ARTICLE

Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data

Lin S. Chen,[1] Carolyn M. Hutter,[2] John D. Potter,[2] Yan Liu,[1] Ross L. Prentice,[2] Ulrike Peters,[2] and Li Hsu[1,*]

Genome-wide association studies (GWAS) have successfully identified susceptibility loci from marginal association analysis of SNPs. Valuable insight into genetic variation underlying complex diseases will likely be gained by considering functionally related sets of genes simultaneously. One approach is to further develop gene set enrichment analysis methods, which are initiated in gene expression studies, to account for the distinctive features of GWAS data. These features include the large number of SNPs per gene, the modest and sparse SNP associations, and the additional information provided by linkage disequilibrium (LD) patterns within genes. We propose a "*g*ene set *r*idge regression in *as*sociation *s*tudies (GRASS)" algorithm. GRASS summarizes the genetic structure for each gene as eigenSNPs and uses a novel form of regularized regression technique, termed group ridge regression, to select representative eigenSNPs for each gene and assess their joint association with disease risk. Compared with existing methods, the proposed algorithm greatly reduces the high dimensionality of GWAS data while still accounting for multiple hits and/or LD in the same gene. We show by simulation that this algorithm performs well in situations in which there are a large number of predictors compared to sample size. We applied the GRASS algorithm to a genome-wide association study of colon cancer and identified nicotinate and nicotinamide metabolism and transforming growth factor beta signaling as the top two significantly enriched pathways. Elucidating the role of variation in these pathways may enhance our understanding of colon cancer etiology.

Introduction

The complete sequence of the human genome, the HapMap Project, and recent advances in genotyping technology have made large-scale genome-wide association studies (GWAS) feasible. As a result, many novel susceptibility loci have been identified from the marginal association analysis of SNPs with disease risk.[1] However, there is far more information in the data that researchers are just beginning to explore. One particular topic of interest is how germline variation from genes with related functions may affect disease risk.

It is well established that functionally related genes can act concordantly[2] and that their action may be influenced by genetic variation in the chromosomal region of the gene (including the coding region, as well as the upstream and downstream sequences). The dense SNP marker panels from GWAS offer an unprecedented opportunity to comprehensively study germline variability of gene sets. Several informatics databases, such as the gene ontology (GO) database,[3] the Kyoto Encyclopedia of Genes and Genomes (KEGG),[4] and the Molecular Signatures Database (MSigDB),[5] have been curated to provide information on functions and relatedness of genes and to classify genes into gene sets with common underlying features. By assigning SNPs to the nearest genes based on genomic location, one can combine information on all of the variation in the same gene set and collectively assess the association with disease risk. Such analysis can provide valuable insights into the biological basis underlying disease risk. Indeed, the successes of using prior knowledge to assess gene set association in GWAS have been reported in studies of diseases including Parkinson disease,[6,7] age-related macular degeneration,[7] multiple sclerosis,[8] and bipolar disorder.[9]

Gene set enrichment analysis was first proposed by Mootha et al.[10] for detecting concerted changes in the expression of genes grouped by their functional relatedness. It has shown great promise in deriving new information from expression data. Many methods have since been developed; see, for example, Goeman et al.,[11] Subramanian et al.,[5] Tian et al.,[12] Efron and Tibshirani,[13] Jiang and Gentleman,[14] and Dinu et al..[15] Two recent papers[16,17] provide a comprehensive review and comparison of these various methods. Two different approaches are usually taken in assessing gene set enrichment for expression data. The first approach investigates whether a gene set of interest is enriched with genes differentially expressed between two biological states in comparison to a random gene set. To generate the null distribution for this approach, one can randomly sample genes from the same data set to form random gene sets and calculate null statistics for these random gene sets. The second approach tests the null hypothesis that the gene set of interest does not contain any gene or genes differentially expressed between two biological states. For this approach, one could permute phenotype labeling in the data and calculate null statistics based on the permuted phenotypes. In either approach, one can obtain

[1]Biostatistics and Biomathematics Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA; [2]Cancer Prevention Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA
*Correspondence: lih@fhcrc.org
DOI 10.1016/j.ajhg.2010.04.014.

nonparametric p values of gene set association by comparing the observed statistics with the null statistics obtained under the respective null hypothesis.

Compared to expression data, there are several distinctive features in genome-wide association data that require different methodological consideration. First, in contrast to gene expression data, in which each gene contains only a few transcripts, in GWAS data, each gene is comprised of a relatively large number of SNPs because of the high density of marker panels. Second, in contrast to gene expression data, in which many genes are up- or downregulated in the diseased group versus the nondiseased group, the SNP associations in GWAS data tend to be modest and somewhat sparse. Third, many SNPs are in linkage disequilibrium (LD), and using information from these SNPs jointly can enhance the power for detecting disease risk variants, including detecting variants that are not directly genotyped.

Methods for gene set enrichment analysis have been developed for GWAS data. For example, Wang et al.[7] extended the method developed by Subramanian et al.[5] to GWAS data. PLINK,[18] a popular software for analyzing GWAS data, offers an option to perform gene set analysis, which we will shorthand by Plink. Holmans et al.[9] proposed an approach called ALIGATOR (*a*ssociation *li*st *g*o *a*nno*ta*t*or*), which does not require individual-level SNP data. Generally speaking, all of these methods are based on marginal association analysis of each SNP and don't make use of LD structure in the data. As such, ungenotyped disease-associated variants may not be best accounted for in the gene set analysis. Plink formulates gene set statistics on the basis of SNPs, whereas Wang et al.'s method and ALIGATOR are based on genes. Wang et al. chooses the most significant associated SNP in each gene, and ALIGATOR chooses the genes that are hit by any of the top SNPs. Because the number of SNPs in a gene can be large, test statistics based on genes that are represented by only one SNP may lose power after adjusting for the gene size. They may also lose power because of not accounting for the potential multiple hits in a gene. Further discussion about these methods and the comparison with the proposed method is given in the Results.

In this paper, we present a gene set association method that accounts for each of these distinctive properties of GWAS data. We first assign SNPs to genes and summarize the variation of a gene by principal components,[19] which we term as "eigenSNPs." The eigenSNPs capture the overall gene structure and reduce the local correlation because of linkage disequilibrium. We then propose to use regularized regression technique[20] to select one or more representative eigenSNPs for each gene and assess their joint association with disease risk. The regularization is necessary because the total number of predictors, here eigenSNPs, is quite large compared to the sample size. We propose an algorithm called "*g*ene set *r*idge regression in *a*ssociation *s*tudies (GRASS)," which performs regularized logistic regression and assesses the gene set association. The underlying framework in the GRASS algorithm is regression based and therefore can be readily extended to incorporate covariates and to include gene-gene and gene-environment interaction effects at the SNP level.

Materials and Methods

Capturing Gene Structure with Principal Component Analysis

The first step of our method is to summarize the genetic variation in each gene. This is performed prior to our gene-set-based analysis. Often, SNPs within the same gene are in LD and may represent redundant information. Furthermore, genotyped SNPs may tag the true risk-associated variants, which may or may not have been genotyped. It is therefore desirable to capture the unique genetic variation within a gene to reduce the dimensionality of a gene and tag the ungenotyped potential risk variants. In addition, because our gene set analysis is regression based, constructing unique (orthogonal) components will also enhance both the interpretability of selected components and the chance of identifying components that are most strongly associated with disease risk.[21]

For each gene, we use principal components analysis (PCA) to decompose the genetic variation into orthogonal components. We perform PCA as follows: let n be sample size and m be the number of SNPs in a gene. We first standardize the $m \times n$ SNP matrix for each gene, as proposed in Price et al.,[22] so that SNPs with different minor allele frequencies (MAFs) are weighted equally. We apply singular value decomposition on the standardized SNP matrix, Z, and obtain $Z = UDV'$. Essentially, the decomposition projects the complete data onto a reduced eigenspace V of dimension l ($l \leq m$). The matrix U is an $m \times l$ orthogonal matrix with columns corresponding to sample SNP variations, and the matrix V is an $n \times l$ column orthogonal matrix, each column of which is defined as an eigenSNP and is a linear combination of all the relevant real SNPs. The matrix D is a diagonal matrix in which the jth diagonal entry d_j is the eigenvalue of the jth eigenSNP. EigenSNPs are decreasingly ordered by $|d_j|$, and $\pi_j = d_j^2 / \sum_{j'} d_{j'}^2$ represents the proportion of variation in the gene accounted for by the jth eigenSNP. Because a gene contains m SNPs, each with unit variance, if the proportion of variation of an eigenSNP i captured is equal to or greater than $1/m$, i.e., $\pi_i \geq 1/m$, we call the eigenSNP a nontrivial eigenSNP. For each gene, we select all the nontrivial eigenSNPs, which altogether explain ~95% of the gene variation. When applying group ridge regression, the nontrivial eigenSNPs from all the genes in the same set will be treated as predictors in the regression.

