Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 4

A Comparison of Multifactor Dimensionality Reduction and L_1-Penalized Regression to Identify Gene-Gene Interactions in Genetic Association Studies

Stacey Winham* Chong Wang†

Alison A. Motsinger-Reif‡

*North Carolina State University, sjwood@ncsu.edu

†North Carolina State University, cwang19@ncsu.edu

‡North Carolina State University, alison.motsinger@gmail.com

Copyright ©2011 Berkeley Electronic Press. All rights reserved.

A Comparison of Multifactor Dimensionality Reduction and L_1-Penalized Regression to Identify Gene-Gene Interactions in Genetic Association Studies*

Stacey Winham, Chong Wang, and Alison A. Motsinger-Reif

Abstract

Recently, the amount of high-dimensional data has exploded, creating new analytical challenges for human genetics. Furthermore, much evidence suggests that common complex diseases may be due to complex etiologies such as gene-gene interactions, which are difficult to identify in high-dimensional data using traditional statistical approaches. Data-mining approaches are gaining popularity for variable selection in association studies, and one of the most commonly used methods to evaluate potential gene-gene interactions is Multifactor Dimensionality Reduction (MDR). Additionally, a number of penalized regression techniques, such as Lasso, are gaining popularity within the statistical community and are now being applied to association studies, including extensions for interactions. In this study, we compare the performance of MDR, the traditional lasso with L_1 penalty (TL1), and the group lasso for categorical data with group-wise L_1 penalty (GL1) to detect gene-gene interactions through a broad range of simulations.

We find that each method has both advantages and disadvantages, and relative performance is context dependent. TL1 frequently over-fits, identifying false positive as well as true positive loci. MDR has higher power for epistatic models that exhibit independent main effects; for both Lasso methods, main effects tend to dominate. For purely epistatic models, GL1 has the best performance for lower minor allele frequencies, but MDR performs best for higher frequencies. These results provide guidance of when each approach might be best suited for detecting and characterizing interactions with different mechanisms.

KEYWORDS: Multifactor Dimensionality Reduction (MDR), Lasso, gene-gene interactions

---

*This work was supported by Grant Number T32GM081057 from the National Institute of General Medical Sciences and the National Institute of Health.

INTRODUCTION

Fueled by rapid technological advancement, research in the area of genetic epidemiology has exploded, creating a wealth of high-dimensional data and new analytical challenges for identifying genetic risk factors for disease. In addition, much evidence suggests that common, complex diseases may be the result of a complex interplay between multiple genetic and environmental factors and that gene-gene and gene-environment interactions, or epistasis, may play an important role in the etiology of these types of diseases [Moore 2003]. Detecting these interactions in high-dimensional data is a difficult variable selection problem, which is exacerbated as the number of markers increases far beyond the sample size, as is common in studies of human genetics. Therefore as an alternative to traditional statistical methods, data-mining approaches designed to sift through large amounts of data are gaining popularity for association studies, performing variable selection and statistical modeling simultaneously [Cordell 2009].

One of the most commonly used data-mining approaches to evaluate potential gene-gene interactions is Multifactor Dimensionality Reduction (MDR), designed specifically to select potentially interacting genetic variables that are most associated with disease in case/control studies [Ritchie, et al. 2001]. In a range of simulation studies, MDR and its various extensions have displayed high power as compared to other methods [Motsinger-Reif, et al. 2008] and in the presence of genotyping error and missing data [Ritchie, et al. 2003]. MDR has been successful in identifying a number of interactions in real data applications, including multiple sclerosis [Brassat, et al. 2006; Motsinger, et al. 2007], breast cancer [Nordgard, et al. 2007], and HIV immunogenetics [Haas, et al. 2006].

Variable selection is also a hot topic in the field of statistics, and a number of more general penalized regression techniques have also emerged, like the Least Absolute Shrinkage and Selection Operator (Lasso) [Tibshirani 1996]. Lasso has exploded in popularity within the statistical community, and is now being applied for variable selection in human genetics. Penalized regression approaches have been developed for generalized linear models such as logistic regression [Park and Hastie 2007b] and for categorical data [Meier, et al. 2008; Yuan and Lin 2006]. In the area of genetics, penalized techniques have emerged to detect interactions, such as stepPLR [Park and Hastie 2008] with L_2 regularization and an adaptive group Lasso [Yang, et al.]. Most recently, extensions for Lasso have been made for GWA studies, such as a two-stage L_1 penalized approach to identify additive, dominant or recessive effects [Wu, et al. 2009] and the Screen and Clean filter approach to examine marginal and interaction effects [Wu, et al. 2010].

In this study our objective is to directly compare the performance of these two very popular data-mining techniques, MDR and Lasso L_1-penalized

regression, to detect gene-gene interactions in a case/control study with a binary disease outcome. We consider two different Lasso approaches, the traditional (ungrouped) L_1 penalty for logistic regression (TL1) and the group L_1 penalty for categorical data (GL1). MDR has previously been compared to a few penalized regression techniques including a brief, but not extensive comparison with an adaptive group Lasso algorithm [Yang, et al.]. Additionally MDR was compared with stepPLR, focusing on L_2 rather than L_1 regularization, showing context dependent results [He, et al. 2009; Park and Hastie 2008]. In stepPLR, L_2 regularization is utilized because it provides stable parameter estimates as the dimensionality increases, even if the number of variables is greater than the sample size. Unlike L_1 regularization, L_2 regularization does not achieve smoothing and variable selection simultaneously and selection must be performed in a separate step. The comparisons of this study differ from those of Park and Hastie [2008] in the use of a single stage penalization method for estimation and selection.

The implementations of our methods in the current study also differ from those of previous studies because they come from an end-users perspective, based on commonly-used implementations. Our primary goal in this study is to compare the relative performance of MDR and Lasso using easily implementable and commonly used versions that are widely available to researchers, with minimal modifications to accurately reflect typical utilization in practice. We take this approach so that analysts can see how the methods compare in detecting interactions as they would be used in real data applications with available software. MDR is implemented in C and JAVA as in the free software available from http://www.epistasis.org [Hahn, et al. 2003], and TL1 and GL1 are implemented with the freely available R software using the packages 'glmpath' and 'grplasso', respectively [Meier 2009; Park and Hastie 2007a]. Through simulation, we compare the performance of MDR, GL1, and TL1 in regards to power to identify gene-gene interactions, the number of identified loci, and true and false positive rates under a wide range of genetic models and effect sizes. We first describe the three methods considered, followed by our simulation design, and then the results for the simulation analysis.

METHODS

NOTATION

Suppose we have n i.i.d. observations $(\mathbf{x}_i , y_i)$, $i = 1, \ldots, n$ where $\mathbf{x}_i$ is a p-dimensional vector of genotype information at a total of p SNPs; for the jth SNP, $x_{ij} \in \{0, 1, 2\}$. The scalar $y_i \in \{0, 1\}$ is a binary response variable corresponding

DOI: 10.2202/1544-6115.1613

to disease status for individual i. Additionally, assume that the p SNPs are far enough apart in the genome such that they can be considered independent. Let $\mathbf{X}$ denote the $n \times p$ matrix of predictor information and $\mathbf{Y}$ denote the $n \times 1$ vector of disease status for all n individuals.

