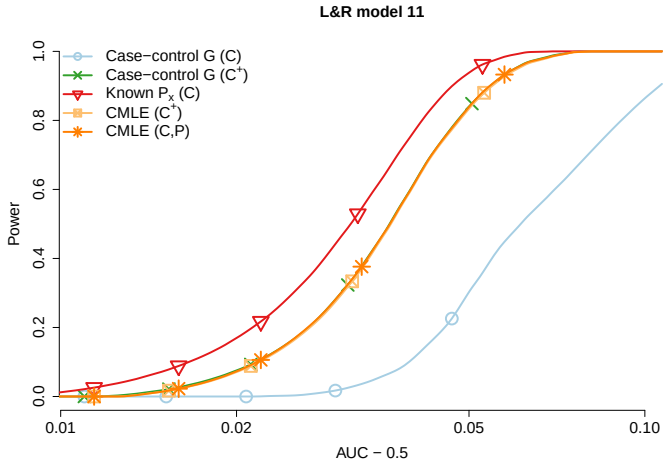
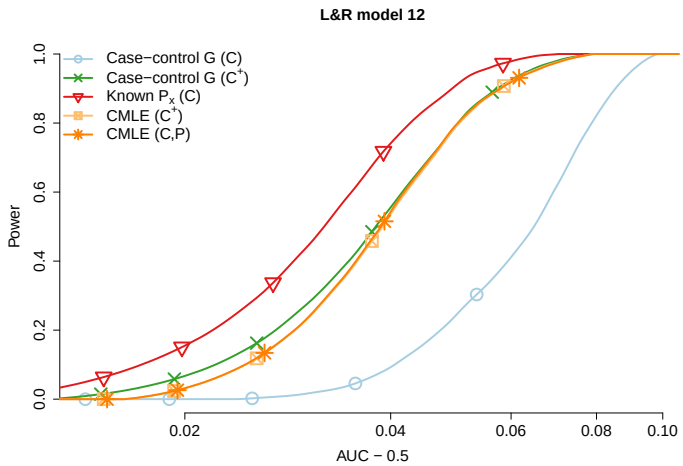


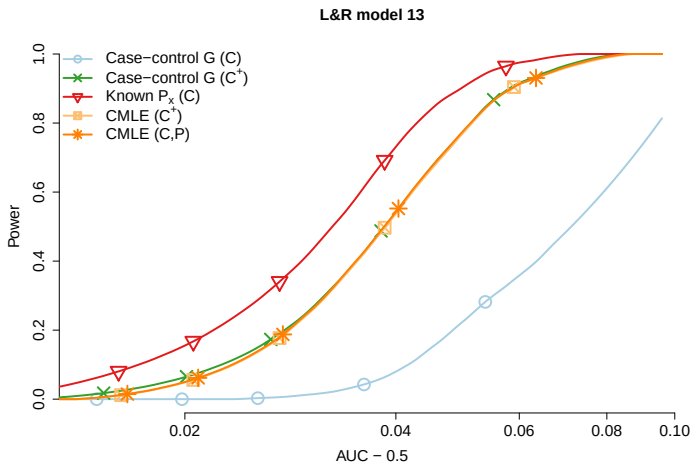
Figure S3: Power simulation results under LD for 0.20 prevalence.

