# Bayesian Selection and Clustering of Polymorphisms in Functionally-Related Genes

**David B. Dunson[1,*], Amy H. Herring[2], and Stephanie A. Mulherin Engel[3]**

[1]*Biostatistics Branch*

*MD A3-03, National Institute of Environmental Health Sciences*

*P.O. Box 12233, Research Triangle Park, NC 27709*

[2]*Department of Biostatistics*

*University of North Carolina at Chapel Hill*

[3]*Department of Community & Preventive Medicine*

*Mount Sinai Medical School*

[*]*E-mail: dunson1@niehs.nih.gov*

This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. The PIN Study was supported by a grant from the National Institute of Child Health and Human Development, National Institutes of Health (HD28684); cooperative agreements from the Association of Schools of Public Health/Centers for Disease Control and Prevention (S455-16/17, S0807-18/20, S1099-19/21); and funds from the Wake Area Health Education Center in Raleigh, NC; and supported in part by a grant (RR00046) from the General Clinical Research Centers Program of the Division of Research Resources, National Institutes of Health. The second author's work was supported in part by grants from the National Institutes of Health (National Institute of Child Health and Human Development 1R03HD045780-01A1, National Institute of Environmental Health Sciences P30ES10126) and the Environmental Protection Agency (RD831843010).

**Summary**. In epidemiologic studies, there is often interest in assessing the relationship between polymorphisms in functionally-related genes and a health outcome. For each candidate gene, single nucleotide polymorphism (SNP) data are collected at a number of locations, resulting in a large number of possible genotypes. Because instabilities can result in analyses that include all the SNPs, dimensionality is typically reduced by conducting single SNP analyses or attempting to identify haplotypes. This article proposes an alternative Bayesian approach for reducing dimensionality. A multi-level Dirichlet process prior is used for the distribution of the SNP-specific regression coefficients within genes, incorporating a variable selection-type mixture structure in the base measure to allow SNPs with no effect. This structure allows simultaneous selection of important SNPs and clustering of SNPs having similar impact on the health outcome. The methods are illustrated using data from a study of pro- and anti-inflammatory cytokine polymorphisms and spontaneous preterm birth.

*Keywords:* Bayesian; Clustering; Dirichlet process; Hierarchical regression; Multiple Testing; Nonparametric Bayes; Shrinkage prior.

## 1. Introduction

In epidemiologic research, there is commonly interest in the association between multiple, closely-related *exposures* and a health outcome. Some examples include drinking water disinfection by-products, agricultural chemicals, and single nucleotide polymorphisms (SNPs) in candidate genes. When the number of exposures is large (e.g., 30+), and the exposures are correlated (e.g., due to linkage disequilibrium between polymorphisms), it is well known that maximum likelihood estimation can result in unstable estimates and inferences. For this reason, analysts typically apply dimensionality reduction techniques, with the most common being (1) consider exposures one at a time in univariate analyses; (2) collapse exposures into class-specific summaries; and (3) run a model selection procedure, such as stepwise selection, to obtain a parsimonious model upon which to base final inferences. There are clear problems with each of these approaches: (1) can produce misleading results by not adjusting for correlated exposures; (2) can discard valuable information on variability in the effect within a class; and (3) can result in overestimation of the regression coefficients due to selection bias.

For these reasons, many authors have proposed hierarchical regression procedures, which shrink the exposure-specific regression coefficients towards a common distribution, using empirical Bayes (Thomas et al., 1985), "semi-Bayes" (Greenland, 1992; 1993) or fully Bayes approaches. Greenland (1993) provides a review and demonstrates improved performance of the empirical and semi-Bayes approaches relative to MLE-based methods. Such hierarchical regression procedures have been used in numerous articles in the epidemiologic literature. For example, De Roos et al. (2001) considered applications to multiple paternal occupational exposures and neuroblastoma in the offspring, and Hung et al. (2004) considered applications to genetic associations studies with multiple markers. For related methods for multiple outcomes, refer to Meng and Dempster (1987) and Coull et al. (2001).

These methods are based on shrinking the exposure-specific regression coefficients to-

wards a normal prior distribution, potentially with unknown mean and variance. Although this shrinkage certainly improves the stability of estimates, many epidemiologists would prefer to avoid the assumption that the regression coefficients for the different exposures follow a normal distribution. In addition, there is typically interest in grouping or clustering the different exposures based on their effects on the outcome. In particular, one wishes to identify exposures having similar effects, including those that are not associated with the outcome, in drawing mechanistic conclusions. Although grouping can be done subjectively based on examination of estimated regression coefficients, it would be appealing to have a formal clustering procedure.

This problem is somewhat related to subset selection in regression, which focuses on identifying predictors with non-zero coefficients from among a potentially high dimensional set of candidates (refer to George and McCulloch, 1997 and Clyde and George, 2004 for reviews of Bayesian approaches). However, following standard epidemiologic practice, we are at least as interested in estimating regression coefficients for the different exposures, and in grouping exposures according to the magnitude of their effect, as we are in identifying exposures that are associated with the response. Hence, the problem is one of clustering the regression coefficients, incorporating information on the exposure class.

The Bayesian approach provides a natural framework for clustering of the exposures in this manner. For a recent article on Bayesian variable selection and clustering in high-dimensional data, refer to Tadesse, Sha and Vannucci (2005). Their focus was on clustering samples of data into groups while simultaneously selecting discriminating variables. In contrast, our focus is on clustering not the data but the unknown exposure effects into groups, while allowing an unknown subset to have no association with the outcome. A related problem was considered by Gopalan and Berry (1998), who used a Dirichlet process (DP) prior (Ferguson, 1973; 1974) to cluster treatment groups in a clinical trial in order to adjust for multiple comparisons. From a Bayesian perspective, the multiple comparison problem can

be considered as an issue of appropriately choosing a prior to account for dependency in multiple, related hypotheses (refer to Westfall, Johnson and Utts, 1997; Berry and Hochberg, 1999; Gonen, Westfall and Johnson, 2003; Berry and Berry, 2004; Dunson, 2005)

Although the Gopalan and Berry (1998) approach could be directly modified to allow clustering of the regression coefficients for the different exposures, such an approach would not incorporate information on exposure class or allow identification of exposures having no effect. To perform simultaneous variable selection and clustering, both within and across exposure classes, the article proposes an alternative approach. In the one class case, the approach relies on DP clustering, with the base measure chosen to have a mixture structure, allowing the incorporation of a null cluster containing exposures with no effect. In the multiple class case, a multi-level DP structure is chosen, allowing common clusters across exposure classes, while also introducing class-specific clusters.

Section 2 motivates the problem through application to the problem of clustering of polymorphisms in functionally-related genes. Section 3 describes the regression model and proposes the hierarchical clustering prior. Section 4 develops methods for posterior computation. Section 5 applies the method to data on cytokine polymorphisms and risk for spontaneous preterm birth, while also presenting results from simulation studies. Section 6 contains a discussion.

## 2. Identifying Polymorphisms Predicting Disease

This article is motivated by the problem of selection and clustering of polymorphisms in functionally-related genes. Using the nomenclature of Section 1, genes correspond to *classes* and *exposures* to single nucleotide polymorphisms (SNPs). For a given gene, SNPs can be collected within the coding region, which consists of the sequence of amino acids that codes directly for the protein product of the gene, within regulatory regions upstream of the coding region, or within intronic sequences. SNPs that occur within regulatory regions are thought

to be much more likely to affect biologic function and gene expression.

For subject $i$ ($i = 1, \ldots, n$) and gene $c$ ($c = 1, \ldots, C$), the SNP data consist of a vector $\mathbf{x}_{ic} = (x_{ic1}, \ldots, x_{ic,p_c})$ of 0/1 indicators. The most commonly observed genotype at a given locus is coded as a 0, while the less common variant is coded as a 1. In cases with more than two variants, additional indicators are introduced, so that $\mathbf{x}_{ic}$ may correspond to less than $p_c$ locations. This dummy coding scheme is flexible in allowing dominant or recessive effects at each locus.

Scientific interest focuses on assessing the relationship between the SNPs for $C$ functionally-related genes, $\mathbf{x}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iC})'$, and a health outcome, $y_i$, adjusting for potential confounders, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iq})'$. For example, we are interested in relating SNPs in cytokine gene regulatory regions to risk for spontaneous preterm birth using data from the Pregnancy, Infection and Nutrition (PIN) study (Savitz et al., 1999), which enrolled women between 24 and 29 weeks gestation, collecting blood at the intake visit. As shown in Table 1, there are 12 cytokines (soluble proteins that mediate and regulate immunity and inflammation) of interest, including IL1$\alpha$, IL1$\beta$, IL2-IL6, IL10, IL13, LTA, TGF$\beta$1 and TNF. The number of sites within regulatory regions at which SNP data are collected ranges from one to three per cytokine (22 total), with 3 genotypes per site, resulting in $p = 66 - 22 = 44$ indicator variables in $\mathbf{x}_i$. The genotypes varied in frequency, with $\overline{x}_{ch} = \sum_{i=1}^{n} x_{ich} \in [8, 221]$, with $n = 447$ (excluding women with missing genotype data). It is straightforward to generalize the approach to account for genotypes that are missing at random by defining a model for the genotype frequencies, and updating the associated parameters along with the missing genotypes within an MCMC algorithm.

Results for one site at a time logistic regression analyses, with spontaneous preterm birth ($y_i = 1$ preterm, $y_i = 0$ full term) as the outcome, are provided in Table 1. Results are stratified on ethnicity (White, African American), because African American women have higher rates of spontaneous preterm birth and potentially-different genetic factors. Although

all three genotypes were represented in the study for each ethnic group, some categories had no women with spontaneous preterm births, so certain genotype-specific odds ratios could not be obtained. In addition, we were unable to obtain convergence for the full model with all SNPs included simultaneously.

Focusing on a smaller group of common proinflammatory cytokines and two genotype categories per SNP, Mulherin Engel et al. (2005a) reported an association with spontaneous preterm birth based on a semi-Bayes analysis (refer also to Mulherin Engel et al., 2005b). Our interest here is in identifying the specific SNPs predictive of spontaneous preterm birth, while also clustering SNPs within and across cytokines that have a similar risk of spontaneous preterm birth. Such clustering is not obvious from subjective examination of one site at a time or semi-Bayes analyses.

