

Using Biological Knowledge to Discover Higher Order Interactions in Genetic Association Studies

Gary K. Chen* and Duncan C. Thomas

Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, California

The recent successes of genome-wide association studies (GWAS) have revealed that many of the replicated findings have explained only a small fraction of the heritability of common diseases. One hypothesis that investigators have suggested is that higher order interactions between SNPs or SNPs and environmental risk factors may account for some of this missing heritability. Searching for these interactions poses great statistical and computational challenges. In this article, we propose a novel method that addresses these challenges by incorporating external biological knowledge into a fully Bayesian analysis. The method is designed to be scalable for high-dimensional search spaces (where it supports interactions of any order) because priors that use such knowledge focus the search in regions that are more biologically plausible and avoid having to enumerate all possible interactions. We provide several examples based on simulated data demonstrating how external information can enhance power, specificity, and effect estimates in comparison to conventional approaches based on maximum likelihood estimates. We also apply the method to data from a GWAS for breast cancer, revealing a set of interactions enriched for the Gene Ontology terms growth, metabolic process, and biological regulation. *Genet. Epidemiol.* 34:863–878, 2010. © 2010 Wiley-Liss, Inc.

Key words: Bayesian analysis; variable selection; pathway; MCMC

Contract grant sponsor: NIH; Contract grant numbers: R01 ES016813; R01 ES015090.

*Correspondence to: Gary K. Chen, University of Southern California, Health Sciences Campus, NRT 1504, M/C 9601, Los Angeles, CA 90089-9601. E-mail: gary.k.chen@usc.edu

Received 20 May 2010; Revised 11 August 2010; Accepted 9 September 2010

Published online 18 November 2010 in Wiley Online Library (wileyonlinelibrary.com).

DOI: 10.1002/gepi.20542

INTRODUCTION

In genome-wide association studies (GWAS), analyses usually entail a hypothesis-free search for correlations between disease status and each SNP featured on the panel among a set of cases and controls. The estimated relative risks (RRs) for the genetic variants reported so far have been modest, with many explaining less than 1% of phenotypic variation in the study population [Hirschhorn, 2009]. It is encouraging that many of the implicated loci have shown overlap with known pathways: for example 11 of 23 loci related to lipid levels are in genes known to be involved in lipid metabolism [Mohlke et al., 2008]. The primary challenge then is to interpret the remaining loci that may show either a statistically significant association but not point unambiguously to a gene with a relevant function or show only a suggestive evidence of association but have potentially more relevance to a disease process. One possibility is that the conventional approach to GWAS, testing one SNP in a model at a time, is underpowered in the context of common diseases, which are thought to be multifactorial, with some proportion showing evidence for statistical epistasis (i.e. an effect on disease risk departing from additivity or log-additivity). When one considers such higher order interactions between genes and/or environmental factors, the number of variables balloons nearly quadratically with each increase in the order. For example, for a typical GWAS data set with 1 million features, one

would need to consider over 500 billion variables if we limited the analysis only to main effects and second-order interactions. Although exhaustive searches are computationally feasible [Marchini et al., 2005] within the scope of pairwise interactions, moving beyond second-order interactions would require more efficient use of the data. Furthermore, it is preferable to model all SNPs and their interactions at once in a multivariate linear model, since many SNPs will be correlated [Hoggart et al., 2008], but this approach is infeasible, given the fact that the problem would be highly underdetermined (i.e. the number of predictors far surpasses the number of observations).

The standard approach to addressing the need to modeling many variables jointly is penalized stepwise regression, a variable selection technique that seeks to find the set of independent variables that best predicts disease risk. One practical concern is that an exhaustive search across all possible models is computationally feasible only for small data sets, since each model is usually fit using iterative logistic regression. Also, interpretation can be difficult since asymptotic *P*-values from the “best” model are not adjusted for the number of previously tested models. More advanced techniques in variable selection have helped to address these shortcomings.

One strategy aimed at facilitating the interpretation of the effect estimates (e.g. β , the log RRs for each SNP) in a multivariate analysis is known as “shrinkage.” The LASSO penalized regression, for example, generates sparse models by shrinking the values of all elements in β toward zero

based on the value of a tuning parameter, so that many elements that are set exactly to zero are considered not to be associated with a trait [Tibshirani, 1996]. By introducing a small amount of bias in exchange for a large reduction in the variance of the effect estimates, the LASSO infers models of disease risk that are more interpretable than their maximum likelihood counterpart. A drawback to the method is that it does not explicitly provide standard errors on the estimates of β , so a Monte Carlo approach such as bootstrapping is sometimes employed to estimate uncertainty.

Alternatives to LASSO in variable selection that do implicitly model uncertainty are known as stochastic search algorithms. A few examples include evolutionary programming [Ritchie et al., 2003], Monte Carlo logic regression [Kooperberg and Ruczinski, 2005], stochastic search variable selection (SSVS) [George and McCulloch, 1993; Conti et al., 2003], and reversible jump Markov Chain Monte Carlo (RJMCMC) [Green, 1995]. Bayesian methods, such as SSVS and RJMCMC, assess model and variable significance through Bayes model averaging (BMA) [Viallefont et al., 2001], where test statistics are averaged across all models sampled from the posterior distribution. Using BMA, one can simultaneously perform variable selection (e.g. through a variable's posterior probability of inclusion in a model) and shrinkage of unstable maximum likelihood estimates (MLE) toward their prior means (e.g. through the use of their prior distributions). BMA is considered to be more robust to multiple comparisons than methods that rely on a single point estimate, since evidence from all possible models are taken into account, so that estimates from inferior models can dampen those driven by a few "interesting" but unlikely models. Furthermore, computational efficiency can be gained if density functions based on informative prior information can guide the algorithm toward spending a greater proportion of computational effort on regions supported by prior biological knowledge.

In this article we describe an RJMCMC variable selection method designed to efficiently sample models across an enormous search space through the use of biologically motivated priors, allowing investigators to discover plausible associations between a trait of interest and candidate variables, such as main effects and/or interactions of any order (e.g. three-way interactions). Model uncertainty is assessed by averaging parameter values across all posterior draws. We demonstrate the utility of our method through two simulated data sets and a real data set from a study of breast cancer. In the first simulation, we demonstrate how models that use informative priors can enhance the power to detect interactions with no loss in specificity when compared to models with uninformative priors. We also show how our model averaging approach better controls for multiple comparisons when compared to an approach that relies solely on MLE. The second simulation example integrates previous work in modeling folate metabolism with our newly proposed method, again demonstrating how informative priors can be a powerful approach in the discovery of biologically plausible interactions. In contrast to the two simulated data sets, which use continuous intermediate phenotypes (e.g. biomarkers) as priors, our final example using a real data set shows how one can incorporate publicly available gene annotation information to guide our SSVS algorithm in a genome-wide search for higher order interactions. C++ source code and binaries that implement

our method is freely available at <http://www-hsc.usc.edu/garykche/>.

METHODS

MODEL SPECIFICATION

In this section, we provide a description of hierarchical modeling, which serves as the backbone of the variable selection algorithm described in this article. This technique stabilizes MLE of association between covariates and an outcome of interest through the use of prior distributions on these estimates, such that the strength of an association is "distributed" across all other covariates that share the same prior mean. The prior distributions are defined through the use of external information such as genomic annotations. Hierarchical modeling has been shown to enhance rankings in a GWAS by incorporating such prior information [Chen and Witte, 2007; Lewinger et al., 2007]. The model can be represented in terms of two or more levels, defined as follows.

First level of the model: a generalized linear model for subject data. Suppose that there are n individuals with available phenotype data in an observational study. The first level is a generalized linear model typically used in genetic epidemiology to measure the association between an outcome of interest Y and p predictors or covariates such as genotypes, environmental exposures, or higher order interactions. Throughout the remainder of this article, we will use the term model variable to correspond to any covariate specified in the first level. Note that model variables are constrained to be "hierarchical" when interactions are considered, so that a higher order interaction can only be included in the model when its constituent main effects are also present. When considering interactions, however, p can be much larger than n , so that the parameter of interest β , measuring association between Y and the p model variables, is not estimable. For example, if 1,000 SNPs and their second-order interactions are considered, $p = 1,000 + ((1,000)(999))/2 = 499,500$. To address this issue at any step of the procedure, we consider only a subset of size m , where $m < n$, of the p model variables. The values for the model variables are then stored in an n by m matrix X and the vector Y of length n stores the outcome variable, such as disease status. The relationship between Y and X is thus specified as:

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_{1..m}X. \quad (1)$$

The unobserved vector β is estimated using logistic regression.

Second level of the model: prior distributions on β . The second level of the hierarchical model defines a prior distribution on the vector β . One may simply assume that all elements of β arise from a common distribution so that they are exchangeable with the same mean and variance. However, it is more likely that most model variables considered are not associated with a trait, where such associations may be due to random error or spurious correlations (e.g. linkage disequilibrium, incomplete population admixture, and confounding by environmental variables). On the other hand, others may be more plausible, such as SNPs in conserved regions, missense mutations, known environmental risk factors, etc. We can harness biological or epidemiological expert knowledge so

that such information can be used to help define the means, variances, and probability of model inclusion at the level of each model variable. Thus, for any given model variable indexed k , β_k can be defined to have distribution:

$$\beta_k \sim \pi^T Z_k + \theta_k + \phi_k. \quad (2)$$

where Equation (2) describes a mixed model as formulated in [Besag et al., 1991] with fixed effect $\pi^T Z_k$ and random effects

$$\theta_k \sim N(0, \sigma^2), \quad \phi_k \sim N\left(\bar{\phi}_{-k}, \frac{\tau^2}{v_k}\right), \quad (3)$$

and

$$\bar{\phi}_{-k} = \frac{\sum_{j=1}^m \phi_j A_{jk}}{\sum_{j=1}^m A_{jk}}. \quad (4)$$

The term $\pi^T Z_k$ quantifies our belief that certain underlying phenotypes (e.g. gene expression, enzyme kinetic rates, protein levels, etc.) or expert knowledge (e.g. annotation) influence disease risk through β . The design matrix denoted Z encodes this information and the vector π empirically estimates the correlation between β and Z . In addition, we may have additional knowledge about the relationship between model variable k and other model variables, which can be specified in a connectivity matrix A such that a matrix element A_{jk} of 1 represents a connection between variable j and k , 0 otherwise. Information in A is modeled using a spatial autoregressive distribution in the random effects component ϕ_k . The mean of this component $\bar{\phi}_{-k}$ is taken across all of k 's neighbors and the variance of this component τ^2 is scaled by $1/v_k$, v_k being the number of neighbors of k (i.e. the sum of row or column k in A). Finally, any residual variation not accounted for by $\pi^T Z_k$ and ϕ_k is captured in the random effect component θ_k with variance σ^2 .