MULTIFACTOR DIMENSIONALITY REDUCTION

Multifactor dimensionality reduction (MDR) is a nonparametric data-mining tool to identify potential gene-gene interactions in case-control genetic association studies using data reduction [Ritchie, et al. 2001]. MDR reduces the full dimensionality of the data by focusing on combinations of loci that may interact, and utilizes these combinations to create a classification rule. Assume our sample consists of n_1 cases and n_0 controls where $n_1 = n_0$ for simplicity, and suppose we are considering potential interactions of size $k = 1,...,K$ loci. Let model m denote a particular combination of k loci where $m = 1,...,\binom{p}{k}$; model m will be a subset of the columns of $\mathbf{X}$ pertaining to the k loci. With 3 possible genotypes per locus, define $G_m = j$ to be genotype combination j for the loci of model m with $j = 1, \ldots, 3^k$. MDR assigns high-risk/low-risk status to the genotype combination $G_m = j$ using the following Naïve Bayes classifier H_{mj}:

$$H_{mj} = \begin{cases} 1 \text{ if } \dfrac{n_{1,G_m=j}}{n_{0,G_m=j}} > \dfrac{n_1}{n_0} \\ 0 \text{ if } \dfrac{n_{1,G_m=j}}{n_{0,G_m=j}} \leq \dfrac{n_1}{n_0} \end{cases} \qquad (1)$$

The classifier H_{mj} is an indicator variable for high-risk status for $G_m = j$. Each possible model m classifies an individual i as a case or a control based on the characterization of that individual's genotype combination as high or low-risk; for model m, $\hat{y}_i \mid (G_m = j) = H_{mj}$. For balanced data, this classification scheme maximizes the posterior probability of y_i under model m.

For a given combination of k loci, the full data is reduced from 3^k -dimensions to a single dimension with two levels, high-risk and low-risk. MDR performs this dimension reduction in an exhaustive fashion over all possible models m and selects a best model over these $\binom{p}{k}$ possible reductions by maximizing classification accuracy, the proportion of individuals correctly classified by model m; for unbalanced studies with $n_1 \neq n_0$, balanced accuracy, the

mean of sensitivity and specificity, is instead utilized [Velez, et al. 2007]. A final best model over all possible sizes of interaction $k = 1, \ldots, K$ is chosen with cross-validation [Ritchie, et al. 2001] and the statistical significance of the prediction accuracy estimate can be assessed nonparametrically with permutation testing [Motsinger-Reif 2008]. The SNPs identified in the final model m represent the active set of predictors, or the subset of loci selected by MDR. For a more detailed explanation of the method see [Hahn, et al. 2003].

TRADITIONAL L1-PENALIZED REGRESSION (LASSO)

Variable selection on a set of SNPs can also be performed under a regression framework considering both main effects and gene-gene interactions. Because our response vector $\mathbf{Y}$ is binary, this suggests the use of logistic regression. To avoid making any assumptions on the genetic mode of inheritance (similar to the nonparametric encoding of MDR), we treat the SNP variables as categorical factors, with levels {0,1,2}. Consider a basis expansion of our predictor matrix $\mathbf{X}$ such that each SNP $j = 1, \ldots, p$ is encoded as a series of indicator variables for each level using reference coding, where without loss of generality we treat genotype 0 as the reference level. This new dummy encoded $n \times (2p)$ matrix of main effects, $\mathbf{Z}$, will therefore consist of a group of (3-1)=2 main effects indicator variables (for genotypes 1 and 2) for each of the p predictor variables. We can further partition $\mathbf{Z} = \left[\mathbf{Z}_1 \cdots \mathbf{Z}_j \cdots \mathbf{Z}_p\right]$ for $j=1,\ldots,p$, where $\mathbf{Z}_j$ is an $n \times 2$ sub-matrix $\mathbf{Z}_j = \left(\mathbf{z}_{j_1}, \mathbf{z}_{j_2}\right)$ defining the jth group of indicators corresponding to the jth predictor locus; $\left(\mathbf{z}_{j_1}, \mathbf{z}_{j_2}\right)$ are indicator vectors for genotypes 1 and 2, respectively.

Additionally, consider expanding our predictor matrix to include all $q = p \cdot (p-1)/2$ pairwise interactions between SNPs. For each of q pairwise interactions we will create (3-1)*(3-1)=2^2 indicator variables (for genotype combinations 11, 12, 21, and 22), resulting in an $n \times (2^2 q)$ expanded design matrix $\mathbf{W}$. We can further partition $\mathbf{W} = \left[\mathbf{W}_1 \cdots \mathbf{W}_k \cdots \mathbf{W}_q\right]$ for $k=1,\ldots,q$, where $\mathbf{W}_k$ is an $n \times 2^2$ sub-matrix $\mathbf{W}_k = \left(\mathbf{w}_{k_{11}}, \mathbf{w}_{k_{12}}, \mathbf{w}_{k_{21}}, \mathbf{w}_{k_{22}}\right)$ defining the kth group of indicators corresponding to the j_1th and j_2th predictor loci; $\left(\mathbf{w}_{k_{11}}, \mathbf{w}_{k_{12}}, \mathbf{w}_{k_{21}}, \mathbf{w}_{k_{22}}\right)$ are indicator vectors for genotype combinations 11, 12, 21 and 22 respectively.

Our full $n \times \left(2p + 2^2 q\right)$ design matrix $\mathbf{D}$ will consist of both main effects and pair-wise interactions, partitioned for ease of notation such that $\mathbf{D} = \left[\mathbf{Z}\ \mathbf{W}\right]$ where $j = 1,\ldots,p$ refers to the collection of indices for the p groups of main effects

DOI: 10.2202/1544-6115.1613

and $k = 1,...,q$ refers to the collection of indices for the q groups of interactions. We could further consider expanding D to include higher order interactions, but for simplicity we consider only two-way interactions.

Similarly, let our $2p + 2^2 q \times 1$ parameter vector $\theta = \left[\beta\ \gamma\right]$ be partitioned into sub-vectors for main effects and interactions. Let $\beta = \left[\beta_1,...,\beta_j,...,\beta_p\right]$ be a parameter vector for main effects where $\beta_j = \left[\beta_{j_1}\ \beta_{j_2}\right]$ is a sub-vector for the *j*th SNP consisting of components for genotypes 1 and 2 respectively and let $\gamma = \left[\gamma_1,...,\gamma_k,...,\gamma_q\right]$ be a parameter vector for pair-wise interactions where $\gamma_k = \left[\gamma_{k_{11}}\ \gamma_{k_{12}}\ \gamma_{k_{21}}\ \gamma_{k_{22}}\right]$ is a sub-vector for the *k*th SNP by SNP interaction consisting of components for genotype combinations 11, 12, 21 and 22 respectively. We now define the following logistic regression model for individual *i*,

$$\log\left(\frac{p_\theta(\mathbf{d}_i)}{1-p_\theta(\mathbf{d}_i)}\right) = \theta_0 + \mathbf{d}_i^T\theta = \theta_0 + \mathbf{z}_i^T\beta + \mathbf{w}_i^T\gamma = \theta_0 + \sum_{j=1}^{p}\mathbf{z}_{i,j}^T\beta_j + \sum_{k=1}^{\mathbf{q}}\mathbf{w}_{i,k}^T\gamma_k \qquad (2)$$

where $p_\theta(\mathbf{d}_i) = \mathrm{P}(y_i = 1 \mid \mathbf{d}_i, \theta)$. This model will be used to define the usual log-likelihood for the model, $l(\theta) = \sum_{i=1}^{n} y_i \cdot \log p_\theta(\mathbf{d}_i) + (1-y_i)\cdot\log(1-p_\theta(\mathbf{d}_i))$.