## 3. Hierarchical Clustering of Genetic Polymorphisms

### *3.1 Model and Background*

For simplicity, we will focus on the logistic regression model:

$$\text{logit}\,\Pr(y_i = 1 \,|\, \mathbf{x}_i, \mathbf{z}_i) = \mathbf{z}_i'\boldsymbol{\kappa} + \mathbf{x}_i'\boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_q)'$ are unknown coefficients including the intercept and slopes for the confounders, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_C')'$, with $\boldsymbol{\beta}_c = (\beta_{c1}, \ldots, \beta_{c,p_c})'$, are regression coefficients for the different SNPs. Clearly, when $p$ is large and $n_j = \sum_{i=1}^{n} x_{ij}$ is small for some $j$ (as is typically the case), maximum likelihood estimation of model (1) runs into difficulties. In particular, the estimated regression coefficients, $\widehat{\boldsymbol{\beta}}$, can be unstable, taking values known to be unreasonable *a priori*, and the MLEs may not exist.

A natural solution to this type of problem is to use a shrinkage estimator for $\boldsymbol{\beta}$, borrowing information across SNPs for functionally-related genes. This can be accomplished by assigning a prior distribution to the elements of $\boldsymbol{\beta}$ as follows:

$$\beta_{ch} \sim G_c, \quad \text{for } h = 1, \ldots, p_c \text{ and } c = 1, \ldots, C. \tag{2}$$

Here, $G_c$ is the unknown distribution of the regression coefficients for SNPs in gene $c$, and $\mathcal{G} = \{G_c, c = 1, \ldots, C\}$ is the collection of unknown distributions for the SNPs in the different genes.

Within a gene, dependency in the regression coefficients is accommodated by assuming that the SNP-specific regression coefficients are drawn from a common distribution. This will tend to shrink the coefficients towards each other in a manner dependent on the variance and shape of the distribution $G_c$. Between genes, dependency is accommodated by assuming that the different distributions in the collection $\mathcal{G}$ have similar features.

*3.2 Simultaneous Variable Selection and Clustering*

We first consider the case in which all the SNPs under study relate to a single gene, so that $C = 1$. In this case, repressing the $c$ subscript, we let $\beta_h \overset{iid}{\sim} G$, for $h = 1, \ldots, p$. Then, to allow for uncertainty in $G$, while clustering the SNPs having identical regression coefficients, we choose a DP prior, $G \sim DP(\alpha_0 G_0)$, with $\alpha_0$ a precision parameter and $G_0$ a base distribution. Following Sethuraman's (1994) stick-breaking representation:

$$G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \text{ where } \frac{\pi_h}{\prod_{l=1}^{h-1}(1-\pi_l)} \overset{iid}{\sim} \text{beta}(1, \alpha_0) \text{ and } \theta_h \overset{iid}{\sim} G_0, \tag{3}$$

for $h = 1, \ldots, \infty$, with $\delta_\theta$ denoting the degenerate distribution with all its mass at $\theta$. The distribution $G$ can be shown to be almost surely discrete with probability one, with the atoms $\boldsymbol{\Theta} = \{\theta_h, h = 1, \ldots, \infty\}$ generated independently from the base distribution $G_0$.

In contrast to common practice, which assumes that $G_0$ is non-atomic so that clustering arises solely from the discrete form of the DP (refer to Blackwell and MacQueen, 1973; Antoniak, 1974; among others), we assume

$$G_0 = \pi_0 \delta_0 + (1 - \pi_0)\text{N}(\mu_0, \sigma_0^2), \tag{4}$$

which is a mixture distribution consisting of a point mass at 0, with probability $\pi_0$, and a normal distribution with mean $\mu_0$ and variance $\sigma_0^2$. Related mixture distributions are used

routinely as priors for regression coefficients in variable selection applications (George and McCulloch 1997; Clyde and George, 2004). By incorporating a point mass at zero in the base distribution of the DP, we assign atoms to $\theta_h = 0$ with probability $\pi_0$ instead of generating distinct atoms with each new draw from $G_0$.

Under (4), we can reexpress (3) as follows:

$$\begin{aligned} G &= \sum_{h=1}^{\infty} \pi_h \{\pi_0 \delta_0 + (1-\pi_0)\delta_{\theta_h^*}\} = \pi_0 \delta_0 \Big\{ \sum_{h=1}^{\infty} \pi_h \Big\} + (1-\pi_0) \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*} \\ &= \pi_0 \delta_0 + (1-\pi_0) \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*} = \pi_0 \delta_0 + (1-\pi_0) G^*, \text{ where } \theta_h^* \overset{iid}{\sim} G_0^*, \end{aligned} \tag{5}$$

where $G^* \sim DP(\alpha_0 G_0^*)$, with $G_0^*$ denoting the non-atomic $\mathrm{N}(\mu_0, \sigma_0^2)$ distribution. Hence, the random distribution $G$ can be formulated as a mixture of a degenerate distribution with all its mass at zero and a DP with non-atomic base measure. Note that we assume $G_0^*$ corresponds to the normal distribution for concreteness, but the same concept could be used for other non-atomic base distributions.

Because we are interested in clustering and previous theoretic results focus primarily on clustering properties of the DP for non-atomic base measures, expression (5) is very useful. Let $p = p_0 + p_1$, with $p_0 = \sum_{h=1}^{p} 1(\beta_h = 0)$ denoting the number of SNPs having zero regression coefficients. Then, for $\beta_h \overset{iid}{\sim} G$, it follows directly from (5) that

$$\Pr(p_0 = h \,|\, \pi_0, \alpha_0, p) = \binom{p}{h} \pi_0^h (1-\pi_0)^{p-h}, \quad h = 1, \ldots, p, \tag{6}$$

so that, conditional on $\pi_0$, the prior distribution for the number of SNPs having zero coefficients is binomial. We refer to the group of SNPs having zero coefficients as the *null cluster*.

*Theorem 1.* Assume $\beta_h \overset{iid}{\sim} G$, for $h = 1, \ldots, p$, with $G$ defined by expressions (3) and (4). Then, letting $k^*$ denote the number of unique, non-zero elements of $\boldsymbol{\beta}$, the prior distribution for $k^*$ is

$$\Pr(k^* = k \,|\, \pi_0, \alpha_0, p) = \sum_{h=0}^{p} \binom{p}{h} (1-\pi_0)^h \pi_0^{p-h} \, \frac{a_h(k) \; \alpha_0^k}{\alpha_0^{(h)}}, \tag{7}$$

where $a_h(k)$ are the absolute values of Stirling numbers of the first kind (refer to Abramowitz and Stegun, 1964, page 833), and $\alpha^{(h)} = \alpha(\alpha+1)\dots(\alpha+h-1)$. The proof is straightforward using expressions (5) - (6) and the result of Antoniak (1974), page 1161.

From theorem 1, it is clear that the number of non-null clusters, which are defined as groups of SNPs have identical non-zero regression coefficients, increases stochastically with $\alpha_0$ and decreases with $\pi_0$. Hence, $\alpha_0$ and $\pi_0$ are key hyperparameters controlling the clustering of SNPs into null and non-null groups. Note also that the number of clusters increases automatically as the number of SNPs under consideration, $p$, increases. It is also apparent that the approach performs simultaneous variable selection and clustering, classifying a subset of SNPs as having no effect while clustering the remaining SNPs. There is a clear biological justification for clustering of the regression coefficients in this manner, because many polymorphisms occur together due to the presence of haplotypes. This constrains the possible unique sequences in $\mathbf{x}_i$, causing the number of possible genotypes to be much less than $2^p$.

To obtain additional insight into the clustering process, we derive prior probabilities of the coefficients $\boldsymbol{\beta}$ falling into different $\mathcal{C}$ classes, with $\boldsymbol{\beta}$ belongs to class $\mathcal{C}(m_0, m_1, \dots, m_p)$ if there are $m_0$ elements of $\boldsymbol{\beta}$ equal to zero, $m_1$ non-zero elements of $\boldsymbol{\beta}$ that occur once, $m_2$ non-zero elements that occur twice, up to $m_p$ non-zero elements that occur $p$ times. It follows that $k = 1(m_0 > 0) + \sum_{h=1}^{p} m_h = 1(m_0 > 0) + k^*$ is the number of unique elements of $\boldsymbol{\beta}$, denoted $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$.

*Theorem 2.* Assume $\beta_h \overset{iid}{\sim} G$, for $h = 1, \dots, p$, with $G$ defined by expressions (3) and (4). Then, the prior probability that $\boldsymbol{\beta}$ belongs to class $C(m_0, m_1, \dots, m_p)$ is

$$\begin{aligned}
&\Pr\{\boldsymbol{\beta} \in \mathcal{C}(m_0, m_1, \dots, m_p)\} \\
&\quad = \left[ \sum_{m_0=0}^{p} \binom{p}{m_0} \pi_0^{m_0} (1-\pi_0)^{p-m_0} \left\{ \frac{(p-m_0)!}{\prod_{h=1}^{p-m_0} h^{m_h} (m_h!)} \right\} \left\{ \frac{\alpha_0^{\sum_{h=1}^{p-m_0} m_h}}{\alpha_0^{(p-m_0)}} \right\} \right] \qquad (8)
\end{aligned}$$

This theorem follows from proposition 3 of Antoniak (1974) after appropriate modification

to allow the non-atomic base measure.

Theorem 2 can be used to derive the prior probabilities corresponding to a number of interesting special cases. For example, the probability that none of the SNPs have an effect is simply $\Pr\{\boldsymbol{\beta} \in \mathcal{C}(p, 0, \ldots, 0)\} = \pi_0^p$. The probability that all the SNPs have an equivalent non-null effect is

$$\Pr\{\boldsymbol{\beta} \in \mathcal{C}(0, \ldots, 0, 1)\} = (1 - \pi_0)^p \frac{\alpha_0 (p-1)!}{\prod_{h=1}^{p} (\alpha_0 + h - 1)}.$$

Other class probabilities corresponding to different numbers of null and non-null SNPs, and various clustering in the non-null SNPs, can be calculated easily. Potentially, $\pi_0$ and $\alpha_0$ can be chosen subjectively based on back-calculating from these probabilities.

