Assessment of Sparse Regression Machine Learning Methods for Genome-Wide Association Studies

Hui Yi

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics and Computational Biology

Ina Hoeschele, Chair
Xinwei Deng
M. A. Saghai Maroof
Hongxiao Zhu

August 11, 2014
Blacksburg, Virginia

Keywords: Genome-wide Association Study, penalized regression, false discovery rate, linkage disequilibrium

Assessment of Sparse Regression Machine Learning Methods for Genome-Wide Association Studies

Hui Yi

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 (including penalties incorporating linkage disequilibrium) in comparison with SMA. PR methods were compared with SMA on realistically simulated GWAS data consisting of genotype data from single and multiple chromosomes and a continuous phenotype and on real data. Based on our 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. PR controlled the FDR conservatively while SMA-BH may not achieve FDR control in all situations. Differences among PR methods seem quite small when the focus is on variable selection with FDR control. Incorporating LD into PR by adapting penalties developed for covariates measured on graphs can improve power but also generate morel false positives or wider regions for follow-up. We recommend using the Elastic Net with a mixing weight for the Lasso penalty near 0.5 as the best method.

ACKNOWLEDGEMENTS

I would like to give special thanks to my committee chair Dr. Ina Hoeschele for her effective guidance through the five years. I thank my committee member Dr. Xinwei Deng, who has taught me Data Analytics course and provided me novel perspectives when thinking about a problem. I thank Dr. M. A. Saghai Maroof, who has taught me Plant Genomics course, one of the most beneficial biology courses I’ve taken here. I also thank Dr. Hongxiao Zhu, for her constructive suggestions and precious experiences sharing which helped me go through all those difficult times.

I appreciate helpful discussions and accompany from my colleagues Ting Guan, Angie Dement, and especially Dr. Netsanet Imam, with whom I had a joyful cooperation for one and a half years. I also thank my parents, sisters, and especially my husband Ning Ni, for their everlasting love and support.

Lastly, I acknowledge funding from GBCB fellowship and NIH grants (R01HG005254 and R01HL101250) for supporting the five years’ Ph.D training.

ATTRIBUTION

Hui Yi drafted Chapter 1 and 3 of this dissertation.

Chapter 2 of this dissertation is a research article which has been submitted to Genetics. Hui Yi ran the simulations and performed the statistical analyses. Dr. Patrick Breheny (an assistant professor in the Department of Biostatistics at the University of Iowa), independently derived the analytic FDR estimator. Dr. Netsanet Imam (a postdoctoral fellow from Dr. Ina Hoeschele's Statistical Genetics group), contributed to the simulation study for multiple chromosomes. Dr. Yongmei Liu (a professor in the Departments of Epidemiology & Prevention and Internal Medicine, Division of Public Health Sciences, and Translational Research Institute at Wake Forest School of Medicine), provided the real data from her Health ABC GWAS project. Dr. Ina Hoeschele (a professor in the Department of Statistics and Virginia Bioinformatics Institute at Virginia Tech) conceived the study, in particular the combination of penalized regression with FDR control, designed and directed the study, interpreted the results and wrote the manuscript.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Genome-wide association studies (GWAS) have identified hundreds of genetic variants, most frequently single nucleotide polymorphisms (SNPs), associated with complex human diseases, such as diabetes, schizophrenia, hypertension, and various forms of cancers, providing insights into their genetic architecture (Altshuler et al., 2008, Frazer et al., 2009, Manolio et al. 2009). While loci with large effect have been identified, the predominant pattern is that multiple loci are identified, each explaining a small portion and their sum less than half of the trait heritability (Gibson 2012). The potential explanations for the unexplained heritability include untested rare variants (Pritchard 2001, Gibson 2012), gene-gene and gene-environment interaction (Cordell 2009, Hunter 2005, Stranger et al. 2010), as well as a lack of statistical power to detect SNPs with small effect sizes (Hayes 2013). The invention of novel analytic approaches is essential to improve power of detecting causal SNPs (especially with small individual heritabilities) that are associated with the common or complex trait of interest, by taking advantage of the rich resources of microarray/resequencing data, such as Hapmap (International HapMap Consortium 2007) and 1000 Genomes Projects (1000 Genomes Project Consortium 2012), from a joint effort of scientists as a national-wide project. The ultimate goal of GWAS is to use genetic risk factors to predict disease risk, progression, and response to therapies and facilitate disease prevention and treatments (Manolio et al. 2009).

The simplest statistical method is the genome-wide association test using single marker regression (SMA; Hindorff et al., 2009). This leads to a large number, typically tens of thousands up to millions of simultaneous hypothesis tests. As a result, a nominal significance level of each test needs to be properly adjusted to achieve the overall error rate control. The disadvantage of this method is that it fits an incorrect model to the phenotype that is affected by more SNPs than one SNP. The alternative methods fit multiple or all markers in a multiple regression model simultaneously (Risch 1990a, 1990b, Hoh and Ott 2003, Moore et al., 2010). The rationale is, by conditioning on the causal SNPs that are already in the model, the marginally uncorrelated causal SNPs have a better chance of being selected while the false positive signals tend to be weakened. This approach is enabled by a form of penalized regression (PR) due to the high dimensionality of the genetic markers, i.e., the number of SNPs $\gg$ the number of samples.

A challenge in implementing and promoting PR methods for GWAS is the determination of the “optimal” value of the penalty parameter(s) and the lack of a measure of error associated with the variable selection. The currently existing approaches for selecting values for the penalty parameters are more appropriate for prediction and model selection than for variable selection. These criteria lead to undesirably high false positive rates because a good predictive or a well-fitting model often contains more variables than needed and may not necessarily contain all the important ones. Previous authors who recognized the absence of the error rate control for PR proposed data permutation, data splitting and sub-sampling based methods, which suffer from high computational costs or reduced power when applied to genome-wide data. Another problem in applying PR to GWAS is the determination of an “optimal” penalty function. In this dissertation, we review PR methods in the context of GWAS, extend them to incorporate FDR control, evaluate different penalty functions which are relevant to GWAS, and assess their performance in comparison with SMA and other multivariate methods on both simulated GWAS data and real data.

REFERENCES

Altshuler, D., Daly, M. J., & Lander, E. S., 2008 Genetic mapping in human disease. science, 322(5903), 881-888.

Cordell, H. J., 2009 Detecting gene–gene interactions that underlie human diseases. Nature Reviews Genetics, 10(6), 392-404.

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J., 2009 Human genetic variation and its contribution to complex traits. Nature Reviews Genetics, 10(4), 241-251.

Gibson, G., 2012 Rare and common variants: twenty arguments. Nature Reviews Genetics, 13(2), 135-145.

Hayes, B., 2013 Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). In Genome-Wide Association Studies and Genomic Prediction (pp. 149-216). Humana Press.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A., 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences, 106(23), 9362-9367.

Hoh, J., & Ott, J., 2003 Mathematical multi-locus approaches to localizing complex human trait genes. Nature Reviews Genetics, 4(9), 701-709.

Hunter, D. J., 2005 Gene–environment interactions in human diseases. Nature Reviews Genetics, 6(4), 287-298.

International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Visscher, P. M., 2009 Finding the missing heritability of complex diseases. Nature, 461(7265), 747-753.

Moore, J. H., Asselbergs, F. W., & Williams, S. M., 2010 Bioinformatics challenges for genome-wide association studies. Bioinformatics, 26(4), 445-455.

1000 Genomes Project Consortium., 2012 An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56-65.

Pritchard, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? The American Journal of Human Genetics, 69(1), 124-137.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., & Lander, E. S., 2001 Linkage disequilibrium in the human genome. Nature,411(6834), 199-204.

Risch, N., 1990 Linkage strategies for genetically complex traits. I. Multilocus models. American journal of human genetics, 46(2), 222.

Risch, N., 1990 Linkage strategies for genetically complex traits. II. The power of affected relative pairs. American journal of human genetics, 46(2), 229.

Stranger, B. E., Stahl, E. A., & Raj, T., 2011 Progress and promise of genome-wide association studies for human complex trait genetics. Genetics, 187(2), 367-383.

CHAPTER 2

Penalized Multi-Marker versus Single-Marker Regression Methods for Genome-Wide Association Studies of Quantitative Traits

A research article submitted in Genetics in 2013.

Hui Yi*,§, Patrick Breheny†, Netsanet Imam*, Yongmei Liu** and Ina Hoeschele*,‡

* Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24061, USA

§ Genetics, Bioinformatics and Computational Biology, Virginia Tech, Blacksburg, VA 24061, USA

† Department of Biostatistics, University of Iowa, Iowa City, IA 52240, USA

**Departments of Epidemiology & Prevention and Internal Medicine, Division of Public Health Sciences, Translational Research Institute, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA

‡ Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA

*,‡ To whom correspondence should be addressed. E-mail: inah@vt.edu

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 (including penalties incorporating linkage disequilibrium) in comparison with SMA. PR methods were compared with SMA on realistically simulated GWAS data consisting of genotype data from single and multiple chromosomes and a continuous phenotype and on real data. Based on our 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. PR controlled the FDR conservatively while SMA-BH may not achieve FDR control in all situations. Differences among PR methods seem quite small when the focus is on variable selection with FDR control. Incorporating LD into PR by adapting penalties developed for covariates measured on graphs can improve power but also generate morel false positives or wider regions for follow-up. We recommend using the Elastic Net with a mixing weight for the Lasso penalty near 0.5 as the best method.

INTRODUCTION

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 association 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 which 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 but not for our 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 multi-stage (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 (including methods with added fusion-type penalties) to incorporate 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.

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.

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 re-sampling 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 dataset, SNPs were removed if they had a Minor Allele Frequency (MAF) below 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 will be 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 ($5.5\times10^{-8}\times3849034/50165$) for chromosome 21 and ($5.5\times10^{-8}\times3849034/(56607+50165+54786)$) 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 QTLs with heritability of 0.10 each and hence a total heritability of 0.20; (2) eight QTLs comprising a group of four QTLs with weak pairwise LD ($0.01 \leq r^2 \leq 0.1$) and two groups of two QTLs each with within-group LD of $r^2 \approx 0.5$. All QTLs had MAF > 0.05. For scenario (2), the four QTLs in one group had a heritability of 0.05, and the remaining four QTLs had a heritability of 0.04. Taking into account LD among the 8 QTLs in scenario (2), the total heritability was 0.48. 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 QTLs. 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, $Y = \sum_{j=1}^{Q} X_j \beta_j + e$. Heritability of a simulated QTL j was computed as $h_j^2 = \beta_j^2 var(X_j)/var(Y)$, where X_j denotes the allelic dose of QTL j, and Y denotes the phenotype. "Effective" heritability (increased due to LD with other QTL) was computed as above but by replacing β_j with

$$\tilde{\beta}_j = cov(\mathrm{Y}, X_j)/var(X_j) = \left[\beta_j var(X_j) + \sum_{\substack{j'=1 \\ j' \neq j}}^{Q} \beta_{j'} cov(X_j, X_{j'})\right] \Big/ var(X_j) \,.$$

For the multi-chromosome simulation, chromosome 19 did not harbor any QTL, chromosome 21 harbored the 8 QTLs 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 cut-off value for the p-values, determined by some method of multiple testing control, most commonly the genome- (family-) wise error rate (GWER), 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.*, The 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 False Discovery Rate (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). The BH procedure orders the P p-values $p_{(P)} \geq p_{(P-1)} \geq \ldots \geq p_{(1)}$ and then finds *k* and rejects all null hypotheses with rank 1 to *k* for

$$k = \max\left\{ j : p_{(j)} \leq \frac{j}{P}\alpha \right\} \qquad [1]$$

where α is the desired level of FDR control. To ensure FDR control under any form of dependence, Benjamini and Yekutieli's (2001) BY approach replaces α with $\alpha \Big/ \sum_{j=1}^{P} \frac{1}{j}$ and is known to be (very) conservative. Additionally, we consider the local FDR (Efron 2005; Efron and Tibshirani 2002), where the p-values, assumed to represent a mixture of null and alternative hypotheses, are transformed to z-scores which are modeled by a mixture distribution, or

