

BAYESIAN MODEL SEARCH AND MULTILEVEL INFERENCE FOR SNP ASSOCIATION STUDIES

BY MELANIE A. WILSON^{*,‡} EDWIN S. IVERSEN^{*,‡} MERLISE A. CLYDE^{*,‡}
SCOTT C. SCHMIDLER AND JOELLEN M. SCHILDKRAUT^{*,‡}

Duke University

Technological advances in genotyping have given rise to hypothesis-based association studies of increasing scope. As a result, the scientific hypotheses addressed by these studies have become more complex and more difficult to address using existing analytic methodologies. Obstacles to analysis include inference in the face of multiple comparisons, complications arising from correlations among the SNPs (single nucleotide polymorphisms), choice of their genetic parametrization and missing data. In this paper we present an efficient Bayesian model search strategy that searches over the space of genetic markers and their genetic parametrization. The resulting method for Multilevel Inference of SNP Associations, MISA, allows computation of multilevel posterior probabilities and Bayes factors at the global, gene and SNP level, with the prior distribution on SNP inclusion in the model providing an intrinsic multiplicity correction. We use simulated data sets to characterize MISA's statistical power, and show that MISA has higher power to detect association than standard procedures. Using data from the North Carolina Ovarian Cancer Study (NCOCS), MISA identifies variants that were not identified by standard methods and have been externally 'validated' in independent studies. We examine sensitivity of the NCOCS results to prior choice and method for imputing missing data. MISA is available in an R package on CRAN.

Key words: AIC; Bayes factor; Bayesian model averaging; BIC; Evolutionary Monte Carlo; false discovery; genetic models; lasso; model uncertainty; single nucleotide polymorphism; variable selection.

1. Introduction. Recent advances in genotyping technology have resulted in a dramatic change in the way hypothesis-based genetic association studies are conducted. While previously investigators were limited by

^{*}Partially supported by National Institute of Health grant NIH/NHLBI R01-HL090559.

[†]Partially supported by the National Science Foundation grants DMS-0342172 and DMS-0406115. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[‡]This work was supported by the Duke SPORE in Breast Cancer, P50-CA068438; the North Carolina Ovarian Cancer Study, R01-CA076016.

costs to investigating only a handful of variants within the most interesting genes, researchers may now conduct candidate–gene and candidate–pathway studies that encompass many hundreds or thousands of genetic variants, often single nucleotide polymorphisms (SNPs). For example, the North Carolina Ovarian Cancer Study (NCOCS) (Schildkraut *et al.* [2008]), an ongoing population–based case–control study, genotyped 2129 women at 1536 SNPs in 170 genes on 8 pathways, where ‘pathway’ is defined as a set of genes thought to be simultaneously active in certain circumstances.

The analytic procedure most commonly applied to association studies of this scale is to fit a separate model of association for each SNP that adjusts for design and confounder variables. As false discoveries due to multiple testing are often a concern, the level of significance for each marginal test of association is adjusted using Bonferroni or other forms of false discovery correction [Storey 2002, Wacholder 2004, Balding 2006]. While these methods have been shown to be effective in controlling the number of false discoveries reported, correlations between the markers may limit the power to detect true associations [Efron 2007]. The NCOCS study provides a case-in-point. When simple marginal methods are applied to the NCOCS data, no SNPs are identified as notable.

Marginal SNP-at-a-time methods do not address directly many of the scientific questions in candidate pathway studies, such as ‘Is there an overall association between a pathway and the outcome of interest?’ and ‘Which genes are most likely to be driving this association?’ The Multilevel Inference for SNP Association (MISA) method we describe here is designed to simultaneously address these questions of association at the level of SNP, gene, and pathway.

MISA, in contrast to the marginal methods, identifies ten SNPs of interest in the NCOCS study. To date, one of these (ranked tenth by MISA) has been validated in external data by a large multi–center consortium [Schildkraut *et al.* 2009]; additional testing is underway for other top SNPs discovered by MISA. To buttress this empirical evidence, we demonstrate using simulation studies (Section 4) that MISA has higher power to detect associations than other simpler procedures, with a modest increase in the false discovery rate (Figure 1).

In the next section, we describe the Bayesian hierarchical model behind MISA and highlight how it addresses many of the key issues in analysis of SNP association studies: identification of associated SNPs and genetic models, missing data, inference for multi–level hypotheses and control of the false discovery rate. Like stepwise logistic regression [Balding 2006], lasso [Park and Hastie 2008, Shi *et al.* 2007, Wu *et al.* 2009] and logic regression

[Ruczinski *et al.* 2003, Kooperberg and Ruczinski 2004, Schwender and Ickstadt 2007], MISA improves upon marginal, SNP-at-a-time methods by modeling the outcome variable as a function of a multivariate genetic profile, which provides measures of association that are adjusted for the remaining markers. MISA uses Bayesian Model Averaging [Hoeting *et al.* 1999] to combine information from multiple models of association to address the degree to which the data support an association at the level of individual SNPs, genes, and pathways, while taking into account uncertainty regarding the best genetic parametrization. By using model averaging, MISA improves upon methods that select a single model, which may miss important SNPs because of LD structure. We show how the prior distribution on SNP inclusion provides a built-in multiplicity correction. Because missing data are a common phenomenon in association studies, we discuss two options for handling this problem.

In Section 3, we present an Evolutionary Monte Carlo algorithm to efficiently sample models of association according to their posterior probabilities. In Section 4 we apply our method to simulated data sets and demonstrate that MISA outperforms less complex and more commonly used alternatives for detecting associations in modestly powered candidate-gene case-control studies. The simulation approach may also be used to guide selection of the prior hyper-parameters given the study design. In Section 5 we return to the NCOCS study and present results from the analysis of a single pathway from that study. We examine the sensitivity of results to prior hyperparameter choice and methods for imputing missing data. We conclude in Section 6 with recommendations and a discussion of future extensions.

2. Models of Association. We consider SNP association models with a binary phenotype, such as presence or absence of a disease as in case-control designs. For $i = 1, \dots, n$, let D_i indicate the disease status of individual i , where $D_i = 1$ represents a disease case and $D_i = 0$ represents a control. For each individual, we have S SNP measurements, where SNP s is either homozygous common ($A_s A_s$), heterozygous ($a_s A_s$ or $A_s a_s$), homozygous rare ($a_s a_s$), or missing and is coded as 0, 1, 2, representing the number of rare alleles, or NA if the SNP is missing for that individual. We will discuss methods for imputing missing SNP data in Section 2.3. In addition to the SNP data, for each individual we have a q -dimensional vector \mathbf{z}_i^T of design and potential confounding variables that will be included in all models, henceforth referred to as ‘design’ variables.

We use logistic regression models to relate disease status to the design variables and subsets of SNPs. We denote the collection of all possible models

by \mathcal{M} . An individual model, denoted by \mathcal{M}_γ , is specified by the S dimensional vector γ , where γ_s indicates the inclusion and SNP-specific genetic parametrization of SNP s in model \mathcal{M}_γ : $\gamma_s = 0$ if $\text{SNP}_s \notin \mathcal{M}_\gamma$, $\gamma_s = 1$ if $\text{SNP}_s \in \mathcal{M}_\gamma$ with a log-additive parametrization, $\gamma_s = 2$ if $\text{SNP}_s \in \mathcal{M}_\gamma$ with a dominant parametrization, and $\gamma_s = 3$ if $\text{SNP}_s \in \mathcal{M}_\gamma$ with a recessive parametrization. When no homozygous rare cases or controls are observed, we fix the genetic parametrization to be log-additive. Under each of these genetic parametrizations, SNP s may be encoded using one degree of freedom. In particular, for the log-additive model, the design variable representing SNP s is a numeric variable equal to the number of copies of the risk allele a_s . For the dominant model, we use an indicator variable of whether allele a_s is present (homozygous rare or heterozygous) and for the recessive model, an indicator variable of whether SNP s has the homozygous rare genotype. For each individual, the logistic regression under model \mathcal{M}_γ assuming complete data is given by

$$(2.1) \quad \text{logit}(p(D_i = 1 | \mathbf{z}_i, \mathbf{x}_{\gamma_i}, \boldsymbol{\theta}_\gamma, \mathcal{M}_\gamma)) = \alpha_0 + \mathbf{z}_i^T \boldsymbol{\alpha} + \mathbf{x}_{\gamma_i}^T \boldsymbol{\beta}_\gamma$$

where \mathbf{x}_{γ_i} represents the coding of SNPs in model \mathcal{M}_γ and $\boldsymbol{\theta}_\gamma$ is the vector of model specific parameters $(\alpha_0, \boldsymbol{\alpha}^T, \boldsymbol{\beta}_\gamma^T)$, with intercept α_0 , vector of design variable coefficients $\boldsymbol{\alpha}$, and log-odds ratios $\boldsymbol{\beta}_\gamma$. Prospective models for disease outcome given multivariate genetic marker data as in equation (2.1) provide measures of association that are adjusted for other markers which can increase the power to detect associations [Balding 2006], however, one is faced with an extremely large collection of possible models. While stepwise selection methods may be used to select a single model [Cordell and Clayton 2002], this leads to difficulty in interpreting the significance of SNPs in the selected model. Bayesian model averaging is an alternative to stepwise selection methods and is an effective approach for identifying subsets of likely associated variables, for prioritizing them and for measuring overall association in the presence of model uncertainty (see the review articles by Hoeting *et al.* [1999] and Clyde and George [2004] and references therein).