Traditionally, Lasso performs variable selection and model fitting simultaneously through penalization. By penalizing the coefficients associated with each of the predictors during model fitting, the unimportant coefficients will shrink towards zero, effectively eliminating these variables from the model and achieving sparsity [Tibshirani 1996]. Specifically, Lasso fits the regression model by minimizing -$l(\theta)$ subject to a constraint equivalent to the L_1-penalty, based on the L_1-norm $\|\theta\|_1 = \sum_l |\theta_l|$. That is,

$$\hat{\theta}_\lambda = \arg\min_\theta\left(-l(\theta) + \lambda\|\theta\|_1\right) = \arg\min_\theta\left(-l(\theta) + \lambda\left(\sum_{j=1}^{p}\sum_{g=1}^{2}|\beta_{j,g}| + \sum_{k=1}^{q}\sum_{g_1=1}^{2}\sum_{g_2=1}^{2}|\gamma_{j,g_1g_2}|\right)\right) \qquad (3)$$

where λ is a tuning parameter. We refer to this traditional model for L_1-penalized logistic regression defined in equation [3] as TL1, which is an extension of the general model described in Park and Hastie 2007 [Park and Hastie 2007b], and we fit this model using the 'glmpath' package in R [Park and Hastie 2007a]. For computational simplicity, we choose the tuning parameter λ using BIC as recommended in [Zou, et al. 2007]. This model penalizes each of the indicator variables associated with either a main effect or pair-wise interaction equally, allowing each individual indicator to be set to either $\theta_l = 0$ or $\theta_l \neq 0$, for

l=1,…,$2p$+2^2q. The set $\{l : \theta_l \neq 0\}$ defines the variables in the active set, and the particular SNPs associated with {l} represent the SNPs select by TL1.

GROUP LASSO FOR LOGISTIC REGRESSION

In situations where we have categorical predictors, we may consider variable selection on the groups of dummy variables associated with particular main effects and interactions rather than the dummy variables themselves. In the traditional Lasso approach (TL1), each indicator variable (and θ_l) is treated individually, and there is no guarantee that all of the indicators corresponding to the same main effect or interaction will be either in or out of {l}. In addition, TL1 is not invariant to the particular parameterization of genotypes, such as choice of reference level, resulting in potentially different sets of factors in {l} [Yuan and Lin 2006]. To alleviate these concerns, we also employ group Lasso for logistic regression, which instead of the usual L_1 penalty that shrinks each indicator individually, shrinks groups of predictors corresponding to a single categorical factor toward zero together [Meier, et al. 2008; Yuan and Lin 2006]. Suppose we fit the logistic regression model of equation [2] considering this grouped structure as follows:

$$\hat{\theta}_\lambda = \arg\min_\theta \left(-l(\theta) + \lambda \left(\sum_{j=1}^{p} \sqrt{df_j} \left\| \beta_j \right\|_2 + \sum_{k=1}^{q} \sqrt{df_k} \left\| \gamma_k \right\|_2 \right) \right) \qquad (4)$$

where $\left\| \beta_j \right\|_2 = \left(\beta_j^T \beta_j \right)^{1/2} = \sqrt{\sum_{g=1}^{2} \beta_{j,g}^2}$ and $\left\| \gamma_k \right\|_2 = \left(\gamma_k^T \gamma_k \right)^{1/2} = \sqrt{\sum_{g_1=1}^{2} \sum_{g_2=1}^{2} \gamma_{k,g_1 g_2}^2}$ with $df_j = 2$ and $df_k = 2^2 \ \forall j,k$. This penalty is the group-wise L_2-norm, intermediate between the L_1-Lasso and the L_2-Ridge penalties, and penalizes main effect groups j=1,…,p and pairwise interaction groups k=1,…,q collectively, encouraging sparsity between but not within factors [Yuan and Lin 2006]. To distinguish this grouped method from the traditional penalization, we refer to this method as GL1. Equation (5) is an extension of the model defined in [Meier, et al. 2008], and we fit this model using the 'grplasso' package in R [Meier 2009]. As with TL1, the particular SNPs associated with the set $\{l : \theta_l \neq 0\}$ define the SNPs select by GL1.

DOI: 10.2202/1544-6115.1613

SIMULATION DESIGN AND ANALYSIS

In order to fairly compare MDR and Lasso (both TL1 and GL1) for variable selection, we designed a Monte Carlo simulation study. Our objective is to compare the performance of the three variable selection approaches in detecting various patterns of epistasis in data with different sample characteristics, including number of total loci p, minor allele frequency, and effect size. For each model scenario, case/control Monte Carlo replicate datasets were generated with 125 cases and 125 controls to represent a small sample size for an association study.

In order to depict different patterns of epistatic interactions, including interactions in both the presence and absence of independent main effects, three complex penetrance patterns were utilized. All three patterns (XOR, BOX, and MOD, described below) display two-locus interactions, characterized by the penetrance at each of the nine two-locus genotype combinations seen in both Table 1 and Figure 1. The XOR pattern represents a two-locus purely epistatic interaction with no marginal effects at either locus; this pattern is a modification of a non-linear exclusive OR function first described by [Li and Reich 2000]. In the XOR model, low risk of disease is dependent on a heterozygous genotype at exactly one of two loci ('Aa' or 'Bb'). Both the BOX and MOD patterns are two-locus interactions with main effects at both loci. The BOX pattern is symmetric and is a variation on the dominant-dominant model described by [Neuman and Rice 1992]. In the BOX model, low risk of disease is dependent on two low-risk alleles ('AA' or 'BB') at either one or both loci. The MOD pattern is assymmetric, and represents a modifying effect model on an exclusive OR function described by [Li and Reich 2000]. In the exclusive OR model, high risk of disease depends on two high risk alleles ('aa' or 'bb') at exactly one of two loci, so genotype combination 'AaBB' is low-risk; under the MOD model, 'AaBB' is modified from low-risk to high-risk, resulting in both marginal effects and a complex interaction. See Table 1 and Figure 1 for more details on all three patterns.

Model	**XOR**			**BOX**			**MOD**		
Genotype	**AA**	**Aa**	**aa**	**AA**	**Aa**	**aa**	**AA**	**Aa**	**aa**
BB	y	x	y	x	x	x	x	y	y
Bb	x	z	x	x	y	y	x	x	y
bb	y	x	y	x	y	y	y	y	x

Table 1 - Penetrance patterns for 2-locus epistatic models. Cells marked *x* represent genotype combinations with lower risk. The values *x*, *y*, and *z* represent penetrance values with $0 < x < y \leq z < 1$ which were chosen to achieve the desired heritability. The baseline penetrance, *x*, was fixed at 0.05. For XOR models with MAF=0.5, $z = y$; for XOR models with MAF=0.25, $z > y$ to achieve no marginal effects at either locus.

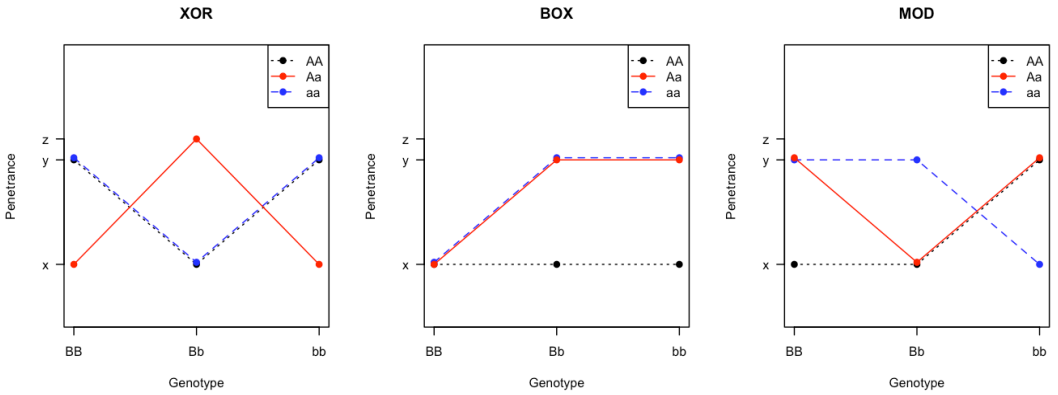

Figure 1 – Penetrance function by genotype for XOR, BOX, and MOD patterns.