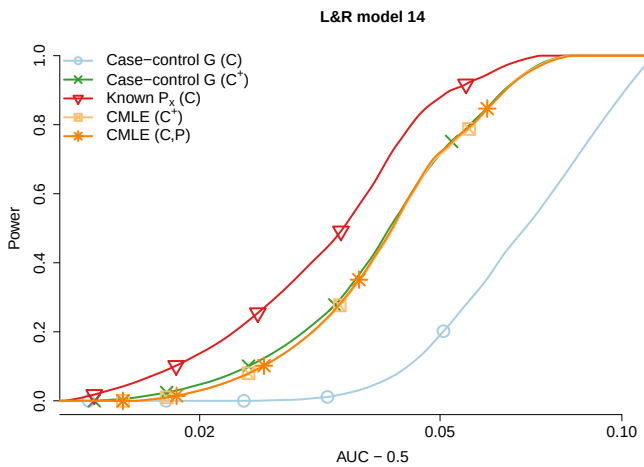


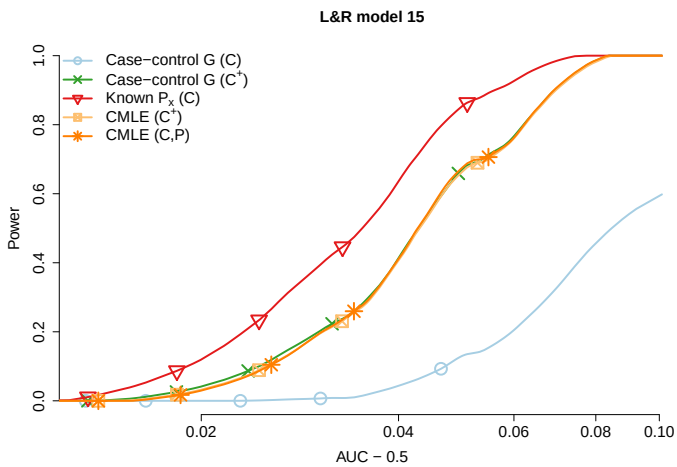


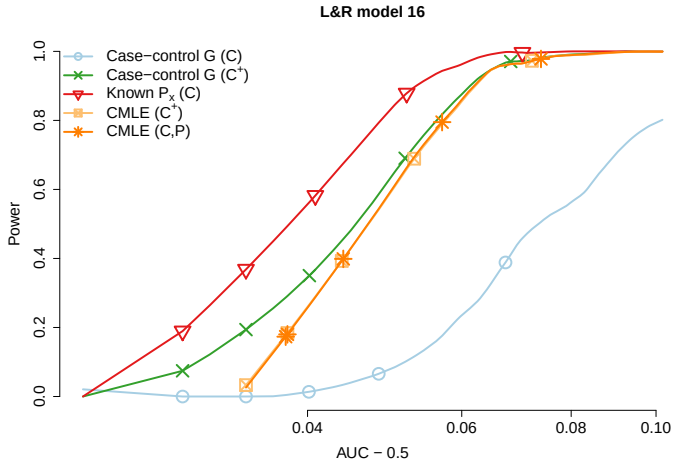


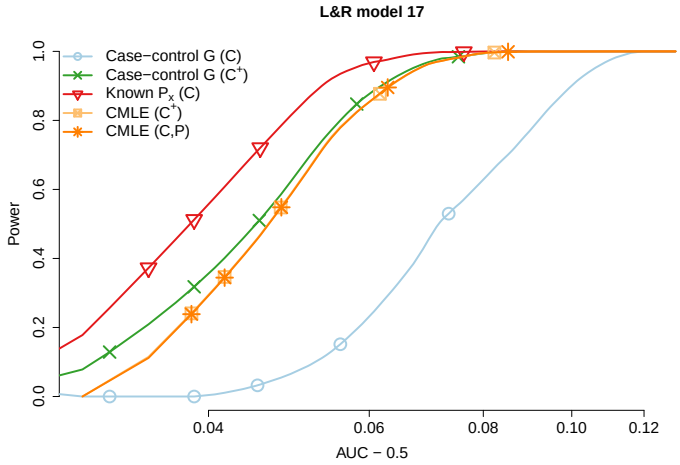


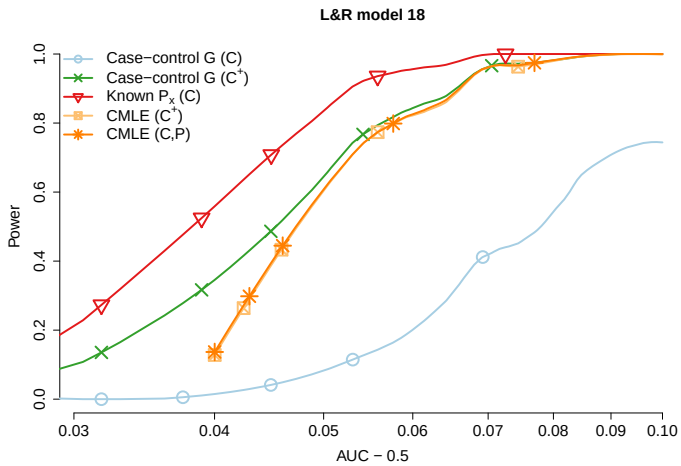


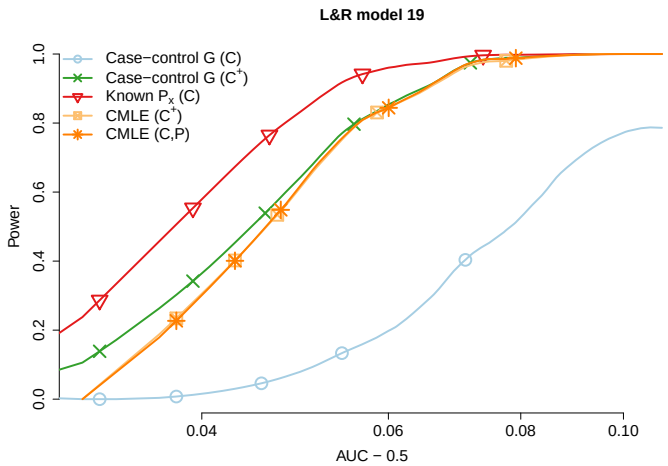


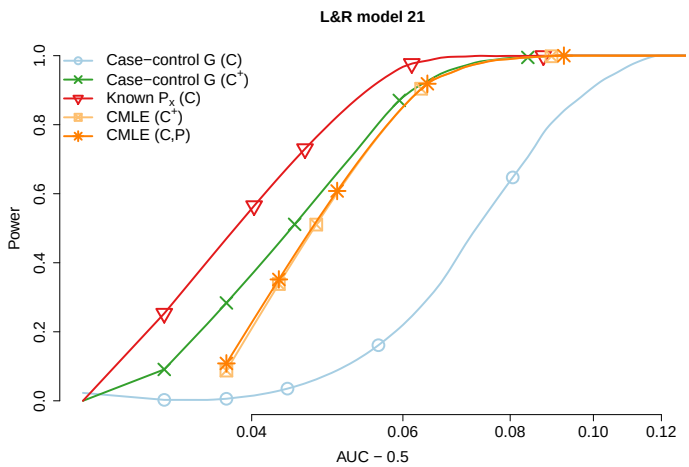


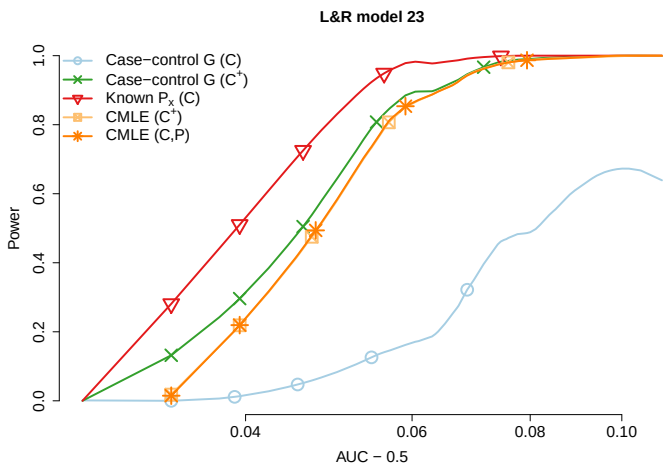


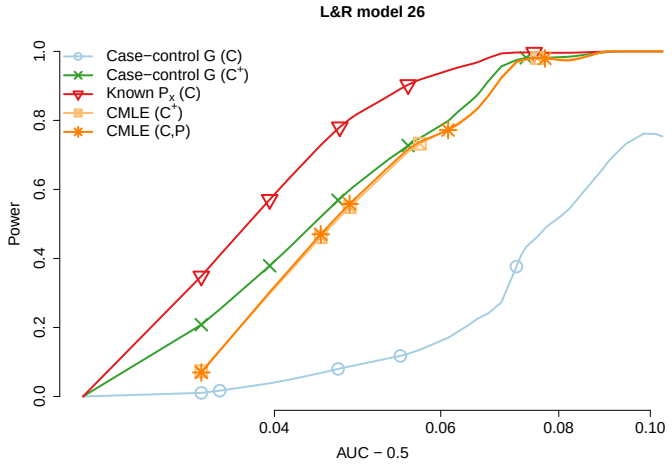


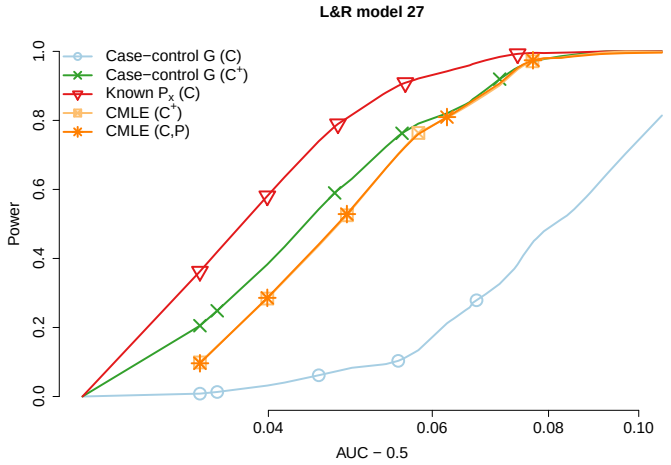


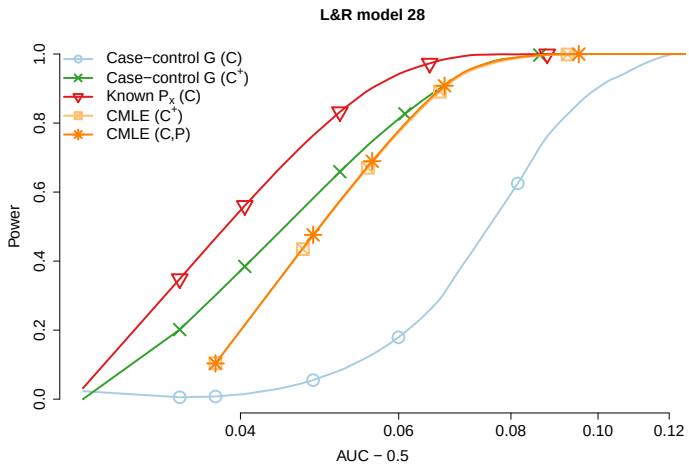


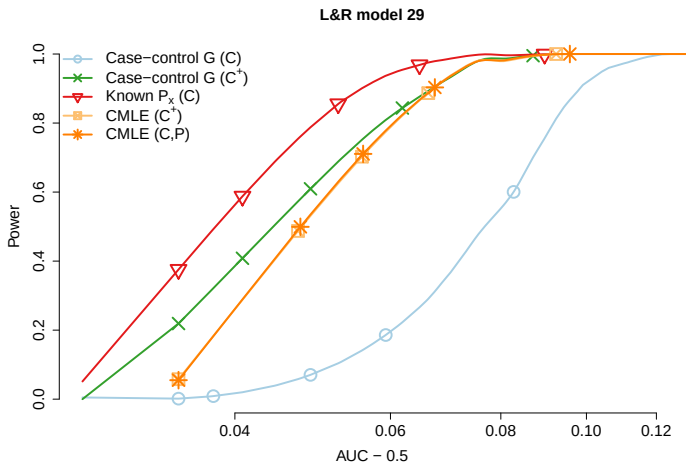


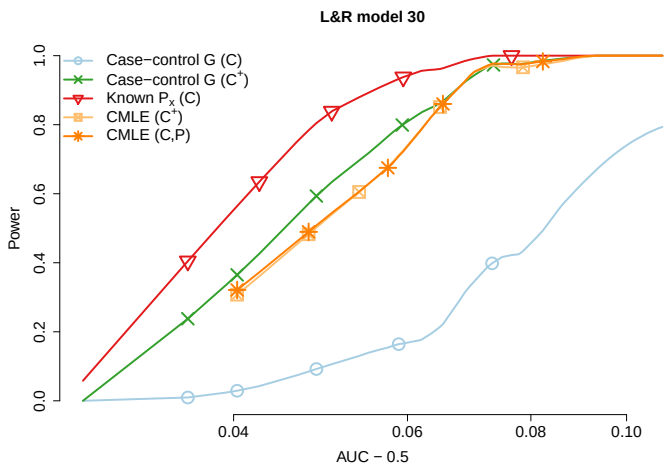