2.1. *Posterior Inference.* Posterior model probabilities measure the degree to which the data support each model in a set of competing models. The posterior model probability of any model \mathcal{M}_γ in the space of models \mathcal{M} is expressed as

$$p(\mathcal{M}_\gamma | D) = \frac{p(D | \mathcal{M}_\gamma)p(\mathcal{M}_\gamma)}{\sum_{\mathcal{M}_\gamma \in \mathcal{M}} p(D | \mathcal{M}_\gamma)p(\mathcal{M}_\gamma)} \quad \text{for } \mathcal{M}_\gamma \in \mathcal{M}$$

where $p(D | \mathcal{M}_\gamma)$ is the (marginal) likelihood of model \mathcal{M}_γ obtained after integrating out model-specific parameters θ_γ with respect to their prior distribution, and $p(\mathcal{M}_\gamma)$ is the prior probability of \mathcal{M}_γ .

While posterior probabilities provide a measure of evidence for hypotheses or models, it is often difficult to judge them in isolation as individual model probabilities may be “diluted” as the space of models grows [Clyde 1999, George 1999, Clyde and George 2004]. Bayes factors (BF) [Kass and Raftery 1995] compare the posterior odds of any two models (or hypotheses) to their prior odds

$$\text{BF}(\mathcal{M}_{\gamma_1} : \mathcal{M}_{\gamma_2}) = \frac{p(\mathcal{M}_{\gamma_1} | D)/p(\mathcal{M}_{\gamma_2} | D)}{p(\mathcal{M}_{\gamma_1})/p(\mathcal{M}_{\gamma_2})}$$

and measures the *change* in evidence (on the log scale) provided by data for one model, \mathcal{M}_{γ_1} , to another, \mathcal{M}_{γ_2} or for pairs of hypotheses. Goodman [1999] and Stephens and Balding [2009] provide a discussion on the usefulness of Bayes factors in the medical context and Wakefield [2007] illustrates their use in controlling false discoveries in genetic epidemiology studies. Below we define Bayes factors for quantifying association at multiple levels (global, gene, and SNP) and assessing the most likely SNP-specific genetic parametrization.

2.1.1. *Global Bayes Factor.* The Bayes factor in favor of H_A , the alternative hypothesis that there is at least one SNP associated with disease, to H_0 , the null hypothesis that there is no association between the SNPs under consideration and disease, measures the relative weight of evidence of H_A to H_0 . The null model corresponding to H_0 is the model which includes only design variables and no SNPs, and is denoted \mathcal{M}_0 . The alternative hypothesis is represented by all of the remaining models in \mathcal{M} . Because the space of models is large, the null model (or any single model in general) may receive small probability (both prior and posterior), even when it is the highest posterior probability model (this illustrates the dilution effect of large model spaces); Bayes factors allow one to judge how the posterior odds compare to one’s prior odds.

The Global Bayes factor for comparing H_A to H_0 may be simplified to

$$(2.2) \quad \text{BF}(H_A : H_0) = \sum_{\mathcal{M}_\gamma \in \mathcal{M}} \text{BF}(\mathcal{M}_\gamma : \mathcal{M}_0) p(\mathcal{M}_\gamma | H_A)$$

which is the weighted average of the individual Bayes factors $\text{BF}(\mathcal{M}_\gamma : \mathcal{M}_0)$ for comparing each model in H_A to the null model with weights given by the prior probability of \mathcal{M}_γ conditional on being in H_A , $p(\mathcal{M}_\gamma | H_A)$. Because

the alternative is a composite hypothesis, the resulting Global Bayes factor is not independent of the prior distribution on the models that comprise the alternative, thus the prior distribution on models will play an important role in controlling the (relative) weights that models of different sizes receive. For a large number of SNPs, it is impossible to enumerate the space of models and posterior summaries are often based on models sampled from the posterior distribution. In equation (2.2), if we replace the average over all models in H_A with the average over the models in \mathcal{S} (the collection of unique models sampled from the posterior distribution), the result

$$\text{BF}(H_A : H_0) > \text{BF}_{\mathcal{S}}(H_A : H_0) \equiv \sum_{\mathcal{M}_{\gamma} \in \mathcal{S}} \text{BF}(\mathcal{M}_{\gamma} : \mathcal{M}_0) p(\mathcal{M}_{\gamma} | H_A)$$

is a lower bound for the Bayes factor for testing global association. If the lower bound indicates evidence of an association, then we can be confident that this evidence will only increase as we include more models.

2.1.2. *SNP Bayes Factors.* While it is of interest to quantify association at the global level, interest is primarily in identifying the gene(s) and variant(s) within those genes that drive the association. We begin by defining SNP inclusion probabilities and associated Bayes factors. These marginal summaries are adjusted for the other potentially important SNPs and confounding variables and provide a measure of the strength of association at the level of individual SNPs. Given each sampled model $\mathcal{M}_{\gamma} \in \mathcal{S}$ and the model specification vectors $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_S)$ previously defined in Section 2, the inclusion probability for SNP s is estimated as:

$$(2.3) \quad p(\gamma_s \neq 0 | D) = \sum_{\mathcal{M}_{\gamma} \in \mathcal{S}} 1_{(\gamma_s \neq 0)} p(\mathcal{M}_{\gamma} | D, \mathcal{S})$$

where $p(\mathcal{M}_{\gamma} | D, \mathcal{S})$ is the posterior probability of a model re-normalized over the sampled model space. The SNP Bayes factor is the ratio of the posterior odds of the SNP being associated to the prior odds of the same, and is defined as:

$$\text{BF}(\gamma_s \neq 0 : \gamma_s = 0) = \frac{p(\gamma_s \neq 0 | D)}{p(\gamma_s = 0 | D)} \div \frac{p(\gamma_s \neq 0)}{p(\gamma_s = 0)};$$

where $p(\gamma_s \neq 0)$ is the prior probability of SNP s being associated. Estimates of the SNP Bayes Factor may be obtained using the estimated SNP inclusion probabilities from (2.3).

2.1.3. *Gene Bayes Factors.* In cases where there are SNPs in Linkage Disequilibrium (LD), SNP inclusion probabilities may underestimate the significance of an association at a given locus. This occurs because SNPs in LD may provide competing explanations for the association, thereby diluting or distributing the probability over several markers. Since the amount of correlation between markers across different genes is typically negligible, calculating inclusion probabilities and Bayes factors at the gene level will not be as sensitive to this dilution. A gene is defined to be associated if one or more of the SNPs within the given gene are associated. Hence we define the gene inclusion probability as:

$$p(\Gamma_g = 1 \mid D) = \sum_{\mathcal{M}_\gamma \in \mathcal{S}} 1_{(\Gamma_g=1)} p(\mathcal{M}_\gamma \mid D, \mathcal{S});$$

where $\Gamma_g = 1$ if at least one SNP in gene g is in model \mathcal{M}_γ and is zero otherwise. The gene Bayes factor is defined as:

$$\text{BF}(\Gamma_g = 1 : \Gamma_g = 0) = \frac{p(\Gamma_g = 1 \mid D)}{p(\Gamma_g = 0 \mid D)} \div \frac{p(\Gamma_g = 1)}{p(\Gamma_g = 0)};$$

where $p(\Gamma_g = 1)$ is the prior probability of one or more SNPs in gene g being associated.

2.1.4. *Interpreting Evidence.* [Jeffreys \[1961, page 432\]](#) presents a descriptive classification of Bayes factors into "grades of evidence" (reproduced in [Table 1](#)) to assist in their interpretation (see [Kass and Raftery \[1995\]](#)). In the context in which he presents the grades, he defined the Bayes factor assuming equal prior odds, making it equivalent to posterior odds and enabling a meaningful interpretation in terms of probabilities. It is not clear whether he intended his descriptive grades to be used more broadly for interpreting Bayes factors or for interpreting posterior probabilities.

Jeffreys was well aware of the issues that arise with testing several simple alternative hypotheses against a null hypothesis [[Jeffreys 1961, Section 5.04](#)], noting that if one were to test several hypotheses separately that by chance one might find one of the Bayes factors to be less than one even if all null hypotheses were true. He suggested that, in this context, the Bayes factors needed to be "corrected for selection of hypotheses" by multiplying by the prior odds.

Experience has shown that detectable SNP associations are relatively infrequent, hence the prior odds of any given SNP being marginally associated in the typical genetic association study should be small. For this reason, [Stephens and Balding \[2009\]](#) suggest that marginal Bayes factors calculated

assuming equal prior odds be interpreted in light of a prior odds more appropriate to the study at hand. Our approach to the problem of exploring multiple hypotheses is to embed each of the potential submodels (corresponding to a subset of SNPs) into a single hierarchical model. Unlike the marginal (one-at-a-time) Bayes factors in [Stephens and Balding \[2009\]](#) that are independent of the prior odds on the hypotheses, our SNP Bayes factors are based on comparing composite hypotheses and hence do depend on the prior distribution over models, which implicitly adjusts for the selection of hypotheses.

While Bayes factors do not provide a measure of *absolute* support for or against a hypothesis (except with even prior odds), the log Bayes factor does provide a coherent measure of how much the data *change* the support for the hypothesis (relative to the prior) [[Lavine and Schervish 1997](#)]. Applying Jeffreys grades to Bayes factors using priors distributions that account for competing hypotheses provides an idea of the impact of the data on changing prior beliefs, but ultimately posterior odds provide a more informative measure of evidence and model uncertainty.