*3.3 Semiparametric Hierarchical Clustering*

We now consider the case in which SNPs occur within different, functionally-related genes, and interest focuses on variable selection and clustering within and across genes. In particular, it is appealing to develop a method that allows SNPs for different genes to be assigned to the same cluster, while also allowing clusters to be gene-specific. Such a structure is consistent with the clustering the would arise from haplotypes within genes, and genes within biological pathways. One approach would be to extend the prior of Section 2.2 to incorporate a $c$ subscript on $G$ and then account for dependency in the elements of $\{G_c : c = 1, \ldots, C\}$ by applying the dependent Dirichlet process (DDP) of MacEachern (1999; 2000) (see also, De Iorio et al., 2004). This could be accomplished by defining parallel sticking breaking formulations for each $G_c$, and modeling dependency through a stochastic process for the atoms.

This DDP approach allows for dependency in the coefficients between genes, but does not allow clustering of SNPs in different genes. Therefore, we propose an alternative formulation:

$$\beta_{ch} \quad = \quad \psi_c + \gamma_{ch}, \; h = 1, \ldots, p_c, c = 1, \ldots, C$$

$$\gamma_{ch} \overset{iid}{\sim} F,\ F \sim DP(\alpha_0 F_0),\ F_0 = \delta_0 N(\pi_0, \mu_0, \sigma_0^2),\ \forall h, c$$
$$\psi_c \overset{iid}{\sim} H,\ H \sim DP(\alpha_1 H_0),\ H_0 = \delta_0 N(\pi_1, \mu_1, \sigma_1^2),\ \forall c, \quad (9)$$

where $\psi_c$ is a gene-specific factor, $\gamma_{ch}$ is a SNP-specific factor, and $\delta_0 \mathrm{N}(\pi, \mu, \sigma^2)$ is shorthand for the mixture distribution consisting of a point mass at zero with probability $\pi$ and a $N(\mu, \sigma^2)$ distribution with probability $1 - \pi$. This multi-level formulation allows clustering of SNPs both within and across genes.

There are $k_\Gamma \leq p = \sum_{c=1}^{C} p_c$ unique values $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_{k_\Gamma})'$ of the SNP-specific factor $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_C')$, with $\boldsymbol{\gamma}_c = (\gamma_{c1}, \ldots, \gamma_{c,p_c})'$ for $c = 1, \ldots, C$. There are also $k_\Psi \leq C$ unique values $\mathbf{\Psi} = (\Psi_1, \ldots, \Psi_{k_\Psi})'$ of the gene-specific factor $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_C)'$. Due to the masses at zero, one of the unique values of each of these vectors will typically correspond to zero, so we let $\Gamma_1 = 0$ and $\Psi_1 = 0$, without loss of generality. From this characterization, SNPs from the same gene (say SNPs $c, h$ and $c, h'$) belong to the same cluster if $\gamma_{ch} = \gamma_{ch'}$, while SNPs from different genes (say SNPs $c, h$ and $c', h'$) belong to the same cluster if $\gamma_{ch} = \gamma_{c'h'}$ and $\psi_c = \psi_{c'}$. In particular, given that the SNPs are not null, the probabilities of belonging to the same cluster are, respectively:

$$\Pr(\beta_{ch} = \beta_{ch'} \,|\, \beta_{ch} \neq 0, \beta_{ch'} \neq 0) = \frac{1}{\alpha_0 + 1},\ \Pr(\beta_{ch} = \beta_{c'h'} \,|\, \beta_{ch} \neq 0, \beta_{c'h'} \neq 0) = \frac{\pi_1^2}{\alpha_0 + 1}. \quad (10)$$

It follows that SNPs from the same gene are clustered together with higher probability than SNPs from different genes. Such within-gene dependency is often biologically reasonable. As a rule of thumb, SNPs in different regulatory regions for the same cytokine should have a higher chance of belonging to the same cluster than SNPs for different cytokines, because the biologic action of cytokines can vary. In addition, linkage disequilibrium can cause clustering within a gene, but would more rarely contribute to clustering in different genes occurring on the same chromosome.

To obtain additional insight into the clustering properties, we focus on the null cluster containing SNPs with zero coefficients. For greater flexibility, we choose hyperprior densities

for $\pi_0$ and $\pi_1$ as follows:

$$\pi_0 \sim \text{beta}(a_0, b_0) \quad \text{and} \quad \pi_1 \sim \text{beta}(a_1, b_1). \tag{11}$$

where $\mathbf{a} = (a_1, a_2)'$ and $\mathbf{b} = (b_1, b_2)'$ are pre-specified hyperparameters. Integrating out $\pi_0$ and $\pi_1$, the probability that SNP $c, h$ is null ($\beta_{ch} = 0$) is

$$\begin{aligned} \Pr(\beta_{ch} = 0 \,|\, \mathbf{a}, \mathbf{b}) &= \int\int \pi_0 \pi_1 \frac{\pi_0^{a_0 - 1}(1-\pi_0)^{b_0 - 1}}{B(a_0, b_0)} \, \frac{\pi_1^{a_1 - 1}(1-\pi_1)^{b_1 - 1}}{B(a_1, b_1)} \, d\pi_0 \, d\pi_1 \\ &= \Big(\frac{a_0}{a_0 + b_0}\Big)\Big(\frac{a_1}{a_1 + b_1}\Big). \end{aligned} \tag{12}$$

Similarly, the probability that all the SNPs in the $c$th gene belong to the null cluster is

$$\Pr(\boldsymbol{\beta}_c = \mathbf{0} \,|\, \mathbf{a}, \mathbf{b}) = \frac{a_0^{(p_c)}}{(a_0 + b_0)^{(p_c)}} \Big(\frac{a_1}{a_1 + b_1}\Big), \tag{13}$$

and the probability that all SNPs in all genes belong to the null cluster is

$$\Pr(\boldsymbol{\beta} = \mathbf{0} \,|\, \mathbf{a}, \mathbf{b}) = \frac{a_0^{(p)}}{(a_0 + b_0)^{(p)}} \frac{a_1^{(C)}}{(a_1 + b_1)^{(C)}}, \tag{14}$$

To illustrate the dependency structure, note that the probability that SNP $h$ in gene $c$ is null given another SNP $h'$ in gene $c$ is null is

$$\Pr(\beta_{ch} = 0 \,|\, \beta_{ch'} = 0, \mathbf{a}, \mathbf{b}) = \frac{\Pr(\beta_{ch} = \beta_{ch'} = 0 \,|\, \mathbf{a}, \mathbf{b})}{\Pr(\beta_{ch'} = 0 \,|\, \mathbf{a}, \mathbf{b})} = \frac{a_0 + 1}{a_0 + b_0 + 1}, \tag{15}$$

which is always higher than $\Pr(\beta_{ch} = 0 \,|\, \mathbf{a}, \mathbf{b})$. If we instead condition on knowledge that a SNP in a different gene is null, we obtain

$$\Pr(\beta_{ch} = 0 \,|\, \beta_{c'h'} = 0, \mathbf{a}, \mathbf{b}) = \frac{\Pr(\beta_{ch} = \beta_{c'h'} = 0 \,|\, \mathbf{a}, \mathbf{b})}{\Pr(\beta_{c'h'} = 0 \,|\, \mathbf{a}, \mathbf{b})} = \Big(\frac{a_0 + 1}{a_0 + b_0 + 1}\Big)\Big(\frac{a_1 + 1}{a_1 + b_1 + 1}\Big), \tag{16}$$

which is also higher than $\Pr(\beta_{ch} = 0 \,|\, \mathbf{a}, \mathbf{b})$ (shown in expression 12), but is lower than the probability in expression (15). Thus, the dependency between SNPs in a gene is higher than the dependency between SNPs in different genes, with the magnitude of the difference depending on the hyperparameters $\mathbf{a}$ and $\mathbf{b}$. In the limit as $a_0 + b_0 \to \infty$ and $a_1 + b_1 \to \infty$, holding $a_0/(a_0 + b_0)$ and $a_1/(a_1 + b_1)$ constant, expressions (12), (15) and (16) are equivalent

and there is no borrowing of information across SNPs about the probability of membership in the null cluster.

3.4 *Prior Elicitation*

Motivated by the cytokine application, we illustrate a strategy for prior elicitation. In particular, in choosing $a_0, b_0, a_1, b_1$, we recommend back-calculating from prior probabilities corresponding to different global and local hypotheses. For example, one can specify (i) the prior probability that none of the SNPs are associated with spontaneous preterm birth, $\Pr(\boldsymbol{\beta} = 0 \,|\, \mathbf{a}, \mathbf{b})$; (ii) the probability that a randomly-selected SNP is null, $\Pr(\beta_{ch} = 0 \,|\, \mathbf{a}, \mathbf{b})$; (iii) the probability that two SNPs within a gene are null, $\Pr(\beta_{ch} = \beta_{ch'} = 0 \,|\, \mathbf{a}, \mathbf{b})$; and (iv) the probability that two SNPs within different genes are null, $\Pr(\beta_{ch} = \beta_{c'h'} = 0 \,|\, \mathbf{a}, \mathbf{b})$. The hierarchical structure implies that the probabilities are ordered (i) $<$ (iv) $<$ (iii) $<$ (ii), so one should choose values consistent with this constraint. Because (i)-(iv) are simple analytic functions of $a_0, b_0, a_1, b_1$, it is straightforward to solve the system of non-linear equations using numerical methods.

In the cytokine application, we let $\Pr(\boldsymbol{\beta} = 0 \,|\, \mathbf{a}, \mathbf{b}) = 0.5$ in order to set the probability of the global null hypothesis equal to 0.5, corresponding to a 50% chance that any of the SNPs are predictive of spontaneous preterm birth. This represents a Bayesian approach to limit false positives that arise in multiple testing. We then let $\Pr(\beta_{ch} = 0 \,|\, \mathbf{a}, \mathbf{b}) = 0.8$, noting that the Bayesian Bonferroni approach (Westfall et al., 1997), which treats local hypotheses as independent, ignoring correlation, would instead choose $0.5^{1/44} = 0.984$. An approximately 1% chance that a SNP is important is unrealistically low, given that we are studying promising candidate SNPs. In addition, such a low prior probability would result in a very conservative procedure, requiring very large sample sizes to detect a health effect of the magnitude that would be expected in this study (e.g., odds ratio between 0.5 and 2). As plausible values for probabilities (iii) and (iv), we choose 0.75 and 0.73, respectively.

