Published in final edited form as:
Genet Epidemiol. 2012 May ; 36(4): 352–359. doi:10.1002/gepi.21628.

A Bayesian Integrative Genomic Model for Pathway Analysis of Complex Traits

Brooke L. Fridley[1,*], **Steven Lund**[2], **Gregory D. Jenkins**[1], and **Liewei Wang**[3]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota

[2]Department of Statistics, Iowa State University, Ames, Iowa

[3]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, Minnesota

Abstract

With new technologies, multiple types of genomic data are commonly collected on a single set of samples. However, standard analysis methods concentrate on a single data type at a time and ignore the relationships between genes, proteins, and biochemical reactions that give rise to complex phenotypes. In this paper, we propose a novel integrative model to incorporate multiple types of genomic data into an association analysis with a complex phenotype. The method combines path analysis and stochastic search variable selection into a Bayesian hierarchical model that simultaneously identifies both direct and indirect genomic effects on the phenotype. Results from a simulation study and application of the Bayesian model to a pharmacogenomic study of the drug gemcitabine demonstrate greater sensitivity to detect genomic effects in some simulation scenarios, when compared to the standard single data type analysis. Further research is required to extend and modify this integrative modeling framework to increase computational efficiency to best capitalize on the wealth of information available across multiple genomic data types.

Keywords

stochastic search variable selection; Markov chain Monte Carlo (MCMC); genetic association; single nucleotide polymorphism (SNPs); mRNA expression; cell lines

INTRODUCTION

With the wealth of data being produced by new technologies, it is becoming common to collect multiple types of genomic data on a set of samples. Currently, standard analysis methods concentrate on a single data type, or experiment, at a time. Few statistical methods have been developed to maximize the use of the enormous amount of genomic data to unravel the etiology of complex disease and drug-related phenotypes. Joint analysis of multiple types of genomic data would likely provide novel insights, especially when the etiology of the disease or phenotype is complex [Reif et al., 2004].

A few methods have been proposed to integrate multiple forms of genomic data, sometimes referred to as “integrative genomics” or “genomic convergence” [Hauser et al., 2003; Schadt et al., 2005]. Other approaches involve a multistep procedure to identify potential key

[*]Correspondence to: Brooke L. Fridley, Harwick 766, Department of Health Science Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905. fridley.brooke@mayo.edu.

drivers of complex traits by integrating genetic variation and mRNA gene expression data [Li et al., 2009; Niu et al., 2010]. Schadt et al. [2005] describe a multistep method that uses conditional correlation measures and a likelihood-based causality model selection to model the relationship between DNA variation, mRNA expression, and clinical traits. Another type of integrative analysis involves approaches that complete a comprehensive analysis, as opposed to a multistep procedure. However, these approaches have mostly been described in the context of a small number of genomic variables, such as those contained within a pathway. Nock et al. [2007] proposed an approach based on structural equation modeling (SEM), whereas Conti et al. [2003] used a Bayesian approach to model the complex relationship within a small, biological pathway of interest.

Similar to these comprehensive approaches, we propose a modeling framework that combines Bayesian pathway analysis [Congdon, 2007] and Bayesian variable selection using stochastic search variable selection (SSVS) [Fridley, 2009; George and McCulloch, 1993] into a novel modeling framework to identify both direct and indirect genomic effects on a complex phenotype. The novel Bayesian integrative genomic model is illustrated using data from mRNA expression and single nucleotide polymorphisms (SNPs) within the gemcitabine drug pathway from a pharmacogenomic study of the drug gemcitabine. We also present simulation studies to assess the ability of the proposed model to detect simulated effects. Finally, we compare results from the Bayesian integrative genomic model to results from the standard “one-at-a-time” analysis using both the pharmacogenomic data and simulated data.

METHODS AND MATERIALS

PHARMACOGENOMIC STUDY OF GEMCITABINE

To understand the pharmacogenomics of gemcitabine drug therapy, the Coriell Human Variation Panel lymphoblastoid cell lines have been utilized extensively [Li et al., 2009; Niu et al., 2010]. EBV-transformed B lymphoblastoid cells were derived from approximately 60 Caucasian American (CA), 60 African American (AA), and 60 Han Chinese-American (HCA) subjects. Gemcitabine cytotoxicity data for the cell lines were collected at drug dosages 1,000, 100, 10, 1, 0.1, 0.01, 0.001, and 0.0001 uM. The quantitative phenotype IC50 (effective dose that kills 50% of the cells) was estimated using a four parameter logistic model [Gallant, 1987], and was used as the drug-response phenotype. A large value of IC50 indicates that a cell line is resistant to the drug, whereas a small value indicates a cell line is sensitive to the drug.

Genotyping of SNPs on the Illumina's Infinium Human-Hap 550K and 510S BeadChips for the cell lines was completed at the Genotyping Shared Resources at the Mayo Clinic in Rochester, Minnesota. Quality control was completed by excluding SNPs with Hardy-Weinberg Equilibrium (HWE) P-value < 0.001 (minimum P-value between exact test for HWE by race [Guo and Thompson, 1992; Wigginton et al., 2005] and stratified test for HWE [Schaid et al., 2006]), minor allele frequency <5%, or call rate <95% from further analyses. Missing genotypes within the gemcitabine pathway were imputed prior to analyses using the program *fastPHASE* [Scheet and Stephens, 2006]. This resulted in 15 genes within the gemcitabine pathway containing 265 SNPs.