In addition to the XOR, BOX, and MOD penetrance patterns, simulation design factors considered were number of total loci *p* (either 25 or 100), minor allele frequency (0.25 or 0.5), and heritability (1.0, 2.5, 5.0, 7.5, or 10.0%), for a total of 60 combinations summarized in the Appendix. For each model scenario, the specific penetrance function was established based on the combination of these simulation factors, such that the baseline penetrance was fixed at 0.05 to ensure a realistic population prevalence rate, and are available from the authors upon request. 100 case/control Monte Carlo replicates were generated under each scenario assuming Hardy-Weinberg proportions and case-control status was ascribed based on the particular penetrance function using the software genomeSIM [Dudek, et al. 2006]. Simulated SNP variables were uncorrelated, representing no linkage disequilibrium between them. While this number of potential predictors is much smaller than current typical genotyping studies, even

DOI: 10.2202/1544-6115.1613

candidate gene studies, the sparsity of exploring higher values made running current implementations of the LASSO methodologies impractical.

All 100 datasets were analyzed with MDR, TL1, and GL1 for each of the 60 factor combinations, and the performance of each method recorded. MDR was implemented using 5-fold cross-validation [Motsinger and Ritchie 2006] considering sizes of interaction k=1,…,4, and the single best model with highest accuracy and cross-validation consistency was chosen as the final selected model. For both TL1 and GL1, the genotype encoded as "0", the lower frequency homozygote, was treated as the reference level. The tuning parameter λ was chosen with BIC rather than cross-validation because this approach is commonly used, easily implemented, and recommended [Zou, et al. 2007] and reduces computation time. Moreover, unlike cross-validation which tends to overfit, BIC is selection consistent. [Wang, et al. 2007; Wang and Leng 2007]. The final model was chosen based on this $\lambda > 0$; for the case with $\lambda = 0$, where only the intercept remains and no loci are selected, the first loci that appear in the solution path were chosen as the final model so as to remain comparable with MDR, because MDR does not allow for null models to be selected.

The performance of each method was measured in terms of power to detect the interaction, the average size of the model identified (i.e. number of active predictor loci), and the true and false positive rates. Power, the proportion of correctly identified models across the 100 replicates, was calculated under two definitions, liberal and conservative. Under the conservative definition, a model was considered correct if both true loci were identified, but no false positive loci; the liberal definition allowed for the inclusion of false positive loci. In this case, the conservative estimate of power counts the number of correctly identified models across the 100 replication when no false positive or false negative loci are allowed. The liberal definition estimate counts the number of correctly identified models across the 100 replication when false positive loci are allowed, but no false negative loci are allowed (representing a situation where models might be followed up in a replication set to eliminate false positives). For the MDR analyses, the final model for each dataset was chosen based on minimal prediction error and maximal cross-validation consistency, where in the case of disagreement in these metrics, the final model is chosen based on the rule of parsimony. For the LASSO analyses (both the TL1 and GL1 results), the genotypes with significant terms were considered to be included in the final model, where to be considered a significant interaction, the interaction term must be included in the model. True positive rate was calculated as the proportion of true positives identified out of the total number of true positive identifications possible (2*100) and the false positive rate was calculated as the proportion of false positives identified out of the total number of false positive identifications possible ((p-2)*100). All simulations were performed on quad-core Core2 Xeon

processors (8 processors, each at 3 GHz and with 4GB of memory) using software for MDR written in C [Hahn, et al. 2003] and the freely available R software for TL1 and GL1 [R Development Core Team 2005]. Final results were statistically analyzed for differences in performance between the three methods with a mixed effects model and pair-wise contrasts between methods using SAS version 9.1 [SAS Institute Inc. 2004]. To account for the repeated analysis on the same datasets by the three variable selection methods, a random effect was specified for each combination of simulation factors.

RESULTS

Figures 2 and 3 display the power results (conservative and liberal, respectively) for all 60 simulated models. To test for significant differences among methods, after visual inspection, statistical models were fit controlling for number of loci, heritability, and the potential main effects, two-way, and three-way interactions between type of analysis method, minor allele frequency, and penetrance pattern. In general, we see that both conservative and liberal power increase with effect size h^2 ($p<0.0001$; $p<0.0001$) and with total number of loci p ($p<0.0001$; $p<0.0001$) after controlling for the effects of MAF, model type, and variable selection method. In general, liberal power is slightly higher than conservative power for MDR and GL1, but much higher for TL1. As can be seen in Figures 2 and 3, for both conservative and liberal power, there is a three-way interaction between model type, MAF, and variable selection approach ($p=0.0025$; $p<0.0001$). Therefore in order to compare variable selection methods, pair-wise contrasts comparing both MDR versus GL1 and TL1 versus GL1 were calculated stratified on each of the six combinations of minor allele frequency and penetrance pattern, for a total of 12 contrasts per power measure. As these results are meant to help interpret the results of the simulation, p-values reported below are unadjusted and adjustment is left to the reader. First we compare the power of GL1 and TL1. For MAF=0.5, GL1 has higher conservative and liberal power than TL1 for all three model types (all three $p<0.0004$; both $p<0.0001$ for XOR and MOD), with the exception of the liberal power of the BOX model ($p=0.1405$). For MAF=0.25, GL1 had significantly higher conservative and liberal power for the XOR model ($p<0.0001$; $p<0.0001$), and significantly higher conservative power for the MOD model ($p=0.0001$). TL1 has higher conservative and liberal power only for the BOX model with MAF=0.25 ($p=0.8377$; $p<0.0001$).

Next we compare the power of GL1 and MDR. For the BOX model, the conservative power of MDR is higher than GL1 across MAF although not statistically significant; but liberal power is significantly higher for MDR for both MAF=0.25 and 0.5 ($p=0.0010$; $p=0.0118$). While the results appear similar for

DOI: 10.2202/1544-6115.1613

the MOD model, the higher observed conservative and liberal power of MDR is not significant (p=0.1654 and p=0.0611 for maf=0.25; p=0.8860 and p=0.8162 for maf=0.5). For the XOR model, MDR has the highest conservative and liberal power for MAF=0.5 (p=0.0002; p<0.0001); conservative power is a substantial 18.5% (SE=0.049) higher for MDR than GL1. GL1 has higher conservative and liberal power for MAF=0.25 (p=0.2536; p=0.0149); under this scenario, we estimate that liberal power is 8.5% (SE=0.034) higher for GL1. Conservative and liberal power results for all factor combinations can be seen in Figures 2 and 3.