We now discuss our model proposal density function that defines the probability of changing the size of the current model by one. Let M' denote the new set of variables after either adding a newly proposed variable (indexed as k) into the current model M or removing an existing variable k from M . Then a proposal density can be written as

$$P(M \rightarrow M') = \begin{cases} \Phi^{-1}(z_k) & \text{if var } k \text{ added to } M \\ 1 - \Phi^{-1}(z_k) & \text{if var } k \text{ deleted from } M \end{cases} \quad (5)$$

$\Phi()$ being the univariate probit function (i.e. the inverse cdf of a standard normal distribution), where the parameter z_k is drawn from a univariate normal distribution

$$z_k \sim N(\mu_k - \mu_{\text{baseline}}, 1), \quad (6)$$

with two tuning parameters defining its mean. If both parameters are zero, then the probability of including a new variable into the model is 0.5. To encourage parsimonious models, we fix the tuning parameter μ_{baseline} to be a constant positive-valued penalty term and allow μ_k to vary for each variable that is considered for inclusion. Thus, higher values of μ_k will reduce the effect of the penalty term. We assign μ_k as:

$$\mu_k = \frac{|\pi^T Z_k + \bar{\phi}_{-k}|}{\sigma^2 + \frac{\tau^2}{v_k}}. \quad (7)$$

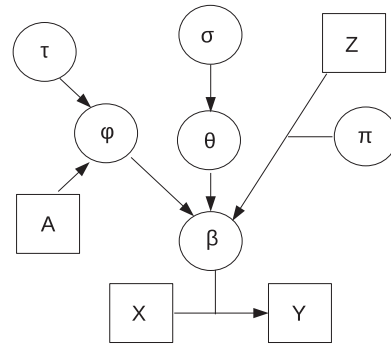


Fig. 1. Directed acyclic graph of the parameters in the hierarchical model.

This combines information from the Z and the A matrices by taking the sum of the two components defining the prior mean of β , standardized by their variances. For the penalty term μ_{baseline} , we chose a value motivated by the BIC statistic:

$$\text{BIC} = -2 \ln(L) + m \ln(n), \quad (8)$$

where $\ln(L)$ is the log-likelihood of the current model M . Since $\ln(L)$ is equal for all models under the null, the difference in the log-likelihood between M and a proposed model M' (with $m+1$ variables) is $\ln(n)$, which is distributed as χ_1^2 . Under the standard BIC, the probability of accepting M' is $F_{\chi_1^2}^{-1}(\ln(n))$, the cdf of the χ_1^2 distribution. We assign μ_{baseline} to reflect this probability by converting back to its quantile on the normal distribution:

$$\mu_{\text{baseline}} = \Phi(F_{\chi_1^2}^{-1}(\ln(n))). \quad (9)$$

Note that these means and variances in Equation (7) are estimated empirically from the data so that information encoded in Z and A that minimally support the data lead to values of μ_{prior} close to zero. Under these circumstances, the algorithm encourages model parsimony resembling those penalized by the standard BIC, but allows some stochastic variation in this proposal probability through the random draw in Equation (6).

The directed acyclic graph in Figure 1 shows how the parameters in the hierarchical model relate to one another through their dependencies. Squares represent observed data and circles represent unobserved latent parameters that are to be estimated based on their full conditional distribution.

POSTERIOR INFERENCE

Based on the hierarchical model described above, a posterior density can be used as the objective function within the framework of a RJMCMC algorithm. The algorithm efficiently samples a large number of realizations from the posterior distribution, using a Metropolis-Hastings ratio proposal step to determine whether to accept or reject a proposed change in the number of variables included in the current model. The steps taken are detailed in Appendix.

We compute posterior estimates of the various parameters of interest (e.g. π , σ^2 , τ^2 , and β) by averaging their values across all models sampled from their posterior distribution. In terms of measuring the importance of

model variable k , we calculate a Bayes Factor (BF), which quantifies how much more significant a variable is in light of observed data (e.g. disease status) than in the prior [Kass and Raftery, 1995]. We estimate the BF by dividing the posterior odds by the prior odds:

$$BF_k = \frac{\tilde{P}(\sigma, \tau, \pi, \beta|Y, X)/(1 - \tilde{P}(\sigma, \tau, \pi, \beta|Y, X))}{P(\sigma, \tau, \pi, \beta)/(1 - P(\sigma, \tau, \pi, \beta))}. \quad (10)$$

Values in the denominator are estimated by sampling from the marginal prior density following the same procedure as for sampling from the posterior (as described in the Appendix), with the exception that the MLE $\hat{\beta}$ from the first level are not calculated by fitting the model to the data, but rather are randomly sampled from a multivariate standard Gaussian at each iteration. We present results where at least one of the scenarios of interest yielded a BF of at least 100, which is conventionally deemed as “definitive” support for one hypothesis over another [Kass and Raftery, 1995].

RESULTS

SIMULATION 1: USING EXTERNAL INFORMATION TO ENHANCE POWER AND SPECIFICITY

We compared several scenarios through simulations to evaluate how external information enhances the search for higher order interactions. Scenarios consisted of varying types and degrees of prior information available to the model as well as differing disease penetrances. For each scenario tested, 100 data replicates were generated and evaluated. We sought to answer three primary questions: (1) How does specification of external information in our proposed hierarchical model impact the power and specificity of the method? (2) How does hierarchical modeling influence posterior estimates of β ? (3) How does our Bayesian model averaging approach perform in comparison to a conventional exhaustive search strategy?

To generate data for each replicate in simulations, we randomly sampled genotypes for 10,000 individuals across 14 independent SNPs based on expected genotype frequencies under Hardy-Weinberg equilibrium. The first six SNPs (indexed from 0 to 5) were considered to be involved in disease risk. We assigned four two-way interactions (0*1, 0*2, 3*4, and 3*5) among these six markers to be jointly causal for disease risk, where each interaction was comprised of a moderate-frequency variant and a second variant of varying allele frequency, as shown in Table I. Each causal interaction was then assumed to influence disease risk through a continuous intermediate endophenotype. These endophenotypes were

not used in the analysis of the case-control data (i.e. the first level of the hierarchical model) but rather treated as external information in a separate data set for constructing prior information in the second level of the hierarchical model. Endophenotypes for each individual were assigned by randomly drawing from a mixture of a univariate normal distribution centered on an interaction between s_a and s_b (genotypes at SNPs a and b) with variance one, and a uniform distribution, so that at a causal interaction k between SNP a and b for individual i in endophenotype y_{ik} was defined as:

$$y_{ik} = (1 - b)N(s_{ia} * s_{ib}, 1) + bU(0, 1) \quad b \sim \text{Bernoulli}(P). \quad (11)$$

The vector y_k was standardized to have mean zero and variance one. For each of the four interactions, we simulated y_k using six different values of P , ranging from 0 to 1 in increments of 0.2 in order to simulate correlated but progressively noisier non-causal phenotypes, for a total of 24 phenotypes. For each of the six phenotypes simulated based on an interaction’s genotype means, Table I indicates the variance explained (R^2) by that interaction.

We next simulated disease status based on a multifactorial risk model where the four causal interactions were jointly associated with disease risk. Using only the four true endophenotypes (i.e. where y_k in Equation (11) was simulated at $P = 0$), we assigned disease status based on the model:

$$\text{logit}(Y_i = 1) = \beta_0 + \beta_1 y_{i01} + \beta_2 y_{i02} + \beta_3 y_{i34} + \beta_4 y_{i35}. \quad (12)$$

β_1 through β_4 were specified to share the same specified RR per unit increase of y_k . For the analyses, genotypes and disease status from 8,000 randomly selected individuals were used to fit the logistic regression in the first level of the hierarchical model as first shown in Equation (1):

$$\text{logit}(Y_i = 1) = \beta_0 + \sum_{a=0}^{13} \beta_a s_{ia} + \sum_{a=0}^{12} \sum_{b=a+1}^{13} \beta_{ab} s_{ia} s_{ib}. \quad (13)$$

Note that Equation (13) defines the saturated model of size p . Our MCMC algorithm however, fits only a subset (of size m) of these p variables at each iteration.

Genotypes and intermediate phenotypes from the remaining 2,000 individuals were used to construct prior information in the Z and A matrices in the second level of the hierarchical model. The Z matrix included m rows corresponding to the ensemble of model variables and 25 columns corresponding to the intercept term plus the 24 intermediate phenotypes. Entries in Z were calculated as

$$Z_{kq} = \text{corr}(g_k, y_q), \quad (14)$$

TABLE I. Properties of the four interactions that were simulated to jointly contribute to disease risk through their respective endophenotypes

| Interaction | Allele frequency | | R^2 at varying levels of noise | | | | | |
|-------------|------------------|-------|----------------------------------|-------|-------|-------|-------|-------|
| | SNP 1 | SNP 2 | 0% | 20% | 40% | 60% | 80% | 100% |
| 0*1 | 0.25 | 0.10 | 0.109 | 1e-03 | 1e-05 | 6e-05 | 3e-05 | 8e-06 |
| 0*2 | 0.25 | 0.20 | 0.208 | 3e-02 | 1e-02 | 5e-03 | 2e-03 | 3e-03 |
| 3*4 | 0.25 | 0.30 | 0.263 | 2e-03 | 7e-04 | 4e-04 | 1e-07 | 2e-05 |
| 3*5 | 0.25 | 0.40 | 0.360 | 7e-02 | 3e-02 | 9e-03 | 2e-03 | 7e-03 |

Variance explained by each interaction are shown for the six endophenotypes that were associated to that interaction.

quantifying correlation between model variable genotypes g_k and intermediate phenotype y_q , taken across the 2,000 individuals.

In specifying A , we defined a model where any two model variables that shared a similar profile of correlations across all 24 simulated endophenotypes were likely to share similar roles, so that for a pair of model variables j and k , its similarity was quantified as

$$A_{jk} = \text{corr}(Z_j, Z_k), \quad (15)$$

taken across the 24 endophenotypes.

To address how the algorithm performs in light of external information, we analyzed various models using six different combinations of information in A and Z , and compared power and specificity across these models. Note that for models that included information from Z , we used only 5 of the 25 columns specified in Equation (14) in order to explore the effects of mis-specification of Z . Under Model 1, we specified β to be normally distributed around zero and uncorrelated so that no information from Z or A was incorporated into the model, which is equivalent to a ridge regression style prior. Model 2 included only the correlations with the four *causal* intermediate phenotypes in Z (i.e. 0% noise) and did not use the A matrix. Model 3 included only the A matrix while not using the Z matrix. Model 4 combined information from Models 2 and 3 by including A and Z . To assess the impact of mis-specification of Z , we varied Model 2 by replacing the four columns in Z with vectors that were computed from the correlated phenotypes with 20% noise (defined as Model 5) and 80% noise (defined as Model 6). Although we evaluated other models as well (i.e. using a Z matrix with 40%, 60% noise), we found that results revealed performance that was intermediate to the 20 and 80% cases as expected, and have omitted them from the presentation of results for simplicity. In order to estimate power and false-discovery rate (FDR), for each BF threshold we tabulated the number of causal variants (4 interactions shown in Table I) and non-causal variants (all other variables) included in the model, taken across all 100 replicates.

Table II summarizes the performance of all the tested models across different disease models by the area under the curve (AUC) statistic. All tested models performed well (AUC > 0.9) for high disease penetrances (RR > 1.3), and had AUC values > 0.8 even with a RR of 1.1. Figures 2–6 illustrate the contrast among the models under the various effect sizes. At an FDR of 0.05 (shown in the ROC figures as a gray vertical line and in Table III), we observed that models with informative priors showed an increase in power over their less informative counterparts as disease penetrance decreased from RR = 1.5, where all models had nearly 100% power to detect the causal interactions.

