Published in final edited form as:

Genet Epidemiol. 2010 July ; 34(5): 418–426. doi:10.1002/gepi.20494.

Bayesian Mixture Models for the Incorporation of Prior Knowledge to Inform Genetic Association Studies

Brooke L. Fridley^{§,1}, Daniel Serie¹, Gregory Jenkins¹, Kristin White¹, William Bamlet¹, John D. Potter², and Ellen L. Goode¹

¹Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, 55905, USA

²Division of Public Health Services, Fred Hutchinson Cancer Research Center, Seattle WA, 98109, USA

Abstract

In the last decade, numerous genome-wide linkage and association studies of complex diseases have been completed. The critical question remains of how to best use this potentially valuable information to improve study design and statistical analysis in current and future genetic association studies. With genetic effect size for complex diseases being relatively small, the use of all available information is essential to untangling the genetic architecture of complex diseases. One promising approach to incorporating prior knowledge from linkage scans, or other information, is to up- or down-weight p-values resulting from genetic association study in either a frequentist or Bayesian manner. As an alternative to these methods, we propose a fully Bayesian mixture model to incorporate previous knowledge into on-going association analysis. In this approach, both the data and previous information collectively inform the association analysis, in contrast to modifying the association results (p-values) to conform to the prior knowledge. By using a Bayesian framework, one has flexibility in modeling, and is able to comprehensively assess the impact of model specification on posterior inferences. We illustrate use of this method through a genome-wide linkage study of colorectal cancer, and a genome-wide association study of colorectal polyps.

Keywords

Bayesian; genetic association; linkage; mixture model; prior information

1. Introduction

Many genome-wide linkage[Amos, et al. 2006; John, et al. 2004; Lange, et al. 2006; Middleton, et al. 2004; Sellick, et al. 2005] and association[Consortium 2007; Easton, et al. 2007; Gudmundsson, et al. 2007; Scott, et al. 2007; Zanke, et al. 2007] studies of other complex diseases have been completed in the last decade. Often these datasets and/or results are available to outside investigators through collaborations or via dbGaP. A critical question is how to use this valuable information to inform current and future studies on study design and statistical analysis. One method, described by Roeder et al.[Roeder, et al. 2006] and Genovese et al.[Genovese, et al. 2006], uses results from linkage scans to up- or

[§]Corresponding Author: Brooke L Fridley, Ph.D. Harwick 766, Division of Biostatistics Mayo Clinic, 200 First Street SW Rochester, MN 55905 Tel: 507-538-3646 Fax: 507-284-9542 fridley.brooke@mayo.edu.

Web Resources

dbGap: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gap>

down-weight p-values in genome-wide association studies (GWAS), with various choices for the weighting function (e.g., cumulative or exponential weighting function). Modification of this approach of up- or down-weighting individual SNPs has also been investigated, where groups of SNPs are up-weighted or down-weighted together [Roeder, et al. 2007]. A method similar to the approach of Roeder et al. has recently been proposed by Eskin [Eskin 2008], which accounts for linkage disequilibrium (LD) between markers. In addition, Ionita-Laza et al. [Ionita-Laza, et al. 2007] propose a method for weighting results based on family data where the between family information is used in the first stage (screening), and the within family information is used in the second or testing stage.

Alternative approaches to the weighting of p-values or significance thresholds, is the use of hierarchical, mixed [McCulloch 1997; McCulloch and Searle 2001], or multi-level models [Snijders and Bosker 1999] to incorporate prior knowledge into the association analysis [Chen and Witte 2007; Conti and Witte 2003; Witte 1997]. Chen and Witte [Chen and Witte 2007] describe a mixed model framework for modeling M SNPs together where the SNP effects are modeled with both the mean and variance of the multivariate normal distribution depending on prior information. Bayesian analysis of case-control studies using power priors to incorporate historical knowledge was proposed by Cheng and Chen [Cheng and Chen 2005], while Lewinger et al. [Lewinger, et al. 2007] proposed a hierarchical Bayes method of weighting single SNP association results in a prior model that incorporates previous knowledge.

The flaw with many of the previously proposed approaches is that an incorrectly specified model is used for the association analysis, with analysis results modified to conform to prior knowledge. This modification of p-values, also changes the fundamental interpretation of a p-value, which is no longer defined as the probability of obtaining a result at least as extreme as the one that was actually observed, given that the null hypothesis is true. As an alternative to these methods, we propose a fully Bayesian mixture model to incorporate previous knowledge into on-going association analysis. In this approach, both the data and the previous information collectively inform the association analysis, as opposed to modifying the association results (p-values) to conform to the prior knowledge. In addition, the fully Bayesian analysis produces interpretive results (e.g., posterior probabilities, 95% credible intervals) and allows for comprehensive assessment of the impact of prior knowledge on statistical inferences, as well, the flexibility of model choice and form of the previous information (e.g., biological, association study, linkage study). The proposed modeling method differs from previously proposed methods in that: prior knowledge is incorporated at the start of the analysis as opposed to at the end of the analysis; SNP effects are modeled using a three-component mixture model with the specification of the strength of the previous information accomplished through the specification of the T parameter in the Dirichlet distribution; the method is applicable to multiple forms of prior knowledge and phenotypes; allows for sensitive analysis on the choice of prior distribution specification; and provides interpretable results in terms of probability statements for parameters of interest.

The proposed Bayesian modeling framework for the incorporation of prior knowledge is illustrated with a genome-wide association study of adenomatous colorectal polyps using previous knowledge from a recently-completed colorectal cancer genome-wide linkage scan. The proposed Bayesian mixture model for incorporating prior knowledge can be easily extended to include multiple forms of biological knowledge beyond information from previous linkage or association studies.