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z) \qquad [2]$$

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 [2], the local FDR is

$$locFDR(z) = \pi_0 f_0(z) / f(z) \qquad [3]$$

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 2005). Sun and Cai (2007) developed an oracle testing procedure that minimizes the false negative rate (the expected proportion of false negatives among all non-rejections) 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 [3] are consistent. To optimally control the FDR at level α, the *locFDR* values are ordered from smallest to largest, and null hypotheses 1 to *m* are rejected, where

$$m = max\left\{j: \frac{1}{j}\sum_{k=1}^{j} locFDR_{(k)} \leq \alpha\right\} \qquad [4]$$

We estimated *f(z)* and π_0 using the simple non-parametric approach described in Storey *et al.* (2005). Despite the correlation structure of the z-values, f_0 was fit well by the N(0,1) null distribution (see 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 [3] in Jin and Cai (2007). We will refer to the thresholding (approximate oracle) rule in [4] 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 just below or at 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 following estimator:

$$\hat{\vec{\beta}} = \arg\min_{\vec{\beta}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{p} x_{ik}\beta_k \right)^2 + P(\vec{\lambda}, \vec{\beta}) \right\} \qquad [5]$$

where $\vec{\beta}$ is a vector of *p* regression coefficients, y_i is the quantitative 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*(.) 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*(.). Variable selection, by setting unimportant coefficients to zero, occurs if the penalty has a singularity at zero. Below we only consider 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 which 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

$$P_{Lasso}(\vec{\lambda},\vec{\beta}) = \lambda \sum_{k=1}^{p} \left|\beta_k\right|; \quad \lambda \geq 0 \qquad [6]$$

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 will select 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 multi-collinearity, or

$$P_{EN}(\vec{\lambda},\vec{\beta}) = \lambda_1 \lambda_2 \sum_{k=1}^{p} \left|\beta_k\right| + \lambda_1 (1-\lambda_2) \sum_{k=1}^{p} \beta_k^2; \quad \lambda_1 \geq 0,\ 0 \leq \lambda_2 \leq 1 \qquad [7]$$

In contrast with the Lasso, the EN may be expected to co-select 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

$$P_{AdaLasso}(\vec{\lambda}, \vec{\beta}) = \lambda \sum_{k=1}^{p} w_k \left|\beta_k\right|; \quad \lambda \geq 0, w_k = \left|\tilde{\beta}_k\right|^{-1} \qquad [8]$$

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. Zou considered Ordinary Least-Squares to obtain the weights for the $p < N$ case and Ridge Regression for the $p > N$ case. Huang *et al.* (2008) suggested univariate regression as the initial estimator and showed that, under a partial orthogonality condition and for the case when $p_N \to \infty$ as $N \to \infty$, the Adaptive Lasso is again an oracle procedure. 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 further away from zero (*e.g.*, Breheny and Huang (2011)), as can be seen from the derivatives of the respective penalty functions. 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 only consider the MCP with penalty function

$$P_{MCP}(\vec{\lambda},\beta_k)=\lambda_1\int_0^{|\beta_k|}\left[1-\frac{x}{\lambda_1\lambda_2}\right]_+ dx=\lambda_1\begin{cases}\left(|\beta_k|-\dfrac{\beta_k^2}{2\lambda_1\lambda_2}\right) & \text{if } |\beta_k|\le\lambda_1\lambda_2\\ \dfrac{\lambda_1\lambda_2}{2} & \text{if } |\beta_k|>\lambda_1\lambda_2\end{cases};\lambda_1\ge 0,\ \lambda_2>1 \quad [9]$$

For the second tuning parameter in MCP (λ_2), four values were evaluated (3, 10, 30, 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 (df) of a model, and the df are not straightforward to obtain in PR (*e.g.*, Ye (1998), Zou *et al*. (2007) and 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 which produces one false positive among 100,000 markers tested. Multi-stage 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 p-values by standard multiple regression. This resulted in B raw p-values $\tilde{P}_j^b$ and B adjusted p-values $P_j^{(b)}=\min\left(\tilde{P}_j^b\left|\tilde{S}^b\right|,1\right)$ for each variable j ($j=1,\dots,p$), where $\tilde{S}^b$ is the set of selected variables from the screening step, and

$|\tilde{S}^b|$ is the number of variables in $\tilde{S}^b$. Then adjusted p-values for each variable were obtained as follows, letting $\gamma \in (\gamma_{min}, 1)$ with $\gamma_{min} = 0.05$ as recommended and q_γ denoting the empirical γ-quantile function:

$$P_j = \min\{(1 - \log(\gamma_{min}))\, inf_{\gamma \in (\gamma_{min},1)}\, Q_j(\gamma)\,, 1\}$$

$$\text{where } Q_j(\gamma) = \min\left\{q_\gamma\left(\left\{\frac{P_j^{(b)}}{\gamma}; b = 1, \dots, B\right\}\right), 1\right\}$$

The FWER can be controlled at level α by thresholding P_j. Asymptotic FDR control can be achieved by straightforward modification of BH or BY in [1] to work with adjusted p-values P_j, rejecting all null hypotheses with rank 1 to k for $k = \max\{j : P_{(j)} \leq j\alpha\}$. Asymptotic FDR control occurs under the condition that the method used in the first step satisfies the screening and sparsity properties

$$\lim_{N\to\infty} \Pr(\tilde{S}^b \supseteq S) = 1, and\ |\tilde{S}^b| < \frac{N}{2}$$

where S is the set of true alternative hypotheses. 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 dataset as in Sun *et al.* (2010) should improve power; these authors used the IAL in the screening step and backwards 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 dataset.

Meinshausen and Bühlmann (2010) introduced Stability Selection (SS) which combines sub-sampling with penalized regression by applying the Lasso repeatedly to different sub-samples of size N/2 of the original data. Variables are chosen by their frequency of being selected across all sub-sampled datasets. Meinshausen and Bühlmann presented a theoretical upper bound for controlling the FWER, which can be used to select the regularization based on a desired value of the bound. The bound, however, relies on theoretical conditions whose applicability to real data sets depends on the correlation structure of the predictors and is difficult to assess. The authors showed that their method produces a sparser set of predictors than CV. When applied to GWAS data, SS was very conservative and had low power when compared to SMA (Alexander and Lange (2011); own results not shown). This finding is expected as the Lasso typically chooses one variable from a set of (strongly) correlated variables, but the variable chosen may differ between sub-samples leading to low selection probabilities. We also found power to be dependent on the choice of the value for the threshold imposed on the selection frequencies (within the limits recommended by the authors).

Recently, an improved SS method was proposed by Shah and Samworth (2013) which controls the expected number of included variables which have low selection probabilities rather than the FWER. Another version of stability selection is the Stability Approach to Regularization Selection (StARS) method (Liu *et al.* 2010) which controls variable selection instability (variance) rather than FWER or FDR. For GWAS, these methods are affected by the correlation structure of the SNPs in the same way as the original SS. Lastly, Ahmed *et al.* (2011) combined SS with BH FDR control by using data permutation to compute empirical p-values associated with the selection probabilities. The need to perform both sub-sampling and permutation would make this method challenging to apply to large GWAS datasets. Additionally, this method does not provide an approach for selecting an optimal value of the tuning parameter, and therefore the analysis must be performed on a grid of tuning parameter values within a chosen range (for which the authors obtained fairly stable results). To avoid or reduce the problem of SS when applied to correlated SNPs (and to reduce the computational requirements), the authors thinned their set of SNPs by tagging, which however may reduce power. For these reasons we did not further evaluate the SS based methods in this paper.

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 (Cai and Sun 2009; Sun and Cai 2007).

The PUMA (Penalized Unified Multiple-locus Association) 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 datasets. It performs heuristic model and SNP selection. Models (defined by a single (e.g. Lasso) or a combination (e.g., MCP) of two penalty parameter values) are limited to the subset where the number of non-zero coefficients does not exceed $1.5\sqrt{N}$. Within this set, an 'optimal' model is identified based on AIC. This is followed by heuristic SNP ranking and selection based on fitting an unpenalized regression model including all SNPs with nonzero coefficients, computing p-values for each SNP by sequentially identifying maximally correlated (above a threshold) pairs of SNPs and dropping the one with the lower coefficient and retaining the smallest p-value for each SNP across these unpenalized models, ranking the SNPs based on their p-values and applying an arbitrary cut-off on the number of SNPs or the p-value. 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 $P_{XY}^{adj} = P_{XX}^{-1/2} P_{XY}$, 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 which 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 sub-sampling. 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} which 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 non-zero coefficients is determined at each λ, denoted by $R(\lambda)$. Then B permuted datasets (with random re-orderings of the continuous vector of phenotypes) are analyzed on the same grid of λ values, and the number of non-zero coefficients is again determined for each permuted dataset b ($b = 1,\ldots,B$) 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 non-zero solutions in the original data, $FDR(\lambda) = \frac{1}{B}\sum_{b=1}^{B} F(b,\lambda) \big/ R(\lambda)$. 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 which 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 [9]).

$$f_{Lasso}(\beta_k) = \frac{1}{2}\left(\beta_k - \beta_k^{OLS}\right)^2 + \lambda|\beta_k|; \quad f_{MCP}(\beta_k) = \frac{1}{2}\left(\beta_k - \beta_k^{OLS}\right)^2 + \lambda|\beta_k| - \frac{1}{2\gamma}\beta_k^2 \qquad [10]$$

where $\beta_k^{OLS} = \frac{1}{N}\sum_{i=1}^{N} x_{ik} r_{i-k}$ and $r_{i-k} = y_i - \sum_{j=1, j\neq k}^{p} x_{ij}\beta_j$. Differentiating [10] with respect to β_k yields the solutions

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

$$\beta_k^{MCP} = \begin{cases} \frac{\beta_k^{OLS} - \lambda}{1 - \frac{1}{\gamma}} & if\ \beta_k^{OLS} > 0\ and\ \left|\beta_k^{OLS}\right| > \lambda \\ \frac{\beta_k^{OLS} + \lambda}{1 - \frac{1}{\gamma}} & if\ \beta_k^{OLS} < 0\ and\ \left|\beta_k^{OLS}\right| > \lambda \\ 0 & if\ \left|\beta_k^{OLS}\right| \leq \lambda \\ \beta_k^{OLS} & if\ \left|\beta_k^{OLS}\right| > \gamma\lambda \end{cases} \qquad [11]$$

Based on [11], we consider the probability of a false positive, or $Pr\left(\left|\beta_k^{OLS}\right| > \lambda | \beta_k = 0\right)$. We again define the FDR as $E(F)/R$ and approximate the numerator with

$$\widehat{E(F)} = \sum_{k=1}^{p} Pr\left(\left|\beta_k^{OLS}\right| > \lambda | \beta_k = 0\right) \qquad [12]$$

In general, FDR control based on [12] is expected to be conservative similar to the BH FDR, because in [12] we sum over all variables rather than the (unknown) true null variables. However, this does 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

$$Pr\left(\left|\beta_k^{OLS}\right| > \lambda|\beta_k = 0\right) = Pr\left(\frac{1}{N}\left|\boldsymbol{x}_k^T\boldsymbol{r}_{-k}\right| > \lambda|\beta_k = 0\right) \qquad [13]$$