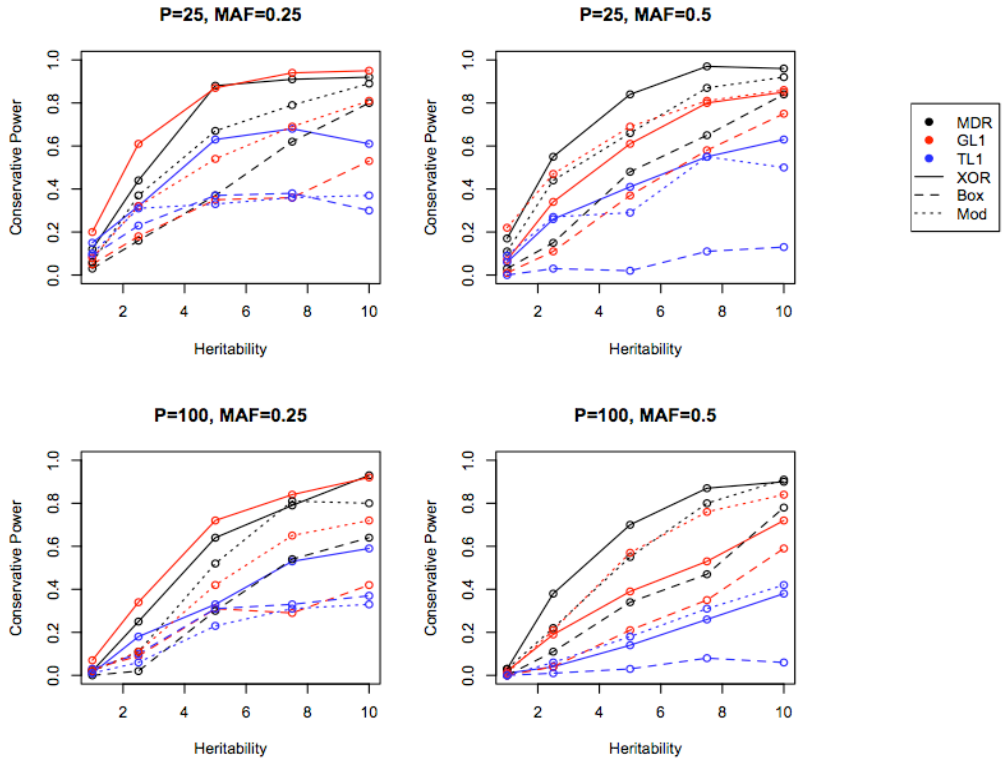

Figure 2 – Conservative power for increasing heritability for MDR, TL1, and GL1.
Conservative power is plotted for the XOR, BOX, and MOD patterns for MAF=0.25 and P=25 (top left), MAF=0.5 and P=25 (top right), MAF=0.25 and P=100 (bottom left), and MAF=0.5 and P=100 (bottom right). Standard errors range from 0.017 to 0.050, 0.000 to 0.050, 0.000 to 0.050, and 0.000 to 0.050, respectively.

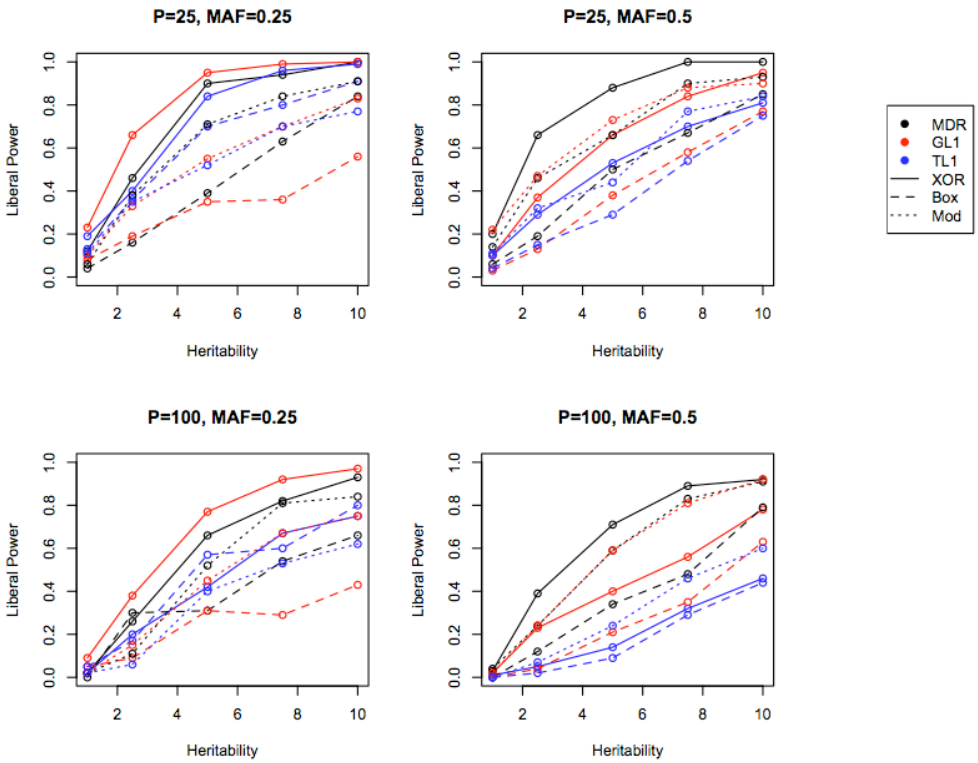

Figure 3 – Liberal power for increasing heritability for MDR, TL1, and GL1.
Liberal power is plotted for the XOR, BOX, and MOD patterns for MAF=0.25 and P=25 (top left), MAF=0.5 and P=25 (top right), MAF=0.25 and P=100 (bottom left), and MAF=0.5 and P=100 (bottom right). Standard errors range from 0.000 to 0.050, 0.000 to 0.050, 0.000 to 0.050, and 0.000 to 0.050, respectively.

The reported true and false positive rates represent a decomposition of our definitions of power into true positive and false positive components. Statistical models for both true and false positive rates were fit controlling for the main effects of number of loci, heritability and the potential two and three-way interactions between type of analysis method, minor allele frequency, and penetrance pattern. The estimated true positive rates of all three variable selection approaches can be seen in Figure 4. As expected, we see the true positive rate increase with effect size ($p<0.0001$) and total number of predictors ($p<0.0001$). Like power, we also observe a three-way interaction between MAF, analysis method, and penetrance pattern on the true positive rate ($p<0.0001$, see Figure 4). To compare analysis methods, pair-wise contrasts were stratified on each of the six combinations of minor allele frequency and penetrance pattern to compare both MDR versus GL1 and TL1 versus GL1, for a total of 12 contrasts. GL1 has a higher true positive rate than TL1 for the XOR models (both $p<0.0001$), BOX model with MAF=0.25 ($p<0.0001$), and MOD model with MAF=0.5 ($p<0.0001$).

DOI: 10.2202/1544-6115.1613

In general, MDR and GL1 have similar true positive rates for the BOX and MOD models; however, for the XOR model the rate is 17.4% higher for MDR with MAF=0.5 ($p<0.0001$) and 8.8% higher for GL1 with MAF=0.25 ($p=0.0018$). In terms of false positive rates, we see a decrease with effect size ($p<0.0001$) and total number of predictors ($p=0.0005$) with can be seen in Figure 5. Additionally, we also observe a three-way interaction between MAF, analysis method, and penetrance pattern on the false positive rate ($p=0.0040$; see Figure 5). Therefore pair-wise contrasts to compare methods were again stratified on each of the six combinations of minor allele frequency and penetrance pattern to compare MDR versus GL1 and TL1 versus GL1. False positive rates are quite similar between MDR and GL1 for all combinations, with the exception of the XOR model with MAF=0.5, where the rate is higher for GL1 ($p<0.0001$). Most notably, the false positive rates are quite high for TL1, and are consistently higher than GL1 (and also MDR) in all cases (all six $p<0.0001$). For TL1, the false positive rate decreases with heritability, but for even the high effect size of h^2=10%, the false