These values are chosen to be slightly lower than 0.8, with a modest degree of within-gene dependency. In simulation studies, we have found a high degree of robustness to the specific values chosen.

The precision parameters $\alpha_0$ and $\alpha_1$ are assigned gamma hyperprior distributions:

$$\alpha_0 \sim \text{gamma}(a_{\alpha_0}, b_{\alpha_0}) \quad \text{and} \quad \alpha_1 \sim \text{gamma}(a_{\alpha_1}, b_{\alpha_1}), \tag{17}$$

where $a_{\alpha_0}, b_{\alpha_0}, a_{\alpha_1}, b_{\alpha_1}$ are pre-specified hyperparameters. In choosing these values, we recommend letting $a_{\alpha_0} = b_{\alpha_0} = a_{\alpha_1} = b_{\alpha_1} = 1$, as a somewhat vague prior for the number of global and local clusters, which favors smaller numbers of clusters.

For the hyperparameters characterizing the base distributions, $\mu_0, \sigma_0^2, \mu_1, \sigma_1^2$, we recommend letting $\mu_0 = \mu_1 = 0$ and $\sigma_0^2 = \sigma_1^2 = 1$. By setting the means of the base distributions, $F_0$ and $H_0$, equal to zero, we express our uncertainty about the directions of the associations between the SNP categories and the risk of spontaneous preterm birth. The variances are chosen to assign a high probability to a plausible range for the SNP category-specific odds ratios. Potentially, hyperprior densities could be chosen for the means and variances for greater flexibility. This may be a useful strategy in cases in which there are very large numbers of candidate genes and SNPs, and less is known about scientifically plausible values for the regression coefficients.

## 4. Posterior Computation

For posterior computation, we propose a data augmentation Gibbs sampling algorithm. First, let $y_i = 1(y_i^* > 0)$, where $y_i^* = \mathbf{z}_i'\boldsymbol{\kappa} + \mathbf{x}_i'\boldsymbol{\beta} + \phi_i\epsilon_i$, with $\phi_i \sim \text{gamma}(\nu/2, \nu/2)$ and $\epsilon_i \sim \text{N}(0, \sigma^2)$, resulting in a $t$ density for $\phi_i\epsilon_i$. Following O'Brien and Dunson (2004), setting $\sigma^2 = \pi(\nu - 2)/3\nu$ and $\nu = 7.3$ produces an almost exact approximation to the logistic density. The algorithm alternates between (1) sampling $y_i^*$ and $\phi_i$ from their respective truncated normal and gamma full conditional posterior distributions, for $i = 1, \ldots, n$; and (2) sampling unknowns related to $\boldsymbol{\beta}$ jointly with $\boldsymbol{\kappa}$, assuming a $\text{N}(\boldsymbol{\kappa}_0, \boldsymbol{\Sigma}_\kappa)$ prior for $\boldsymbol{\kappa}$. The

first step is straightforward, so we focus our attention on step 2.

Letting the $(ch)$ superscript denote a quantity obtained excluding element $c, h$, the conditional prior distribution of $\gamma_{ch}$ given $\boldsymbol{\gamma}^{(ch)}$ is

$$\Big(\frac{\alpha_0(1-\pi_0)}{\alpha_0+p-p^{(ch)}_{\Gamma_0}-1}\Big)N(\mu_0,\sigma_0^2)+\pi_0\delta_0+\sum_{l=2}^{k^{(ch)}_\Gamma}\Big(\frac{p^{(ch)}_{\Gamma_l}(1-\pi_0)}{\alpha_0+p-p^{(ch)}_{\Gamma_0}-1}\Big)\delta_{\Gamma^{(ch)}_l}, \tag{18}$$

where $p^{(ch)}_{\Gamma_l}$ is the number of elements of $\boldsymbol{\gamma}^{(ch)}$ equal to $\Gamma^{(ch)}_l$, $\boldsymbol{\Gamma}^{(ch)}=(\Gamma^{(ch)}_l, l=1,\ldots,k^{(ch)}_\Gamma)'$, $\Gamma^{(ch)}_1=0$, $\Gamma^{(ch)}_l$ for $l=2,\ldots,k^{(ch)}_\Gamma$ denotes the unique non-zero values of $\boldsymbol{\gamma}^{(ch)}$, and $k^{(ch)}_\Gamma$ is the number of atoms in (18). Expression (18) follows by applying the Pólya urn scheme of Blackwell and MacQueen (1973), after placing $F$ in the form of expression (5).

The conditional distribution of $\psi_c$ given $\boldsymbol{\psi}^{(c)}=(\psi_1,\ldots,\psi_{c-1},\psi_{c+1},\ldots,\psi_C)'$ is

$$\Big(\frac{\alpha_1(1-\pi_1)}{\alpha_1+C-p^{(c)}_{\Psi_0}-1}\Big)N(\mu_1,\sigma_1^2)+\pi_1\delta_0+\sum_{l=2}^{k^{(c)}_\Psi}\Big(\frac{p^{(c)}_{\Psi_l}(1-\pi_1)}{\alpha_1+C-p^{(c)}_{\Psi_0}-1}\Big)\delta_{\Psi^{(c)}_l}, \tag{19}$$

where $p^{(c)}_{\Psi_l}$ is the number of elements of $\boldsymbol{\psi}^{(c)}$ equal to $\Psi^{(c)}_l$, $\boldsymbol{\Psi}^{(c)}=(\Psi^{(c)}_l, l=1,\ldots,k^{(c)}_\Psi)'$, $\Psi_1=0$, $\Psi_l$ for $l=2,\ldots,k^{(c)}_\Psi$ denotes the unique non-zero values of $\boldsymbol{\psi}^{(c)}$, and $k^{(c)}_\Psi$ is the number of atoms in (19).

As shorthand, let $\mathbf{u}^{(ch)}=(u^{(ch)}_0,u^{(ch)}_1,\ldots,u^{(ch)}_{k^{(ch)}_\Gamma})'$ and $\mathbf{w}^{(c)}=(w^{(c)}_0,w^{(c)}_1,\ldots,w^{(c)}_{k^{(c)}_\Psi})'$ denote the probability weights on the respective mixture components in expressions (18) and (19). Updating conditional priors (18) and (19) using information in the data, we obtain the following full conditional posterior distributions:

$$(\gamma_{ch}\,|\,-) \;=\; U^{(ch)}_0\mathrm{N}(\gamma_{ch};E^{(ch)}_\gamma,V^{(ch)}_\gamma)+\sum_{l=1}^{k^{(ch)}_\Gamma}U^{(ch)}_l\delta_{\Gamma^{(ch)}_l}, \tag{20}$$

$$(\psi_c\,|\,-) \;=\; W^{(c)}_0\mathrm{N}(\psi_c;E^{(c)}_\psi,V^{(c)}_\psi)+\sum_{l=1}^{k^{(c)}_\Psi}W^{(c)}_l\delta_{\Psi^{(c)}_l}, \tag{21}$$

where $\widetilde{y}^{(c)}_i=y^*_i-\mathbf{z}_i'\boldsymbol{\kappa}-\mathbf{x}_i\boldsymbol{\gamma}-\sum_{c'\neq c}x_{ic'}\psi_{c'}$, $x_{ic}=\sum_{h=1}^{p_c}x_{ich}$, $\widetilde{y}^{(ch)}_i=y^*_i-\mathbf{z}_i'\boldsymbol{\kappa}-\mathbf{x}^{(ch)'}_i\boldsymbol{\beta}^{(ch)}-x_{ich}\psi_c$, the conditional posterior means and variances in the normal components are

$$V^{(ch)}_\gamma=\Big(\sigma_0^{-2}+\sum_{i=1}^n\sigma_i^{-2}x^2_{ich}\Big)^{-1},\quad E^{(ch)}_\gamma=V^{(ch)}_\gamma\Big(\sigma_0^{-2}\mu_0+\sum_{i=1}^n\sigma_i^{-2}x_{ich}\widetilde{y}^{(ch)}_i\Big),$$

$$V_\psi^{(c)} = \Big(\sigma_1^{-2} + \sum_{i=1}^{n} \sigma_i^{-2} x_{ic}^2\Big)^{-1}, \quad E_\psi^{(c)} = V_\psi^{(c)}\Big(\sigma_1^{-2}\mu_1 + \sum_{i=1}^{n} \sigma_i^{-2} x_{ic}\widetilde{y}_i^{(c)}\Big),$$

and the updated mixture weights are defined as follows:

$$\begin{aligned} U_0^{(ch)} &= c_u \times \frac{u_0^{(ch)}\mathrm{N}(0;\mu_0,\sigma_0^2)\prod_{i=1}^n \mathrm{N}(\widetilde{y}_i^{(ch)};0,\sigma_i^2)}{\mathrm{N}(0;E_\gamma^{(ch)},V_\gamma^{(ch)})},\ U_l^{(ch)} = c_u \times u_l^{(ch)} \prod_{i=1}^{n} \mathrm{N}(\widetilde{y}_i^{(ch)};\Gamma_l^{(ch)},\sigma_i^2), \\ W_0^{(c)} &= c_w \times \frac{w_0^{(c)}\mathrm{N}(0;\mu_1,\sigma_1^2)\prod_{i=1}^n \mathrm{N}(\widetilde{y}_i^{(c)};0,\sigma_i^2)}{\mathrm{N}(0;E_\psi^{(c)},V_\psi^{(c)})},\ W_l^{(c)} = c_w \times w_l^{(c)} \prod_{i=1}^{n} \mathrm{N}(\widetilde{y}_i^{(c)};\Psi_l^{(c)},\sigma_i^2), \end{aligned}$$

where $c_u$ and $c_w$ are normalizing constants.