The distribution of the estimated residuals (*r*) is unknown and complicated but may be approximated by the normal distribution $\boldsymbol{r}_{-k} \sim N(\boldsymbol{0}, \boldsymbol{V}_k)$ or even simpler by $\boldsymbol{r}_{-k} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_k^2)$ such that $\frac{\boldsymbol{1}}{\boldsymbol{N}}\boldsymbol{x}_k^T\boldsymbol{r}_{-k} \sim N\left(0, \frac{1}{N}\sigma_k^2\right)$. 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

$$Pr\left(\frac{1}{N}\left|\boldsymbol{x}_k^T\boldsymbol{r}_{-k}\right| > \lambda|\beta_k = 0\right) = 1 - \Phi\left(\frac{\sqrt{\mathrm{N}}\lambda}{\sigma_\mathrm{k}}\right) + \Phi\left(-\frac{\sqrt{\mathrm{N}}\lambda}{\sigma_\mathrm{k}}\right) = 2\Phi\left(-\frac{\sqrt{\mathrm{N}}\lambda}{\sigma_\mathrm{k}}\right) \qquad [14]$$

Using the "natural" estimator $\hat{\sigma}_k^2 = \frac{1}{N}\boldsymbol{r}^T\boldsymbol{r}$, [14] becomes constant for all *k* and

$$\widehat{FDR} = 2\mathrm{p}\Phi\left(-\frac{\lambda\mathrm{N}}{\sqrt{\boldsymbol{r}^T\boldsymbol{r}}}\right)/\mathrm{R} \qquad [15]$$

The FDR estimation [15] 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 [15] needs to be modified to

$$\widehat{FDR} = 2p\Phi\left(-\frac{\lambda_1\lambda_2\mathrm{N}}{\sqrt{\boldsymbol{r}^T\boldsymbol{r}}}\right)/\mathrm{R} \qquad [16]$$

Penalized Regression With Fusion-Type Penalties

It is well-known that the Lasso tends to select a single predictor from a group of predictors which have strong pair-wise 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 following form (referred to as LD2lasso):

$$\frac{\varphi}{2}\sum_{k=1}^{p}\sum_{m\in S_k} h(r_{km})\cdot\left(|\beta_k|-|\beta_m|\right)^2;\ \ \varphi\geq 0 \qquad [17]$$

where S_k is a set of SNPs which 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} which 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 [17] 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 [17]. Using [5] with the Lasso penalty and the added penalty in [17], the LD2lasso objective function for a single coefficient k can be written as

$f_{LD2lasso}(\beta_k)$

$= \frac{1}{2}\left(\beta_k - \beta_k^{OLS}\right)^2 + \lambda\varphi\left|\beta_k\right| + \lambda(1-\varphi)\frac{1}{2}\sum_{m\in S_k} h(r_{km})\left(\left|\beta_k\right| - \left|\beta_m\right|\right)^2\ ;\ \lambda \geq 0, 0 \leq \varphi \leq 1$ [18]

where

$\beta_k^{OLS} = \frac{1}{N}x_k^T\left(y_i - \sum_{j=1,j\neq k}^{p} x_{ij}\,\beta_j^{LD2lassso}\right)$

Then one can show that the LD2lasso solution satisfies

$$\beta_k^{LD2lassso} = sign(\beta_k^{OLS})\,\frac{\left(\left|\beta_k^{OLS}\right| - \lambda\left(\varphi - (1-\varphi)\sum_{m\in S_k} h(r_{km})\cdot\left|\beta_m^{LD2lasso}\right|\right)\right)_+}{1 + \lambda(1-\varphi)\sum_{m\in S_k} h(r_{km})} \qquad [19]$$

For LD2lasso with analytic FDR control, the probability of a false positive analogous to [14] is

$Pr\left(\frac{1}{N}\left|\boldsymbol{x}_k^T\boldsymbol{r}_{-k}\right| > \lambda\left(\varphi - (1-\varphi)\sum_{m\in S_k} h(r_{km})\cdot\left|\beta_m^{LD2lasso}\right|\right) | \beta_k = 0\right)$

$= 2\Phi\left(-\frac{\sqrt{\mathrm{N}}\kappa}{\sigma_\mathrm{k}}\right);\quad \kappa = \lambda\left(\varphi - \lambda(1-\varphi)\sum_{m\in S_k} h(r_{km})\cdot\left|\beta_m^{LD2lasso}\right|\right)$ [20]

where in contrast with [14] 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 [17] is the following penalty which we refer to as the LD1lasso penalty:

$$\varphi\sum_{k=1}^{p}\sum_{m\in S_k} h(r_{km})\cdot\left\|\beta_k\right| - \left|\beta_m\right\|;\quad \varphi \geq 0 \qquad [21]$$

Penalty [21] is essentially identical to the extra penalty term employed in the General Fused Lasso and Graphical Fused Lasso (*e.g.*, Kim and Xing (2009)), which includes terms $\left|\beta_k - sign(r_{km})\beta_m\right|$ to induce sparsity in the differences among the values of the regression coefficients.

Computation of the LD2lasso solution ([19]) was implemented with a straight-forward path-wise coordinate descent algorithm. It can be verified that for the objective function in [5] with the Lasso penalty plus the penalty in [17] the coordinate descent algorithm will converge to its minimizer. An important result is that if the objective function can be partitioned into a function $g(\vec{\beta})$ which is continuously differentiable and convex and a term which is separable (*i.e.* in the form $\sum_{k=1}^{p} P_k(\beta_k)$ where the β_k are mutually exclusive scalars or vectors and the $P_k(.)$ are convex functions), then the coordinate descent algorithm will converge to the minimizer of the objective function. Relevant references for this result are Friedman *et al.* (2007) who cited Tseng (1988) and Tseng (2001). This result applies to the LD2lasso if we define $g(\vec{\beta})$ as the residual sums of squares term in [5] plus the penalty in [17] and the second term $\left(\sum_{k=1}^{p} P_k(\beta_k)\right)$ as the Lasso penalty. However, for the LD1lasso it is not possible to find both a first term that is continuously differentiable and a second term which is separable. Specialized algorithms to deal with this problem are available (Chen *et al.* 2010; Friedman *et al.* 2007; Kim and Xing 2009; Tseng 2001). For evaluation purposes, we focus on the LD2lasso which was simpler to implement and seemed to perform better (results not shown).

Consider a group of SNPs in strong LD and containing a single SNP which 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 non-zero 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. However, as we show below if there are several causal SNPs (QTLs) in moderate to low LD on the same chromosome, a rather wide region with non-zero solutions may be identified by LD2lasso and contain SNPs which are only in low LD with any individual causal SNP.

Analyses of Multiple Chromosomes

The analysis of multiple chromosomes may be regarded as a multi-group 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 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 multi-marker model, potentially lowering detection power and diminishing the postulated advantage of an all-markers over single marker analysis. Lastly, 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.

<u>SMA</u>

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 in [4] 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 procedure in [4], (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 the resulting method in the *R* package *PLIS*. The HMM models the hidden association status of each SNP, where the observed data (z-statistics for each SNP) were assumed independent conditional on the association status and to follow a two-component mixture distribution with the non-associated component being standard normal and the associated component represented by a mixture of normal distributions with unknown means and variances and an unknown number of components (chosen by the BIC criterion). We tested this method on the single chromosome 21 simulation with 8 true QTLs and *2*08 SNPs linked to true QTLs at a threshold of 0.5 for the (absolute) correlation between causal and linked SNPs (see below the section Comparison Among Methods). The associated mixture of normals had four components as determined by BIC. This method identified 2400 SNPs as significant at the FDR threshold of 0.05, which included the eight true causal SNPs and 195 out of the 208 true linked SNPs (with absolute correlation > 0.5). The other significant SNPs were nearly uniformly distributed along the chromosome. We then replaced the mixture of four normal components with a single normal which reduced the number of significant SNPs only slightly to 2329 and increased the number of linked SNPs to 198 out of 208. Due to this very high number of false positive results 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) which is significant. We define a Linked True Positive (LTP) as a SNP which 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 a LTP. Methods are compared, across the replicate simulations, based on (1) True Positive Rate TPR1 = number of CTPs / number of QTLs, (2) TPR2 = number of QTLs detected by at least one CLTP (CTP or LTP) / number of QTLs, and (3) threshold-specific empirical tFDR = number of significant SNPs which are not a CLTP at a given T / 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, 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).

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 100kb 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 SNPs. 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 Supplement A; the tables include standard errors for all average (across replicates) tFDR and TPR1/2 values which 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 QTLs: Single Marker Analysis

For two isolated QTLs, SMA results are summarized in Figure 1 (and Table S1 in Supplement A). For SMA with GWER control (referred to as Genome-Wide Threshold GWT in Figure 1) which 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 above 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 (above 0.5).

For the eight-QTL scenario (Figure 2 and Table S2 in Supplement A), TPR was increased for the FDR based SMA methods relative to the two-QTL case (0.7 and above for BH, LFDR1 and LFDR2; around 0.6 for BY compared with 0.4 or below for GWT). All FDR methods controlled the FDR in the sense of producing at least one tFDR value which was not significantly higher than 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 QTLs: 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 (above 0.5; results not shown). The results for PR with FDR control for this scenario are presented in Figure 3 and Tables S3 and S4 in Supplement A. 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 λ_2=3 to λ_2=100) produced almost identical results. All PR methods were able to control the FDR by producing tFDR values below 0.05 for all T. In terms of power and using the EN50 to represent PR, in comparison with SMA-BH, the average TPR1/2 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 TPR1/2 values of EN50 were higher with all p-values around 0.002. The Lasso with permutation FDR control had slightly higher tFDR values compared with the Lasso with analytic FDR control; TPR2 of the permutation Lasso was similar to that of SMA-BH control and higher than that of the Lasso with analytic FDR control (smallest p-value for the difference equal to 0.01). However, the TPR1 of the permutation Lasso was lower than that of the analytic Lasso (EN50) with a p-value of 0.03 (0.003) for the difference. The results for Lasso with permutation FDR were stable across different sets of permutations (100 to 400 permutations).

Single Chromosome Simulation with Eight QTLs: PR with FDR Control

The results for this scenario are presented in Figure 4 and Tables S5 and S6 in Supplement A. Most of the differences in empirical FDR 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 TRP1 and TRP2 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 (λ_2= 3, 10) had lower TPR1 (p-values 1.3×10^{-14} and 0.003, respectively) and TPR2 than the Lasso (p-values 2×10^{-15} to 0.03 and 0.003

to 0.33, respectively, with the upper end of each range representing TPR2 at T = 0.5). All PR methods were able to control the FDR based on their average tFDR values. Power was expectedly higher relative to the 2-QTL scenario (for EN50, TPR2 at T= 0.5 increased from 0.43 to 0.73). Comparing the EN50 with SMA-BH, the average TPR1/2 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 versus analytic FDR estimation, the permutation Lasso had lower TPR1 and TPR2 with p-values for the differences ranging from 6×10^{-17} to 0.0006. This surprising finding appears to be a problem of the permutation method; one possible explanation is that in contrast with the permuted data, in the real data (with eight QTLs) an effect needs to be detected in the presence of multiple other true effects which may require a smaller λ for detection.

The eight-QTL simulation included two groups of two QTLs each with within-group $r^2 \approx 0.5$ (absolute correlation among SNP doses of 0.71). Because of the Lasso's tendency to only select a single SNP per group and the EN's ability to deal with co-linearity and potential co-selection 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 0 SNPs in 22%, 1 SNP in 43%, 2 SNPs in 26% and >2 SNPs in 9% of cases. By comparison, the EN selected 0 SNPs in 19%, 1 SNP in 37%, 2 SNPs in 31% and >2 SNPs in 13% of cases. Expectedly the EN selected 2 and >2 SNPs more frequently but the difference was not substantial.

Single Chromosome Simulation with Two/Eight QTLs: 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 to 5 and range 0 to 39) or $|r| < 0.50$ (median number of SNP ‘neighbors’ equal to 14 with IQR 6 to 25 and range 0 to 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 which were randomly selected from the 200 replicates. For the chromosome 21 data with two isolated QTLs, the LD2lasso achieved higher power than Lasso and EN50 while maintaining FDR control (Figure 5 and Table S7 in Supplement A). The largest difference was the increase in TPR1 from 0.35 for the Lasso to 0.5 for the best LD2lasso analysis with $\varphi = 0.5$ and $h(r) = r^2$ (p-value < 0.05), while all other differences in TPR1/2 were smaller.