2. Materials AND Methods

2.1. Bayesian Mixture Model

Bayesian association analysis is similar to the frequentist association approach, except that it includes the addition of prior distributions on the parameters in the model. Inferences about the parameters of interest are based on the posterior probability distribution $p(\theta|Y) \propto p(Y|\theta)p(\theta)$ of the parameters, which is comprised of models of the data, $p(Y|\theta)$, and the parameters (prior distribution), $p(\theta)$ [Gelman, et al. 2000]. In fitting the Bayesian model, Markov chain Monte Carlo (MCMC) methods are used to approximate the posterior probability distributions [Bennett, et al. 1996; Gelfand and Smith 1990].

Bayesian mixture models can be thought of as hierarchical models with the inclusion of a latent variable. Bayesian mixture and latent variable models became widely used after the seminal papers by Tanner and Wong [Tanner and Wong 1987], and Gelfand and Smith [Gelfand and Smith 1990] which outlined the use of data augmentation for analysis of models with latent variables or missing data. Since then, data augmentation and analysis with latent variables has become a mainstay in Bayesian analysis where Gibbs sampling [Geman and Geman 1984] is often used to estimate parameters in the mixture model [Gilks, et al. 1996].

In the realm of genomic studies, mixture models have been used to classify sets of differentially expressed genes in the analysis of gene expression data [Do, et al. 2005; Efron, et al. 2001; Kendzioriski, et al. 2006; Lee, et al. 2000; Lewin, et al. 2007; Medvedovic, et al. 2004]. We propose a novel modeling framework for the incorporation of previously obtained information using a three-component mixture model.

2.2. Bayesian mixture model for incorporation of previous knowledge

Let Y_i represent the affection status of subject i , with Y_i equal to 1 if the subject has the disease (case) and 0 if not (control), with Y_i following a Bernoulli distribution with parameter p_i . The probability of being affected is then modeled as a function of the SNP genotype, $\log(p_i/(1-p_i)) = \beta_0 + \beta_1(SNP_i)$, where p_i is the probability that subject i has the disease and SNP_i is the genotype (0, 1, 2 representing the number of minor alleles) for subject i . The next level of the model is the specification of the mixture model for the SNP effect, β_1 . In specifying the mixture model, there are three components to represent a SNP effect, β_1 , that is either negative (e.g., minor allele is “protective”), “null” or no impact of the SNP alleles on disease risk (e.g., “neutral”), or positive (e.g., minor allele is “deleterious”). The mixture prior for β_1 has the form $p(\beta_1) = \sum_{k=1}^3 q_k p_k(\beta_1)$, where $p_1(\beta_1)$, $p_2(\beta_1)$, $p_3(\beta_1)$ represents the distribution of the SNP effect. Therefore, the posterior distribution of β_1 is a mixture of the respective posterior distributions under each prior assumption. In specifying the mixture distributions, we allow both the mean and variance to differ with respect to the mixing distributions, with specification of the weights (q_1 , q_2 and q_3 , with $\sum_{k=1}^3 q_k = 1$) based on information from previously completed linkage or association studies, or other biologic knowledge.

There are many choices for modeling the prior distribution, $p_k(\beta_1)$, $k = 1, 2, 3$, including choice of distributional form (i.e., normal, gamma), and specification of the hyper-priors for the parameters in the prior distribution. We have chosen to use a mixture of three normal distributions as follows. The prior distribution representing the case when the minor allele of the SNP is “protective” is specified as $p_1(\beta_1|\mu_1, \tau_1^2)\mathcal{N}(\mu_1, \tau_1^2)$ with the hyper-prior for the mean being a normal distribution $\mu_1 \sim \mathcal{N}(\psi_1, \nu_1^2)$ and the hyper-prior for the precision (inverse

of the variance) being a gamma distribution $1/\tau_1^2 \text{Gamma}(\delta_1, \pi_1)$. Values for the hyper-parameters are then set to reflect one's previous knowledge about the SNP effect. With many replicated SNP effects for complex diseases found to be relatively small, one possible model would be $\psi_1 = -0.69$, $v_1 = 0.10$, $\delta_1 = 20$ and $\pi_1 = 4$, resulting in a somewhat diffuse normal prior distribution for β_1 centered at -0.69 (odds ratio of 0.50). Similarly, the prior distribution representing the case when the SNP alleles have no effect on disease risk, is

specified as $p_2(\beta_1|\mu_2, \tau_2^2)\mathcal{N}(\mu_2, \tau_2^2)$ with $\mu_2 \sim N(\omega, v_2)$ and $1/\tau_2^2 \text{Gamma}(\delta_2, \pi_2)$. For the prior distribution for a “null” SNP, one may wish to specify a nugget prior, where the nugget prior is similar to a point mass prior, but with variance > 0 (i.e., $< \text{variance} < \varepsilon$, with $\varepsilon \rightarrow 0$), for “shrinkage” with $\psi_2 = 0$, $v_2 = 0.001$, $\delta_2 = 1000$ and $\pi_2 = 10$, resulting in a prior distribution for β_1 centered at zero (odds ratio of 1). Lastly, the prior distribution for the case of a positive SNP effect (“deleterious” effect of the minor allele) is specified as

$p_3(\beta_1|\mu_3, \tau_3^2)\mathcal{N}(\mu_3, \tau_3^2)$ with $\mu_3 \sim N(\omega_3, v_3)$ and $1/\tau_3^2 \text{Gamma}(\delta_3, \pi_3)$ specification of all hyper-parameters in the priors is completed in a similar manner as for the case when the SNP minor allele is “protective” (i.e., diffuse normal prior distribution for β_1 centered at $+0.69$ [odds ratio of 2.0]).