TABLE II. Area under the curve, estimated across 100 simulated replicates

| Relative risk | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 |
|-----------------------|-------|-------|-------|-------|-------|
| Non-informative prior | 0.826 | 0.953 | 0.964 | 0.995 | 0.999 |
| Informative Z only | 0.917 | 0.973 | 0.992 | 0.997 | 0.999 |
| Informative A only | 0.886 | 0.955 | 0.987 | 0.996 | 0.999 |
| Informative A and Z | 0.925 | 0.975 | 0.992 | 0.997 | 0.999 |
| 20% noise in Z | 0.885 | 0.959 | 0.983 | 0.995 | 0.998 |
| 80% noise in Z | 0.872 | 0.953 | 0.983 | 0.994 | 0.999 |

Specifically, at a RR of 1.4, all the tested models had >90% power to detect the causal variants. At RR = 1.3, Models 2–4, which included the three most informative priors, maintained approximately 90% power while the remaining models provided approximately 85% power. All the models began to suffer in terms of power (<80%) at RR < 1.3, but even at RR = 1.2, the three most informative models still maintained at least a 5% improvement in power over the other three models. Across all RRs, the Z-only model with information from only the causal phenotypes (Model 2) outperformed the A-only model (Model 3), which includes substantially more noise from the 20 additional phenotypes, except for RR = 1.4. In contrast, adding 20% or

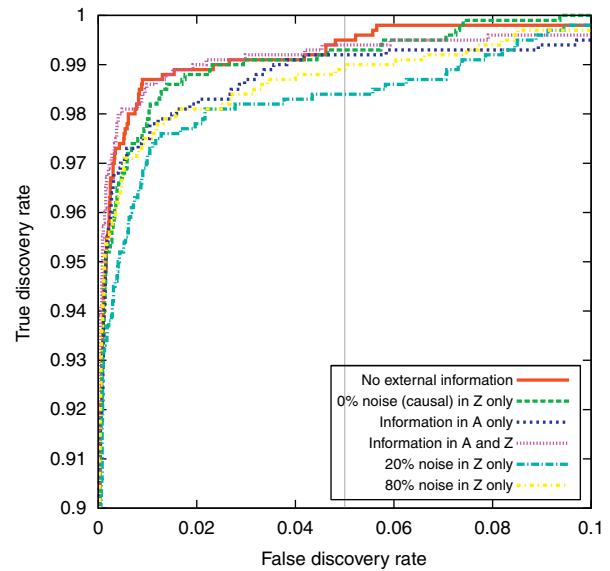


Fig. 2. Receiver operating curve under disease penetrance RR = 1.5. RR, relative risk.

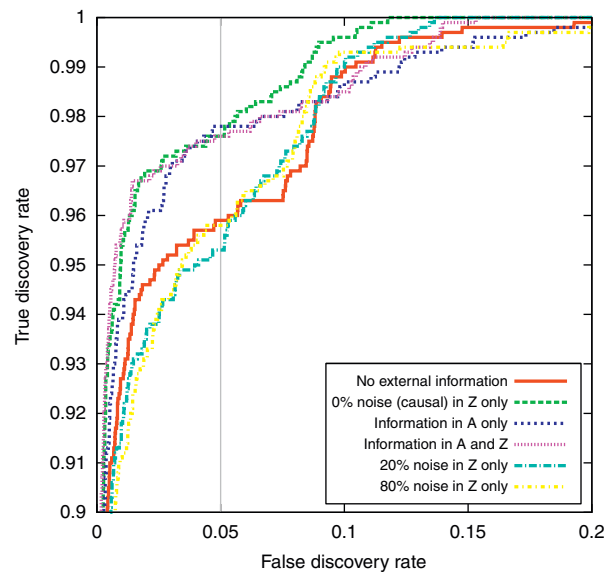


Fig. 3. Receiver operating curve for under disease penetrance RR = 1.4. RR, relative risk.

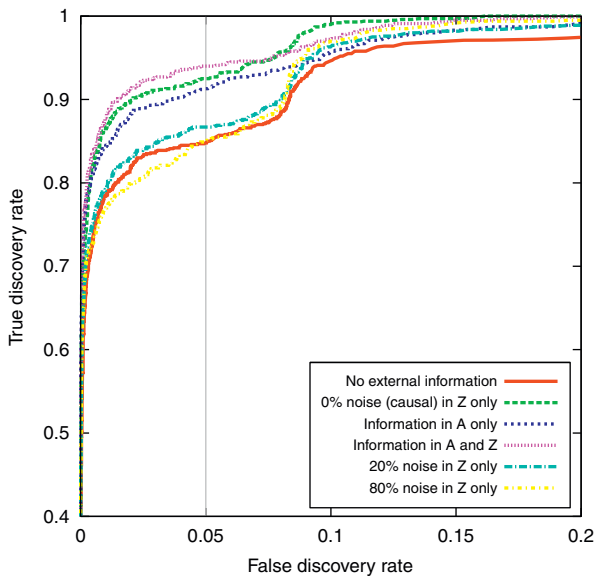


Fig. 4. Receiver operating curve under disease penetrance RR = 1.3. RR, relative risk.

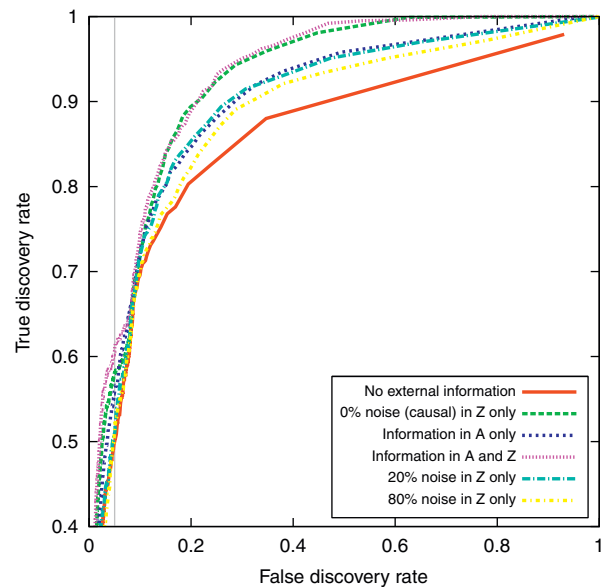


Fig. 6. Receiver operating curve under disease penetrance RR = 1.1. RR, relative risk.

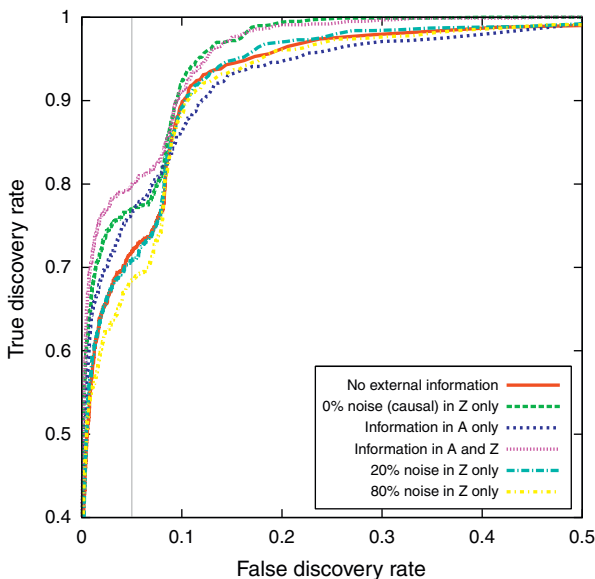


Fig. 5. Receiver operating curve under disease penetrance RR = 1.2. RR, relative risk.

80% noise into the Z-only model (Models 5 and 6) caused the method to perform more poorly than the A-only model across all RRs. In summary, the results show that the priors can be most helpful under moderate RRs, but at the extremes of the range of tested disease penetrances, they appear slightly less useful.

We assessed the potential of hierarchical modeling in providing improved posterior estimates of association between SNPs and disease by comparing posterior estimates of β (using Model 4, the fully informative model) extracted from our method to MLE from a multivariate logistic regression that fits all 14 SNPs and their interactions in a single model. Although β is

TABLE III. Power at an FDR of 0.05, estimated across 100 simulated replicates

| Relative risk | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 |
|-----------------------|-------|-------|-------|-------|-------|
| Non-informative prior | 0.500 | 0.719 | 0.848 | 0.959 | 0.995 |
| Informative Z only | 0.581 | 0.770 | 0.925 | 0.976 | 0.993 |
| Informative A only | 0.559 | 0.763 | 0.913 | 0.978 | 0.992 |
| Informative A and Z | 0.606 | 0.799 | 0.940 | 0.976 | 0.994 |
| 20% noise in Z | 0.510 | 0.708 | 0.867 | 0.953 | 0.984 |
| 80% noise in Z | 0.510 | 0.685 | 0.849 | 0.958 | 0.989 |

shrunk toward zero under our method, Figure 7 shows that these posterior estimates of β for the non-causal interactions are shrunk closer to zero than the MLE, so that causal interactions are better distinguished from the non-causal interactions. Posterior standard errors for β estimated from our method are less dispersed than in Model 1, as shown in Figure 8.

To consider how Bayesian model averaging compares to a conventional approach when challenged with multiple comparisons, we expanded the null distribution (which had eight SNPs and 87 pairwise interactions that were null already) by including an additional 100 null SNPs along with all their possible pairwise interactions for a total of 114 SNPs, or 6,441 interactions. We also added an additional 1,000 null SNPs leading to a total of 1,014 SNPs or 513,591 interactions. For each of the 100 simulated replicate data sets, we compared the rank of the causal interactions among all pairwise interactions between our method (based on the posterior probability of inclusion in the model) to those from an exhaustive search based on the P-value of the interaction coefficient β_3 in the model:

$$\text{logit}(Y_i = 1) = \beta_0 + \beta_1 s_{ia} + \beta_2 s_{ib} + \beta_3 s_{ia} s_{ib}. \quad (16)$$

Disease status was simulated under a penetrance of RR = 1.3 for each of the 100 simulated replicates. Figures 9 and 10 compare the rankings between our method and the

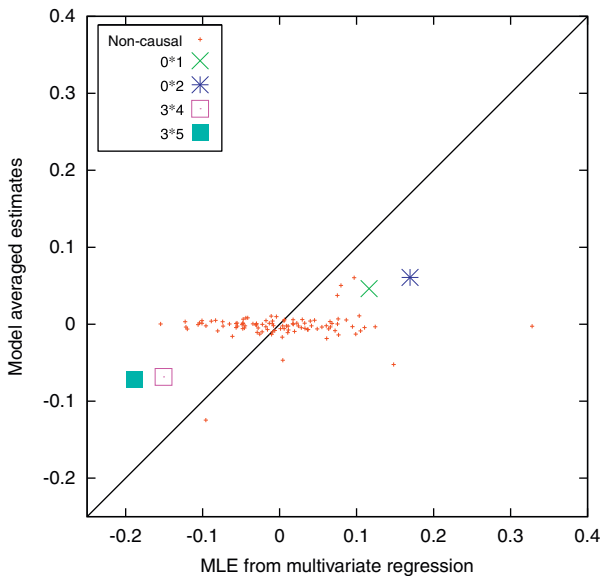


Fig. 7. Shrinkage of β relative to maximum likelihood for the four causal and 87 null interactions via Bayes model averaging.