For the chromosome 21 data with eight QTLs (Figure 6 and Table S8 in Supplement A), 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 below 4×10^{-4}). We recall that the eight-QTL data contained a group of four QTLs with weak pairwise LD ($0.01 \leq r^2 \leq 0.1$); the high tFDR values (0.2 to 0.26) for LD2lasso were mostly due to SNPs in LD with the group of four causal SNPs at small absolute correlation values. In terms of TPR, the difference in TPR1 between the best LD2lasso analysis with $\varphi = 0.5$, $h(r) = r^2$ 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 above 0.05 due to smaller differences (less than half) and the small number of replicates.

Single Chromosome Simulation with Eight QTLs: PR versus SMA and CAR

Because all previous comparisons focused on differences among methods in empirical TPR and tFDR for a single FDR cut-off of 0.05, here we present figures depicting TPR1 versus tFDR at T = 0.25 and T = 0.50 (Figure 7) and TPR2 (T = 0.50) versus tFDR at T = 0.25 and T = 0.50 (Figure 8). These figures were generated by varying the p-value cut-off 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 LD2lassso and EN50, with Lasso and MCP (λ_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 above 0.05.

In terms of power to detect QTLs 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 for higher tFDR(T=0.25) values SMA 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 QTLs: Previous Approaches

The multi-split analysis of Meinshausen *et al.* (2009) was evaluated based on a random subset of 20 replicates of the chromosome 21 2-QTL data and on another random subset of 20 replicates of the chromosome 21 8-QTL data (Table S9 in Supplement A). 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 below 0.18 for the two-QTL data and below 0.21 for the eight-QTL data. P-values for the differences in TPR between the multi-split 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 Supplement A). R code for executing this method was generously provided to us by the first author. For our data, the local FDR Adaptive Lasso controlled the FDR at the local FDR threshold of 0.1 (tFDR values below 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 above 0.05) but was controlled for the eight-QTL data (tFDR values below 0.05) where the TPR values however remained significantly below the values of the Lasso.

PUMA was run on the eight-QTL data with 200 replicates, choosing the 2D-MCP method, omitting the SNP pre-selection step and otherwise using default parameter values. The results are in Table S11 (Supplement A) for four different p-value thresholds for SNP selection ranging from 1×10^{-07} 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 recommended a p-value threshold of 1×10^{-07} for 2D-MCP, which controlled the FDR in terms of the tFDR values in Table S11; the next lower threshold of 1×10^{-06} had tFDR values above 0.06 whose one-sided test for exceeding level 0.05 had p-values from 0.055 to 0.11. Overall, TPR1/2 values for 2D-MCP were significantly below those for MCP with analytic FDR control (p-values for differences from 9×10^{-05} to 8×10^{-29} for 2D-MCP SNP selection threshold 1×10^{-07}, and from 0.032 to 2×10^{-22} for 2D-MCP SNP selection threshold 1×10^{-06}). To explain these differences, we looked at the values for λ_2 for the models chosen in PUMA by AIC over the 200 datasets, 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 TPR1/2 (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 versus 0.310) This improvement still underperformed the results in Table S6 because PUMA uses an (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.

Analyses of Multiple Chromosomes

Here we analyzed the simulated data including chromosomes 19 (no QTL), 21 (8 QTLs) and 22 (2 QTLs) with 200 replicates. For single marker analysis (SMA), the average tFDR values for pooled and separate BH analyses and for the local FDR grouping procedure were all above 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 Supplement A).

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 Supplement A). We therefore compared the separate and pooled analyses for individual chromosomes. From Table S13, we can see that for chromosome 21 (with eight QTLs), TPR1/2 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 QTLs), in contrast, TPR1/2 increased significantly as expected from separate to pooled analysis with all p-values below 4×10^{-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 Supplement B 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), SMA with BH FDR control (SMA-BH), and Elastic Net with 50% Lasso penalty

(EN50). 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 which 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 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 about 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 re-ran 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 QTLs were identified with the stringent SMA-GWT, while SMA-BH and EN50-sep identified the same QTLs (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 five EN50-sep SNPs, and for chromosome 19, the three SMA-BH SNPs were all in LD with the one 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.

Lastly, the joint analysis of all chromosomes by EN50 with FDR control (EN50-joint), compared to the separated analyses (EN50-sep), produced fewer significant findings (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 which 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-value) being much smaller than those of all other identified (by any method) SNPs (Figure 10). The EN50-joint SNP on chromosome 3 however was in LD (absolute correlation above 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) 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. Somewhat surprisingly, the permutation-based method which we had originally designed as a check on the analytic approximation, did not perform well for some data scenarios, while the analytic method performed consistently well although being somewhat conservative. 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 model and SNP selection using PR for GWAS, and this approach should be appealing to practitioners who desire a measure of error, in particularly in terms of FDR, associated with the selected SNPs. Lastly we note that our simple analytic FDR method can also be applied to penalized logistic regression.

While SMA is justifiably criticized on theoretical grounds resulting from its use of an incorrect statistical model, it is not easy to develop a multi- or all-marker method that consistently outperforms SMA, when the focus is purely 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, 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 datasets for joint (pooled) analysis of all chromosomes, pre-screening 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 pre-screening is performed, then for tuning parameter value selection by our analytic FDR method the number of markers (p) in [15] should be set to the total number of markers (prior to pre-selection) to maintain control of the FDR (which should produce the same result as what would be obtained without pre-selection). It would however be prudent to perform PR with FDR control both to the entire set of SNPs from all chromosomes (pooled analysis) as well as separately to 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.

Lastly, code used for data simulation and code for EN50 analysis with analytic FDR control are provided in Supplement C. An example simulated dataset with eight QTLs on chromosome 21 is available at Dryad (*to be uploaded after conditional acceptance*).

References

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. Automatic Control, IEEE Transactions on **19:** 716-723.

Akaike, H., 1977 On entropy maximization principle. Application of statistics.

Alexander, D. H., and K. Lange, 2011 Stability selection for genome-wide association. Genetic epidemiology **35:** 722-728.

Ayers, K. L., and H. J. Cordell, 2010 SNP Selection in genome-wide and candidate gene studies via penalized logistic regression. Genetic epidemiology **34:** 879-891.

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

Benjamini, Y., and D. Yekutieli, 2001 The control of the false discovery rate in multiple testing under dependency. Annals of statistics**:** 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. The annals of applied statistics **5:** 232.

Breheny, P. J., 2009 Regularized methods for high-dimensional and bi-level variable selection. Theses and Dissertations**:** 325.

Cai, T. T., and W. Sun, 2009 Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. Journal of the American Statistical Association **104**.

Carbonetto, P., and M. Stephens, 2012 Integrated analysis of variants and pathways in genome-wide association studies using polygenic models of disease. arXiv preprint arXiv:1208.4400.

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 preprint arXiv:1005.3579.

Dudbridge, F., and A. Gusnanto, 2008 Estimation of significance thresholds for genomewide association scans. Genetic epidemiology **32:** 227-234.

Efron, B., 2005 *Local false discovery rates*. Division of Biostatistics, Stanford University.

Efron, B., 2008 Simultaneous inference: When should hypothesis testing problems be combined? The annals of applied statistics**:** 197-223.

Efron, B., and R. Tibshirani, 2002 Empirical Bayes methods and false discovery rates for microarrays. Genetic epidemiology **23:** 70-86.

Fan, J., and R. Li, 2001 Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association **96:** 1348-1360.

Fan, J., and H. Peng, 2004 Nonconcave penalized likelihood with a diverging number of parameters. The annals of statistics **32:** 928-961.

Friedman, J., T. Hastie, H. Höfling and R. Tibshirani, 2007 Pathwise coordinate optimization. The annals of applied statistics **1:** 302-332.

Friedman, J., T. Hastie and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. Journal of statistical software **33:** 1.

Gauderman, W. J. and J. Morrison, 2001 QUANTO documentation (Technical report no. 157). Department of Preventive Medicine, University of Southern California.

Hoffman, G.E., B.A. Logsdon and J.G. Mezey (2013) PUMA: A Unified Framework for Penalized Multiple Regression Analysis of GWAS Data. PLoS Computational Biology 9(6): e1003101.

Hu, J. X., H. Zhao and H. H. Zhou, 2010 False discovery rate control with groups. Journal of the American Statistical Association **105**.

Huang, J., S. Ma and C.-H. Zhang, 2008 Adaptive Lasso for sparse high-dimensional regression models. Statistica Sinica **18:** 1603.

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. Journal of the American Statistical Association **102:** 495-506.

Kim, S., and E. P. Xing, 2009 Statistical estimation of correlated genome associations to a quantitative trait network. PLoS genetics **5:** e1000587.

Kim, Y., H. Choi and H.-S. Oh, 2008 Smoothly clipped absolute deviation on high dimensions. Journal of the American Statistical Association **103:** 1665-1673.

Kohavi, R., 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection, pp. 1137-1145 in *IJCAI*.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet **22**:139–144.

Langaas, M., B. H. Lindqvist and E. Ferkingstad, 2005 Estimating the proportion of true null hypotheses, with application to DNA microarray data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67:** 555-572.

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. The Annals of Applied Statistics **4:** 1498-1516.

Liu, H., K. Roeder and L. Wasserman, 2010 Stability approach to regularization selection (stars) for high dimensional graphical models. arXiv preprint arXiv:1006.3316.

Liu, J., K. Wang, S. Ma and J. Huang, 2011 Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method, pp. Technical report, Department of Statistics and Actuarial Science, University of Iowa.

Marttinen, P., J. Gillberg, A. Havulinna, J. Corander and S. Kaski, 2012 Genome-wide association studies with high-dimensional phenotypes. Statistical applications in genetics and molecular biology**:** 1-19.

Meinshausen, N., and P. Bühlmann, 2010 Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72:** 417-473.

Meinshausen, N., L. Meier and P. Bühlmann, 2009 P-values for high-dimensional regression. Journal of the American Statistical Association **104**.

Pritchard, J.K. and M. Przeworski, 2001 Linkage disequilibrium in humans: Models and Data. American Journal of Human Genetics 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.

Schwarz, G., 1978 Estimating the dimension of a model. The annals of statistics **6:** 461-464.

Shah, R. D., and R. J. Samworth, 2013 Variable selection with error control: another look at stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75:** 55-80.

Smyth, G. K., 2005 Limma: linear models for microarray data, pp. 397--420.

Storey, J. D., J. M. Akey and L. Kruglyak, 2005 Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biology **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. Journal of the American Statistical Association **102:** 901-912.

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.

Sun, W., and T. Tony Cai, 2009 Large-scale multiple testing under dependence. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **71:** 393-424.

The International HapMap, C., 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological)**:** 267-288.

Tseng, P., 1988 *Coordinate ascent for maximizing nondifferentiable concave functions*. Massachusetts Institute of Technology, Laboratory for Information and Decision Systems.

Tseng, P., 2001 Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of optimization theory and applications **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. Frontiers in Genetics **4**: Article 270.

Wang, H., R. Li and C.-L. Tsai, 2007 Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika **94:** 553-568.

Wasserman, L., and K. Roeder, 2009 High dimensional variable selection. Annals of statistics **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. Journal of the American Statistical Association **93:** 120-131.

Zhang, C.-H., 2010 Nearly unbiased variable selection under minimax concave penalty. The annals of statistics **38:** 894-942.

Zou, H., 2006 The adaptive lasso and its oracle properties. Journal of the American Statistical Association **101:** 1418-1429.

Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67:** 301-320.

Zou, H., T. Hastie and R. Tibshirani, 2007 On the “degrees of freedom” of the lasso. The annals of statistics **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.