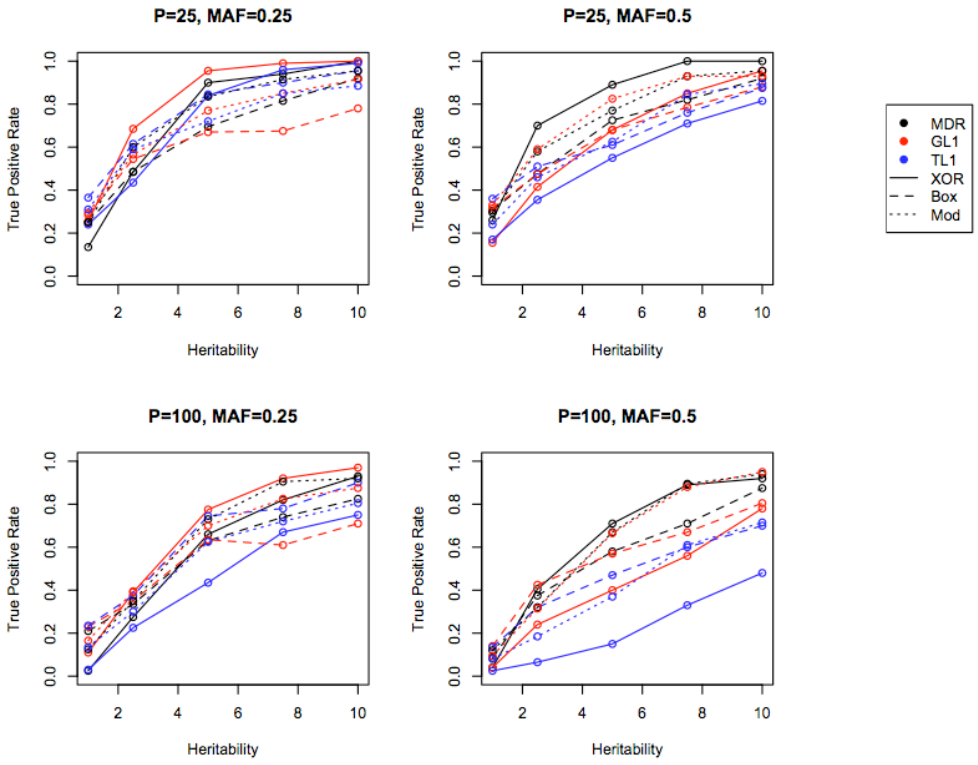

Figure 4 – True positive rates for increasing heritability for MDR, TL1, and GL1.
The true positive rate is plotted for the XOR, BOX, and MOD patterns for MAF=0.25 and P=25 (top left), MAF=0.5 and P=25 (top right), MAF=0.25 and P=100 (bottom left), and MAF=0.5 and P=100 (bottom right). Standard errors range from 0.000 to 0.050, 0.000 to 0.050, 0.016 to 0.050, and 0.016 to 0.050, respectively.

positive rate is still much greater than zero (see Figure 5). True and false positive rates for all 60 factor combinations can be seen in Figures 4 and 5.

The patterns we observe in both true and false positive rates are validated by the results for the average number of active predictors identified by each method, detailed in Figure 6. A statistical model for average model size was fit controlling for the main effects of number of loci, heritability, type of analysis method, minor allele frequency, and penetrance pattern. In general, the three variable selection methods display significantly different average model sizes ($p<0.0001$). For all methods, identified model size increases with effect size ($p<0.0001$) and total number of predictors ($p=0.0030$), consistent with the power and true positive rate findings. There was no observable interaction between analysis method and any of the other design factors, so in order to compare variable selection techniques, all three pair-wise contrasts were computed between methods without further stratification. MDR and GL1 have similar

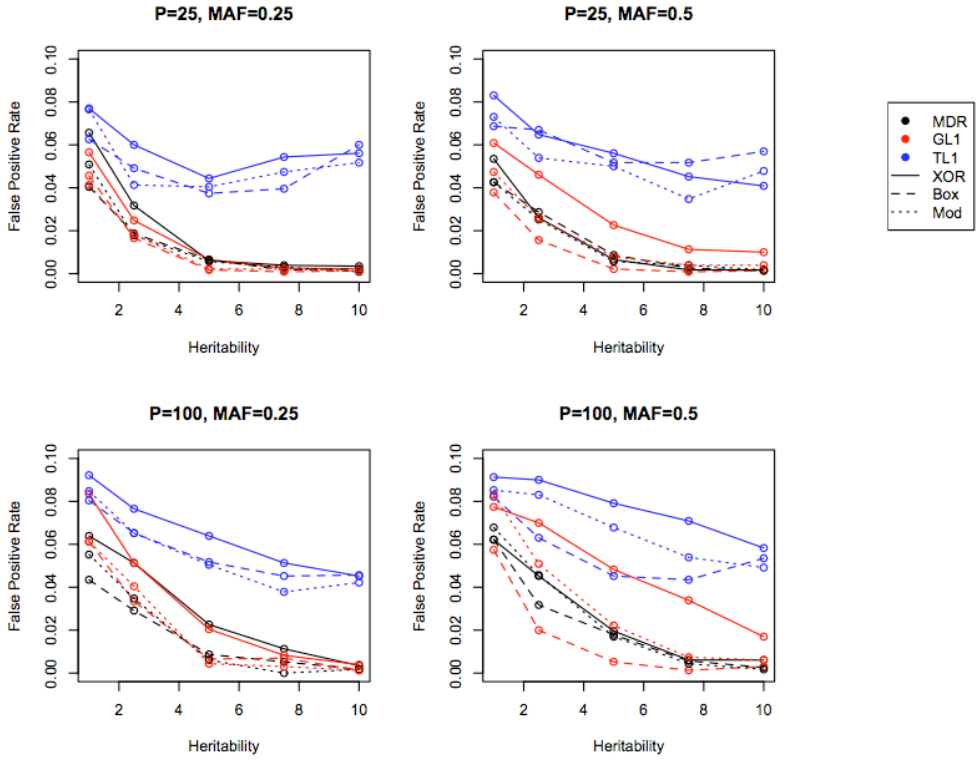

Figure 5 – False positive rates for increasing heritability for MDR, TL1, and GL1.
The false positive rate is plotted for the XOR, BOX, and MOD patterns for MAF=0.25 and P=25 (top left), MAF=0.5 and P=25 (top right), MAF=0.25 and P=100 (bottom left), and MAF=0.5 and P=100 (bottom right). Standard errors range from 0.003 to 0.027, 0.003 to 0.028, 0.000 to 0.029, and 0.004 to 0.029, respectively.

DOI: 10.2202/1544-6115.1613

model size results (p=0.7658), where the average number of active predictors approaches the target value of 2, the true number of causative loci, as heritability and predictor size increase. TL1 has significantly higher model size than both MDR and GL1 (p<0.0001; p<0.0001), and the observed model size moves further from the target 2 as heritability and predictor size increase. For MDR and GL1, the number of active predictors is below 2 in most cases, but is highest (and closest to 2) for the XOR model and furthest below 2 for the BOX model. For TL1, the model size is consistently above 2, displaying that TL1 is over-fitting. Results for average model size can be seen for all 60 factor combinations in Figure 6.

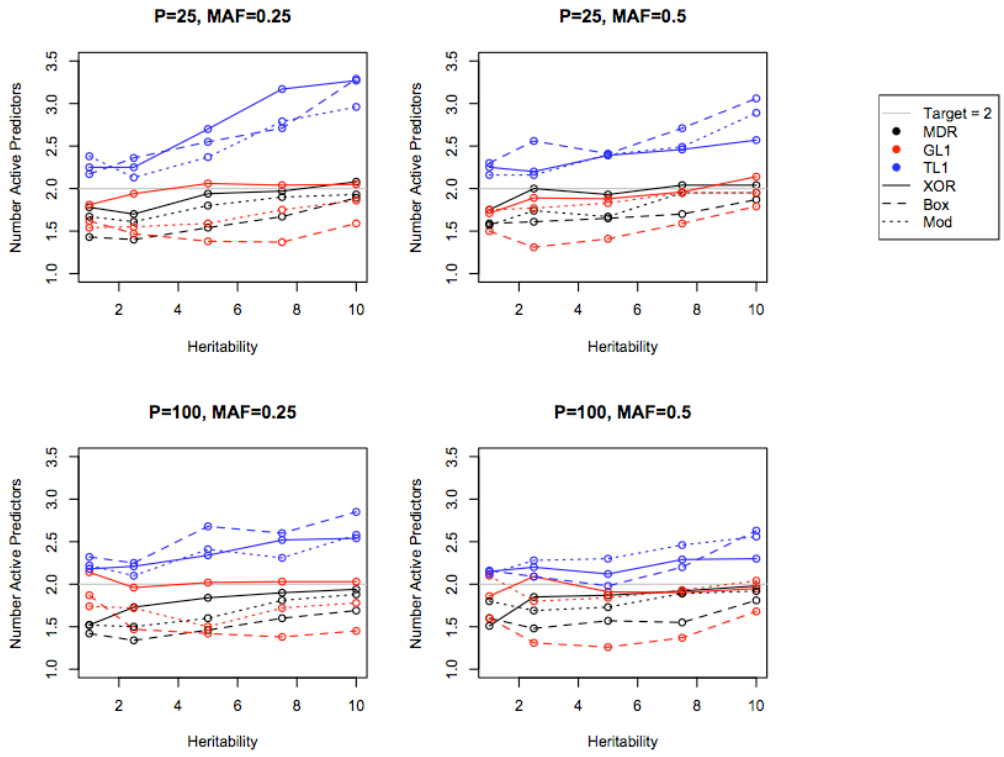

Figure 6 – Average number of active predictors versus heritability for MDR, TL1, and GL1.
The average model size is plotted for the XOR, BOX, and MOD patterns for MAF=0.25 and P=25 (top left), MAF=0.5 and P=25 (top right), MAF=0.25 and P=100 (bottom left), and MAF=0.5 and P=100 (bottom right). Standard errors range from 0.022 to 0.960, 0.036 to 0.914, 0.026 to 0.870, and 0.040 to 0.995, respectively.

