

Penalized Multimarker vs. Single-Marker Regression Methods for Genome-Wide Association Studies of Quantitative Traits

Hui Yi,^{*,†} Patrick Breheny,^{*} Netsanet Imam,^{*} Yongmei Liu,[§] and Ina Hoeschele^{*,**,*1}

^{*}Virginia Bioinformatics Institute, [†]Ph.D. Program in Genetics, Bioinformatics, and Computational Biology, and ^{**}Department of Statistics, Virginia Tech University, Blacksburg, Virginia 24061, [‡]Department of Biostatistics, University of Iowa, Iowa City, Iowa 52240, and [§]Departments of Epidemiology and Prevention and Internal Medicine, Division of Public Health Sciences, Translational Research Institute, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157

ABSTRACT The data from genome-wide association studies (GWAS) in humans are still predominantly analyzed using single-marker association methods. As an alternative to single-marker analysis (SMA), all or subsets of markers can be tested simultaneously. This approach requires a form of penalized regression (PR) as the number of SNPs is much larger than the sample size. Here we review PR methods in the context of GWAS, extend them to perform penalty parameter and SNP selection by false discovery rate (FDR) control, and assess their performance in comparison with SMA. PR methods were compared with SMA, using realistically simulated GWAS data with a continuous phenotype and real data. Based on these comparisons our analytic FDR criterion may currently be the best approach to SNP selection using PR for GWAS. We found that PR with FDR control provides substantially more power than SMA with genome-wide type-I error control but somewhat less power than SMA with Benjamini–Hochberg FDR control (SMA-BH). PR with FDR-based penalty parameter selection controlled the FDR somewhat conservatively while SMA-BH may not achieve FDR control in all situations. Differences among PR methods seem quite small when the focus is on SNP selection with FDR control. Incorporating linkage disequilibrium into the penalization by adapting penalties developed for covariates measured on graphs can improve power but also generate more false positives or wider regions for follow-up. We recommend the elastic net with a mixing weight for the Lasso penalty near 0.5 as the best method.

THE goal of genome-wide association studies (GWAS) in humans and model organisms is to select a small subset of DNA markers, typically single-nucleotide polymorphisms (SNPs), which are in strong linkage disequilibrium (LD) with functional polymorphisms affecting a biomedical/clinical trait of interest. The selected markers are then replicated in other GWAS, fine-mapped, and further validated. GWAS may be viewed as a large-scale variable selection problem, with several millions of common SNPs, measured directly or imputed, available in current studies in humans.

GWAS practitioners still strongly rely on single-marker analysis (SMA), including linear regression for continuous

phenotypes, with control of the genome-wise error rate (GWER) [a special case of the family-wise error rate (FWER)], which accounts for the multiplicity of the entire genome. Assuming that a biomedical trait of interest is affected by multiple polymorphisms with “detectable” effects, SMA fits an incorrect model, and it is sensible to consider alternative methods that test all or subsets of markers simultaneously. This approach requires a form of penalized regression (PR) as the number of SNPs is much larger than the sample size.

A major practical issue with the use of penalized regression is the determination of “optimal” values for the tuning parameter(s) and the lack of an error rate associated with the selection of SNPs. Common approaches for tuning parameter value determination include cross-validation (CV) (e.g., Kohavi 1995) and the use of a model selection criterion such as Akaike’s information criterion (AIC) (Akaike 1974, 1977), the Bayesian information criterion (BIC) (Schwarz 1978), or the extended Bayesian information criterion (EBIC) (Chen and Chen 2008). These approaches work well for prediction

Copyright © 2015 by the Genetics Society of America

doi: 10.1534/genetics.114.167817

Manuscript received June 28, 2014; accepted for publication October 21, 2014; published Early Online October 28, 2014.

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167817/-/DC1>.

¹Corresponding author: Virginia Bioinformatics Institute, Virginia Tech University, Blacksburg, VA 24061-0477. E-mail: inah@vbi.vt.edu

but not for the goal of variable selection as most of these criteria lead to an unacceptable number of false positives and do not provide a measure of error associated with the selected SNPs. This distinction is important as the best predictive models may contain covariates that are not important, while the model with all important covariates may not be the best predictive model.

The absence in PR of the control of an appropriate error rate such as the FWER or the false discovery rate (FDR) (Benjamini and Hochberg 1995; Sabatti and Freimer 2003) has been recognized by several authors who have employed different strategies, including data permutation (Ayers and Cordell 2010), stability selection (Meinshausen and Bühlmann 2010) with FDR control (Ahmed *et al.* 2011), and multistage (data splitting) approaches (Meinshausen *et al.* 2009; Wasserman and Roeder 2009) to control FWER or FDR. However, these approaches require much more computation and may not be feasible on a genome-wide scale, and/or the provision of an error rate estimate comes at the expense of reduced power.

A second practical issue is that a number of PR methods differing in the penalty function have been proposed, and hence it is unclear whether any and which of these methods should be preferred in the context of variable selection in GWAS. Moreover, some authors (Kim and Xing 2009; Liu *et al.* 2011) have suggested that penalties should incorporate LD, but it is unclear whether such penalties produce any gain in power and/or a reduction in the false positive findings.

The purpose of this contribution is to review penalized regression methods in the context of GWAS, to extend selected PR methods by incorporating FDR control and to assess their performance in comparison with single-marker (SNP) analysis, and to provide recommendations on the use of these methods in GWAS practice. PR methods are compared with SMA on realistically simulated GWAS, consisting of genotype data on single and multiple chromosomes and a continuous phenotype, and on real data.

Materials and Methods

Data simulation

To simulate GWAS data with realistic patterns of LD, we used the software HAPGEN2 (Su *et al.* 2011), which produces genotyped individuals by resampling from a set of reference haplotypes. We used the haplotypes of the 60 Caucasians of European origin (CEU) in HapMap2 (International HapMap Consortium 2007). For most simulation scenarios, SNP genotypes were simulated for a single chromosome (chromosome 21). For an additional simulation scenario, SNP genotypes were simulated for three chromosomes (19, 21, and 22). The number of SNPs genotyped across all 22 autosomes in the HapMap2 population is 3,849,034, and the numbers of SNPs on chromosomes 19, 21, and 22 are 56,607, 50,165, and 54,786, respectively. Following the

simulation of each SNP data set, SNPs were removed if they had a minor allele frequency (MAF) < 0.01 or if their absolute correlation with another SNP exceeded 0.999, reducing the number of SNPs on average to 25,033, 21,519, and 22,199 for chromosomes 19, 21, and 22, respectively.

To determine suitable sample sizes for the data simulation, we performed power calculations, using QUANTO (Gauderman and Morrison 2001). A functional or causal SNP affecting the phenotype is referred to as a quantitative trait locus (QTL). The sample sizes were determined as those needed to detect an isolated QTL by SMA, with a heritability (variance explained by the QTL over total variance) of 0.10, with a power of 50%, and with a P -value-based significance threshold of $(5.5 \times 10^{-8} \times 3,849,034/50,165)$ for chromosome 21 and $(5.5 \times 10^{-8} \times 3,849,034/(56,607 + 50,165 + 54,786))$ for the three-chromosome simulation, using the GWER P -value threshold of 5.5×10^{-8} (see below). The required sample sizes were $N = 201$ and $N = 222$, respectively.

For the chromosome 21 simulations, two scenarios were considered: (1) two isolated QTL with heritability of 0.10 each and hence a total heritability of 0.20 and (2) eight QTL comprising a group of four QTL with weak pairwise LD ($0.01 \leq r^2 \leq 0.1$) and two groups of two QTL each with within-group LD of $r^2 \approx 0.5$. All QTL had $MAF > 0.05$. For scenario 2, the four QTL in one group had a heritability of 0.05, and the remaining four QTL had a heritability of 0.04. Taking into account LD among the eight QTL in scenario 2, the total heritability was 0.48 [computed as $h^2 = \text{Var}(\sum_{k=1}^8 X_k \beta_k) / (\text{Var}(\sum_{k=1}^8 X_k \beta_k) + \sigma_e^2)$, where X_k is the allelic dose of SNP k in individual i , β_k is the regression coefficient for SNP k , and σ_e^2 is the residual variance]. Moreover, for scenario 2 the “effective” heritability of each individual QTL in the context of the single-marker model was ≈ 0.10 due to LD among QTL. Hence, SMA had the same expected power for QTL detection under both scenarios. A total of 200 replicate simulations were performed, in which the QTL positions and effects were kept constant. The phenotype was simulated based on an additive model with Q QTL, $Y = \sum_{j=1}^Q X_j \beta_j + e$, where Y denotes the phenotype. Heritability of a single QTL j was computed as $h_j^2 = \beta_j^2 \text{var}(X_j) / \text{var}(Y)$. Effective heritability (increased due to LD with other QTL) was computed as above but by replacing β_j with

$$\tilde{\beta}_j = \frac{\text{cov}(Y, X_j)}{\text{var}(X_j)} = \frac{\beta_j \text{var}(X_j) + \sum_{j' \neq j}^Q \beta_{j'} \text{cov}(X_j, X_{j'})}{\text{var}(X_j)}.$$

For the multichromosome simulation, chromosome 19 did not harbor any QTL, chromosome 21 harbored the eight QTL of scenario 2, and chromosome 22 harbored two

isolated QTL with heritability of 0.10 and MAF > 0.05 as in scenario 1. The total heritability was 0.68.

Single-marker regression

For single-marker regression, variable selection is performed by choosing a cutoff value for the P -values, determined by some method of multiple-testing control, most commonly the GWER (or FWER), which accounts for the multiplicity of the entire genome. The GWER-based P -value threshold is obtained by estimating an “effective number of independent tests” (e.g., International HapMap 2005; Dudbridge and Gusnanto 2008) by permutation or analytic approximation. The International HapMap Consortium permutation-based estimated significance threshold is 5.5×10^{-8} for two-sided tests of SNPs, which we used here. As an alternative to the stringent GWER threshold, we investigate the FDR (Benjamini and Hochberg 1995). For FDR control, we use the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg 1995) shown to control the FDR under positive-regression dependence (Benjamini and Yekutieli 2001), and this condition is believed to hold for GWAS (Sabatti and Freimer 2003). We also use Benjamini and Yekutieli’s (2001) BY approach, which ensures FDR control under any form of dependence but can be very conservative. Additionally, we consider the local FDR (Efron and Tibshirani 2002; Efron 2003), where the P -values, assumed to represent a mixture of null and alternative hypotheses, are transformed to z -scores that are modeled by a mixture distribution, or

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z), \quad (1)$$

where $f_0(z)$ is the density of z under the null hypothesis, $f_1(z)$ is the density under the alternative hypothesis, $f(z)$ is the (overall) mixture density of z , and π_0 is the proportion of true null hypotheses. Based on (1), the local FDR is

$$\text{locFDR}(z) = \frac{\pi_0 f_0(z)}{f(z)}, \quad (2)$$

which is also a posterior probability that the null hypothesis is true given z . The FDR is a lower bound on locFDR because it is the expectation of locFDR within a tail area (Efron 2003). Sun and Cai (2007) developed an oracle testing procedure that minimizes the false negative rate (the expected proportion of false negatives among all nonrejections) subject to a constraint on the FDR. This procedure is a thresholding rule based on the locFDR and is implemented with an adaptive procedure that is asymptotically valid and optimal if the estimates of π_0 , f_0 , and f in (2) are consistent. We estimated $f(z)$ and π_0 , using the simple nonparametric approach described in Storey *et al.* (2005). Despite the correlation structure of the z -values, f_0 was fitted well by the $N(0, 1)$ null distribution (see Supporting Information, Figure S1 as a typical representative of the data simulation replicates). We also used the consistent estimators of π_0 , f_0 , and f in equation 2 of Jin and Cai (2007). We refer to this procedure with the two sets of estimators for π_0 , f_0 , and f as

LFDR1 and LFDR2. The proportion of true nulls π_0 is expected to be ≤ 1.0 in the context of GWAS.

Penalized regression methods

All penalized regression methods for the multiple-regression model $y_i = \sum_{k=1}^p x_{ik} \beta_k + e_i$ (assuming a centered response variable y and standardized SNP covariates x) can be represented by the estimator

$$\hat{\vec{\beta}} = \arg \min_{\vec{\beta}} \left\{ \frac{1}{2N} \sum_{i=1}^N \left(y_i - \sum_{k=1}^p x_{ik} \beta_k \right)^2 + P(\vec{\lambda}, \vec{\beta}) \right\}, \quad (3)$$

where $\vec{\beta}$ is a vector of p regression coefficients, y_i is the continuous trait measurement on individual i , x_{ik} is allelic dose of SNP k in individual i , e_i is a residual, N is sample size, $P(\cdot)$ is a penalty function, and $\vec{\lambda}$ is a vector of tuning or penalty parameters with typically one or two components. PR methods differ in the specification of $P(\cdot)$. Variable selection, by setting unimportant coefficients to zero, occurs if the penalty has a singularity at zero. Below we consider only such penalties. Given the extensive literature, we do not review these methods in detail but rather present their penalty functions and discuss some properties relevant to their application to GWAS.

Desirable properties for penalized regression methods have been established by several authors and include sparsity (variable selection is enabled by automatically setting small coefficients to zero), continuity (an estimator that is continuous in data avoids instability in model prediction), asymptotic unbiasedness (bias should be low, in particular for large true coefficients), and the (strong) oracle property (Fan and Li 2001; Fan and Peng 2004; Kim *et al.* 2008; Zhang 2010). The Lasso (Tibshirani 1996) performs poorly with regard to the unbiasedness property, prompting the development of other penalties to overcome this problem. The adaptive Lasso (Zou 2006), the smoothly clipped absolute deviation (SCAD) (Fan and Li 2001), and the minimax concave penalty (MCP) (Zhang 2010) possess the (strong) oracle property.

Lasso, elastic net, and adaptive lasso: The Lasso (Tibshirani 1996) employs the L1 penalty, or

$$P_{\text{Lasso}}(\vec{\lambda}, \vec{\beta}) = \lambda \sum_{k=1}^p |\beta_k|; \quad \lambda \geq 0. \quad (4)$$

If there are groups of variables with strong pairwise correlations, it has been empirically observed that the Lasso tends to select only one variable from each group. This behavior is in contrast with single-marker regression, which selects all markers in sufficiently strong LD with a QTL.

The elastic net (EN) (Zou and Hastie 2005) combines the Lasso (L1) penalty, enabling it to perform variable selection, with the ridge (L2) penalty, enabling it to deal with multicollinearity, or

$$P_{\text{EN}}(\vec{\lambda}, \vec{\beta}) = \lambda_1 \lambda_2 \sum_{k=1}^p |\beta_k| + \lambda_1 (1 - \lambda_2) \sum_{k=1}^p \beta_k^2; \quad (5)$$

$$\lambda_1 \geq 0, \quad 0 \leq \lambda_2 \leq 1.$$

In contrast with the Lasso, the EN may be expected to coselect SNPs in strong LD. Here, EN analysis was performed with three different values for λ_2 : 0.9, 0.5, and 0.3. Lasso and EN are implemented in the R package glmnet (Friedman *et al.* 2010).

Zou (2006) proposed the adaptive Lasso to obtain an oracle procedure, which has the penalty

$$P_{\text{AdaLasso}}(\vec{\lambda}, \vec{\beta}) = \lambda \sum_{k=1}^p w_k |\beta_k|; \quad \lambda \geq 0, \quad w_k = |\hat{\beta}_k|^{-1}, \quad (6)$$

where the weights w_k are computed from an initial set of solutions for the coefficients. The adaptive Lasso makes sense intuitively by placing smaller weights on the important regressors to reduce the shrinkage applied to their coefficients. Here we used SMA, Lasso with CV, and ridge regression with CV as the initial estimators.

MCP and SCAD: The MCP and SCAD penalties depend on two tuning parameters; both begin with the same penalization as the Lasso but increasingly reduce the penalization farther away from zero (*e.g.*, Breheny and Huang 2011). Because in preliminary studies we found little difference between the MCP and SCAD results and SCAD is expected to perform somewhere between MCP and Lasso (Breheny and Huang 2011), here we consider only the MCP with penalty function

$$P_{\text{MCP}}(\vec{\lambda}, \beta_k) = \lambda_1 \int_0^{|\beta_k|} \left[1 - \frac{x}{\lambda_1 \lambda_2} \right]_+ dx$$

$$= \lambda_1 \begin{cases} \left(|\beta_k| - \frac{\beta_k^2}{2\lambda_1 \lambda_2} \right) & \text{if } |\beta_k| \leq \lambda_1 \lambda_2 \\ \frac{\lambda_1 \lambda_2}{2} & \text{if } |\beta_k| > \lambda_1 \lambda_2; \end{cases} \quad (7)$$

$$\lambda_1 \geq 0, \quad \lambda_2 > 1.$$

For the second tuning parameter in MCP (λ_2), four values were evaluated (3, 10, 30, and 100), with the largest value approaching the Lasso. MCP is implemented in the R package ncvreg (Breheny and Huang 2011).

Selection of tuning parameter values: Current state: A major practical issue with the use of penalized regression for GWAS is the need to determine optimal values for the tuning parameter(s). The “usual” approaches include CV and the use of model selection criteria such as AIC, BIC, and EBIC. While CV and AIC are generally useful for prediction rather than model/variable selection, BIC tends to select a true sparse model but does not achieve sufficient

sparsity in high-dimensional very sparse settings such as linkage (QTL) mapping and GWAS for which modified BIC criteria have been proposed (*e.g.*, Bogdan *et al.* 2004). Moreover, the BIC criterion is a function of the degrees of freedom (d.f.) of a model, and the d.f. are not straightforward to obtain in PR (*e.g.*, Ye 1998; Zou *et al.* 2007; Zhang 2010).

Given the shortcomings of model selection criteria, a superior approach should be obtained by combining PR with the control of an error rate such as FWER or FDR. Ayers and Cordell (2010) used data permutation to determine the value of the tuning parameter. Multistage strategies have also been suggested as an approach to performing error control with PR (Meinshausen *et al.* 2009; Wasserman and Roeder 2009). Meinshausen *et al.* randomly split the data in half multiple (B) times. Each time the first half was used for variable screening using Lasso with CV and related methods, and the second half was used to compute adjusted P -values, based on which the FWER and the FDR can be controlled asymptotically. We used Lasso with CV for tuning parameter value selection in the screening step. We expected this method to have reduced power relative to our single-step FDR-controlling methods (see below) due to the data splitting and the use of multiple regression, which reduces the significance of two (highly) correlated variables representing alternative hypotheses. Performing the two steps of screening and variable selection both on the entire data set as in Sun *et al.* (2010) should improve power; these authors used their Iterative Adaptive Lasso in the screening step and backward selection instead of multiple regression in the second (cleaning) step. However, there is no analytical proof of (asymptotic) FDR (or FWER) control, and we have therefore not further considered a two-step approach performed on the entire data set.