We follow West et al. (1994) and MacEachern (1994) in alternating between updating (i) the cluster allocation; and (ii) the cluster-specific parameters. Let $\mathcal{S}_{ch} = l$ if $\gamma_{ch} = \Gamma_l^{(ch)}$, for $l = 1, \ldots, k_\Gamma^{(ch)}$, and $\mathcal{T}_c = l$ if $\psi_c = \Psi_l^{(c)}$, for $l = 1, \ldots, k_\Psi^{(c)}$, index the allocation of $\gamma_{ch}$ and $\psi_c$ to clusters. In addition, let $\mathcal{S}_{ch} = 0$ denote that $\gamma_{ch} \notin \mathbf{\Gamma}^{(ch)}$, so that SNP $c, h$ cannot be grouped with the other SNPs and a new cluster needs to be introduced. Also, $\mathcal{T}_c = 0$ denotes that $\psi_c \notin \mathbf{\Psi}^{(c)}$, so that a new cluster is introduced for gene $c$. The conditional posterior distributions of $\mathcal{S}_{ch}$ and $\mathcal{T}_c$ are respectively:

$$(\mathcal{S}_{ch} \,|\, \mathbf{S}^{(ch)}, \mathbf{T}, \mathbf{\Gamma}^{(ch)}, \mathbf{\Psi}, \text{data}) = \text{Multinomial}\Big(0, 1, \ldots, k_\Gamma^{(ch)}; U_0^{(ch)}, U_1^{(ch)}, \ldots, U_{k_\Gamma^{(ch)}}^{(ch)}\Big), \quad (22)$$

$$(\mathcal{T}_c \,|\, \mathbf{S}, \mathbf{T}^{(c)}, \mathbf{\Gamma}, \mathbf{\Psi}^{(c)}, \text{data}) = \text{Multinomial}\Big(0, 1, \ldots, k_\Psi^{(c)}; W_0^{(c)}, W_1^{(c)}, \ldots, W_{k_\Psi^{(c)}}^{(c)}\Big). \quad (23)$$

In step 2 (i), we sample from these multinomial distributions. When $\mathcal{S}_{ch} = 0$ a new value for $\gamma_{ch}$ is drawn from $\mathrm{N}(E_\gamma^{(ch)}, V_\gamma^{(ch)})$, while when $\mathcal{T}_c = 0$ a new value for $\psi_c$ is drawn from $\mathrm{N}(E_\psi^{(c)}, V_\psi^{(c)})$.

Then, in step 2 (ii) we update $\boldsymbol{\kappa}$ jointly with the cluster-specific parameters $\mathbf{\Gamma}^* = (\Gamma_2, \ldots, \Gamma_{k_\Gamma})'$ and $\mathbf{\Psi}^* = (\Psi_2, \ldots, \Psi_{k_\Psi})'$ by sampling from the conditional posterior distribution given the cluster allocation indicators, $\mathbf{S}, \mathbf{T}$, and number of clusters, $k_\Gamma, k_\Psi$:

$$(\mathbf{\Theta} \,|\, \mathbf{S}, k_\Gamma, k_\Psi, \text{data}) = \mathrm{N}(\widehat{\mathbf{\Theta}}, \widehat{V}_\Theta), \quad (24)$$

where $\mathbf{\Theta} = (\boldsymbol{\kappa}', \mathbf{\Gamma}^{*'}, \mathbf{\Psi}^{*'})'$, $\Sigma_\Theta = \text{block-diag}(\Sigma_\kappa, \sigma_0^2\mathbf{I}_{k_\Gamma - 1}, \sigma_1^2\mathbf{I}_{k_\Psi - 1})$, $\mathbf{\Theta}_0 = (\boldsymbol{\kappa}_0', \mu_0\mathbf{1}_{k_\Gamma-1}', \mu_1\mathbf{1}_{k_\Psi-1}')$,

$$\widehat{V}_\Theta = \Big(\Sigma_\Theta^{-1} + \sum_{i=1}^{n} \sigma_i^{-2}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i\Big)^{-1}, \quad \widehat{\mathbf{\Theta}} = \widehat{V}_\Theta\Big(\Sigma_\Theta^{-1}\mathbf{\Theta}_0 + \sum_{i=1}^{n} \sigma_i^{-1}\widetilde{\mathbf{x}}_i y_i^*\Big),$$

$\widetilde{\mathbf{x}}_i = (\mathbf{z}_i', \mathbf{x}_{\Gamma,i}', \mathbf{x}_{\Psi,i}')'$, $\mathbf{x}_{\Gamma,i} = (x_{\Gamma,im}, m = 2, \ldots, k_\Gamma)'$, $\mathbf{x}_{\Psi,i} = (x_{\Psi,il}, l = 2, \ldots, k_\Psi)'$, $x_{\Gamma,im} = \sum_{c=1}^{C}\sum_{h=1}^{p_c} x_{ich} 1(\mathcal{S}_{ch} = m)$, and $x_{\Psi,il} = \sum_{c=1}^{C}\sum_{h=1}^{p_c} x_{ich} 1(\mathcal{T}_c = l)$. This block updating procedure should improve computational efficiency substantially over one at a time updating.

The full conditional posterior distributions of $\pi_0$ and $\pi_1$ are, respectively:

$$(\pi_0 \,|\, \mathbf{S}, \boldsymbol{\Gamma}, \mathbf{T}, \boldsymbol{\Psi}, \boldsymbol{\kappa}, \mathbf{y}^*) = \text{beta}\Big(a_0 + \sum_{c=1}^{C}\sum_{h=1}^{p_c} 1(\mathcal{S}_{ch} = 1), b_0 + p - \sum_{c=1}^{C}\sum_{h=1}^{p_c} 1(\mathcal{S}_{ch} = 1)\Big), \quad (25)$$

$$(\pi_1 \,|\, \mathbf{S}, \boldsymbol{\Gamma}, \mathbf{T}, \boldsymbol{\Psi}, \boldsymbol{\kappa}, \mathbf{y}^*) = \text{beta}\Big(a_1 + \sum_{c=1}^{C} 1(\mathcal{T}_c = 1), b_1 + C - \sum_{c=1}^{C} 1(\mathcal{T}_c = 1)\Big). \quad (26)$$

In step 2 (iii) we sample from these conditionals. Finally, in step 2 (iv) we update $\alpha_0$ and $\alpha_1$ by applying the approach of Escobar and West (1995),

## 5. Cytokines and Preterm Birth Application

### 5.1 *Real Data Results*

We applied the approach to data from the cytokines and spontaneous preterm birth application introduced in Section 2. Using the approach to prior elicitation proposed in Section 3.4, we obtained $a_0 = 0.55, b_0 = 0.10, a_1 = 0.82, b_1 = 0.049$, which implies $\Pr(\beta_{ch} = 0) = 0.8$ and $\Pr(\boldsymbol{\beta} = 0) = 0.5$. We ran analyses separately for African Americans ($n = 195$) and Caucasians ($n = 252$), following standard epidemiologic practice in this area. In each case, the MCMC algorithm was run for 100,000 iterations, discarding a burn-in of 5,000 iterations and collecting every 10th sample to thin the chain. Samples converged quickly to a stationary distribution and mixing was rapid, suggesting that the proposed algorithm is efficient.

For whites, the posterior probability of the global alternative hypothesis that any of the SNPs were predictive of spontaneous preterm birth was $\Pr(H_1 \,|\, \text{data}) = 0.224$, with the corresponding Bayes factor being BF = 0.288. For African Americans, the values were $\Pr(H_1 \,|\, \text{data}) = 0.149$ and BF = 0.176. Hence, the data support the null hypothesis that cytokine polymorphisms are not predictive of spontaneous preterm birth. These results are robust to moderate changes in the prior, and we repeated the analysis with $\Pr(\boldsymbol{\beta} = 0) = 0.2$

and $\Pr(\beta_{ch} = 0) = 0.5$ without change in the conclusion. Conclusions were also unchanged running the analysis for blacks and whites combined. SNP-specific BFs ($\times$ marks) and posterior means given inclusion in the model ($o$ marks) are provided in Figures 1 and 2 for Whites and African Americans. In each case, the SNP-specific BFs were well below one. In general, the estimated coefficients given inclusion tend to parallel the results shown in Table 1, with the extreme estimates occurring at similar locations. However, the model-averaged estimates (not shown) are all approximately zero, reflecting the low inclusion probabilities.

5.2 *Simulation Study*

A potential concern is that the approach may be overly-conservative, particularly in cases in which there are effects only for one or two of the SNPs. In such cases, borrowing of information across the SNPs regarding the probability of inclusion in the model can conceivably cause SNP-specific effects to be obscured. To assess whether the approach is capable of detecting SNP-specific effects, we ran a small simulation study. In particular, using the sample size and genotype data for the women in the PIN study but randomly permuting the assignment of SNPs to subjects, we simulated the spontaneous preterm birth outcome variable under four different scenarios: (i) null model ($\beta_{ch} = 0$ for all $c, h$); (ii) one SNP positive (cytokine IL13, location -1112, genotype TC) positive ($\beta_{83} = 1$, $\beta_{ch} = 0$ for all $c, h \neq 8, 3$); (iii) two SNPs positive ($\beta_{83} = \beta_{91} = 1$); and finally a more interesting case (iv) in which there were two non-null clusters at $\beta = 1$ [SNPs $\{(2,5),(3,1),(5,1),(7,1),(9,3)\}$] and $\beta = 1.5$ [SNPs $\{(4,3),(6,1),(8,3),(10,3)\}$].

Under each scenario, we simulated 25 data sets, implementing the MCMC algorithm as for the real data example in each case, but with the algorithm run for 10,000 iterations with a 1,000 burn-in. Summaries of the results for cases (i)-(iii) are presented in Table 2. In case (i), only 1/25 data sets had estimated Bayes factors greater than one for either the global or local alternatives, suggesting that the approach does not tend to produce many

false positives. In case (ii), 24/25 of the data sets had $BF_{8,3} > 1$ and 20/25 data sets had a global $BF > 1$, suggesting that the approach has good power to detect SNP-specific effects. In addition, the posterior mean for the coefficient for the positive SNP was close to the true value. Similar results were obtained for case (iii).

For the more complex case (iv), results are presented in Figures 3 and 4. Figure 3 shows the estimated marginal inclusion probabilities for each of the SNPs, ordered so that 1-35 are in the null cluster, 36-40 are in the $\beta = 1$ cluster, and 41-44 are in the $\beta = 1.5$ cluster. Values for each of the 25 simulated data sets are shown, with the horizontal line representing the average inclusion probability across SNPs in a cluster and across data sets. Clearly, the approach tends to assign substantially higher inclusion probabilities to the SNPs in the two non-null clusters. Figure 4 shows the posterior means for $\beta_{ch}$ for all $c, h$ across the different simulated data sets, using the means conditional on inclusion in the model. On average, the null SNPs have estimated coefficients close to zero given inclusion in the model (the model-averaged estimates are all very close to zero), while the non-null SNPs have estimated coefficients close to the true value, with some evidence of shrinkage towards zero.