Figure Legends

Figure 1: Single Marker Analysis (SMA) of chromosome 21 data with two isolated QTLs. 200 replicates, sample size N = 201. Empirical power (defined as True Positive Rate TPR) on the left: Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTLs), and light blue represents TPR2 (TPR also including significant SNPs linked to QTLs 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 and 0.5 on the right: red horizontal line represents the tFDR value of 0.05. CWT: Chromosome-Wide Threshold which represents the p-value threshold $5.5\times10^{-8}\times3849034/50165$; GWT: Genome-Wide Threshold which 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).

Figure 2: Single Marker Analysis (SMA) of chromosome 21 data with eight QTLs. 200 replicates, sample size N = 201. Empirical power (defined as True Positive Rate TPR) on the left: Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTLs), and light blue represents TPR2 (TPR also including significant SNPs linked to QTLs 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 and 0.5 on the right: red horizontal line represents the tFDR value of 0.05. CWT: Chromosome-Wide Threshold which represents the p-value threshold $5.5\times10^{-8}\times3849034/50165$; GWT: Genome-Wide Threshold which 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).

Figure 3: Penalized Regression (PR) analyses of chromosome 21 data with two isolated QTLs. 200 replicates, sample size N = 201. Empirical power (defined as True Positive Rate TPR) on the top: Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTLs), and light blue represents TPR2 (TPR also including significant SNPs linked to QTLs 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 and 0.5 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, Ridge Regression with CV) produced the same result; EN_Lambda2 = x: Elastic Net with Lasso portion of λ_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 λ_2 = x (x = 3, 10, 30, 100).

Figure 4: Penalized Regression (PR) analyses of chromosome 21 data with eight QTLs. 200 replicates, sample size N = 201. Empirical power (defined as True Positive Rate TPR) on the top: Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTLs), and light blue represents TPR2 (TPR also including significant SNPs linked to QTLs 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 and 0.5 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, Ridge Regression with CV) produced the same result; EN_Lambda2 = x: Elastic Net with Lasso portion of λ_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 λ_2 = x (x = 3, 10, 30, 100).

Figure 5: Empirical power (True Positive Rate TPR) and thresholded False Discovery Rate (tFDR) for LD2lasso versus Lasso and Elastic Net (EN) Penalized Regression (PR) analyses of chromosome 21 data with two isolated QTLs. 20 randomly selected replicates, sample size N = 201. Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTLs), and light blue represents TPR2 (TPR also including significant SNPs linked to QTLs at absolute correlation threshold 0.5, 0.7 or 0.9). The empirical tFDR was computed at three absolute correlation thresholds between causal and linked SNPs (T = 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 λ_2 = 0.5; LD2lasso_0.9_r^2 (LD2lasso_0.9_|r|) represents the LD2lasso with weight on the Lasso penalty equal to φ = 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.

Figure 6: Empirical power (True Positive Rate TPR) and thresholded False Discovery Rate (tFDR) for LD2lasso versus Lasso and Elastic Net (EN) Penalized Regression (PR) analyses of chromosome 21 data with eight QTLs. 20 randomly selected replicates, sample size N = 201. Dark blue represents TPR1 as defined in the main text (TPR including only the significant QTLs), and light blue represents TPR2 (TPR also including significant SNPs linked to QTLs at absolute correlation threshold 0.5, 0.7 or 0.9). The empirical tFDR was computed at three absolute correlation thresholds between causal and linked SNPs (T = 0.25, 0.30, 0.50). The red horizontal line represents the tFDR value of 0.05. EN_Lambda2 = 0.5 denotes the Elastic Net with Lasso portion of λ_2 = 0.5; LD2lasso_0.9_r^2 (LD2lasso_0.9_|r|) represents the LD2lasso with weight on the Lasso penalty equal to φ = 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.

Figure 7: Plot of True Positive Rate TPR1 versus thresholded False Discovery Rate (tFDR): TPR1 versus tFDR (T=0.25) and TPR1 versus tFDR (T=0.50). TPR1 and tFDR are defined in the main text. MCP was run with λ_2 = 10. CAR represents correlation-adjusted marginal correlations.

Figure 8: Plot of True Positive Rate TPR2 versus thresholded False Discovery Rate (tFDR): TPR2 (T=0.50) versus tFDR (T=0.25) and TPR2 (T = 0.50) versus tFDR (T=0.50). TPR2 and tFDR are defined in the main text. MCP was run with λ_2 = 10. CAR represents correlation-adjusted marginal correlations.

Figure 9: Plot of – $\log_{10}$(p-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 grey 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.

Figure 10: Plot of – $\log_{10}$(p-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, respectively. The y-axis represents raw p-values from SMA.

Figures

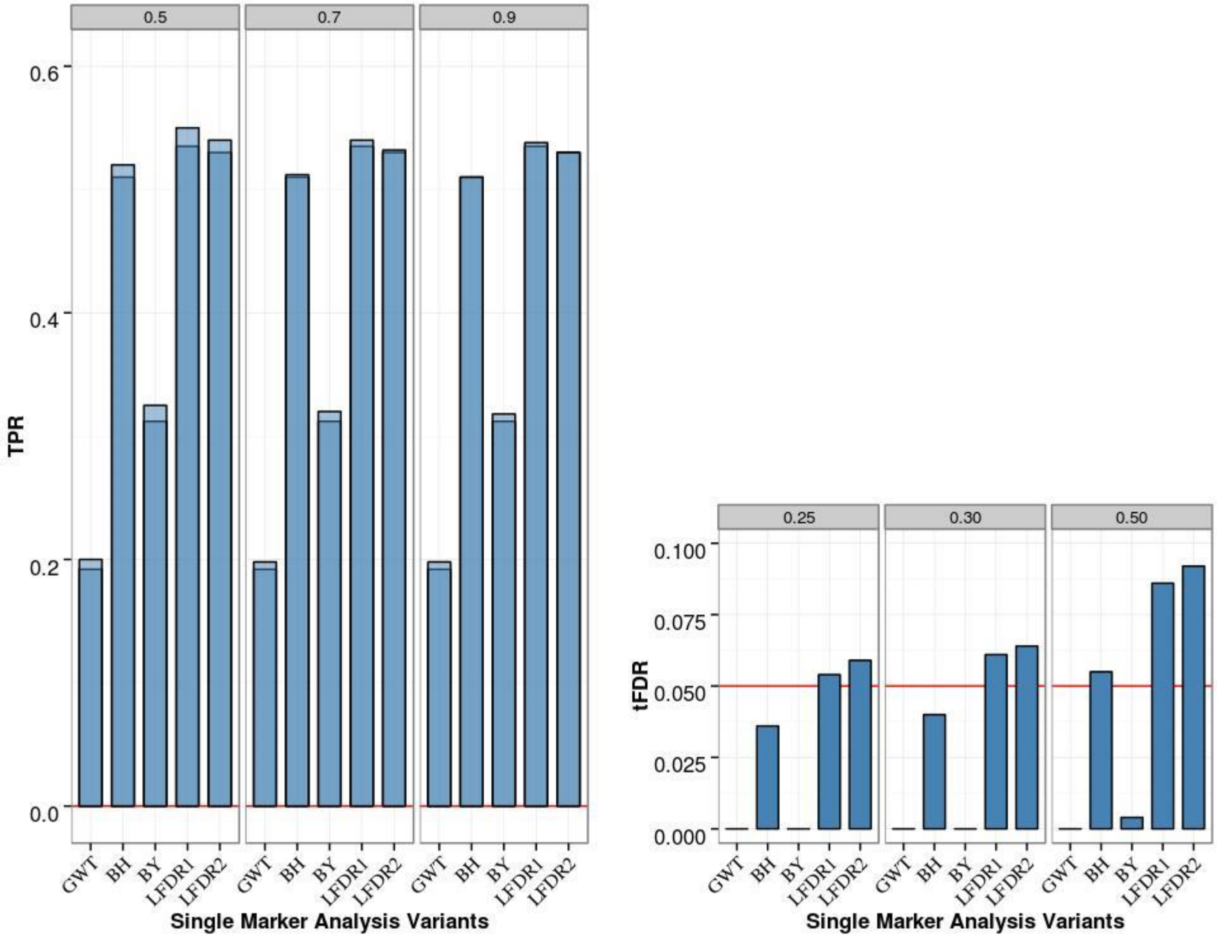


Figure 1. Single Marker Analysis (SMA) of chromosome 21 data with two isolated QTLs