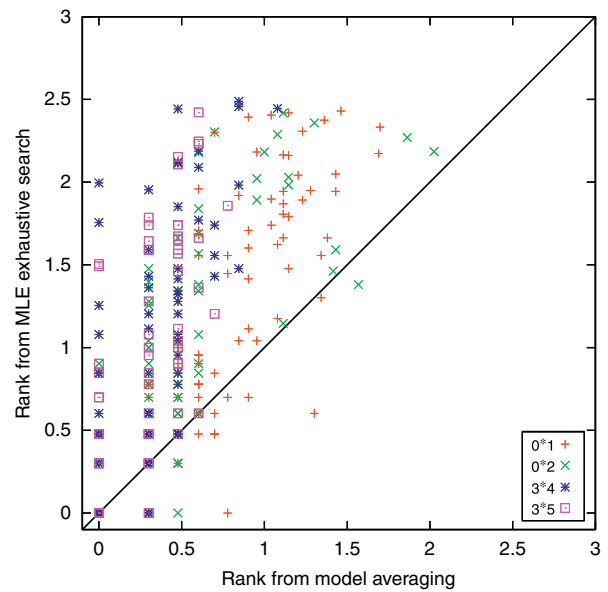


Fig. 9. \log_{10} Rankings based on four causal interaction among 6,441 total interactions.

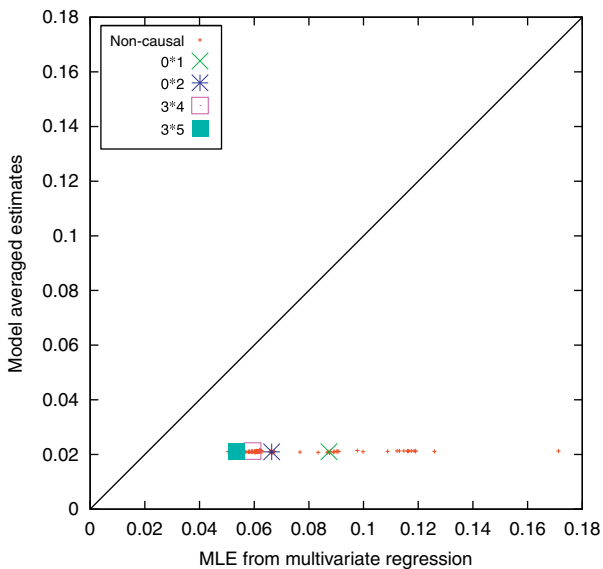


Fig. 8. Shrinkage for $SE(\beta)$ relative to maximum likelihood for the four causal and 87 null interactions via Bayes model averaging.

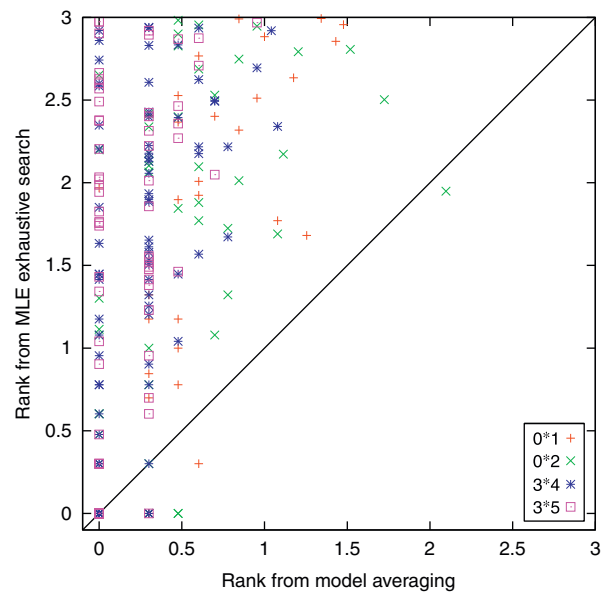


Fig. 10. \log_{10} Rankings based on four causal interactions among 513,591 total interactions.

exhaustive search where each data point represents the rank determined by both methods based on one simulation replicate. The improvement in rankings by model averaging over use of maximum likelihood based P -values is more pronounced as the number of additional null interactions is included in the analysis, as seen in Table IV.

SIMULATION 2: DISCOVERING INTERACTIONS IN A KNOWN PATHWAY

To demonstrate how our approach might be applied to a more realistic but complex setting, we applied this

approach to data simulated using a mechanistic model of the folate metabolism pathway described previously [Thomas et al., 2009; Reed et al., 2006]. Folate metabolism is a complex pathway, with two main loops: the folate cycle and the methionine cycle. Figure 11 illustrates the enzymes and metabolites involved in the pathway [Reed et al., 2006]. The folate cycle influences pyrimidine synthesis, while the methionine cycle influences DNA methylation. These two mechanisms, along with purine synthesis and homocysteine levels, have been suggested to have connections to carcinogenesis. The goal behind this analysis was to determine whether use of prior information could help the method to discover $G \times G$ or $G \times E$

interactions that are biologically plausible under these four alternative models of etiology.

We simulated genotypes and biomarker levels for 10,000 individuals at 14 genes and two environmental exposures (intracellular folate and methionine intake) in the same manner as described earlier [Thomas et al., 2009] through the use of a physiologically based pharmacokinetic (PBPK) model [Reed et al., 2006]. This model predicts steady-state concentrations of 10 intermediate metabolites and 14 reaction rates through differential equations that take into account genotype-specific K_m and V_{max} values at each of the 14 genes. Predictions from the PBPK model were previously confirmed through biological experiments [Reed et al., 2006]. From among the 24 continuous biomarker variables (i.e. intermediate metabolite and reaction rates), we considered homocysteine concentration, pyrimidine synthesis, purine synthesis, and DNA methylation separately as plausible disease mechanisms. Disease status for the population of 10,000 individuals was assigned using a logistic model with a RR of 1.2 per unit increase of the standardized causal variable. We assumed a modest RR of 1.2 for each of the mechanisms studied.

TABLE IV. Comparison of rankings of pairwise interactions only between Bayes model averaging and an exhaustive search after adding additional null SNPs into the data set

| Total tested interactions | Causal interaction | Proportion of simulation replicates where $BMA_{rank} < MLE_{rank}$ (%) |
|---------------------------|--------------------|---|
| 6,441 | 0*1 | 82 |
| | 0*2 | 84 |
| | 3*4 | 95 |
| | 3*5 | 97 |
| | 0*1 | 97 |
| 513,591 | 0*2 | 92 |
| | 3*4 | 97 |
| | 3*5 | 99 |

Case-control status, genotypes, and environmental exposures for the first 4,000 cases and 4,000 controls were considered as observed data, while genotypes, environmental exposures, and data across the 24 intermediate metabolites for the remaining 2,000 individuals were reserved as prior data. The A and the Z matrices were constructed from the prior data in the same manner as described previously in Simulation 1. Although we simulated affection status from four mutually exclusive disease mechanisms (simulation scenarios), for each scenario we included the same four columns in Z in the model, each column corresponding to the metabolite that was causal in one of the simulated disease mechanisms, with the expectation that only the metabolites relevant to the simulated mechanism would be correlated to disease risk as quantified in the corresponding element of the vector π . Genotype-metabolite correlations across all 24 metabolites were used in defining A .

We allowed the MCMC algorithm to run for 1 million iterations, discarding the first 50,000 iterations as burn-in. We tabulated statistics based on the remaining 950,000 realizations from the posterior distribution. The program completed within 6.2hr. To assess whether our model converged, we compared the distribution of variables drawn from two chains that began with different initial values. The presence of a variable in the current model was tabulated at (thinning) intervals of 100. A Kolmogorov-Smirnov test confirmed that draws from each of the two chains were from the same distribution ($P = 0.9509$). For each of the four causal mechanisms simulated, we summarized the median of π and their standard errors. As shown in Table V, posterior estimates for π approximated the log RRs simulated under each of the four scenarios. Estimates of σ^2 was slightly larger than for τ^2 , indicating increased shrinkage toward the component of the prior mean of β encoded by the A matrix.

Table VI presents the main effects with the highest BF from the analysis for each of the four causal mechanisms studied when including both A and Z in the model for the

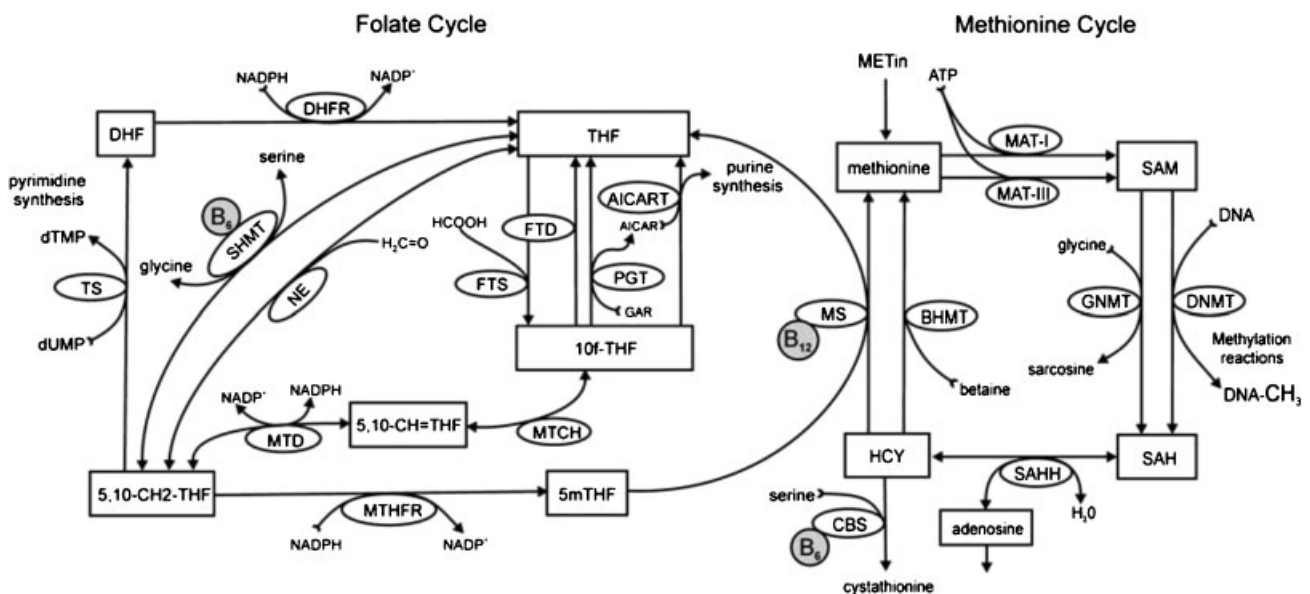


Fig. 11. Topology of the folate pathway as presented originally in [Nicolae et al., 2010].

TABLE V. Bayes model averaged estimates of second-level means $\tilde{\pi}$ and variances τ^2, σ^2 across different causal mechanisms in folate pathway

| Simulated mechanism | Prior variance | | Second-level coefficients π | | | |
|---------------------|----------------|------------|---------------------------------|--------------|-------------|-------------|
| | τ^2 | σ^2 | Homocysteine | Pyrimidine | Purine | Methylation |
| Homocysteine | 0.03 | 0.06 | 0.18(0.13) | -0.09(0.536) | 0.002(0.38) | 0.04(0.58) |
| Pyrimidine | 0.01 | 0.02 | -0.04(0.22) | 0.22(0.066) | -0.01(0.06) | -0.02(0.21) |
| Purine | 0.01 | 0.03 | -0.01(0.36) | 0.16(0.327) | 0.19(0.07) | -0.22(0.45) |
| Methylation | 0.01 | 0.03 | -0.09(0.22) | -0.02(0.16) | 0.025(0.16) | 0.23(0.12) |

Standard errors are enclosed in parentheses.

