Biostatistics (2020) **21**, 2, *pp.* 319–335
doi:10.1093/biostatistics/kxy049
Advance Access publication on September 21, 2018

Bayesian model averaging for the X-chromosome inactivation dilemma in genetic association study

BO CHEN, RADU V. CRAIU, LEI SUN*

Department of Statistical Sciences, University of Toronto, Sidney Smith Hall, 100 St. George Street, Toronto, ON M5S3G3, Canada

sun@utstat.toronto.edu

Summary

X-chromosome is often excluded from the so called "whole-genome" association studies due to the differences it exhibits between males and females. One particular analytical challenge is the unknown status of X-inactivation, where one of the two X-chromosome variants in females may be randomly selected to be silenced. In the absence of biological evidence in favor of one specific model, we consider a Bayesian model averaging framework that offers a principled way to account for the inherent model uncertainty, providing model averaging-based posterior density intervals and Bayes factors. We examine the inferential properties of the proposed methods via extensive simulation studies, and we apply the methods to a genetic association study of an intestinal disease occurring in about 20% of cystic fibrosis patients. Compared with the results previously reported assuming the presence of inactivation, we show that the proposed Bayesian methods provide more feature-rich quantities that are useful in practice.

Keywords: Bayes factors; Bayesian methods; Bayesian model averaging; Genome-wide association studies; Markov chain Monte Carlo; Model uncertainty; Ranking; X-chromosome.

1. Introduction

In the search for genetic markers that are responsible for heritable complex human traits, whole-genome scans including the genome-wide association studies (GWAS) and the next generation sequencing (NGS) studies have made tremendous progress; see `www.genome.gov/gwastudies` for the most recent summary of GWAS findings by the National Human Genome Research Institute (Welter *and others*, 2014). The "whole-genome" nature of these studies, however, is often compromised by the omission of the X-chromosome (Heid *and others*, 2010; Teslovich *and others*, 2010). In fact, it was found that "only 33% (242 out of 743 papers) reported including the X-chromosome in analyses" based on the NHGRI GWAS Catalog (Wise *and others*, 2013). The exclusion of X-chromosome from GWAS and NGS is due to it being fundamentally different between females and males. In contrast to the 22 autosomal chromosomes where both females and males have two copies, females have two copies of X-chromosome (XX), whereas males have only one X coupled with one Y-chromosome (XY). Thus, statistical association methods well developed for analyzing autosomes require additional considerations for valid and powerful application to X-chromosome.

*To whom correspondence should be addressed.

Focusing on the single nucleotide polymorphisms (SNPs) as the genetic markers of interest here and without loss of generality, let d and D be the two alleles of a SNP and D be the risk allele. An X-chromosome SNP in females has three possible (unordered) genotypes, dd, dD, and DD, in contrast to d and D in males. Suppose each copy of the D allele has an effect size of β on the outcome of interest; this β is the coefficient in linear regression for studying (approximately) normally distributed outcomes, or the log odds ratio in logistic regression for analyzing binary traits. To ensure "dosage compensation for X-linked gene products between XX females and XY males", X-chromosome inactivation (XCI) may occur so that one of the two alleles in females is randomly selected to be silenced (Gendrel and Heard, 2011). In other words, the effects of dd, dD, and DD in females are now respectively 0, $\beta/2$ and β on average after XCI vs. 0, β and 2β without XCI. However, without collecting additional biological data the status of XCI is unknown.

Previous work on developing association methods for X-chromosome SNPs mostly focused on issues other than XCI, including the assumptions of Hardy–Weinberg equilibrium (HWE) and equal allele frequencies or sample sizes between females and males (Zheng *and others*, 2007; Clayton, 2008). In his classic review article, Clayton (2009) also discussed analytical strategies for multi-population or family-based studies. In each of these cases, either the XCI or no-XCI model is assumed, and naturally, these methods work well only if the underlying assumption about the XCI status is correct (Hickey and Bahlo, 2011; Loley *and others*, 2011; Konig *and others*, 2014).