Grade	BF($H_A : H_0$)	Evidence against H_0
1	1 to 3.2	Indeterminate
2	3.2 to 10	Positive
3	10 to 31.6	Strong
4	31.6 to 100	Very Strong
5	> 100	Decisive

TABLE 1
Jeffrey's grades of evidence [[Jeffreys 1961](#), page 432].

2.2. Prior Distributions, Laplace Approximations and Marginal Likelihoods. We assume normal prior distributions for the coefficients θ_γ with a covariance matrix that is given by a constant $1/k$ times the inverse Fisher Information matrix. For logistic regression models, analytic expressions for $p(D | \mathcal{M}_\gamma)$ are not available and Laplace approximations or the Bayes Information Criterion are commonly used to approximate the marginal likelihood [[Raftery 1986](#), [Wakefield 2007](#), [Burton et al. 2007](#)]. Using a Laplace approximation with the normal prior distribution (see Appendix A), the posterior probability of model \mathcal{M}_γ takes the form of a penalized likelihood

$$(2.4) \quad p(\mathcal{M}_\gamma | D) \propto \exp\left\{-\frac{1}{2}[\text{dev}(\mathcal{M}_\gamma; D) + \text{pen}(\mathcal{M}_\gamma)]\right\}$$

where $\text{dev}(\mathcal{M}_\gamma; D) = -2 \log(p(D | \hat{\theta}_\gamma, \mathcal{M}_\gamma))$ is the model deviance, and the penalty term $\text{pen}(\mathcal{M}_\gamma)$ encompasses a penalty on model size induced

by the choice of k in the prior distribution on coefficients θ_γ and the prior distribution over models. Because we expect that effect sizes will be small, we calibrate the choice of k based on the Akaike information criterion (Appendix A), leading to

$$\text{pen}(\mathcal{M}_\gamma) = 2(1 + q + s_\gamma) - 2 \log(p(\mathcal{M}_\gamma)).$$

2.3. Missing Data. The expression in (2.4) assumes complete data on all SNPs. Missing SNP data, unfortunately, are the norm rather than the exception in association studies. Removing all subjects with any missing SNP genotype data will typically result in an unnecessary loss of information and potential bias of estimated effects if the missing data are non-ignorable. It is possible, however, to exploit patterns in LD to efficiently impute the missing genotypes given observed data [Balding 2006]. We use fastPHASE [Stephens *et al.* 2001, Servin and Stephens 2007] to sample haplotypes and missing genotypes (G^m) given the observed unphased genotypes (G^o). This assumes that the pattern of missing data is independent of case-control status, which, if not true may lead to serious biases [Clayton *et al.* 2005]. This assumption may be examined by using indicator variables of missingness as predictors in MISA.

The posterior probabilities of models given the data are obtained by averaging the marginal likelihood of a model over imputed genotype data:

$$\begin{aligned} p(\mathcal{M}_\gamma | D) &\propto \int \exp\left\{-\frac{1}{2}[\text{dev}(\mathcal{M}_\gamma; D, G^o, G^m) + \text{pen}(\mathcal{M}_\gamma)]\right\} p(G^m | G^o) dG^m \\ (2.5) \quad &\approx \frac{1}{M} \sum_{i=1}^I \exp\left\{-\frac{1}{2}[\text{dev}(\mathcal{M}_\gamma; G^o, G_i^m) + \text{pen}(\mathcal{M}_\gamma)]\right\} \equiv \Psi(\mathcal{M}_\gamma) \end{aligned}$$

where I is the number of imputed data sets, $\text{dev}(\mathcal{M}_\gamma; D, G^o, G^m)$ is the deviance based on the completed data, and $\Psi(\mathcal{M}_\gamma)$ is an estimate of the un-normalized posterior model probability for model \mathcal{M}_γ . We have found that the number of imputed sets must be on the order of $I = 100$ to provide accurate estimates of posterior quantities. This has a significant computational impact in the model search algorithm described in Section 3. As a simple alternative, we approximate (2.5) by a modal approximation, where the missing genotypes are imputed with the mode of the sampled genotypes using fastPHASE. While it is well known that plugging in a single estimate for the missing data under-estimates uncertainty, the modal approximation provides dramatic computational savings. In Section 5 we examine the sensitivity of results to the method of imputing missing data and find that the modal approximation gives comparable results for SNP BF's.

2.4. *Choice of Prior Distribution on Models.* The prior distribution on the space of models \mathcal{M} , $p(\mathcal{M}_\gamma)$, completes our model specification. The frequentist approach for SNP association studies usually involves some form of adjustment for multiple-testing, which can, in effect, penalize the researcher who looks beyond single-SNP models of association to multiple SNP models or models of interactions. Under the Bayesian approach, posterior evidence in the data is judged against the prior odds of an association using Bayes factors, which should not be affected by the number of tests that an investigator chooses to carry out [Balding 2006].

While it has been common practice to adopt a “non-informative” uniform distribution over the space of models for association (this is after marginalizing over the possible genetic models for each SNP), this choice has the potentially undesirable “informative” implication that $\frac{1}{2}$ of the SNPs are expected to be associated *a priori*, and the prior odds of at least one SNP being included (which is used in the global Bayes factor) depends on the number of tests (2^S) (Table 2).

A recommended alternative is the Beta-Binomial distribution on the model size, which provides over-dispersion, added robustness to prior misspecification, and multiplicity corrections as a function of the number of variables [Ley and Steel 2009, Scott and Berger 2008, Cui and George 2008]. We construct a hierarchical prior distribution over the space of models defined by subsets of SNPs and their genetic parametrizations as follows. For any SNP included in the model, we assign a uniform distribution over the possible genetic parametrizations. The prior distribution on the model size s_γ is $\text{Bin}(S, \rho)$ conditional on ρ , and for the last stage, ρ is assigned a $\text{Beta}(a, b)$ distribution. Integrating over the distribution on ρ , leads to the $\text{BetaBinomial}(a, b)$ distribution on model size,

$$(2.6) \quad p(s_\gamma) = \frac{B(s_\gamma + a, S - s_\gamma + b)}{(S + 1)B(s_\gamma + 1, S - s_\gamma + 1)B(a, b)}$$

and the following distribution on models,

$$(2.7) \quad p(\mathcal{M}_\gamma) = \left(\frac{1}{3}\right)^{s_\gamma} \frac{B(s_\gamma + a, S - s_\gamma + b)}{B(a, b)}$$

where $B(\cdot, \cdot)$ is the beta function and the factor of $1/3$ accounts for the distribution over genetic parametrizations.

2.4.1. *Default Hyper-Parameter Choice.* Following Ley and Steel [2009] and Scott and Berger [2008], we recommend $a = 1$ as a default, so that the prior distribution on model size is non-increasing in s_γ . The hyper-parameter b can then be chosen to reflect the expected model size, global

prior probability of at least one association, or the marginal prior odds that any SNP is associated (Table 2). A default choice is to set $b = 1$, leading to

	Binomial ($S, 1/2$)	Beta-Binomial ($1, 1$)	Beta-Binomial ($1, \lambda S$)
Expected Model Size	$\frac{S}{2}$ (∞)	$\frac{S}{2}$ (∞)	$\frac{S}{\lambda S + 1}$ ($\frac{1}{\lambda}$)
Global Prior Odds of an Association	$\frac{2^{2S}}{2^S + 1}$ (∞)	S (∞)	$\frac{1}{\lambda}$
Marginal Prior Odds of an Association	1	1	$\frac{1}{\lambda S}$ (0)
Prior Odds of Adding a Variable	1	$\frac{s_\gamma + 1}{S - s_\gamma}$ (0)	$\frac{s_\gamma + 1}{(\lambda + 1)S - s_\gamma - 1}$ (0)

TABLE 2

General prior characteristics and limiting behavior (in parentheses) of the $\text{Bin}(S, 1/2)$, $\text{BetaBinomial}(1, 1)$ and $\text{BetaBinomial}(1, \lambda S)$ distribution on model size.

a uniform distribution on model size [Ley and Steel 2009, Scott and Berger 2008]. Like the binomial distribution, the $\text{BetaBinomial}(1, 1)$ distribution results in an expected model size of $\frac{S}{2}$ (Table 2), although the $\text{BetaBinomial}(1, 1)$ distribution has a larger variance than the $\text{Bin}(S, 1/2)$. Alternatively, if b is proportional to S , $b = \lambda S$ the expected model size approaches a limit of $\frac{1}{\lambda}$ as S approaches infinity.

The choices for hyperparameters have implications for the global Bayes factor. The $\text{BetaBinomial}(1, 1)$ has a global prior odds of association equal to the number of SNPs, S , and would be appropriate for the case where increasing the number of SNPs under consideration reflects increased prior certainty that an overall (global) association can be detected. Under the $\text{BetaBinomial}(1, \lambda S)$, the global prior odds are constant, $1/\lambda$, reflecting a prior odds for overall association that is independent of the number of genes/SNPs tagged. Also, with both Beta-Binomial prior distributions, the prior odds of incorporating an additional SNP in any model decreases with model size s_γ and approaches 0 in the limiting case as the number of SNPs, S , increases. This provides an implicit multiple testing correction in the number of SNPs (rather than tests) that are included in the study of interest. The $\text{BetaBinomial}(1, \lambda S)$ achieves this by keeping the global (pathway) prior odds of an association constant while decreasing the marginal prior odds of any one of the SNPs being associated as the number of SNPs increases. As a skeptical “default” prior, we suggest the hyper-parameters $a = 1$ and $b = S$ which leads to the global prior odds of there being at least one association

of 1 and the marginal prior odds of any single SNP being associated of $1/S$.