Most recently, Sampson *et al.* (2013) have proposed a local FDR-based method for selecting the value of the tuning parameter specifically for the adaptive Lasso. The advantage of this approach is that it uses the local FDR based on which optimal oracle procedures were developed for SMA (Sun and Cai 2007; Cai and Sun 2009).

The penalized unified multiple-locus association (PUMA) method (and software) implements some penalized regression methods for GWAS with an efficient algorithm for generalized linear models and with pre-selection of SNPs based on SMA (Hoffman *et al.* 2013) to make these methods applicable to large GWAS data sets. It performs heuristic model and SNP selection. PUMA implements MCP with a two-dimensional model search across its two tuning parameters (2D-MCP), and we chose this method here.

A different type of multivariate method for SNP selection in GWAS was proposed by Zuber *et al.* (2012). For a continuous phenotype, this method computes correlation-adjusted marginal correlations (CAR scores) as $P_{XY}^{\text{adj}} = P_{XX}^{-1/2} P_{XY}$, where P_{XY} is a vector of marginal correlations between each SNP dose (X) and the phenotype (Y), and P_{XX} is the correlation matrix among the SNPs, which is estimated by

a computationally efficient shrinkage method. Note that if there are non-SNP study covariates, the phenotype (Y) must be replaced with the residuals from a linear model containing these covariates. For SNP selection, CAR scores are provided as input to the `fdrtools` R package that computes P -values, FDR, and local FDR values based on its empirical null modeling (Strimmer 2008). We refer to this method as CAR.

Selection of tuning parameter values: Control of the false discovery rate

Here we propose two approaches for combining penalized regression with FDR control, which can be applied to all penalized regression methods considered here and do not require data splitting or subsampling. The first approach is based on data permutation and the second on an analytic approximation.

FDR control using data permutation: Let λ denote the single (Lasso) or the first (MCP, EN) tuning parameter. Penalized regression [with coordinate descent algorithms (Friedman *et al.* 2007)] is efficiently performed by computing the coefficient solutions on a grid of (say, 1000) λ -values, starting from a maximum value λ_{\max} at which all solutions are zero and ending at a minimum value λ_{\min} that is zero or produces an excessively large model, with the solutions from any previous λ -value serving as a warm start for the next λ . Therefore, the original data set is analyzed on this grid of λ -values, and the number of nonzero coefficients is determined at each λ , denoted by $R(\lambda)$. Then B permuted data sets (with random reorderings of the continuous vector of phenotypes) are analyzed on the same grid of λ -values, and the number of nonzero coefficients is again determined for each permuted data set b ($b = 1, \dots, B$) and λ -value, denoted by $F(b, \lambda)$. Then for each λ -value, we compute the permutation FDR as the average number of nonzero solutions in the permuted data over the number of nonzero solutions in the original data, $FDR(\lambda) = (1/B) \sum_{b=1}^B F(b, \lambda) / R(\lambda)$. We note that here we define the FDR as $E(F)/R$, where F denotes the number of false positives and R the number of rejections (recall that in PR a null hypothesis is rejected if the corresponding coefficient solution is nonzero). This definition is different from the original definition $E(F/R)$ and thus does not take into account the dependency between F and R , but is easier to work with.

Analytic FDR control: We now present an approximate, analytic FDR method that was originally proposed in Breheny (2009). This approach can be applied to all penalized regression methods considered here, but we illustrate it for the Lasso and the MCP. We assume that the predictors have been standardized such that $\sum_{i=1}^N x_{ik} = 0$ and $\sum_{i=1}^N x_{ik}^2 = N$. We write the objective function in terms of a single predictor (k), as in Friedman *et al.* (2007) for the Lasso and in Breheny and Huang (2011) for the MCP [setting $\lambda = \lambda_1$ and $\gamma = \lambda_2$ in (7)],

$$f_{\text{Lasso}}(\beta_k) = \frac{1}{2}(\beta_k - \beta_k^{\text{OLS}})^2 + \lambda|\beta_k|; \tag{8}$$

$$f_{\text{MCP}}(\beta_k) = \frac{1}{2}(\beta_k - \beta_k^{\text{OLS}})^2 + \lambda|\beta_k| - \frac{1}{2\gamma}\beta_k^2,$$

where $\beta_k^{\text{OLS}} = (1/N) \sum_{i=1}^N x_{ik} r_{i-k}$ and $r_{i-k} = y_i - \sum_{j=1, j \neq k}^p x_{ij} \beta_j$. Differentiating (8) with respect to β_k yields the solutions

$$\beta_k^{\text{Lasso}} = \begin{cases} \beta_k^{\text{OLS}} - \lambda & \text{if } \beta_k^{\text{OLS}} > 0 \text{ and } |\beta_k^{\text{OLS}}| > \lambda \\ \beta_k^{\text{OLS}} + \lambda & \text{if } \beta_k^{\text{OLS}} < 0 \text{ and } |\beta_k^{\text{OLS}}| > \lambda \\ 0 & \text{if } |\beta_k^{\text{OLS}}| \leq \lambda \end{cases}$$

$$\beta_k^{\text{MCP}} = \begin{cases} \frac{\beta_k^{\text{OLS}} - \lambda}{1 - 1/\gamma} & \text{if } \beta_k^{\text{OLS}} > 0 \text{ and } |\beta_k^{\text{OLS}}| > \lambda \\ \frac{\beta_k^{\text{OLS}} + \lambda}{1 - 1/\gamma} & \text{if } \beta_k^{\text{OLS}} < 0 \text{ and } |\beta_k^{\text{OLS}}| > \lambda \\ 0 & \text{if } |\beta_k^{\text{OLS}}| \leq \lambda \\ \beta_k^{\text{OLS}} & \text{if } |\beta_k^{\text{OLS}}| > \gamma\lambda. \end{cases} \tag{9}$$

Based on (9), we consider the probability of a false positive, or $Pr(|\beta_k^{\text{OLS}}| > \lambda | \beta_k = 0)$. We again define the FDR as $E(F)/R$ and approximate the numerator with

$$\widehat{E(F)} = \sum_{k=1}^p Pr(|\beta_k^{\text{OLS}}| > \lambda | \beta_k = 0). \tag{10}$$

In general, FDR control based on (10) is expected to be conservative, similar to the BH FDR, because in (10) we sum over all variables rather than the (unknown) true null variables. However, this should not matter in the context of GWAS where the proportion of null variables is very close to 1.

We rewrite the probability of a false positive as

$$Pr(|\beta_k^{\text{OLS}}| > \lambda | \beta_k = 0) = Pr\left(\frac{1}{N} |\mathbf{x}_k^T \mathbf{r}_{-k}| > \lambda \mid \beta_k = 0\right). \tag{11}$$

The distribution of the estimated residuals (r) is unknown and complicated but may be approximated by the normal distribution $\mathbf{r}_{-k} \sim N(\mathbf{0}, \mathbf{V}_k)$ or even simpler by $\mathbf{r}_{-k} \sim N(\mathbf{0}, \mathbf{I}\sigma_k^2)$ such that $(1/N)\mathbf{x}_k^T \mathbf{r}_{-k} \sim N(0, (1/N)\sigma_k^2)$. This is an approximation for multiple reasons, including the normality assumption, the mean of zero (correct for least squares where the coefficients are estimated unbiasedly but not for PR), and the independence or zero covariances. Then

$$Pr\left(\frac{1}{N} |\mathbf{x}_k^T \mathbf{r}_{-k}| > \lambda \mid \beta_k = 0\right) = 1 - \Phi\left(\frac{\sqrt{N}\lambda}{\sigma_k}\right) + \Phi\left(-\frac{\sqrt{N}\lambda}{\sigma_k}\right) = 2\Phi\left(-\frac{\sqrt{N}\lambda}{\sigma_k}\right). \tag{12}$$

Using the “natural” estimator $\hat{\sigma}_k^2 = (1/N)\mathbf{r}^T \mathbf{r}$, (12) becomes constant for all k and

$$\widehat{\text{FDR}} = \frac{2p\Phi\left(-\left(\lambda N/\sqrt{\mathbf{r}^T \mathbf{r}}\right)\right)}{R}. \quad (13)$$

The FDR estimation (13) is noisy when the model becomes saturated; this is not a problem in practice when FDR evaluation starts from a model with all coefficients set to zero (λ_{\max}) and terminates once the desired level of the FDR has been stably exceeded.

When applying the FDR estimation described for the Lasso and the MCP to the EN, Equation 13 needs to be modified to

$$\widehat{\text{FDR}} = \frac{2p\Phi\left(-\left(\lambda_1\lambda_2 N/\sqrt{\mathbf{r}^T \mathbf{r}}\right)\right)}{R}. \quad (14)$$

Penalized regression with fusion-type penalties

It is well known that the Lasso tends to select a single predictor from a group of predictors that have strong pairwise correlations (in the extreme the Lasso does not have a unique solution when two variables are perfectly collinear). For this reason, we prefer the EN (see below), but some authors (Kim and Xing 2009; Liu *et al.* 2011) have suggested that fused Lasso-type penalties developed for covariates measured on graphs would be more appropriate than the Lasso in such situations. An expectation was that such penalties may increase power and decrease the false positive rate. Fusion-type penalties impose pairwise similarity on the coefficient solutions of highly correlated predictors or encourage sparsity in the differences among the values of the regression coefficients. Pairwise similarity can be imposed on the effects of correlated (SNP) regressors by adding an extra penalty to the existing penalty term, depending on the type of PR method used (here the Lasso). We first consider an added penalty of the form (referred to as LD2lasso)

$$\frac{\phi}{2} \sum_{k=1}^p \sum_{m \in S_k} h(r_{km}) \cdot \left(|\beta_k| - |\beta_m|\right)^2; \quad \phi \geq 0, \quad (15)$$

where S_k is a set of SNPs that are correlated to SNP k , r_{km} denotes the (Pearson) correlation coefficient among the allelic doses of SNPs k and m , and $h(r_{km})$ is a function of r_{km} that we specified as $h(r_{km}) = |r_{km}|$ or $h(r_{km}) = (r_{km})^2$. Set S_k can be limited to adjacent SNPs (Liu *et al.* 2011) or contain all SNPs whose absolute correlation with SNP k exceeds a certain threshold (here). Taking absolute values of the coefficients accounts for the fact that two correlated SNPs can have similar effect sizes but opposite signs due to the arbitrary coding of the SNP alleles. We note that Li and Li (2008, 2010) used a similar penalty for the general case of (genomics) variables whose dependency structure can be represented as a graph. Their penalty is induced by the Laplace matrix of the graph and differs from (15) by dividing the regression coefficients by the square root of the degree of the corresponding variables, which is justified by the

argument that variables with more connections should be allowed to have larger coefficients. This argument does not apply in the GWAS context, and we therefore use (15). Using (3) with the Lasso penalty and the added penalty in (15), the LD2lasso objective function for a single coefficient k can be written as

$$f_{\text{LD2lasso}}(\beta_k) = \frac{1}{2}(\beta_k - \beta_k^{\text{OLS}})^2 + \lambda\phi|\beta_k| + \lambda(1 - \phi) \times \frac{1}{2} \sum_{m \in S_k} h(r_{km}) \left(|\beta_k| - |\beta_m|\right)^2; \quad \lambda \geq 0, 0 \leq \phi \leq 1, \quad (16)$$

where

$$\beta_k^{\text{OLS}} = \frac{1}{N} \mathbf{x}_k^T \left(y_i - \sum_{j=1, j \neq k}^p x_{ij} \beta_j^{\text{LD2lasso}} \right).$$

Then one can show that the LD2lasso solution satisfies

$$\begin{aligned} \beta_k^{\text{LD2lasso}} &= \text{sign}(\beta_k^{\text{OLS}}) \\ &\times \frac{\left(|\beta_k^{\text{OLS}}| - \lambda \left(\phi - (1 - \phi) \sum_{m \in S_k} h(r_{km}) \cdot |\beta_m^{\text{LD2lasso}}| \right) \right)_+}{1 + \lambda(1 - \phi) \sum_{m \in S_k} h(r_{km})}. \end{aligned} \quad (17)$$

For LD2lasso with analytic FDR control, the probability of a false positive analogous to (12) is

$$\begin{aligned} Pr\left(\frac{1}{N} |\mathbf{x}_k^T \mathbf{r}_{-k}| > \lambda \left(\phi - (1 - \phi) \sum_{m \in S_k} h(r_{km}) \cdot |\beta_m^{\text{LD2lasso}}| \right) \mid \beta_k = 0\right) \\ = 2\Phi\left(-\frac{\sqrt{N}\kappa}{\sigma_k}\right); \quad \kappa = \lambda \left(\phi - \lambda(1 - \phi) \sum_{m \in S_k} h(r_{km}) \cdot |\beta_m^{\text{LD2lasso}}| \right), \end{aligned} \quad (18)$$

where in contrast with (12) the threshold (κ) applied to β_k^{OLS} is now a function of the LD2lasso solutions of other SNPs in strong LD with the current SNP.

An alternative to (15) is a penalty that we refer to as the LD1lasso penalty and is essentially identical to a penalty term in the general and graphical fused Lasso (*e.g.*, Kim and Xing 2009):

$$\phi \sum_{k=1}^p \sum_{m \in S_k} h(r_{km}) \cdot \left| |\beta_k| - |\beta_m| \right|; \quad \phi \geq 0. \quad (19)$$

Computation of the LD2lasso solution (17) was implemented with a straightforward pathwise coordinate descent algorithm, because its objective function consists of a convex, a continuously differentiable, and a separable term (of convex functions) (Tseng 1988, 2001; Friedman *et al.* 2007; Chen *et al.* 2010). This does not apply to the LD1lasso, and hence

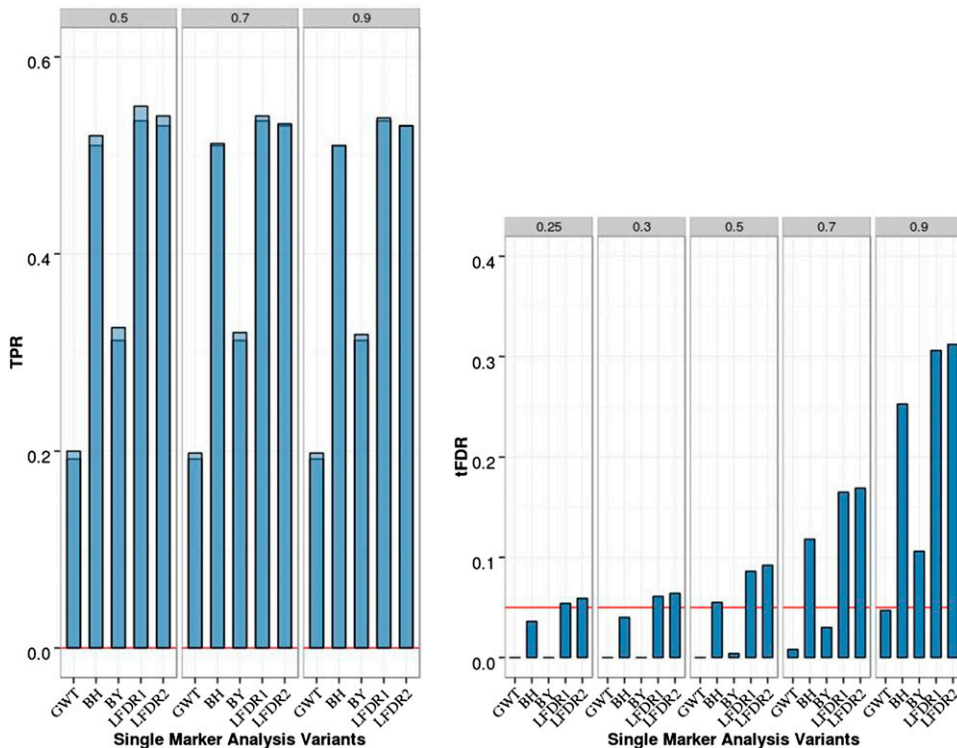


Figure 1 Single-marker analysis (SMA) of chromosome 21 data with two isolated QTL, 200 replicates, and sample size $N = 201$. Empirical power is shown [defined as true positive rate (TPR)] on the left: Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTL), and light blue represents TPR2 (TPR also including significant SNPs linked to QTL at absolute correlation threshold 0.5, 0.7, or 0.9). Empirical thresholded false discovery rate (tFDR) with absolute correlation thresholds 0.25, 0.3, 0.5, 0.7, and 0.9 is shown on the right: Red horizontal line represents the tFDR value of 0.05. GWMT, genome-wide threshold that represents the P -value threshold 5.5×10^{-8} ; BH, Benjamini–Hochberg; BY, Benjamini–Yekutieli; LFDR1 and LFDR2, local FDR-based thresholding using different estimators of the proportion of true null hypotheses (see main text for details).

we focused on the LD2lasso, which seemed to perform better (results not shown).

Consider a group of SNPs in strong LD and containing a single SNP that is causal for the phenotype. In this situation, the Lasso (and other penalties not including fusion-type penalties) will tend to select a single or few SNPs, while LD2lasso/LD1lasso would have nonzero solutions for several or all SNPs in this group. Hence for PR without fusion-type penalties, a QTL region for subsequent fine-mapping may be identified as the SNP(s) with nonzero solution plus other SNPs in LD with these SNP(s) above some threshold (e.g., $|r| > 0.5$). For LD2lasso/LD1lasso, a QTL region may be identified as including all (contiguous) SNPs with nonzero coefficients.