Group Ridge Regression with Lasso Penalty within the Group

The essence of the proposed GRASS algorithm is to identify association signals based on predefined gene sets via regularized regression. The intuition behind regularized regression is as follows: ordinary maximum likelihood estimation is sometimes not achievable or not efficient, particularly when the sample size is small relative to the number of predictors. In such cases, we can choose to trade bias for efficiency in estimation by maximizing the log likelihood function, subject to some constraint, or equivalently by minimizing the negated log likelihood plus the penalty function, which has a one-to-one relationship with the constraint. Many penalty functions have been proposed for regularizing

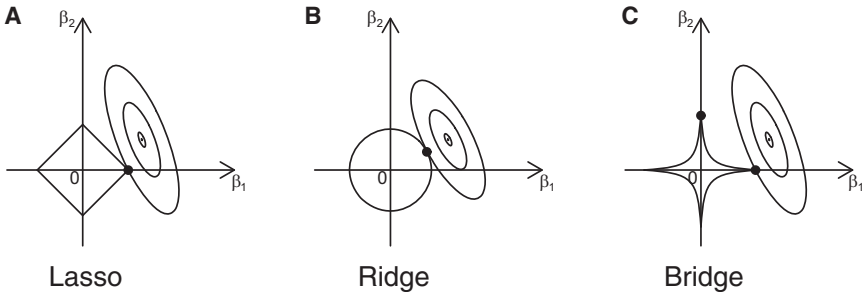

Figure 1. A Graphic Illustration of the Properties of Different Penalty Functions
A graphic illustration of the properties of three different penalty functions. The eclipses represent the likelihood contours. (A–C) The square, round, and star shapes represent the lasso, ridge, and bridge constraint, respectively. The dots are the points where likelihood contours are "tangent" to the constraints, i.e., the penalized likelihood estimates. Note that in lasso (A) or bridge (C), the constraint is discontinuous at zero. If the likelihood contour first touches the constraint at point zero, the corresponding parameter estimate is zero, and variable selection is achieved.

parameter values β_i for $i = 1, \ldots, p$ predictors. For example, the lasso penalty[20] is the ℓ_1 norm of parameter values, in which the ℓ_γ norm of parameter vector β is defined as $\|\boldsymbol{\beta}\|_\gamma = (\sum_{i=1}^{p}|\beta_i|^\gamma)^{\frac{1}{\gamma}}$, $\gamma > 0$. Therefore, the lasso penalty function is defined as $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p}|\beta_i|$. The ridge[23] and bridge penalty[24] take the form of the ℓ_2 and ℓ_γ norm with $0 < \gamma < 1$, respectively. As expected, different penalty functions affect the estimation in different ways. Figure 1 shows the behavior of these three penalty functions in a two-parameter case, β_1 and β_2. To obtain maximum constrained likelihood estimators, we essentially seek the points at which the log likelihood contour first "hits" the constraint. Lasso, ridge, and bridge penalty functions have constraints shaped like a square, circle, and star, respectively. As a consequence of the different shapes, lasso is likely to involve variable selection ($\beta_1 = 0$ or $\beta_2 = 0$), as well as parameter estimate shrinkage, and ridge yields mainly parameter estimate shrinkage; in contrast, bridge induces an even higher chance of variable selection than lasso, because the star shape of bridge makes the likelihood contour even more likely to hit one of the points ($\beta_1 = 0$ or $\beta_2 = 0$) than does the diamond shape of lasso.[25]

For gene set association analysis, we would like the estimation to capture the effects of all genes in the same pathway by using the most representative eigenSNPs in each gene. Note, in this context, that "the most representative eigenSNPs" refers to those that are most associated with disease risk. They are not necessarily the eigenSNPs that explain the most variation in a gene. This requires variable selection within a gene while shrinking the parameter estimates for the gene effects across genes. Clearly, none of the penalty functions discussed above meet these objectives. We therefore propose a group ridge penalty that is specifically tailored to the gene set association analysis via GWAS data. The group ridge combines ℓ_2-norm regularization at the gene level and ℓ_1-norm regularization within each gene. It shrinks (by ridge) the contribution to disease risk from each gene to down-weigh genes that may otherwise possibly exhibit extreme associations because of stochastic variation while performing eigenSNP variable selection (by lasso) within each gene simultaneously. This penalty function ensures that each gene contributes to the association score of the gene set. The amount of the contribution from each gene is determined flexibly by selecting the most associated eigenSNPs, which could be more than one, within the gene. In comparison, a simple lasso penalty would primarily impose variable selection among all eigenSNPs, regardless of which gene they belong to. This may result in gene set statistics being dominated by a single eigenSNP or by eigenSNPs from only a few genes. This somewhat contradicts the essence of the gene set association analysis, in which interest is on the assessment of the collective effect of a gene set on disease risk. Two other possible group-based penalty functions have also been proposed in the literature: group lasso[26] and group bridge.[27] In our context, group lasso does considerable variable selection at the gene level while retaining all eigenSNPs in the selected genes. As such, it also deviates from the motivation of gene set association analysis. Group bridge does considerable variable selection at both the gene and eigenSNP levels and can therefore be overly selective and miss useful, albeit modest, association signals in GWAS data.

Let $X = \{V_1, \ldots, V_G\}$ be an $n \times p$ pooled eigenSNP matrix of a gene set S with G total genes, in which $V_g (g = 1, \ldots, G)$ is an $n \times k_g$ eigenSNP matrix for gth gene and n, k_g, and $p = \sum_{g=1}^{G} k_g$ are the number of samples, the number of eigenSNPs in the gth gene, and the total number of eigenSNPs in the gene set, respectively. Let the parameter vector be denoted by $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_G^T)^T$, in which β_0 is the intercept and $\boldsymbol{\beta}_g (g = 1, \ldots, G)$ is the vector of corresponding regression coefficients for V_g. To simplify the representation of the log likelihood, we recode the disease status $y = 1$ for diseased and –1 for nondiseased. The log likelihood function can then be written as

$$\ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \ln\{1 + \exp(-\mathbf{X}_i\boldsymbol{\beta} \cdot y_i)\}. \qquad \text{(Equation 1)}$$

The group ridge (GRASS) estimator, $\widehat{\boldsymbol{\beta}}_\lambda$, can be obtained by minimizing the penalized likelihood function, which is the negated log likelihood plus the penalty term, given by

$$S_\lambda(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} w_g \left(\|\boldsymbol{\beta}_g\|_1\right)^2, \qquad \text{(Equation 2)}$$

in which $\|\boldsymbol{\beta}_g\|_1 = \sum_{j \in \text{gene}_g} |\boldsymbol{\beta}_j|$, and λ is a penalty parameter that governs how much penalty will be imposed on the parameter estimators. As one can see, the penalty function is the sum of squares of the ℓ_1 norm $\|\boldsymbol{\beta}_g\|_1$ over all genes weighted by a weight function w_g. It applies the ridge penalty among genes and the lasso penalty within each gene. Note that the intercept is not "regularized."