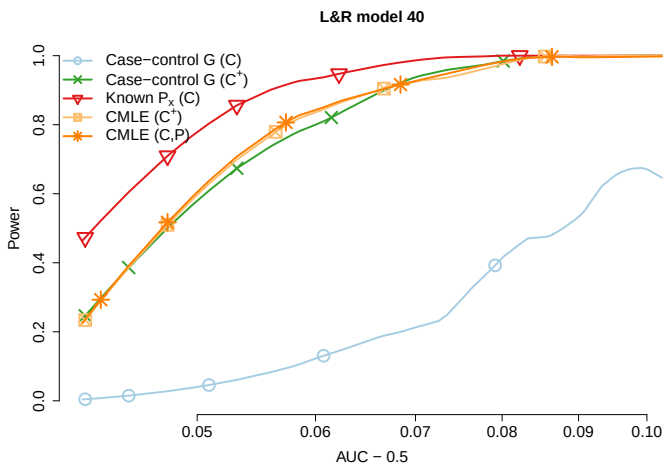


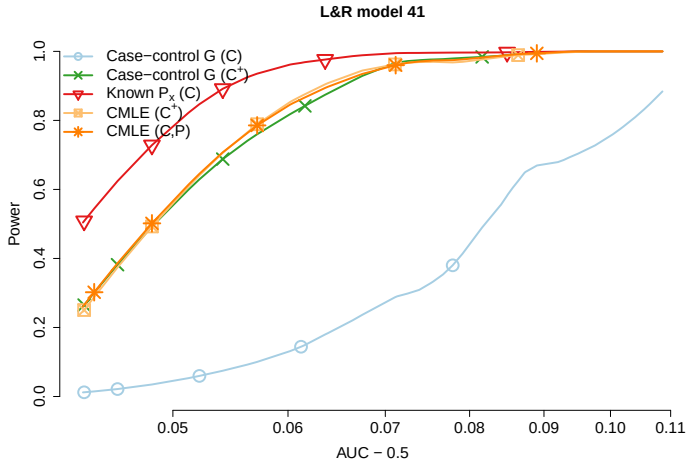


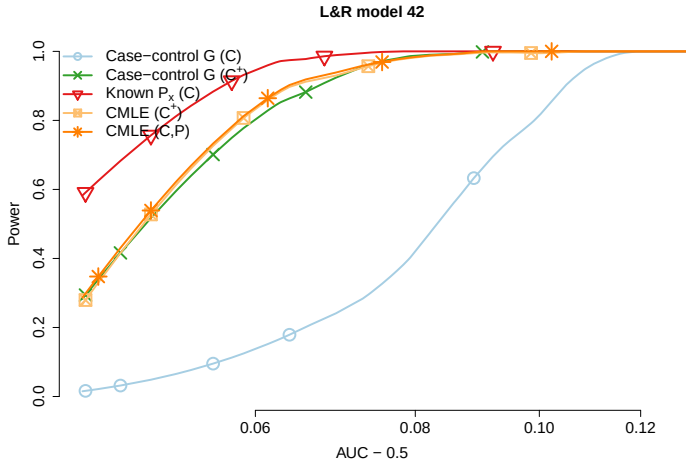


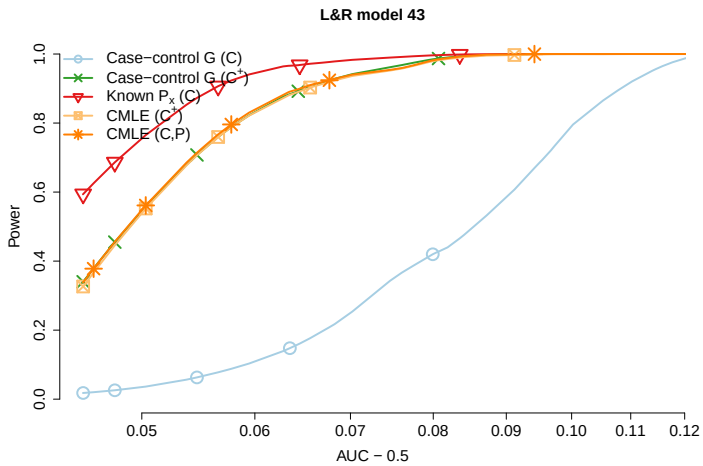


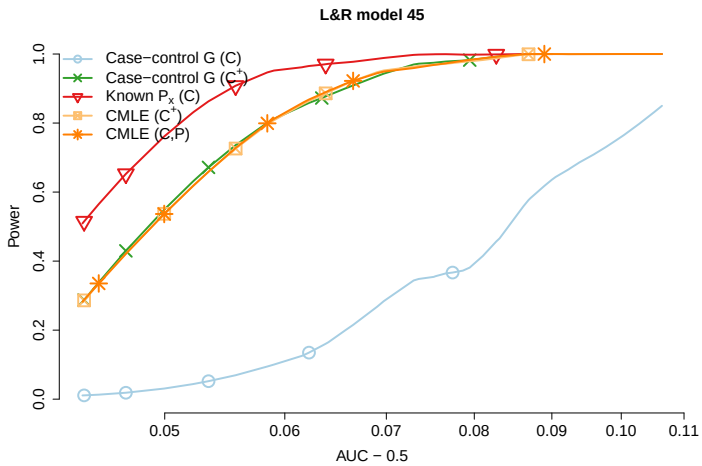


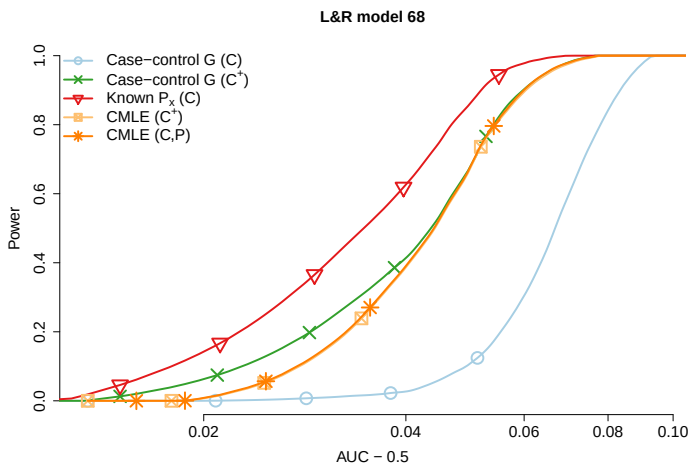


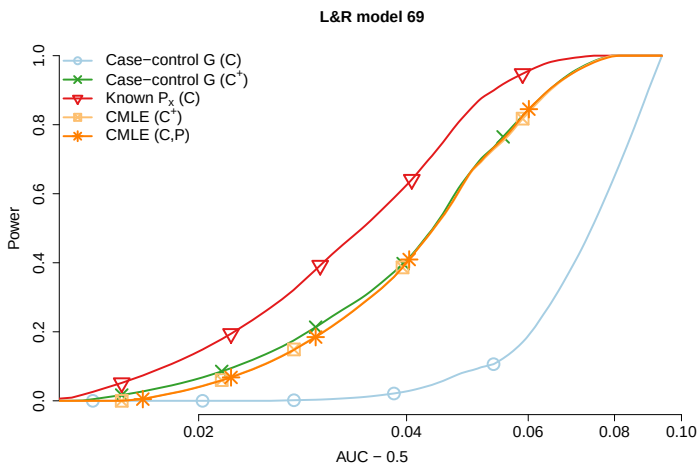


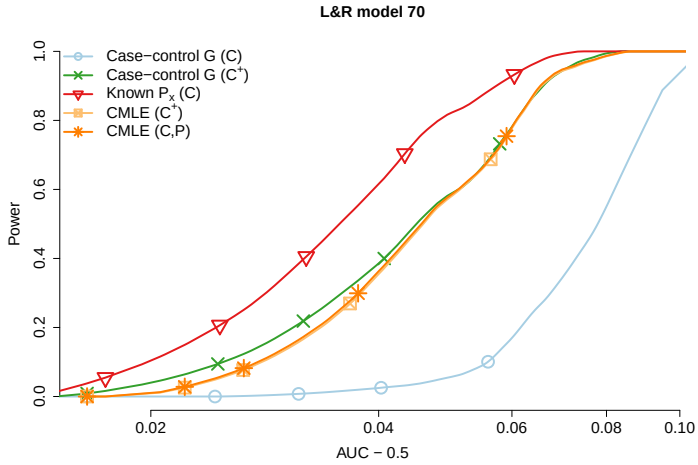


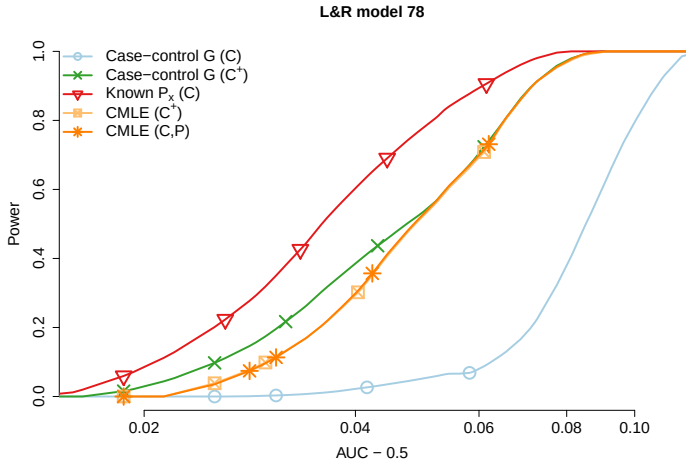


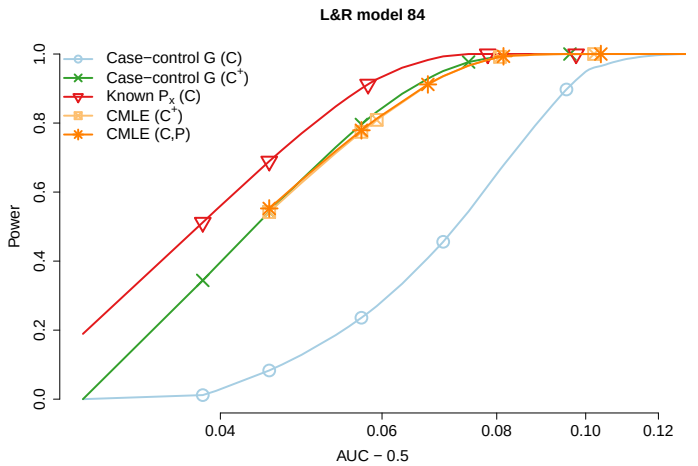


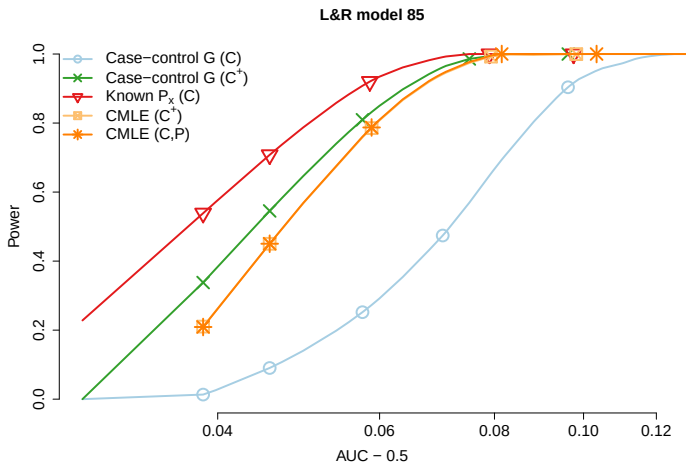


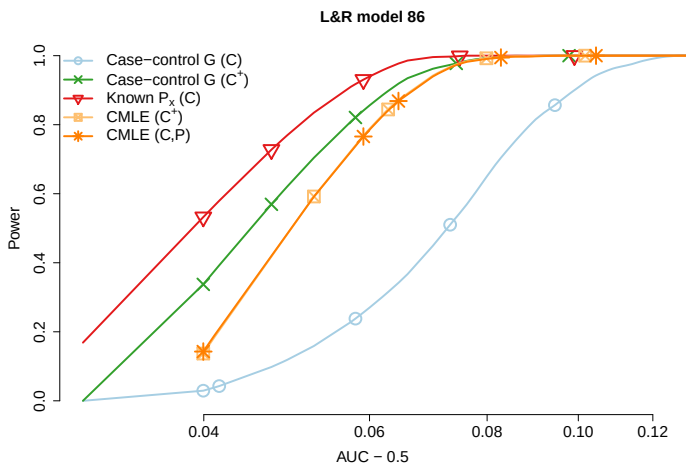


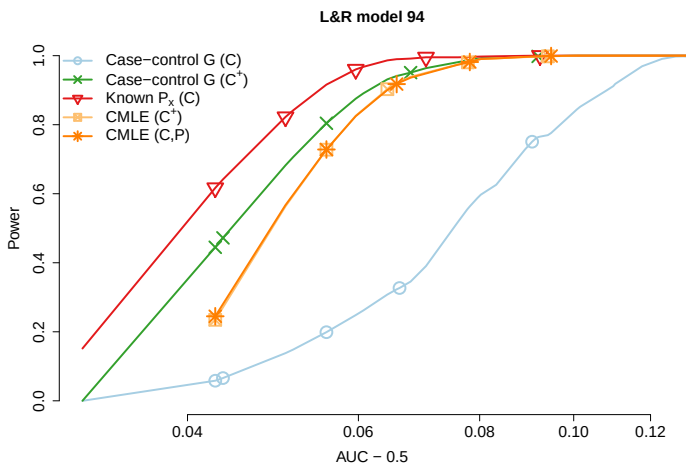