3. Stochastic Search for SNPs. Given the number of SNPs under consideration, enumeration of all models for S greater than 25–30 is intractable. While it is possible to enumerate all single variable SNP models, the number of models with 2 or 3 SNPs allowing for multiple genetic parametrizations is in the millions or more for a typical modern hypothesis-oriented study. Stochastic variable selection algorithms [see [Clyde and George 2004](#), for a review] provide a more robust search procedure than stepwise methods, but also permit calculation of posterior probabilities and Bayes factors based on a sample of the most likely candidate models from the posterior distribution.

MISA makes use of a stochastic search algorithm based on the Evolutionary Monte Carlo (EMC) algorithm of [Liang and Wong \[2000\]](#). EMC is a combination of parallel tempering [[Geyer 1991](#)] and a genetic algorithm [[Holland 1975](#)] and samples models based on their “fitness”. While originally designed to find optimal models based on AIC, in our application the fitness of the models is given by $\psi(\mathcal{M}_\gamma)$

$$\psi(\mathcal{M}_\gamma) = \log(\Psi(\mathcal{M}_\gamma))$$

where $\Psi(\mathcal{M}_\gamma)$ is defined in equation (2.5) and is equal to the log of the unnormalized posterior model probability. This results in models being generated according to their posterior probability.

The EMC algorithm requires that we specify the number of parallel chains that are run and the associated temperature for each chain that determines the degree of annealing. If the temperatures are too spread out for the number of chains, then the algorithm may exhibit poor mixing and slow convergence. [Liang and Wong \[2000\]](#) show that even with all chains run at a temperature of 1 (no annealing), EMC outperforms alternative sampling methods such as Gibbs sampling and Reversible Jump MCMC in problems where strong correlations among the predictor variables lead to problems with exploring multiple modes in the posterior distribution. We have found that a constant temperature ladder with 5 parallel chains provides good mixing and finds more unique models than using a custom temperature ladder based on the prescription in [Liang and Wong \[2000\]](#), and recommend the constant temperature ladder as a default. To assess convergence, we take two independent EMC runs using randomly chosen starting points and examine trace plots of the fitness function. We use the marginal likelihoods from the set of unique models in the sample for inference and compute estimates of marginal posterior inclusion probabilities for each run. We continue

running the two instances of the EMC algorithm until the posterior probabilities derived from each are sufficiently close. This leads to longer running times than those suggested by conventional convergence diagnostic such as Gelman-Rubin [Gelman and Rubin 1992].

Efficiency of stochastic algorithms often diminishes as the total number of models increases. For this reason, we have found it useful to reduce the number of SNPs included in the EMC search using a screen when S is large. Such a screen will typically be fairly permissive, leaving only the weakest candidates out of the stochastic search. The screen should be quick to calculate, adjust for the same design variables and consider the same genetic parametrizations as in the full analysis. In our analyses, we calculated marginal (i.e. SNP-at-a-time) Bayes Factors for each of the log-additive, dominant and recessive models of association against the model of no association. We ordered SNPs according to the maximum of the three marginal Bayes factors and retained those with a maximum marginal BF greater than or equal to one. More details are available in Appendix B.

4. Simulation Comparison. We used the 124 simulated case – control data sets, details of the simulation can be found in Appendix C, to estimate true and false positive rates for MISA and seven other alternative procedures:

Bonferroni We fit a logistic regression model for each SNP under the log-additive parametrization and calculate the p-value for testing association using a Chi-Squared test. We use a Bonferroni corrected level $\alpha = 0.05$ test to declare a SNP associated.

Adjusted Bonferroni We fit a logistic regression model for each SNP under the log-additive parametrization and calculate the p-value for testing association using a Chi-Squared test. We use a Bonferroni corrected level α test to declare a SNP associated where α is chosen so that the proportion of false positives detected is the same as in MISA default (1) and custom (2).

Benjamini-Hochberg We fit the same SNP-at a time logistic regression as above, but declare a SNP to be associated if it has a Benjamini-Hochberg false discovery rate of less than 0.05.

Marginal BF This also utilizes the single SNP at a time logistic regression, but calculates a BF for association under each of the three genetic models. If the maximum BF over the three genetic models is greater than 3.2, we declare the SNP associated. See Appendix C for more detail.

Stepwise LR (AIC) We use a stepwise multiple logistic regression proce-

ture to select SNPs based on AIC. Each SNP is coded using 2 degrees of freedom to select among the three genetic models. SNPs in the final model are called associated.

Stepwise LR (BIC) Same as above but using BIC to select models.

Lasso We use the `Lasso2` package in R [Lokhorst *et al.* 2009] that is based on the algorithm developed by Osborne *et al.* [1999] to select SNPs based on the least absolute shrinkage and selection operator. Each SNP is coded using 2 degrees of freedom to represent the three genetic models and all SNPs in the final model with coefficients greater than zero are called associated.

MISA We reduced the number of SNPs using the marginal Bayes factor method above to eliminate SNPs with a marginal BF ≥ 1 . We ran MISA using the default `BetaBinomial(1, S)` prior distribution on the models using two runs of 400,000 iterations based on convergence of the marginal inclusion probabilities. SNPs are called associated if their MISA SNP BF is greater than 3.2. All SNPs that did not pass the marginal screen step in MISA were declared not associated.

The first four are single SNP methods, while the last three are multi-SNP methods that take into account the genetic parametrization for each SNP.

Figure 1 shows the proportion of SNPs detected by each of the methods as a function of the assumed true odds ratio. Thus, at an odds ratio of 1.00 we plot the proportion of SNPs that were falsely declared associated by each of the methods. While both Bonferroni and Benjamini-Hochberg have the smallest false positive rates, they have much lower power to detect true associations than any of the other methods; the marginal BF has the highest power out of the three marginal methods, and is comparable to lasso, a multi-SNP method. Stepwise model selection using BIC has the lowest power of the multiple SNP model selection procedures. Stepwise logistic regression using AIC to select a model, on the other hand, has high power to detect associations, but an unacceptably high false positive rate (44%). With the exception of stepwise/AIC, the MISA methods have higher power than the alternatives at all odds ratios (ORs) in the simulation, with the gain in power most noticeable for the smaller ORs, those encompassing the range, 1.25 – 1.75 typically seen in practice [Flint and Mackay 2009]. This increase in power comes at the cost of only a slight increase in the false positive rate. Overall, MISA using the default `BetaBinomial(1, S)` prior distribution is able to detect 9% as many associations at the SNP level and 13% as many at the gene level than the marginal BF method used alone. In addition, MISA is able to detect 19% as many true associations at the SNP level and 27% as many at the gene level as the calibrated Bonferroni method (the two

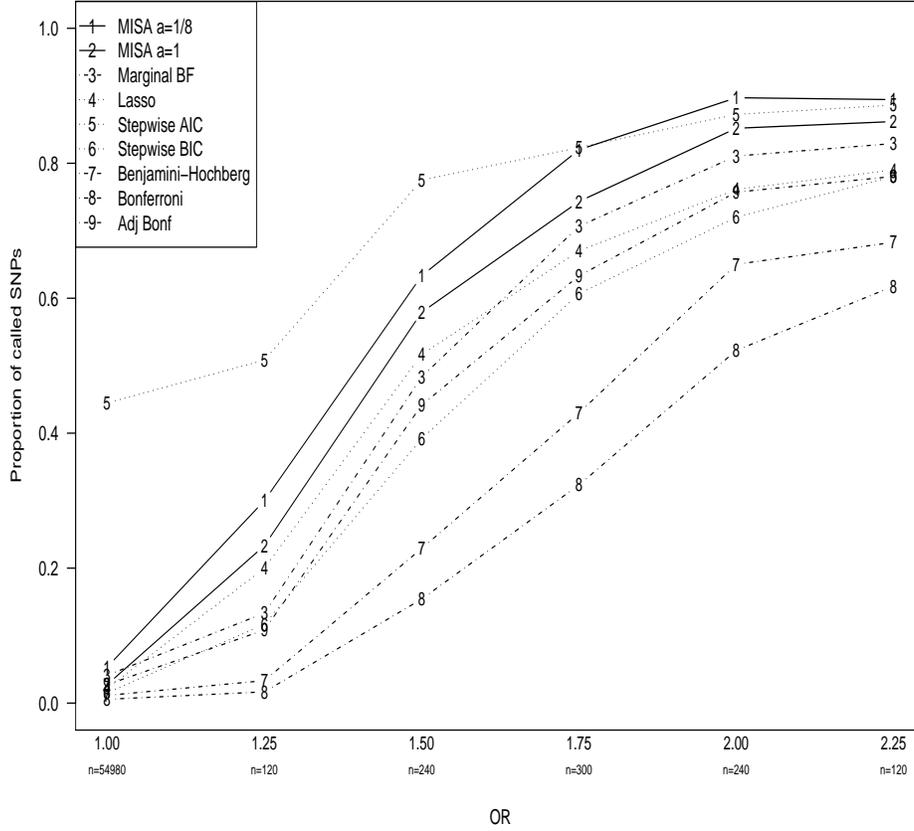


FIG 1. True and False positive rates of MISA versus alternative methods.

methods have the same Type I error rate).