As with the choices for the prior distributions, there are many choices for the weighting function. We have chosen to use a Dirichlet prior for the weights q_k , $k = 1, 2, 3$, where the Dirichlet prior involves the previous linkage information. Let $\mathbf{q} = (q_1, q_2, q_3)^T \sim \text{DIR}(T * \mathbf{w})$, where $\mathbf{w} = (w_1, w_2, w_3)^T$ with $w_2 = 1/\exp\{-\log_{10}(p - \text{value})\}$ and $w_1 = w_3 = (1 - w_2)/2$, in which the p-value is determined from a previously completed nonparametric linkage analysis using the Kong and Cox LOD scores [Kong and Cox 1997] computed from the S_{all} statistic [Whittemore and Halpern 1994]. The strength of the Dirichlet prior is based on the specification of T . As $T \rightarrow 0$, the prior for the weights becomes non-informative, and as $T \rightarrow \infty$, the prior for the weights becomes informative. Thus, choice of T reflects how much strength the previous information represents.

Various mixture priors for SNP effects are presented in Figure 1. The figure depicts two different prior models for various values of $\mathbf{q} = (q_1, q_2, q_3)^T$, where parameters μ_k and τ_k^2 in the distributions $p_k(\beta_k|\mu_k, \tau_k^2)\mathcal{N}(\mu_k, \tau_k^2)$ are set equal to -0.69 , 0 , and 0.69 , respectively for μ_k , $k = 1, 2, 3$ in both models. For the first model in Figure 1, the values of τ_k^2 are set to 0.20 , 0.01 , 0.20 for $k = 1, 2$, and 3 , whereas, in the second model, the values of τ_k^2 are set to 0.40 , 0.10 , 0.40 for $k = 1, 2$, and 3 . The Bayesian mixture model for incorporation of prior knowledge is fit for each individual SNP marker using Markov Chain Monte Carlo (MCMC) methods [Gilks, et al. 1996; Smith and Roberts 1993] with statistical inference based on the posterior distribution for parameters of interest [Gelfand and Smith 1990].

2.3. Genetic Epidemiology Studies for Colorectal Cancer and Polyps

Approximately 77,000 individuals are estimated to be diagnosed with colorectal cancer in the U.S. annually, and more than 24,000 individuals will die of the disease [Jemal, et al. 2008]. There are several known genetic syndromes that lead to increased risk of colorectal cancer including Lynch Syndrome (MIM 120435 caused by mutations in mismatch repair genes such as *MSH2* and *MSH6* on chromosome 2), Peutz-Jeghers Syndrome (MIM 175200), *MUTYH*-associated polyposis (MIM 608456), familial adenomatous polyposis (MIM 175100), and juvenile polyposis (MIM 174900) [Abdel-Rahman, et al. 2006; Mecklin 2008]. Beyond these known syndromes, residual evidence for the existence of additional colorectal cancer loci [Jenkins, et al. 2002] has led to the initiation of single nucleotide polymorphism (SNP)-based genome-wide studies using linkage [Kemp, et al. 2006; Papaemmanuil, et al. 2008] or association [Houlston, et al. 2008; Zanke, et al. 2007]. One

reported colorectal cancer association locus resides on chromosome 8q24 and has been replicated in numerous populations[Berndt, et al. 2008; Schafmayer, et al. 2009; Tenesa, et al. 2008; Zanke, et al. 2007]. Below we describe two studies, one linkage study and genetic association study, to determine novel loci involved in colon cancer risk. These studies were used to illustrate the incorporation of linkage information into a genetic association study using the proposed Bayesian mixture modeling framework.

2.3.1. Linkage Study of Colorectal Cancer—Genome-wide linkage analysis was conducted using families collected by the Colon Cancer Family Registry (Colon CFR), an NCI-supported international consortium for colorectal cancer genetic epidemiology which collects probands and their family members, along with blood and extensive epidemiologic data[Newcomb, et al. 2007]. Population-based and clinic-based recruitment of families took place at six sites within the U.S., Canada, and Australia[Newcomb, et al. 2007] and at an additional non-CFR site in Newfoundland, Canada which used identical protocols. Analysis focused on families that were linkage-informative with at least two affected individuals for the phenotype of invasive colorectal cancer, and with no evidence of known genetic syndromes.

Individuals were genotyped using the Affymetrix GeneChip® Human Mapping 10K 2.0 Array, containing 10,204 single nucleotide polymorphisms (SNPs); 1,530 subjects in 287 white non-Hispanic families had a SNP call rate > 95%. Extensive quality-control filters were applied to detect and remove Mendelian errors within families and poorly performing SNPs. SNPs were also removed with minor allele frequency (MAF) less than 1% to avoid possible bias due to rare alleles. Because multipoint linkage analysis assumes linkage equilibrium, we used ldSelect[Carlson, et al. 2004] to reduce linkage disequilibrium (LD) to r^2 less than 0.10, retaining the more common SNP where multiple tagSNPs within an LD bin were observed. Allele frequencies for all SNP genotypes were estimated across the pool of all subjects, ignoring the genetic relationships within families, and the genetic map was created by Affymetrix using the deCODE genetic map[Kong, et al. 2002]. Nonparametric multipoint linkage analyses were implemented in MERLIN, version 1.0.1[Abecasis, et al. 2002] to compute Kong & Cox LOD scores[Kong and Cox 1997] from the linear model based on the S_{all} statistics[Whittemore and Halpern 1994].