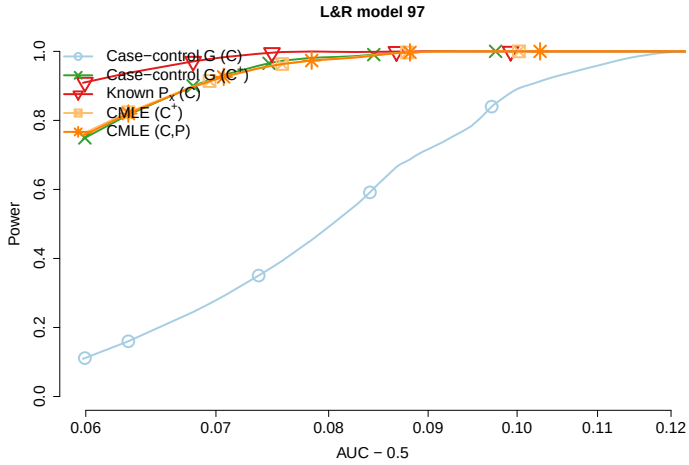


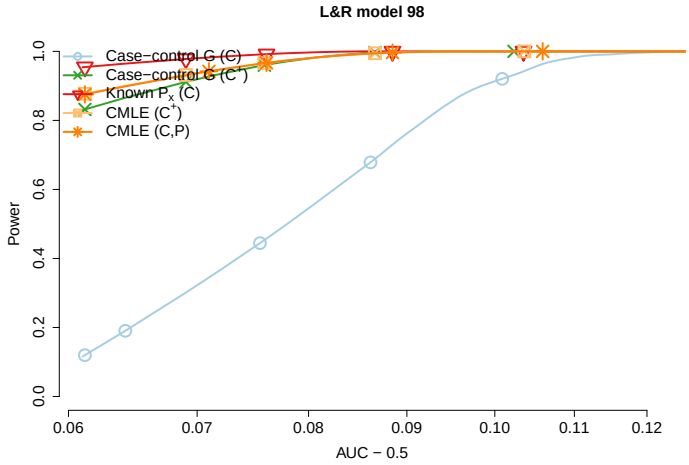


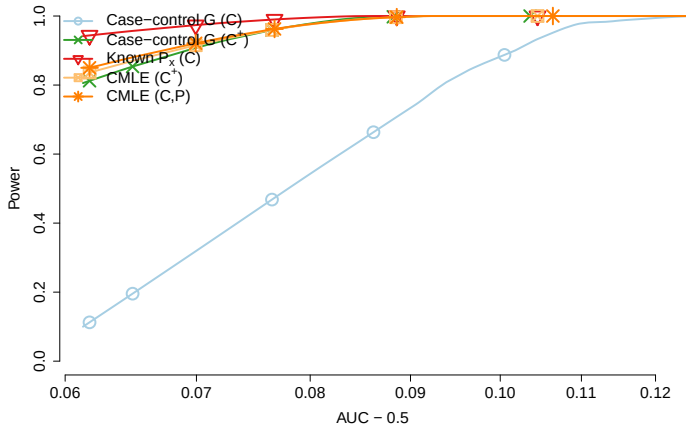


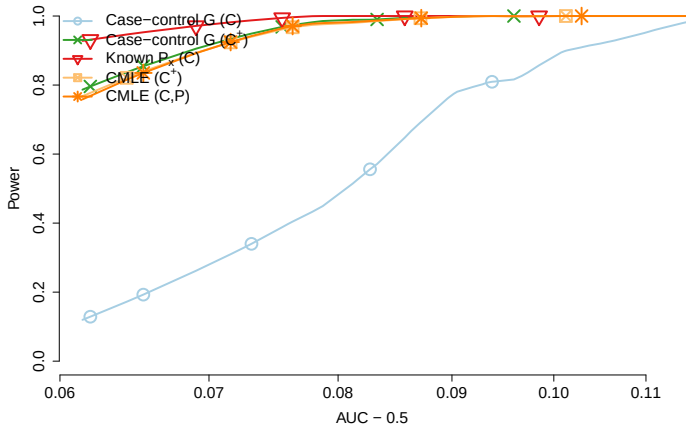


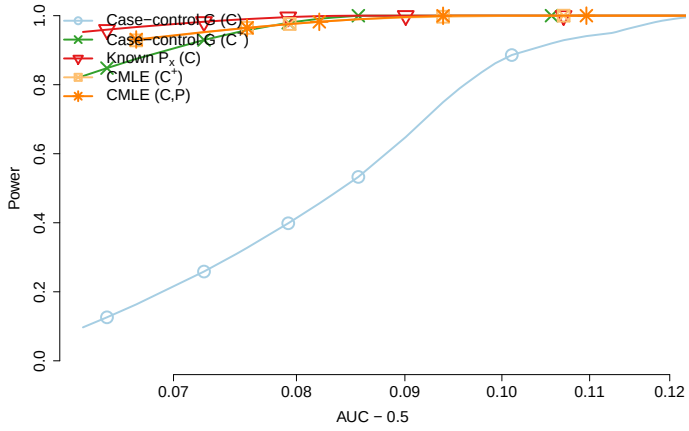


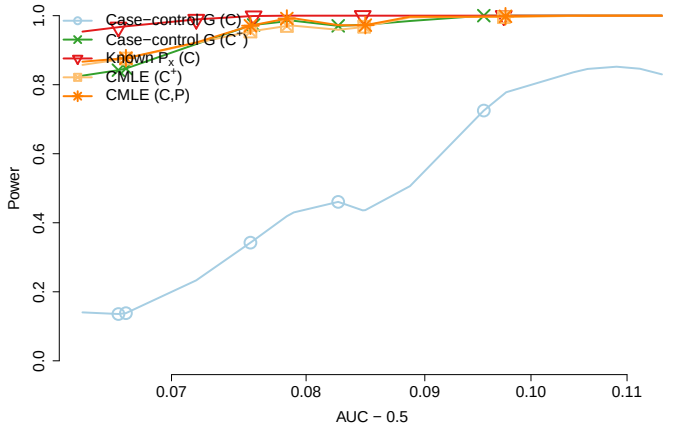












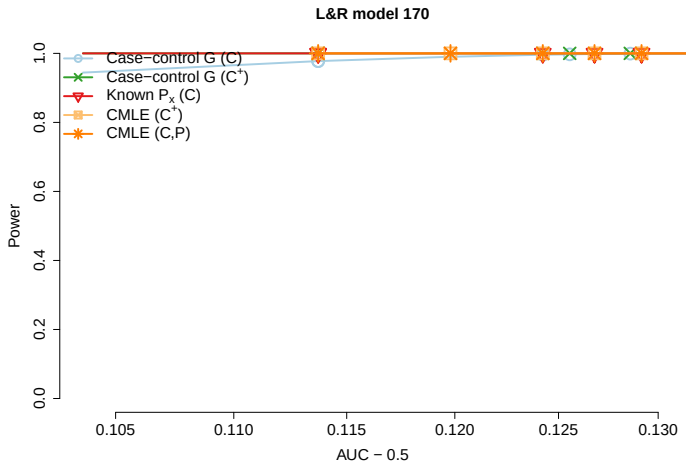
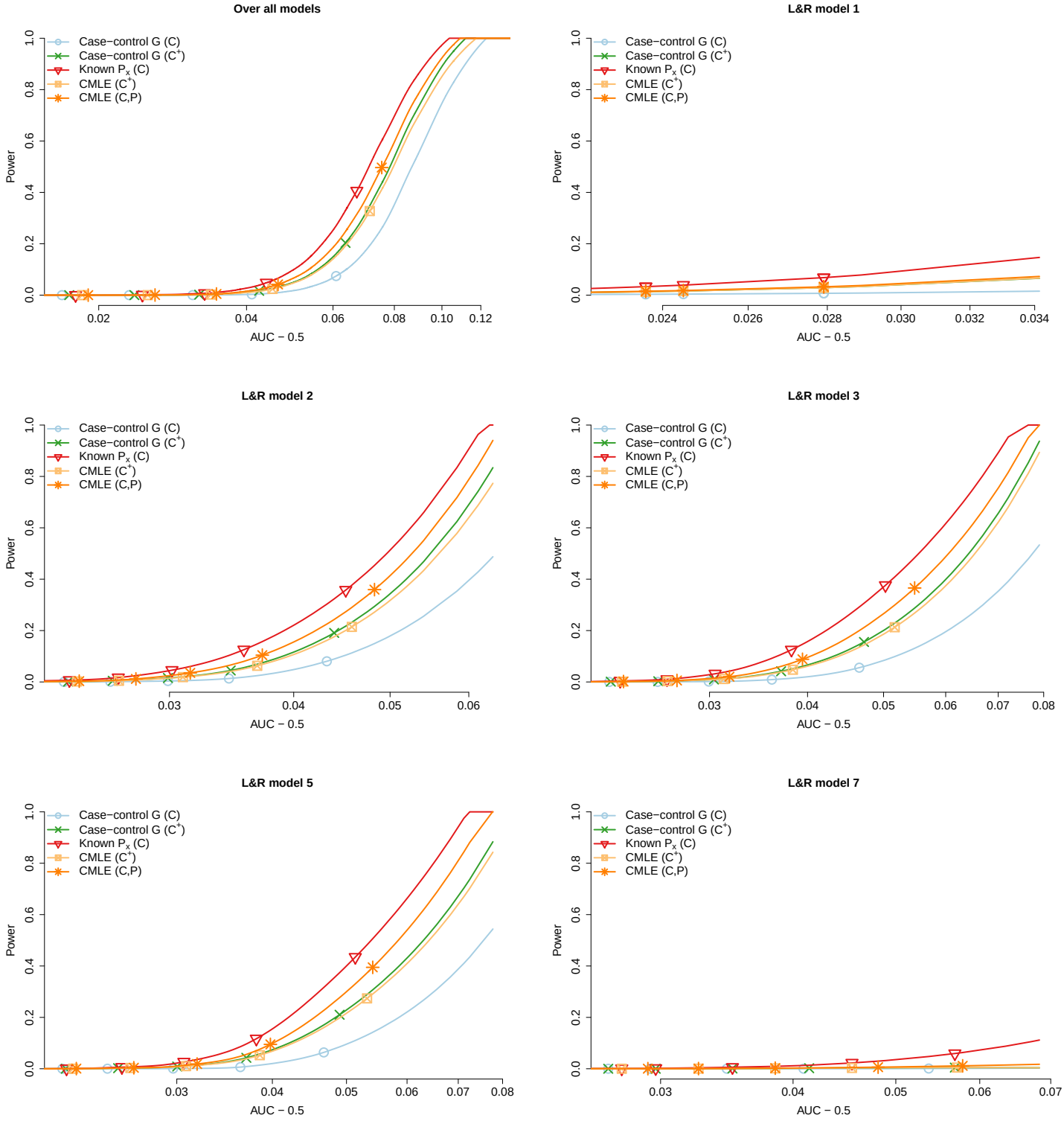
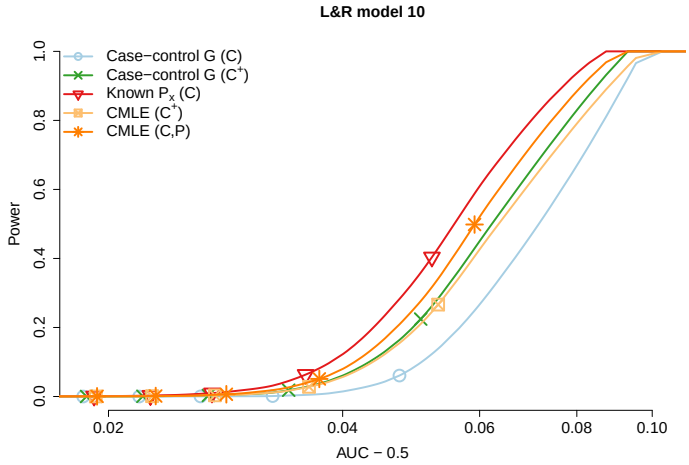
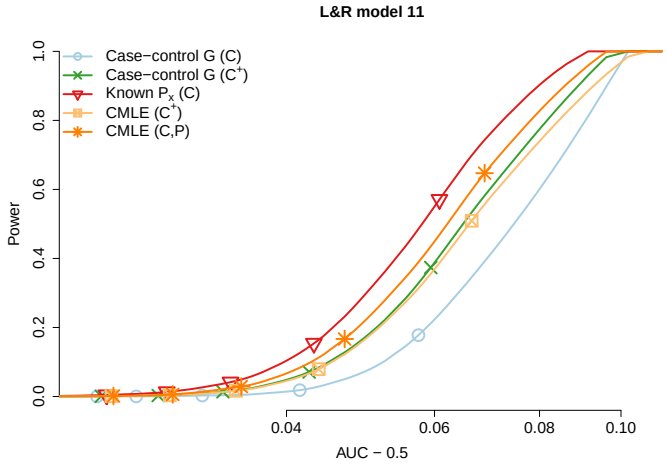
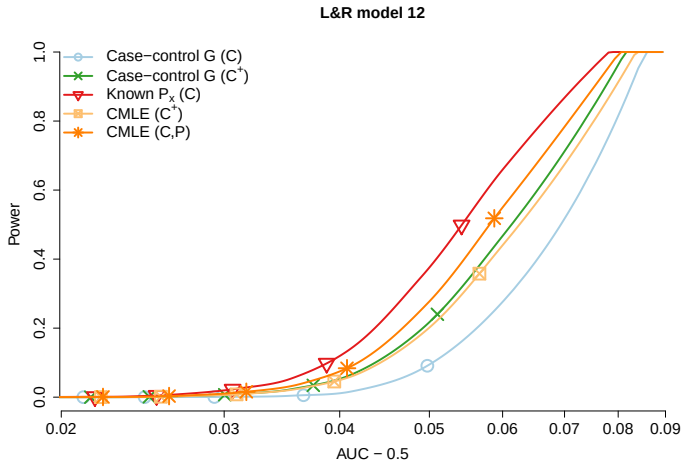


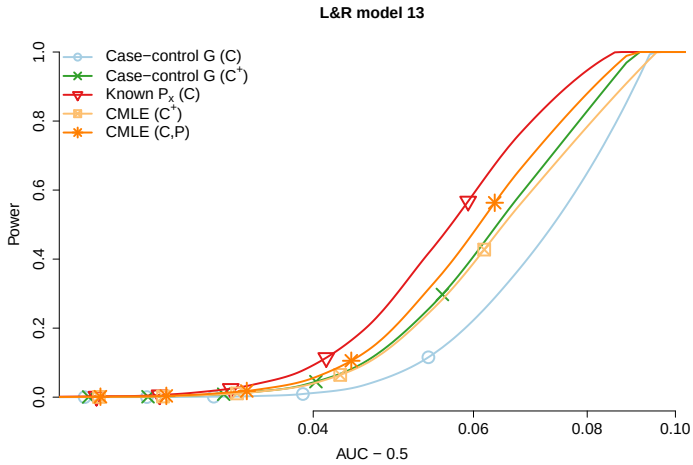
Figure S4: Power simulation results under LD for 0.20 prevalence.