4.1. *Sensitivity to Hyperparameters.* We examined a range of parameters (a and b) for the Beta-Binomial prior distribution on model size (Table 3) to assess sensitivity of true positive and false positive rates. In practice, this may be done by reweighting the MCMC output using the new prior distribution, without resorting to additional MCMC runs, as long as high posterior probability models receive adequate support under both prior distributions.

Over the range of values for (a, b) , MISA has a higher gene and SNP true positive rate than any of the other simpler procedures, with the exception

of Stepwise AIC. In general, decreasing a leads to higher true positive rates, but at the expense of higher false positive rates. The SNP false positive rate is modest, ranging from 0.025 to 0.099, providing effective control of the experiment wide error rate. While these rates are higher than the false positive rates under Bonferroni or Benjamini-Hochberg, eliminating a SNP from consideration that truly is associated has a higher scientific cost than continuing to collect data to confirm that a SNP is really a null finding. Because the NCOCS will follow-up apparent associations, a higher true positive rate with a modest increase in false positives was preferable.

The hyper-parameters $a = 1/8$ and $b = S$, highlighted in bold in Table 3 were selected for comparison with the default choice ($a = 1$, $b = S$) in the analysis of the NCOCS data presented in the next section. MISA using the `BetaBinomial(1/8, S)` is able to detect 19% as many true associations at the SNP level and 26% as many at the gene level as the marginal BF method used alone. In addition, MISA with the `BetaBinomial(1/8, S)` prior is able to detect 14% as many true associations at the SNP level and 24% as many at the gene level as a calibrated Bonferroni method, (the two methods have the same Type I error rate).

5. Ovarian Cancer Association Analysis. In this section, we describe a MISA candidate pathway analysis of data from the ongoing NCOCS ovarian cancer case-control association study. The NCOCS is a population based study that covers a 48 county region of North Carolina [Schildkraut *et al.* 2008]. Cases are between 20 and 74 years of age and were diagnosed with primary invasive or borderline epithelial ovarian cancer after January 1, 1999. Controls are frequency matched to the cases by age and race and have no previous diagnosis of ovarian cancer. In the analysis we present, we focus on self-reported Caucasians and a specific histological subtype of the cancer, leaving us a total of 397 cases and 787 controls. Because the ovarian cancer results have not yet been published, we have anonymized the pathway, the genes chosen to represent it and the IDs of the SNPs tagging variation in those genes. The pathway is comprised of 53 genes tagged by 508 tag SNPs.

All models fit in the screen and by MISA included the patient's age as a design variable. We used the modal approximation to fill in missing SNP data. We screened 508 SNPs using marginal Bayes factors, retaining $S = 70$ SNPs that exceeded the threshold of 1 in favor of an association. Using the default hyperparameters $a = 1$ and $b = S$, we ran two independent runs of the algorithm from independent starting points for a total of 1.2 million iterations — the point at which the SNP marginal inclusion probabilities from the two independent runs were determined to be in sufficiently close

Method	True Positive		False Positive		PO of Assoc.		
	Gene (se)	SNP (se)	Gene (se)	SNP (se)	Global	SNP	
n	1020	1020	5546	54980			
MISA							
<i>a</i>	<i>b</i>						
1	$\frac{1}{2}S$.77 (.006)	.669 (.007)	.128 (.001)	.025 (.0001)	2.00	.04
1/2	.	.809 (.005)	.704 (.007)	.166 (.001)	.031 (.0001)	.74	.020
1/4	.	.846 (.004)	.729 (.006)	.189 (.001)	.041 (.0002)	.32	.009
1/8	.	.874 (.003)	.739 (.006)	.259 (.001)	.048 (.0002)	.15	.005
1/16	.	.896 (.003)	.746 (.006)	.341 (.001)	.065 (.0003)	.07	.002
1/32	.	.904 (.003)	.746 (.006)	.437 (.001)	.090 (.0003)	.04	.001
1	<i>S</i>	.784 (.005)	.685 (.007)	.150 (.001)	.027 (.0001)	1.00	.020
1/2	.	.821 (.005)	.716 (.006)	.185 (.001)	.035 (.0001)	.42	.009
1/4	.	.855 (.004)	.736 (.006)	.207 (.001)	.044 (.0002)	.19	.005
1/8	.	.877 (.003)	.743 (.006)	.280 (.001)	.053 (.0002)	.09	.002
1/16	.	.899 (.003)	.746 (.006)	.368 (.001)	.073 (.0003)	.04	.001
1/32	.	.904 (.003)	.746 (.006)	.465 (.001)	.098 (.0004)	.02	.001
1	$\frac{3}{2}S$.791 (.005)	.696 (.007)	.169 (.001)	.029 (.0001)	.67	.01
1/2	.	.825 (.005)	.722 (.006)	.190 (.001)	.037 (.0002)	.29	.006
1/4	.	.855 (.004)	.735 (.006)	.222 (.001)	.048 (.0002)	.14	.003
1/8	.	.878 (.003)	.744 (.006)	.291 (.001)	.057 (.0002)	.07	.002
1/16	.	.898 (.003)	.746 (.006)	.377 (.001)	.075 (.0003)	.03	.001
1/32	.	.902 (.003)	.746 (.006)	.474 (.001)	.099 (.0004)	.02	.0004
Marg. BF		.695 (.007)	.627 (.007)	.171 (.001)	.041 (.0002)	–	1.00
lasso		.708 (.007)	.607 (.008)	.158 (.001)	.022 (.0001)	–	–
Step. AIC		.993 (.000)	.794 (.005)	.969 (.0001)	.445 (.001)	–	–
Step. BIC		.680 (.007)	.547 (.008)	.122 (.001)	.015 (.0001)	–	–
BH		.439 (.008)	.419 (.008)	.013 (.0001)	.011 (.0001)	–	–
Bonf.		.337 (.007)	.330 (.008)	.003 (.00001)	.006 (.00002)	–	–
Adj. Bonf. 1		.618 (.007)	.574 (.008)	.069 (.0003)	.027 (.0001)	–	–
Adj. Bonf. 2		.708 (.007)	.644 (.007)	.184 (.001)	.053 (.0002)	–	–

TABLE 3

Estimated overall false and true positive rates with standard errors and posterior odds (PO) of association at the gene and SNP levels. The values in bold characterize the method selected for use in the analysis of the NCOCS ovarian cancer example.

agreement.

On basis of this analysis, we estimate a lower bound on the pathway-wide Bayes factor for association to be $\text{BF}(H_A : H_0) = 7.67$ (which is also the posterior odds for this prior). This constitutes “positive” evidence in favor of an association between the pathway and ovarian cancer based on Jeffreys’ grades of evidence and corresponds to a posterior probability that the pathway is associated of roughly 0.89. Figure 2 summarizes the associations of the ten SNPs that had a SNP BF greater than 3.2, while Figure 3 illustrates the seven genes that contained these SNPs and two

others that received comparable support. SNPs and genes in the pathway are denoted by a two level name (e.g. S1 and G1) where the number represents the rank of the SNP or gene by its respective Bayes factor. These plots provide a graphical illustration of the top 100 models $\mathcal{M}_\gamma \in \mathcal{M}$ selected on basis of their posterior model probabilities. Models are ordered on the x-axis in descending probability and the width of the column associated with a model is proportional to that probability. SNPs (Figure 2) or genes (Figure 3) are represented on the y-axis. The presence of a SNP or gene in a model is indicated by a colored block at the intersection of the model’s column and the SNP’s or gene’s row. In Figure 2, the color of the block indicates the parametrization of the SNP: purple for log-additive, blue for recessive and red for dominant. The “checkerboard” pattern (as opposed to the presences of more vertical bars) suggests substantial model uncertainty.

The top five models depicted in Figure 2 include only a single SNP in addition to age at diagnosis (the design variable is omitted in the figure as it is included in all models). The top model includes SNP S1 in gene G1 under the log-additive genetic parametrization, which is estimated to have an odds ratio (OR) of approximately 1.42 (the posterior mode). The second ranked model includes only SNP S2 in gene G1 under the log-additive genetic parametrization with an estimated OR of 1.37. Note that the study has relatively low power to detect effects of this magnitude (Figure 1).

Figure 2 also illustrates that many of the top models beyond the first five include multiple SNPs. This suggests that if we were to restrict our attention to single SNP models we would potentially lose substantial information regarding their joint effects. For example, model six is comprised of both SNP S3 from gene G3 and SNP S2 from gene G1, while model 12 is comprised of both SNP S3 from gene G3 and SNP S1 from gene G1. In both cases, SNP S3 is included in models with a SNP from gene G1. This may indicate that not only are SNPs S1, S2, and S3 important as single effects in the top four models, but that their combined effects may be of interest. Note that, in cases where the disease variant is unmeasured but ‘tagged,’ several tagged SNPs may be required to explain variation at that locus.