Different weighting options, w_g, for the penalty term can impact the estimation. One can weigh each gene equally by using $w_g = 1$. Alternatively, one can assign weights based on gene length. For example, genes with more eigenSNPs may be penalized more by employing a weight, $w_g = \sqrt{k_g}$, suggested by Yuan and Lin.[26] This weight function rescales the penalty function with respect to the dimensionality (number of eigenSNPs) of the parameter

vector, $\boldsymbol{\beta}_g$. In our analysis, we choose $w_g = 1$, so that genes with different numbers of eigenSNPs are evaluated equally in the set. However, to ensure that each gene contributes equally, we standardize the statistic contributed by each gene when forming the gene set statistics (see below).

Group Ridge Estimation Algorithm

We minimize the function in Equation 2 to obtain the group ridge estimator, $\widehat{\boldsymbol{\beta}}_\lambda$. When the number of genes and eigenSNPs is large, minimizing $S_\lambda(\boldsymbol{\beta})$ can be challenging. Our GRASS algorithm adopts ideas from several algorithms in the literature.[28–30] We use a block coordinate descent method to search for the optimal estimate for each block (group and gene), and, within each block, we employ a cyclic coordinate descent algorithm[28,29] (see Appendix A, Algorithm A1 for the general idea, and Algorithm S1 and S2 available online for detailed procedures).

Briefly, the algorithm decreases the objective function one coordinate (parameter) at a time while fixing other parameters at the current values. The procedure is repeated until some convergence criterion is met. To find the tentative next step, we use a one-step Newton-Raphson algorithm. The Newton-Raphson algorithm requires the objective function to be convex and smooth in order to find the minimum. With the group ridge penalty, the objective function is convex and smooth everywhere except for when some component of the regression vector equals zero as a result of the lasso penalty within a gene. Therefore, at each tentative step, the algorithm checks whether the estimate crosses zero and, if so, the estimate is set to zero. When the current estimate is zero, the algorithm tries both directions to see whether either direction improves the objective function and, if not, then the estimate remains zero. Because of convexity, it is not possible for both directions to improve the objective function.

We choose λ by Akaike's information criterion (AIC) for each gene set. AIC is defined as AIC $= 2p - 2\ell$, in which p is the number of parameters in the model and ℓ is the log likelihood for the estimated model.[31]

Gene Set Association Analysis

After λ is chosen based on AIC, we obtain the regression estimates at the optimal λ value and use these estimates to measure the strength of the association of the gene set with disease risk. We first summarize the association of each gene and then summarize the association of all genes together. By doing this, we can avoid potential bias as a result of varying gene size. Specifically, the gene-level association is estimated by

$$\|\widehat{\boldsymbol{\beta}}_{\lambda g}\| = \sqrt{\widehat{\beta}_{g1}^2 + \widehat{\beta}_{g2}^2 + \ldots + \widehat{\beta}_{gk_g}^2}, \qquad \text{(Equation 3)}$$

where $\widehat{\beta}_{g1},\ldots,\widehat{\beta}_{gk_g}$ are the estimated regularized log odds ratios for eigenSNPs in the gene g, obtained from minimizing $S_\lambda(\boldsymbol{\beta})$ with the optimal λ value. Note that because of the variable selection feature of group ridge within each gene, many of the estimated regularized log odds ratios are zero.

To adjust for gene size, one can choose a weighting function w_g in Equation 2 based on the structure of the data; however, the choice can be ad hoc. Here we use a permutation-based approach to adjust for gene size. Specifically, we use $w_g = 1$ in Equation 2 to obtain $\widehat{\boldsymbol{\beta}}_{\lambda g}$ and to standardize the gene-level statistic, $\|\widehat{\boldsymbol{\beta}}_{\lambda g}\|$, by

$$\tilde{\boldsymbol{\beta}}_g = \frac{\|\widehat{\boldsymbol{\beta}}_{\lambda g}\| - \widehat{\mu}_g}{\widehat{\sigma}_g}, \qquad \text{(Equation 4)}$$

in which $\widehat{\mu}_g$ and $\widehat{\sigma}_g$ are the mean and standard deviation estimates of $\|\boldsymbol{\beta}_{\lambda g}\|$ under the null hypothesis that the gene g is not associated with disease risk. In this way, every gene, regardless of its size, contributes equally to the gene set association statistics. To estimate $\widehat{\mu}_g$ and $\widehat{\sigma}_g$, we permute the case and control status B times, and for each permutation we obtain $\|\widehat{\boldsymbol{\beta}}^0_{\lambda g}\|$ via the same λ value as the original data set. We then calculate $\widehat{\mu}_g$ and $\widehat{\sigma}_g$ by the mean and standard deviation of $\|\widehat{\boldsymbol{\beta}}^0_{\lambda g}\|$ s over B permutations.

The gene set association statistic for a gene set S is then defined as

$$T_\lambda = \|\tilde{\boldsymbol{\beta}}\| = \sqrt{\tilde{\beta}_1^2 + \ldots + \tilde{\beta}_G^2}, \qquad \text{(Equation 5)}$$

in which $\tilde{\beta}_g, g = 1,\ldots,G$ are the standardized estimates (Equation 4) for each gene g in the set. Via the same permutation, we can standardize the gene set statistics under the null hypothesis and obtain the B null statistics $T^0{}_b(\lambda)$, $b = 1, 2, \ldots, B$.

To test whether the gene set S is associated with disease risk, we compare the observed statistic T_λ with the null statistics and calculate the p value as

$$\text{p value} = \frac{\{T_b^0(\lambda) \geq T_\lambda; b = 1,2,\ldots,B\}}{B}. \qquad \text{(Equation 6)}$$

Another option is to approximate the p value based on a normal distribution for T_λ under the null hypothesis by $1 - \Phi^{-1}\left\{\frac{T_\lambda - m_G(\lambda)}{sd_G(\lambda)}\right\}$. One could estimate the mean, $m_G(\lambda)$, and standard deviation, $sd_G(\lambda)$, of T_λ under the null hypothesis based on fewer number of permutations than that for nonparametric p values in Equation 6 and thus save on computation time substantially. We summarize GRASS in Algorithm A2 in the Appendix.

Results

Simulations

Comparison with Other Penalty Functions

We simulated different scenarios to evaluate the performance of group ridge logistic regression under the null and alternative hypotheses. In each simulation set, we compared the proposed group ridge penalty with several commonly used penalties: lasso,[20] group lasso,[26,30] group bridge,[27] and conventional logistic regression with no penalty. The lasso penalty is the ℓ_1 norm of parameter values $\sum_{i=1}^{p}|\beta_i|$, in which p is the total number of predictors in the regression. The group lasso penalty uses ℓ_1 norm at the group level and ℓ_2 norm within each group, and it is defined as $\sum_{g=1}^{G}\|\boldsymbol{\beta}_g\|_2$, in which G is the number of groups. The group bridge penalty uses ℓ_γ norm $(0 < \gamma < 1)$ at the group level and ℓ_1 norm within each group. Here we chose $\gamma = 0.5$, and the group bridge penalty is then defined as $\sum_{g=1}^{G}(\|\boldsymbol{\beta}_g\|_1)^{0.5}$.