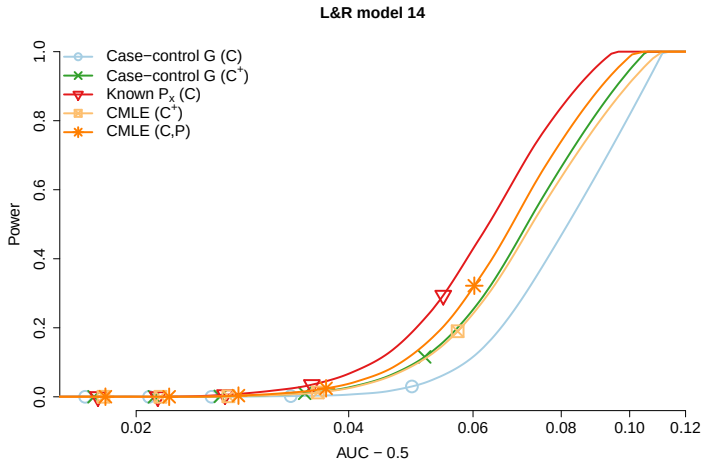


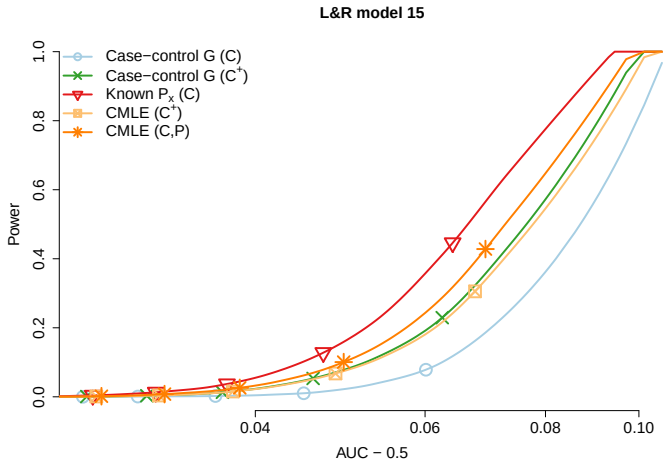


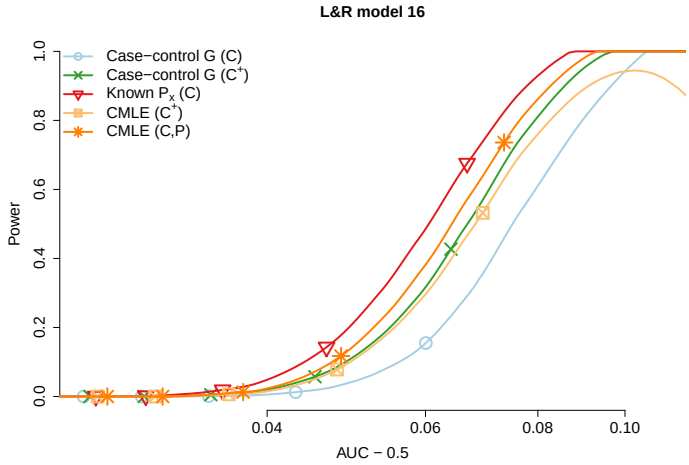


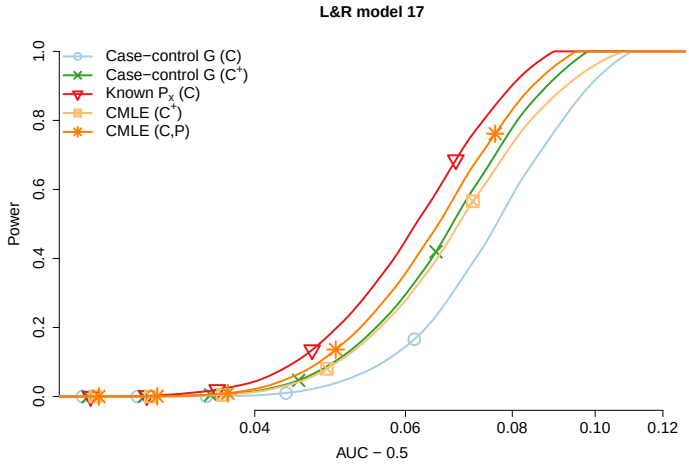


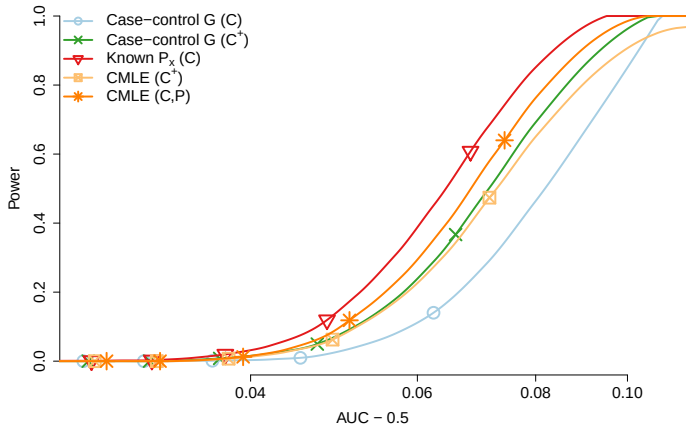


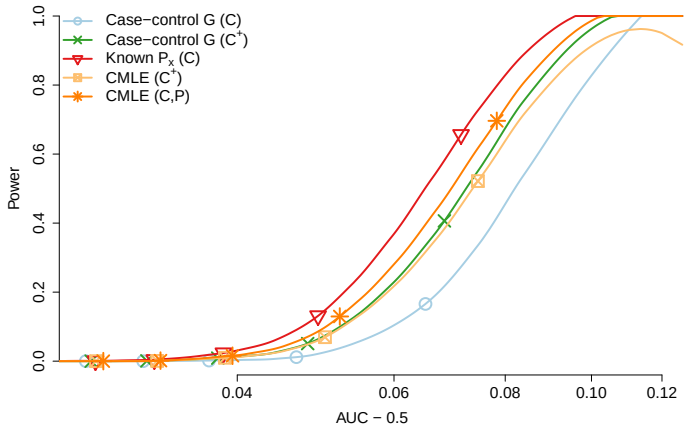


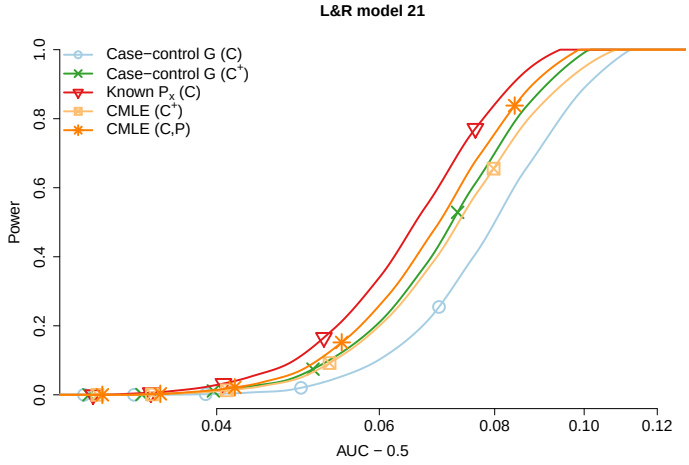


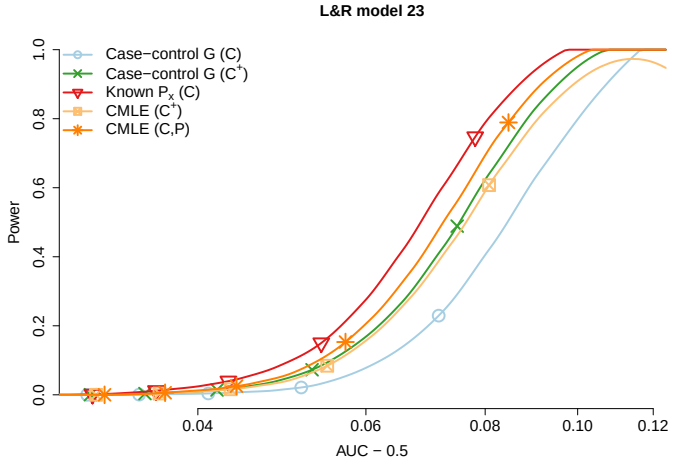


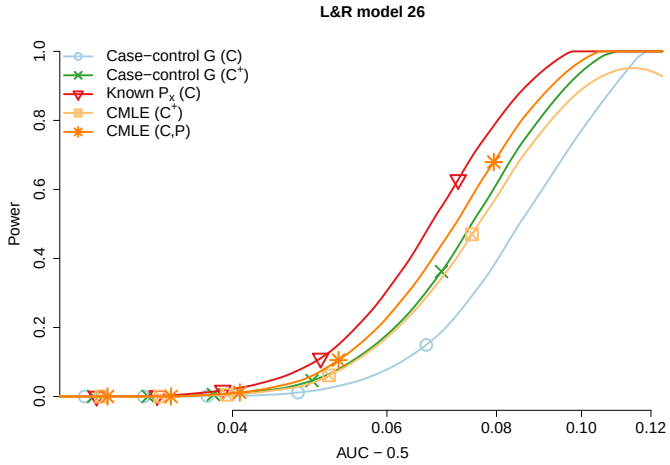


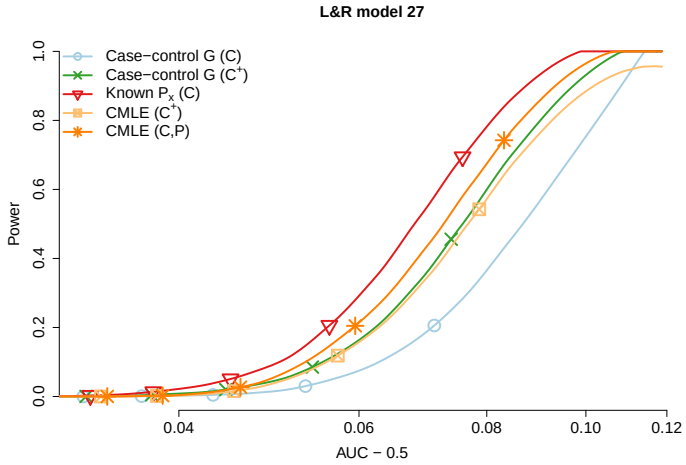


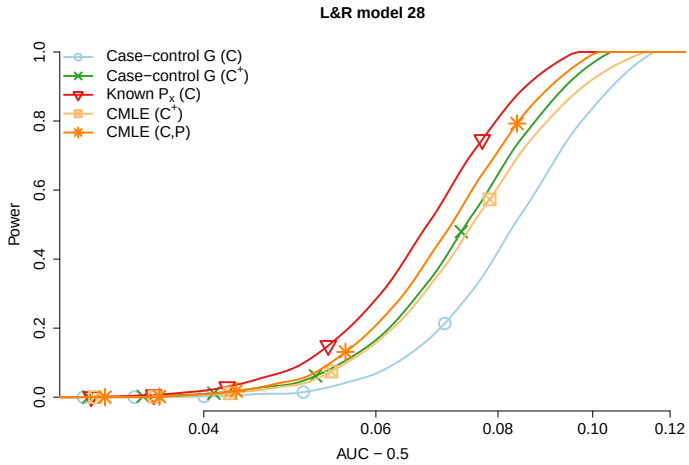


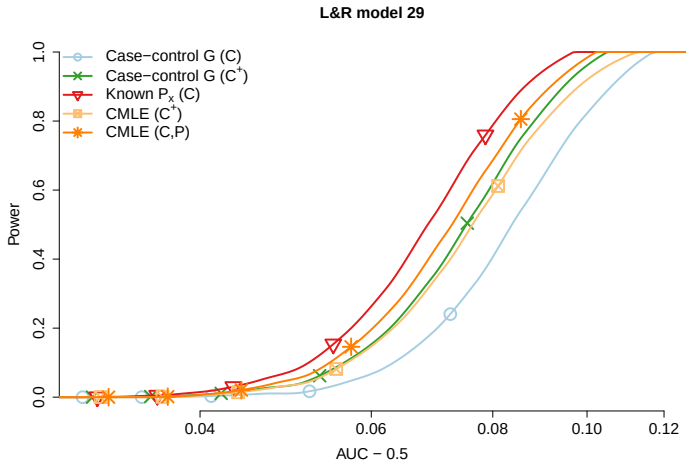


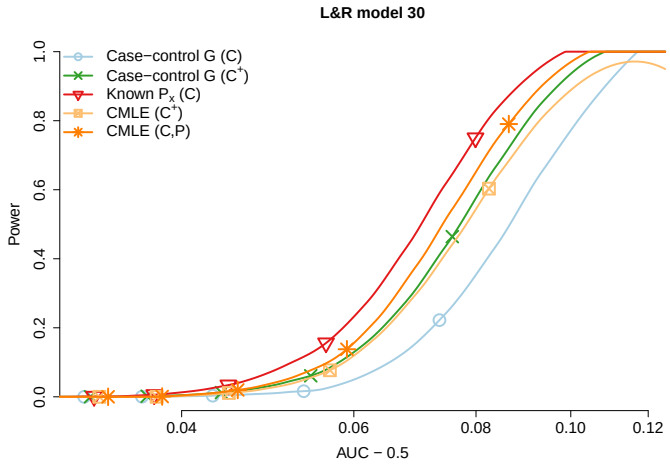