DISCUSSION

In the current study, we evaluate the relative performance of three variable selection techniques for detecting gene-gene interactions in case/control genetic association studies: Multifactor Dimensionality Reduction, a commonly-used data-mining approach, and two L_1-penalized regression methods, the traditional ungrouped and the group Lasso for logistic regression. As expected, we find that none of the three approaches is optimal for all genetic models, but rather that the highest performing method is context dependent, with interactions based on both type of penetrance pattern and minor allele frequency. By and large, both MDR and GL1 identify more true positive loci than TL1. In terms of the two Lasso approaches, we observe that GL1 is frequently superior to TL1, which tends to over-fit, identifying false positive as well as true positive loci. Overall, MDR has higher power to detect interactions for models that also exhibit independent main effects, such as the BOX and MOD patterns. Upon further investigation, it appears that for both types of Lasso, the main effects of these models tend to dominate; often only one of the two loci is identified, and the interaction effects are overlooked. This is not surprising, since for a regression-based approach, main effects are typically more easily detectable than interactions because they require fewer degrees of freedom and are less affected by the curse of dimensionality [Bellman 1961]. The parameters associated with main effects may be more precisely estimated since the data is less sparse, resulting in less penalization than the less precise interaction. A constructive induction technique such as MDR treats both the main effects and interactions between loci collectively rather than separately, so the main effects will be less likely to overshadow the presence of the interaction.

When the model is purely epistatic, such as the case of the XOR model, both GL1 and MDR perform better than TL1. For lower minor allele frequencies, GL1 outperforms MDR, whereas MDR outperforms GL1 for higher frequencies. For the two Lasso approaches we see an interesting trend of improved performance for the lower minor allele frequency of 0.25 for this particular model, arising from the parametric nature of the approach. Typically we expect that as the minor alleles become more common in the population, that the causative loci would be easier to identify. However, when the MAF is 0.5 for the XOR pattern, a large number of the expected cases and controls have the double heterozygous genotype, with few observations having any of the four double homozygous genotypes. Therefore when one of the homozygous genotypes is treated as the reference level in the regression parameterization (as would commonly be the case), we expect fewer observations in the four multi-locus genotypes representing the interactive effects then what we see with MAF of 0.25. This results in unstable parameter estimation for these interactive effects, and

DOI: 10.2202/1544-6115.1613

because of the absence of independent main effects, the causative loci may be missed. A similar phenomenon for this penetrance pattern has been reported in other studies [He, et al. 2009]. This challenge could potentially be avoided by a direct parameterization of the nine two-locus genotype combinations instead of considering the main and interactive effects separately, although this needs further investigation.

In general, MDR and GL1 identify fewer false positive loci than TL1. The large false positive rate of TL1 is also reflected in the high average number of active predictors, indicating that TL1 is over-fitting. Because TL1 has a reasonable liberal power and rate of true positives, it seems that the correct loci are being identified, but that additional nuisance loci are also appearing in the final model. One possible explanation is that TL1 is not properly accounting for the categorical structure of the data, increasing the difficulty to distinguish true associations from those due to chance alone. By not penalizing terms dealing with the same effect together, we fail to draw strength across groups and the individual dummy variables are not penalized enough. Additionally, the 'glmpath' algorithm employed internally computes the grid of values considered for the tuning parameter λ, and these values cannot be easily manipulated by the user; potentially, a finer grid of stronger λ values may be more appropriate.

The results of our study are relatively consistent with those of previous comparisons of MDR with the two-stage L_2 regularization approach, stepPLR. Park and Hastie [2008] explore only four different model scenarios, and find that MDR has slightly lower power than stepPLR for a heterogeneity model and a purely epistatic model; however their comparisons considered a smaller number of simulation factors and their results were based on only 30 replicate datasets. In a more extensive study He et al [2009] found that the comparison of stepPLR and MDR depends on allele frequency and model pattern, where stepPLR performs better for additive effects and worse for purely epistatic effects, and the results were relatively consistent with those of this study. The present comparison of MDR with TL1 and GL1 helps us to glean new information on how a non-parametric data-mining method such as MDR compares with a parametric penalization approach, particularly for one-stage variable selection techniques implemented in a realistic fashion; these two popular approaches are directly compared as they would typically be used by a researcher.

Based on these simulation results, we provide a few recommendations of when each variable selection approach might be most suitable for detecting and characterizing interactions with different mechanisms. For data with lower minor allele frequencies, penalized regression approaches such as the Lasso may be more appropriate, particularly if main effects are not expected. For models that may exhibit independent main effects within interaction models or for data with higher allele frequencies, MDR may be more appropriate. For categorical

genotypic data, such as SNPs, it is important to account for the natural grouped structure of the predictors and GL1 may be better suited for detecting interactions than TL1, particularly if p is large.

While this study is useful in informing researchers about choosing an analysis strategy for detecting epistasis under a number of different scenarios, it is not a fully comprehensive comparison. While we consider a broad range of scenarios including various penetrance patterns of interaction, heritabilities, minor allele frequencies, and number of total SNPs, we only consider complete genotype data without error. The goal of any methods comparison should be to guide researchers in how to choose an analysis method for real data application; but real data includes various types of error such as genotyping error, missing data, phenocopy, and genetic heterogeneity, and these types of error should be incorporated into the comparison. We compare MDR, TL1, and GL1, but as the application of penalized regression for variable selection in genetic epidemiology becomes more popular, other variations of Lasso and other penalized approaches should be considered. The adaptive Lasso has proved a promising technique with attractive theoretical properties [Zou 2006], and while it cannot be directly implemented if the sample size is less than the total number of predictors ($2p+2^2q$), other adaptive strategies could be explored. Additionally, recently developed novel data-mining approaches for epistasis should be investigated, such as the Evaporative Cooling feature selection technique, based on the thermodynamic process of evaporation [McKinney, et al. 2007].

In the current study we consider the effect of analyzing datasets with 25 to 100 SNPs, a small choice of p compared to what is frequently seen in real data. All methods become much less computationally efficient as p increases, so comparisons in terms of computation time in addition to performance are of paramount concern to analysts. MDR is typically faster than GL1 and TL1, particularly if the grid of λ considered is large and tuned with cross-validation rather than BIC. Of course, MDR is written in C, a faster language than R, so a thorough comparison of computational efficiency involves additional research. On the same note, as genotyping technology improves and p increases, approaches for GWAS analysis become more relevant. In order to reflect this trend, expanding our comparison to include scenarios on a genome-wide scale rather than a candidate gene study requires further investigation. Because of the high-dimensionality of GWAS data, filter approaches are gaining popularity to address this analytical challenge. Both MDR and penalized regression can be used as filters, and recently screening approaches have emerged for Lasso such as the Screen and Clean [Wu, et al. 2010]. Future studies should consider comparisons at the level of full-genome data.