For group ridge, lasso, group lasso, and group bridge penalty functions, we estimated the βs by minimizing the negated log likelihood plus the corresponding term defined by different penalty functions, similar to Equation 2. We chose the penalty parameter λ for each simulated gene set by AIC criterion. For no penalty, we obtained the β estimates by applying univariate logistic regression

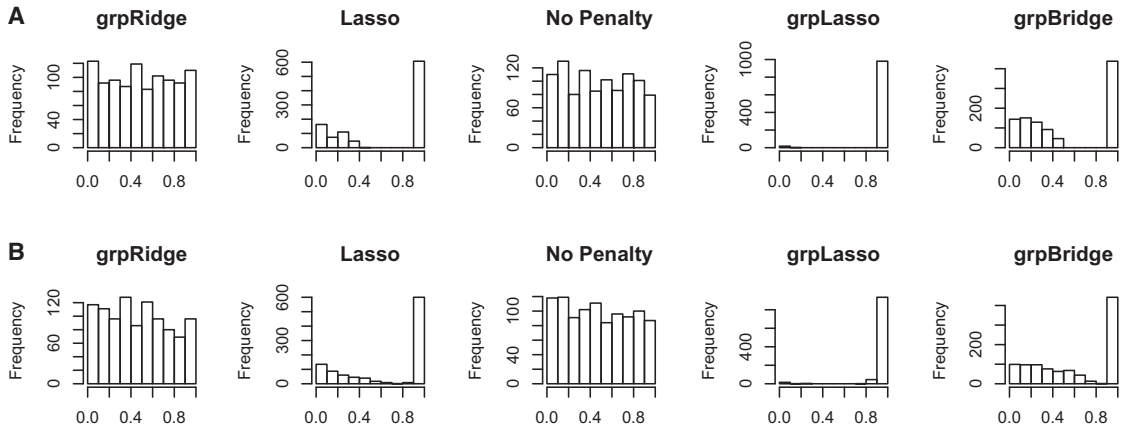

Figure 2. p Values under the Null for Different Penalty Functions with Fixed and Selected λ
(A and B) p value histograms from different penalty functions under the null, with penalty parameter value for null statistics chosen by AIC for each permuted data set (A) and penalty parameter value fixed by the value of λ in the corresponding original pathway (B).

on each of the simulated eigenSNPs. To calculate the p values for each penalty function, we used the permuting phenotype scheme proposed above. We then summarized and standardized gene level statistics according to Equation 4 and formulated the gene set level statistics with Equation 5. To save computation time, we used $B = 100$ permutations to evaluate the general trend of the performance for each method (we used $B = 1000$ in the real data analysis described below to increase accuracy). For each permutation, we fixed the penalty parameter value to the corresponding λ used in calculating the observed statistic for the gene set being tested. We also performed a set of simulations, in which we chose the optimal penalty parameter value by AIC for each permuted data set, and found that the two choices were generally comparable (see Figure 2). Because fixing λ saves substantial computing time, we performed the rest of the simulation study with fixed λ.

In all of the simulations, we simulated 500 cases and 500 controls. For each set of simulations, we generated 100 pathways, with 20 genes each. For each gene, we generated a group of eigenSNPs in which each eigenSNP is normally distributed with effect size being zero under the null (simulation A) and moderate additive effect size, (0.20, 0.12, and 0.10 for simulations B, C, and D, respectively). In simulations B, C, and D, the alternative gene sets consist of 1, 5, and 10 genes, respectively, each harboring one eigenSNP associated with disease risk. Thus, the three alternative simulations have decreasing effect sizes but with increasing amount of signals. In simulations A1, B1, C1, and D1, each gene has k eigenSNPs, in which k is randomly selected from 3 to 20. Thus, the total number of eigenSNPs, p, in a pathway is ~200, which is less than the sample size, $N = 1000$. In A2, B2, C2, and D2, we increase the maximum of gene size from 20 to 100 eigenSNPs and generate k randomly from 3 to 100, so that $p \geq N$.

Figure 2 shows the p value histograms under the null hypothesis for different penalty functions when $p < N$, with varying gene size (each gene harbors 3–20 eigenSNPs) based on 1000 simulations. The histograms show that group ridge gives a nearly uniform distribution. An advantage of p values being uniformly distributed is that we can accurately estimate false discovery rates (FDR) when accounting for multiple hypothesis testing. In comparison, the distributions for lasso, group lasso and group bridge are more like a mixture of a continuous distribution and a point mass at 1. This is because these penalty functions do considerable variable selection, and for many null pathways they don't select any genes. For these cases, the test statistics would be 0, which yields $p = 1$. We also observed a slightly inflated type I error rate under the null, especially for lasso penalty when $p \geq N$ (see Table 1, simulations A1 and A2). When we set all gene sizes equal or increased the number of permutations to $B \geq 1000$, all of the methods controlled the type I error rate. This suggests that when adjusting for gene size in the test statistics, the standard deviations in Equation 4 are not well estimated, particularly if it is a mixture distribution of lasso type of statistics. Therefore, for lasso penalty when adjusting for gene size, a larger number of permutations is required to better control the type I error rate.

In Table 1, when both $p < N$ and $p \geq N$, group ridge is the most powerful among all of the methods, with lasso closely behind. Group lasso deviates from the goal of pathway analysis and does not behave well in any situation. The no-penalty and group bridge approaches have intermediate performances. Lasso penalty often yields a smaller number of selected variables than group ridge and tends to favor pathways with a single SNP, having a large effect. Group bridge penalty selects even fewer variables than lasso and incurs a substantial loss of power. For the no-penalty approach, the performance deteriorates as the signal-to-noise ratio decreases. This is probably because the no-penalty approach, which is really just the ordinary logistic regression, includes many nonassociated SNPs in the test statistic and thus has reduced power to detect disease risk-associated gene sets when the signal is low.

Table 1. Type I Error Rates and Power for Different Penalty Functions

					Type I Error or Power ($\alpha = 0.05$)				
Set	**N[1]**	**p[2]**	**G^*/G[3]**	**Effect Size**	**Group Ridge (GRASS)**	**Lasso**	**No Penalty**	**Group Lasso**	**Group Bridge**
A1	1000	~200	0/20	0	0.06	0.07	0.07	0.01	0.06
A2	1000	~1000	0/20	0	0.07	0.09	0.08	0.06	0.04
B1	1000	~200	1/20	0.2	0.73	0.69	0.29	0.24	0.22
B2	1000	~1000	1/20	0.2	0.61	0.55	0.18	0.08	0.11
C1	1000	~200	5/20	0.12	0.72	0.66	0.37	0.19	0.22
C2	1000	~1000	5/20	0.12	0.51	0.43	0.14	0.06	0.12
D1	1000	~200	10/20	0.1	0.80	0.76	0.54	0.09	0.53
D2	1000	~1000	10/20	0.1	0.51	0.45	0.18	0.09	0.13

Summary of type I error rates (set A) and power (sets B, C, and D) for group ridge, lasso, no penalty, group lasso, and group bridge under various simulation scenarios. The significance level $\alpha = 0.05$.
[1] Sample size.
[2] Total number of eigenSNPs in each pathway.
[3] Number of genes associated with disease risk (numerator) out of the total number of genes (denominator) in each set.

In summary, group ridge appears to have uniformly distributed p values under the null. The uniform distribution doesn't seem to change whether we fix or optimize λ in the null statistic calculation and thus can be implemented with fixed λ to increase computational efficiency. It provides good power when the number of signals is relatively small, and the effect size is moderate even in the "large p small n" situation. We have also examined the approximate p values based on the asymptotic normal distribution with a moderate number of permutations ($B = 100$). The performance is comparable with nonparametric p values, though with much less computation time. However, the asymptotic theory of group ridge has not yet been established. Hence, in applications in which the test statistics may not follow a normal distribution, it would be prudent to use nonparametric p values.