Analyses of multiple chromosomes

The analysis of multiple chromosomes may be regarded as a multigroup analysis, which we first discuss in the context of SMA. For a given biomedical trait of interest, many chromosomes may show no signal, while other chromosomes may show sparse signals or even extended stronger signals. To control the GWER such a grouping structure is irrelevant. However, for FDR control there is no unique way to proceed in the presence of a group structure. As argued by Efron (2008), the usual pooled FDR analysis, where the group structure is simply ignored, can be overly conservative or overly liberal for any particular group. Efron also established that a separate analysis, where FDR control (at the same level) is performed separately for each group and subsequently the results are simply combined, is valid in terms of achieving overall (*i.e.*, across groups) FDR control (at that

same level). Efron’s concern about joint FDR analysis of all groups (here chromosomes) within an experiment or study applies directly to SMA with FDR control in GWAS. While this also applies to PR with FDR control, here a separate analysis of each chromosome implies that QTL on other chromosomes are not included in the PR multimarker model, potentially lowering detection power and diminishing the postulated advantage of an all-markers over single-marker analysis. Finally, for most GWAS, one expects that the proportions of true null hypotheses are close to 1 and very similar across groups (chromosomes), so there may be little difference between joint and separate analyses and group-based analyses described below.

SMA: Besides these two basic strategies of pooled and separate analyses, some specialized methods that take the group structure into account have recently been proposed in the context of univariate analysis (SMA) (Cai and Sun 2009; Hu *et al.* 2010). Cai and Sun estimated the local FDR within each group and then applied the (approximate oracle) thresholding procedure to the combined set of local FDR values. For SMA of multiple chromosomes, we compared (i) pooled analysis based on the BH method or the local FDR thresholding procedure, (ii) separate analysis based on BH or the local FDR procedure, and (iii) the local FDR-based grouping procedure of Cai and Sun (2009).

Sun and Cai (2009) extended their earlier method to dependent test statistics, using a hidden Markov model (HMM). Wei *et al.* (2009) combined this method with the group-based method of Cai and Sun (2009) in the context of GWAS and implemented it in the R package PLIS. We tested

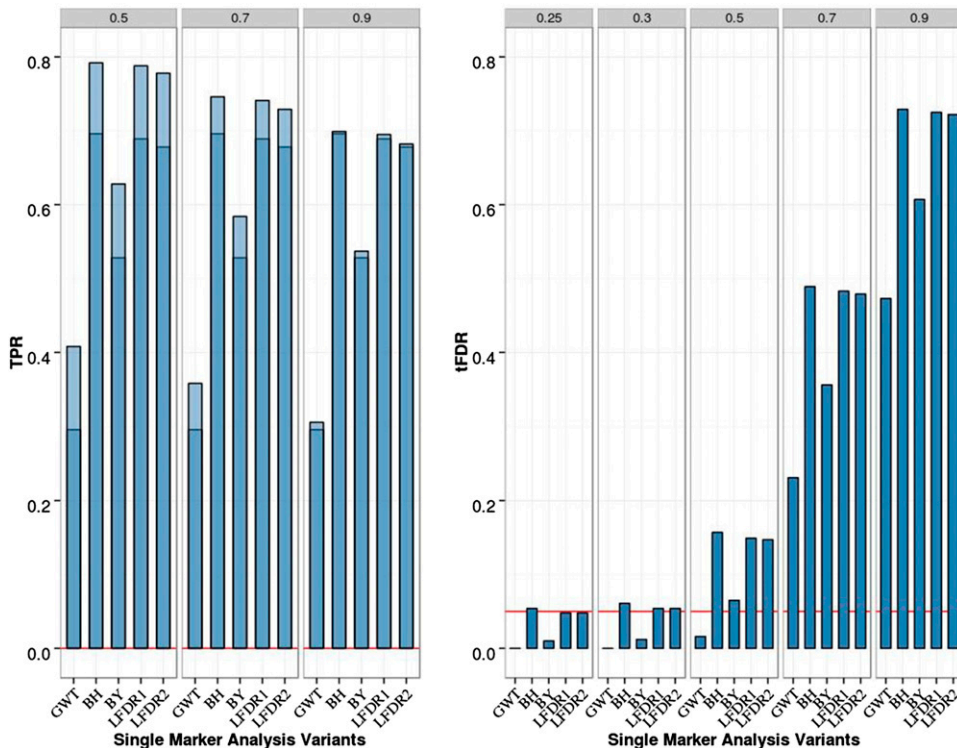


Figure 2 Single-marker analysis (SMA) of chromosome 21 data with eight QTL, 200 replicates, and sample size $N = 201$. Empirical power [defined as true positive rate (TPR)] is on the left: Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTL), and light blue represents TPR2 (TPR also including significant SNPs linked to QTL at absolute correlation threshold 0.5, 0.7, or 0.9). Empirical thresholded false discovery rate (tFDR) with absolute correlation thresholds 0.25, 0.3, 0.5, 0.7, and 0.9 is shown on the right: Red horizontal line represents the tFDR value of 0.05. GWTL, genome-wide threshold that represents the P -value threshold 5.5×10^{-8} ; BH, Benjamini-Hochberg; BY, Benjamini-Yekutieli; LFDR1 and LFDR2, local FDR-based thresholding using different estimators of the proportion of true null hypotheses (see main text for details).

this method on the single chromosome 21 simulation with eight true QTL, where it produced a very high number of false positives (irrespective of whether the alternative component was fitted with a mixture or a single normal distribution), and hence we did not further pursue the PLIS method.

Penalized regression: For PR with FDR control, we also compared separate and pooled analyses. For the pooled analysis, the SNPs from all three chromosomes were fitted jointly with the Lasso penalty, and the penalty parameter value was chosen based on controlling the FDR at level 0.05. For the separate analysis, the Lasso analysis was performed separately for each chromosome by omitting the SNPs on the other chromosomes and choosing a separate penalty parameter value to control the FDR at level 0.05 for each chromosome.

Comparison among methods

We define a causal true positive (CTP) as a QTL (defined earlier as a functional or causal SNP affecting the phenotype) that is significant (significant means identified by any of the methods considered). We define a linked true positive (LTP) as a SNP that is significant and has an absolute correlation with a QTL above a certain threshold (T) (note that T is a threshold on the absolute correlation coefficient, not the squared coefficient). We define a CLTP as being either a CTP or an LTP. Methods are compared, across the replicate simulations, based on (1) true positive rate (TPR) $1 = \text{number of CTPs}/\text{number of QTL}$, (2) $\text{TPR2} = \text{number of QTL detected by at least one CLTP (CTP or LTP)}/\text{number}$

of QTL, and (3) $\text{threshold-specific empirical tFDR} = \text{number of significant SNPs that are not a CLTP at a given } T/\text{number of significant SNPs}$. For criterion 2, we used $T = 0.9, 0.7$, and 0.5 for an increasingly relaxed definition of TP. For criterion 3, we used $T = 0.5, 0.3$, and 0.25 . The first value ($T = 0.5$) was chosen because our simulation was designed such that a QTL had a power of 0.5 to be detected and computed using Quanto for a given heritability, sample size, and chromosome-wide P -value threshold, and this power decreased to < 0.01 for a SNP correlated with the QTL at 0.5 (assuming that SNP and QTL have the same allele frequency). The second value ($T = 0.3$) is just below the threshold of “useful LD” ($r = 0.316$ or $r^2 = 0.1$) as defined by Kruglyak (1999) and Pritchard and Przeworski (2001) as the value at which the sample size is increased at most 10-fold. The last value ($T = 0.25$) is close to the threshold used in a previous comparison study (Ayers and Cordell 2010). For completeness, we also provide tFDR values at the remaining thresholds used for TPR2 (0.9, 0.7).

It is not straightforward to evaluate the empirical FDR in GWAS simulations due to LD. We experimented with alternative ways of defining the tFDR, such as partitioning a chromosome into 100-kb windows and computing tFDR (for a given T) as the number of windows in which the significant SNPs are all false positives divided by the number of windows that contain at least one significant SNP. Such an alternative approach did not change any results significantly, so we report tFDR for the first definition.

The results are presented in figures while more detailed results in the form of tables are available in [File S1](#); the tables include standard errors for all average (across

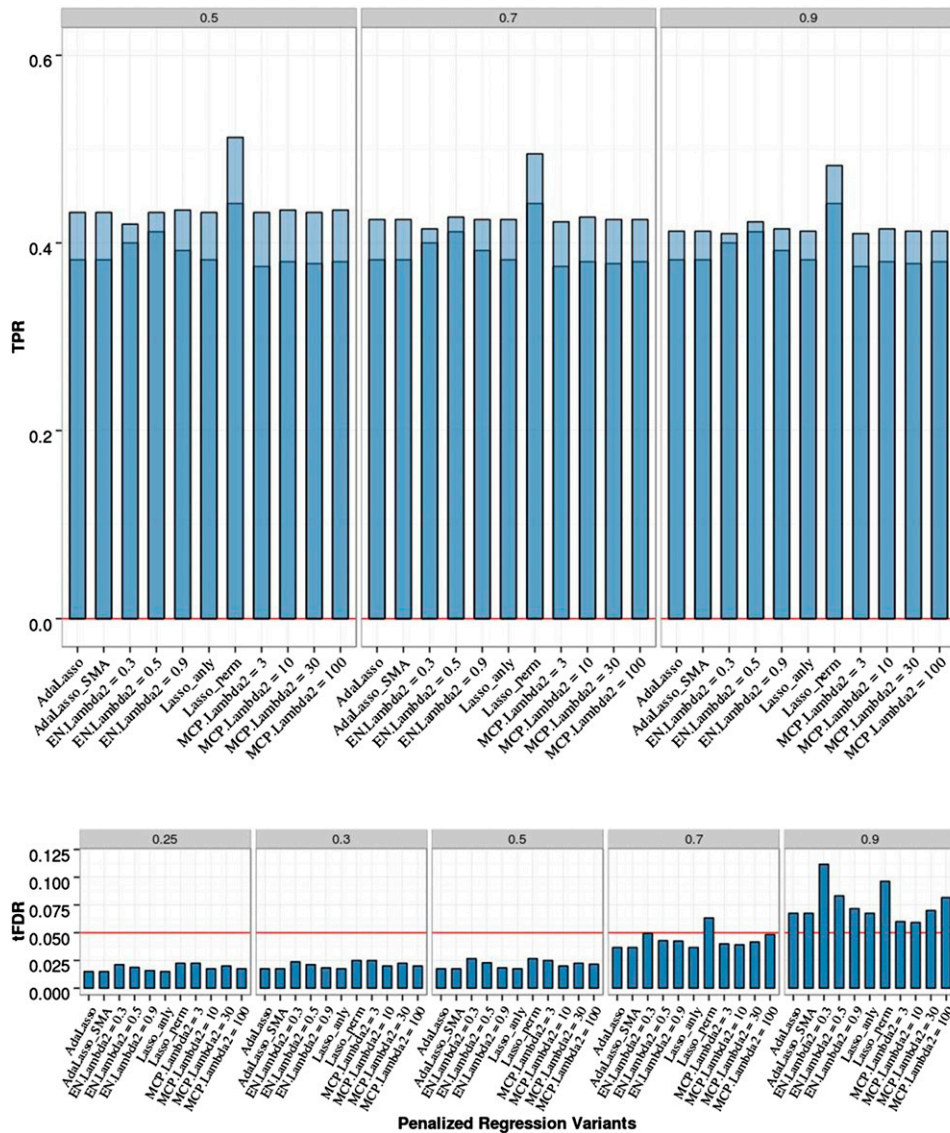


Figure 3 Penalized regression (PR) analyses of chromosome 21 data with two isolated QTL, 200 replicates, and sample size $N = 201$. Empirical power [defined as true positive rate (TPR)] is shown on the top: Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTL), and light blue represents TPR2 (TPR also including significant SNPs linked to QTL at absolute correlation threshold 0.5, 0.7, or 0.9). Empirical thresholded false discovery rate (tFDR) with absolute correlation thresholds 0.25, 0.3, 0.5, 0.7, and 0.9 is shown on the bottom: Red horizontal line represents the tFDR value of 0.05. The following penalized regression methods (in the order shown) are represented: AdaLasso, adaptive Lasso where all three initial estimators (SMA, Lasso with CV, and ridge regression with CV) produced the same result; EN_Lambda2 = x , elastic net with Lasso portion of $\lambda_2 = x$ ($x = 0.3, 0.5, 0.9$); Lasso_anly, Lasso with analytical FDR control; Lasso_perm, Lasso with permutation FDR control; MCP_lambda2 = x , minimax concave penalty with $\lambda_2 = x$ ($x = 3, 10, 30, 100$).

replicates) tFDR, TPR1, and TPR2 values that were used for computing P -values for differences in these values between methods or for a one-sided test of average tFDR to exceed level 0.05.

Results and Discussion

Single-chromosome simulation with two or eight QTL: Single-marker analysis

For two isolated QTL, SMA results are summarized in Figure 1 (and Table S1 in File S1). For SMA with GWER control [referred to as genome-wide threshold (GWT) in Figure 1] that controlled the FWER conservatively, the TPR was low (0.2) as expected. For SMA with FDR control, BH and BY controlled the FDR at level 0.05 based on two (three) of the three empirical threshold-specific FDRs (tFDRs), while the local FDR-based approximate oracle procedures LFDR1 and LFDR2 had tFDR values all > 0.05 although not significantly

(at nominal P -value of 0.05). The TPR was somewhat higher (0.31) for BY compared with the GWT, while it was substantially higher for BH, LFDR1, and LFDR2 (> 0.5).

For the eight-QTL scenario (Figure 2 and Table S2 in File S1), TPR was increased for the FDR-based SMA methods relative to the two-QTL case (≥ 0.7 for BH, LFDR1, and LFDR2; ~ 0.6 for BY compared with ≤ 0.4 for GWT). All FDR methods controlled the FDR in the sense of producing at least one tFDR value that was not significantly > 0.05 (nominal P -value). The smallest P -value for exceeding level 0.05 by a tFDR value at $T = 0.3$ was $P = 0.034$ for SMA-BH.

Single-chromosome simulation with two isolated QTL: PR with FDR control

We verified the expectation that Lasso with standard tuning parameter value selection based on CV produces an unacceptably high empirical FDR (> 0.5 ; results not shown). The results for PR with FDR control for this scenario are presented in Figure 3 and Table S3 and Table S4 in File S1.

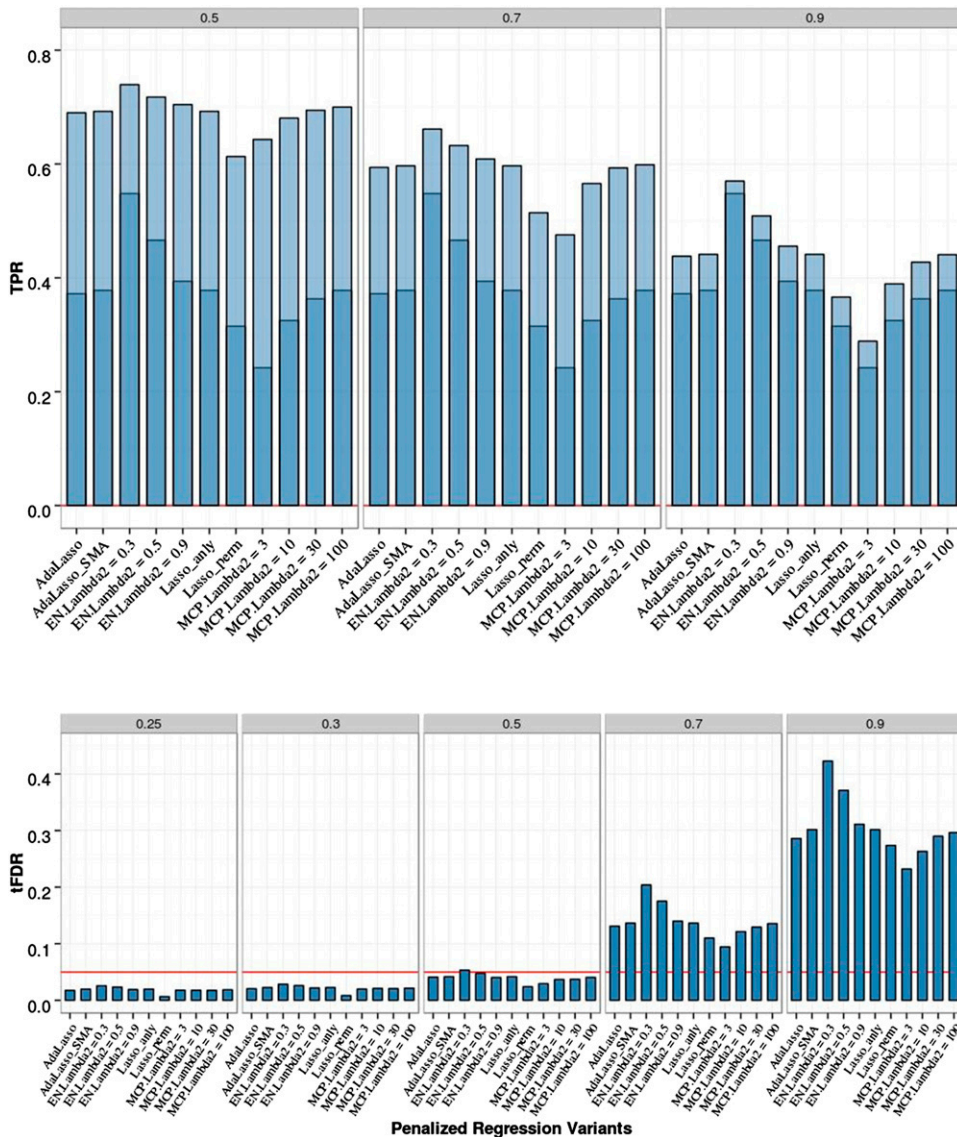


Figure 4 Penalized regression (PR) analyses of chromosome 21 data with eight QTL, 200 replicates, and sample size $N = 201$. Empirical power [defined as true positive rate (TPR)] is shown on the top: Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTL), and light blue represents TPR2 (TPR also including significant SNPs linked to QTL at absolute correlation threshold 0.5, 0.7, or 0.9). Empirical thresholded false discovery rate (tFDR) with absolute correlation thresholds 0.25, 0.3, 0.5, 0.7, and 0.9 is shown on the bottom: Red horizontal line represents the tFDR value of 0.05. The following penalized regression methods (in the order shown) are represented: Adalasso, adaptive Lasso where all three initial estimators (SMA, Lasso with CV, and ridge regression with CV) produced the same result; EN_Lambda2 = x , elastic net with Lasso portion of $\lambda_2 = x$ ($x = 0.3, 0.5, 0.9$); Lasso_anly, Lasso with analytical FDR control; Lasso_perm, Lasso with permutation FDR control; MCP_lambda2 = x , minimax concave penalty with $\lambda_2 = x$ ($x = 3, 10, 30, 100$).