In the current study we compare and contrast the performance of three variable selection strategies to identify epistatic interactions and provide general

DOI: 10.2202/1544-6115.1613

recommendations for their usage. We focus primarily on implementations of these techniques that are easily accessible to the general researcher, emphasizing relative performance in a realistic as opposed to an ideal setting. Although these comparisons of analytical approaches are extensive but not exhaustive, we do gain a better understanding of the strengths and weaknesses of each approach and some insight as to when each method might be most appropriate. No approach had consistently high performance, and therefore researchers will need to tailor their analysis to the particular application at hand. In the future, as both technological and methodological advancements are made in this area, the investigation of gene-gene and gene-environment interactions for common complex disease in high-dimensional data will become more widespread; the researcher's selection of an appropriate analytical strategy will be imperative to properly identifying these complex genetic etiologies, and comparisons such as this will be an important tool towards this end.

APPENDIX

Specifications for the 60 simulated 2-locus epistatic models, including number of loci, penetrance pattern, heritability, and minor allele frequency.

Model Number	Number of loci (*p*)	Model Type	h^2 (%)	MAF
1	25	XOR	1.0	0.25
2	25	XOR	1.0	0.50
3	25	XOR	2.5	0.25
4	25	XOR	2.5	0.50
5	25	XOR	5.0	0.25
6	25	XOR	5.0	0.50
7	25	XOR	7.5	0.25
8	25	XOR	7.5	0.50
9	25	XOR	10.0	0.25
10	25	XOR	10.0	0.50
11	25	BOX	1.0	0.25
12	25	BOX	1.0	0.50
13	25	BOX	2.5	0.25
14	25	BOX	2.5	0.50
15	25	BOX	5.0	0.25
16	25	BOX	5.0	0.50
17	25	BOX	7.5	0.25
18	25	BOX	7.5	0.50
19	25	BOX	10.0	0.25
20	25	BOX	10.0	0.50
21	25	MOD	1.0	0.25
22	25	MOD	1.0	0.50
23	25	MOD	2.5	0.25
24	25	MOD	2.5	0.50
25	25	MOD	5.0	0.25
26	25	MOD	5.0	0.50
27	25	MOD	7.5	0.25

Appendix, continued

28	25	MOD	7.5	0.50
29	25	MOD	10.0	0.25
30	25	MOD	10.0	0.50
31	100	XOR	1.0	0.25
32	100	XOR	1.0	0.50
33	100	XOR	2.5	0.25
34	100	XOR	2.5	0.50
35	100	XOR	5.0	0.25
36	100	XOR	5.0	0.50
37	100	XOR	7.5	0.25
38	100	XOR	7.5	0.50
39	100	XOR	10.0	0.25
40	100	XOR	10.0	0.50
41	100	BOX	1.0	0.25
42	100	BOX	1.0	0.50
43	100	BOX	2.5	0.25
44	100	BOX	2.5	0.50
45	100	BOX	5.0	0.25
46	100	BOX	5.0	0.50
47	100	BOX	7.5	0.25
48	100	BOX	7.5	0.50
49	100	BOX	10.0	0.25
50	100	BOX	10.0	0.50
51	100	MOD	1.0	0.25
52	100	MOD	1.0	0.50
53	100	MOD	2.5	0.25
54	100	MOD	2.5	0.50
55	100	MOD	5.0	0.25
56	100	MOD	5.0	0.50
57	100	MOD	7.5	0.25
58	100	MOD	7.5	0.50
59	100	MOD	10.0	0.25
60	100	MOD	10.0	0.50

REFERENCES

Bellman R. 1961. Adaptive Control Processes. Princeton: Princeton University Press.

Brassat D, Motsinger AA, Caillier SJ, Erlich HA, Walker K, Steiner LL, Cree BA, Barcellos LF, Pericak-Vance MA, Schmidt S and others. 2006. Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans. Genes Immun 7(4):310-5.

Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics 10(6):392-404.

Dudek SM, Motsinger AA, Velez DR, Williams SM, Ritchie MD. 2006. Data simulation software for whole-genome association and other studies in human genetics. Pac Symp Biocomput:499-510.

DOI: 10.2202/1544-6115.1613

Haas DW, Geraghty DE, Andersen J, Mar J, Motsinger AA, D'Aquila RT, Unutmaz D, Benson CA, Ritchie MD, Landay A. 2006. Immunogenetics of CD4 lymphocyte count recovery during antiretroviral therapy: An AIDS Clinical Trials Group study. J Infect Dis 194(8):1098-107.

Hahn LW, Ritchie MD, Moore JH. 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19(3):376-382.

He H, Oetting WS, Brott MJ, Basu S. 2009. Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene Interaction in a case-control study. Bmc Medical Genetics 10.

Li W, Reich J. 2000. A complete enumeration and classification of two-locus disease models. Hum Hered 50(6):334-49.

McKinney BA, Reif DM, White BC, Crowe JE, Moore JH. 2007. Evaporative cooling feature selection for genotypic data involving interactions. Bioinformatics 23(16):2113-2120.

Meier L. 2009. grplasso: Fitting user specified models with Group Lasso penalty [R package]. Version 0.4-2.

Meier L, van de Geer SA, Buhlmann P. 2008. The group lasso for logistic regression. Journal of the Royal Statistical Society Series B-Statistical Methodology 70:53-71.

Moore JH. 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56(1-3):73-82.

Motsinger-Reif AA. 2008. The effect of alternative permutation testing strategies on the performance of multifactor dimensionality reduction. BMC Res Notes 1:139.

Motsinger-Reif AA, Reif DM, Fanelli TJ, Ritchie MD. 2008. A Comparison of Analytical Methods for Genetic Association Studies. Genetic Epidemiology 32(8):767-778.

Motsinger AA, Brassat D, Caillier SJ, Erlich HA, Walker K, Steiner LL, Barcellos LF, Pericak-Vance MA, Schmidt S, Gregory S and others. 2007. Complex gene-gene interactions in multiple sclerosis: a multifactorial approach reveals associations with inflammatory genes. Neurogenetics 8(1):11-20.

Motsinger AA, Ritchie MD. 2006. The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. Genet Epidemiol 30(6):546-55.

Neuman RJ, Rice JP. 1992. TWO-LOCUS MODELS OF DISEASE. Genetic Epidemiology 9(5):347-365.

Nordgard SH, Ritchie MD, Jensrud SD, Motsinger AA, Alnaes GI, Lemmon G, Berg M, Geisler S, Moore JH, Lonning PE and others. 2007. ABCB1 and GST polymorphisms associated with TP53 status in breast cancer. Pharmacogenet Genomics 17(2):127-36.

Park MY, Hastie T. 2007a. glmpath: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Models. [R package]. Version 0.94.

Park MY, Hastie T. 2007b. L-1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society Series B-Statistical Methodology 69:659-677.

Park MY, Hastie T. 2008. Penalized logistic regression for detecting gene interactions. Biostatistics 9(1):30-50.

R Development Core Team. 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ritchie MD, Hahn LW, Moore JH. 2003. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24(2):150-7.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69(1):138-47.

SAS Institute Inc. 2004. Cary, NC.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B-Methodological 58(1):267-288.

Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH. 2007. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet Epidemiol 31(4):306-15.

Wang H, Li R, Tsai CL. 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94(3):553-568.

Wang HS, Leng CL. 2007. Unified LASSO estimation by least squares approximation. Journal of the American Statistical Association 102(479):1039-1048.

Wu J, Devlin B, Ringquist S, Trucco M, Roeder K. 2010. Screen and Clean: A Tool for Identifying Interactions in Genome-Wide Association Studies. Genetic Epidemiology 34(3):275-285.

Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25(6):714-721.

DOI: 10.2202/1544-6115.1613

Yang C, Wan X, Yang Q, Xue H, Yu WC. Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. Bmc Bioinformatics 11.

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B-Statistical Methodology 68:49-67.

Zou H. 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(476):1418-1429.

Zou H, Hastie T, Tibshirani R. 2007. On the "degrees of freedom" of the lasso. Annals of Statistics 35(5):2173-2192.