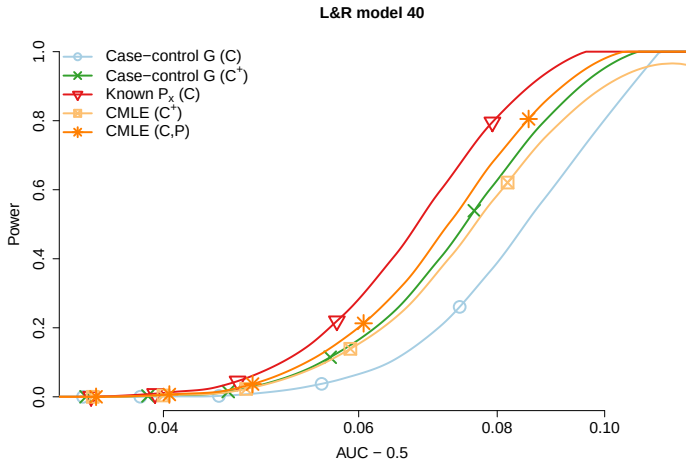


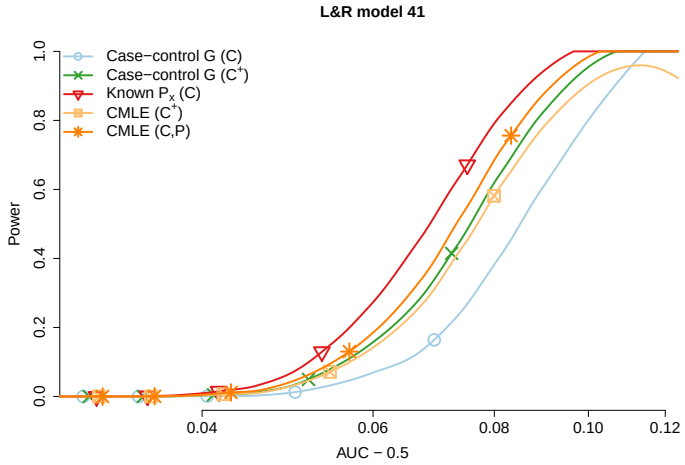


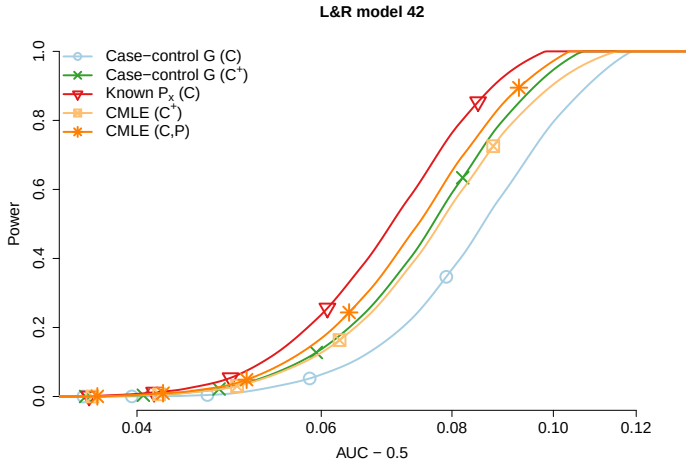


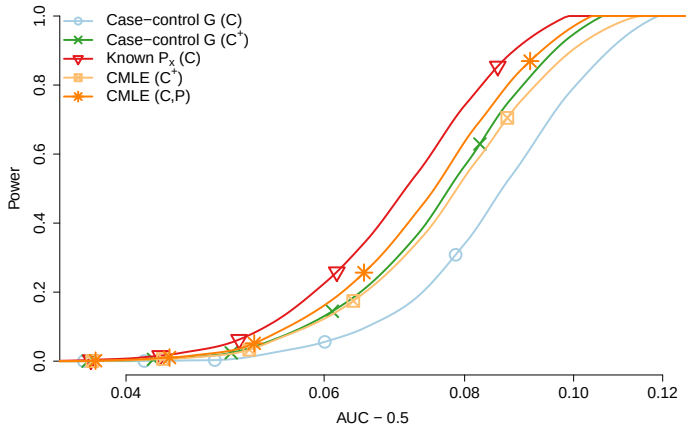


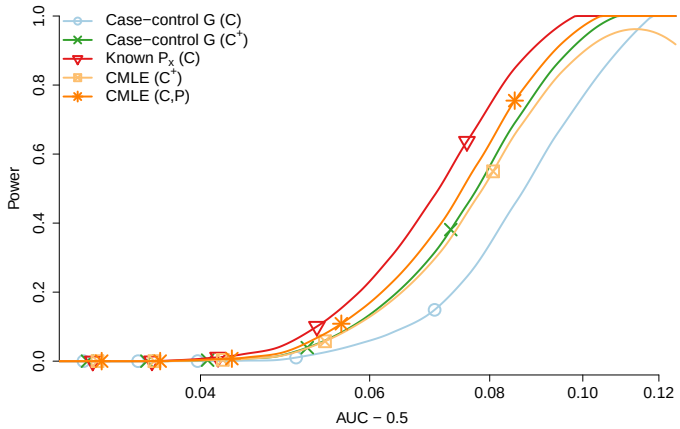


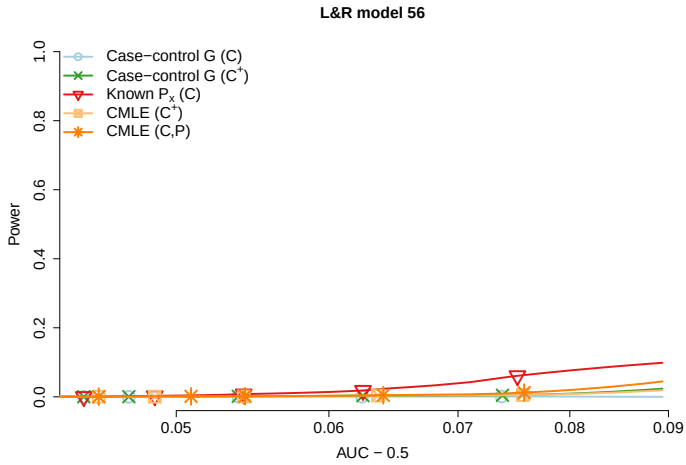


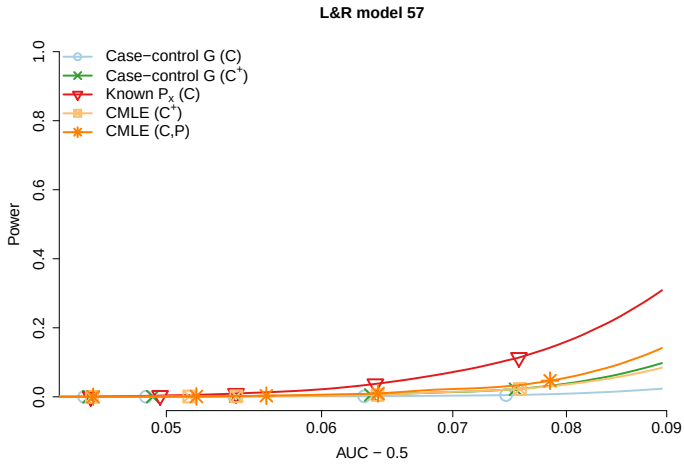


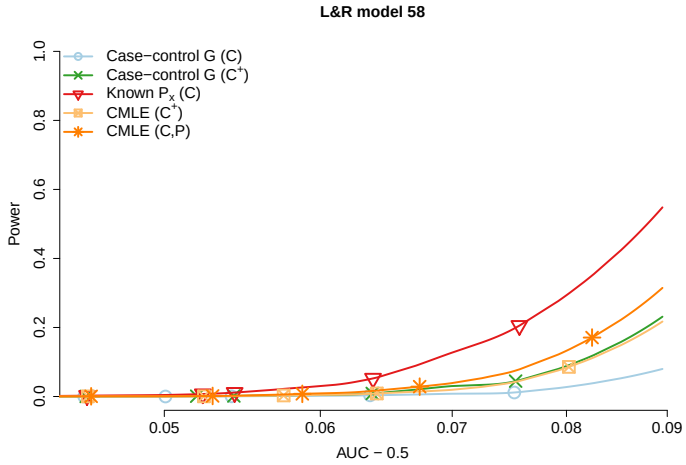


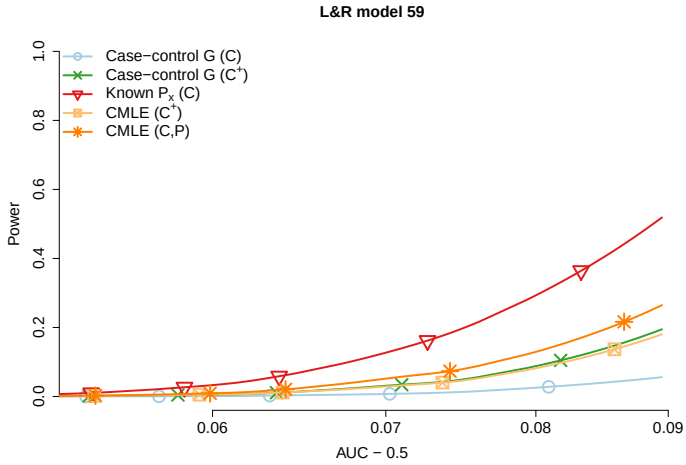


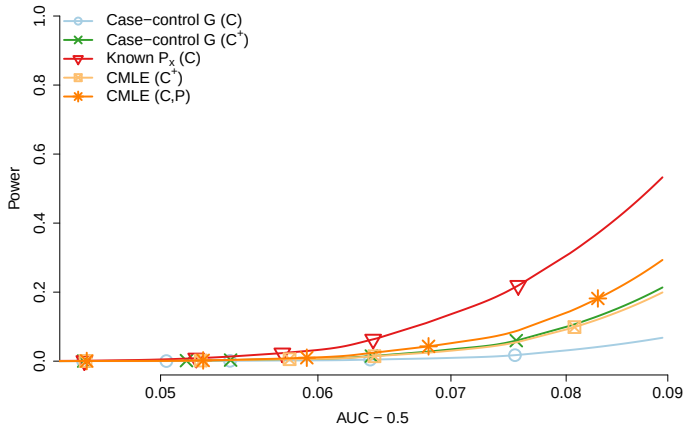


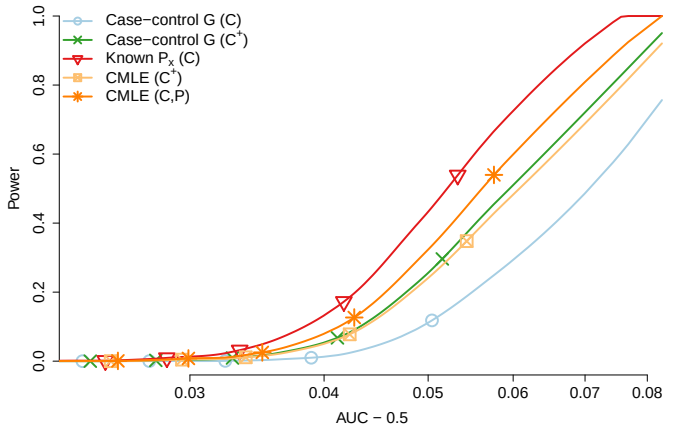


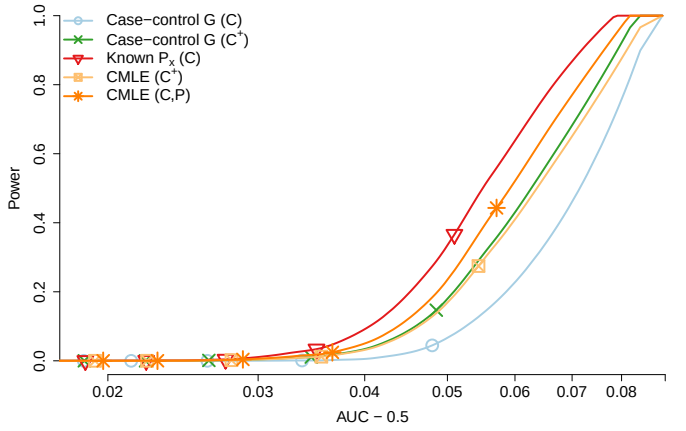