Differences in empirical tFDR and TPR among all PR methods with analytic FDR control were small; there was a very slight advantage for the elastic net with the mixing parameter set to 0.5 (EN50); Lasso and MCP with different values for the second tuning parameter (from $\lambda_2 = 3$ to $\lambda_2 = 100$) produced almost identical results. All PR methods were able to control the FDR by producing tFDR values < 0.05 for all $T \leq 0.5$. In terms of power and using the EN50 to represent PR, in comparison with SMA-BH, the average TPR1 and TPR2 values of EN50 were lower with P -values for the differences ranging from 0.004 to 0.013. Comparing the EN50 with SMA-BY, the average TPR1 and TPR2 values of EN50 were higher with all P -values ~ 0.002 . The Lasso with permutation FDR control had slightly higher tFDR values compared with the Lasso with analytic FDR control; TPR1 and TPR2 of the permutation Lasso were similar to those of SMA-BH and higher than those of the Lasso with analytic FDR control (smallest P -value for the difference = 0.01). The results for Lasso with permutation FDR

were stable across different sets of permutations (100–400 permutations).

Single-chromosome simulation with eight QTL: PR with FDR control

The results for this scenario are presented in Figure 4 and Table S5 and Table S6 in File S1. All PR methods were able to control the FDR by producing tFDR values below (or not significantly above) 0.05 for all $T \leq 0.5$. Most of the differences in empirical tFDR and TPR among different PR methods with analytic FDR control remained small but some were more pronounced for the eight-QTL data relative to the two-QTL data: EN50 had a higher TPR1 and TPR2 than the Lasso, with a P -value of 2×10^{-5} for the difference in TPR1 and $P = 0.001$ for TPR2 at $T = 0.9$; MCP with low values for the second tuning parameter ($\lambda_2 = 3, 10$) had lower TPR1 (P -values 1.3×10^{-14} and 0.003, respectively) and TPR2 (P -values from 2×10^{-15} to 0.03 and from 0.003 to 0.33, respectively, with the upper end of each range

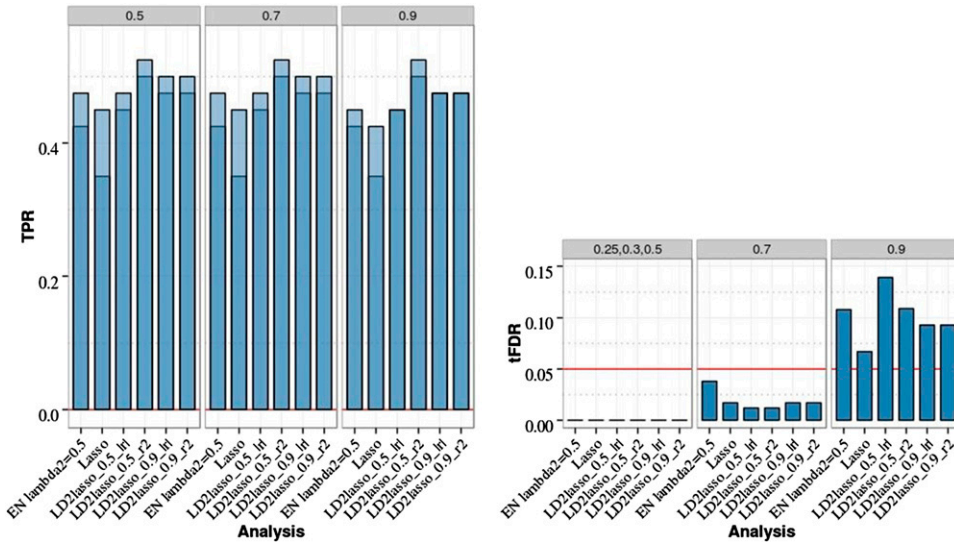


Figure 5 Empirical power [true positive rate (TPR)] and thresholded false discovery rate (tFDR) for LD2lasso vs. Lasso and elastic net (EN) penalized regression (PR) analyses of chromosome 21 data with two isolated QTL, 20 randomly selected replicates, and sample size $N = 201$. Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTL), and light blue represents TPR2 (TPR also including significant SNPs linked to QTL at absolute correlation threshold 0.5, 0.7, or 0.9). The empirical tFDR was computed at five absolute correlation thresholds between causal and linked SNPs ($T = 0.25, 0.3, 0.5, 0.7, 0.9$). The red horizontal line represents the tFDR value of 0.05. EN_Lambda2 = 0.5 denotes the elastic net with Lasso portion of $\lambda_2 = 0.5$; LD2lasso_0.9_r² (LD2lasso_0.9_lr) represents

the LD2lasso with weight on the Lasso penalty equal to $\phi = 0.9$ and LD function $h(r) = r^2$ ($h(r) = |r|$). The LD function was set to zero when the absolute correlation between two SNPs was $|r| < 0.85$.

representing TPR2 at $T = 0.5$) than the Lasso. Power was expectedly higher relative to the two-QTL scenario (for EN50, TPR2 at $T = 0.5$ increased from 0.43 to 0.73). Comparing the EN50 with SMA-BH, the average TPR1 and TPR2 values of EN50 were lower with P -values for the differences ranging from 5×10^{-22} to 0.001. Comparing the EN50 with SMA-BY, EN50 had a significantly higher value for TPR2 at $T = 0.5$ (P -value 0.0005) while EN50 had a lower value for TPR1 (P -value 0.005). In general, differences among PR methods were largest for TPR1 and TPR2 at $T = 0.9$ and were diminished for TPR2 with $T = 0.5$. Comparing Lasso with permutation vs. analytic FDR estimation, the permutation Lasso had lower TPR1 and TPR2 with P -values for the differences ≤ 0.0006 . It is surprising that the permutation method, which was the most powerful approach in the two-QTL scenario, is so conservative in the eight-QTL scenario. This appears to be due to the fact that permuting the phenotypes, while justified under the global null hypothesis of no association between any SNPs and the phenotypes, is conservative when evaluating the significance of a QTL in the presence of substantial signal from other QTL. By permuting the phenotype, which consists of both fixed signal and random error, the contribution of the random errors is overestimated. This does not affect the validity of the permutation approach in terms of properly controlling the FDR but does diminish power in cases like the eight-QTL scenario.

The eight-QTL simulation included two groups of two QTL each with within-group $r^2 \approx 0.5$ (absolute correlation among SNP doses of 0.71). Because of the Lasso's tendency to select only a single SNP per group and the EN's ability to deal with collinearity and potential coselection of correlated SNPs (SNPs in LD), we took a closer look at the behavior of Lasso and EN in this situation. Averaged over both groups and the 200 replicate simulations and including LTPs ($T = 0.5$), the Lasso

selected zero SNPs in 22%, one SNP in 43%, two SNPs in 26%, and more than two SNPs in 9% of cases. By comparison, the EN selected zero SNPs in 19%, one SNP in 37%, two SNPs in 31%, and more than two SNPs in 13% of cases. Expectedly the EN selected two and more than two SNPs more frequently but the difference was not substantial.

Single-chromosome simulation with two/eight QTL: Fusion-type penalties

Here we focused on the LD2lasso with ϕ (the relative weight given to the Lasso penalty) fixed at two different values (0.9, 0.5), with the two LD functions $h(r) = |r|$ and $h(r) = r^2$, and with $h(r) = 0$ if $|r| < 0.85$ [for this threshold, the median number of SNP “neighbors” was 2 with interquartile range (IQR) 0–5 and range 0–39] or $|r| < 0.50$ (median number of SNP neighbors = 14 with IQR 6–25 and range 0–102). Because we did not optimize the implementation of the LD2lasso for computational speed, we compared the LD2lasso with the Lasso and EN50 based on only 20 data replicates that were randomly selected from the 200 replicates. For the chromosome 21 data with two isolated QTL, the LD2lasso achieved higher power than Lasso and EN50 while maintaining FDR control (Figure 5 and Table S7 in File S1). The largest difference was the increase in TPR1 from 0.35 for the Lasso to 0.5 for the best LD2lasso analysis with $\phi = 0.5$ and $h(r) = r^2$ (P -value < 0.05), while all other differences in TPR1 and TPR2 were smaller.

For the chromosome 21 data with eight QTL (Figure 6 and Table S8 in File S1), however, all LD2lasso variants were unable to control the FDR with average tFDR values at the lowest threshold of $T = 0.25$ near 0.2 (P -value for the differences to FDR level $0.05 < 4 \times 10^{-4}$). We recall that the eight-QTL data contained a group of four QTL with weak pairwise LD ($0.01 \leq r^2 \leq 0.1$); the high tFDR values (0.2–0.26) for LD2lasso were mostly due to SNPs in LD with the

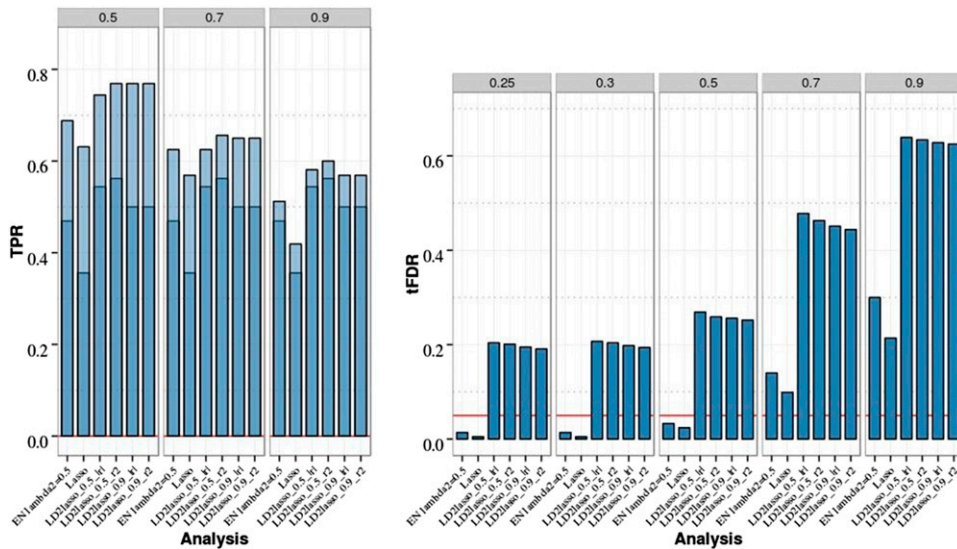


Figure 6 Empirical power [true positive rate (TPR)] and thresholded false discovery rate (tFDR) for LD2lasso vs. Lasso and elastic net (EN) penalized regression (PR) analyses of chromosome 21 data with eight QTL, 20 randomly selected replicates, and sample size $N = 201$. Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTL), and light blue represents TPR2 (TPR also including significant SNPs linked to QTL at absolute correlation threshold 0.5, 0.7, or 0.9). The empirical tFDR was computed at five absolute correlation thresholds between causal and linked SNPs ($T = 0.25, 0.3, 0.5, 0.7, 0.9$). The red horizontal line represents the tFDR value of 0.05. EN_Lambda2 = 0.5 denotes the elastic net with Lasso portion of $\lambda_2 = 0.5$; LD2lasso_0.9_r² (LD2lasso_0.9_rl) represents the LD2lasso with weight on the Lasso penalty equal to $\phi = 0.9$ and LD function $h(r) = r^2$ ($h(r) = |r|$). The LD function was set to zero when the absolute correlation between two SNPs was $|r| < 0.85$.

represents the LD2lasso with weight on the Lasso penalty equal to $\phi = 0.9$ and LD function $h(r) = r^2$ ($h(r) = |r|$). The LD function was set to zero when the absolute correlation between two SNPs was $|r| < 0.85$.

group of four causal SNPs at small absolute correlation values. In terms of TPR, the difference in TPR1 between the best LD2lasso analysis with $\phi = 0.5$, $h(r) = r^2$, and $|r| < 0.85$ ($|r| < 0.50$) and the Lasso had a P -value of 5.4×10^{-5} (1.7×10^{-5}), while the difference in TPR1 between the best LD2lasso and the EN50 had a larger P -value of 0.10 (0.008). TPR2 values were also higher, but most P -values for the differences in TPR2 between the same methods were > 0.05 due to smaller differences (less than half) and the small number of replicates.

Single-chromosome simulation with eight QTL: PR vs. SMA and CAR

Because all previous comparisons focused on differences among methods in empirical TPR and tFDR for a single FDR cutoff of 0.05, here we present TPR1 vs. tFDR at $T = 0.25$ and $T = 0.50$ (Figure 7) and TPR2 ($T = 0.50$) vs. tFDR at $T = 0.25$ and $T = 0.50$ (Figure 8). Figure 7 and Figure 8 were generated by varying the P -value cutoff for SMA and CAR and varying the penalty parameter for PR. In terms of power to detect causal SNPs (TPR1) and when using the most relaxed definition of tFDR (with $T = 0.25$), SMA dominated all other methods with CAR being second followed by LD2lasso and EN50, with Lasso and MCP ($\lambda_2 = 10$) performing worst (Figure 7). When using a more stringent definition of false positives with tFDR at $T = 0.50$, SMA/CAR dominated in terms of TPR1 only for tFDR values > 0.05 .

In terms of power to detect QTL with causal or linked SNPs (TPR2 at $T = 0.5$) (Figure 8) and when using the most relaxed definition of tFDR (with $T = 0.25$), EN50 and SMA (the latter for higher tFDR ($T = 0.25$) values) slightly dominated the other methods, with the LD2lasso now performing worst for the most relevant range of tFDR ($T = 0.25$) from 0.01 to 0.10. When replacing tFDR ($T = 0.25$) with

tFDR ($T = 0.50$), EN50 performed best (slightly better than Lasso and MCP), with SMA, CAR, and LD2lasso now separated as the worst performers.

Single-chromosome simulation with two/eight QTL: Previous approaches

The multisplit analysis of Meinshausen *et al.* (2009) was evaluated based on a random subset of 20 replicates of the chromosome 21 two-QTL data and on another random subset of 20 replicates of the chromosome 21 eight-QTL data (Table S9 in File S1). It controlled the FDR quite conservatively (all tFDR values equal to zero) and had low power relative to the single-step PR methods with FDR control, as expected. Power (TPR) was < 0.18 for the two-QTL data and < 0.21 for the eight-QTL data. P -values for the differences in TPR between the multisplit analysis and the Lasso ranged from 0.01 to 0.04 for the two-QTL data and from 2.4×10^{-7} to 4×10^{-4} for the eight-QTL data.

The adaptive Lasso with local FDR control of Sampson *et al.* (2013) was evaluated using all 200 replicates of the chromosome 21 two-QTL and eight-QTL data (Table S10 in File S1). R code for executing this method was generously provided to us by the first author (H. Yi). For our data, the local FDR adaptive Lasso controlled the FDR at the local FDR threshold of 0.1 (tFDR values < 0.05), but its power (TPR) was much lower than that of the Lasso for both the two-QTL and the eight-QTL data (P -values for the differences in TPR to the Lasso ranged from 1×10^{-12} to 2×10^{-13} and from 2.4×10^{-67} to 2.5×10^{-83} , respectively). When raising the local FDR threshold to 0.5, the FDR was not controlled for the two-QTL data (tFDR values highly significantly > 0.05) but was controlled for the eight-QTL data (tFDR values < 0.05) where the TPR values, however, remained significantly below the values of the Lasso.

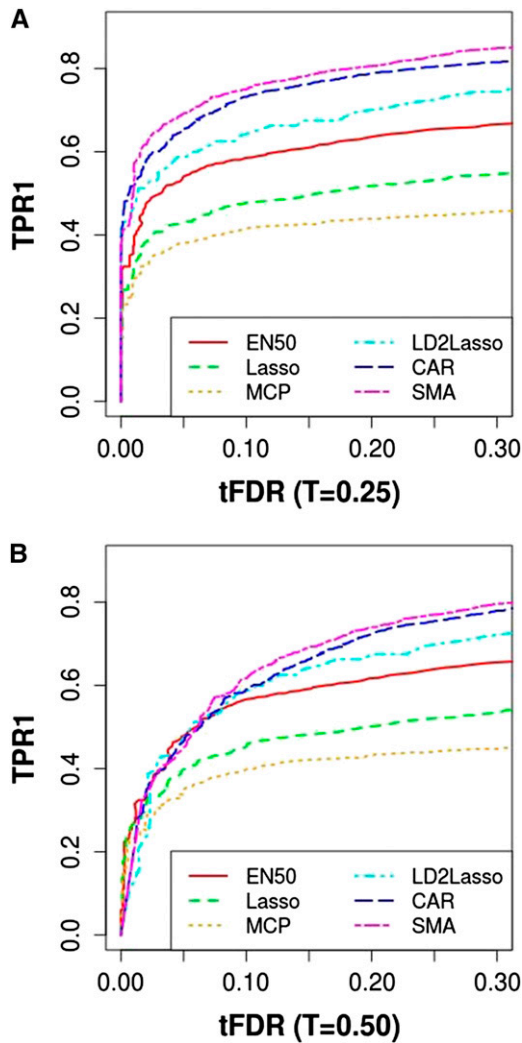


Figure 7 Plot of true positive rate TPR1 vs. thresholded false discovery rate (tFDR): TPR1 vs. tFDR ($T = 0.25$) and TPR1 vs. tFDR ($T = 0.50$). TPR1 and tFDR are defined in the main text. MCP was run with $\lambda_2 = 10$. CAR represents correlation-adjusted marginal correlations.