Comparison with Other Gene-Set-Based Approaches

We also simulated different scenarios to compare the proposed GRASS algorithm with recently published gene-set-based approaches. These include the method by Wang et al.,[7] the PLINK gene set option,[18] and ALIGATOR by Holmans et al.[9] Briefly, the idea behind the method by Wang et al. is to assign the top individual SNP association statistic within the gene as the statistic of the gene and to rank all the genes by significance. The method then compares the distribution of the ranks of genes from a given pathway to that of the remaining genes via a weighted Kolmogorov-Smirnov test, with greater weight given to genes with more extreme statistic values. PLINK[18] offers an option to perform gene set analysis (we termed this "Plink"). For each pathway, Plink selects up to N_{snp} "independent" SNPs (here independent is defined as the pair-wise r^2s all below a certain threshold) with marginal association p values less than a predetermined threshold. The method calculates the pathway statistic as the average of the test statistics from the selected SNPs. The significance levels for pathways are determined by permuting the phenotypes, and it compares the observed pathway statistics to the null statistics calculated from permutation. ALIGATOR[9] is similar to Plink in the sense that both use a preselected p value threshold to define a set of significantly associated SNPs. Plink averages the test statistics over all of these SNPs, whereas ALIGATOR counts the number of genes in a pathway that contains these SNPs, with each gene counted only once regardless of the number of significant SNPs in the gene. Instead of permuting phenotypes to establish the null distribution as in Plink, ALIGATOR uses resampling of SNPs. ALIGATOR thus only requires a p value or summary statistic from each SNP as input.

Our simulation is based on real colon cancer GWAS data from the Women's Health Initiative (WHI) study.[32] A more detailed description of the data set can be found in the next section. We chose a random KEGG pathway with 20 genes, the HSA00534 heparan sulfate biosynthesis pathway, as our basic pathway to simulate different scenarios. Different methods have slightly different ways of defining a gene region and assigning SNPs to genes. In this simulation, we adopted the definition of a gene region suggested by Holmans et al.[9] to make the SNP assignments consistent for all methods: SNPs that are within 20 kb of the exons of a gene are assigned to the gene. In simulation E1, we chose the five smallest genes (ranging from 4 to 11 SNPs) from this pathway, and for each gene we randomly selected one SNP (we call it the "tagging" SNP) and simulated one causal SNP, which is in LD, with the tagging SNP (maximum $r^2 = 0.8$). We then simulated the case-control status based on the model $\text{logit}\{\Pr(Y = 1)\} = \sum_{i=1}^{5} \beta_i \text{SNP}_i + \varepsilon$, in which SNP_is are the simulated causal SNPs with the log odds ratios β_is generated from $U[1.3, 1.4]$ and ε follows a standard normal distribution. These simulated casual SNPs are not included in the SNP data,

Table 2. Power Comparison with Existing Pathway Approaches

Set	G[1]	p[2]	Causal SNPs	α Level	Power			
					Wang	Plink	ALIGATOR	GRASS
E1	20	444	5[3]	0.05	0.48	0.68	0.41	0.87
				0.01	0.24	0.35	0.24	0.79
E2	20	444	5[4]	0.05	0.54	0.92	0.55	0.91
				0.01	0.18	0.50	0.16	0.80
F1	30	2286	same as in E1	0.05	0.29	0.44	0.18	0.87
				0.01	0.20	0.23	0.02	0.80
F2	30	2286	same as in E2	0.05	0.32	0.80	0.33	0.87
				0.01	0.17	0.61	0.13	0.73

Summary of power comparison of Wang et al.,[7] Plink,[18] ALIGATOR,[9] and the proposed GRASS under various simulated scenarios.
[1] Number of genes.
[2] Total number of SNPs.
[3] One causal SNP in each of the five smallest genes (range: 4–11 SNPs).
[4] One causal SNP in each of the five largest genes (range: 28–91 SNPs).

nor are they in the subsequent gene set analysis. However, the tagging SNPs are kept. For genes in which SNPs are in LD, many of them may be associated with disease risk. With this simulation, we kept the real LD structures in the data and the potential moderate correlation structures of a real biological pathway. In simulation E2, we chose the five largest genes (ranging from 28 to 91 SNPs) to embed the causal loci and simulated the disease status, similar to E1. Note that here the definition of large versus small genes is defined by the number of SNPs in the gene. Therefore, a larger gene likely has more SNPs in LD with the causal locus than a smaller gene. In simulations F1 and F2, we kept the same structures as in E1 and E2, respectively, but modified the pathway size by adding ten random very large genes with 105–290 SNPs to the pathway. The total number of SNPs in the pathway HAS00534 heparan sulfate biosynthesis is 444 (in E1 and E2), and with the ten added genes, the total number of SNPs in the pathway is 2286 (in F1 and F2). These simulations represent four different scenarios: in E1, the signals reside in small genes with few SNPs in LD, whereas in E2, the signals are in larger genes with more SNPs in LD. In simulations F1 and F2, both the number of nonassociated genes and the number of nonassociated SNPs are increased. This will reduce the signal-to-noise ratio in F1 and F2. For each simulation scenario, we generated 100 data sets. We want to point out that we didn't simulate scenarios with which the pathway has only one significantly associated SNP in one gene, because we think such signals would likely be picked up by marginal association analysis and do not need pathway analysis for detection.

We evaluate the performance of Wang et al., Plink, ALIGATOR, and our proposed GRASS algorithm. For Plink and ALIGATOR, $p < 0.01$ was used to define the significance of the SNPs. For Plink, SNPs with LD $r^2 > 0.5$ are filtered out. All methods control the type I error rate (data not shown). Because of the discrete nature of the test statistic of ALIGATOR, which is defined as the number of genes "hit" by significant SNPs, ALIGATOR can be conservative when the number of genes in a pathway is small or when SNP significance threshold is set to be high.

Table 2 shows the power comparison of these methods in the four simulated scenarios. It can be seen that for signals residing in small genes (simulations E1 and F1), GRASS is more powerful than all other methods. For signals residing in large genes, Plink and GRASS have comparable power when $\alpha < 0.05$, and both perform better than the Wang et al. method and ALIGATOR. When we choose a more stringent significant cutoff, $\alpha < 0.01$, the power of GRASS is much higher than all other methods. As the number of nonassociated genes increases (in F1 and F2), all methods lose power, with GRASS losing the least. This is because GRASS selects disease-associated eigenSNPs within a gene while shrinking the contributions of genes to the pathway statistic. So when there are more nonassociated genes added to the pathway, GRASS will shrink the contributions of these genes to a small amount compared to those associated with disease risk, regardless of the number of SNPs in the gene. Thus, those large and nonassociated genes did not hurt as much to the power of the GRASS method as to other approaches.

It is interesting to see that Plink has good power when the signals reside in large genes. This is probably because in large genes, more SNPs are in LD with the causal locus than in small genes, which makes Plink less likely to miss the region that harbors the causal locus. Another possible reason may be that even though Plink filters out SNPs in high LD (here $r^2 > 0.5$), the remaining selected SNPs are not absolutely "independent"; there might still be small to moderate LDs among the selected SNPs. Because Plink pathway statistic is defined as the average statistic of all the selected SNPs, the small to moderate LDs among those

Table 3. Top-Ranking KEGG Pathways Associated with Colon Cancer Risk in the Women's Health Initiative Sample

Rank	KEGG Number	Pathway Name	No. Genes/No. eigenSNPs	p Value
1	HSA00760	Nicotinate and nicotinamide metabolism	23/602	0.015
2	HSA04350	TGF-beta signaling	89/2912	0.035

Top-ranking KEGG pathways that are associated with colon cancer risk at significance level $\alpha = 0.05$. p values are calculated from Equation 6 based on 1000 permutations.

SNPs, particularly if they are also in LD with the causal SNP, may help boost the power.

The powers of Wang et al. and ALIGATOR are comparable, regardless of whether the signals are in small genes or in large genes. This is because the gene set statistics for both methods use essentially only the strongest signal within each gene. Wang et al. chooses one SNP within each gene that has the maximum association test statistic. ALIGATOR counts a gene only once, even if there are multiple SNPs in the gene passing the p value threshold. Neither method makes use of the LD information, and as a result, they may lose power compared to other methods in situations. For example, multiple SNPs in a gene are in LD with the causal SNP or multiple independent causal SNPs associated with disease risk in a gene, although we didn't simulate the latter case. In contrast, GRASS is more flexible in terms of number of SNPs (or eigenSNPs) being selected in a gene, and thus it is less likely to miss these multiple disease-associated SNPs in the gene. Plink's test statistic is based on SNPs, not on genes, and therefore is also able to account for multiple associated SNPs in a single gene.