Whole genome expression data for the cell lines was obtained with Affymetrix U133 Plus 2.0 expression array chip, which contains over 54,000 probe sets. The mRNA expression array data were normalized on the $\log_2$ scale using GCRMA methodologies [Bolstad et al., 2003; Irizarry et al., 2003; Wu et al., 2004]. This data contained 20 probe sets for genes within the gemcitabine pathway. The goal of this study is to complete a comprehensive integrative analysis of both genetic variation (in the form of SNPs) and mRNA gene

expression variation to determine their collective impact on gemcitabine response as measured by the IC_{50}.

STANDARD ANALYSIS APPROACH

The standard analysis approach for the analysis of multiple types of genomic data is to complete, for each data type, the univariate association of the genomic variable with the phenotype. Following the analysis of each genomic data type with the phenotype, to understand these association results, expression quantitative trait loci (eQTLs) analysis is then completed to determine the association of each SNP with the level of gene expression. *P*-values for each effect type are adjusted according to Bonferroni's multiple testing procedure. Throughout this paper, we will refer to this analysis approach as the standard all-pairwise correlation analysis, denoted as APCA.

BAYESIAN INTEGRATIVE MODEL

To model the joint relationship of mRNA gene expression and SNP genotypes on a quantitative phenotype (e.g. drug cytotoxicity), we propose the following Bayesian integrative model. Let Y_i be the phenotypic value for subject i (i = 1, …, N), and let $\underline{SNP_i} = (SNP_{i1}, SNP_{i2}, \ldots, SNP_{iK})^T$ represent the SNP genotypes of subject i where SNP_{ik} is the genotype (under additive coding in terms of the number of minor alleles) for SNP k (k = 1, …, K). Next, let $\underline{GE_i} = (GEi1, \ldots, GE_{iJ})^T$ represent a vector of gene expression values for subject i, where GE_{ij} is the standardized mRNA expression value for gene j (j = 1, …, J) for subject i. The first step in specifying the Bayesian integrative model is to specify the direct effects of the SNPs and mRNA expression on the quantitative phenotype, where the phenotype follows a normal distribution whose mean is a function of both the SNP genotypes and the mRNA expression levels. Let $Y_i \tilde{} N\left(\mu_{Yi}, \sigma_Y^2\right)$ for i = 1, …, N, where $\mu_{yi} = b_0 + \underline{b}^T\underline{SNP}_i + \underline{c}^T\underline{GE}_i$, $\underline{b} = (b_1, \ldots, b_K)^T$ represents a vector of K direct SNP effects with b_k representing the direct effect of SNP k on the phenotype, $\underline{c} = (c_1, \ldots, c_J)^T$ represents a vector of J mRNA expression effects with c_j representing the direct effect of the mRNA expression value for gene j on the phenotype.

Next, we model the effect of SNPs on the phenotype via the mRNA expression levels. In assessing the association between the SNPs and the mRNA expression levels, there are two possibilities (1) assessment of only *cis*-acting SNPs (i.e. SNPs within the 5′ r 3′ genomic region of a gene); or (2) to assess both *cis*- and *trans*-acting SNPs (i.e. SNPs not near or within the 5′ or 3′ genomic region of the gene). Incorporating only *cis*-relationships in a model reduces the model complexity, which eases model evaluation; however, this limits the types of relationships that can be discovered and assessed. For models examining both *cis*- and *trans*-relationships between SNPs and mRNA expression (i.e. all possible SNP-mRNA pairs), for every subject i 1, …, N and gene j = 1, …, J, let $GE_{ij} \tilde{} N\left(\mu_{GEij}, \sigma_{GEj}^2\right)$ where $\mu_{GEij} = a_{j0} + \underline{a}_j^T \underline{SNP}_i$ and $\underline{a}_j = (a_{j1}, \ldots, a_{jK})^T$ where $a_{jk} a_{j1}$ represents the effect of SNP k on mRNA expression value for gene j.

In contrast, for models examining only *cis*-acting SNPs, let $\underline{SNP_{ij}}$ be the subset of $\underline{SNP_i}$ containing SNPs within gene j and N_j be the number of SNPs within gene j (i.e. length of the vector $\underline{SNP_{ij}}$) with $GE_{ij} \tilde{} N\left(\mu_{GEij}, \sigma_{GEj}^2\right)$ where $\mu_{GEij} = a_{j0} + \underline{a}_j^T \underline{SNP}_{ij}$ and $\underline{a}_j = (a_{j1}, \ldots, a_{jN_j})^T$, where a_{jk} represents the *cis*-acting effect of SNP k on the mRNA expression for gene j. The effect of SNPs on mRNA expression modeled in this fashion could be indirect effects of SNPs on the phenotype acting through mRNA expression, or less desirable, eQTLs not related to the phenotype. For the remainder of this paper, we will not make the distinction between SNP effects on mRNA expression that are associated or not associated with the

phenotype. The effects of SNPs and mRNA on the phenotype are referred to as "direct" effects.