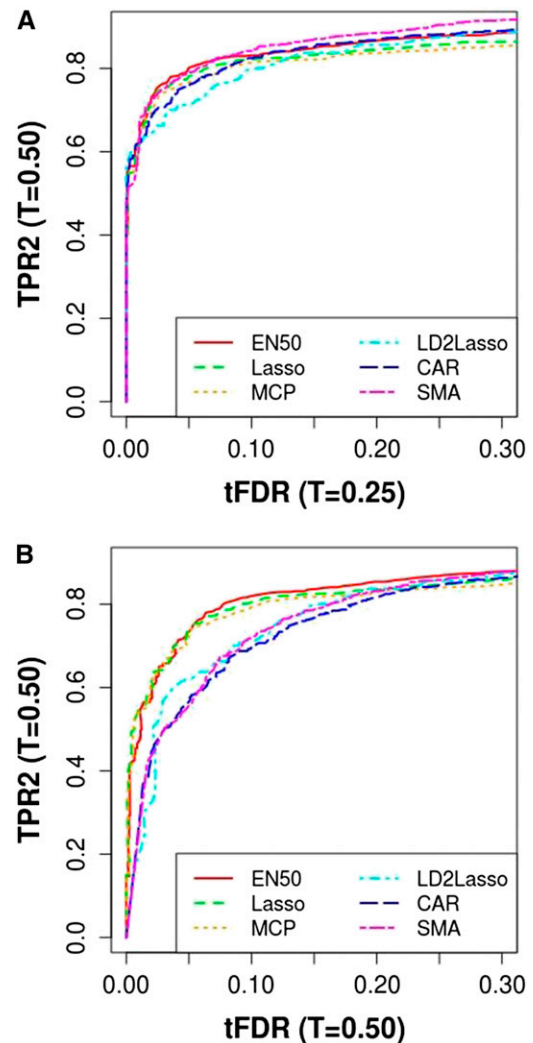


Figure 8 Plot of true positive rate TPR2 vs. thresholded false discovery rate (tFDR): TPR2 ($T = 0.50$) vs. tFDR ($T = 0.25$) and TPR2 ($T = 0.50$) vs. tFDR ($T = 0.50$). TPR2 and tFDR are defined in the main text. MCP was run with $\lambda_2 = 10$. CAR represents correlation-adjusted marginal correlations.

PUMA was run on the eight-QTL data with 200 replicates, choosing the 2D-MCP method, omitting the SNP preselection step and otherwise using default parameter values. The results are in Table S11 in File S1 for four different P -value thresholds for SNP selection ranging from 1×10^{-7} to 0.05, and these results need to be compared with our MCP with analytic FDR control [and four fixed values of the second tuning parameter λ_2 in (9)] in Table S6. The authors (Hoffman *et al.* 2013) recommended a P -value threshold of 1×10^{-7} for 2D-MCP, which controlled the FDR in terms of the tFDR values in Table S11; the next lower threshold of 1×10^{-6} had tFDR values >0.06 whose one-sided test for exceeding level 0.05 had P -values from 0.055 to 0.11. Overall, TPR1 and TPR2 values for 2D-MCP were significantly below those for MCP with analytic FDR control (P -values for differences from 9×10^{-5} to 8×10^{-29} for 2D-MCP SNP selection threshold 1×10^{-7} and from 0.032 to 2×10^{-22} for 2D-MCP SNP selection threshold

1×10^{-6}). To explain these differences, we looked at the values for λ_2 for the models chosen in PUMA by AIC over the 200 data sets, which surprisingly had a median of 3.46, with almost all values well below 10. For our own MCP analysis with FDR control we had chosen four values for the second tuning parameter (3, 10, 30, and 100), of which the lowest value performed the worst. We then replaced the heuristic AIC model selection followed by SNP selection of PUMA with our analytic FDR tuning parameter selection, which significantly improved TPR1 and TPR2 (last column of Table S11). The analytic FDR criterion selected models with larger values for λ_2 with a median of 33.6 (as well as larger values for λ_1 , median 0.166 vs. 0.310) This improvement still underperformed the results in Table S6 because PUMA uses a (internally determined) grid of values for λ_1 and λ_2 in 2D-MCP, which was too sparse for optimal performance of the analytic FDR approach.

Table 1 Analysis of the real data with single-marker analysis, using the genome-wide threshold P -value $< 5.5 \times 10^{-8}$ (SMA-GWT) or Benjamini–Hochberg FDR control (SMA-BH) applied separately to the data on each chromosome and elastic net penalized regression with analytic FDR control applied separately to each chromosome (EN50s) or jointly across all five chromosomes (EN50j)

Analysis Method	Significant SNPs	Chromosome				
		1	3	4	6	19
SMA-GWT	No. sig	74	0	8	0	0
	\cap EN50s	9/17/42/61	0/0/0/0	6/8/8/8	0/0/0/0	0/0/0/0
SMA-BH	No. sig	161	30	10	0	3
	\cap EN50s	13/23/49/101	5/30/30/30	6/9/9/9	0/0/0/0	1/1/3/3
EN50j	No. sig	14	3	3	0	0
	\cap EN50s	14/14/14/14	2/2/2/2	2/2/2/2	0/0/0/0	0/0/0/0
EN50s	No. sig	15	5	8	0	1
	\cap SMA-GWT	9/9/9/9	0/0/0/0	6/6/6/6	0/0/0/0	0/0/0/0
	\cap SMA-BH	13/13/13/14	5/5/5/5	6/6/6/6	0/0/0/0	1/1/1/1

No. sig denotes the number of significant SNPs, and \cap denotes overlap with. The four numbers for overlap (e.g., of SMA-GWT with EN50s in row 3) represent the number of significant SMA-GWT SNPs that are identical to a significant EN50s SNP / correlated ≥ 0.8 with a significant EN50s SNP / correlated ≥ 0.5 with a significant EN50s SNP / correlated ≥ 0.25 with a significant EN50s SNP.

Analyses of multiple chromosomes

Here we analyzed the simulated data including chromosomes 19 (no QTL), 21 (eight QTL), and 22 (two QTL) with 200 replicates. For SMA, the average tFDR values for pooled and separate BH analyses and for the local FDR grouping procedure were all > 0.05 for $T = 0.25$ and 0.3 but most were insignificantly higher (largest P -value of 0.001 for separate BH). Differences in TPR between SMA-based pooled and separate analyses and the grouping procedure of Cai and Sun (2009) were small and insignificant, although there was a trend for increasing TPR from pooled analysis to separate analysis to the grouping procedure (see Table S12 in File S1).

For penalized regression using Lasso PR with analytic FDR control, pooled and separate analyses controlled the FDR in terms of their average tFDR values, and differences in TPR between the pooled and the separate Lasso were small and insignificant (Table S13 in File S1). We therefore compared the separate and pooled analyses for individual chromosomes. From Table S13, we can see that for chromosome 21 (with eight QTL), TPR1 and TPR2 decreased somewhat from separate to pooled analysis with P -values for the differences ranging from 0.02 to 0.05 . For chromosome 22 (with only two QTL), in contrast, TPR1 and TPR2 increased significantly as expected from separate to pooled analysis with all P -values $< 4 \times 10^{-10}$.

Analysis of real data

We applied PR and SMA methods to data from the Health ABC GWAS of interleukin 6 soluble receptor (IL-6 SR). The analysis included 786 Health ABC Caucasians who had both GWAS and serum IL-6 SR measurements (a continuous phenotype, approximately normally distributed) available (see File S2 for details). The genotype data included 750,424 SNPs (after standard edits) from chromosomes 1, 3, 4, 6, and 19. Covariates included in the analysis models were age, gender, site of data collection, and one principal component score obtained using Eigenstrat. We analyzed

these data by SMA with GWER threshold (SMA-GWT), by SMA with BH FDR control (SMA-BH), and with the elastic net with 50% Lasso penalty (EN50) and analytic FDR control. SMA-GWT is commonly used in practice, and SMA-BH performed best among all SMA methods with FDR control on the simulated data. The elastic net was chosen as the single PR method because it performed best on the simulated data among all PR methods combined with our novel method for FDR control and because differences among PR methods were generally small. For SMA-BH and EN50 (EN50-sep), chromosomes were analyzed separately, and for EN50, a joint analysis of all chromosomes was also performed (EN50-joint).

Chromosome 1 represents a special situation in that it contains a region with an extremely strong QTL signal. The very stringent SMA-GWT threshold identified 74 SNPs for this chromosome, many more than the 15 EN50-sep SNPs (see Table 1). Of these 74 SNPs, 61 were in LD above the absolute correlation threshold of 0.25 with EN50-sep SNPs, and the remaining 13 SNPs had largest absolute correlations with EN50-sep SNPs between 0.20 and 0.24 and therefore likely still overlap with EN50-sep SNPs (given the very strong signal) rather than represent independent signals. The EN50-sep selected a small subset of the peak area SNPs that had the smallest SMA P -values (Figure 9). While the SMA-GWT SNPs were all located in the area of the major peak, several EN50-sep SNPs were located in regions clearly separated from the peak, which mostly overlapped with SNPs detected by SMA-BH except for one EN50-sep SNP (see Figure 9). Expectedly, SMA-BH identified the largest number of SNPs (161), of which 101 were in LD above the absolute correlation threshold of 0.25 with EN50-sep SNPs. The remaining 60 SMA-BH SNPs had largest absolute correlations with EN50-sep SNPs between 0.07 and 0.24 , indicating together with Figure 9 that SMA-BH selected some SNPs physically well separated from and likely not in LD with any EN50 SNPs. To check on the interpretation of the SNP–SNP correlation values, we simulated independent SNPs for $N = 786$ representing 11 different MAF values between 0.01 and 0.5 (assuming Hardy–Weinberg

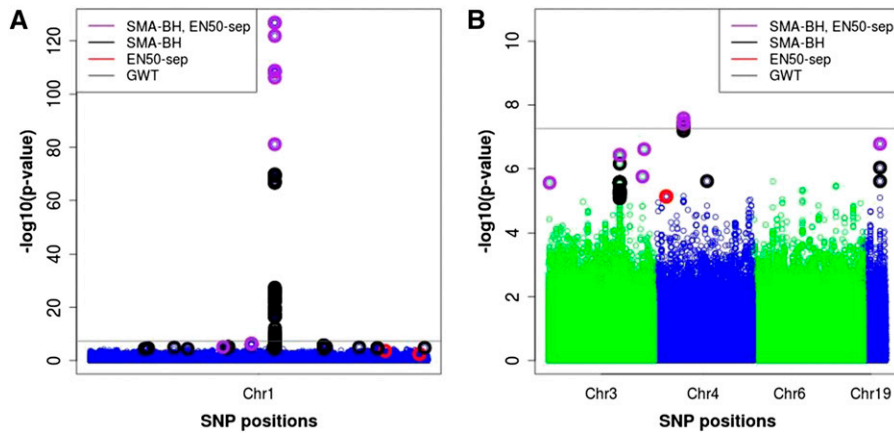


Figure 9 Plot of $-\log_{10}(P\text{-value})$ for SNPs on chromosomes 1 (left), 3, 4, 6, and 19 (right) for the real data. The purple, black, and red circles represent significant SNPs selected by both SMA with BH FDR control (SMA-BH) and EN50 (elastic net with 50% weight on L1), by SMA-BH only, and by EN50 only, respectively. EN50-sep denotes EN50 applied to each chromosome separately. The gray line is the genome-wide threshold of 5.5×10^{-8} , above which SNPs were selected by SMA-GWT. The y-axis represents raw P -values from SMA.

equilibrium) and computed all pairwise absolute correlations over 100,000 replications, yielding 75th, 95th, and 99th percentiles just below 0.025, 0.07, and 0.13, respectively, with a maximum value of 0.31. Based on our simulation studies (in particular the eight-QTL simulation), SMA-BH may have higher power than EN50-sep (by $\sim 10\%$) but also may not completely control the FDR, so the 60 SMA-BH SNPs with very small correlations to EN50-sep SNPs may represent some additional true signals and some false positives. We also reran EN50-sep on chromosome 1 by excluding all peak area SNPs but including a single peak area SNP (with the smallest P -value) in the model without shrinkage, but this analysis did not identify any additional SNPs (such as any of the 60 SMA-BH SNPs) compared to the previous EN50-sep analysis.

On chromosome 4, SMA-GWT, SMA-BH, and EN50-sep identified similar numbers of SNPs, 8 for SMA-GWT (all correlated with EN50 SNPs), 10 for SMA-BH (all but 1 correlated with EN50 SNPs), and 8 for EN50 (6 of which were correlated with SMA-GWT or SMA-BH SNPs).

On chromosomes 3 and 19, no QTL were identified with the stringent SMA-GWT, while SMA-BH and EN50-sep identified the same QTL (Table 1 and Figure 9). For chromosome 3, the 30 SMA-BH SNPs were all in LD (absolute correlation threshold of 0.25) with the 5 EN50-sep SNPs, and for chromosome 19, the 3 SMA-BH SNPs were all in LD with the 1 EN50-sep SNP. No SNPs were identified by any method on chromosome 6. Overall, these results show that SMA-BH and EN50-sep identify very similar candidate regions in a genome scan; that EN50 (and other PR methods) will select a subset of SNPs in a cluster of correlated SNPs, here those with the smallest SMA P -values; and that for follow-up (fine-mapping) studies, all EN50 (or other PR) identified SNPs plus a set of SNPs correlated with the identified SNPs (above a certain threshold) should be considered.

Finally, the joint analysis of all chromosomes by EN50 with FDR control (EN50-joint), compared to the separated analyses (EN50-sep), identified fewer SNPs (20 SNPs for EN50-joint compared with 29 SNPs for EN50-sep). EN50-joint identified one SNP on chromosome 3 and another on chromosome 4 that were not in LD with EN50-sep SNPs. The

EN50-joint SNP on chromosome 3 is likely a false positive due to its $-\log_{10}(P\text{-value})$ being much smaller than those of all other identified (by any method) SNPs (Figure 10). The EN50-joint SNP on chromosome 4, however, was in LD (absolute correlation > 0.9) with a SMA-BH SNP.

Concluding remarks

The goals of this study were to provide a review and an independent comparison of penalized regression methods with single-marker analysis for variable (SNP) selection in GWAS and to implement and evaluate penalty/tuning parameter value selection by FDR control. A recent application of PR to GWAS (Waldmann *et al.* 2013, p. 1) arrived at this conclusion: “Hence, we can conclude that it is important to analyze GWAS data with both the lasso and the elastic net and *an alternative tuning criterion to minimum MSE is needed for variable selection.*” Here we provided such an alternative criterion.

We provided both a simple, approximate analytic method and a permutation-based method for FDR control in PR. The analytic method performed consistently well although being somewhat conservative. The permutation method expectedly had higher power (was less conservative) than the analytic method in the two-QTL scenario but was more conservative than the analytic method in the eight-QTL scenario. This initially unexpected behavior of the permutation method appears to be due to the fact that permuting the phenotypes means permuting some signal along with the random error and thus overestimates the amount of random error in the data. This causes the permutation method to become increasingly conservative with increasing numbers and effect sizes of causal SNPs (QTL). Because our simulated data had small sample sizes (~ 200) and contained QTL jointly explaining 20 or 48% of the phenotypic variance (for the two- or eight-QTL data, respectively), this problem of the permutation method is more pronounced in our simulation than in real GWAS data characterized by very large sample sizes (thousands to tens of thousands) and small SNP effects (explaining $\leq 1\%$ of the overall variation). However, the problem may manifest itself also in real data, when there are larger numbers of causal SNPs or causal SNPs with larger

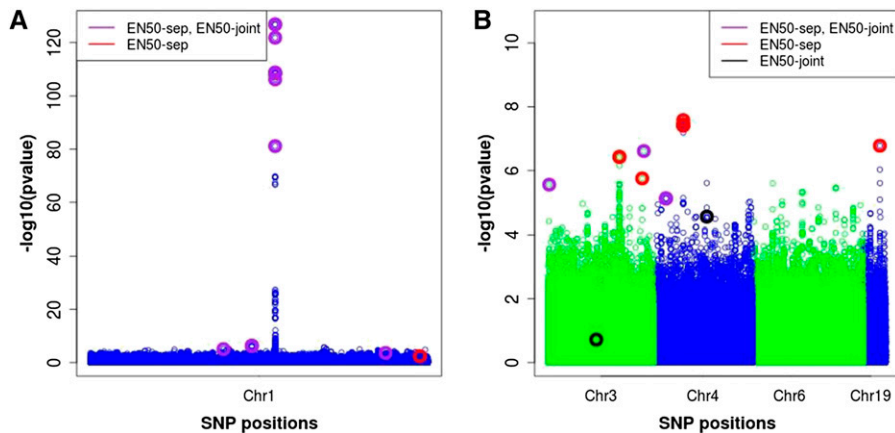


Figure 10 Plot of $-\log_{10}(P\text{-value})$ for SNPs on chromosomes 1 (left), 3, 4, 6, and 19 (right) for the real data. EN50-sep (EN50-joint) denotes the elastic net with 50% weight on L1 applied to each chromosome separately (to all chromosomes jointly). The purple, red, and black circles represent significant SNPs selected by both EN50-sep and EN50-joint, by EN50-sep only, and by EN50-joint only, respectively. The y-axis represents raw P -values from SMA.

effect sizes as in our Health ABC GWAS. This is unfortunate because the main advantage of multimarker over single-marker approaches is that they are able to use the signal from other markers in determining significance. It may be possible to overcome this limitation by estimating the residuals using the fitted model and then permuting the residuals instead of the phenotype; we are currently investigating the performance of this approach.

Based on our comparisons with previous strategies for SNP selection in PR, it appears that our analytic FDR criterion is currently the best approach to SNP selection when using PR for GWAS, and this approach should be appealing to practitioners who desire a measure of error, in particular in terms of FDR, associated with the selected SNPs. Finally we note that our simple analytic FDR method can also be applied to penalized logistic regression.

When the focus is on variable selection (here the identification of a causal SNP by itself or a linked SNP) and not on estimation or prediction, our results indicate that among all PR methods investigated here, a version of the elastic net (EN50) performs better than the other PR methods in most situations, although differences were often small. Based on our analyses of simulated and real data, the EN50 should identify QTL regions that are very similar to those identified by SMA combined with BH FDR control. In contrast with SMA-BH, EN50 identifies a QTL region with a single SNP or few SNPs, and hence subsequent fine-mapping should include the EN50 SNPs plus additional SNPs in LD with the EN50 SNPs above a certain threshold. Expectedly, EN50 and SMA-BH have substantially more power than SMA with the genome-wide type-I error threshold.

Incorporating fusion-type penalties developed for covariates measured on graphs may improve power but can also generate more false positives or rather wide QTL regions for follow-up, in situations when there are multiple, moderately correlated causal SNPs located on the same chromosome and false positives are defined as any significant SNPs correlated with a causal SNP below an absolute value of 0.25.