2.3.2. Association Study of Colorectal Polyps—A pilot-scale GWAS was conducted using cases and controls from the well-established Minnesota Cancer Prevention Research Unit Polyp Study[Goode, et al. 2007; Goode, et al. 2004; Potter, et al. 1996; Ulrich, et al. 1999]. Cases with colorectal adenomatous and polyp-free control subjects were recruited through a large multi-clinic private gastroenterology practice in metropolitan Minneapolis. Individuals aged 30 to 74 years, who were scheduled for a colonoscopy between April 1991 and April 1994 were recruited prior to colonoscopy so that patients and recruiters were blinded to the final diagnosis. The pilot-scale GWAS study primarily sought to assess genotyping feasibility; however, even though power was minimal, it was designed to maximize the likelihood of genetic differences between cases and controls while matching on potential confounding factors. Cases consisted of 20 individuals with first-degree family history who were found to have adenomatous polyps; 20 controls had no family history, were found to be free of polyps, and were individually-matched to the cases on gender (65% female) and age (range 47 to 73 years). All individuals were non-Hispanic white.

Individuals were genotyped at 116,204 SNPs (median inter-marker distance of 8.5 kb) using the Affymetrix GeneChip® Human Mapping 100K Set. Only chromosomes 2 and 8 were analyzed here for illustration of the proposed Bayesian model for incorporation of prior knowledge. These chromosomes were selected because chromosome 2 encompasses genes known to be associated with colorectal cancer (*MSH2*, *MSH6*, and *PSM1*), and chromosome

8 has a recently identified region (8q24) covering SNPs associated with breast[Easton, et al. 2007], prostate[Gudmundsson, et al. 2007] and colorectal cancers[Zanke, et al. 2007]. Due to low power, SNPs with MAF less than 10% or call rates less than 90% were removed prior to analysis, leaving 5,901 SNPs on chromosome 2 and 3,820 SNPs on chromosome 8 for analysis.

3. Results from the Colorectal Cancer and Polyps Studies

In fitting the Bayesian mixture model to assess the association of genome-wide SNPs with occurrence of colorectal polyps in a manner that included the previous linkage information for colorectal cancer, p-values from the linkage analysis 10K array needed to be mapped to the association analysis 100K data to determine w . Thus, the following pre-processing of the data was completed: SNPs on both panels were mapped to genome build 36, and SNPs with unknown physical position or without rsids were removed from further analyses.

Nonparametric linkage results were linearly interpolated to assign a Kong and Cox LOD score and p-value for each of the SNPs on the 100K panel. The few SNPs on the Affymetrix 100K panel that did not fall between two SNPs on the Affymetrix 10K linkage panel were removed because of the inability to assign extrapolated linkage results.

For the association analysis of colorectal polyps with the inclusion of previous information from the linkage study, the following hyper-priors were used (model 1): $\mu_1 \sim N(-0.69, 0.1)$ and $1/\tau_1^2 \text{Gamma}(20, 4)$; $\mu_2 \sim N(0, 0.001)$ and $1/\tau_2^2 \text{Gamma}(1000, 10)$; and $\mu_3 \sim N(0.69, 0.1)$ and $1/\tau_3^2 \text{Gamma}(20, 4)$. In fitting the Bayesian mixture model, the parameter T for the Dirichlet distribution was set to $T=100$ to reflect the large sample size for the linkage study. Sensitivity analysis was conducted for the specification of T , resulting in similar SNP association outcomes based on models with $T = 100$ and $T = 10$.

In addition to running the Bayesian mixture model outlined in Section 2.1 for incorporation of previous information into the association analysis, the same colorectal polyp study was analyzed using a Bayesian non-mixture model with non-informative priors for all parameters in the model (i.e., $\beta_1 \sim N(0, 1000)$). All analyses were completed in WinBUGS[Spiegelhalter, et al. 2004] using the R package R2WinBUGS[Sturtz, et al. 2005], in which three chains were run for 10,000 iterations, removing the first 1,000 for burn-in. Convergence was checked using $\sqrt{R}$ [Gelman, et al. 1995], with $\sqrt{R} < 1.10$ for all parameters in the model. The computing time for the Bayesian models were moderate when analyzed on a Dell 1950, dual CPU (dual core) machine, with approximately four markers analyzed in under a minute. The computing time reduced dramatically, with approximately 100 markers analyzed in a minute, when the analyses were completed on the Mayo Research Computing Facility, which runs a Beowulf-style Linux cluster: a set of Linux-based systems which can work together to complete complex, processor-intensive tasks with over 400 CPUs.

The posterior probability (PP) was defined as the maximum proportion of the simulated posterior distribution greater than 0, and the proportion of the simulated posterior distribution less than 0. With this definition, posterior probabilities (PP) ranged from 0.50 to 1.0, with 1.0 representing the situation in which the posterior distribution does not contain 0, and 0.50 representing the situation in which the posterior distribution is centered at 0. Figures 2 and 3 display the $-\log_{10}(1-PP)$ for Bayesian analyses with and without incorporation of previous information from the colorectal cancer linkage study for chromosomes 2 and 8. Point estimates for SNP effects (median of posterior distribution) for chromosomes 2 and 8 are also displayed in Figure 4, with the summary of point estimates for SNP effects on chromosomes 2 and 8 presented in Table 1. The SNP effects, for both chromosomes 2 and 8, are shrunk towards zero when the mixture model is used. This is due

to the nugget distribution being centered at 0 for the case in which the SNP is neutral or null. Figures 2 and 3 show that, in regions with larger $-\log_{10}(\text{linkage p-values})$, the shrinkage of the SNP effects towards zero was not as large as in regions with no evidence of linkage (i.e., small $-\log_{10}(\text{linkage p-values})$), as expected.