PRIOR DISTRIBUTIONS FOR THE INTEGRATIVE MODEL

Bayesian approaches to model selection and shrinkage are useful for regression-based analysis of genomic data since the uncertainty in model choice can be incorporated into the inferences for genomic effects. Additionally, Markov chain Monte Carlo (MCMC) and stochastic search methods efficiently interrogate the model space without fitting all possible models. To model SNP effects on mRNA expression and a quantitative phenotype, we use SSVS for Bayesian model averaging with "shrinkage" of SNP effects toward zero [Conti et al., 2003; Conti and Gauderman, 2004; Fridley, 2009; George and McCulloch, 1993; Kim et al., 2004]. Under this SSVS prior model for the SNP effects, the prior distribution of each SNP coefficient is a mixture of two normal distributions that represent two cases: the SNP is selected for inclusion in the model or the SNP is not selected for inclusion in the model. Coefficients for SNPs included in the model have a normal prior distribution centered at zero with a large variance, while coefficients for SNPs not included in the model have a normal prior distribution centered at zero but with a small variance (and thus will be estimated close to zero).

For modeling the direct SNP effects (b_k `s), let $\phi = (\phi_1, \phi_2, \ldots, \phi_K)$ be a vector of latent indicator variables for every SNP $k = 1, \ldots, K$, where $\phi_k = 1$ if SNP_k is included in the modeling of the phenotype or $\phi_k = 0$ if SNP_k is omitted from modeling the phenotype. Next, the latent indicators are modeled as $\phi_k | p_b \sim Bernoulli\ (p_b)$ with $p_b \sim Beta(\alpha_1, \beta_1)$. To determine appropriate prior values for the parameters α_1 and β_1, we followed an approach similar to one outlined by Chen et al. [2004]. We assumed a priori that only 10% of the K SNPs will be related to the phenotype, which suggests the mode of the prior distribution is 0.1. We further assumed that the prior mean will be 0.12, 20% larger than the prior mode, resulting in values of $\alpha_1 = 4.8$ and $\beta_1 = 35.2$, where $\alpha_1 = (\text{mean} - 2 \times \text{mode} \times \text{mean})/(\text{mean} - \text{mode})$ and $\beta_1 = [(1 - 2 \times \text{mode})(1 - \text{mean})]/(\text{mean} - \text{mode})$. Other possible priors for p_b include $p_b \sim Uniform(0,1)$ or fixing p_b to a constant (e.g. $p^b = 0.10$). Similarly, for indirect SNP effects on the phenotype via the mRNA expression levels (ajk `s), let $\gamma_j = (\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jK})$ be a vector of latent indicator variables, where $\gamma_{jk} = 1$ if SNP K is included in the modeling of the mRNA expression level for gene j or $\gamma_{jk} = 0$ if SNP K is omitted from the modeling of mRNA gene expression level for each gene $j = 1, \ldots, J$.

Using this prior SSVS model for the=SNP effects, the conditional distributions for the direct SNP effects (b_k `s) and the indirect SNP effects on the phenotype via mRNA gene expression ($a_{jk}$ `s) given their respective latent indicators are then given by

$$b_k|\varphi_k \tilde{} (1 - \varphi_k)\, N\left(0, w^2\right) + \varphi_k N\left(0, z^2 w^2\right) \text{ for } j=1, \ldots, J \text{ and } a_{jk}|\gamma_{jk} \tilde{} (1 - \gamma_{jk})\, N\left(0, w^2\right) + \gamma_{jk} N\left(0, z^2 w^2\right) \text{ for } j=1, \ldots, \mathrm{J}$$

for $k = 1, \ldots, K$, where z and w are constants specified such that w^2 is small and $z^2\, w^2$ is large. For our analyses, we chose to shrink coefficient estimates less than $\delta = 0.005$ to zero and chose $z = 50$. This results in the value of w equal to 0.00179, using the formula

$w = \sqrt{\frac{z^2 - 1}{2z^2 \log(z)}} \times \delta$ [Gilks et al., 1996]. The reader is referred to George and McCulloch [1993] and Wakefield and Bennett [1996] for further insight and guidance on choice of *z* and *w*.

For the remaining parameters in the Bayesian integrative genomic model, diffuse proper prior distributions were placed on the model parameters. In particular, $b_0 \sim N(0, 1{,}000)$, $\sigma_Y^{-2} \sim Gamma(0.01, 0.01)$, $\sigma_{GE_j}^{-2} \sim Gamma(0.01, 0.01)$, $a_{10}, \ldots, a_{J0} \sim N(0, 1{,}000)$, $\underline{c} \sim MVN(\underline{\mu}_{GE}, \mathbf{V})$, $\underline{\mu}_{GE} \sim MVN(\underline{\mu}, \mathbf{V}_\mu)$ and $\mathbf{V} \sim Wishart(\mathbf{R}, J)$ with $\underline{\mu} = (0, \ldots, 0)^T$ and $\mathbf{R} = 0.1^*\mathbf{I}$, where *J* is the number of gene expression variables in the model and **I** is the identity matrix.