## 6. Discussion

This article has proposed a semiparametric Bayesian approach for simultaneous variable selection and clustering in applications involving many, related predictors. There is a rapidly expanding literature on methods for identifying important predictors from an extremely high-dimensional set of candidates, primarily motivated by gene expression data (Efron et al., 2001; Newton et al., 2001; Ibrahim et al., 2002 among many others). Our motivation is somewhat different in that we are interested in more focused genetic studies that collect genotype data at a moderate number of locations (e.g., 30+), corresponding to regulatory or coding regions for functionally-related genes. Such studies are potentially conducted as a second stage after preliminary identification of promising candidate genes through gene

expression studies. Because we have a more modest number of predictors, we can be more ambitious in attempting to address questions about overall significance and clustering of effect sizes. Our method should also be useful in epidemiologic studies collecting information for environmental exposures, such as pesticides or nutrients, that can be grouped into pre-specified classes.

Our motivation was genetic epidemiology studies in which investigators preselect SNPs based on presumed functionality judged from the literature. This extremely common strategy tends to limit the number of SNPs that need to be genotyped, but can potentially misrepresent variation within a gene. Technological advances now allow one to use a dense collection of *tagSNPs* that may have no function in themselves, but are instead markers of variability within a gene. TagSNPs can be selected to be approximately evenly spaced across a gene, or they can be selected on the basis of estimates of linkage disequilibrium (LD) which will result in a denser set of markers covering areas of low LD. Our proposed method is promising as an approach for identifying regions of a gene that may contain a functional SNP(s) from a field of anonymous tagSNPs. Otherwise, by relying on preselection of a small number of SNPs, there is always the possibility that important variability exists at other locations. Hence, inferences are necessarily limited by the SNPs chosen and one can not make general conclusions about the importance of a particular gene in predicting a health outcome.

We have proposed a particular strategy of prior elicitation that treats the different SNPs as exchangeable within a gene, while also treating the genes as exchangeable. Although this is a reasonable default strategy for many studies, in certain cases there may be information available to suggest that certain genes and SNPs are particularly promising candidates, while little or no information is available for others. In such cases, as noted by Wacholder et al. (2004), the exchangeability assumption is implausible. Fortunately, it is straightforward to modify our procedure to allow the prior probabilities of inclusion to vary for the different

genes and SNPs under study.

## References


Antoniak, C.E. (1974) Mixtures of Dirichlet processes with application to nonparametric problems. *The Annals of Statistics*, **2**, 1152-1174.

Berry, D.A. and Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82, 215-227.

Berry, S.M. and Berry, D.A. (2004). Accounting for multiplicities in assessing drug safety: A three level hierarchical mixture model. *Biometrics*, 60, 418-426.

Blackwell, D. and MacQueen, J.B. (1973) Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, **1**, 353-355.

Coull, B.A., Hobert, J.P., Ryan, L.M. and Holmes, L.B. (2001). Crossed random effect models for multiple outcomes in a study of teratogenesis. *Journal of the American Statistical Association*, 96, 1194-1204.

De Iorio, M., Müller, P., Rosner, G.L. and MacEachern, S.N. (2004) An Anova model for dependent random measures. *Journal of the American Statistical Association*, **99**, 205-215.

De Roos, A.J., Poole, C., Teschke, K. and Olshan, A.F. (2001). An application of hierarchical regression in the investigation of multiple paternal occupational exposures and neuroblastoma in offspring. *American Journal of Industrial Medicine*, 39, 477-486.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V.G. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96, 1151-1160.

Dunson, D.B. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, 100, 618-627.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209-230.

Ferguson, T.S. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615-629.

Gonen, M., Westfall, P.H. and Johnson, W.O. (2003). Bayesian multiple testing for two-sample multivariate endpoints. *Biometrics*, 59, 76-82.

Gopalan, R. and Berry, D.A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association*, 93, 1130-1139.

Greenland, S. (1992). A Semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Statistics in Medicine*, 11, 219-230.

Greenland, S. (1993). Methods for epidemiologic analyses of multiple exposures - A review and comparative-study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Statistics in Medicine*, 12, 717-736.

Greenland, S. (1994). Hierarchical regression for epidemiologic analyses of multiple exposures. *Environmental Health Perspective*, 102, 33-39.

Hung, R.J., Brennan, P., Malaveille, C., Porru, S., Donato, F., Boffetta, P. and Witte, J.S. (2004). Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer. *Cancer Epidemiology Biomarkers & Prevention*, 13, 1013-1021.

Ibrahim, J.G., Chen, M.-H., and Gray, R.J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97, 88-99.

MacEachern, S.N. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, **23**, 727-741.

MacEachern, S.N. (1999) Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.

MacEachern, S.N. (2000) Dependent Dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University.

Meng, C.Y.K. and Dempster, A.P. (1987). A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Biometrics*, 43, 301-311.

Mulherin Engel, S.A., Ericksen, H.C., Savitz, D.A., Thorp, J., Chanock, S.J. and Olshan, A.F. (2005a), "Risk of Spontaneous Preterm Birth is Associated with Common Proinflammatory Cytokine Polymorphisms," *Epidemiology*, 16, 469-477.

Mulherin Engel, S.A., Olshan, A.F., Savitz, D.A., Thorp, J., Ericksen, H.C. and Chanock, S.J. (2005b), "Risk of Small-for-Gestational Age is Associated with Common Anti-Inflammatory Cytokine Polymorphisms," *Epidemiology*, 16, 478-486.

Newton, M.A., Kendziorski, C.M., Richmond, C.C., Blattner, F.R. and Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8, 37-52.

Savitz, D.A., Dole, N., Williams, J. et al. (1999), "Determinants of Participation in an Epidemiological Study of Preterm Delivery," *Paediatric and Perinatal Epidemiology*,

13, 114-125.

Sethuraman, J. (1994), "A Constructive Definition of the Dirichlet Process Prior," *Statistica Sinica*, 2, 639-650.

Tadesse, M.G., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100, 602-617.

Thomas, D.C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M. and Armstrong, B.G. (1985). The problem of multiple inference in studies designed to generate hypotheses. *American Journal of Epidemiology*, 122, 1080-1095.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. and Rothman (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96, 434-442.

West, M., Müller, P. and Escobar, M.D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *A Tribute to D. V. Lindley* (A.F.M. Smith and P.R. Freeman). John Wiley and Sons.

Westfall, P.H., Johnson, W.O. and Utts, J.M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, 84, 419-427.

Witte, J.S., Greenland, S., Haile, R.W. and Bird, C.L. (1994). Hierarchical regression-analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology*, 5, 612-621.

**Table 1**. Summary of one site at a time logistic regression analyses with spontaneous preterm birth as the outcome variable.

| Cytokine | Site | Genotype | Odds Ratio for Spontaneous Preterm Birth | |
|---|---|---|---|---|
| | | | White | African American |
| *IL1α* | +4845 | GG | 1.0 | 1.0 |
| | | GT | 1.1 (0.6, 1.9) | 0.8 (0.5, 1.5) |
| | | TT | 1.8 (0.6, 5.2) | -† |
| | IVS5-109 | AA | 1.0 | 1.0 |
| | | AC | 0.7 (0.4, 1.3) | 1.1 (0.6, 1.9) |
| | | CC | 0.4 (0.1, 1.7) | 1.1 (0.4, 3.2) |
| *IL1β* | 1061 | CC | 1.0 | 1.0 |
| | | TC | 1.8 (1.0, 3.3) | 0.7 (0.3, 1.6) |
| | | TT | 1.1 (0.4, 3.0) | 1.1 (0.5, 2.5) |
| | +3594 | CC | 1.0 | 1.0 |
| | | CT | 1.2 (0.7, 2.1) | 0.9 (0.5, 1.7) |
| | | TT | -† | -† |
| | -581 | TT | 1.0 | 1.0 |
| | | TC | 2.0 (1.1, 3.8) | 0.6 (0.3, 1.4) |
| | | CC | 1.0 (0.4, 2.8) | 1.1 (0.5, 2.4) |
| *IL2* | -385 | TT | 1.0 | 1.0 |
| | | TG | 1.6 (0.9, 2.9) | 0.7 (0.3, 1.6) |
| | | GG | 1.1 (0.4, 2.9) | -† |
| *IL4* | -589 | CC | 1.0 | 1.0 |
| | | CT | 0.8 (0.4, 1.6) | 0.7 (0.3, 1.7) |
| | | TT | 28.4 (3.3, 241.5) | 0.9 (0.4, 2.0) |
| | -1099 | TT | 1.0 | 1.0 |
| | | GT | 0.7 (0.3, 1.9) | 1.8 (1.0, 3.3) |
| | | GG | -† | 0.5 (0.1, 3.9) |
| | -33 | CC | 1.0 | 1.0 |
| | | TC | 0.8 (0.4, 1.7) | 1.9 (1.0, 3.7) |
| | | TT | 14.9 (2.9, 76.5) | 1.7 (0.7, 4.0) |
| *IL5* | -746 | TT | 1.0 | 1.0 |
| | | TC | 1.4 (0.4, 5.0) | 0.9 (0.5, 1.6) |
| | | CC | 2.5 (0.7, 8.7) | 0.7 (0.2, 3.6) |
| *IL6* | -174 | GG | 1.0 | 1.0 |
| | | CG | 1.1 (0.6, 2.1) | 1.1 (0.5, 2.6) |
| | | CC | 1.2 (0.5, 2.7) | -† |