More recently, Wang *and others* (2014) recognized the problem and proposed a maximum likelihood approach. In essence, the proposed method calculates multiple association statistics for testing the effect of a X-chromosome SNP under XCI and no-XCI models, then uses the maximum. To adjust for the inherent selection bias, the method uses a permutation-based procedure to obtain the empirical distribution for the maximal test statistic and assess its significance. Although Wang *and others* (2014) method appears to be adequate in terms of association testing, in the presence of model uncertainty it is not clear how to construct a point estimate or confidence interval for effect size β, or, what is a suitable measure of evidence for supporting one model over the other. Thus, an alternative paradigm that directly accounts for the inherent model uncertainty is desirable.

To close this gap, we propose a Bayesian approach that can handle in a principled manner the uncertainty about the XCI status. The use of Bayesian methods for genetic association studies is not new. Stephens and Balding (2009) and Craiu and Sun (2014) provide reviews in the context of studying autosome SNPs. Herein, we consider the posterior distributions generated from Bayesian regression models for analyzing X-chromosome SNPs under the XCI and no-XCI assumptions. We combine the estimates from the two models following the Bayesian model averaging (BMA) principle that has long been recognized as a proper method for incorporating model uncertainty in a Bayesian analysis (Draper, 1995; Hoeting *and others*, 1999). We calculate the BMA-based highest posterior density (HPD) region for the parameter of interest. The BMA posterior distribution is directly interpretable as a weighted average for β, averaged over the XCI and no-XCI models with more weight given to the one with stronger support from the data. To rank multiple SNPs, we compare the lower bounds of the HPD regions for each SNP.

In Section 2, we present the theory of BMA for handling the XCI uncertainty issue. We first consider linear regression models for studying continuous traits where closed-form solutions can be derived. We then discuss extension to logistic models for analyzing binary outcomes where Markov chain Monte Carlo (MCMC) methods are used for inference. In this setting, the calculation of Bayes factors is no longer possible analytically so we implement numerical approximations that have been reliably used in computing ratios of normalizing constants. In Section 3, we conduct extensive simulation studies to evaluate the performance of the proposed Bayesian approach comparing with Wang *and others* (2014) method. In Section 4, we apply our method to a X-chromosome association study of meconium ileus, an intestinal disease present in cystic fibrosis patients, providing further evidence of method performance. In Section 5, we discuss possible extensions and future work.

2. Methods

2.1. *Normally distributed outcomes*

The methodology development here focuses on linear models, studying association relationship between a (approximately) normally distributed trait/outcome Y and a X-chromosome SNP. Let (dd, dD, DD) and (d, D) be the genotypes of a SNP, respectively, for females and males. For autosome or X-chromosome SNPs in females, genotypes dd, dD, and DD are typically coded additively as 0, 1, and 2, representing the number of copies of a reference allele, assumed to be D here. Under the X-chromosome inactivation (XCI) assumption, one of the two alleles of a female is randomly selected to have no effect on the outcome. Thus, the XCI (M_1) and no XCI (M_2) assumptions lead to two different coding schemes, respectively, G_1 and G_2 as summarized below.

		Female			Male	
Model	Coding	dd	dD	DD	d	D
M_1: XCI	G_1	0	0.5	1	0	1
M_2: no XCI	G_2	0	1	2	0	1

Let Y be the vector of outcome measures of sample size n, and G_k be the vector of genotype values for the n individuals coded under the two models M_k, $k = 1$ and 2. In addition, sex may have an effect on the outcome and should be included as a covariate in X-chromosome association studies. We use S to denote sex, where $S = 0$ for females and $= 1$ for males as in convention. For each model M_k, we consider a linear regression model $Y = X_k\boldsymbol{\theta}_k + \epsilon_k$, where $X_k = (\mathbf{1}_n, G_k, S)$ is the design matrix, $\boldsymbol{\theta}_k = (\alpha_k, \beta_k, \gamma_k)'$ and $\epsilon_k \sim N(0, \sigma^2 I_n)$. Here, β_k represents the genetic effect of one copy of D under model $M_k, k = 1, 2$, accounting for the effects of sex and other covariates $Z \in \mathcal{R}^p$ such as age, smoking status, and population information. For notation simplicity but without loss of generality for implementing the following Bayesian model average framework, the additional covariate Z vector is omitted from the regression model. The coding of 0.5 for genotype dD under M_1 reflects the fact that the effect of dD under the XCI assumption is the average of zero effect of d (if D was silenced) and β effect of D (if d was silenced). In addition, ϵ_1 and ϵ_2 have the same variance $\sigma^2 I_n$ because both models are based on same response variable Y.