SIMULATION STUDY

To assess the ability of the Bayesian integrative model to detect both direct and indirect effects, a simulation study was completed, with the simulation scenarios depicted in Figure 1. The first simulation scenario, Scenario A (Fig. 1A), includes two SNP effects on the phenotype and on the gene expression of the corresponding gene with the gene expression of these genes also associated with the phenotype. Scenario B (Fig. 1B) has two *cis*-SNP effects on gene expression with the gene expression associated with the phenotype. The third simulation scenario, Scenario C (Fig. 1C), consists of a SNP with both a direct-SNP effect on the phenotype and a *cis*-acting effect on gene expression, with the gene expression of this gene also associated with the phenotype. In addition to these effects, the gene expression of a second gene is also associated with the phenotype. The last simulation scenario is the null scenario (Scenario D) with no genomic effects on the phenotype (Fig. 1D).

Genotypic data for SNPs within five genes were simulated based on the genotypic data collected on 60 Caucasian cell lines from the gemcitabine pharmacogenomic study. SNPs were tagged using *Haploview* (http://www.broad.mit.edu/mpg/haploview/) [Barrett et al., 2005] using a threshold for r^2 of 0.64 and minor allele frequency of 0.05. This resulted in seven SNPs within Gene 1, 15 SNPs within Gene 2, five SNPs within Gene 3, eight SNPs within Gene 4, and four SNPs within Gene 5, with a total of 39 SNPs in the five genes. Haplotype frequencies were estimated using the *haplo.em* function within the *haplo.stats* R library (http://cran.r-project.org/web/packages/haplo.stats/index.html). For each cell line, the haplotype with the greatest estimated frequency was used as the “true” haplotype, creating an underlying population of 120 haplotypes. Six hundred haplotypes were simulated using the *hapsim* library in R (http://cran.r-project.org/web/packages/hapsim/index.html) based on this underlying haplotype population. These simulated haplotypes were then assigned to 300 simulated individuals.

Conditioned on the simulated genotypic data, mRNA expression data for five genes were simulated for each individual according to $GE_{ij} \sim N\left(\underline{a}_j^T \underline{SNP}_i, \sigma_{GE}^2 = 1\right)$. Finally, a quantitative phenotype for each individual was simulated based on both SNP and mRNA values according to $Y_i \sim N\left(\underline{b}^T \underline{SNP}_i + \underline{c}^T \underline{GE}_i, \sigma_Y^2 = 10\right)$. For the simulation scenarios, non-zero coefficients for gene expression effects on the phenotype ($\underline{c}$) were chosen to be 0.4, whereas non-zero coefficients for SNP direct ($\underline{b}$) and indirect ($\underline{a}_j$) effects were set to 0.88 and 0.3, respectively. For the simulations, Gene 3 and Gene 5 were considered to have mRNA expression effects on the phenotype; SNP 2 in Gene 3 and SNP 4 in Gene 5 were consider to have genetic effects on the phenotype, with these two SNPs also having *cis*-acting effects on Gene 3 and Gene 5, respectively. The various scenarios are also depicted in Figure 1 and Table I. For each simulation, the mRNA expression data were standardized for

each gene prior to analysis. Data for each scenario were simulated with n = 300 subjects and 50 replicates.

Following the simulation of the synthetic datasets, each simulated data set was analyzed using both the standard APCA and the Bayesian integrative model assessing only *cis*-relationships. Each simulated data set included simulated mRNA expression data for five genes and genotypes for 39 SNPs within the five genes. For the APCA, significant effects were identified using a two-sided, Bonferroni-corrected *P*-value to account for multiple testing. For the Bayesian integrative model, SNP effects were considered significant (or selected) using the median model decision rule [Barbieri and Berger, 2004; Swartz et al., 2008] (i.e. $P(\gamma_{jk}|\text{ data}) \geq 0.50$ for indirect SNP effects or $P(\phi_k|\text{ data}) \geq 0.50$ for direct SNP) and significant gene expression effects were identified according to their 95% posterior credible interval (i.e. if posterior credible interval does not contain zero). In addition to using the median model decision rule to determine if SNP effects were significant, we also considered significance of the SNP effects if the $P(\gamma_{jk} | \text{ data}) \geq 0.25$ or 0.35 and $P(\phi_k|\text{ data}) \geq$ 0.10, 0.15, or 0.25. While power and type I error rates are not clearly defined in the Bayesian framework, each coefficient was either declared to be significant or not. Power for the Bayesian integrative model was defined as the proportion of truly non-zero coefficients that were declared significant, while the type I error rates were defined as the proportion of truly zero coefficients declared significant.

RESULTS

SIMULATION STUDY

The type I error rate estimates and power estimates for Bayesian integrative model and APCA are presented in Tables II and III, respectively. Due to the difference in significance testing and philosophy of these two approaches, it is not possible to complete a fair "head-to-head" comparison of the approaches. For example, in the Bayesian analysis, the choice of prior can impact the error rates, along with there being no concept of adjustment for multiple testing. Therefore, we present the results for these two approaches in separate tables, noting the difference in interpretation of the results from these two contrasting approaches.