When applying PR to large genome-wide data sets for joint (pooled) analysis of all chromosomes, prescreening SNPs based on SMA P -values is still necessary due to the high memory requirements for holding the entire SNP genotype

matrix in memory (see Hoffman *et al.* 2013 for more details). We note that if prescreening is performed, then for tuning parameter value selection by our analytic FDR method the number of markers (p) in (13) should be set to the total number of markers (prior to preselection) to maintain control of the FDR (which should produce the same result as what would be obtained without preselection). It would, however, be prudent to perform PR with FDR control both on the entire set of SNPs from all chromosomes (pooled analysis) and separately on each chromosome (separate analysis). Our results from the simulated and real data representing several chromosomes indicate that the pooled analysis does not necessarily provide better power for all chromosomes.

We note that currently there is substantial interest in extending GWAS of single phenotypes to high-dimensional phenotypes (*e.g.*, Marttinen *et al.* 2012). While high-dimensional phenotypes allow aspects of modeling that are not feasible with a single phenotype, the results of the current study still provide useful information for the design of analysis methods for GWAS of high-dimensional phenotypes.

Finally, code used for data simulation and code for EN50 analysis with analytic FDR control are provided in File S3. An example simulated data set with eight QTL on chromosome 21 is available at the Dryad Digital Repository (<http://doi.org/10.5061/dryad.hc445>).

Acknowledgments

This work was supported by the National Institute of General Medical Sciences grant R01HG005254 (to I.H.), by the National Institute on Aging grant 1R01AG032098-01A1 and the National Heart, Lung, and Blood Institute grant R01HL101250 (to Y.L.), and by a Virginia Tech Genetics, Bioinformatics, and Computational Biology Ph.D. program fellowship (to H.Y.). The Health ABC study was supported by the National Institute on Aging contracts N01AG62101, N01AG62103, and N01AG62106. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract HHSN268200782096C.

Literature Cited

- Ahmed, I., A.-L. Hartikainen, M.-R. Järvelin, and S. Richardson, 2011 False discovery rate estimation for stability selection: application to genome-wide association studies, pp. 1–20 in *Statistical Applications in Genetics and Molecular Biology*.
- Akaike, H., 1974 A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19: 716–723.
- Akaike, H., 1977 On entropy maximization principle, pp. 27–41 in *Application of Statistics*, edited by P. R. Krishnaiah. North-Holland, Amsterdam, The Netherlands.
- Ayers, K. L., and H. J. Cordell, 2010 SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* 34: 879–891.
- Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57: 289–300.
- Benjamini, Y., and D. Yekutieli, 2001 The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29: 1165–1188.
- Bogdan, M., J. K. Ghosh, and R. Doerge, 2004 Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* 167: 989–999.
- Breheny, P., and J. Huang, 2011 Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* 5: 232.
- Breheny, P. J., 2009 Regularized methods for high-dimensional and bi-level variable selection. Ph.D. Thesis, University of Iowa, Iowa City, IA.
- Cai, T. T., and W. Sun, 2009 Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. *J. Am. Stat. Assoc.* 104: 1467–1481.
- Chen, J., and Z. Chen, 2008 Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 95: 759–771.
- Chen, X., S. Kim, Q. Lin, J. G. Carbonell, and E. P. Xing, 2010 Graph-structured multi-task regression and an efficient optimization method for general fused Lasso. arXiv:1005.3579.
- Dudbridge, F., and A. Gusnanto, 2008 Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* 32: 227–234.
- Efron, B., 2003 Robbins, empirical Bayes, and microarrays. *Annals of Statistics* 31: 366–378.
- Efron, B., 2008 Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Stat.* 2: 197–223.
- Efron, B., and R. Tibshirani, 2002 Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* 23: 70–86.
- Fan, J., and R. Li, 2001 Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 96: 1348–1360.
- Fan, J., and H. Peng, 2004 Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* 32: 928–961.
- Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani, 2007 Pathwise coordinate optimization. *Ann. Appl. Stat.* 1: 302–332.
- Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33: 1.
- Gauderman, W. J., and J. Morrison, 2001 QUANTO documentation. Technical report no. 157. Department of Preventive Medicine, University of Southern California, Los Angeles.
- Hoffman, G. E., B. A. Logsdon, and J. G. Mezey, 2013 PUMA: a unified framework for penalized multiple regression analysis of GWAS data. *PLoS Comput. Biol.* 9(6): e1003101.
- Hu, J. X., H. Zhao, and H. H. Zhou, 2010 False discovery rate control with groups. *J. Am. Stat. Assoc.* 105: 1215–1227.
- International HapMap Consortium, 2005 A haplotype map of the human genome. *Nature* 437: 1299–1320.
- International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861.
- Jin, J., and T. T. Cai, 2007 Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Am. Stat. Assoc.* 102: 495–506.
- Kim, S., and E. P. Xing, 2009 Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet.* 5: e1000587.
- Kim, Y., H. Choi, and H.-S. Oh, 2008 Smoothly clipped absolute deviation on high dimensions. *J. Am. Stat. Assoc.* 103: 1665–1673.
- Kohavi, R., 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, Vol. 2, pp. 1137–1145.
- Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22: 139–144.
- Li, C., and H. Li, 2008 Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 24: 1175–1182.
- Li, C., and H. Li, 2010 Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.* 4: 1498–1516.
- Liu, J., K. Wang, S. Ma, and J. Huang, 2011 Accounting for linkage disequilibrium in genome-wide association studies: a penalized regression method. Technical report. Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA.
- Marttinen, P., J. Gillberg, A. Havulinna, J. Corander, and S. Kaski, 2012 Genome-wide association studies with high-dimensional phenotypes. *Stat. Appl. Genet. Mol. Biol.* 12: 413–431.
- Meinshausen, N., and P. Bühlmann, 2010 Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 72: 417–473.
- Meinshausen, N., L. Meier, and P. Bühlmann, 2009 P-values for high-dimensional regression. *J. Am. Stat. Assoc.* 104: 1671–1681.
- Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69: 1–14.
- Sabatti, C., and N. Freimer, 2003 False discovery rate in linkage and association genome screens for complex disorders. *Genetics* 164: 829–833.
- Sampson, J. N., N. Chatterjee, R. J. Carroll, and S. Müller, 2013 Controlling the local false discovery rate in the adaptive Lasso. *Biostatistics* 14: 653–666.
- Schwarz, G., 1978 Estimating the dimension of a model. *Ann. Stat.* 6: 461–464.
- Storey, J. D., J. M. Akey, and L. Kruglyak, 2005 Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.* 3: e267.
- Strimmer, K., 2008 A unified approach to false discovery rate estimation. *BMC Bioinformatics* 9: 303.
- Su, Z., J. Marchini, and P. Donnelly, 2011 HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* 27: 2304–2305.
- Sun, W., and T. T. Cai, 2007 Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc.* 102: 901–912.
- Sun, W., and T. T. Cai, 2009 Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 71: 393–424.
- Sun, W., J. G. Ibrahim, and F. Zou, 2010 Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics* 185: 349–359.
- Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58: 267–288.

- Tseng, P., 1988 *Coordinate Ascent for Maximizing Nondifferentiable Concave Functions*. Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, Cambridge, MA.
- Tseng, P., 2001 Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* 109: 475–494.
- Waldmann, P., G. Mészáros, B. Gredler, C. Fuerst, and J. Sölkner, 2013 Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.* 4: 270.
- Wasserman, L., and K. Roeder, 2009 High dimensional variable selection. *Ann. Stat.* 37: 2178.
- Wei, Z., W. Sun, K. Wang, and H. Hakonarson, 2009 Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics* 25: 2802–2808.
- Ye, J., 1998 On measuring and correcting the effects of data mining and model selection. *J. Am. Stat. Assoc.* 93: 120–131.
- Zhang, C.-H., 2010 Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* 38: 894–942.
- Zou, H., 2006 The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101: 1418–1429.
- Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67: 301–320.
- Zou, H., T. Hastie, and R. Tibshirani, 2007 On the “degrees of freedom” of the lasso. *Ann. Stat.* 35: 2173–2192.
- Zuber, V., A. P. D. Silva, and K. Strimmer, 2012 A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies. *BMC Bioinformatics* 13: 284.

Communicating editor: F. Zou

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167817/-/DC1>

Penalized Multimarker *vs.* Single-Marker Regression Methods for Genome-Wide Association Studies of Quantitative Traits

Hui Yi, Patrick Breheny, Netsanet Imam, Yongmei Liu, and Ina Hoeschele

File S1 Supplement S1

Table S1 Single Marker Analysis (SMA) of 200 replicates of chromosome 21 data with two isolated QTLs, and with sample size N = 201. GWT, CWT, BH, BY and LFDR are different SNP selection criteria for SMA. GWT is Genome-Wide Threshold which represents the p-value threshold 5.5×10^{-8} ; CWT is Chromosome-Wide Threshold which represents the p-value threshold $5.5 \times 10^{-8} \times 3849034/50165$; BH denotes Benjamini-Hochberg FDR control; BY represents Benjamini-Yekutieli FDR control; LFDR1 and LFDR2 represent the approximate oracle procedure of Sun and Cai (2007) based on Efron's Local FDR with two different sets of parameter estimates (see main text). For the FDR methods, a cut-off value of 0.05 was used. Comparison criteria are FWER, tFDR, TPR1, and TPR2. FWER is the empirical family-wise error rate (proportion of replicates with ≥ 1 false positives); tFDR is an empirical thresholded false discovery rate mean(n_{FP}/n_{Sig}) with n_{FP} (n_{Sig}) being the number of false positives (significances), where a false positive is a SNP which is not in LD above a threshold T with any QTL (causal SNP); TPR1 = n_{CTP}/Q is the first true positive rate with n_{CTP} denoting the number of true QTLs identified and Q denoting the number of true QTL (2×200) over all replicates; TPR2 = n_{CLTP}/Q is the second true positive rate with n_{CLTP} denoting the number of true QTLs identified with the causal SNP or a linked SNP according to threshold T; T is a threshold, on the absolute correlation between the allelic doses of the causal and a linked SNP, above which a linked SNP is counted as a true positive.

Criterion	T	GWT	BH	BY	LFDR1 LFDR2
tFWER	0.25	0	n.a.	n.a.	n.a.
	0.3	0	n.a.	n.a.	n.a.
	0.5	0	n.a.	n.a.	n.a.
tFDR	0.25	0	0.036 ± 0.008	0	0.054 ± 0.010 0.059 ± 0.012
	0.3	0	0.040 ± 0.009	0	0.061 ± 0.010 0.064 ± 0.012
	0.5	0	0.055 ± 0.010	0.004 ± 0.003	0.086 ± 0.012 0.092 ± 0.014
TPR1		0.192 ± 0.020	0.510 ± 0.026	0.312 ± 0.025	0.535 ± 0.027 0.530 ± 0.027
TPR2	0.5	0.200 ± 0.021	0.520 ± 0.026	0.325 ± 0.025	0.550 ± 0.026 0.540 ± 0.027
	0.7	0.198 ± 0.020	0.512 ± 0.026	0.320 ± 0.025	0.540 ± 0.027 0.532 ± 0.027
	0.9	0.198 ± 0.020	0.510 ± 0.026	0.318 ± 0.025	0.538 ± 0.027 0.530 ± 0.027

Table S2 SMA of 200 replicates of chromosome 21 data with eight QTLs, and with sample size N = 201. See Table S1 for abbreviations and details.

Criterion	T	GWT	BH	BY	LFDR1 LFDR2
tFWER	0.25	0	n.a	n.a	n.a
	0.3	0	n.a	n.a	n.a
	0.5	0.135 + 0.024	n.a	n.a	n.a
tFDR	0.25	0	0.054 ± 0.006	0.010 ± 0.005	0.048 ± 0.006 0.048 ± 0.006
	0.3	0	0.061 ± 0.006	0.012 ± 0.005	0.054 ± 0.006 0.054 ± 0.006
	0.5	0.016 ± 0.004	0.157 ± 0.008	0.065 ± 0.008	0.149 ± 0.008 0.147 ± 0.009
TPR1		0.296 ± 0.012	0.696 ± 0.016	0.528 ± 0.018	0.689 ± 0.016 0.678 ± 0.017
TPR2	0.5	0.408 ± 0.015	0.792 ± 0.016	0.628 ± 0.020	0.788 ± 0.016 0.778 ± 0.017
	0.7	0.358 ± 0.015	0.746 ± 0.016	0.584 ± 0.019	0.741 ± 0.016 0.729 ± 0.017
	0.9	0.306 ± 0.013	0.699 ± 0.016	0.537 ± 0.018	0.695 ± 0.016 0.682 ± 0.017

Table S3 Penalized regression analysis using Lasso, Adaptive Lasso and Elastic Net, with FDR based selection of the penalty parameter values, of 200 replicates of chromosome 21 data with two isolated QTLs, and with sample size N = 201. FDR control was at the 0.05 level, and all methods used the analytic FDR except Lasso perm which used the permutation FDR. EN represents the Elastic Net with lasso weight (λ_2) set to 0.3, 0.5 or 0.9. AdaLasso represents the Adaptive Lasso using weights obtained by Lasso with CV (CV), Ridge Regression with CV (RR), and SMA (results for these three AdaLasso varieties were identical here). See Table S1 for other definitions.

Criterion	T	Lasso perm	Lasso	AdaLasso (CV,RR, SMA)	EN ($\lambda_2 = 0.3$)	EN ($\lambda_2 = 0.5$)	EN ($\lambda_2 = 0.9$)
tFDR	0.25	0.022 ± 0.01	0.015 ± 0.008	0.015 ± 0.008	0.021 ± 0.010	0.019 ± 0.009	0.016 ± 0.008
	0.3	0.025 ± 0.01	0.018 ± 0.008	0.018 ± 0.008	0.024 ± 0.010	0.021 ± 0.010	0.018 ± 0.008
	0.5	0.027 ± 0.010	0.018 ± 0.008	0.018 ± 0.008	0.027 ± 0.010	0.023 ± 0.010	0.018 ± 0.008
TPR1		0.442 ± 0.024	0.382 ± 0.025	0.382 ± 0.025	0.400 ± 0.025	0.412 ± 0.026	0.392 ± 0.026
TPR2	0.5	0.512 ± 0.024	0.432 ± 0.027	0.432 ± 0.027	0.420 ± 0.026	0.432 ± 0.027	0.435 ± 0.027
	0.7	0.495 ± 0.024	0.425 ± 0.027	0.425 ± 0.027	0.415 ± 0.026	0.428 ± 0.027	0.425 ± 0.027
	0.9	0.482 ± 0.025	0.412 ± 0.027	0.412 ± 0.027	0.410 ± 0.026	0.422 ± 0.026	0.415 ± 0.026

Table S4 Penalized regression analysis using MCP with several fixed values of the second tuning parameter (λ_2) and with FDR based selection of the value for the first tuning parameter, of 200 replicates of chromosome 21 data with two isolated QTLs, and with sample size $N = 201$. FDR control was at the 0.05 level, and all methods used the analytic FDR. MCP represents Minimax Concave Penalty. See Table S1 for other definitions.

Criterion	T	MCP ($\lambda_2 = 3$)	MCP ($\lambda_2 = 10$)	MCP ($\lambda_2 = 30$)	MCP ($\lambda_2 = 100$)
tFDR	0.25	0.022 \pm 0.010	0.018 \pm 0.008	0.020 \pm 0.009	0.018 \pm 0.008
	0.3	0.025 \pm 0.010	0.020 \pm 0.009	0.022 \pm 0.009	0.020 \pm 0.009
	0.5	0.025 \pm 0.010	0.020 \pm 0.009	0.022 \pm 0.009	0.022 \pm 0.009
TPR1		0.375 \pm 0.025	0.380 \pm 0.025	0.378 \pm 0.025	0.380 \pm 0.025
TPR2	0.5	0.432 \pm 0.027	0.435 \pm 0.027	0.432 \pm 0.027	0.435 \pm 0.027
	0.7	0.422 \pm 0.027	0.428 \pm 0.027	0.425 \pm 0.027	0.425 \pm 0.027
	0.9	0.410 \pm 0.027	0.415 \pm 0.027	0.412 \pm 0.027	0.412 \pm 0.027

Table S5 Penalized regression analysis using Lasso, Adaptive Lasso and Elastic Net, with FDR based selection of the penalty parameter values, of 200 replicates of chromosome 21 data with eight QTLs, and with sample size N = 201. FDR control was at the 0.05 level, and all methods used the analytic FDR except Lasso perm which used the permutation FDR. EN represents the Elastic Net with Lasso weight (λ_2) set to 0.3, 0.5 or 0.9. AdaLasso represents the Adaptive Lasso using weights obtained by Lasso with CV (CV), Ridge Regression with CV (RR), and SMA. See Table S1 for other definitions.

Criterion	T	Lasso perm	Lasso	AdaLasso (CV)	AdaLasso (RR,SMA)	EN ($\lambda_2 = 0.3$)	EN ($\lambda_2 = 0.5$)	EN ($\lambda_2 = 0.9$)
tFDR	0.25	0.006 + 0.003	0.020 + 0.005	0.019 + 0.005	0.020 + 0.005	0.026 ± 0.005	0.023 ± 0.005	0.019 ± 0.005
	0.3	0.008 + 0.003	0.023 + 0.005	0.022 + 0.005	0.023 + 0.005	0.028 ± 0.005	0.026 ± 0.005	0.022 ± 0.005
	0.5	0.024 ± 0.006	0.042 ± 0.007	0.041 ± 0.006	0.042 ± 0.007	0.053 ± 0.006	0.048 ± 0.006	0.040 ± 0.006
TPR1		0.315 ± 0.011	0.378 ± 0.014	0.372 ± 0.014	0.378 ± 0.014	0.548 ± 0.017	0.466 ± 0.016	0.394 ± 0.014
TPR2	0.5	0.613 ± 0.016	0.692 ± 0.018	0.690 ± 0.018	0.692 ± 0.018	0.739 ± 0.018	0.718 ± 0.018	0.704 ± 0.018
	0.7	0.514 ± 0.015	0.597 ± 0.018	0.594 ± 0.018	0.597 ± 0.018	0.661 ± 0.018	0.632 ± 0.018	0.609 ± 0.018
	0.9	0.366 ± 0.012	0.441 ± 0.015	0.438 ± 0.015	0.441 ± 0.015	0.570 ± 0.017	0.509 ± 0.016	0.456 ± 0.015

Table S6 Penalized regression analysis using MCP with different fixed values of the second tuning parameter (λ_2) and with FDR based selection of the value of the first tuning parameter, of 200 replicates of chromosome 21 data with eight QTLs, and with sample size $N = 201$. FDR control was at the 0.05 level, and all methods used the analytic FDR. MCP represents Minimax Concave Penalty. See Table S1 for other definitions.