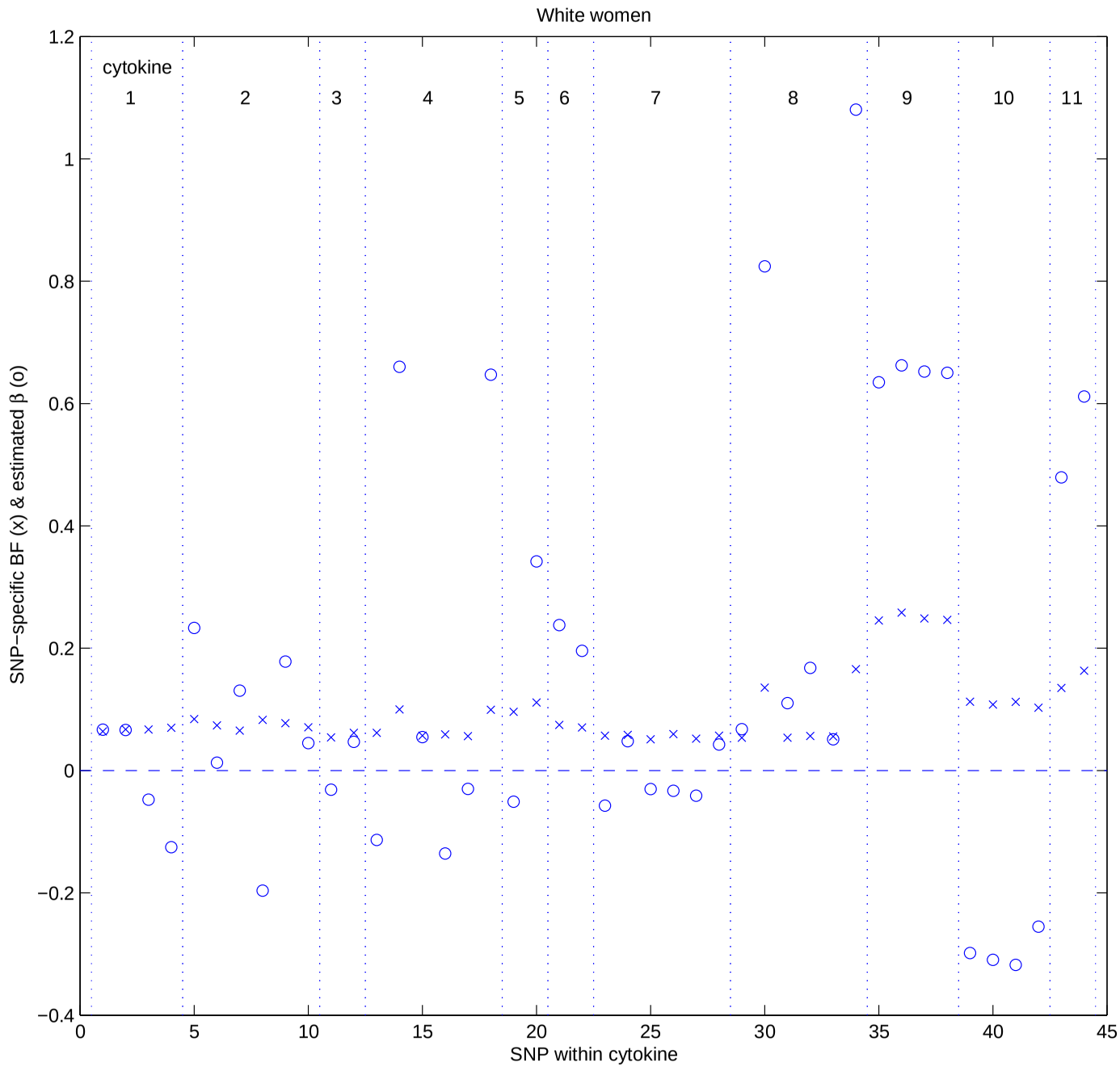
**Table 1, continued**

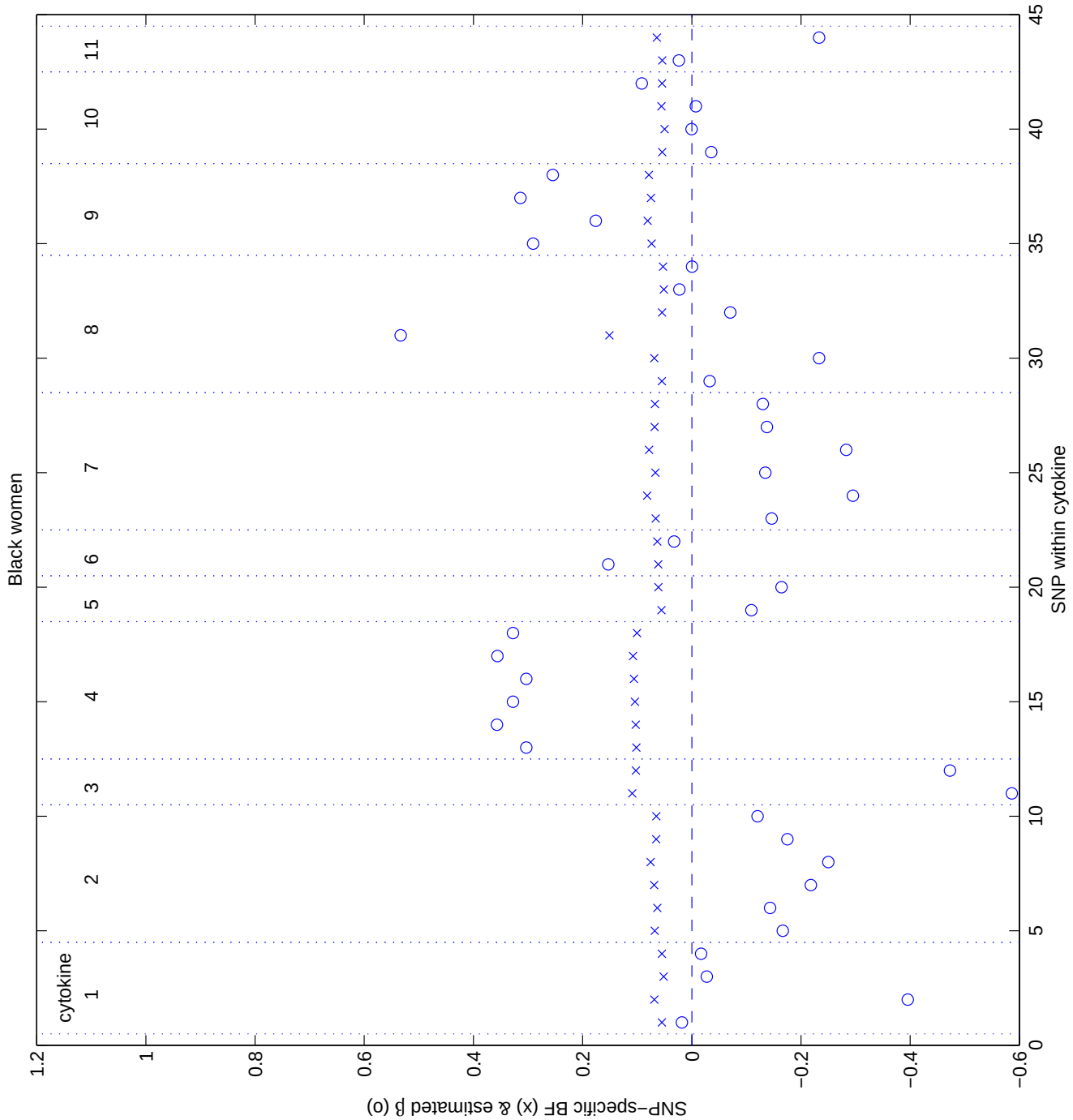
| Cytokine | Site | Genotype | Odds Ratio for Spontaneous Preterm Birth | |
|---|---|---|---|---|
| | | | White | African American |
| *IL10* | -854 | CC | 1.0 | 1.0 |
| | | TC | 0.8 (0.5, 1.5) | 0.9 (0.5, 1.7) |
| | | TT | 1.0 (0.4, 2.9) | 0.5 (0.2, 1.1) |
| | -627 | CC | 1.0 | 1.0 |
| | | | 0.8 (0.5, 1.5) | 1.0 (0.5, 1.8) |
| | | | 0.7 (0.2, 2.2) | 0.5 (0.2, 1.3) |
| | -1082 | AA | 1.0 | 1.0 |
| | | AG | 1.5 (0.7, 3.1) | 1.1 (0.6, 2.0) |
| | | GG | 1.5 (0.7, 3.5) | 1.2 (0.5, 3.1) |
| *IL13* | +2034 | GG | 1.0 | 1.0 |
| | | AG | 1.3 (0.7, 2.3) | 0.8 (0.5, 1.5) |
| | | AA | 5.4 (1.8, 16.3) | 0.4 (0.0, 3.0) |
| | -1112 | CC | 1.0 | 1.0 |
| | | TC | 1.6 (0.9, 2.9) | 2.2 (1.2, 4.2) |
| | | TT | 2.6 (0.9, 8.1) | 1.1 (0.4, 3.0) |
| | IVS3-24 | CC | 1.0 | 1.0 |
| | | TC | 1.3 (0.7, 2.3) | 1.1 (0.4, 2.6) |
| | | TT | 4.7 (1.6, 13.9) | 0.9 (0.4, 2.2) |
| *LTA* | IVS1+90 | AA | 1.0 | 1.0 |
| | | AG | 1.3 (0.7, 2.3) | 1.4 (0.7, 2.8) |
| | | GG | 1.8 (0.7, 4.4) | 0.9 (0.4, 2.1) |
| | IVS1-82 | GG | 1.0 | 1.0 |
| | | CG | 1.7 (0.9, 3.2) | 1.3 (0.7, 2.3) |
| | | CC | 1.3 (0.5, 3.2) | 0.9 (0.2, 3.5) |
| *TGFβ1* | L10P | TT | 1.0 | 1.0 |
| | | TC | 0.8 (0.4, 1.4) | 1.1 (0.6, 2.0) |
| | | CC | 0.5 (0.2, 1.3) | 1.2 (0.5, 2.7) |
| | -1347 | CC | 1.0 | 1.0 |
| | | CT | 0.8 (0.4, 1.5) | 1.2 (0.7, 2.1) |
| | | TT | 0.5 (0.1, 1.7) | 1.4 (0.3, 5.5) |
| *TNF* | -308 | GG | 1.0 | 1.0 |
| | | GA | 1.6 (0.9, 2.9) | 1.0 (0.5, 2.0) |
| | | AA | 3.3 (0.9, 11.9) | 0.5 (0.1, 3.9) |