3.1. Prior Sensitivity Analyses

Incorporation of previous knowledge into an association study allows informative priors to be specified. However, the degree to which the information is incorporated can vary from very strong priors, to those that are more moderate or diffuse. One way this can be varied is by modifying the values of the parameters that are used in the hyper-priors. To illustrate the difference in strength of previous knowledge, we have fitted three additional models for a region of chromosome 2, in which a moderate LOD score was observed in the region of genes *MSH2* and *MSH6*. The second model (model 2) involves the hyper-prior specification:

$\mu_1 \sim N(-0.69, 0.5)$, $1/\tau_1^2 \text{Gamma}(20, 4)$, $\mu_2 \sim N(0, 0.1)$, $1/\tau_2^2 \text{Gamma}(100, 10)$, $\mu_3 \sim N(0.69, 0.5)$, and $1/\tau_3^2 \text{Gamma}(20, 4)$, whereas model 3 involves the hyper-prior specification: $\mu_1 \sim N(-0.69, 0.5)$, $1/\tau_1^2 \text{Gamma}(10, 4)$, $\mu_2 \sim N(0, 0.1)$, $1/\tau_2^2 \text{Gamma}(100, 10)$, $\mu_3 \sim N(0.69, 0.5)$, and $1/\tau_3^2 \text{Gamma}(10, 4)$. The last model (model 4) is similar to model 1, the original model used in the analysis of chromosome 2, with normal distributions centered at -1.0 and 1.0 , in contrast to -0.69 and 0.69 .

Figure 5 displays the posterior probabilities from a non-mixture model (black), a mixture (model 1) (red), and a more diffuse mixture model (model 2) (blue) for the region of chromosome 2 that contains the linkage signal for *MSH2* and *MSH6*. As Figure 5 illustrates, in regions where there is no evidence of linkage, we observe more shrinkage of posterior probabilities under the informative mixture-model prior (model 1) and less shrinkage under the more diffuse mixture prior (model 2). Table 2 shows the summary of the posterior probabilities and parameter estimates for the five different models, including the four different hyper-priors for the mixture model (for the region on chromosome 2). Specifying a more diffuse prior model (model 2 and model 3) resulted in less shrinkage of the parameter effects and posterior probabilities, as compared to the informative mixture model (model 1). The specification of the distribution for the “deleterious” and “protective” situations in model 4 centered at 1.0 or -1.0 (odds ratio 2.72 and 0.368, respectively) as compared to -0.69 or 0.69 in model 1 (odds ratio of 2 and 0.50, respectively) resulting in little difference in posterior probabilities or parameter estimates for the SNP effects.

4. Discussion

We have described a fully Bayesian mixture model incorporating previous knowledge into genetic association studies. In doing so, we have outlined the use of the model for incorporation of previous linkage information into genetic association studies. With genetic effect size for complex diseases being relatively small, the use of all available information is crucial to untangling the genetic architecture of complex disease.

As with any analysis, care is needed to determine whether to include information from previous studies, especially association studies, in that both study populations need to be from similar ethnic backgrounds and have similar phenotype definitions. If the studies are not comparable in these ways, the applicability of the previous knowledge will be compromised. Likewise, if the study providing the previous knowledge is under powered, this knowledge may be weak, non-informative, or incorrect and may dilute or diminish a “true” signal. Specification of the appropriate strength of the previous information can be accomplished through the specification of the T parameter in the Dirichlet distribution. If the

previous study is well powered, a larger value of T can be set, whereas for a small previous study, the value of T would be correspondingly smaller.

This Bayesian model can be extended to incorporate previous knowledge from genetic association studies (from collaborations or dbGaP) or other biologic information (e.g., known functional polymorphism) to inform current association studies. When results from previous genetic association studies are available, a prior distribution based on the effect size, as opposed to the p-value, is recommended. The method is also flexible, in that a quantitative trait or phenotype could be modeled with a Gaussian distribution (as opposed to the Bernoulli distribution for a binary trait). In addition to flexibility regarding phenotype, the method can be varied by using different mixture distributions; for example, using gamma rather than normal distributions[Lewin, et al. 2007] or modification of hyper-priors.

The association study that we utilized to demonstrate the mixture model is limited in its power since only 40 subjects were included. Well-powered, case-control genome-wide association studies of colorectal cancer within the Colon CFR and other study populations are underway. Analysis of these GWAS data, incorporating information from an appropriately-powered linkage study is of high relevance, and will serve as a critical next step.

To incorporate previous knowledge into an analysis, a Bayesian approach is the natural choice. In this approach, both the model for the data (i.e., likelihood) and the model for the previous information (i.e., prior distribution) collectively determine the posterior distribution for which statistical inferences are determined. It is clear that an informative prior will influence the inferences made from the subsequent posterior distribution. However, in a Bayesian framework there is also the ability to comprehensively assess the impact of prior distribution specification on the posterior inferences (i.e., sensitivity analysis). In essence, a Bayesian analysis incorporating previous knowledge can be thought of as a “pooled” analysis in which the prior distribution is acting like the “data” from a previous study.