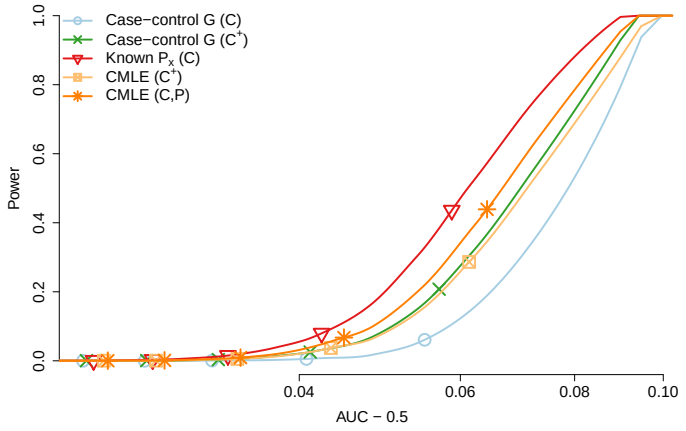


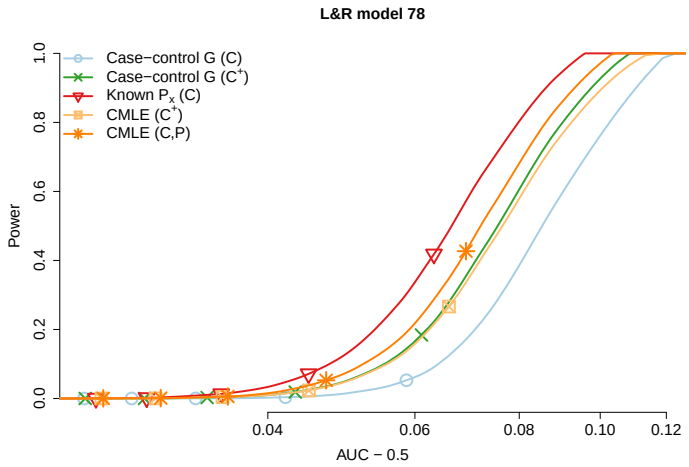


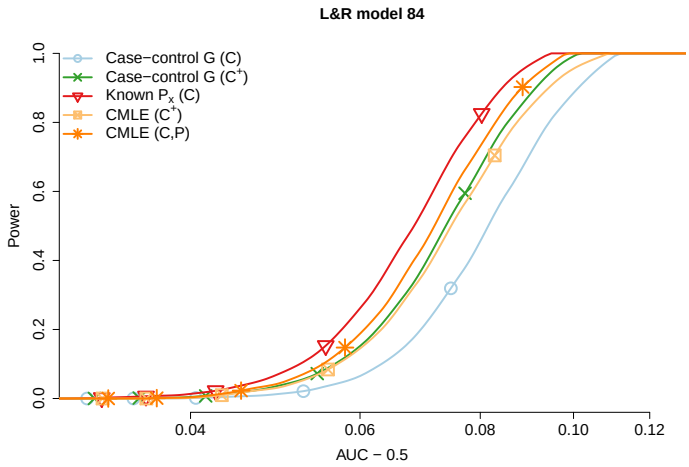


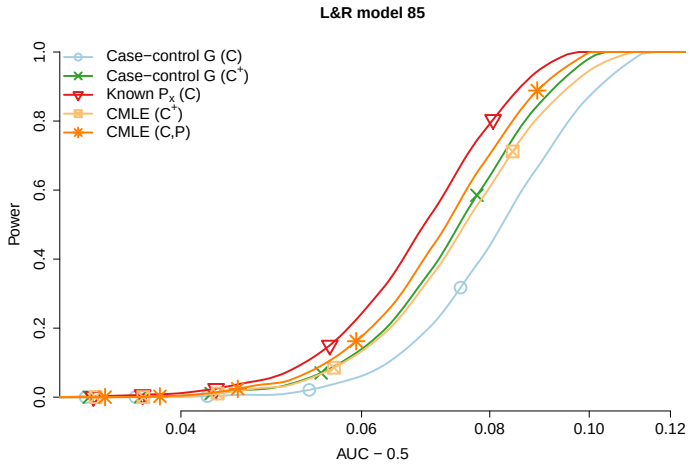


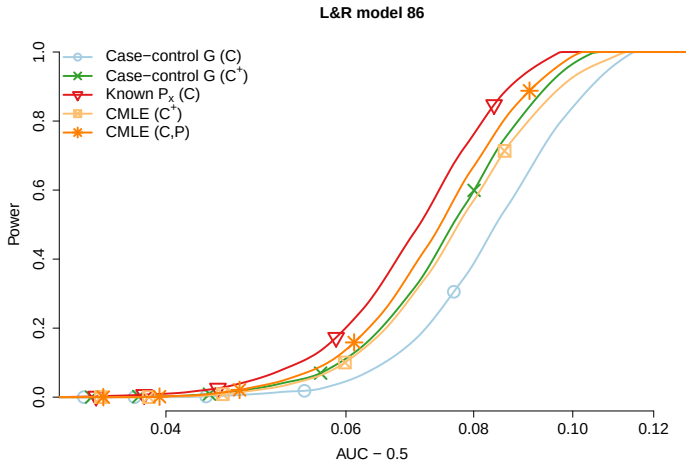


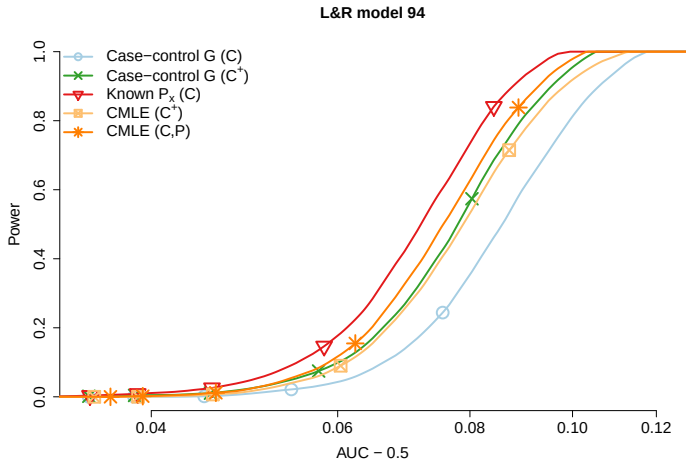


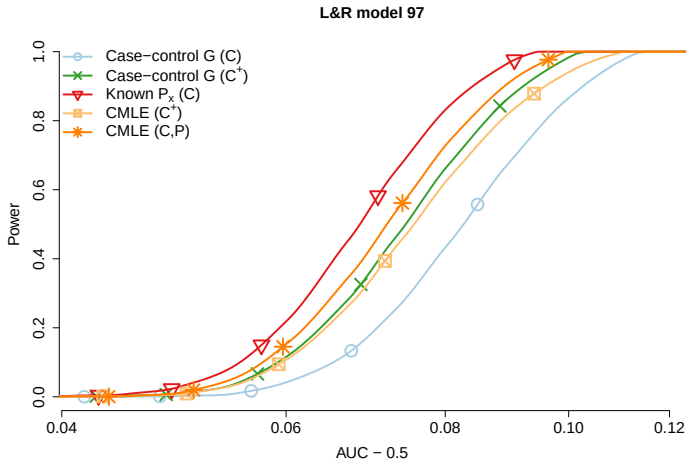


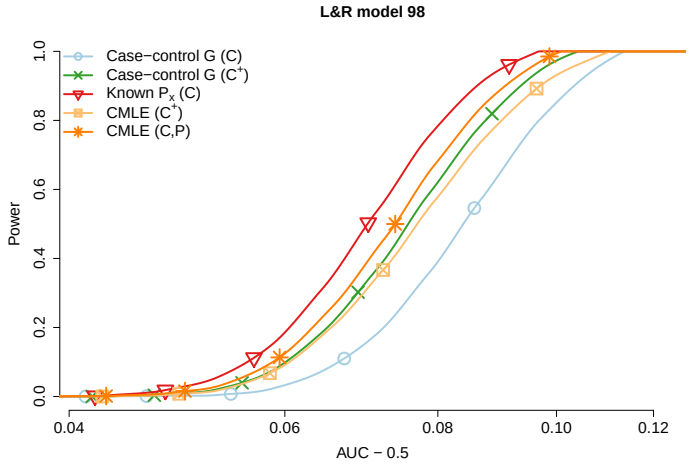


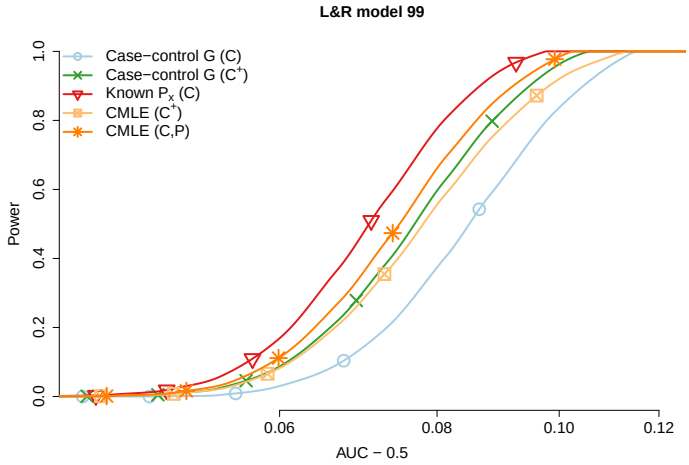


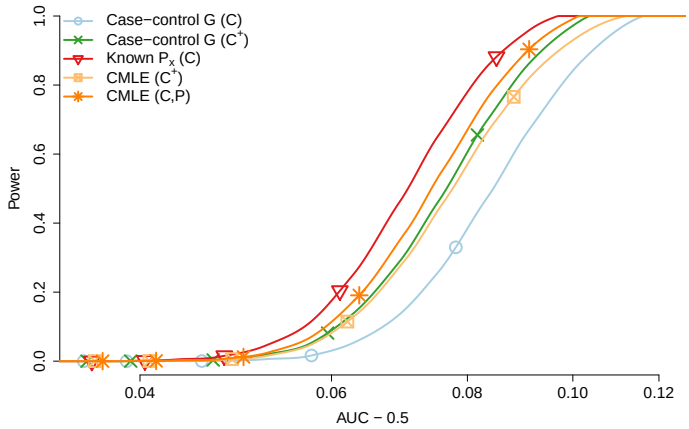


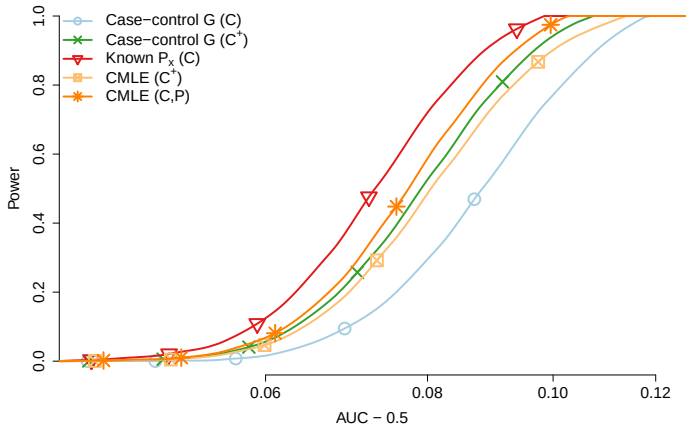


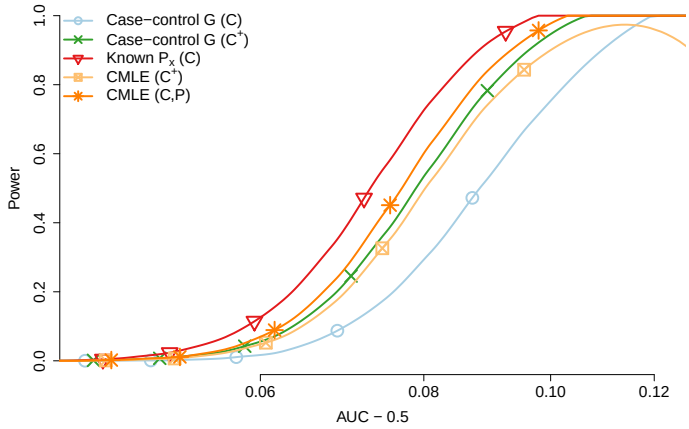


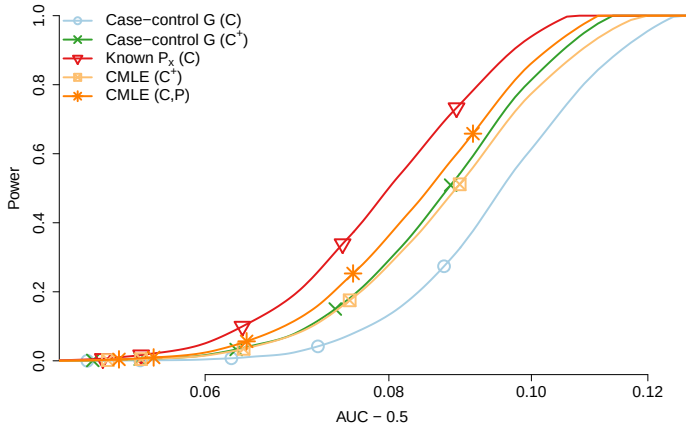