A Colon Cancer Genome-wide Association Study

We applied the GRASS algorithm to a colon cancer case-control study nested within the multicenter WHI study.[32] The data set contains 483 cases and 530 controls, frequency matched based on age. All participants are female and self-reported as white. Samples were genotyped with the Illumina HumanHap550 Genotyping BeadChip.[33] Samples with call rate < 98% were excluded; SNP exclusion criteria was call rate < 98%, MAF < 0.05, or deviations from Hardy Weinberg equilibrium ($p < 0.0001$), resulting in 392,361 SNPs. The quantile-quantile plot shows that the p values for marginal association of log additive model adjusting for age and the first three major principal components derived by using EIGENSTRAT[22] closely follow the 45° line (see Figure S1), with a genomic control value of 1.01. Therefore, we only used the first three major principal components to account for any potentially hidden structure in the pathway analysis. SNPs were assigned to nearby genes by relative distance (see Supplemental Gene Definition). Pathways were defined by using the KEGG database.[4] There are a total of 200 KEGG human disease pathways (details can be found on the KEGG website). We restricted our analysis to pathways with at least ten genes, in line with Efron and Tibshirani.[13] After exclusions, there are 170 KEGG pathways with 10–253 genes. The significance level 0.05 is used throughout the analysis.

Two pathways were identified as significant via GRASS (Table 3). A list of the top ten pathways, all of which have $p < 0.1$, is given in Table S1. The highest ranked pathway is the nicotinate and nicotinamide metabolism pathway ($p = 0.015$), followed by the transforming growth factor beta (TGF-beta) signaling pathway ($p = 0.035$). Neither is significant at level 0.05 after adjusting for multiple comparison with the Bonferroni correction. Both pathways have a potential biological role in colon cancer etiology. Nicotinamide and nicotinate are the two main forms of niacin (vitamin B3) and are precursors of nicotinamide adenine dinucleotide (NAD) and nicotinimide phosphate adenine dinucleotide (NADP). NAD and NADP are cofactor enzymes involved in cellular redox reactions.[34] Furthermore, NAD has been shown to play a role in signaling pathways involved in DNA repair, intracellular calcium signaling, and transcriptional regulation.[34,35] Genes in the nicotinate and nicotinamide metabolism pathway have been shown to be differentially expressed in colon cancer cells.[34] The TGF-beta signaling pathway is commonly altered in human cancers.[36] The pathway signals through the TGF-beta serine or threonine kinase receptors and downstream intercellular proteins of the SMAD transcription factor family[37] to inhibit cell proliferation and induce apoptosis; the pathway also induces tumor progression via cell differentiation, migration, and adhesion.[36] See Supplemental Gene Lists for the top two pathways by GRASS algorithm.

We investigated the sensitivity of the GRASS algorithm to the choice of convergence criterion and found that the results of our colon cancer analysis are largely unchanged with various choices of convergence criterion unless the estimated penalty parameter, λ, changes dramatically under different criteria. With a relaxed convergence criterion, the λ estimates tend to be larger, leading to more penalization and more conservative p value estimation.

We also applied Wang et al.,[7] Plink,[18] and ALIGATOR[9] pathway methods to the same GWAS data. The Wang et al. approach identified a total of 15 significant pathways (see Table S3). None was significant after Bonferroni correction, and the minimum FDR was 0.49. The top pathway is valine leucine and isoleucine degradation. There is some observation that the expression of genes in the valine leucine and isoleucine degradation pathway may be decreased in metastatic tissue from colon cancer cases.[38]

For Plink, we chose the default $r^2 > 0.5$ as the SNP LD filtering criterion and 0.001 as the SNP p value threshold. Eight pathways were found to be significant (see

Table S3), and none were significant after multiple testing correction. Among the eight pathways, the top one is the notch signaling pathway, which plays a key role in cell fate determination. Similar to the TGF-beta pathway, several genes in this pathway are up- or downregulated in colon cancer tissue. Interestingly, one mechanism of notch signaling is to inhibit TGF-beta signaling.[39,40] We tried two other thresholds, 0.005 and 0.0001, for p values. A total of eight and five pathways, respectively, were identified. Among these, five pathways were identified by all three p value thresholds (and they are the top five pathways). We also tried another r^2 filtering criterion of 0.2. The results were largely not changed.

When using ALIGATOR to analyze our data, we chose the same p value threshold, 0.001, as in Plink and found only one pathway, O-glycan biosynthesis, as significant (see Table S4). The O-linked mucin type glycans are often altered in colonic disease, including colon cancer. Alterations in O-glycans lead to changes in the interactions between the intestinal cells and their surrounding microenvironment. These changes may have oncogenic effects.[41,42] We also used 0.005 and 0.0001 as the thresholds for p values and found four and five pathways, respectively; none are overlapped.

In terms of computation efficiency, ALIGATOR is the fastest and takes about 1.5 hrs to finish the analysis of the WHI data, using 12 computer nodes with 8 processors each. As a comparison, for the same data set under the same computing power, the proposed GRASS method takes ~5 hr and the Wang et al. approach, based on logistic regression, takes ~24 hr. Plink, as a software, is less amenable to parallel computing. If it can be parallelized, the computation time should be similar to GRASS.

Discussion

We have developed GRASS, an algorithm that performs a novel form of regularized logistic regression to assess the concerted association of genes with disease risk. The regularization has a dual function: selecting SNPs within a gene by lasso penalty while simultaneously shrinking the estimates of the genes by ridge penalty. The method is most powerful when there are several genes in the pathway associated with disease risk.

The GRASS algorithm tests the null hypothesis that none of the genes in a gene set harbor SNPs associated with disease risk. To test this hypothesis, we estimated the null statistics from permuting phenotype labeling. A related null hypothesis[12] is that a gene set is not more enriched with disease-associated genes than a randomly sampled set of genes. To test the latter hypothesis of enrichment, one can adapt the resampling procedure to randomly sampling genes from the genome. In gene expression data analysis, there are some potential differences between testing the two hypotheses.[12] This is because many genes are differentially expressed and gene-gene correlations are relatively common in expression data. When testing the latter hypothesis of enrichment by resampling genes, there is also a concern about inflated type I error rate[12,15] if gene-gene correlation is not adequately taken into account. Unlike expression data, GWAS signals are much more sparse, and intergene correlations are rarely observed. The weak gene-gene correlation is because long-range intergene LD is relatively uncommon. In our study of GWAS data, we found that the difference between testing the two hypotheses is relatively small.

In testing either hypothesis, it is important to adjust for gene size. Because of multiple comparisons, a larger gene is more likely to produce significant associated SNPs than a smaller gene. If gene size is not properly adjusted, a bias is likely to occur, and p values of gene set statistics may be correlated with gene sizes.[7] In our algorithm for gene set association, we permuted phenotypes and standardized the statistic contributed by each gene while testing the significance of pathways with the same permutation data sets. Interestingly, even after adjustment of gene size, we found a modest correlation (Pearson correlation = 0.191) between our estimated pathway p values and the number of eigenSNPs in a pathway. This nonzero correlation suggests that there might be more causal variants in larger pathways with longer genes. In the resampling gene procedure for testing enrichment, we can additionally permute the phenotype to standardize the test statistic to adjust for gene size. However, this is rather costly in computation time. Alternatively, a weight function may be imposed in the regularize regression in Equation 2 so that longer genes are subject to a larger penalty term; however, the choice of the weight function may become ad hoc.