TABLE VI. Genes and interactions detected in the folate pathway either through Bayes model averaging (BF > 1,000) or stepwise regression ($P < 0.05$) across four alternative disease mechanisms with RR = 1.2

| Main effect/interaction | Homocysteine | | Pyrimidine | | Purine | | Methylation | |
|-------------------------|--------------|---------|------------|---------|--------|---------|-------------|---------|
| | BF | P-value | BF | P-value | BF | P-value | BF | P-value |
| DHFR | 36 | 0.015 | 35 | 0.251 | 49 | 0.397 | 20 | 0.660 |
| Fol | 17 | 0.631 | 1,737 | 3e-4 | 3,017 | 5e-5 | 173 | 0.263 |
| FTD | 176 | 0.075 | 47 | 0.104 | 1,616 | 0.071 | 140 | 0.143 |
| Met | ∞ | 3e-4 | 58 | 0.932 | 105 | 0.628 | 166 | 0.578 |
| MS | 263 | 0.048 | 35 | 0.952 | 130 | 0.034 | 46 | 0.101 |
| MTD | 24 | N/S | 442,799 | 0.006 | 188 | 0.015 | 1,243 | 0.023 |
| MTHFR | 18 | 0.292 | 22 | 0.377 | 75 | 0.211 | 4,139 | 0.777 |
| PGT | 17 | 0.613 | 10 | 0.618 | 40,768 | 0.002 | 17 | 0.310 |
| SHMT | 42 | 0.267 | 2,422 | 0.878 | 452 | 0.425 | 90 | 0.859 |
| TS | 23 | 0.183 | 344,893 | 0.148 | 76 | 0.050 | 22 | 0.131 |
| CBS*MAT-II | 77 | 0.045 | 45 | 0.062 | 28 | N/S | 19 | 0.151 |
| CBS*Met | 1,072 | N/S | 51 | N/S | 33 | N/S | 106 | N/S |
| DHFR*AICART | 1 | N/S | 6 | 0.108 | 5 | 0.116 | 16 | 0.027 |
| FTD*MAT-II | 38 | 0.045 | 15 | 0.038 | 334 | 0.031 | 36 | 0.057 |
| FTD*MTHFR | 213 | 0.015 | 20 | 0.046 | 301 | 0.086 | 1,814 | 0.035 |
| MS*Met | 1,129 | N/S | 16 | N/S | 67 | N/S | 63 | N/S |
| MTCH*CBS | 112 | 0.139 | 281 | 0.092 | 551 | 0.050 | 16 | N/S |
| MTCH*FTS | 11 | 0.112 | 19 | 0.092 | 412 | 0.079 | 16 | 0.040 |
| MTCH*MS | 978 | 0.006 | 534 | 0.006 | 1,130 | 0.008 | 130 | 0.016 |
| MTCH*PGT | 10 | 0.145 | 5 | N/S | 1,416 | 0.026 | 9 | 0.125 |
| MTD*MTHFR | 7 | N/S | 118 | N/S | 63 | N/S | 2,203 | N/S |
| MTHFR*Met | 143 | N/S | 4 | N/S | 54 | N/S | 1,038 | 0.101 |
| PGT*CBS | 34 | N/S | 32 | 0.056 | 1,022 | 0.069 | 56 | 0.097 |
| PGT*MS | 75 | 0.044 | 14 | 0.018 | 2,851 | 0.007 | 36 | 0.025 |
| SHMT*CBS | 90 | N/S | 1,254 | 0.133 | 319 | 0.134 | 137 | N/S |
| SHMT*Fol | 11 | 0.108 | 2,324 | 0.036 | 1,398 | 0.022 | 121 | N/S |
| SHMT*MAT-II | 8 | 0.139 | 160 | 0.062 | 646 | 0.012 | 39 | 0.083 |
| TS*MTHFR | 41 | 0.022 | 227 | 0.022 | 57 | 0.024 | 349 | 0.019 |
| TS*SHMT | 17 | N/S | 1,091 | N/S | 147 | N/S | 23 | N/S |

Variables that were not selected via stepwise regression are shown as N/S. RR, relative risk; BF, Bayes Factors.

prior. There is some overlapping evidence shared across each of the causal mechanisms for some variables, but the strongest evidence for each of the causal mechanisms show consistency with the pathway topology (Fig. 11). In particular, primary drivers included methionine intake (Met) when homocysteine levels were the simulated causal factor (BF = ∞), thymidylate synthase and methylenetetrahydromethanopterin dehydrogenase (MTD) in the pyrimidine synthesis simulation (BF = 344,893 and 442,799 respectively), and phosphoribosyl glycinamide-transformylase (PGT) in the purine synthesis simulation (BF = 40,768). Evidence for a genetic or environmental

effect was substantially weaker when DNA methylation was considered causal, although 5,10-methylenetetrahydrofolate reductase (MTHFR) attained a BF of 4,140.

We also observed several plausible second-order interactions with strong evidence (BF > 1,000) based on our method that had mild to no evidence based on an analysis using AIC penalized stepwise regression (Table VI). Under the model with homocysteine levels as causal, we detected the strongest evidence between MS and Met (BF = 1,129), adjacent neighbors immediately upstream of homocysteine, respectively. The strongest evidence under the pyrimidine synthesis model was between SHMT and

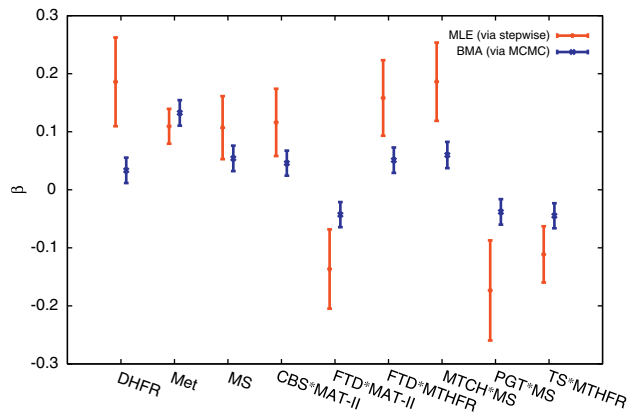


Fig. 12. MLE and model-averaged estimates of β for homocysteine disease model. MLE, maximum likelihood estimates.

folate with a BF of 2,324. Tetrahydrofolate is a substrate for folate and the immediate downstream neighbor of SHMT. The purine synthesis model revealed the strongest evidence as an interaction between MS and PGT (BF = 2,851). The link between these two genes is less clear based on the pathway topology in Figure 11. Under the DNA methylation model, the strongest evidence (BF = 2,203) for an interaction was between two neighboring genes: MTD and MTHFR.

To assess the effects of shrinkage on estimates of β and its standard errors, we compared the MLE of β from stepwise regression to the posterior estimates of β from our method. Figure 12 presents β and its standard errors for both methods. The MLE of β were generally more dispersed compared to the posterior estimates. Evidence for methionine levels was the highest among all variables based on results from our method (BF = ∞) and the stepwise regression ($P < 3^{-4}$). The MLE for methionine levels had a modest, albeit more stable, estimated effect size compared to the other variables. The BMA estimates of β were shrunk closer to zero than the MLE for all variables except methionine levels.

APPLICATION TO REAL DATA: USING GENE ANNOTATIONS TO INFORM A SEARCH FOR INTERACTIONS

We applied our approach to real data from a genome-wide association scan for breast cancer. The data set, publicly available from NCI (<https://caintegrator.nci.nih.gov/cgems/>), includes 1,145 cases and 1,142 controls of European ancestry. Rather than consider all 546,631 SNPs that were genotyped in the original study, we focused on the subset of SNPs that contained annotations in Gene Ontology (GO) under the “Biological Process” schema, reducing the set of SNPs to 6,078, each SNP chosen based on having the lowest marginal P -value within the gene it is contained in.

We constructed our priors from the 22 GO terms (Level 3) immediately under the “Biological Process” category. Genes with more granular annotations were assigned to their ancestral annotation on Level 3 so that genes were more readily comparable. In coding the 23 column Z matrix, we coded presence of an annotation as 1 and absence as 0. When coding the A matrix, we defined pairwise similarity scores as a weighted average by taking into account the

TABLE VII. Estimates for the second-level regression coefficient π measuring correlation of β to Z in the breast cancer data set

| GO term | π | SE(π) |
|---|--------|-------------|
| Intercept | -0.035 | 0.086 |
| Metabolic process | 0.018 | 0.059 |
| Developmental process | 0.026 | 0.086 |
| Cellular process | 0.067 | 0.094 |
| Localization | -0.019 | 0.127 |
| Positive regulation of biological process | 0.005 | 0.082 |
| Growth | -0.141 | 0.120 |
| Response to stimulus | 0.050 | 0.088 |
| Reproductive process | 0.048 | 0.079 |
| Multicellular organismal process | 0.005 | 0.078 |
| Pigmentation | -0.080 | 0.138 |
| Locomotion | -0.028 | 0.175 |
| Cell killing | -0.085 | 0.132 |
| Viral reproduction | -0.091 | 0.124 |
| Rhythmic process | 0.187 | 0.198 |
| Biological regulation | -0.001 | 0.121 |
| Regulation of biological process | -0.103 | 0.123 |
| Negative regulation of biological process | 0.075 | 0.090 |
| Immune system process | -0.001 | 0.134 |
| Multi-organism process | -0.053 | 0.135 |
| Reproduction | 0.041 | 0.075 |
| Biological adhesion | -0.006 | 0.131 |
| Establishment of localization | 0.017 | 0.138 |

frequency of annotations across the set of genes, so that the score between gene j and k is

$$A_{j,k} = \frac{\sum_{t=1}^g \frac{m - c_t + 1}{m} I(Z_{jt} = 1 \wedge Z_{kt} = 1)}{\sum_{t=1}^g \frac{m - c_t + 1}{m}}, \quad (17)$$

where g is the total number of GO terms (22 GO annotations in this example) and c_t is the number of genes in the model with annotation term t .