Criterion	T	MCP ($\lambda_2 = 3$)	MCP ($\lambda_2 = 10$)	MCP ($\lambda_2 = 30$)	MCP ($\lambda_2 = 100$)
tFDR	0.25	0.018 \pm 0.006	0.018 \pm 0.005	0.018 \pm 0.004	0.019 \pm 0.005
	0.3	0.020 \pm 0.007	0.021 \pm 0.005	0.021 \pm 0.005	0.021 \pm 0.005
	0.5	0.029 \pm 0.007	0.037 \pm 0.006	0.037 \pm 0.006	0.040 \pm 0.006
TPR1		0.242 \pm 0.010	0.325 \pm 0.013	0.363 \pm 0.014	0.378 \pm 0.014
TPR2	0.5	0.643 \pm 0.018	0.681 \pm 0.018	0.694 \pm 0.018	0.700 \pm 0.018
	0.7	0.476 \pm 0.016	0.566 \pm 0.017	0.593 \pm 0.018	0.599 \pm 0.018
	0.9	0.289 \pm 0.011	0.389 \pm 0.014	0.428 \pm 0.015	0.441 \pm 0.015

Table S7 LD2lasso analyses of 20 randomly selected replicates of chromosome 21 data with two isolated QTLs, and with sample size N = 201. Lasso and EN ($\lambda_2 = 0.5$) results are presented for comparison. For LD2lasso, φ is the relative weight on the Lasso penalty, and the LD function $h(r)$ was set to zero when the absolute correlation between two SNPs was $|r| < 0.85$ or 0.50. All methods used the analytic FDR method with control at the 0.05 level. See Table S1 for other definitions.

Criterion	T	Lasso	EN ($\lambda_2 = 0.5$)	LD2lasso $\varphi = 0.9$ $h(r) = r^2$ $ r < 0.85$	LD2lasso $\varphi = 0.5$ $h(r) = r^2$ $ r < 0.85$	LD2lasso $\varphi = 0.9$ $h(r) = r $ $ r < 0.85$	LD2lasso $\varphi = 0.5$ $h(r) = r $ $ r < 0.85$
tFDR	0.25	0	0	0	0	0	0
	0.3	0	0	0	0	0	0
	0.5	0	0	0	0	0	0
TPR1		0.350 ± 0.089	0.425 ± 0.092	0.475 ± 0.098	0.500 ± 0.101	0.475 ± 0.098	0.450 ± 0.101
	TPR2	0.5	0.450 ± 0.095	0.475 ± 0.098	0.500 ± 0.101	0.525 ± 0.098	0.500 ± 0.101
	0.7	0.450 ± 0.095	0.475 ± 0.098	0.500 ± 0.101	0.525 ± 0.098	0.500 ± 0.101	0.475 ± 0.098
	0.9	0.425 ± 0.098	0.450 ± 0.095	0.475 ± 0.098	0.525 ± 0.098	0.475 ± 0.098	0.450 ± 0.101

Criterion	T	LD2lasso $\varphi = 0.9,$ $h(r) = r^2, r $ < 0.50	LD2lasso $\varphi = 0.5,$ $h(r) = r^2,$ $ r < 0.50$
tFDR	0.25	0	0
	0.3	0	0
	0.5	0	0
TPR1		0.475 ± 0.098	0.450 ± 0.101
	TPR2	0.5	0.500 ± 0.101
	0.7	0.500 ± 0.101	0.475 ± 0.098
	0.9	0.475 ± 0.098	0.450 ± 0.101

Table S8 LD2lasso analyses of 20 randomly selected replicates of chromosome 21 data with eight QTLs, and with sample size N = 201. Lasso and EN ($\lambda_2 = 0.5$) results are presented for comparison. For LD2lasso, φ is the relative weight on the Lasso penalty, and the LD function $h(r)$ was set to zero when the absolute correlation between two SNPs was $|r| < 0.85$. All methods used the analytic FDR method with control at the 0.05 level. See Table S1 for other definitions.

Criterion	T	Lasso	EN ($\lambda_2 = 0.5$)	LD2lasso $\varphi = 0.9$ $h(r) = r^2$ $ r = 0.85$	LD2lasso $\varphi = 0.5$ $h(r) = r^2$ $ r = 0.85$	LD2lasso $\varphi = 0.9$ $h(r) = r $ $ r = 0.85$	LD2lasso $\varphi = 0.5$ $h(r) = r $ $ r = 0.85$
tFDR	0.25	0.005 ± 0.005	0.014 ± 0.006	0.191 ± 0.041	0.201 ± 0.041	0.195 ± 0.041	0.204 ± 0.041
	0.3	0.005 ± 0.005	0.014 ± 0.006	0.194 ± 0.041	0.204 ± 0.041	0.198 ± 0.041	0.207 ± 0.044
	0.5	0.024 ± 0.014	0.033 ± 0.016	0.252 ± 0.048	0.259 ± 0.048	0.256 ± 0.048	0.269 ± 0.051
TPR1		0.269 ± 0.041	0.469 ± 0.044	0.500 ± 0.051	0.562 ± 0.054	0.500 ± 0.051	0.544 ± 0.06
TPR2	0.5	0.631 ± 0.057	0.688 ± 0.060	0.769 ± 0.057	0.769 ± 0.057	0.769 ± 0.057	0.744 ± 0.006
	0.7	0.569 ± 0.057	0.625 ± 0.057	0.650 ± 0.057	0.656 ± 0.057	0.65 ± 0.057	0.625 ± 0.067
	0.9	0.419 ± 0.047	0.512 ± 0.048	0.569 ± 0.054	0.600 ± 0.054	0.569 ± 0.054	0.581 ± 0.06

Criterion	T	LD2lasso $\varphi = 0.9,$ $h(r) = r^2, r = 0.50$	LD2lasso $\varphi = 0.5,$ $h(r) = r^2, r = 0.50$
tFDR	0.25	0.194 ± 0.041 (0.00058)	0.166 ± 0.041 (0.0037)
	0.3	0.199 ± 0.041 (0.00041)	0.169 ± 0.041 (0.0031)
	0.5	0.255 ± 0.048 (6.271e-05)	0.223 ± 0.048 (0.0045)
TPR1		0.556 ± 0.054	0.650 ± 0.057
TPR2	0.5	0.756 ± 0.060	0.775 ± 0.060
	0.7	0.650 ± 0.057	0.694 ± 0.060
	0.9	0.594 ± 0.057	0.662 ± 0.057

Table S9 Multi-split analysis using the method of Meinshausen et al. (2009) of 20 randomly chosen replicates of chromosome 21 data with two isolated QTLs or eight QTLs, and with sample size N=201. The analysis used Benjamini-Hochberg (BH) or Benjamini-Yekutieli (BY) FDR control. See Table S1 for other definitions.

Criterion	T	Two QTLs			Eight QTLs		
		Lasso	BH	BY	Lasso	BH	BY
tFDR	0.25	0	0	0	0.005 ± 0.005	0	0
	0.3	0	0	0	0.005 ± 0.005	0	0
	0.5	0	0	0	0.024 ± 0.014	0	0
TPR1		0.350 ± 0.089	0.150 ± 0.073	0.050 ± 0.035	0.269 ± 0.041	0.094 ± 0.025	0.038 ± 0.016
TPR2	0.5	0.450 ± 0.095	0.175 ± 0.076	0.050 ± 0.035	0.631 ± 0.057	0.206 ± 0.051	0.069 ± 0.032
	0.7	0.450 ± 0.095	0.175 ± 0.076	0.050 ± 0.035	0.569 ± 0.057	0.144 ± 0.041	0.056 ± 0.029
	0.9	0.425 ± 0.098	0.175 ± 0.076	0.050 ± 0.035	0.419 + 0.047	0.100 ± 0.025	0.044 ± 0.019

Table S10 Adaptive Lasso with Local FDR estimation with bootstrap size 100 and different cut-off values for the local FDR (locFDR), of 200 replicates of chromosome 21 data with two isolated QTLs, and with sample size N=201. For comparison, results for the Lasso with analytic FDR control are also shown. See Table S1 for other definitions.

Criterion	T	Two QTLs			Eight QTLs			
		Lasso	AdaLasso locFDR=0.1	AdaLasso locFDR=0.5	Lasso	AdaLasso locFDR=0.1	AdaLasso locFDR=0.5	AdaLasso locFDR=0.9
tFDR	0.25	0.015 ± 0.008	0.020 ± 0.006	0.328 ± 0.025	0.020 ± 0.005	0	0.033 ± 0.006	0.573 ± 0.014
	0.3	0.018 ± 0.008	0.020 ± 0.006	0.353 ± 0.025	0.023 ± 0.005	0	0.033 ± 0.006	0.589 ± 0.015
	0.5	0.018 ± 0.008	0.020 ± 0.006	0.354 ± 0.025	0.042 ± 0.007	0.033 ± 0.011	0.086 ± 0.009	0.627 ± 0.014
TPR1		0.382 ± 0.025	0.150 ± 0.020	0.400 ± 0.025	0.378 ± 0.014	0.056 ± 0.006	0.188 ± 0.010	0.344 ± 0.009
TPR2	0.5	0.432 ± 0.027	0.175 ± 0.021	0.575 ± 0.021	0.692 ± 0.018	0.125 ± 0.014	0.350 ± 0.016	0.738 ± 0.014
	0.7	0.425 ± 0.027	0.175 ± 0.021	0.500 ± 0.023	0.597 ± 0.018	0.094 ± 0.010	0.281 ± 0.013	0.538 ± 0.014
	0.9	0.412 ± 0.027	0.175 ± 0.021	0.475 ± 0.024	0.441 ± 0.015	0.069 ± 0.007	0.212 ± 0.010	0.381 ± 0.010

Table S11 Two dimensional (2D) MCP implemented in PUMA, across 200 replicates of chromosome 21 data with eight QTLs, and with sample size N=201. Final SNP selection was based on p-values with different cut-offs.

Criterion	T	2D MCP with different p-value thresholds				2D MCP analytic FDR
		1×10^{-07}	1×10^{-06}	1×10^{-05}	0.05	
tFDR	0.25	0.025 ± 0.006	0.061 ± 0.009	0.138 ± 0.013	0.729 ± 0.005	0.013 ± 0.004
	0.3	0.025 ± 0.006	0.061 ± 0.009	0.139 ± 0.013	0.74 ± 0.004	0.015 ± 0.004
	0.5	0.025 ± 0.006	0.066 ± 0.01	0.142 ± 0.013	0.756 ± 0.004	0.032 ± 0.006
TPR1		0.191 ± 0.009	0.217 ± 0.009	0.244 ± 0.009	0.356 ± 0.01	0.312 ± 0.014
TPR2	0.5	0.516 ± 0.017	0.581 ± 0.016	0.632 ± 0.015	0.839 ± 0.01	0.644 ± 0.021
	0.7	0.386 ± 0.013	0.433 ± 0.013	0.473 ± 0.013	0.638 ± 0.012	0.528 ± 0.019
	0.9	0.231 ± 0.009	0.261 ± 0.009	0.289 ± 0.009	0.414 ± 0.01	0.365 ± 0.015

Table S12 Joint analysis of chromosomes 19, 21 and 22 by single marker analysis. Compared are pooled and separate analyses using the local FDR (locFDR) based thresholding procedure or the BH procedure, and the locFDR based grouping procedure of Cai and Sun (2009) referred to as Group locFDR. See Table S1 for other definitions.

Criterion	T	Pooled locFDR	Separate locFDR	Group locFDR	Pooled BH	Separate BH
tFDR	0.25	0.026 ± 0.004	0.036 ± 0.004	0.052 ± 0.005	0.050 ± 0.004	0.058 ± 0.005
	0.3	0.029 ± 0.004	0.039 ± 0.004	0.058 ± 0.005	0.056 ± 0.005	0.065 ± 0.005
	0.5	0.087 ± 0.006	0.117 ± 0.007	0.135 ± 0.007	0.127 ± 0.007	0.156 ± 0.007
TPR1		0.576 ± 0.014	0.610 ± 0.013	0.634 ± 0.013	0.636 ± 0.013	0.659 ± 0.012
TPR2	0.5	0.664 ± 0.015	0.694 ± 0.013	0.724 ± 0.013	0.720 ± 0.014	0.737 ± 0.012
	0.7	0.624 ± 0.014	0.654 ± 0.013	0.682 ± 0.013	0.682 ± 0.014	0.699 ± 0.012
	0.9	0.582 ± 0.014	0.616 ± 0.013	0.642 ± 0.013	0.644 ± 0.013	0.665 ± 0.012

Table S13 Analyses of chromosomes 19, 21 and 22 by separate and pooled Lasso PR with analytic FDR control at the 0.05 level. Also compared are the results from separate and pooled analyses of the three chromosomes for individual chromosomes 21 and 22. See Table S1 for other abbreviations / definitions.

Criterion	T	Separate	Pooled	C21 Separate	C21 Pooled	C22 Separate	C22 Pooled
tFDR	0.25	0.028 ± 0.005	0.023 ± 0.004	0.021 ± 0.004	0.010 ± 0.003	0.025 ± 0.009	0.024 ± 0.007
	0.3	0.028 ± 0.005	0.024 ± 0.004	0.021 ± 0.004	0.010 ± 0.003	0.025 ± 0.009	0.024 ± 0.007
	0.5	0.041 ± 0.005	0.035 ± 0.005	0.034 ± 0.005	0.024 ± 0.005	0.028 ± 0.009	0.028 ± 0.008
TPR1		0.399 ± 0.011	0.414 ± 0.013	0.422 ± 0.013	0.390 ± 0.014	0.305 ± 0.022	0.510 ± 0.024
TPR2	0.5	0.656 ± 0.013	0.664 ± 0.016	0.733 ± 0.016	0.687 ± 0.017	0.348 ± 0.023	0.572 ± 0.025
	0.7	0.577 ± 0.012	0.586 ± 0.015	0.636 ± 0.015	0.591 ± 0.016	0.340 ± 0.023	0.568 ± 0.025
	0.9	0.456 ± 0.011	0.472 ± 0.014	0.486 ± 0.014	0.449 ± 0.015	0.340 ± 0.023	0.565 ± 0.025

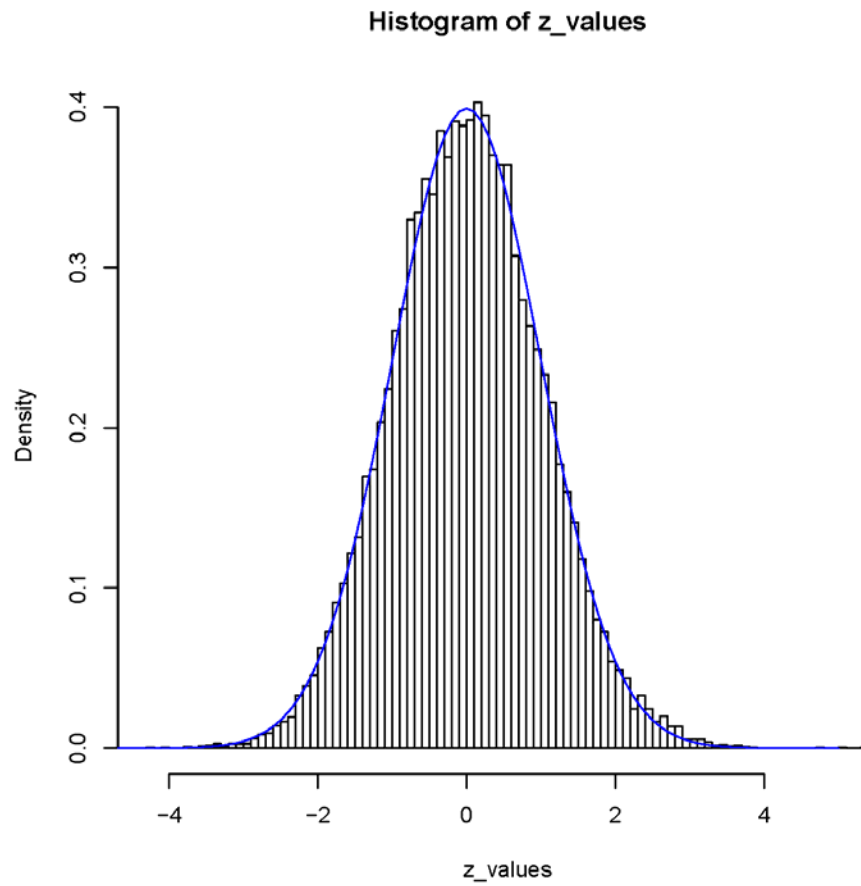


Figure S1 Histogram of the z-values pertaining to 21,530 SNPs on chromosome 21. The blue solid curve represents the $N(0,1)$ distribution, which fits the empirical z-value distribution well.

File S2 SUPPLEMENT S2

HABC Study

The Health ABC study is a prospective cohort study investigating the associations between body composition, weight-related health conditions, and incident functional limitation in older adults. Health ABC enrolled well-functioning, community-dwelling black (n=1281) and white (n=1794) men and women aged 70-79 years between April 1997 and June 1998. Participants were recruited from a random sample of white and all black Medicare eligible residents in the Pittsburgh, PA, and Memphis, TN, metropolitan areas. The present study sample consists of 786 white participants with available genotyping and IL-6 SR data.

Phenotypic Information

To measure the level of IL-6 SR, venipuncture was performed for each of the participants after an overnight fast of at least 8 h, and serum samples were then frozen at -70°C. IL-6 SR levels were measured by ultrasensitive ELISA (R&D Systems) and had CVs of 3.5%–5.2%.

Genotyping and Imputation

Genomic DNA was extracted from buffy coat collected using PUREGENE® DNA Purification Kit during the baseline exam. Genotyping was performed by the Center for Inherited Disease Research (CIDR) using the Illumina Human1M-Duo BeadChip system. Samples were excluded from the dataset for the reasons of sample failure, genotypic sex mismatch, and first-degree relative of an included individual based on genotype data. Genotyping was successful for 1,151,215 SNPs in 2,802 unrelated individuals (1663 Caucasians and 1139 African Americans). Imputation was done for the autosomes using the MACH software version 1.0.16. SNPs with minor allele frequency $\geq 1\%$, call rate $\geq 97\%$ and HWE $p \geq 10^{-6}$ were used for imputation. HapMap II phased haplotypes were used as reference panels. For Caucasians, genotypes were available on 914,263 high quality SNPs for imputation based on the HapMap CEPH reference panel (release 22, build 36).