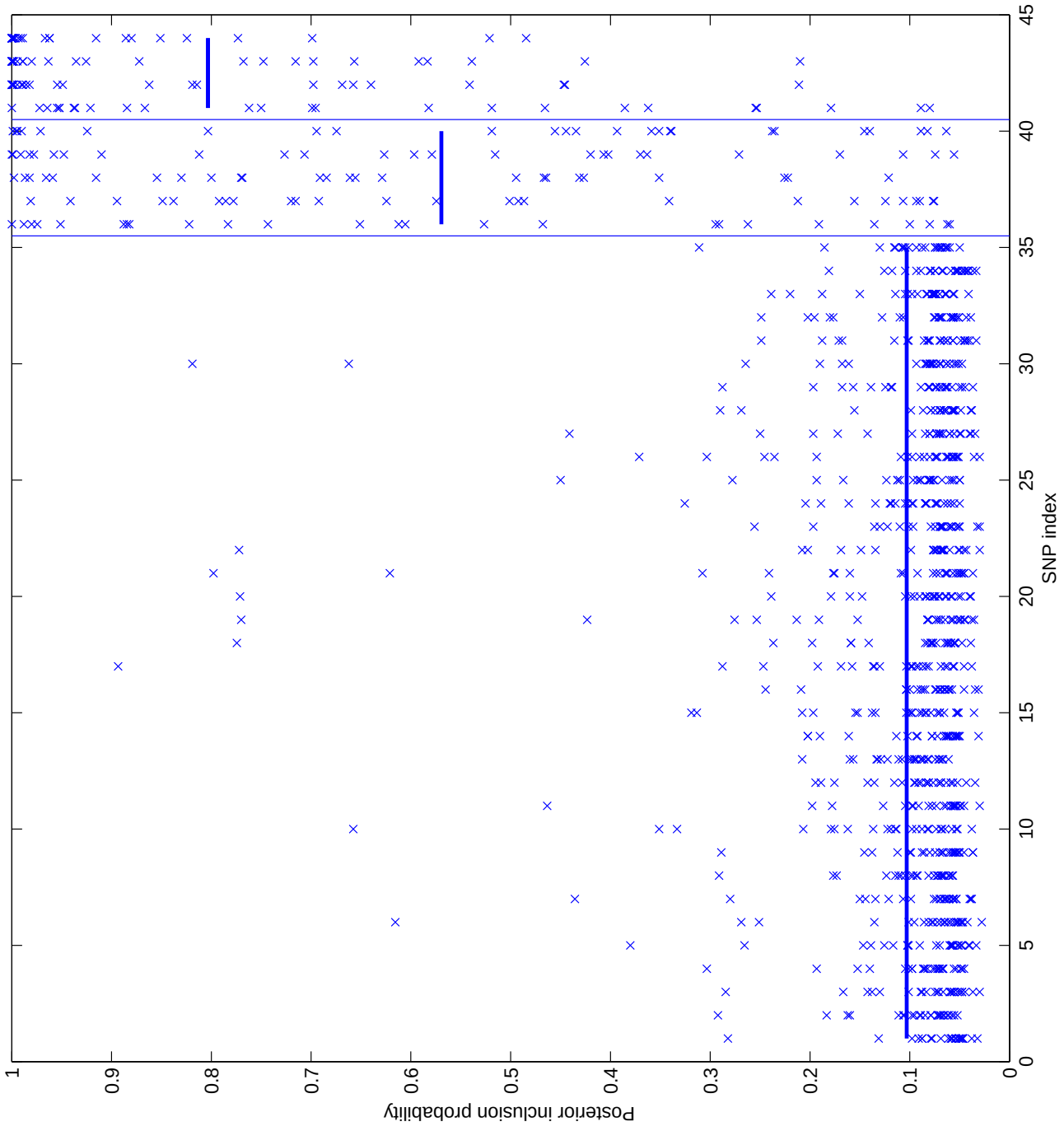
† no women with spontaneous preterm births in this category

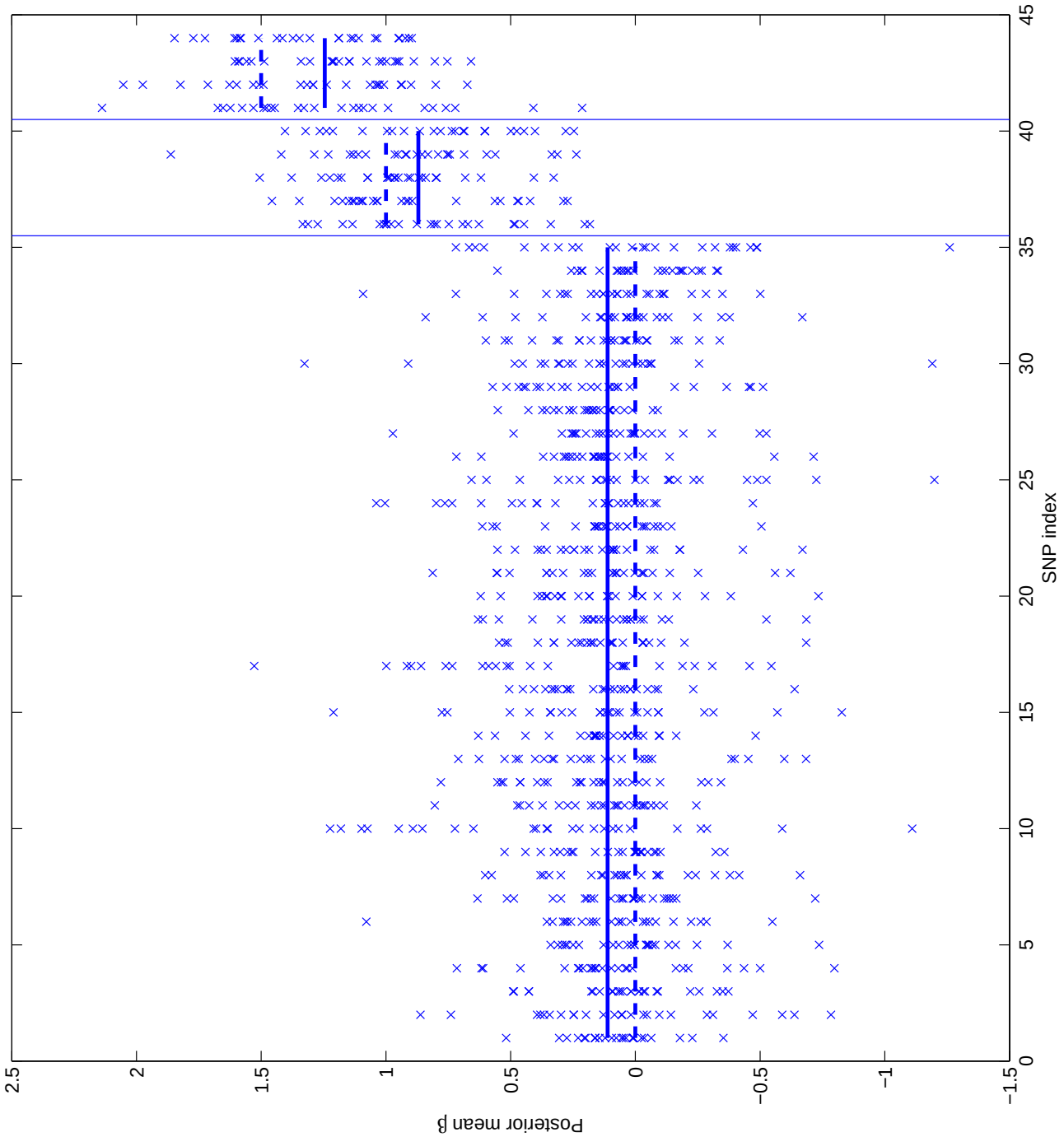
**Table 2**. Simulation results. [$BF$=global Bayes factor in favor of $H_1$, $BF_{c,h}$=local Bayes factor in favor of $H_{1,ch}$, ($\overline{\beta}_-, \overline{BF}_-$)=average coefficients and BFs for negative SNPs]

| | | Summary across simulations | | | |
|---|---|---|---|---|---|
| Case | Quantity | Mean | Median | [25th,75th] | Proportion $> 0$ |
| (i) | $\log BF$ | -1.92 | -2.10 | [-2.31 , -1.68] | 0.04 |
| | $\log \overline{BF}_-$ | -3.11 | -3.25 | [-3.52 , -2.83] | 0.00 |
| | $\overline{\beta}_-$ | -0.028 | -0.014 | [-0.08 , 0.03] | 0.44 |
| (ii) | $\log BF$ | $> 10$ | 3.52 | [0.73, 4.89] | 0.80 |
| | $\log BF_{8,3}$ | $> 10$ | 4.86 | [1.96, 6.03] | 0.96 |
| | $\log \overline{BF}_-$ | -2.51 | -2.53 | [-2.70, -2.27] | 0.00 |
| | $\beta_{8,3}$ | 0.97 | 0.98 | [0.85, 1.08] | 1.00 |
| | $\overline{\beta}_-$ | -0.02 | -0.02 | [-0.06, 0.03] | 0.28 |
| (iii) | $\log BF$ | $> 10$ | $> 10$ | $[> 10, > 10]$ | 1.00 |
| | $\log BF_{8,3}$ | $> 10$ | 8.18 | $[3.14, > 10]$ | 1.00 |
| | $\log BF_{9,1}$ | $> 10$ | 6.80 | $[4.32, > 10]$ | 0.96 |
| | $\log BF_-$ | -2.07 | -2.11 | [-2.27, -1.88] | 0.00 |
| | $\beta_{8,3}$ | 0.98 | 1.01 | [0.79, 1.14] | 1.00 |
| | $\beta_{9,1}$ | 0.99 | 0.96 | [0.87, 1.11] | 1.00 |
| | $\beta_-$ | -0.01 | 0.00 | [-0.05, 0.03] | 0.52 |


White women
cytokine
1
2
3
4
5
6
7
8
9
10
11
SNP−specific BF (x) & estimated β (o)
SNP within cytokine
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
0
5
10
15
20
25
30
35
40
45


Black women
cytokine
1
2
3
4
5
6
7
8
9
10
11
SNP−specific BF (x) & estimated β (o)
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
0
5
10
15
20
25
30
35
40
45
SNP within cytokine


Posterior inclusion probability
SNP index
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
35
40
45


Posterior mean β
SNP index
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
0
5
10
15
20
25
30
35
40
45