Two independent chains were run for assessing convergence, where each chain was randomly initialized with model variables most likely to be distant from the posterior mode (i.e. variables with a marginal P -values ≥ 0.5). The program was run for 1 million iterations with the first 50,000 discarded as burn-in. Computation time was approximately 23 hr. The Kolmogorov-Smirnov test for assessing convergence, as described in the folate simulation example, implied that the two chains had converged to the same distribution ($P = 0.25$). Estimates for σ^2 and τ^2 based on the posterior median were 0.07 and 0.02, respectively. For each model sampled, we fitted only a subset of the Z matrix through multiple regression, using as many linearly independent columns as available, since for certain sampled models none (or all) of the model variables associated to a given annotation or GO annotations were concordant between two or more GO terms across all the variables (e.g. 4 of the 22 GO terms were related to the concept of biological regulation). Table VII lists the posterior median estimates for each element in π corresponding to the 22 GO terms, calculated across only the MCMC samples where a particular term was included in Z . Based on these standard errors, there was slight evidence for a negative correlation between β and the GO term “growth”. We noted that

TABLE VIII. Main effects in the breast cancer data set with at least prior specification having BF > 100

| Main effect | Single SNP analysis | | | Non-informative prior | | | Informative Z only | | | Informative A only | | | Full-informative prior | | |
|-------------|---------------------|---------------------|------|-----------------------|---------------|-------|--------------------|---------------|----------|--------------------|---------------|-------|------------------------|---------------|----------|
| | $\hat{\beta}$ | SE($\hat{\beta}$) | Rank | β | SE(β) | BF | β | SE(β) | BF | β | SE(β) | BF | β | SE(β) | BF |
| FGFR2 | -0.284 | 0.060 | 1 | -0.208 | 0.044 | 1,145 | -0.238 | 0.046 | ∞ | -0.210 | 0.044 | 5,326 | -0.239 | 0.046 | ∞ |
| HCN1 | -0.256 | 0.059 | 2 | -0.187 | 0.044 | 98 | -0.198 | 0.046 | 116 | -0.182 | 0.044 | 82 | -0.197 | 0.046 | 114 |
| AK5 | 0.284 | 0.069 | 3 | 0.192 | 0.044 | 266 | 0.212 | 0.046 | 507 | 0.196 | 0.044 | 220 | 0.215 | 0.047 | 580 |
| BMPR1B | 0.248 | 0.061 | 4 | 0.183 | 0.044 | 84 | 0.206 | 0.046 | 512 | 0.175 | 0.043 | 112 | 0.210 | 0.046 | 380 |
| ABO | -0.249 | 0.063 | 5 | -0.173 | 0.044 | 29 | -0.192 | 0.046 | 113 | -0.174 | 0.044 | 75 | -0.184 | 0.047 | 70 |
| PARK2 | 0.321 | 0.082 | 7 | 0.190 | 0.045 | 68 | 0.213 | 0.048 | 326 | 0.188 | 0.045 | 109 | 0.213 | 0.048 | 169 |
| CNGA3 | -0.260 | 0.067 | 8 | -0.177 | 0.044 | 56 | -0.196 | 0.046 | 127 | -0.170 | 0.044 | 98 | -0.199 | 0.046 | 134 |
| FAM129A | 0.225 | 0.059 | 11 | 0.176 | 0.044 | 48 | 0.191 | 0.046 | 104 | 0.168 | 0.043 | 45 | 0.196 | 0.046 | 103 |
| ABCA1 | -0.302 | 0.080 | 12 | -0.169 | 0.044 | 30 | -0.194 | 0.046 | 173 | -0.167 | 0.043 | 33 | -0.188 | 0.047 | 129 |
| PTPRD | 0.320 | 0.090 | 20 | 0.171 | 0.044 | 23 | 0.192 | 0.047 | 105 | 0.176 | 0.045 | 30 | 0.196 | 0.047 | 96 |
| SORCS1 | -0.391 | 0.110 | 23 | -0.196 | 0.046 | 81 | -0.224 | 0.050 | 253 | -0.205 | 0.047 | 74 | -0.223 | 0.050 | 135 |
| PRKCQ | -0.214 | 0.060 | 25 | -0.170 | 0.044 | 34 | -0.194 | 0.046 | 151 | -0.168 | 0.044 | 61 | -0.194 | 0.046 | 124 |
| BARD1 | 0.200 | 0.061 | 64 | 0.160 | 0.044 | 12 | 0.177 | 0.046 | 27 | 0.154 | 0.043 | 16 | 0.185 | 0.046 | 114 |

BF, Bayes factors.

overall, standard errors for π were relatively large compared to the median estimates of π . To investigate whether these large values may have been due to possible collinearity among some of the GO term annotations, we re-ran the analyses, including only the intercept and one GO term for each analysis. The standard errors and median values of π under the univariate analyses were of similar magnitude to those estimated under the multivariate models, with none of the GO terms significantly correlated to β .

We revisited the first four prior models described in our first simulation to assess how information from GO influenced posterior rankings and evidence for association. Table VIII lists posterior estimates of β and the BF for each gene across the four models where at least one model had a minimum BF of 100. For comparison, MLE and rankings from a single SNP-by-SNP analysis for marginal effects are also shown. Evidence for FGFR2 is compelling regardless of the analysis used. Rankings differed slightly from the MLE analysis. At the BF > 100 threshold, the top 12 genes (except for those ranked 6, 9, and 10) from the MLE analysis are recovered. However, genes ranking lower based on the MLE analysis showed increased evidence when information from Z was included (Models 2 and/or 4). For example, PTPRD (rank 20) was found exclusively under Model 2, and BARD1 (rank 64) was found exclusively under Model 4.

Table IX, which is ordered in the descending order of BFs under Model 4, reveals a larger number of G x G interactions found at a BF threshold of 100 compared to main effects. The interaction between PARK2 and SORCS1 had the strongest evidence across the four models, with BF of 70,000 under Model 2, 50,000 under Model 4, 20,000 under Model 3, and 10,000 under Model 1. As with other interactions listed in Table IX, we also observed a contrast in the degree of shrinkage across the four models. For example, under Model 1, where no prior information was incorporated, shrinkage of this first interaction toward zero was most pronounced ($\beta = 0.22$). The other three models demonstrated shrinkage of posterior estimates of β toward non-zero prior means informed by the GO: Model 3 with $\beta = 0.23$, Model 2 with $\beta = 0.26$, and Model 4 with $\beta = 0.27$. The model-averaged estimates for the non-zero valued components of the prior mean of this interaction

were $\pi^T Z_k = 0.06$ and $\bar{\phi}_{-k} = 0.06$, which appears to explain the reduced shrinkage toward zero under Models 2–4 when compared to Model 1.

We annotated these interactions to assess the distribution of their associated GO terms. Table X lists the annotations common to both genes in the interaction for the interactions of interest listed in Table IX. Although at first glance, the list appears to suggest that certain terms such as biological regulation or cellular process are enriched in this list of interactions, the probability a gene is associated with a term varies greatly across terms. Furthermore, the list includes numerous interactions that are not truly independent, since in some cases one of the two genes in an interaction is shared across several iterations (e.g. 11 interactions contain FGFR2). We calculated an empirical P-value to estimate the probability that a term could be associated with a random set of interactions at a proportion that was equal to or greater than the observed proportion. We took into account the non-independence of the interaction by defining bins representative of the distribution of observed interactions, where each bin contained interactions with a shared first gene, but distinct second genes. The bins were then pooled into one data set, and the number of times a particular term's frequency exceeded the observed frequency in the pool was tallied and divided by 1 million (the total number of replicates). The first row of Table X lists these empirical P-values, revealing a significant enrichment for biological regulation ($P = 0.008$), growth ($P = 1e^{-6}$, metabolic process ($P = 0.008$), and regulation of biological process ($P = 0.003$).

DISCUSSION

We have presented several examples in which our proposed variable selection algorithm can simultaneously use information from observed data as well as prior biological knowledge toward the goal of discovering genes or higher level interactions that might escape a one-variable-at-a-time maximum-likelihood-based scan for marginal effects. Rather than fixing the influence of our priors a priori, our Bayesian method allows the data to inform the influence of the priors through the second-level

TABLE IX. Interactions in the breast cancer data set with at least prior specification having BF > 100

| Interaction | Non-informative Prior | | | Informative Z only | | | Informative A only | | | Full-informative Prior | | |
|-----------------|-----------------------|---------------|-----|--------------------|---------------|-------|--------------------|---------------|-----|------------------------|---------------|-----|
| | β | SE(β) | BF | β | SE(β) | BF | β | SE(β) | BF | β | SE(β) | BF |
| PARK2*SORCS1 | 0.222 | 0.060 | 1e4 | 0.262 | 0.064 | 7e4 | 0.226 | 0.060 | 2e4 | 0.267 | 0.064 | 5e4 |
| AK5*ARHGAP26 | 0.159 | 0.045 | 427 | 0.165 | 0.047 | 197 | 0.161 | 0.045 | 294 | 0.165 | 0.047 | 903 |
| FGFR2*MAML2 | -0.113 | 0.044 | 1 | -0.157 | 0.046 | 91 | -0.122 | 0.044 | 81 | -0.155 | 0.046 | 686 |
| SHC3*KIF13B | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.171 | 0.048 | 621 |
| PCLO*ME3 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.182 | 0.051 | 528 |
| CNGA3*CNN1 | -0.158 | 0.052 | 41 | -0.185 | 0.054 | 608 | N/A | N/A | N/A | -0.170 | 0.054 | 462 |
| FGFR2*CDT1 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | -0.157 | 0.045 | 445 |
| SHC3*CXCL16 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | -0.177 | 0.048 | 403 |
| FGFR2*ABCA1 | -0.096 | 0.046 | 158 | -0.104 | 0.048 | 380 | -0.090 | 0.046 | 133 | -0.105 | 0.048 | 268 |
| CYP2J2*SORCS1 | -0.113 | 0.050 | 74 | -0.136 | 0.054 | 187 | -0.119 | 0.051 | 27 | -0.140 | 0.054 | 266 |
| FGFR2*SCG5 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.208 | 0.049 | 235 |
| PARK2*COL13A1 | 0.097 | 0.044 | 57 | 0.074 | 0.047 | 0 | 0.105 | 0.045 | 81 | 0.124 | 0.046 | 231 |
| CYP2J2*KLK14 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | -0.164 | 0.049 | 212 |
| TCF7L2*VAV3 | N/A | N/A | N/A | -0.104 | 0.047 | 0 | N/A | N/A | N/A | -0.130 | 0.048 | 170 |
| FGFR2*BMPR1B | -0.035 | 0.044 | 63 | -0.046 | 0.046 | 62 | -0.047 | 0.045 | 42 | -0.051 | 0.046 | 146 |
| PARK2*BARD1 | N/A | N/A | N/A | -0.042 | 0.046 | 4 | N/A | N/A | N/A | -0.044 | 0.048 | 146 |
| NLGN1*HEYL | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | -0.145 | 0.047 | 144 |
| FGFR2*NTSR1 | 0.084 | 0.044 | 115 | 0.088 | 0.046 | 43 | 0.087 | 0.044 | 55 | 0.092 | 0.046 | 141 |
| CNGA3*BARD1 | N/A | N/A | N/A | 0.057 | 0.046 | 16 | 0.060 | 0.043 | 42 | 0.050 | 0.046 | 138 |
| TCF7L2*DNAJC8 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | -0.147 | 0.047 | 133 |
| DFFA*MAN2B1 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.175 | 0.049 | 122 |
| PARK2*TCF7L2 | N/A | N/A | N/A | -0.052 | 0.048 | 45 | -0.044 | 0.044 | 5 | -0.063 | 0.048 | 117 |
| SORCS1*RPL6 | N/A | N/A | N/A | -0.155 | 0.056 | 57 | N/A | N/A | N/A | -0.138 | 0.054 | 114 |
| FGFR2*BARD1 | 0.031 | 0.045 | 11 | 0.033 | 0.046 | 11 | N/A | N/A | N/A | 0.033 | 0.046 | 104 |
| FGFR2*PRKCQ | N/A | N/A | N/A | -0.021 | 0.046 | 59 | -0.010 | 0.044 | 43 | -0.013 | 0.046 | 103 |
| CACNG3*BARD1 | N/A | N/A | N/A | N/A | N/A | N/A | 0.085 | 0.045 | 54 | 0.090 | 0.047 | 102 |
| ABO*TCF7L2 | N/A | N/A | N/A | 0.143 | 0.048 | 420 | 0.120 | 0.045 | 66 | 0.135 | 0.048 | 67 |
| SORCS1*GLIS3 | 0.141 | 0.048 | 33 | 0.169 | 0.052 | 359 | 0.169 | 0.050 | 487 | 0.149 | 0.052 | 65 |
| ABO*SORCS1 | 0.100 | 0.045 | 99 | 0.089 | 0.048 | 139 | 0.097 | 0.046 | 24 | 0.092 | 0.047 | 53 |
| SORCS1*GLTP | -0.105 | 0.046 | 43 | -0.127 | 0.049 | 121 | -0.105 | 0.046 | 30 | -0.111 | 0.050 | 42 |
| ABCA1*HHATL | N/A | N/A | N/A | -0.152 | 0.046 | 133 | N/A | N/A | N/A | -0.145 | 0.046 | 41 |
| FGFR2*HCN1 | 0.054 | 0.045 | 86 | 0.078 | 0.047 | 71 | 0.064 | 0.044 | 120 | 0.072 | 0.046 | 24 |
| FAM129A*ABCA1 | N/A | N/A | N/A | 0.068 | 0.047 | 149 | 0.068 | 0.045 | 18 | 0.079 | 0.047 | 22 |
| SORCS1*TRIO | 0.125 | 0.048 | 24 | 0.137 | 0.051 | 115 | 0.140 | 0.048 | 39 | 0.132 | 0.050 | 17 |
| PRKCQ*IRF2 | N/A | N/A | N/A | 0.158 | 0.047 | 122 | N/A | N/A | N/A | 0.133 | 0.048 | 16 |
| PARK2*KIR3DL2 | N/A | N/A | N/A | 0.154 | 0.047 | 6 | 0.153 | 0.045 | 185 | 0.134 | 0.047 | 11 |
| FGFR2*ASL | 0.138 | 0.045 | 112 | 0.139 | 0.047 | 13 | 0.125 | 0.045 | 39 | 0.107 | 0.046 | 7 |
| PARK2*CAP2 | 0.090 | 0.044 | 9 | 0.110 | 0.047 | 119 | 0.093 | 0.044 | 20 | 0.101 | 0.047 | 5 |
| BMPR1B*ZMPSTE24 | N/A | N/A | N/A | N/A | N/A | N/A | -0.174 | 0.045 | 257 | N/A | N/A | N/A |
| AK5*KREMEN1 | N/A | N/A | N/A | 0.196 | 0.046 | 2,012 | N/A | N/A | N/A | N/A | N/A | N/A |
| PIGN*HNRPDL | -0.176 | 0.049 | 139 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| PTPRD*CLCN6 | -0.147 | 0.045 | 42 | -0.179 | 0.048 | 798 | N/A | N/A | N/A | N/A | N/A | N/A |
| SLC5A7*ANLN | N/A | N/A | N/A | 0.153 | 0.050 | 123 | N/A | N/A | N/A | N/A | N/A | N/A |
| FGFR2*OR2F1 | -0.137 | 0.046 | 157 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| PARK2*DSCAM | 0.153 | 0.047 | 160 | 0.166 | 0.050 | 87 | 0.143 | 0.046 | 18 | N/A | N/A | N/A |
| ABO*TGIF1 | N/A | N/A | N/A | N/A | N/A | N/A | 0.147 | 0.045 | 125 | N/A | N/A | N/A |
| AK5*SYCP1 | N/A | N/A | N/A | N/A | N/A | N/A | -0.163 | 0.045 | 288 | N/A | N/A | N/A |
| PMM2*ATP6V0A4 | 0.153 | 0.044 | 44 | 0.178 | 0.046 | 184 | N/A | N/A | N/A | N/A | N/A | N/A |
| NRG1*BCKDHB | -0.138 | 0.049 | 102 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| PARK2*CNTF | -0.125 | 0.045 | 9 | -0.130 | 0.047 | 106 | N/A | N/A | N/A | N/A | N/A | N/A |