The SNP Bayes factors of S1 (BF = 42.2) and S2 (BF = 17.8) provide “strong evidence” of changes in prior beliefs, however, the marginal posterior probabilities of association with ovarian cancer are 0.38 and 0.20, respectively. Figure 2 illustrates that when one of SNP S1 or S2 is included in a model, the other is often not (at least in the top 50 models). This trade off often arises when SNPs are correlated (i.e. in high linkage disequilibrium). In this case, R^2 of 0.5 suggests fairly strong LD between SNPs S1 and S2, in which case the joint inclusion probabilities are more meaningful than

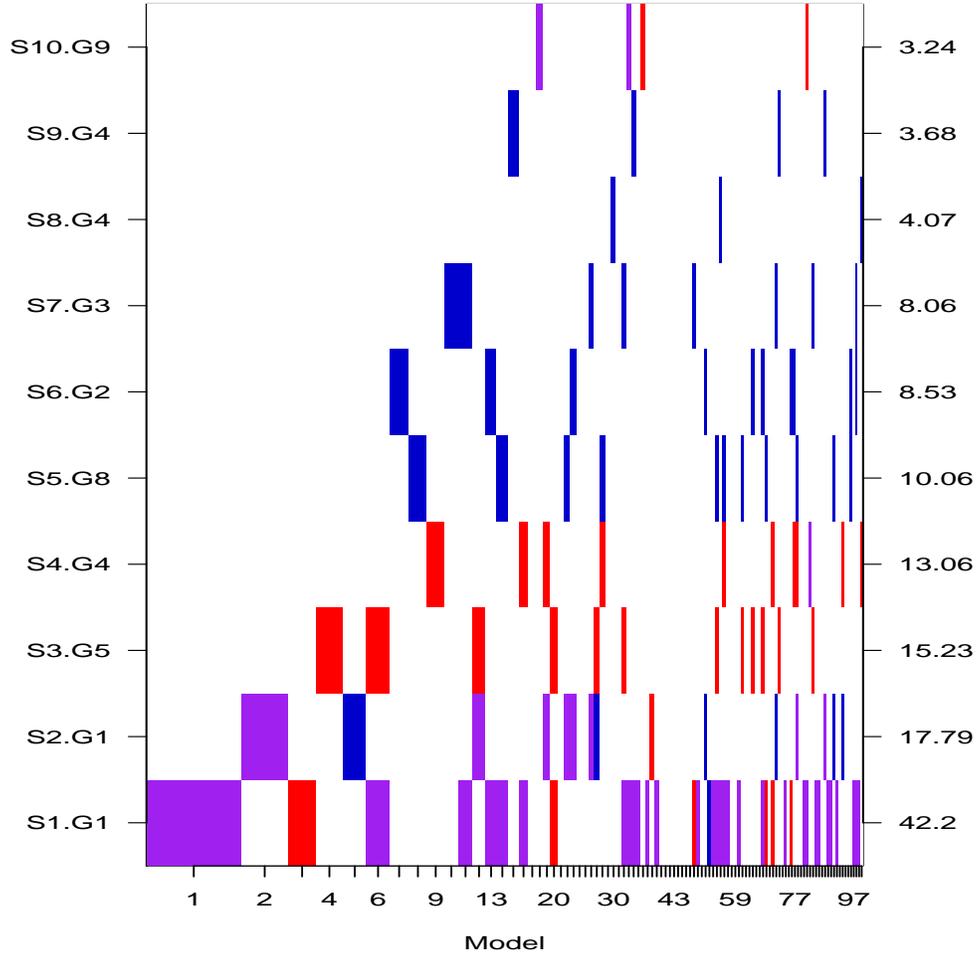


FIG 2. Image plot of the SNP inclusion indicators for the SNPs with marginal Bayes factors greater than 3.2 and the top 100 Models. The color of the inclusion block corresponds to the genetic parametrization of the SNP in that model. Purple corresponds to a log-additive parametrization, red to a dominant parametrization and blue to a recessive parametrization. SNPs are ordered on basis of their marginal SNP Bayes Factors which are plotted on the right axis across from the SNP of interest. Width of the column associated with a model is proportional to its estimated model probability.

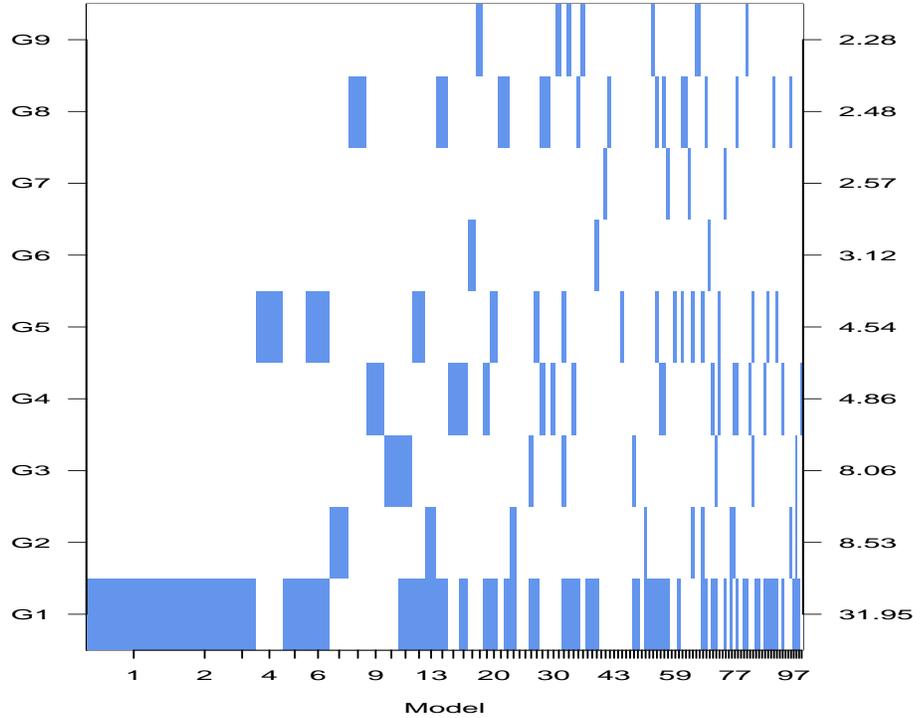


FIG 3. Image plot of the gene inclusion indicators for the top 100 Models. Genes are ordered based on their marginal gene Bayes Factors which are plotted on the right axis. Columns correspond to models and have width proportional to the estimated model probability, models are plotted in descending order of posterior support. The color is chosen to be neutral since the genetic parametrizations are not defined at the gene level.

marginal probabilities. Both SNP 1 and SNP 2 are in gene G1 which has a gene Bayes factor of 31.95 (Figure 3) and posterior probability of association of 0.58. These probabilities need to be interpreted in the context of model uncertainty; conditional on the pathway being associated with ovarian cancer, the probability that gene G1 is driving the association is $0.58/0.89 = 0.65$. However, there remains substantial uncertainty regarding which genes and SNPs may explain it as the posterior mass is spread over competing models/hypotheses. The positive support for an association suggests the continuation of data accrual to refine these posterior probabilities.

Gene G1 and other genes in Figure 3 highlight a caution regarding the

interpretation of Bayes factors as a measure of absolute support with composite hypotheses. The gene Bayes factor for G1 is 31.95, which is smaller than the SNP Bayes factors for S1 (42.2). The posterior probability that gene G1 is associated is based on summing the probabilities of all models that include at least one SNP from that gene (S1, S2, and S51) hence the *posterior probability* for gene inclusion is always greater than or equal to the probability that any one SNP is included (i.e. posterior probabilities observe a monotonicity property with composite hypotheses). Bayes factors (and p-values) for composite hypotheses do not share this monotonicity property [Lavine and Schervish 1997]. Bayes factors for comparing composite hypotheses may be expressed as the ratio of the weighted average (with respect to the prior distribution) of marginal likelihoods conditional on the hypotheses, which may decrease the evidence in favor of a composite hypothesis when a subset of the individual hypotheses have low likelihood. As mentioned in Section 2.1.4, while Bayes factors do not provide a coherent measure of *absolute* support because of their non-monotonicity property, Lavine and Schervish [1997] show that the log Bayes factor does provide a coherent measure of how much the data *change* the support for the hypothesis (relative to the prior). Hence, they do provide useful summaries of changes in prior beliefs of association in large association studies with many competing models/hypotheses.

	df	Sum Sq	Mean Sq	F value	Pr(>F)
snp	69	1635891.00	23708.57	208.04	$< 2 \times 10^{-16}$
prior	1	169641.66	169641.66	1488.60	0.0000
impute	1	134.41	134.41	1.18	0.28
prior:impute	1	53.16	53.16	0.47	0.50
Residuals	207	23589.77	113.96		

TABLE 4

Analysis of variance for the ranked SNP Bayes factors contrasting the prior hyperparameters (default $a = 1$ versus $a = 1/8$) and method of imputation (full imputation with 100 data sets versus a modal estimate of the missing genotypes) for the 70 SNPs in the NCOCS pathway that passed the marginal screen.

5.1. *Sensitivity Analysis.* In this section, we consider sensitivity of the results in the NCOCS study to the prior distribution on the models and to the method of imputation. The simulation study suggests that priors with smaller values of a may identify more associated SNPs. We estimated that the $\text{BetaBinomial}(1/8, S)$ prior distribution on model size has a false positive rate comparable to the marginal BF method, but a much higher true positive rate, in the scenarios we considered. Full data imputation, achieved

by averaging over the distribution of missing SNPs, is probabilistically correct, but computationally expensive. Thus, if the use of modal imputation provides an accurate approximation to BF calculated using full imputation, the computational efficiency of MISA can be greatly improved at small cost.