Notably, the GRASS algorithm can be applied regardless of how one groups SNPs to genes. To group SNPs to genes, one can use absolute genetic distance (e.g., SNPs within 50 kb of the exons of a gene are allocated to the gene[7]) or relative distance that allocates each SNP to the nearest gene. We chose the latter, because it is more comprehensive and flexible. Although grouping by relative distance could result in misclassification and/or false enrichment with similar SNP representation for nearby genes, this is less of a concern when considering functional gene sets, because nearby genes do not often belong to the same functional set. The GRASS algorithm can also be applied regardless of how a gene set is defined, for example, whether the definition is based on other informatics databases (e.g., GO, MsigDB) or gene networks constructed from other biologic information (e.g., gene-expression or protein-protein-interaction networks).

Several approaches have recently been proposed for pathway analysis that use GWAS data. We compared our GRASS algorithm with three of them: Wang et al.,[7] Plink,[18] and ALIGATOR.[9] Through simulations, we showed that our GRASS method generally has good power compared to other approaches, even when the signal-to-noise ratio

is low or when signals reside in smaller genes with modest LD. In the real data analysis, we can see that different methods identified different pathways. This finding is not unexpected, because different methods are powerful in detecting different types of pathways. Ideally, there would be a benchmark data set with known pathway effects to compare all of these methods. However, to the best of our knowledge, such a data set is currently not available. Alternatively, one can validate these findings in an independent data set, which also may need to be developed.

Wang et al. uses a weighted Kolmogorov-Smirnov test to assess whether the distribution of the ranks of genes from the pathway differs from the rest of the genes. Thus, a significant pathway may not necessarily indicate that the pathway is associated with disease risk. In addition, Wang et al.'s choice of weight may favor pathways in which one or a few SNPs have very large test statistics. Other than choosing the weight, the Wang et al. approach, in fact, does not need any other choices to perform the test.

Both Plink and ALIGATOR, on the other hand, require that one preselect associated SNPs by using a threshold. Plink simply averages the test statistics of SNPs that exceed the threshold, whereas ALIGATOR counts the number of genes that contain such SNPs. As expected, both methods can be sensitive to the choice of the threshold. An advantage of ALIGATOR is that it only needs p values or summary statistics from tests of SNP associations and thus is particularly useful when individual-level SNP data are not available, for example, in a large-scale meta analysis. However, because it does not take into account the sample variation as in other approaches, the test can be sensitive to the SNP significance threshold if the sample size is small to moderate, as in the case of WHI data. Plink is quite powerful if signals reside in genes with dense SNPs. When the signal-to-noise ratio decreases, the power also does not reduce as much as the Wang et al. approach or ALIGATOR, particularly for large genes. This is probably because Plink makes use of the small to moderate LDs that are often present in large genes. The GRASS algorithm, which does SNP selection within genes, appears to be the least affected when the signal-to-noise ratio is reduced.

We applied the GRASS algorithm to a colon cancer GWAS data set, using our GRASS algorithm to identify disease risk-associated gene sets defined by KEGG pathways. Although none of the pathways were significant after adjustment for multiple testing, the top pathways have putative functional connection to colon cancer. In particular, the second ranked pathway, the TGF-beta signaling pathway, involves signal transduction and regulation of cell proliferation. Based on our gene definition, the TGF-beta pathway includes three chromosomal regions previously identified in GWAS studies as being associated with risk of colorectal cancer: *SMAD7* (rs4939827),[43] 8q24/*MYC* (rs6983267),[44–46] and *BMP4* (rs4444253).[47] In fact, the TGF-beta pathway was recently implicated in colorectal cancer based on the ten common genetic variants identified from previous GWAS.[48]

It is clear that findings that use the GRASS algorithm or other approaches need to be replicated in an independent data set via exactly the same approach as performed in the original data set. Alternatively, replication could be done via targeted genotyping, in which specific SNPs are selected to validate the pathway findings. For example, we can select SNPs that are most correlated with the eigenSNPs that have nonzero effects in the test statistic of an identified pathway. Another way to select SNPs for replication is that we can first detect significant genes within an identified pathway by comparing the gene statistics in Equation 4 with the null statistics obtained from the same permutation as in the gene set association analysis. Then we can choose the most significant and nonredundant SNPs in these significant genes for replication. Taking our WHI study as an example, there are ten genes that are significant at level 0.1 in the TGF-beta signaling pathway. To replicate this finding in an independent replication study, we suggest to choose the most significant and nonredundant SNPs, e.g., $r^2 \leq 0.8$, per significant gene.

The statistical framework presented by GRASS is general and flexible. Because it is regression based, it can easily accommodate other covariates such as age, gender, and center, as well as major principal components,[22] to account for population substructure. It can also be extended to other high-dimensional settings, when the number of predictors is large and a priori information is available, to allow the grouping of predictors. Both the direction of effects in pathways and the relationship among the genes within a pathway may be incorporated in studying gene-gene interactions. Another extension may include the joint analysis of multiple related pathways. Given that signals are usually sparse in GWAS, such joint analysis could be powerful and illuminating. Research in pathway analysis is rapidly evolving, and many methods have been proposed to assess the association to disease risk of potential factors based on gene sets. We believe that with the fast-growing number of available GWAS data, gene-set-based methods will soon be more fully utilized for identifying pathways associated with disease risk. Identification of such pathways could potentially improve our biological understanding of disease processes and help inform clinical decisions for disease prevention and treatment.

Appendix A

Algorithm A1. A Block Coordinate Descent Algorithm for Minimizing $\mathbf{S}_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$

Initialize $\boldsymbol{\beta}$ to be a zero vector
repeat
  $\beta_0 \leftarrow \arg\min_{\beta_0} S_\lambda(\boldsymbol{\beta})$
  for each gene $g = 1, \ldots, G$ **do**
    find the optimal $\boldsymbol{\beta}_g$ while fixing other $\boldsymbol{\beta}_{g'} (g' \neq g)$,
    $\boldsymbol{\beta}_g \leftarrow \arg\min_{\beta_g} S_\lambda(\boldsymbol{\beta})$
  end for
until some convergence criterion is met

Algorithm A2. GRASS

for any given candidate gene set S **do**
  for each gene $g = 1, \ldots, G$ **do**
    apply PCA on the standardized SNP matrix and obtain the $n \times k_g$ eigenSNP matrix $\mathbf{V}_g$ that represents most, if not all, of the genetic variation in the gene
  end for
  $\mathbf{X} \leftarrow \{\mathbf{V}_1, \ldots, \mathbf{V}_G\}$
  apply group ridge logistic regression on the eigenSNP matrix $\mathbf{X}$ and the phenotype y (Algorithm A1), choose λ by using AIC, and obtain the regularized β estimates
  for $b = 1, \ldots, B$ **do**
    permute phenotype y and obtain y_b, apply Algorithm A1 with the same λ chosen using the original data set, and obtain the estimates $\boldsymbol{\beta}^0$ under the null
  end for
  for $g = 1, \ldots, G$ **do**
    calculate the mean and the standard deviation of $\|\boldsymbol{\beta}_g\|$ under the null, as defined in Equation 3, from $\boldsymbol{\beta}_g^0$
    standardize the observed and the null gene estimates by Equation 4
  end for
  compute the observed and null association statistics, T_λ and $T^0{}_b(\lambda)$, $b = 1, \ldots, B$, for the gene set S by Equation 5
  calculate p value by comparing the observed gene set association statistic with the null statistics by Equation 6
end for

Supplemental Data

Supplemental Data include Supplemental Experimental Procedures, two algorithms, four tables, and one figure and can be found with this article online at http://www.ajhg.org.

Acknowledgments

The authors are grateful to Chris Carlson for helping us with the gene definition and with the possible biological mechanism of germline variation for gene functions, David Duggan and his colleagues at TGen for generating the genotype data, Pei Wang for helpful discussions on sparse regression for high-dimensional data, and Ellen Wijsman for the critical reading and many helpful suggestions on the paper. The authors are also grateful to the editor and the anonymous reviewers for their suggestions and comments, which have greatly improved the paper. The work is partially supported by grants from the National Institutes of Health (AG014358, CA053996, CA011821, CA059045, R25CA94880, and U01ES015089). The Women's Health Initiative is supported by the National Heart, Lung and Blood Institute Contracts (N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-19, 32122, 42107-26, 42129-32, and 44221). The authors thank the WHI investigators and staff for their dedication and the study participants for making the WHI program possible. A listing of WHI investigators can be found at the WHI website. The authors declare no conflict of interest.