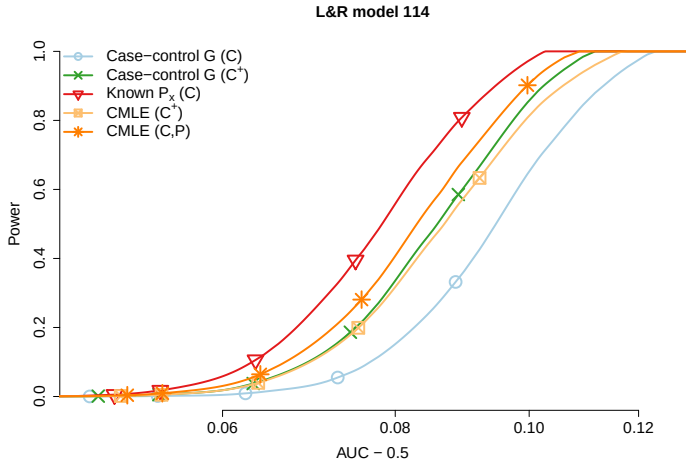


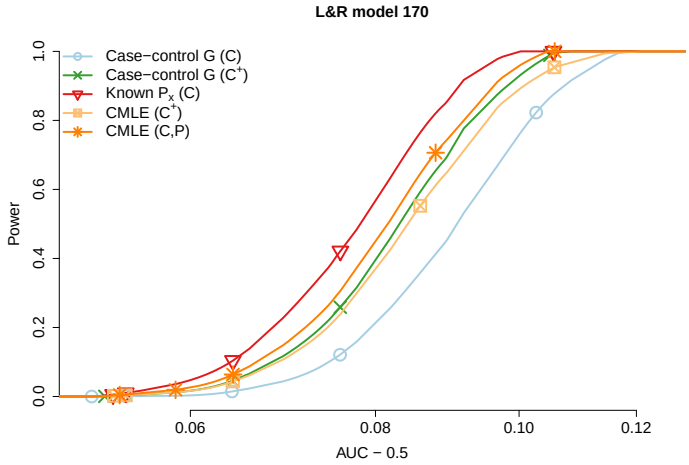


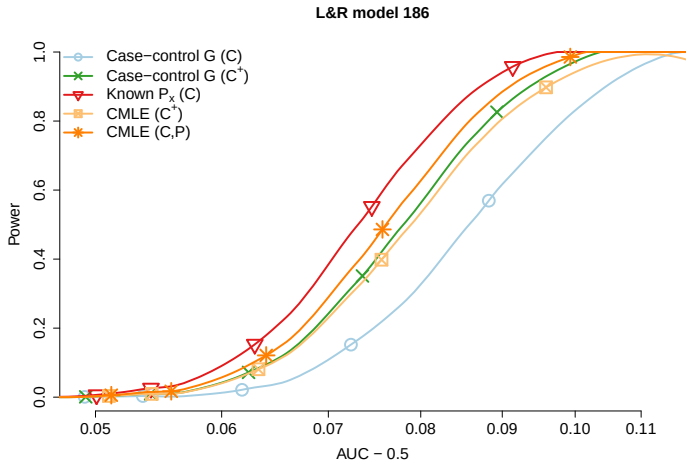












References

W. Li and J. Reich. A complete enumeration and classification of two-locus disease models. *Human heredity*, 50(6):334–349, 2000.

R.J. Neuman, J.P. Rice, and A. Chakravarti. Two-locus models of disease. *Genetic Epidemiology*, 9(5): 347–365, 2005.

File S1

Results Dataset

Available for download as a .zip file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162511/-/DC1