Interactions that were not sampled are shown as N/A.

variance parameters σ^2 , τ^2 , and regression coefficients π . This behavior was demonstrated in the sensitivity analysis across the six different prior specifications, where the informative priors improved power over the less informative priors under moderate-to-weak disease penetrances, but showed little difference in power once evidence in the

observed data was very weak (RR = 1.1). In simulations, we showed that the Z-only model that contains only the covariates causal for disease risk performs better than the A-only model. Although similar information can be captured through either component of the prior (i.e. information from the same covariates enter both matrices),

TABLE X. Annotations and enrichment P-values for the same G × G interactions shown in Table IX

| Enrichment P-value | Variable | | | | | | | | | |
|--------------------|-----------------------|------------------|-----------------------|-------------------------------|--------|--------------|-------------------|----------------------------------|---|----------------------------------|
| | Biological regulation | Cellular process | Developmental process | Establishment of localization | Growth | Localization | Metabolic process | Multicellular organismal process | Positive regulation of biological process | Regulation of biological process |
| | 0.008 | 0.05 | 0.09 | 0.36 | 6e-5 | 0.53 | 0.008 | 0.06 | 0.19 | 0.003 |
| PARK2*SORCS1 | X | | | | | | | | | |
| AK5*ARHGAP26 | X | | | | | | | | | |
| FGFR2*MAML2 | X | X | | | | | X | | | X |
| SHC3*KIF13B | X | X | | | | | | | | X |
| PCLO*ME3 | X | X | | | | | | X | | X |
| CNGA3*CNNA1 | X | X | | | | | X | | | X |
| FGFR2*CDT1 | X | X | | | | | X | | | X |
| SHC3*CXCL16 | X | X | | | | | X | | | X |
| FGFR2*ABCA1 | X | X | | | | | X | | | X |
| CYP2J2*SORCS1 | X | X | | | | | X | | | X |
| FGFR2*SCG5 | X | X | | | | | X | | | X |
| PARK2*COL13A1 | X | X | X | | | | X | | | X |
| CYP2J2*KLK14 | X | X | | | | | X | | | X |
| TCF7L2*VAV3 | X | X | | | | | X | | | X |
| FGFR2*BMPRI1B | X | X | | | X | | X | | | X |
| PARK2*BARD1 | X | X | X | | | | X | X | | X |
| NLGN1*HEYL | X | X | X | | | | X | X | | X |
| FGFR2*NTSR1 | X | X | | | | | X | | | X |
| CNGA3*BARD1 | X | X | | | | | X | | | X |
| TCF7L2*DNAJC8 | X | X | | | | | X | | | X |
| DDFA*MAN2B1 | X | X | | | | | X | | | X |
| PARK2*TCF7L2 | X | X | X | | | | X | | | X |
| SORCS1*RPL6 | X | X | | | | | X | | | X |
| FGFR2*BARD1 | X | X | | | | | X | | | X |
| FGFR2*PRKCQ | X | X | | | X | | X | | | X |
| CACNG3*BARD1 | X | X | | X | | | X | | | X |
| ABO*TCF7L2 | X | X | | | | | X | | | X |
| SORCS1*GLIS3 | X | X | | | | | X | | | X |
| ABO*SORCS1 | X | X | | | | | X | | | X |
| SORCS1*GLTP | X | X | | | | | X | | | X |
| ABCA1*HHATL | X | X | | | | | X | | | X |
| FGFR2*HCN1 | X | X | | | | | X | | | X |
| FAM129A*ABCA1 | X | X | | | | | X | | X | X |
| SORCS1*TRIO | X | X | | | | | X | | | X |
| PRKCQ*IRF2 | X | X | | | | | X | | | X |
| PARK2*KIR3DL2 | X | X | | | | | X | | | X |
| FGFR2*ASL | X | X | | | | | X | | | X |
| PARK2*CAP2 | X | X | X | | | | X | | | X |
| PARK2*DSCAM | X | X | X | | | | X | | | X |
| FGFR2*OR2F1 | X | X | | | | | X | | | X |

TABLE X. Continued

| Variable | Variable | | | | | | | | | | |
|----------------|--------------------|-----------------------|------------------|-----------------------|-------------------------------|--------|--------------|-------------------|----------------------------------|---|----------------------------------|
| | Enrichment P-value | Biological regulation | Cellular process | Developmental process | Establishment of localization | Growth | Localization | Metabolic process | Multicellular organismal process | Positive regulation of biological process | Regulation of biological process |
| PIGN*HNRPDL | | | X | | | | | | | | |
| NRG1*BCKDHB | | | X | | | | | X | | | |
| AK5*KREMEN1 | | | X | | | | | X | | | |
| PTPRD*CLCN6 | | | X | | | | | | | | X |
| PMM2*ATP6V0A4 | | X | | | | | | | | | |
| SLC5A7*ANLN | | X | | | | | | | | | |
| PARK2*CN1F | | | X | | | | | | | | |
| AK5*SYCP1 | | | X | | | | | | | | |
| BMP1B*ZMPSTE24 | | | X | | | | | | | | |
| ABO*TGIF1 | | | X | | | | | | | | |

an A-only model may sometimes be the most sensible prior for two practical reasons. First, a large number of covariates in Z places a lower bound constraint on the size of models whose means, variances, and probabilities will be informed through the hyperparameters, since π is only estimable when m is larger than the number of covariates in Z. Thus, it makes sense to keep Z as parsimonious as possible if we wish to properly sample the space of small models. The four-fold increase in computation time in the breast cancer example clearly demonstrates how model size negatively impacts run-time performance. Second, given that a small set of covariates are chosen in Z, when these covariates are not correctly specified, the A only model can be more powerful as we demonstrated in simulations. Third, whereas LD can create a large number of spurious signals in single SNP analyses, incorporating correlation measures into the A matrix of the hierarchical model allows these spurious signals to be dampened by neighboring null SNPs (provided that these null SNPs are not in perfect LD). Conversely, at disease loci, where one might expect multiple independent mutation events at a gene to predispose one to disease risk, incorporation of correlation information can be advantageous. It seems plausible that this source of prior information can also aid in convergence by defining a smoother likelihood surface. We plan to investigate this question further.

Several classes of Bayesian methods applicable in genetic association studies have been reviewed in the literature, particularly BMA, SSVS, and RJMCMC [Fridley, 2009]. The “spike and slab” approach used in the standard SSVS method assumes a fixed model size (where only a subset of the model is assumed to have non-zero values of β) and updates the parameters from their full conditional distribution. This can be computationally prohibitive even with a modest number of main effects and their higher order interactions, because all possible variables need to be enumerated in the entire model. In contrast, our RJMCMC approach is designed to be scaleable, by allowing the “model” (defined in the context of “spike and slab” as the set of variables with non-zero β) to change dimensions and the variables it contains at each MCMC sample. In our simulations, we carried out analyses allowing for three-way interactions, observing several large BFs for three-way interactions. However, due to computational limitations from the conventional MLE methods (e.g. exhaustive search for interactions, stepwise regression), we limited the comparison of our results to second-order interactions. Like all Bayesian methods, one must remain vigilant of the issue of convergence. We applied a similar approach to the one that was recommended by [Brooks et al., 2001] for assessing convergence in large search spaces. We are currently investigating convergence properties for data sets with much larger dimensionality (e.g. a panel of 1 million SNPs plus interactions), and based on some simulations we have observed that higher quality priors (i.e. reflect disease risk better) can improve the rate of convergence. Our approach is analogous in some respects to the LASSO [Tibshirani, 1996] in that the method seeks to unify variable selection and shrinkage through a common set of tuning parameters. There are some key advantages in our method however. First, we can customize these tuning parameters for each variable through informative priors (weighted by the data) rather than apply a uniform constant across all the variables (i.e. λ in the standard LASSO). Second, an MCMC-based approach allows us to

quantify the uncertainty in estimates of β and the hyperparameters (e.g. σ , τ , π , etc) of the model. Finally, the LASSO does not have provisions to enforce a hierarchical constraint on interactions, which is handled naturally in our RJMCMC algorithm.