For the Bayesian integrative model, the type I error rates ranged from 0.000 to 0.068 for the effects determined significant based on 95% credible intervals, while the type I error rates for the APCA approach had error rates that ranged from 0.000 to 0.002. The more conservative error rates for APCA are attributed to the Bonferroni adjustment for multiple testing, where effects were determined to be significant if the Bonferroni adjusted *P*-value was less than 0.05. However, as depicted in Table II, the choice of significance threshold for the SNP effects modeled with the SSVS mixture prior (e.g. $P(\gamma|\text{ data}) > 0.25$, 0.35, or 0.50) greatly impacts the error rates. As the threshold in the posterior probability of the indicator given the data is lower, more variables are selected resulting in an increase in the type I error rate. In addition, the posterior distribution of the indicator variables is tied to the choice of normal priors used in the SSVS (e.g. choice of z and w) and the prior used for modeling p_b and p_b (e.g. $p_b \sim \text{Unif}(0,1)$ or $p_b \sim Beta(\alpha_1, \beta_1)$). If more "stringent" priors for the SNP effects were used, a larger amount of "shrinkage" would be completed resulting in fewer variables selected for the model.

Power estimates for the Bayesian integrative model, broken down for the various parameters and simulation scenarios are presented in Table II. Power for the mRNA effects on the phenotype (c_3 and c_5) ranged from 0.50 to 0.80. We have chosen to focus the rest of our discussion of power on the SNP effects, where significance is determined by $P(\gamma|\text{ data}) >$ 0.25 or $P(\phi|\text{ data}) > 0.15$, as these thresholds produced reasonable type I error rates. For the direct effects of SNPs on the phenotype, the power ranged from 0.56 to 0.70, while for the

indirect SNP effects via mRNA, the power ranged from 0.68 to 0.96. The power for the standard APCA approach is presented in Table III. In contrast, power (controlling for type I error rate after adjustment for multiple testing at the 0.05 level) for the standard APCA approach was considerably lower, mostly due to the adjustment for multiple testing. The highest power was for detecting mRNA expression effects on the phenotype, as only five genes were tested (power ranged from 0.14 to 0.30). The power of detection for either SNP-phenotype association or *cis*-acting SNP-mRNA associations ranged from 0.00 to 0.12, with multiple testing adjustments for 39 SNP association tests. In an attempt to make a "fair" comparison of power for the APCA method and the Bayesian integrative model for detecting SNP effects (direct or indirect), the power must be observed where the type I error rates are similar. In this case, the significance for SNP effects would be determined using $P(\gamma|\ \text{data}) > 0.35$ and $P(\phi|\ \text{data}) > 0.25$. Using this condition, both methods had low power to detect SNP-mRNA associations, along with the effect of SNP 4 in gene 5 on the phenotype (Bayesian power = 0.06, APCA power = 0.12 for Scenario A). However, for Scenario A, power to detect a SNP effect for SNP 2 in gene 3 was 0.60 for the Bayesian analysis and 0.10 for APCA.

PHARMACOGENOMIC STUDY OF GEMCITABINE

Next, we applied the Bayesian integrative model and the standard APCA approach to the gemcitabine study using the 171 cell lines that had both SNPs (265 SNPs in 15 genes) and mRNA expression variables (20 probe sets) measured within the gemcitabine pathway. All analyses were adjusted for race and gender, with the phenotype gemcitabine IC50 transformed to the log scale. To enable the assessment of both *cis*- and *trans*-acting SNPs on mRNA gene expression in the Bayesian integrative model, we further reduced the dimension of the genotype data and the complexity of the model space as follows. We partitioned SNPs within a gene into bins based on their correlation using a hierarchical clustering method [Rinaldo et al., 2005] with a liberal threshold of 0.05, followed by principal components analysis [Gauderman et al., 2007; Mardia et al., 1979] for SNPs within each bin. The first principal component for each bin of SNPs was used in the model as the "genotypic" variable, as opposed to the individual SNP genotypes. This dimension reduction approach resulted in 38 genotypic factors used in both the Bayesian integrative model and the APCA.

Significant associations of the phenotype IC50 with the genetic factors and mRNA expressions are presented in Table IV for both the Bayesian integrative model and the APCA approaches. Based on results from the simulation study, we chose to determine significant SNP effects if $P(\gamma_{jk}|\ \text{data}) \geq 0.25$ for indirect SNP effects and $P(\phi_k|\ \text{data}) \geq 0.15$ for direct SNP. After using the Bonferroni adjustment to account for multiple testing within each effect type, the APCA identified three direct gene expression effects, zero direct SNP effects, and three indirect SNP effects. The Bayesian integrative model identified two direct gene expression effects (both were among the three identified by the APCA), zero direct SNP effects, and five indirect SNP effects (including the three identified by the APCA).