Received: October 14, 2009
Revised: April 9, 2010
Accepted: April 28, 2010
Published online: June 3, 2010

Web Resources

The URLs for data presented herein are as follows:

Kyoto Encyclopedia of Genes and Genomes (KEGG) Database, http://www.genome.jp/kegg/pathway.html
Software Package for GRASS Algorithm, http://linchen.fhcrc.org/grass.html
Women's Health Initiative list of investigators, http://www.whiscience.org/publications/WHI_investigators_shortlist.pdf

References

1. Manolio, T.A., Brooks, L.D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. J. Clin. Invest. *118*, 1590–1605.
2. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA *96*, 6745–6750.
3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al; The Gene Ontology Consortium. (2000). Gene ontology: Tool for the unification of biology. Nat. Genet. *25*, 25–29.
4. Kanehisa, M. (2002). The KEGG database. Novartis Found. Symp. *247*, 91–101.
5. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.
6. Lesnick, T.G., Papapetropoulos, S., Mash, D.C., Ffrench-Mullen, J., Shehadeh, L., de Andrade, M., Henley, J.R., Rocca, W.A., Ahlskog, J.E., and Maraganore, D.M. (2007). A genomic pathway approach to a complex disease: Axon guidance and Parkinson disease. PLoS Genet. *3*, e98.
7. Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genome-wide association studies. Am. J. Hum. Genet. *81*, 1278–1283.
8. Baranzini, S.E., Galwey, N.W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B.M., Kappos, L., Polman, C.H., et al; GeneMSA Consortium. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum. Mol. Genet. *18*, 2078–2090.
9. Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Owen, M.J., O'Donovan, M.C., and Craddock, N.; Wellcome Trust Case-Control Consortium. (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am. J. Hum. Genet. *85*, 13–24.
10. Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.
11. Goeman, J.J., van de Geer, S.A., de Kort, F., and van Houwelingen, H.C. (2004). A global test for groups of genes: Testing association with a clinical outcome. Bioinformatics *20*, 93–99.
12. Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S., and Park, P.J. (2005). Discovering statistically significant

pathways in expression profiling studies. Proc. Natl. Acad. Sci. USA *102*, 13544–13549.
13. Efron, B., and Tibshirani, R. (2007). On testing the significance of sets of genes. Annals of Applied Statistics *1*, 107–129.
14. Jiang, Z., and Gentleman, R. (2007). Extensions to gene set enrichment. Bioinformatics *23*, 306–313.
15. Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A.J., Jhangri, G.S., Einecke, G., Famulski, K.S., Halloran, P., and Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics *8*, 242.
16. Song, S., and Black, M.A. (2008). Microarray-based gene set analysis: A comparison of current methods. BMC Bioinformatics *9*, 502.
17. Ackermann, M., and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. BMC Bioinformatics *10*, 47.
18. Purcell, S.M., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
19. Gauderman, W.J., Murcray, C., Gilliland, F., and Conti, D.V. (2007). Testing association between disease and multiple SNPs in a candidate gene. Genet. Epidemiol. *31*, 383–395.
20. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B *58*, 267–288.
21. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc., Ser. B *67*, 301–320.
22. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
23. Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics *12*, 55–67.
24. Frank, I.E., and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion). Technometrics *35*, 109–148.
25. Knight, K., and Fu, W. (2000). Asymptotics for lasso-type estimators. Ann. Stat. *28*, 1356–1378.
26. Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. R. Stat. Soc., Ser. B *68*, 49–67.
27. Huang, J., Ma, S., Xie, H., and Zhang, C. (2007). A group bridge approach for variable selection. Biometrika *96*, 339–355.
28. Zhang, T., and Oles, F. (2001). Text categorization based on regularized linear classifiers. Inf. Retrieval *4*, 5–31.
29. Genkin, A., Lewis, D.D., and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. Technometrics *49*, 291–304.
30. Meier, L., van de Geer, S., and Buhlmann, P. (2008). The group lasso for logistic regression. J. R. Stat. Soc., Ser. B *70*, 53–71.
31. Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Automat. Contr. *19*, 716–723.
32. Hays, J., Hunt, J.R., Hubbell, F.A., Anderson, G.L., Limacher, M., Allen, C., and Rossouw, J.E. (2003). The Women's Health Initiative recruitment methods and results. Ann. Epidemiol. *13*, S18–S77.
33. Steemers, F.J., and Gunderson, K.L. (2007). Whole genome genotyping technologies on the BeadArray platform. Biotechnol. J. *2*, 41–49.
34. Garten, A., Petzold, S., Körner, A., Imai, S., and Kiess, W. (2009). Nampt: Linking NAD biology, metabolism and cancer. Trends Endocrinol. Metab. *20*, 130–138.
35. Bogan, K.L., and Brenner, C. (2008). Nicotinic acid, nicotinamide, and nicotinamide riboside: A molecular evaluation of NAD+ precursor vitamins in human nutrition. Annu. Rev. Nutr. *28*, 115–130.
36. Xu, Y., and Pasche, B. (2007). TGF-β signaling alterations and susceptibility to colorectal cancer. Hum. Mol. Genet. *16*(Spec No 1), R14–R20.
37. Akhurst, R.J. (2004). TGF β signaling in health and disease. Nat. Genet. *36*, 790–792.
38. Gmeiner, W.H., Hellmann, G.M., and Shen, P. (2008). Tissue-dependent and -independent gene expression changes in metastatic colon cancer. Oncol. Rep. *19*, 245–251.
39. Qiao, L., and Wong, B.C. (2009). Role of Notch signaling in colorectal cancer. Carcinogenesis *30*, 1979–1986.
40. Katoh, M., and Katoh, M. (2007). Notch signaling in gastrointestinal tract (review). Int. J. Oncol. *30*, 247–251.
41. Brockhausen, I. (2006). Mucin-type O-glycans in human colon and breast cancer: Glycodynamics and functions. EMBO Rep. *7*, 599–604.
42. Rhodes, J.M., Campbell, B.J., and Yu, L.G. (2008). Lectin-epithelial interactions in the human colon. Biochem. Soc. Trans. *36*, 1482–1486.
43. Broderick, P., Carvajal-Carmona, L., Pittman, A.M., Webb, E., Howarth, K., Rowan, A., Lubbe, S., Spain, S., Sullivan, K., Fielding, S., et al. (2007). A genome-wide association study shows that common alleles of smad7 influence colorectal cancer risk. Nat. Genet. *39*, 1315–1317.
44. Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., et al; CORGI Consortium. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat. Genet. *39*, 984–988.
45. Haiman, C.A., Le Marchand, L., Yamamato, J., Stram, D.O., Sheng, X., Kolonel, L.N., Wu, A.H., Reich, D., and Henderson, B.E. (2007). A common genetic risk factor for colorectal and prostate cancer. Nat. Genet. *39*, 954–956.
46. Zanke, B.W., Greenwood, C.M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S.M., Prendergast, J., Olschwang, S., Chiang, T., Crowdy, E., et al. (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat. Genet. *39*, 989–994.
47. Houlston, R.S., Webb, E., Broderick, P., Pittman, A.M., Di Bernardo, M.C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S., et al; Colorectal Cancer Association Study Consortium; CoRGI Consortium; International Colorectal Cancer Genetic Association Consortium. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat. Genet. *40*, 1426–1435.
48. Tenesa, A., and Dunlop, M.G. (2009). New insights into the aetiology of colorectal cancer from genome-wide association studies. Nat. Rev. Genet. *10*, 353–358.