For purposes of this analysis, we used the set of unique models identified by the EMC search with modal imputations and $a = 1$ and calculated 3 additional sets of BFs. First, we obtained marginal likelihoods for each of these models using 100 imputed data sets with missing SNPs filled in based on their estimated distribution. Second, we calculated BFs using the $\text{BetaBinomial}(1/8, S)$ and $\text{BetaBinomial}(1, S)$ prior distributions using the marginal likelihoods under the full and modal imputations. We applied ANOVA to these four sets of BFs to compare the effects of prior hyperparameters and imputation methods after adjusting for SNP using the ranked SNP BFs.¹

Table 4 shows that the method of imputation has no significant effect on the ranking of SNP BFs. This suggests that, for purposes of model search and calculation of BFs, we may use the modal imputed genotypes in place of full imputation, with significant computational savings. For purposes of parameter estimation, we suggest the use of full imputation using a subset of the top models and top SNPs as using a plug-in approach for imputation is known to underestimate uncertainty.

We anticipated that the prior distribution would have a significant effect based on the higher true positive and false positive rates estimated from the simulation study and by considering differences in the prior odds. While Table 4 suggests that overall the rankings are different between the two prior distributions, the top 20 SNPs have the same rank under each of the four methods, leading to no qualitative differences in our conclusions about the top SNPs. The prior odds for any given SNP's inclusion in a model are 8 times lower under the $\text{BetaBinomial}(1/8, S)$ prior distribution than under to the $\text{BetaBinomial}(1, S)$ prior distribution; the resulting SNP BFs are 2.8 times higher under the $\text{BetaBinomial}(1/8, S)$ prior distribution than those under the $\text{BetaBinomial}(1, S)$ prior distribution. As a result, eight more SNPs are above the 3.2 threshold used by the NCOCS to determine SNPs worthy of additional study.

5.2. *External Validation and Comparison.* To provide a basis of comparison, we applied the methods described in the simulation study (Section 4) to the NCOCS data. We omitted stepwise logistic regression using AIC

¹Ranks were used as residuals on the log scale still exhibited strong departures from normality.

because of its poor operating characteristics. The marginal FDR methods of Bonferroni and Benjamini–Hochberg failed to identify any significant SNPs. Lasso, which accounts for correlation among SNPs, also failed to identify any SNPs. Stepwise logistic regression using BIC selected a model with three of the top four SNPs identified by MISA — S1.G1, S3.G5 and S4.G4 — but failed to identify S2.G1, which has correlation 0.71 with SNP S1.G1. This highlights a problem with selection methods that ignore model uncertainty.

The NCOCS proposed two SNPs — S10 and S14 in G9 — for external validation by the Ovarian Cancer Association Consortium (OCAC), a large international multi-center consortium of ovarian cancer case-control studies. The decision to focus on these variants was made on basis of results from an earlier version of the NCOCS data set and on basis of the strong prior interest NCOCS researchers had in the gene (and not on basis of the analysis described above). Under the default $\text{BetaBinomial}(1, S)$ prior distribution, only SNP S10 in G9 exceeds the 3.2 threshold and the G9 BF is only 2.28. In contrast, under the $\text{BetaBinomial}(1/8, S)$ prior distribution, both SNPs S10 and S14 (LD 0.62) in G9 have SNP BFs greater than 3.2 (8.70 and 5.99, respectively) and the gene BF is 6.18. An additional three SNPs in the same gene were proposed by another member of the consortium on the basis of uncorrected p-values. Of the five SNPs proposed for validation, only SNPs S10 and S14 were confirmed to be associated with serous invasive ovarian cancer by OCAC [Schildkraut *et al.* 2009].

6. Discussion. In this paper, we describe MISA, a natural framework for multi-level inference with an implicit multiple comparisons correction for hypothesis based association studies. MISA allows one to quantify evidence of association at three levels: global (e.g. pathway-wide), gene, and SNP, while also allowing for uncertainty in the genetic parametrization of the markers. We have evaluated MISA against established, simple to implement and more commonly used methods and demonstrated that our methodology does have higher power than these methods in detecting associations in modestly powered candidate pathway case-control studies. The improvement in power is most noticeable for odds ratios of modest (real world) magnitude and comes at the cost of only a minimal increase in the false positive rate. Like stepwise logistic regression, lasso and logic regression, MISA improves upon marginal, SNP-at-a-time methods by considering multivariate adjusted associations. By using model averaging, MISA improves upon these multivariate methods that select a single model, which may miss important SNPs because of LD structure. These improvements have concrete implications for data analysis: MISA identified SNPs in the NCOCS data that were

subsequently externally validated; none of the less complex methods considered here highlighted these SNPs to be of interest. Currently, other top ranked SNPs in genes identified by MISA are undergoing external validation. Finally, we note that while MISA was developed for binary outcomes in case-control studies, MISA is readily adaptable to accommodate other forms of outcome variables (e.g. quantitative traits or survival) that are naturally modeled within a GLM framework.

APPENDIX A: IMPLIED PRIOR DISTRIBUTION UNDER AIC

Given that a closed-form expression for the marginal likelihood is not available for logistic regression, we have used the AIC to approximate the likelihood. In what follows, we determine a prior distribution on model coefficients that is consistent with AIC.

We assume a normal prior distributions on the d_γ -dimensional vector of regression coefficients (log odds ratios) of the form

$$p(\boldsymbol{\theta}_\gamma | \mathcal{M}_\gamma) \sim \text{N} \left(\mathbf{t}_\gamma, \frac{1}{k} \mathbf{J}_\gamma^{-1} \right),$$

where \mathbf{J}_γ is the observed Fisher information under model \mathcal{M}_γ evaluated at the maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\theta}}_\gamma$. Setting the covariance matrix to be proportional to the inverse Fisher information ensures that the correlation structure in the prior distribution matches that of the likelihood.

In order to approximate the marginal likelihood we used a Laplace approximation based on expanding the log-likelihood in a second-order Taylor's series expansion about $\hat{\boldsymbol{\theta}}_\gamma$:

$$\mathcal{L}(\boldsymbol{\theta}_\gamma | \mathcal{M}_\gamma) \approx \mathcal{L}(\hat{\boldsymbol{\theta}}_\gamma | \mathcal{M}_\gamma) - \frac{1}{2}(\boldsymbol{\theta}_\gamma - \hat{\boldsymbol{\theta}}_\gamma)^T \mathbf{J}_\gamma (\boldsymbol{\theta}_\gamma - \hat{\boldsymbol{\theta}}_\gamma)$$

leading to the approximate marginal likelihood

$$\begin{aligned} p(D | \mathcal{M}_\gamma) &\approx \exp\{\mathcal{L}(\hat{\boldsymbol{\theta}}_\gamma | \mathcal{M}_\gamma)\} \times \\ &\int K_{\boldsymbol{\theta}_\gamma}(\hat{\boldsymbol{\theta}}_\gamma, \mathbf{J}_\gamma^{-1}) \frac{1}{(2\pi)^{\frac{d_\gamma}{2}}} |k\mathbf{J}_\gamma|^{\frac{1}{2}} K_{\boldsymbol{\theta}_\gamma}(\mathbf{t}_\gamma, \frac{1}{k}\mathbf{J}_\gamma^{-1}) d\boldsymbol{\theta}_\gamma \\ &= \left(\frac{k}{k+1} \right)^{\frac{d_\gamma}{2}} \exp\{\mathcal{L}(\hat{\boldsymbol{\theta}}_\gamma | \mathcal{M}_\gamma)\} K_{\hat{\boldsymbol{\theta}}_\gamma}(\mathbf{t}_\gamma, \frac{k+1}{k}\mathbf{J}_\gamma^{-1}); \end{aligned}$$

where $K_{\boldsymbol{\theta}_\gamma}(\hat{\boldsymbol{\theta}}_\gamma, \mathbf{J}_\gamma^{-1}) = \exp\{-\frac{1}{2}(\boldsymbol{\theta}_\gamma - \hat{\boldsymbol{\theta}}_\gamma)^T \mathbf{J}_\gamma (\boldsymbol{\theta}_\gamma - \hat{\boldsymbol{\theta}}_\gamma)\}$. Setting this approximate $\log(p(D | \mathcal{M}_\gamma))$ equal to -0.5AIC we have equality when the prior mean \mathbf{t}_γ is set to $\hat{\boldsymbol{\theta}}_\gamma$ where the right-most term vanishes, and $k = \frac{1}{\exp(2)-1}$.

Roughly speaking, this implies that the prior standard deviation of any standardized log odds ratio is about 2.5. This suggests that the approximation of the marginal likelihood under AIC is reasonable for prior distributions with mean zero, as this provides enough dispersion to cover the range of log odds ratios anticipated.

APPENDIX B: MARGINAL BAYES FACTOR SCREEN

We used Laplace approximations to estimate the marginal Bayes Factors (BFs) used to screen the SNPs [Kass and Raftery 1995]. In particular, we estimated the marginal likelihood of each of the three genetic models of association (log-additive, dominant and recessive) and under the null model (model of no genetic association). The BF for a model of association is defined as the ratio of the marginal likelihood of that model of association to the marginal likelihood of the null model.