Our second demonstration of the method, using the folate metabolism pathway, revealed that interactions with the strongest evidence under each disease mechanism generally appeared to be more biologically plausible based on the topology of the pathway than those found using a more conventional MLE approach like stepwise regression. Also, the MLE of β from a stepwise regression were more unstable than our model-averaged estimates. Although this example used a small data set, the results of this analysis provide proof of principle that such an analysis approach can be carried forward to a larger context such as GWAS and gene expression studies. One could construct eQTL-based priors by making use of gene expression and genotype data from a group of individuals independent from the case-control pool. Such analyses are appealing, given the evidence showing that confirmed hits in GWAS are enriched for eQTLs [Nicolae et al., 2010]. Publicly available data to construct such priors is now available on-line for lymphoblastoid cell lines (<http://www.sanger.ac.uk/humgen/genevar/>) and will be released in the future for a diverse array of tissues (e.g. <http://nihroadmap.nih.gov/GTEX/>). For computational efficiency and ease of interpretation, one would likely first pre-process the expression data using some form of dimensionality reduction across the vast space of expression phenotypes [Zhang and Horvath, 2005]. We are beginning to prototype extensions to our software that enable "low-level" parallelization at either the CPU or the GPU (graphics processing unit) to speed up code at bottlenecks (e.g. fitting a logistic regression model), as well as "high-level" parallelization, thus enabling multiple chains to simultaneously explore the search space for the purposes of faster mixing.

Our last example, using a subset of the data from a real GWAS, shows that our method is flexible enough to accommodate public annotation information as priors when obtaining biomarker or gene expression measurements is not practical. We chose the GO based on its coverage of gene annotations, but investigators can use alternative databases that are more pathway-focused such as KEGG, BioCarta, or PANTHER as these other databases expand their breadth of coverage. Although the GO terms we used in defining the Z matrix for the analysis of the breast cancer data set were not significantly associated with disease risk based on standard errors of π , inclusion of this prior information nevertheless appeared to guide variable selection toward interactions that shared common features, possibly uncovering pathway effects. Our permutation strategy revealed that many of the interactions showing the strongest support based on a BF threshold of 100 or greater were significantly enriched with the GO terms growth, metabolic process, biological regulation, and regulation of biological process. We examined the literature for some of the genes among these interactions, noting a plausible connection to breast cancer. For example, *parkin* (PARK2) has been implicated as a potential tumor suppressor gene in both ovarian and breast cancer based on an observation of down-regulation of gene expression in tumors as well as loss-of-heterozygosity (LOH) at the gene locus [Cesari et al., 2003]. *GRAF* (ARHGAP26), which maps

to 5q31.3, may also be involved in tumor suppressor activity, as 43% of a sample of breast tumors were shown to have LOH at 5q31.3 [Johannsdottir et al., 2006]. We tested both the PARK2*SORCS1 and AK5*ARHGAP26 interaction in an independent breast cancer data set of African American women as described in [Chen et al., 2010] and observed the modest evidence for the AK5*ARHGAP26 interaction ($P < 0.05$).

ACKNOWLEDGMENTS

We thank David Conti and James Baurley for their helpful suggestions. This work was supported in part through NIH grants R01 ES016813 and R01 ES015090, and the California Breast Cancer Program Grant 15UB-8402.

REFERENCES

- Besag J, York J, Mollié A. 1991. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 43:1-20. DOI: 10.1007/BF00116466.
- Brooks SP, Giudici P, Philippe A. 2001. Nonparametric convergence assessment for MCMC model selection. *J Comput Graph Stat* 12:1-22.
- Cesari R, Martin ES, Calin GA, Pentimalli F, Bichi R, McAdams H, Trapasso F, Drusco A, Shimizu M, Masciullo V, D'Andrilli G, Scambia G, Picchio MC, Alder H, Godwin AK, Croce CM. 2003. *Parkin*, a gene implicated in autosomal recessive juvenile parkinsonism, is a candidate tumor suppressor gene on chromosome 6q25-q27. *Proc Natl Acad Sci USA* 100:5956-5961.
- Chen GK, Witte JS. 2007. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet* 81:397-404.
- Chen GK, Millikan RC, John EM, Ambrosone CB, Bernstein L, Zheng W, Hu JJ, Ziegler RG, Henderson BE, Haiman CA, Stram DO. 2010. The potential for enhancing the power of genetic association studies in African Americans through the reuse of existing genotype data. *PLoS Genet* 6:e1001096. DOI: 10.1371/journal.pgen.1001096.
- Conti DV, Cortes V, Molitor J, Thomas DC. 2003. Bayesian modeling of complex metabolic pathways. *Hum Hered* 56:83-93.
- Fridley BL. 2009. Bayesian variable and model selection methods for genetic association studies. *Genet Epidemiol* 33:27-37.
- George E, McCulloch R. 1993. Variable selection via Gibbs sampling. *J Am Stat Assoc* 88:881-889.
- Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711-732.
- Hirschhorn JN. 2009. Genomewide association studies—illuminating biological pathways. *N Engl J Med* 360:1699-1701.
- Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet* 4:e1000130.
- Johannsdottir HK, Jonsson G, Johannesdottir G, Agnarsson BA, Eerola H, Arason A, Heikkilä P, Egilsson V, Olsson H, Johannsson OT, Nevanlinna H, Borg A, Barkardottir RB. 2006. Chromosome 5 imbalance mapping in breast tumors from BRCA1 and BRCA2 mutation carriers and sporadic breast tumors. *Int J Cancer* 119:1052-1060.
- Kass R, Raftery A. 1995. Bayes factors. *J Am Stat Assoc* 90:773-795.
- Kooperberg C, Ruczinski I. 2005. Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* 28:157-170.
- Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. 2007. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol* 31:871-882.

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

Mohlke KL, Boehnke M, Abecasis GR. 2008. Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Hum Mol Genet* 17:R102–R108.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 6:e1000888.

Reed MC, Nijhout HF, Neuhaus ML, Gregory JF, Shane B, James SJ, Boynton A, Ulrich CM. 2006. A mathematical model gives insights into nutritional and genetic aspects of folate-mediated one-carbon metabolism. *J Nutr* 136:2653–2661.

Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. 2003. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics* 4:28.

Thomas D, Conti D, Baurley J, Nijhout F, Reed M, Ulrich CM. 2009. Use of pathway information in molecular epidemiology. *Hum Genomics* 4.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B* 58:267–288.

Viallefont V, Raftery AE, Richardson S. 2001. Variable selection and Bayesian model averaging in case-control studies. *Stat Med* 20:3215–3230.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 4:17.

APPENDIX

The following steps describe the SSVS algorithm in more detail:

- (1) Initialize the model using a random or user-specified set of model variables (SNPs and/or environmental covariates).
- (2) Compute the log-likelihood of the first level of the hierarchical model using logistic regression.

$$L(Y|\beta, X) = \sum_{i=1}^N \log\left(\frac{1}{1 + \exp(-\mu - X_i\beta)}\right)^{Y_i} + \log\left(\frac{\exp(-\mu - X_i\beta)}{1 + \exp(-\mu - X_i\beta)}\right)^{1-Y_i}. \quad (A1)$$

- (3) Sample π from its conditional distribution:

$$\pi = \text{MVN}(\hat{\pi}, \Sigma_{\pi}), \quad (A2)$$

where $\hat{\pi}$ is computed by regressing $\hat{\beta}$ on the Z matrix and Σ_{π} denotes the covariance matrix of $\hat{\pi}$.

- (4) Compute $\bar{\phi}_{-k}$, the mean of the spatial autoregressive random effect for SNP k based on a weighted average of the values taken across row k of A .

$$\bar{\phi}_{-k} = \frac{\sum_{j=1}^m \phi_j A_{jk}}{\sum_{j=1}^m A_{jk}}, \quad (A3)$$

where ϕ_j is assigned the MLE β_j calculated in Step 2. Each pairwise element in A is calculated after retrieving data from a repository (e.g. hash-table, SQL database). While the nature of the data is application-specific, this

calculation generally entails quantifying the similarity (e.g. a Pearson correlation coefficient) between two vectors of values where these values describe a set of attributes (e.g. presence in different pathways) pertaining to the two model variables in question. Similarity scores for a pair of variables calculated in previous models can then be cached in a repository for future use.

- (5) Estimate the residual variance parameters σ^2 and τ^2 from the data as:

$$\sigma^2 = \frac{\sum_{j=1}^M (\hat{\beta}_j - \pi^T Z_j)^2}{\sum_{j=1}^M z_j^2}, \quad \tau^2 = \frac{\sum_{j=1}^M (\hat{\beta}_j - \bar{\phi}_{-j})^2}{\sum_{j=1}^M z_j^2}, \quad z \sim N(0, 1). \quad (A4)$$

- (6) Sample β_k from its conditional distribution:

$$\beta_k \sim N\left(\pi^T Z_k + \bar{\phi}_{-k}, \sigma^2 + \frac{\tau^2}{v_k}\right). \quad (A5)$$

- (7) Draw $\tilde{\beta}$ from the posterior distribution

$$\tilde{\beta}_k \sim N(E(\tilde{\beta}_k), \text{Var}(\tilde{\beta}_k)), \quad (A6)$$

where posterior means and variances are defined as a variance weighted average between the likelihood and the prior component:

$$E(\tilde{\beta}_k) = \frac{\frac{\hat{\beta}_k}{\text{Var}(\hat{\beta}_k)} + \frac{\beta_k}{\text{Var}(\beta_k)}}{\frac{1}{\text{Var}(\hat{\beta}_k)} + \frac{1}{\text{Var}(\beta_k)}}, \quad (A7)$$

$$\text{Var}(\tilde{\beta}_k) = \frac{1}{\frac{1}{\text{Var}(\hat{\beta}_k)} + \frac{1}{\text{Var}(\beta_k)}}. \quad (A8)$$

- (8) Propose a change to the model where a SNP can be either deleted, added, or swapped with another SNP that is not in the current model. The probability of selecting a SNP for an add or swap move is governed by the proposal kernel in Equation (6). To maintain the hierarchical structure of interactions, delete and swap moves are restricted to variables of any order in the current model not referenced by a higher order interaction. However, for an add move, we allow the addition of a new main effect or a higher order interaction comprised of variables already in the model. Because our likelihood function is not conjugate to its prior, we rely on the following Metropolis-Hastings ratio:

$$r = \frac{L(Y|\beta', X, M')P(\beta'|Z, \pi, A, \tau, \sigma, M')P(M \rightarrow M')}{L(Y|\beta, X, M)P(\beta|Z, \pi, A, \tau, \sigma, M)P(M' \rightarrow M)}, \quad (A9)$$

and accept with probability $\min(1, r)$.

- (9) Repeat steps 2–8 for some number of iterations where posterior realizations from the burn-in period are discarded to avoid correlation to initial parameter values.