Table IV also includes results from a sensitivity analysis for the prior distribution of the latent variable inclusion probabilities. Alternative priors examined included p_a and $p_b \sim Uniform(0,1)$ and $p_a = p_b = 0.2$. The analysis in which p_a and $p_b \sim Uniform(0,1)$ was most conservative (posterior inclusion probabilities indirect SNP effects range from 0.13 to 1.00), while the analysis where $p_a = p_b = 0.2$ was the most liberal (posterior inclusion probabilities for indirect SNP effects range from 0.51 to 1.00). The top results (ranked by posterior probability) were similar across the three prior distributions used to model p_a and p_b, with the most significant effect for expression of NT5C3L detected by all models (P = 1.42e–18, posterior probability of selection of 1.00 for all three prior models).

DISCUSSION

Phenotypes related to an individual's response to a drug are likely the products of complex networks involving genetic variation, mRNA, proteins, etc. along with external environmental factors. Advances in technology are now allowing the acquisition of these multiple types of genomic data from a single set of samples. When combined with statistical methods appropriate for analyzing such complex networks, this data may generate substantial new insights into the genomic basis of complex traits and phenotypes, such as response to cancer therapeutics. In this paper, we proposed a novel Bayesian integrative model that combines ideas of path analysis with SSVS for a comprehensive modeling strategy. Results from applying this model to the pharmacogenomic study of the drug gemcitabine and simulation studies show the Bayesian integrative model is able to capture the relationship between SNP variation, mRNA expression variation, and the complex phenotype, if one exists.

In the specification of the prior model, we chose to use SSVS priors only for the SNP effects (direct and indirect) and not the mRNA effect, due to the small number of expression probe sets (in relation to the number of SNPs) included in the analysis. The model could be extended to also include SSVS priors on the mRNA expression effects. However, this would add an additional layer of complexity and computation time. Additional care in model specification and interpretation of results from this more complex model would also be needed, particularly in the context of the SNP indirect effects on the phenotype via mRNA.

Implementing the Bayesian integrative model for a comprehensive analysis offers several advantages. First, prior biological knowledge can be easily incorporated into the model thought the prior model specification. The incorporation of biological information will aid in the integrative model's ability to detect biologically relevant loci related to complex traits. Second, inference for a function of effects, such as the collective indirect effect of SNP *k* on the phenotype ($\sum_{j=1}^{J} a_{jk}c_j$), can be conducted easily using MCMC and posterior credible intervals. Third, uncertainty related to selection of important genotype variables is accounted for as a form of Bayesian model averaging. The implementation of the Bayesian integrative model via MCMC is straightforward, however it comes with a high computational cost. The computational burden of the MCMC simulation limits the number of SNP or mRNA expression variables that can be included in association studies. Similar to previous proposed approaches involving SEM [Nock et al., 2007] or Bayesian approaches [Conti et al., 2003], our Bayesian integrative model is only able to handle a few hundred SNP and mRNA expression variables, such as those within a given pathway. However, as demonstrated with the gemcitabine study presented in this paper, one can reduce the dimensionality of genetic datasets to create manageably complex models. To follow-up the results from the Bayesian integrative model, standard regression techniques can be applied using the significant, possibly modified, genotype or gene-expression variables identified.

In conclusion, this paper presented a novel Bayesian integrative model for modeling the complex relationship between genetic variation in the form of SNPs and mRNA expression variation, and their collective impact on a complex phenotype. Further research is needed to develop integrative analysis approaches for the study of complex diseases and phenotypes. In particular, further research is required to extend and modify this integrative modeling framework to increase computational efficiency and extend this model to other data types to capitalize on the wealth of information available across multiple genomic data types.

Acknowledgments

This research was supported by the NIH GM61388, CA140879, CA130828, CA138461, GM86689, Minnesota Partnership for Biotechnology and Medical Genomics, and the Mayo Foundation.

REFERENCES

Barbieri MM, Berger JO. Optimal predictive model selection. Ann Stat. 2004; 32:870–897.

Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21(2):263–265. [PubMed: 15297300]

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003; 19(2):185–193. [PubMed: 12538238]

Chen, W.; Ghosh, D.; Raghuanthan, TE.; Kardia, S. A Bayesian Method for Finding Interactions in Genomic Studies. Vol. 48. Berkeley Electronic Press; University of Michigan School of Public Health: 2004. p. 31

Congdon, P. Bayesian Statistical Modelling. John Wiley & Sons, Ltd.; West Sussex, UK: 2007.

Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. Hum Hered. 2003; 56(1–3):83–93. [PubMed: 14614242]

Conti DV, Gauderman WJ. SNPs, haplotypes, and model selection in a candidate gene region: the SIMPle analysis for multilocus data. Genet Epidemiol. 2004; 27(4):429–441. [PubMed: 15543635]

Fridley BL. Bayesian variable and model selection methods for genetic association studies. Genet Epidemiol. 2009; 33(1):27–37. [PubMed: 18618760]

Gallant, AR. Nonlinear Statistical Models. Wiley; New York: 1987.

Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol. 2007; 31(5):383–395. [PubMed: 17410554]

George EI, McCulloch RE. Variable selections via Gibbs sampling. J Am Stat Assoc. 1993; 88(423): 881–889.

Gilks, WR.; Richardson, S.; Spiegelhalter, DJ. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC; Boca Raton, FL: 1996.

Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics. 1992; 48(2):361–372. [PubMed: 1637966]

Hauser MA, Li YJ, Takeuchi S, Walters R, Noureddine M, Maready M, Darden T, Hulette C, Martin E, Hauser E, Xu H, Schmechel D, Stenger JE, Dietrich F, Vance J. Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. Hum Mol Genet. 2003; 12(6):671–677. [PubMed: 12620972]

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003; 31(4):e15. [PubMed: 12582260]

Kim KW, Suh YJ, Park WY, Jhoo JH, Lee DY, Youn JC, Lee KH, Seo JS, Woo JI. Choline acetyltransferase G +4 A polymorphism confers a risk for Alzheimer's disease in concert with Apolipoprotein E epsilon4. Neurosci Lett. 2004; 366(2):182–186. [PubMed: 15276243]

Li L, Fridley BL, Kalari K, Jenkins G, Batzler A, Weinshilboum RM, Wang L. Gemcitabine and arabinosylcytosin pharmacogenomics: genome-wide association and drug response biomarkers. PLoS One. 2009; 4(11):e7765. [PubMed: 19898621]

Mardia, KV.; Kent, JT.; Bibby, JM. Multivariate Analysis. Academic Press; London: 1979.

Niu N, Qin Y, Fridley BL, Hou J, Kalari KR, Zhu M, Wu TY, Jenkins GD, Batzler A, Wang L. Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines. Genome Res. 2010; 20(11):1482–1492. [PubMed: 20923822]

Nock NL, Larkin EK, Morris NJ, Li Y, Stein CM. Modeling the complex gene x environment interplay in the simulated rheumatoid arthritis GAW15 data using latent variable structural equation modeling. BMC Proc. 2007; 1(Suppl 1):S118. [PubMed: 18466459]

Reif DM, White BC, Moore JH. Integrated analysis of genetic, genomic and proteomic data. Expert Rev Proteomics. 2004; 1(1):67–75. [PubMed: 15966800]

Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, Roeder K. Characterization of multilocus linkage disequilibrium. Genet Epidemiol. 2005; 28(3):193–206. [PubMed: 15637716]

Schaid DJ, Batzler AJ, Jenkins GD, Hildebrandt MA. Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata. Am J Hum Genet. 2006; 79(6):1071–1080. [PubMed: 17186465]

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005; 37(7):710–717. [PubMed: 15965475]

Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Human Genet. 2006; 78(4):629–644. [PubMed: 16532393]

Swartz MD, Yu RK, Shete S. Finding factors influencing risk: comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. Stat Med. 2008; 27(29):6158–6174. [PubMed: 18937224]

Wakefield JC, Bennett JE. The Bayesian modeling of covariates for population pharmacokinetic models. J Am Stat Assoc. 1996; 91:917–927.

Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005; 76(5):887–893. [PubMed: 15789306]

Wu Z, Irizarry R, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligobucleotide expression arrays. J Am Stat Assoc. 2004; 99(468):909–917.

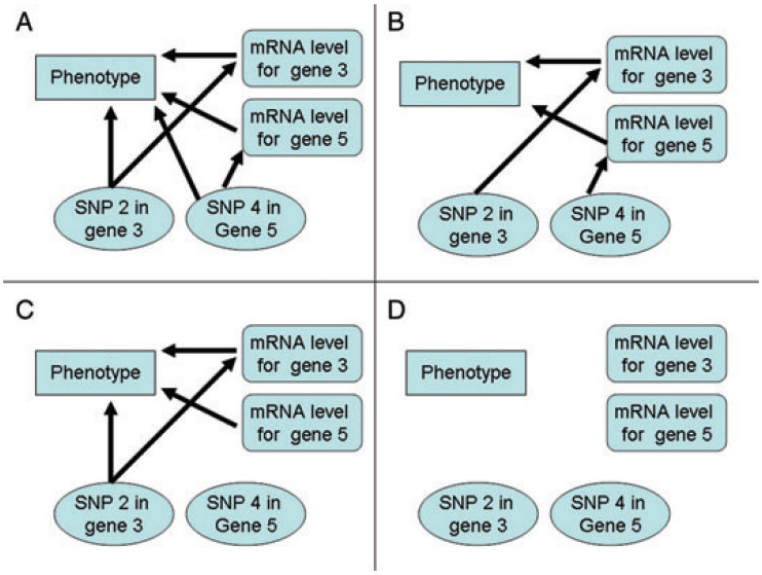

Fig. 1.

Diagram depicting the four simulation scenarios. (A) Scenario A includes two *cis*-acting SNP effects on gene expression and the phenotype; (B) Scenario B has two *cis*-acting SNPs on gene expression, with the gene expression associated with the phenotype; (C) Scenario C consists of a *cis*-acting SNP effect on gene expression and the phenotype with the expression of this gene also associated with the phenotype, along with the gene expression of a second gene associated with the phenotype; (D) Scenario D is the null scenario (no genomic effects on the phenotype).