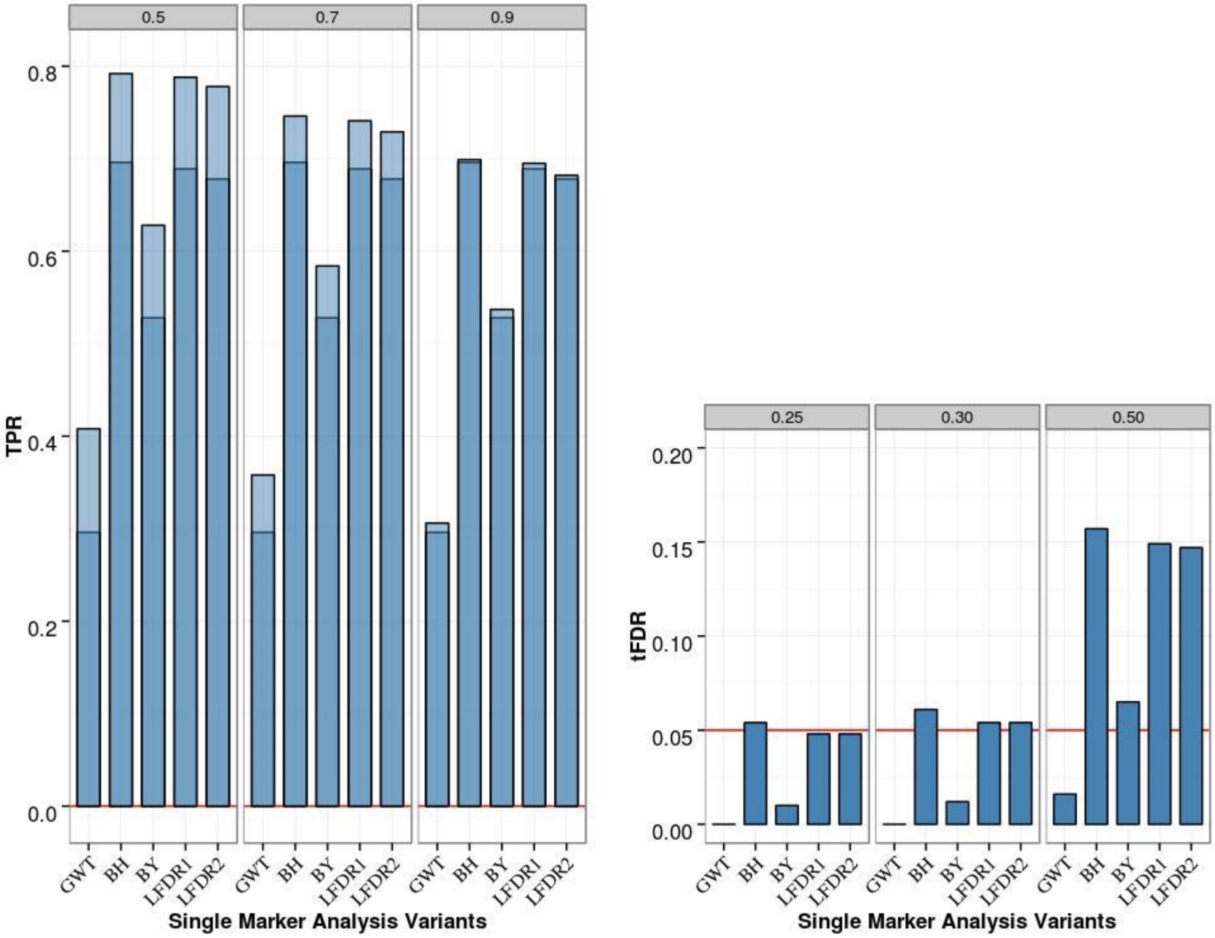


Figure 2. Single Marker Analysis (SMA) of chromosome 21 data with eight QTLs

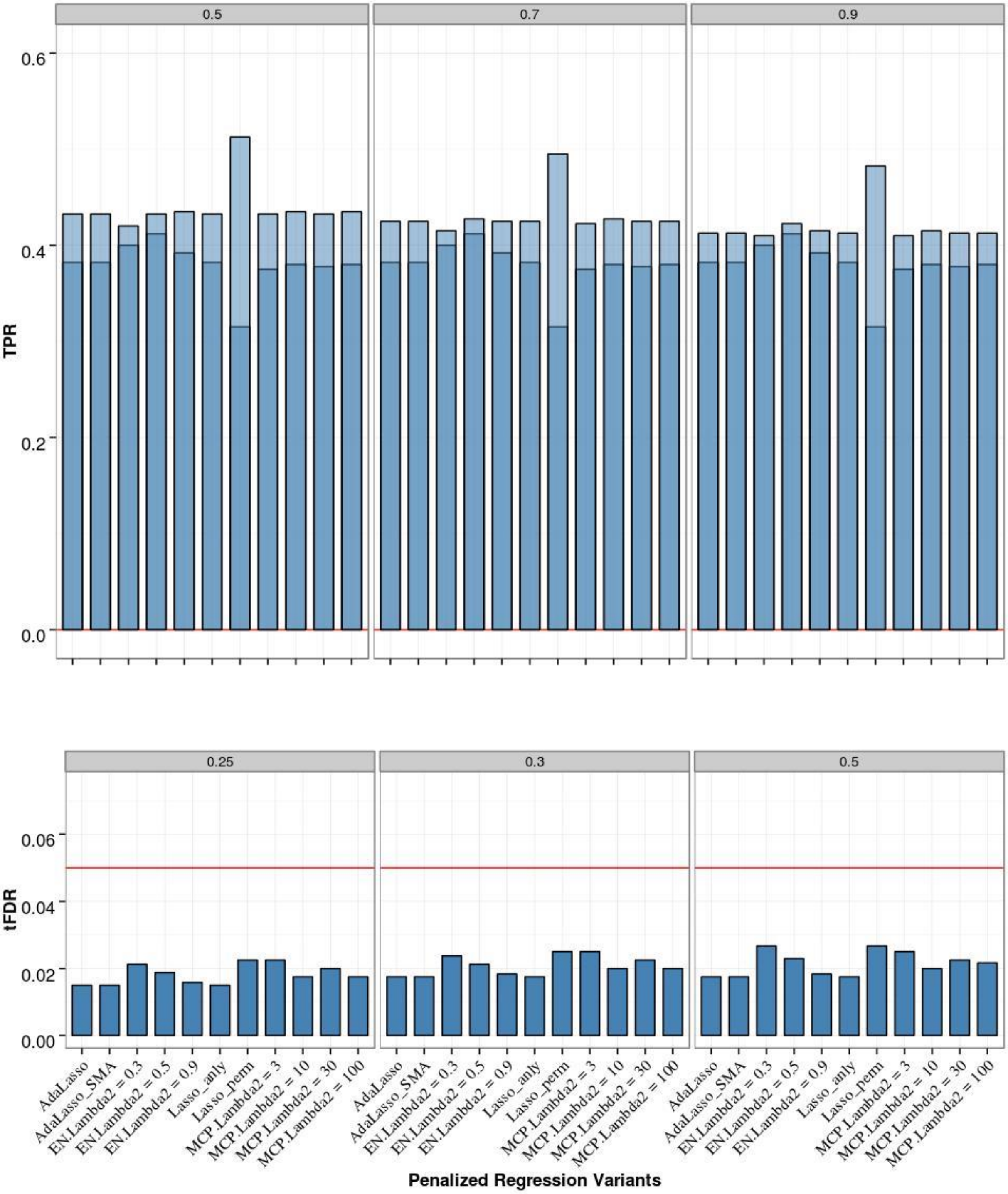


Figure 3. Penalized Regression (PR) analyses of chromosome 21 data with two isolated QTLs

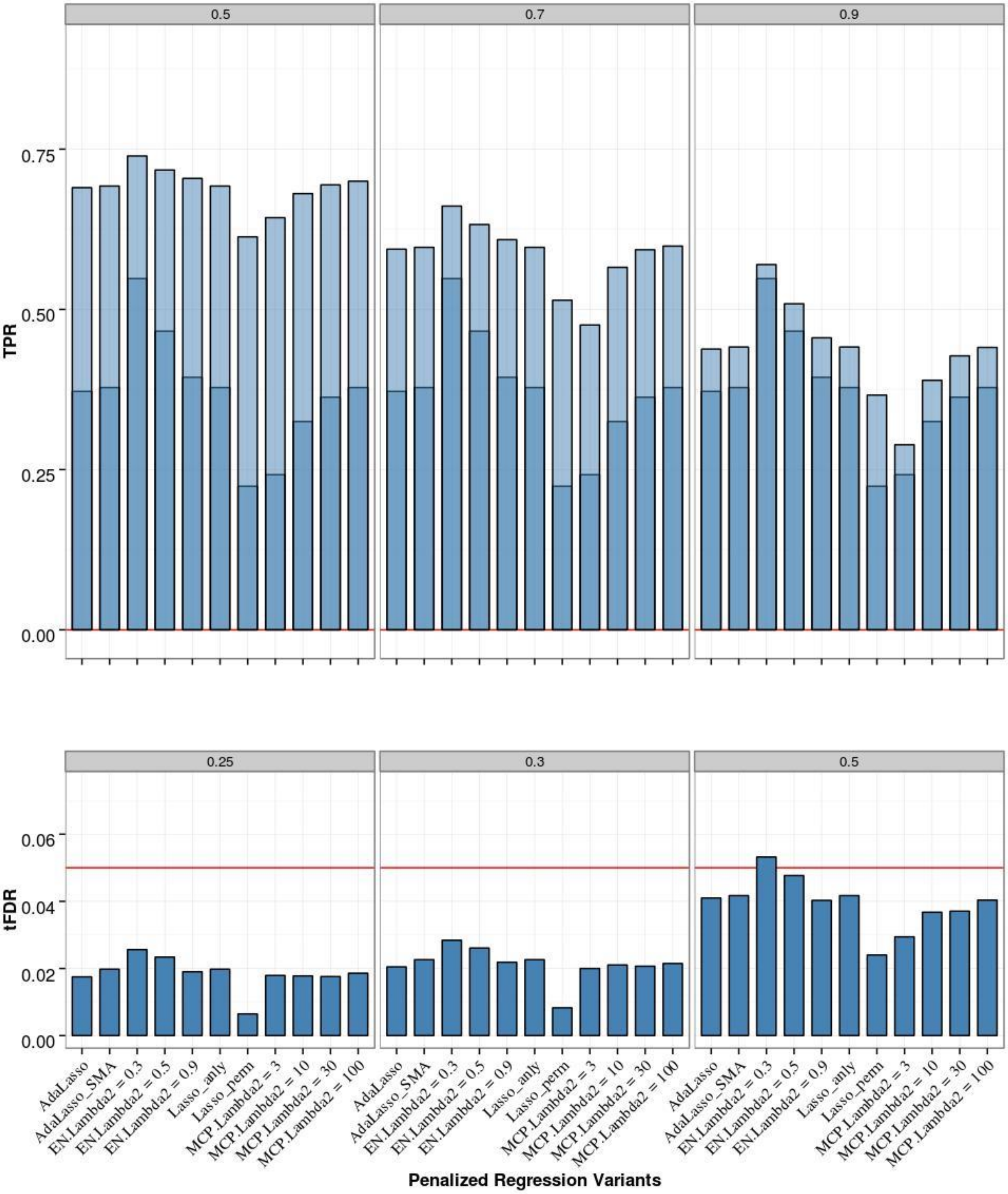


Figure 4. Penalized Regression (PR) analyses of chromosome 21 data with eight QTLs

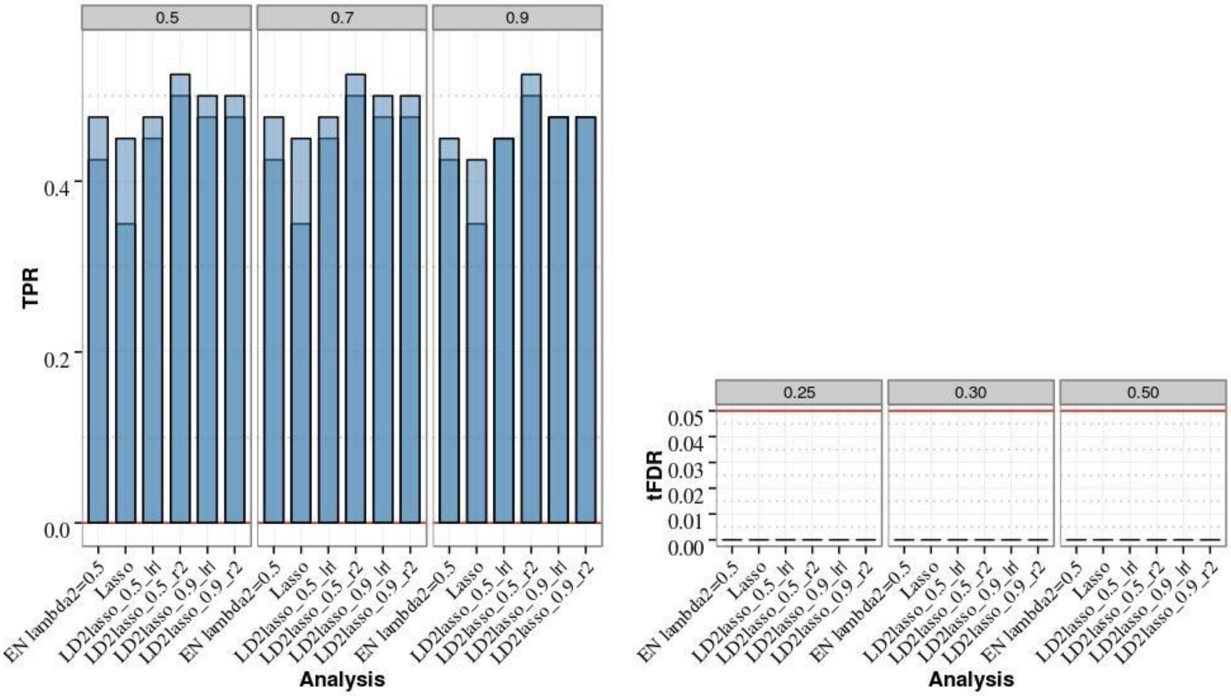


Figure 5. Empirical power (True Positive Rate TPR) and thresholded False Discovery Rate (tFDR) for LD2lasso versus Lasso and Elastic Net (EN) Penalized Regression (PR) analyses of chromosome 21 data with two isolated QTLs