In conclusion, understanding the complex relationship between genetic variation and complex disease is at the heart of “personalized medicine”. In order to increase our knowledge about the etiology of complex diseases, scientific investigation needs to be sequential, with knowledge gained from each step of the discovery process carried forward into the subsequent phases of the study. By including the wealth of knowledge that already exists for many complex diseases, we may increase our chances of unraveling the complex relationship between the human genome, environment and complex disease. The proposed Bayesian model is one tool that is available to researchers to aid in reaching the goal of “personalized medicine”.

Acknowledgments

The research was supported by the NIH National Cancer Institute (NCI) grants R01 CA104667 “Genetic Linkage in Colorectal Cancer Families” and R21 CA140879 “Integrative Genomic Models for Pharmacogenomic Studies”, a Minnesota Partnership for Biotechnology and Medical Genomics grant H9046000431 and a strategic grant from the Mayo Clinic Cancer Center’s Genetic Epidemiology Risk Assessment Program. In addition, this work was supported by the National Cancer Institute, National Institutes of Health under RFA # CA-95-011 and RFA-CA-08-502 and through the following cooperative agreements with members of the Colon Cancer Family Registry (CFR) and P.I.s: U01 CA097735 and U24 CA097735 “Australasian Colorectal Cancer Family Registry”, U01 CA074799 and U24 CA074799 “Familial Colorectal Neoplasia Collaborative Group”, U01 CA074800 and U24 CA074800 “Mayo Clinic Cooperative Family Registry for Colon Cancer Studies”, U01 CA074783 and U24 CA074783 “Ontario Registry for Studies of Familial Colorectal Cancer”, U01 CA074794 and U24 CA074794 “Seattle Colorectal Cancer Family Registry”, U01 CA074806 and U24 CA074806 “University of Hawaii Colorectal Cancer Family Registry”, and U01 CA078296 “University of California, Irvine Informatics Center”. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any

of the collaborating centers in the CFRs, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CFR.

References

- Abdel-Rahman WM, Mecklin JP, Peltomaki P. The genetics of HNPCC: Application to diagnosis and screening. *Critical Reviews in Oncology/Hematology*. 2006; 58:208–220. [PubMed: 16434208]
- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*. 2002; 30(1):97–101. [PubMed: 11731797]
- Amos CI, Chen WV, Lee A, Li W, Kern M, Lundsten R, Batliwalla F, Wener M, Remmers E, Kastner DA. High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. *Genes Immun*. 2006; 7(4):277–86. and others. [PubMed: 16691188]
- Bennett, JE.; Racine-Poon, A.; Wakefield, JC. MCMC for nonlinear hierarchical models. In: Gilks, WR.; Richardson, S.; Spiegelhalter, DJ., editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall; London: 1996. p. 339–357.
- Berndt SI, Potter JD, Hazra A, Yeager M, Thomas G, Makar KW, Welch R, Cross AJ, Huang WY, Schoen RE. Pooled analysis of genetic variation at chromosome 8q24 and colorectal neoplasia risk. *Hum Mol Genet*. 2008; 17(17):2665–72. and others. [PubMed: 18535017]
- Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*. 2004; 74(1):106–20. [PubMed: 14681826]
- Chen GK, Witte JS. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet*. 2007; 81(2):397–404. [PubMed: 17668389]
- Cheng KF, Chen JH. Bayesian models for population-based case-control studies when the population is in Hardy-Weinberg equilibrium. *Genet Epidemiol*. 2005; 28(2):183–92. [PubMed: 15593279]
- Consortium TWTC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447:661–678. [PubMed: 17554300]
- Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am J Hum Genet*. 2003; 72(2):351–63. [PubMed: 12525994]
- Do K, Muller P, Tang F. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Series C*. 2005; 53(3):627–644.
- Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struwing JP, Morrison J, Field H, Luben R. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. 2007; 447(7148):1087–93. and others. [PubMed: 17529967]
- Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*. 2001; 96(456):1151–1160.
- Eskin E. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Res*. 2008; 18(4):653–60. [PubMed: 18353808]
- Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*. 1990; 85(410):398–409.
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. Chapman & Hall; London: 1995.
- Gelman, AB.; Carlin, JS.; Stern, Hs; Rubin, DB. *Bayesian Data Analysis*. Chatfield, C.; Zidek, JV., editors. Chapman & Hall/CRC; Boca Raton: 2000.
- Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*. 1984; 6(6):721–741.
- Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. *Biometrika*. 2006; 93:509–524.
- Gilks, WR.; Richardson, S.; Spiegelhalter, DJ. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC; Boca Raton, FL: 1996.
- Goode EL, Potter JD, Bamlet WR, Rider DN, Bigler J. Inherited variation in carcinogen-metabolizing enzymes and risk of colorectal polyps. *Carcinogenesis*. 2007; 28(2):328–41. [PubMed: 16926176]