File S3 Supplement S3

```
#####  
##### Simulate 200 SNP genotype data sets using Hapgen2  
#####  
  
# Change work directory to the directory containing "hapgen2" software  
and HapMap2 Ref#24 haplotype data sets  
cd ~/.../Hapgen2  
  
# Do the loop to generate 200 SNP genotype data sets.  
# Note that the seed is the current time by default.  
# So it's better to generate a few data sets in each loop,  
# and do it for a few times with time gaps.  
for i in {1..10}; do  
./Hapgen2 -h hapmap_r24_b36_fwd.consensus.qc.poly.chr21_ceu.phased -l  
chr21.ceu.r24.legend -m genetic_map_chr21_CEU_b36.txt -dl 34238344 1 1  
1 -n 201 0 -o ./Results2/$i; done  
  
#####  
##### R Code to pre-process SNP genotype data  
#####  
  
iteration <- 200 # Number of SNP data replicates  
N <- 201 # Sample size  
# Record number of SNPs remaining after filtering by  
# MAF <= 0.01 and correlation >= 0.999.  
nSNPMAF <- nSNPMAFVar <- nSNPVarCor <- numeric(iteration)  
# Record SNP positions for each of the 200 SNP data.  
SNP.pos <- matrix(0, iteration, 30000)  
  
for (its in 1:iteration) {  
### Read each SNP data set  
setwd("../Hapgen2/Results2")  
dataName <- paste(its, ".controls.gen", sep="")  
Data <- read.table(dataName)  
Genotype0 <- Data[, -c(1:5)] # first 5 columns are info  
  
### compute allelic dose from three-column SNP data  
Genotype1 <- matrix(0, nrow(Genotype0), N) # P*N SNP matrix  
for (t in 1:N){  
index <- which(Genotype0[, (t-1)*3+1] ==1)  
Genotype1[index,t] <- 0  
index <- which(Genotype0[, (t-1)*3+2] ==1)  
Genotype1[index,t] <- 1  
index <- which(Genotype0[, t*3] ==1)  
Genotype1[index,t] <- 2  
}  
}
```

```

row.names(Genotype1) <- Data[, 3] # SNP positions

### Remove SNPs with MAF <= 0.01
MAF <- rowSums(Genotype1) / (2 * ncol(Genotype1))
id1 <- c(which(MAF <= 0.01), which(MAF >= 0.99))
Genotype2 <- t(Genotype1[-id1,])
colnames(Genotype2) <- rownames(Genotype2)[-id1]
nSNPMAF[its] <- ncol(Genotype2) # no. SNPs after MAF filter

## Remove SNPs with variance 0
id2 <- c(which(colSums(Genotype2) == 0), which(colSums(Genotype2)
== N), which(colSums(Genotype2) == (2*N)), which(colSums(Genotype2) ==
(3*N)))
Genotype3 <- Genotype2[,-id2]
colnames(Genotype3) <- colnames(Genotype2)[-id2]
nSNPMAFVar[its] <- ncol(Genotype3) # Record number of SNPs after
MAF <= 0.01 filtering
X <- Genotype3

### Remove SNPs with correlation > 0.999
for (t in 1:10) {
  P <- ncol(X)
  id2 <- c()
  for (i in 1:(P-1)){
    if ((i+500) <= P) {
      COR <- cor(X[,i], X[, (i+1):(i+500)])
      temp <- which( abs(COR) > 0.999)
      id2 <- c(id2, i+temp)
    }
    if ((i+500) > P) {
      COR <- cor(X[,i], X[, (i+1):P])
      temp <- which( abs(COR) > 0.999)
      id2 <- c(id2, i+temp)
    }
  }
  if (length(unique(id2)) > 0) {
    tmp <- X[, -unique(id2)]
    X <- tmp
  }
  if (length(unique(id2)) == 0) {
    nSNPVarCor[its] <- ncol(X)
    break
  }
}

### Change data to minor alleles consistent for all SNPs
tmp <- colSums(X) / (2*nrow(X))
id <- which(tmp >= 0.5)
X[, id] <- 2 - X[, id]

### Record SNPs positions
SNP.pos[its, (1:ncol(X))] <- colnames(X)

```

```

### Save final, clean SNP data sets
setwd("../Hapgen2/SNP_matrix")
dataName2 <- paste(its, ".txt", sep="") # Define data name
write.table(X, dataName2, row.names=FALSE, col.names=FALSE)
}

### Save SNP positions for each SNP data replicate
write.table(SNP.pos, '../Hapgen2/SNP_matrix/position.txt',
row.names=FALSE, col.names=FALSE)
### Save number of SNPs after MAF and correlation scanning
write.table(nSNPCor, '../Hapgen2/SNP_matrix/nSNPCor2.txt',
row.names=FALSE, col.names=FALSE)

#####
#### Get the common SNPs across 200 data replicates
#####

### Get the common SNP positions across 200 data replicates
SNP.pos <- read.table('../Hapgen2/SNP_matrix/position.txt')
pos1 <- SNP.pos[1, -which(SNP.pos[1,] == 0)]
pos2 <- SNP.pos[2, -which(SNP.pos[2,] == 0)]
comSNP <- pos1[ which(is.na(match(pos1, pos2)) == F) ]
for (i in 3:iteration) {
  tmp <- SNP.pos[i, -which(SNP.pos[i, ] == 0)]
  comSNP <- tmp[ which(is.na(match(tmp, comSNP)) == F) ]
}
length(comSNP) # 9224

### Save common SNP matrix across 200 data replicates
X.all <- matrix(0, N*iteration, length(comSNP))
for (its in 1:iteration) {
  dataName2 <- paste(its, ".txt", sep="")
  X <- read.table(dataName2)
  X <- as.matrix(X)
  id <- match(comSNP, SNP.pos[its, ])
  X.all[((its-1)*N + 1):(its*N), ] <- X[, id]
}
write.table(X.all, '../Hapgen2/SNP_matrix/commonSNPs.txt',
row.names=FALSE, col.names=FALSE)
### Save common SNPs positions
write.table(comSNP, '../Hapgen2/SNP_matrix/commSNPnames.txt',
row.names=FALSE, col.names=FALSE)

#####
##### Select two isolated QTL loci
#####

### Select two QTL loci
loc <- c(174, 8566)
Q <- length(loc) # Number of QTLs
MAF <- colSums(X.all) / (2*nrow(X.all))

```

```

MAF[loc]
locQTL <- comSNP[loc]          #locations: 14516982 44540242
X.2QTL <- X.all[, loc]

### Calculate variance and correlation of 2 QTLs
fi <- colSums(X.2QTL) / (2*nrow(X.2QTL))
2*fi*(1-fi)                    # SNPs variance based on HWE
diag(var(X.2QTL))              # SNPs variance
cor(X.2QTL)                    # correlation of two QTLs

### Test if the correlation of 2 QTLs is significant or not
require("psych")
r.test(N*iteration, cor(X.2QTL)[1,2])$p      # Not significant

### Save QTL data across 200 data replicates
write.table(X.2QTL, '../Hapgen2/SNP_matrix/twoQTLs.txt',
row.names=FALSE, col.names=FALSE)

#####
##### Get two QTL locations for each SNP data replicates
#####

SNP.pos <- read.table("../Hapgen2/SNP_matrix/position.txt")
N <- 201
Q <- length(loc)
locAll <- matrix(0, iteration, Q)

for (its in 1:iteration) {
  setwd("../Hapgen2/SNP_matrix")
  dataName2 <- paste(its, ".txt", sep="")
  X <- read.table(dataName2)

  # Get location of the two QTLs in each SNP data
  loc <- which(is.na(match(SNP.pos[its, -which(SNP.pos[its, ] ==
0)], locQTL)) == F)
  locAll2[its, ] <- loc
}

# Save two QTL locations across 200 SNP data replicates
write.table(locAll2, '../Hapgen2/SNP_matrix/locAll2.txt',
row.names=FALSE, col.names=FALSE)

#####
##### Generate phenotype data Y with two QTLs
#####

X.2QTL <- read.table("../Hapgen2/SNP_matrix/twoQTLs.txt")
locAll <- read.table("../Hapgen2/SNP_matrix/locAll2.txt")
h <- c(0.1, 0.1)                # Heritability of each QTL
beta <- sqrt( h/diag(var(X.2QTL)) ) # var(Y) = 1
var_res <- 1 - sum(h)          # residual variance

```

```

for (its in 1:iteration) {
  setwd("../Hapgen2/SNP_matrix")
  dataName2 <- paste(its, ".txt", sep="")
  X <- read.table(dataName2)
  Y <- as.matrix( X[, unlist(locAll[its, ])] ) %*% beta + rnorm(N, 0,
sqrt(var_res))
  setwd("../Hapgen2/Response_2QTL")
  Yname <- paste("Y", its, ".txt", sep="")
  write.table(Y, Yname, row.names=FALSE, col.names=FALSE)
}

#####
##### Select eight correlated QTL loci
#####

### Select eight QTL loci: 2 groups of 2 SNPs with  $r^2=0.5$ , one
### group of 4 SNPs with  $0.01 \leq r^2 \leq 0.1$ .
loc <- c(195,198, 3337,3341,3343,3344, 8803,8814)
Q <- length(loc)
# Location of eight QTLs:
# 1st group: 14562752, 14569224
# 2nd group: 27048038, 27053105, 27068060, 27069599
# 3rd group: 45312259, 45360242
MAF <- colSums(X.all) / (2*nrow(X.all))
locQTL <- comSNP[loc]
X.8QTL <- X.all[, loc]

### variance and correlation of eight QTLs
fi <- colSums(X.8QTL) / (2*nrow(X.8QTL))
2*fi*(1-fi)
diag(var(X.8QTL))
cor(X.8QTL)

### Test if correlation between groups is significant or not
require("psych")
r.test(N*iteration, cor(X.8QTL)[1,3])$p # not significant

### Save eight QTL data across 200 data replicates
write.table(X.8QTL, '../Hapgen2/SNP_matrix/eightQTLs.txt',
row.names=FALSE, col.names=FALSE)

#####
##### Get eight QTL locations for all SNP data replicates
#####

SNP.pos <- read.table('../Hapgen2/SNP_matrix/position.txt')
N <- 201
Q <- length(loc)
locAll <- matrix(0, iteration, Q)

for (its in 1:iteration) {

```

```

        setwd("../Hapgen2/SNP_matrix")
        dataName2 <- paste(its, ".txt", sep="")
        X <- read.table(dataName2)
        loc <- which(is.na(match(SNP.pos[its, -which(SNP.pos[its, ] ==
0)], locQTL)) == F)
        locAll8[its, ] <- loc
    }
write.table(locAll8, '../Hapgen2/SNP_matrix/locAll8.txt',
row.names=FALSE, col.names=FALSE)

#####
##### Generate phenotype data Y with 8 QTLs
#####

X.8QTL <- read.table("../Hapgen2/SNP_matrix/eightQTLs.txt")
locAll <- read.table("../Hapgen2/SNP_matrix/locAll8.txt")
h <- c(rep(0.04, 2), rep(0.05, 4), rep(0.04, 2)) # QTL h2
sum(h) # Total h2 = 0.68
E <- c(1, 1, 1, -1, 1, -1, -1, -1) # effect signs
beta <- sqrt(h/diag(var(X[, loc]))) * E # var(Y) = 1
var_res <- 1 - sum(h) # residual variance

for (its in 1:iteration) {
    setwd("../Hapgen2/SNP_matrix")
    dataName2 <- paste(its, ".txt", sep="")
    X1 <- read.table(dataName2)
    Y <- as.matrix( X1[, locAll[its, ]])%*%beta + rnorm(N, 0,
sqrt(var_res))
    setwd("../Hapgen2/Response_8QTL")
    Yname <- paste("Y", its, ".txt", sep="")
    write.table(Y, Yname, row.names=FALSE, col.names=FALSE)
}

#####
### actual heritability explained by each QTL by considering LD
#####

R <- cov(X[, loc])
H <- indH <- numeric(Q)
for (i in 1:Q) {
    temp <- 0
    for (j in 1:Q) {
        H[i] <- H[i] + R[i, j]*beta[i]*beta[j]
        if (j != i) temp <- temp + R[i, j]*beta[j]/R[i,i]
    }
    indH[i] <- R[i,i]*(beta[i] + temp)^2
}
sumH <- sum(H)
totH <- sumH / (sumH + var_res)
totH # total heritability by taking into account LD
individualH <- indH / (sumH + var_res)
round(individualH, 3) # Actual individual heritability with LD

```

```

#####
### required design/covariate matrix standardization function
#####
# first subtracts column mean and then divides column by X(T)X/N;
#   N = number of rows of X (sample size)
standardize <- function(X)
{
  n <- nrow(X)
  center <- colMeans(X)
  X.c <- sweep(X, 2, center)
  scale <- sqrt(apply(X.c,2,crossprod)/n)
  val <- sweep(X.c, 2, scale, "/")
  attr(val,"center") <- center
  attr(val,"scale") <- scale
  val
}

#####
### this function performs Elastic Net Penalized Regression
### with analytic FDR control
#####
### Arguments:
###   X.D: N x p.D matrix of design covariates
###   X.SNP: N x p.SNP matrix of SNP covariates (gene doses)
###   N: sample size
###   p.D: number of design covariates
###   p.SNP: number of SNPs
###   y: N x 1 vector of continuous phenotypes (assumed normal)
###   alpha: weight on the lasso proportion of the penalty
###   nlambda: number of lasso/L1 tuning parameter values (lambda)
###   FDR.level: desired level at which to control the FDR
### Strong recommendations:
###   Give meaningful names to the columns of X.D.
###   Name the columns of X.SNP with the corresponding SNP names
###   (rsxxx).
###   p.SNP should be a large value (at least several thousand).
###   alpha should be in the range of 0.5 to 0.7.
###   nlambda should be at least 100, we recommend 1000
###   (required for achieving FDR control near the desired level)
###   (and small nlambda is not computationally advantageous).
###   FDR control method is conservative, we therefore recommend
###   setting FDR.level to 0.05.
### Additional information on matrix X.D:
###   If your design includes factors in addition to covariates,
###   please create matrix X.D by using the model.matrix function
###   of the stats R package as shown below:
###   X.D <- model.matrix(~chip+site+age+pc1)[,-1]
###   where chip and site are defined as factors and age and pc1
###   are covariates. E.g.,
###   chip <- as.factor(chip), chip <- factor(chip,labels=temp)
###   In the penalized regression, each column of X.D will be

```

```

### treated as a covariate to which no shrinkage will be applied.
### Requires:
### current version of R package glmnet
### Output: EN.results <- EN.FDR(X.D,X.SNP,y,alpha,nlambd,FDR.level)
### EN.results$alpha: alpha as provided in argument list
### EN.results$nlambd: nlambd as provided in argument list
### EN.results$FDR.level: FDR.level as provided in argument list
### EN.results$n.SNP: number of SNPs selected (nonzero coefficients)
### EN.results$which.SNP: identifiers of selected SNPs
### EN.results$betas.SNP: coefficient estimates of selected SNPs
### EN.results$FDR.achieved: nominal level of FDR control (should
### be just below the desired level)
### EN.results$check.FDR: should be zero (if set to 1 then FDR
### control was not achieved - something went wrong!)

```

```

EN.FDR <- function(X.D,X.SNP,y,alpha,nlambd,FDR.level)
{
  library(glmnet)

  if(is.null(X.D)) {
    X.D.s <- NULL
    n.fix <- 0
  } else {
    X.D.s <- standardize(X.D)
    n.fix <- dim(X.D.s)[2]
  }
  X.SNP.s <- standardize(X.SNP)
  scale <- attr(X.SNP.s, "scale")
  nz <- which(scale > 1e-6)
  X.SNP.s <- X.SNP.s[,nz,drop=FALSE]
  n.Xsnp <- dim(X.SNP.s)[2]
  X <- cbind(X.D.s,X.SNP.s)
  penalty.factor <- c(rep(0,n.fix),rep(1,n.Xsnp))
  idx <- seq(1,nlambd)

  # perform EN penalized regression on grid of lambda values
  EN.fit <-
glmnet(X,y,penalty.factor=penalty.factor,alpha=alpha,nlambd=nlambd)
  lseq <- EN.fit$lambda
  beta <- as.matrix(EN.fit$beta)
  if(is.null(X.D)) {
    beta.SNP <- beta
  } else {
    beta.SNP <- beta[-(1:n.fix),] }
  c.nzero <- colSums(beta.SNP!=0)
  R.vec <- rep(0,length(lseq))
  R.vec[which(c.nzero!=0)] <- 1/c.nzero[which(c.nzero!=0)]

  # analytic FDR calculation
  yy <- y - mean(y)
  sigk <- sqrt( colSums( (

```



```

matrix(rep(yy,length(lseq)),byrow=F,length(y))-X**%beta )^2 ) )
  FDR <- 2*n.Xsnp*pnorm(q=(-
length(y)*lseq*alpha/sigk),mean=0,sd=1,lower.tail=TRUE,log.p=FALSE)*R.
vec
  # find lambda value closest from below to the desired FDR level
diff <- FDR.level - FDR
  ### this way:
lim <- min( (sum(diff>0)+10),nlambda )
diff <- abs(diff[1:lim])
lbest <- which.min(diff)[1]
  ### or this way (ensures the selected lambda has FDR <=
FDR.level):
diff <- diff[diff>0]
idx <- idx[diff>0]
lbest <- idx[which.min(diff)[1]]
  ### end lbest

  # output:
check.FDR <- 0
if(FDR[length(FDR)]<FDR.level) check.FDR <- 1    #should never
happen
  which.SNP <- which( beta.SNP[,lbest] !=0 )
betas.SNP <- beta.SNP[which.SNP,lbest]
which.SNP <- rownames(beta.SNP)[which.SNP]
n.SNP <- length(which.SNP)
FDR.achieved <- FDR.level - diff[lbest]
list(alpha=alpha,nlambd=nlambda,FDR.level=FDR.level,n.SNP=n.SNP,
which.SNP=which.SNP,betas.SNP=betas.SNP,FDR.achieved=FDR.achieved,check.FDR=check.FDR)
}

```