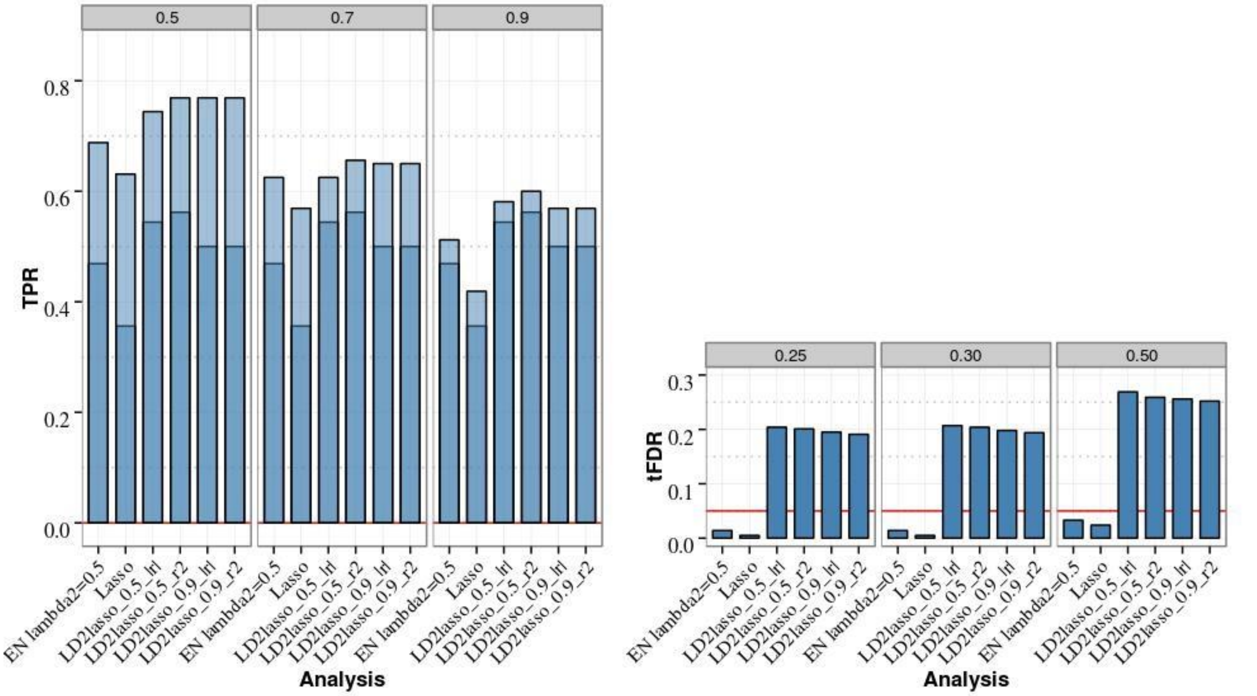
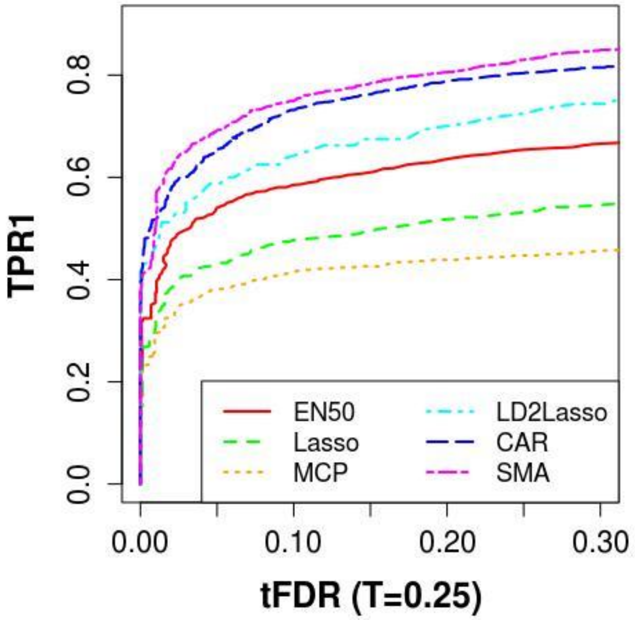


Figure 6. Empirical power (True Positive Rate TPR) and thresholded False Discovery Rate (tFDR) for LD2lasso versus Lasso and Elastic Net (EN) Penalized Regression (PR) analyses of chromosome 21 data with eight QTLs



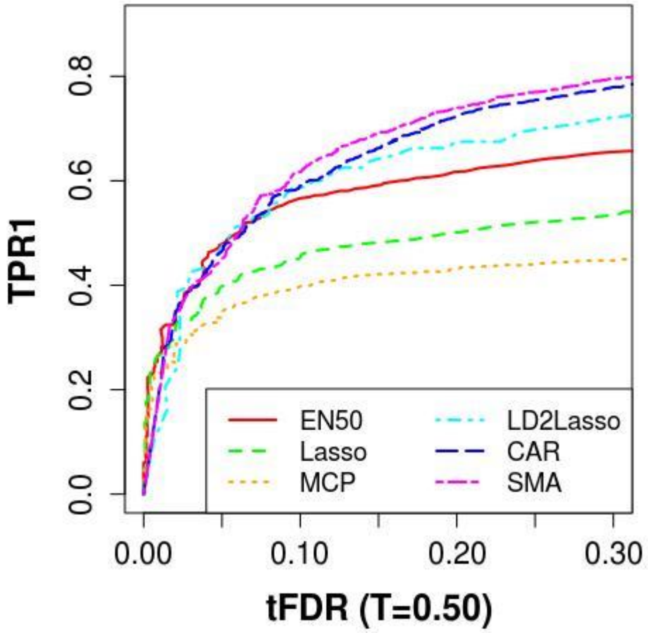
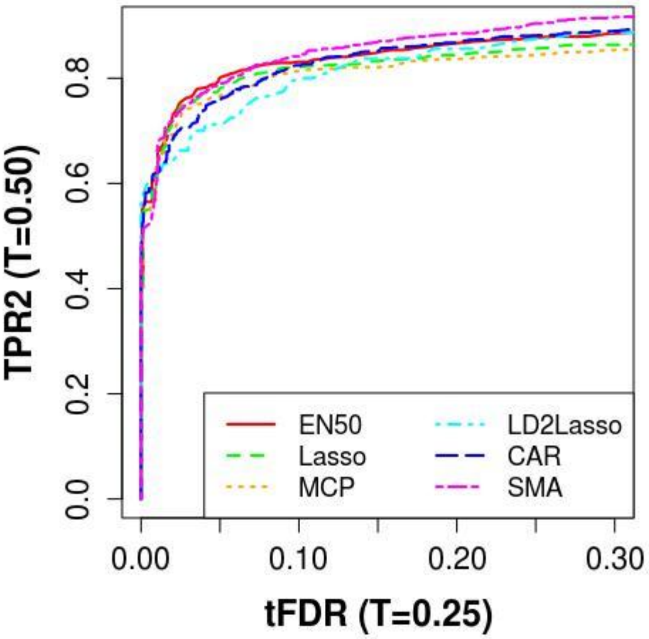


Figure 7. Plot of True Positive Rate TPR1 versus thresholded False Discovery Rate (tFDR)



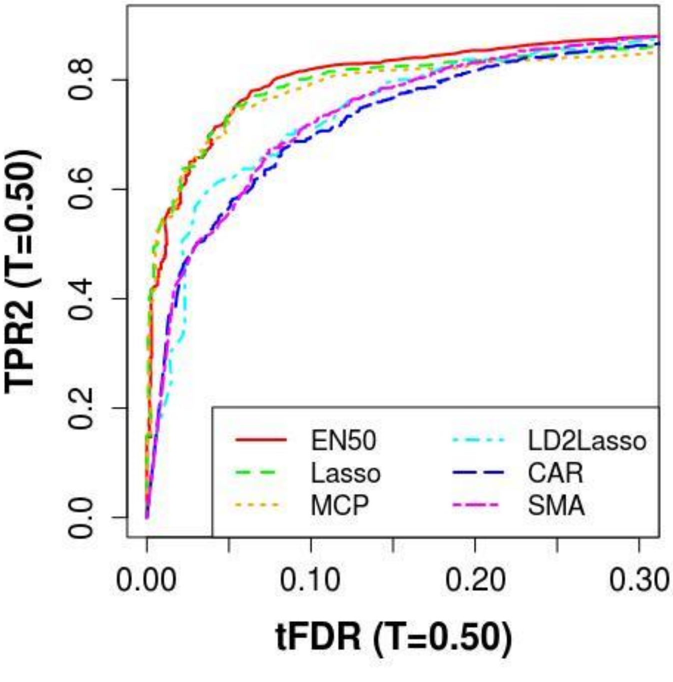


Figure 8. Plot of True Positive Rate TPR2 versus thresholded False Discovery Rate (tFDR)

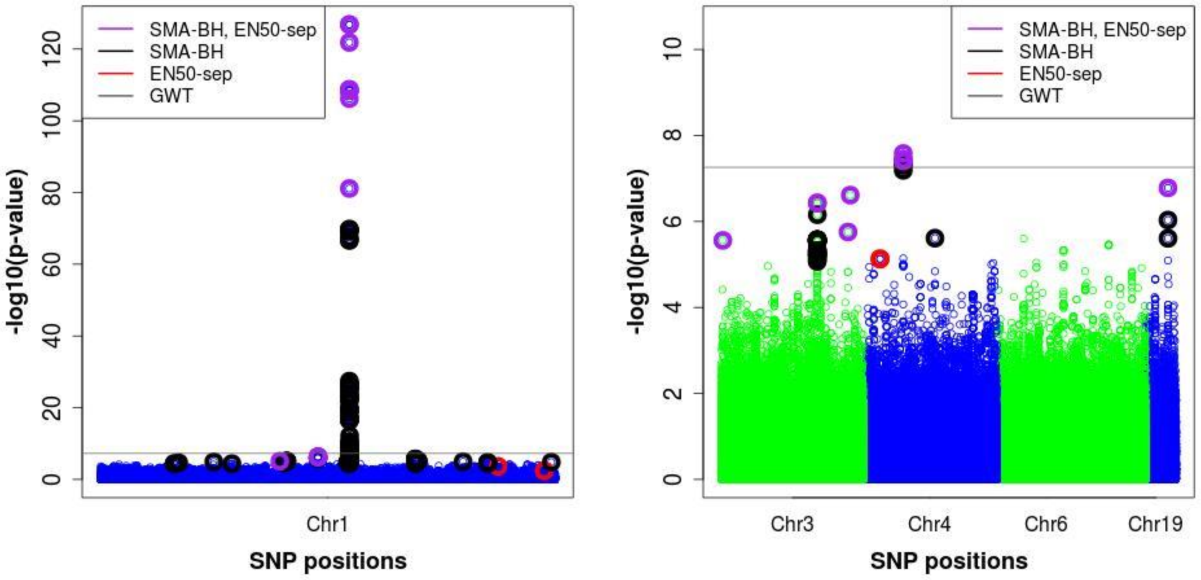


Figure 9. Plot of – $\log_{10}$(p-value) for SNPs on chromosomes 1 (left), 3, 4, 6, and 19 (right) for the real data

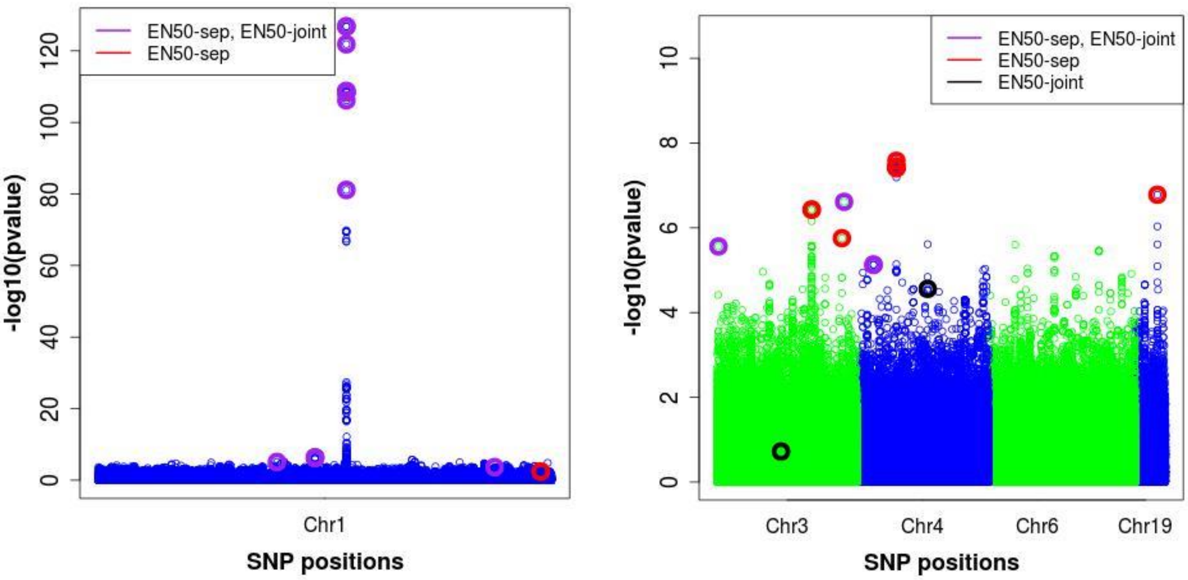


Figure 10. Plot of -$\log_{10}$(p-value) for SNPs on chromosomes 1 (left), 3, 4, 6, and 19 (right) for the real data

Tables

Table 1: Analysis of real data with Single Marker Analysis using the Genome-Wide Threshold p-value < 5.5×10^{-8} (SMA-GWT) or using Benjamini-Hochberg FDR control (SMA-BH) applied separately to the data on each chromosome, Elastic Net penalized regression with analytic FDR control applied separately to each chromosome (EN50s) or jointly across all five chromosomes (EN50j). No. sig denotes number of significant SNPs, ∩ denotes overlap with. The three 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 at or above 0.8 with a significant EN50s SNP / correlated at or above 0.5 with a significant EN50s SNP/ correlated at or above 0.25 with a significant EN50s SNP.