We accounted for missing genetic data by averaging marginal likelihoods over the $M = 100$ imputed genetic data sets. This affected only the calculations under the three genetic models of association, but not the null model. Hence the BF for an association was computed as the average of imputation-specific BFs.

In the ovarian cancer analysis, the model for each SNP was a logistic regression for disease status given the variable age and the model-specific genotype variable. Age was included in all models, including the 'null' model of no association. The simulation models were unadjusted as no design or confounder variables were simulated. We placed a normal, mean zero, standard deviation two prior on the parameter of the genetic effect variable and flat, improper priors on the remaining log odds ratio parameters. We ordered SNPs according to the maximum of the three Bayes factors and considered those with a maximum greater than or equal to one in the MISA model search. Our software for calculating marginal Bayes factors is included in the MISA R package.

APPENDIX C: GENETIC SIMULATIONS

We used simulated case-control data to compare MISA and other commonly used procedures for genetic association studies. The simulated data sets were structured so as to reflect the details — genes, tag SNPs, LD structure, and sample size — of a NCOCS candidate pathway study comprised of 53 genes tagged by 508 tag SNPs. Genotypes were simulated in two stages. First, for each of the 53 genes represented in the data set, we phased the NCOCS control SNP genotype data and estimated recombination rates using PHASE [Stephens *et al.* 2001], which provides estimates of the population

haplotype distribution. Phase is a Bayesian method that obtains approximate samples from the posterior distribution of all possible haplotype pairs (H) given the observed genotypes (G) using Gibbs sampling and estimates recombination rates empirically from this sample. Second, given a model of association and the PHASE output, we generated case-control data at the selected tags using HAPGEN [Marchini and Su 2006]. Hapgen is a program that simulates haplotypes for a case-control sample of individuals given a set of population haplotypes and recombination rates for the regions of interest and choice of the hypothetical associated SNP and its allele-specific odds ratios.

We generated 124 simulated data sets as follows. Ten of the simulations are null; there are no associations in the genes of interest. The remaining 114 simulations assume that a randomly chosen subset of 9 genes are associated and that within the associated genes, a single, randomly chosen SNP is the source of the association. Within the 114 associated simulations, the associated tag SNPs were accorded an odds ratio (OR) of 1.25, 1.5, 1.75, 2.0, or 2.25 and assumed to have either a dominant genetic parametrization, log-additive genetic parametrization or a recessive genetic parametrization. The marginal distribution over odds ratios is given in Figure 1. The marginal distribution over genetic models was uniform. The simulations used for the power analysis can be found at the URL for the software.

We have also developed a software package, SimGbyE, that creates simulated case/control or survival data sets with one or more of the following assumed effects: genetic main effects (G), environmental main effects (E), Gene by Gene interactions (GbyG), Gene by environment interactions (GbyE). The assumed genetic one and two locus models of epistasis are chosen randomly from a set of models chosen from Li and Reich [2000]. Then given a set of assumed coefficients on the effects mentioned above, an outcome variable is simulated (case/control or survival) based on a set user specified distribution parameters. This package differs slightly from the method used to develop the simulations within this article by estimating the population haplotype distribution from HapMap instead of using PHASE to estimate the distribution from the set of control SNP genotypes in the NCOCS data.

The main function calls Hapgen to simulate one replicate from a specified chromosomal region given data from one of the HapMap II populations. The code generates samples of genotypes in a contiguous range of DNA using Hapmap release 21 (NCBI build 35) data. The position range may encompass an entire chromosome or simply bracket a gene or locus of interest. That function can also be used to simulate data from multiple independent regions

to generate a candidate gene/pathway sample or a genome-wide sample. The default is to generate population-based genetic samples. However, to build genetic simulations with main effects only, parameters can be set so that Hapgen will randomly choose a variant in the specified region as the disease allele and generate a case-control sample. To build more complex associations we have written a wrapper function to take the genetic samples produced by Hapgen and simulate an outcome variable based on genetic main effects with multiple genetic parametrizations, environmental main effects, Gene by Gene interactions, and Gene by environment interactions.

WEB RESOURCES

The URL for the software for the methodology and simulations presented in this paper is:

<http://www.isds.duke.edu/gbye/packages.html>.

REFERENCES

- BALDING, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature* **7** 781–791.
- BURTON, P. R. *et al.* (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.
- CLAYTON, D. G. *et al.* (2005). Population structure differential bias and genomic control in a large-scale case-control association study. *Nature Genet.* **37** 1243–1246.
- CLYDE, M. (1999). Bayesian model averaging and model search strategies (with discussion). In *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, 157–185.
- CLYDE, M. and GEORGE, E. I. (2004). Model uncertainty. *Statist. Sci.* **19** 81–94.
- CORDELL, H. J. and CLAYTON, D. G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in type 1 diabetes. *AJHG* **70** 124–141.
- CUI, W. and GEORGE, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Stat. Planning and Inference* **138** 888–900.
- EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of American Statistical Association* **102** 93–103.
- FLINT, J. and MACKAY, T. F. C. (2009). Genetic architecture of quantitative traits in mice, flies and humans. *Genome Research* **19** 723–733.
- GELMAN, A. and RUBIN, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7** 457–472.
- GEORGE, E. (1999). Discussion of “Model averaging and model search strategies” by M. Clyde. In *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*.
- GEYER, C. J. (1991). Markov chain monte carlo maximum likelihood. In *Proc. 23rd Symp. Interface*, 156–163. Computing Science and Statistics.
- GOODMAN, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annal Intern. Med.* **130** 1005–1013.
- HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statist. Sci.* **14** 382–401. Corrected version at <http://www.stat.washington.edu/www/research/online/hoeting1999.pdf>.
- HOLLAND, J. H. (1975). *Adaptation in Natural and Artificial Systems*. The U. of Michigan Press.
- JEFFREYS, H. (1961). *Theory of Probability*. Oxford Univ. Press, third edn.
- KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.
- KOOPERBERG, C. and RUCZINSKI, I. (2004). Identifying interacting SNPs using monte carlo logic regression. *Genetic Epidemiology* **28** 157–170.
- LAVINE, M. and SCHERVISH, M. J. (1997). Bayes factors: What they are and what they are not. *The American Statistician* **53** 119–122.

- LEY, E. and STEEL, M. F. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Applied Econometrics* **24** 651–674.
- LI, W. and REICH, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity* **50** 334.
- LIANG, F. and WONG, W. H. (2000). Evolutionary monte carol: Applications to c_p model sampling and change point problem. *Statistica Sinica* **10** 317–342.
- LOKHORST, J., VENABLES, B., PORT TO R, B. T. and TESTS ETC: MARTIN MAECHLER (2009). *lasso2: L1 constrained estimation aka 'lasso'*. R package version 1.2-10.
- MARCHINI, J. and SU, Z. (2006). Hapgen, a c++ program for simulating case and control snp haplotypes.
- OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (1999). On the LASSO and its dual. *Journal of Comp. and Graph. Stat.* **9** 319–337.
- PARK, M. Y. and HASTIE, T. (2008). Penalized logistic regression for detecting gene interactions. *Bioinformatics* **9** 30–50.
- RAFTERY, A. E. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *J. Roy. Statist. Soc. Ser. B* **48** 249–250.
- RUCZINSKI, I., KOOPERBERG, C. and LEBLANC, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics* **12** 475–511.
- SCHILDKRAUT, J. M. *et al.* (2008). Cyclin E overexpression in epithelial ovarian cancer characterizes an etiologic subgroup. *Cancer Epidemiology Biomarkers and Prevention* **17** 585–593.
- SCHILDKRAUT, J. M. *et al.* (2009). Single nucleotide polymorphisms in the TP53 region and susceptibility to invasive epithelial ovarian cancer. *Cancer Research* **69** 2349–2357.
- SCHWENDER, H. and ICKSTADT, K. (2007). Identification of SNP interactions using logic regression. *Biostatistics* **9** 187–198.
- SCOTT, J. G. and BERGER, J. O. (2008). Multiple testing, empirical Bayes, and the variable-selection problem. Discussion Paper 2008-10, Duke University Department of Statistical Science.
- SERVIN, B. and STEPHENS, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLOS Genetics* **3**.
- SHI, W., LEE, K. and WAHBA, G. (2007). Detecting disease-causing genes by lasso-patternsearch algorithm. *BMC Proceedings* **1** S60.
- STEPHENS, M. and BALDING, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Genet.* **10** 681–690.
- STEPHENS, M., SMITH, N. and DONNELLY, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics* **68** 978–989.
- STOREY, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society* **64** 479–498.
- WACHOLDER, S. (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.
- WAKEFIELD, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *The American Journal of Human Genetics* **81** 208–227.
- WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. and LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25** 714–721.

MERLISE A. CLYDE
EDWIN S. IVERSEN
SCOTT C. SCHMIDLER
MELANIE A. WILSON
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NC, 27708-0251
E-MAIL: clyde@stat.duke.edu

iversen@stat.duke.edu
scs@stat.duke.edu
maw27@stat.duke.edu

JOELLEN M. SCHILDKRAUT
DEPARTMENT OF COMMUNITY AND FAMILY MEDICINE
DUKE UNIVERSITY
DURHAM, NC; 27713
E-MAIL: schil001@mc.duke.edu