- Goode EL, Potter JD, Bigler J, Ulrich CM. Methionine synthase D919G polymorphism, folate metabolism, and colorectal adenoma risk. *Cancer Epidemiol Biomarkers Prev.* 2004; 13(1):157–62. [PubMed: 14744749]
- Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet.* 2007; 39(5):631–7. and others. [PubMed: 17401366]
- Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, Chandler I, Vijayakrishnan J, Sullivan K, Penegar S. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet.* 2008; 40(12):1426–35. and others. [PubMed: 19011631]
- Ionita-Laza I, McQueen MB, Laird NM, Lange C. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am J Hum Genet.* 2007; 81(3):607–14. [PubMed: 17701906]
- Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ. Cancer statistics, 2008. *CA Cancer J Clin.* 2008; 58(2):71–96. [PubMed: 18287387]
- Jenkins MA, Baglietto L, Dite GS, Jolley DJ, Southey MC, Whitty J, Mead LJ, John DJ, Macrae FA, Bishop DT. After hMSH2 and hMLH1--what next? Analysis of three-generational, population-based, early-onset colorectal cancer families. *Int J Cancer.* 2002; 102(2):166–71. and others. [PubMed: 12385013]
- John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W, Vasavda N, Mills T, Barton A, Hinks A. Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am J Hum Genet.* 2004; 75(1):54–64. and others. [PubMed: 15154113]
- Kemp ZE, Carvajal-Carmona LG, Barclay E, Gorman M, Martin L, Wood W, Rowan A, Donohue C, Spain S, Jaeger E. Evidence of linkage to chromosome 9q22.33 in colorectal cancer kindreds from the United Kingdom. *Cancer Res.* 2006; 66(10):5003–6. and others. [PubMed: 16707420]
- Kendzierski CM, Chen M, Yuan M, Lan H, Attie AD. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics.* 2006; 62(1):19–27. [PubMed: 16542225]
- Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet.* 1997; 61(5):1179–88. [PubMed: 9345087]
- Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G. A high-resolution recombination map of the human genome. *Nat Genet.* 2002; 31(3):241–7. and others. [PubMed: 12053178]
- Lange EM, Ho LA, Beebe-Dimmer JL, Wang Y, Gillanders EM, Trent JM, Lange LA, Wood DP, Cooney KA. Genome-wide linkage scan for prostate cancer susceptibility genes in men with aggressive disease: significant evidence for linkage at chromosome 15q12. *Hum Genet.* 2006; 119(4):400–7. [PubMed: 16508751]
- Lee M, Kuo F, Whitmore G, Sklar J. Importance of Replication in Microarray Gene Expression Studies: Statistical Methods and Evidence from a Single cDNA Array Experiment. *Proc Natl Acad Sci U S A.* 2000; 96:9834–9839. [PubMed: 10963655]
- Lewin A, Bochkina N, Richardson S. Fully Bayesian mixture model for differential gene expression: simulations and model checks. *Stat Appl Genet Mol Biol.* 2007; 6 Article36.
- Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol.* 2007; 31(8):871–82. [PubMed: 17654612]
- McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association.* 1997; 92(437):162–170.
- McCulloch, CE.; Searle, SR. *Generalized, Linear, and Mixed Models.* John Wiley & Sons, Inc.; New York, NY: 2001.
- Mecklin JP. The implications of genetics in colorectal cancer. *Annals of Oncology.* 2008; 19:v87–v90. [PubMed: 18611908]
- Medvedovic M, Yeung KY, Bumgarner RE. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics.* 2004; 20(8):1222–32. [PubMed: 14871871]

- Middleton FA, Pato MT, Gentile KL, Morley CP, Zhao X, Eisener AF, Brown A, Petryshen TL, Kirby AN, Medeiros H. Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Am J Hum Genet.* 2004; 74(5): 886–97. and others. [PubMed: 15060841]
- Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R, Hall D, Hopper JL, Jass J, Le Marchand L. Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev.* 2007; 16(11):2331–43. and others. [PubMed: 17982118]
- Papaemmanuil E, Carvajal-Carmona L, Sellick GS, Kemp Z, Webb E, Spain S, Sullivan K, Barclay E, Lubbe S, Jaeger E. Deciphering the genetics of hereditary non-syndromic colorectal cancer. *Eur J Hum Genet.* 2008; 16(12):1477–86. and others. [PubMed: 18628789]
- Potter JD, Bostick RM, Grandits GA, Fosdick L, Elmer P, Wood J, Grambsch P, Louis TA. Hormone replacement therapy is associated with lower risk of adenomatous polyps of the large bowel: the Minnesota Cancer Prevention Research Unit Case-Control Study. *Cancer Epidemiol Biomarkers Prev.* 1996; 5(10):779–84. [PubMed: 8896888]
- Roeder K, Bacanu SA, Wasserman L, Devlin B. Using Linkage Genome Scans to Improve Power of Association in Genome Scans. *American Journal of Human Genetics.* 2006; 78:243–252. [PubMed: 16400608]
- Roeder K, Devlin B, Wasserman L. Improving power in genome-wide association studies: weights tip the scale. *Genet Epidemiol.* 2007; 31(7):741–7. [PubMed: 17549760]
- Schafmayer C, Buch S, Volzke H, von Schonfels W, Egberts JH, Schniewind B, Brosch M, Ruether A, Franke A, Mathiak M. Investigation of the colorectal cancer susceptibility region on chromosome 8q24.21 in a large German case-control sample. *Int J Cancer.* 2009; 124(1):75–80. and others. [PubMed: 18839428]
- Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science.* 2007; 316(5829):1341–5. and others. [PubMed: 17463248]
- Sellick GS, Webb EL, Allinson R, Matutes E, Dyer MJ, Jonsson V, Langerak AW, Mauro FR, Fuller S, Wiley J. A high-density SNP genomewide linkage scan for chronic lymphocytic leukemia-susceptibility loci. *Am J Hum Genet.* 2005; 77(3):420–9. and others. [PubMed: 16080117]
- Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B.* 1993; 55:3–23.
- Snijders, TAB.; Bosker, RJ. *Multilevel Analysis.* Sage Publications Ltd; London, UK: 1999.
- Spiegelhalter, D.; Thomas, A.; Best, N.; Lunn, D. *WinBUGS Version 2.0 User Manual:* MRC biostatistics Unit. Cambridge: 2004.
- Sturtz S, Ligges U, Gelman A. R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software.* 2005; 12(3):1–16.
- Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association.* 1987; 82(398):528–640.
- Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, Barnettson RA, Theodoratou E, Cetnarskyj R, Cartwright N. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet.* 2008; 40(5):631–7. and others. [PubMed: 18372901]
- Ulrich CM, Kampman E, Bigler J, Schwartz SM, Chen C, Bostick R, Fosdick L, Beresford SA, Yasui Y, Potter JD. Colorectal adenomas and the C677T MTHFR polymorphism: evidence for gene-environment interaction? *Cancer Epidemiol Biomarkers Prev.* 1999; 8(8):659–68. [PubMed: 10744125]
- Whittemore AS, Halpern J. A class of tests for linkage using affected pedigree members. *Biometrics.* 1994; 50(1):118–27. [PubMed: 8086596]
- Witte JS. Genetic analysis with hierarchical models. *Genet Epidemiol.* 1997; 14(6):1137–42. [PubMed: 9433637]

Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature Genetics*. 2007; 39(8):989–994. and others. [PubMed: 17618283]

Figure 1.

Example of two different mixture priors. The parameters μ_k and τ_k^2 in the distributions $p_k(\beta_k|\mu_k, \tau_k^2)\mathcal{N}(\mu_k, \tau_k^2)$ set equal to $-0.69, 0$, and 0.69 , respectively for $\mu_k, k = 1, 2, 3$ for both models. Figures (A), (B) and (C) display the prior distribution where τ_k^2 set to $0.20, 0.01, 0.20$ for $k = 1, 2$, and 3 . Figures (D), (E) and (F) display the prior distribution where τ_k^2 set to $0.40, 0.10, 0.40$ for $k = 1, 2$, and 3 . Priors displayed in Figures (A) and (D) had mixing proportion of $\mathbf{q} = (0.05, 0.90, 0.05)^T$ (probability SNP is neutral is 0.90); Figures (B) and (E) had mixing proportions $\mathbf{q} = (0.20, 0.60, 0.20)^T$ (probability that SNP is neutral is 0.60); Figures (C) and (F) had mixing proportions $\mathbf{q} = (0.45, 0.10, 0.45)^T$ (probability SNP is neutral is 0.10).

Figure 2.

(Bottom) Plot of the $-\log_{10}(1-PP)$ for Bayesian analyses with (red) and without (black) incorporation of linkage analysis results in prior distributions for chromosome 2. Lines from a kernel smoother are also displayed. PP is defined as the maximum proportion of the simulated posterior distribution greater than 0, and the proportion of the simulated posterior distribution less than 0. (Top) Plot of the $-\log_{10}(P\text{-value})$ from the nonparametric linkage analysis are displayed.

Figure 3.
 (Bottom) Plot of the $-\log_{10}(1-PP)$ for Bayesian analyses with (red) and without (black) incorporation of linkage analysis results in prior distributions for chromosome 8. Lines from a kernel smoother are also displayed. PP is defined as the maximum proportion of the simulated posterior distribution greater than 0, and the proportion of the simulated posterior distribution less than 0. (Top) Plot of the $-\log_{10}(P\text{-value})$ from the nonparametric linkage analysis are displayed.

Figure 4. Plot of point estimates for SNP effects (median of posterior distribution) from Bayesian analyses with (red) and without (black) incorporation of linkage analysis results in prior distributions for (A) chromosome 2 and (B) chromosome 8.

Figure 5.

(Bottom) Plot of $-\log_{10}(1-PP)$ from the model without incorporation of linkage results (black), the model with the “strong” incorporation of the linkage results (model 1) (red), and the model with the “weaker” incorporation of the linkage results (model 2) (blue) for region of chromosome 2 that contains linkage signal for *MSH2* and *MSH6*. Lines from a kernel smoother are also displayed. PP is defined as the maximum proportion of the simulated posterior distribution greater than 0, and the proportion of the simulated posterior distribution less than 0. (Top) Plot of the $-\log_{10}(P\text{-value})$ from the nonparametric linkage analysis for region on chromosome 2.

Table 1
Summary of Point Estimates of SNP Associations on chromosomes 2 and 8.

Chromosome	Prior Knowledge	Min	Q1	Median	Mean	Q3	Max
2	No	-3.12	-0.39	0.00	0.02	0.43	8.32
2	Yes	-1.41	-0.04	0.00	0.01	0.04	1.32
8	No	-8.42	-0.41	0.00	0.01	0.42	3.65
8	Yes	-1.21	-0.03	0.00	0.00	0.03	1.34

Table 2
Summary of Point Estimates and Posterior Probabilities of SNP Associations in the Chromosome 2 *MSH2* and *MSH6* Linkage Region.

Point Estimates						
Mixture Model	Min	Q1	Median	Mean	Q3	Max
None	-2.556	-0.324	0.0618	0.087	0.520	8.321
1	-1.084	-0.054	0.008	0.038	0.117	1.301
2	-1.257	-0.188	0.032	0.048	0.292	1.715
3	-1.318	-0.184	0.035	0.0495	0.286	1.867
4	-1.295	-0.051	0.007	0.051	0.117	1.542

Posterior Probabilities						
Mixture Model	Min	Q1	Median	Mean	Q3	Max
None	0.500	0.652	0.774	0.775	0.919	1.00
1	0.500	0.576	0.650	0.692	0.814	0.994
2	0.500	0.617	0.715	0.738	0.867	0.995
3	0.500	0.615	0.717	0.737	0.866	0.995
4	0.500	0.577	0.659	0.701	0.838	0.996