Chromosome		**1**	**3**	**4**	**6**	**19**
SMA-GWT	**No. sig**	74	0	8	0	0
	∩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
	∩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
	∩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
	∩SMA-GWT	9/9/9/9	0/0/0/0	6/6/6/6	0/0/0/0	0/0/0/0
	∩SMA-BH	13/13/13/14	5/5/5/5	6/6/6/6	0/0/0/0	1/1/1/1

Supplementary Files A

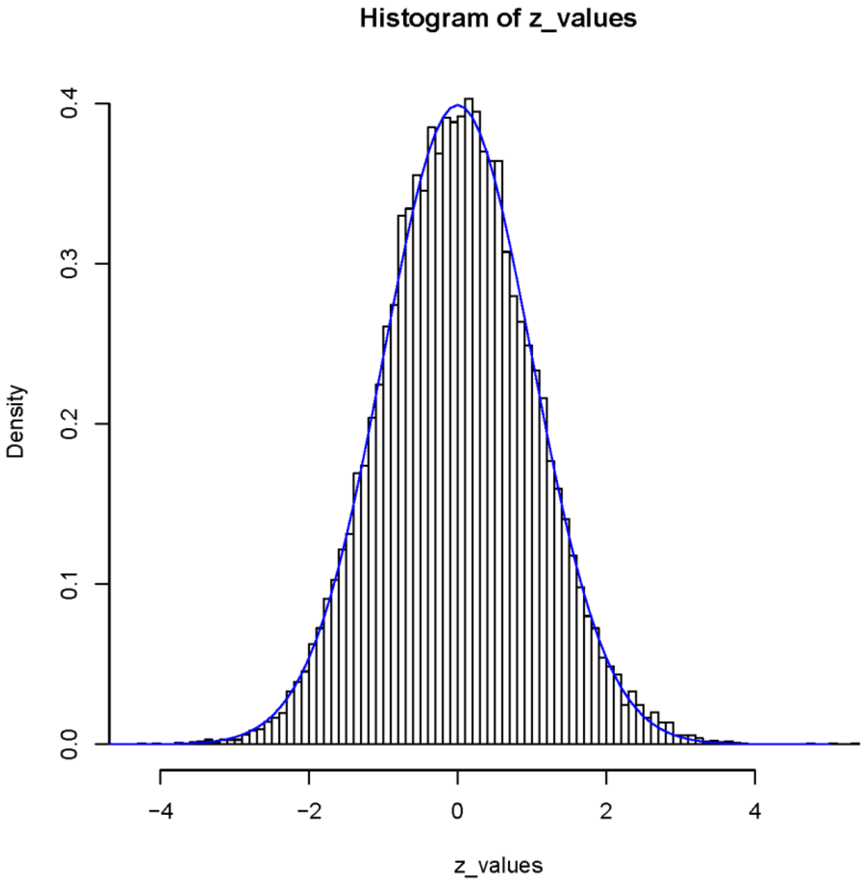


Figure S1: Histogram of the z-values pertaining to 21,530 SNPs on chromosome 21. The blue solid curve represents the N(0,1) distribution, which fits the empirical z-value distribution well.

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 $5.5\times10^{-8}\times3849034/50165$; 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 $\geq$ 1 false positives); tFDR is an empirical thresholded false discovery rate mean(nFP/nSig) with nFP (nSig) 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 = nCTP/Q is the first true positive rate with nCTP denoting the number of true QTLs identified and Q denoting the number of true QTL (2*200) over all replicates; TPR2 = nCLTP/Q is the second true positive rate with nCLTP 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 (λ_2 = 0.3)**	**EN (λ_2 = 0.5)**	**EN (λ_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		0.315 ± 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 (λ_2 = 3)**	**MCP (λ_2 = 10)**	**MCP (λ_2 = 30)**	**MCP (λ_2 = 100)**
tFDR	**0.25**	0.022 ± 0.010	0.018 ± 0.008	0.020 ± 0.009	0.018 ± 0.008
	0.3	0.025 ± 0.010	0.020 ± 0.009	0.022 ± 0.009	0.020 ± 0.009
	0.5	0.025 ± 0.010	0.020 ± 0.009	0.022 ± 0.009	0.022 ± 0.009
TPR1		0.375 ± 0.025	0.380 ± 0.025	0.378 ± 0.025	0.380 ± 0.025
TPR2	**0.5**	0.432 ± 0.027	0.435 ± 0.027	0.432 ± 0.027	0.435 ± 0.027
	0.7	0.422 ± 0.027	0.428 ± 0.027	0.425 ± 0.027	0.425 ± 0.027
	0.9	0.410 ± 0.027	0.415 ± 0.027	0.412 ± 0.027	0.412 ± 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 (λ_2 = 0.3)**	**EN (λ_2 = 0.5)**	**EN (λ_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		0.224 ± 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** **(λ_2 = 3)**	**MCP** **(λ_2 = 10)**	**MCP** **(λ_2 = 30)**	**MCP** **(λ_2 = 100)**
tFDR	**0.25**	0.018 ± 0.006	0.018 ± 0.005	0.018 ± 0.004	0.019 ± 0.005
	0.3	0.020 ± 0.007	0.021 ± 0.005	0.021 ± 0.005	0.021 ± 0.005
	0.5	0.029 ± 0.007	0.037 ± 0.006	0.037 ± 0.006	0.040 ± 0.006
TPR1		0.242 ± 0.010	0.325 ± 0.013	0.363 ± 0.014	0.378 ± 0.014
TPR2	**0.5**	0.643 ± 0.018	0.681 ± 0.018	0.694 ± 0.018	0.700 ± 0.018
	0.7	0.476 ± 0.016	0.566 ± 0.017	0.593 ± 0.018	0.599 ± 0.018
	0.9	0.289 ± 0.011	0.389 ± 0.014	0.428 ± 0.015	0.441 ± 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 (λ_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 (λ_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	0.475 ± 0.098
	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	0.475 ± 0.098
	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 (λ_2 = 0.5)**	**LD2lasso** ***φ = 0.9*** ***$h(r) = r^2$*** ***$\|r\|$ = 0.85***	**LD2lasso** ***φ = 0.5*** ***$h(r) = r^2$*** ***$\|r\|$ = 0.85***	**LD2lasso** ***φ = 0.9*** ***$h(r) = \|r\|$*** ***$\|r\|$ = 0.85***	**LD2lasso** ***φ = 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** ***φ = 0.9,*** ***$h(r) = r^2$, $\|r\|$ = 0.50***	**LD2lasso** ***φ = 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.

		Two QTLs			**Eight QTLs**		
Criterion	**T**	**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.

		Two QTLs			**Eight QTLs**			
Criterion	**T**	**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.

		2D MCP with different p-value thresholds				**2D MCP analytic FDR**
Criterion	**T**	**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

Supplementary Files B

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 ≥ 1%, call rate ≥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).

CHAPTER 3

SUMMARY

This dissertation makes a thorough contribution to the evaluation of Penalized Regression (PR) for Genome-Wide Association Studies (GWAS) of common SNPs. A main innovation is the combination of PR with control of the False Discovery Rate (FDR). Our results indicate that PR with FDR based selection of the penalty parameter value conservatively controls the FDR, while Single Marker Analysis (SMA) may not always achieve FDR control. FDR based analysis expectedly increases power substantially over SMA with genome-wide type-I error control. Preliminary results (not reported here) indicate that the Bayesian counterparts of PR (at least when implemented with variational algorithms) are conservative and have less power than PR with FDR control. In future research, the utility of PR with group structure and with FDR control for the identification of rare variants should be investigated. Of particular interest are methods that allow both for group selection and selection of variants within groups and methods that can fit both common and rare variants jointly (current methods for rare variants ignore or strongly down-weigh common variants). While the Group-Lasso (GL; Yuan and Lin 2007) only selects among groups by applying the lasso penalty to the L2 norm of a group, the Sparse Group Lasso adds a lasso penalty to the GL penalty (Simon et al. 2013). Alternatively, a group penalty can be applied to the L1 (instead of the L2) norm of each group, as for example in the Group Exponential Lasso (Breheny 2014). For the latter, FDR-based penalty parameter value selection can be implemented in a similar way as described in this dissertation. These methods can potentially perform variable selection both within and across groups and fit common and rare variants jointly. Therefore, they should be studied with simulated and real data and compared to current state-of-the art methods for rare variants.

Lastly, it is becoming increasingly common for researchers to collect multiple types of genome-wide omics data (transcriptomics, methylomics etc.) in addition to genetic variant data on the same samples in studies of complex human diseases. The problem now becomes one of determining associations between pairs of variable types in a larger system. Bayesian high-dimensional regression for multivariate phenotypes (Richardson et al. 2010), Lasso with fusion-penalty on the effects of the same SNP on multiple correlated expression profiles (Chen et al., 2010), Sparse Canonical Correlation analysis (Witten et al. 2009, Witten and Tibshirani 2010, Waaijenborg and Winderman 2009), Sparse Partial Least-Squares (Chun and Keles 2010), and Sparse Gaussian Graphical Models (Yin and Li 2011) have been proposed to date for these data. These methods need to be further developed and compared to each other under different well-designed scenarios similar to the method evaluation presented in the present dissertation.

REFERENCES

Bodmer, W. & Bonilla, C. 2008 Common and rare variants in multifactorial susceptibility to common diseases. Nature Genet. 40, 695–701.

Breheny, P., 2014 The group exponential lasso for bi-level variable selection. Biometrics (in revision).

Chen, X., Kim, S., Lin, Q., Carbonell, J. G., & Xing, E. P., 2010 Graph-structured multi-task regression and an efficient optimization method for general fused Lasso. arXiv preprint arXiv:1005.3579.

Chun, H., & Keleş, S., 2010 Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(1), 3-25.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Visscher, P. M., 2009 Finding the missing heritability of complex diseases. Nature, 461(7265), 747-753.

Pritchard, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. 69, 124–137.

Richardson, S., Bottolo, L., & Rosenthal, J. S., 2010 Bayesian models for sparse regression analysis of high dimensional data. Bayesian Statistics, 9, 539-569.

Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J., 2009 Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev. 19, 212–219.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R., 2013 A sparse-group lasso. Journal of Computational and Graphical Statistics, 22(2), 231-245.

Waaijenborg, S., & Zwinderman, A. H., 2009 Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks.BMC bioinformatics, 10(1), 315.

Witten, D. M., Tibshirani, R., & Hastie, T., 2009 A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, kxp008.

Witten, D. M., & Tibshirani, R. J., 2009 Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology, 8(1), 1-27.

Yin, J., & Li, H., 2011 A sparse conditional gaussian graphical model for analysis of genetical genomics data. The annals of applied statistics, 5(4), 2630.

Yuan M, Lin Y., 2006 Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol 68:49–67.