TABLE I

Parameter values used in the simulation scenarios. Scenarios A, B, and C had genomic effects, while Scenario D was the null scenario (no genomic effects simulated)

		Scenario			
Effect type	**Parameter**	**A**	**B**	**C**	**D**
Direct SNP	SNP 2 in Gene 3; b_{24}	0.88	0	0.88	0
	SNP 4 in Gene 5; b_{39}	0.88	0	0	0
Indirect SNP	*cis*-acting SNP 2 in Gene 3; $a_{3,2}$	0.3	0.3	0.3	0
	cis-acting SNP 4 in Gene 5; $a_{5,4}$	0.3	0.3	0	0
Direct mRNA	Gene 3; c_3	0.4	0.4	0.4	0
	Gene 5; c_5	0.4	0.4	0.4	0

TABLE II

Power and type I error rate estimates for the Bayesian integrative model for the three genetic models (Scenarios A, B, and C) and a null simulation scenario (Scenario D). Estimates are based on 50 simulations

Relationship assessed	Type I error rates					Power			
		Scenario					Scenario		
	Parameter	A	B	C	D	Parameter	A	B	C
Phenotype and mRNA expression	c's	0.053	0.053	0.060	0.068	c_3	0.62	0.50	0.72
						c_5	0.80	0.64	0.56
Indicator for SNP and mRNA expression relationship[a]	γ					$\gamma_{3,2}$ (SNP 24)			
	0.25	0.053	0.045	0.043	0.013	0.25	0.84	0.78	0.68
	0.35	0.000	0.000	0.000	0.000	0.35	0.02	0.00	0.00
	0.50	0.000	0.000	0.000	0.000	0.50	0.00	0.00	0.00
						$\gamma_{5,4}$ (SNP 39)			
						0.25	0.90	0.96	–
						0.35	0.00	0.00	–
						0.50	0.00	0.00	–
Indicator for SNP and phenotype relationship[b]	ϕ					ϕ_{24}			
	0.10	0.999	0.998	0.999	0.997	0.10	1.00	–	1.00
	0.15	0.039	0.005	0.027	0.011	0.15	0.70	–	0.56
	0.25	0.001	0.000	0.000	0.000	0.25	0.60	–	0.00
	0.50	0.000	0.000	0.000	0.000	0.50	0.00	–	0.00
						ϕ_{39}			
						0.10	1.00	–	–
						0.15	0.64	–	–
						0.25	0.06	–	–
						0.50	0.00	–	–

[a] Significance determined if P(γ| data) > 0.25, 0.35, or 0.50.

[b] Significance determined if P(ϕ| data) > 0.10, 0.15, 0.25, or 0.50.

TABLE III

Power and type I error rate estimates for the APCA for the three genetic models (Scenarios A, B, and C) and a null simulation scenario (Scenario D). Estimates are based on 50 simulations and the Bonferroni adjusted *P*-values

Relationship assessed	Type I Error Rates				Power			
	Scenario					Scenario		
	A	B	C	D	Parameter	A	B	C
Phenotype and mRNA expression	0.000	0.000	0.000	0.000	c_3	0.18	0.14	0.26
					c_5	0.30	0.16	0.16
SNP and mRNA expression	0.001	0.001	0.001	0.000	$a_{3,2}$	0.04	0.02	0.04
					$a_{5,4}$	0.10	0.00	–
Phenotype and SNP	0.002	0.000	0.000	0.000	b_{24}	0.10	–	0.06
					b_{39}	0.12	–	–

TABLE IV

Associations detected in the gemcitabine study using the Bayesian integrative model or the all-APCA approaches. Results from different prior model specifications for p_a and p_b are presented for the Bayesian integrative model. Effects determined to be significant at the 0.05 level from the APCA (after adjustment for multiple testing) are included with bolded P-values. For the Bayesian integrative model, genomic variables are included if (1) the 95% credible interval for the direct mRNA expression effects did not contain zero; (2) the $P(\gamma_{jk} \mid \text{data}) \geq 0.25$ with the Beta prior for the indirect SNP effect; or (3) the $P(\phi_k \mid \text{data}) \geq 0.15$ with Beta prior for the direct SNP effect

Effect type	mRNA gene expression variation	Genetic variation	APCA *p*-value	*p* ~ *Beta* posterior probability or 95% CI	*p* ~ *Uniform* posterior probability or 95% CI	*p* = 0.2 posterior probability or 95% CI
SNP—mRNA association or indirect SNP effect	SLC28A1	CMPK1	**4.94e–4**	0.35	0.22	0.65
	SLC28A2	NT5C1B	5.77e–3	0.25	0.13	0.53
	CMPK1	CMPK1	**4.06e–4**	0.37	0.20	0.70
	CDA	RRM2	3.97e–3	0.28	0.17	0.51
	NT5C3L	NT5C3L	**1.42e–18**	1.00	1.00	1.00
mRNA expression-—IC50 association or direct mRNA effect	SLC29A1	—	**2.79e–3**	(−0.225, 0.066)	(−0.225, 0.065)	(−0.226, 0.067)
	DCK	—	**2.65e–3**	(−0.266, −0.004)	(−0.267, −0.003)	(−0.267, −0.003)
	NT5C3	—	**9.75e–6**	(0.035, 0.266)	(0.034, 0.266)	(0.035, 0.267)