Before we present the Bayesian approach, we make several important remarks here. First, the regression model above studies the genotype of a SNP, thus it does not require the assumption of HWE; only methods based on allele counts are sensitive to the equilibrium assumption (Sasieni, 1997). Similarly, allele-frequency affects only the efficiency of genotype-based association methods but not the accuracy. In addition, although other types of genetic architecture are possible, e.g. dD and DD having the same effect as in a dominant model or dd and dD having the same effect as in a recessive model, the additive assumption has its theoretical justification and sufficiently approximates many other models (Hill *and others*, 2008). Therefore, we focus on the additive models in this section and the simulation studies. In application study, however, we will study the genotypic model and compare the results with that obtained from the additive assumption.

2.2. *A Bayesian model averaging approach*

In practice, it is unknown which of the two models (M_1 for XCI and M_2 for no XCI) is true. Instead of performing inference based on only one of the two models or choosing the maximum one, the BMA framework naturally aggregates information from both M_1 and M_2. Central to BMA is the Bayes factor (BF) defined as

$$BF_{12} = \frac{P(Y|M_1)}{P(Y|M_2)},$$

where $P(Y|M_k) = \int f(Y|\boldsymbol{\theta}, \sigma^2, M_k)\pi(\boldsymbol{\theta}|\sigma^2, M_k)\pi(\sigma^2|M_k)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\sigma^2$ is the marginal probability of the data under model M_k. Herein, we used the outcome variable Y to denote all available data; meaning should be clear from the context.

We consider conjugate priors for $\pi(\sigma^2|M_k)$ and $\pi(\boldsymbol{\theta}|\sigma^2, M_k)$ for each model, $\pi(\sigma^2|M_k) = \pi(\sigma^2) = IG(a_0, b_0)$ where $IG(a_0, b_0)$ is the inverse gamma distribution with density function

$$p(\sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-a_0-1}\exp\left(-\frac{b_0}{\sigma^2}\right).$$

As noted before, Y is common between M_1 and M_2 so the prior distributions of σ^2 for the two models are the same. For $\pi(\boldsymbol{\theta}|\sigma^2, M_k) = \pi(\boldsymbol{\theta}_k)$,

$$\pi(\boldsymbol{\theta}_k) = N(\boldsymbol{\mu_0}, \sigma^2\Lambda_{0k}^{-1}),$$

where Λ_{0k} is the precision matrix (Wright, 2008). For hyperparameter Λ_{0k}, we adopt the g-prior (Zellner, 1986) that has $\Lambda_{0k} = \frac{\lambda}{n}X_k'X_k$. We note that here the female component of G_1 is half of that of G_2. Thus, if we naïvely use $\pi(\boldsymbol{\theta}_k) = N(\boldsymbol{\mu_0}, \sigma^2\lambda^{-1}I_2)$, this scaling factor can affect the Bayes factor and the ensuing model average quantities; the model with smaller covariate values is always preferred even if rescaling is the only difference. We discuss further in Section 5 the importance of using the g-prior in this setting.

When estimating the posterior distribution of $\boldsymbol{\theta}$ under each model, we find the hyperparameters, namely λ, $\boldsymbol{\mu_0}$, a_0, and b_0, have little effects on the posterior distributions in general. We use $\lambda = 1$ for precision parameter following the recommendations in Kass and Raftery (1995). For other hyperparameters, naturally $\boldsymbol{\mu_0} = \mathbf{0}$ unless there is prior information about association between the SNP under the study (or sex) and the trait of interest. In the absence of additional information for σ^2, we let $a_0 = b_0 = 0.1$; setting $a_0 = b_0 = 0$ in our simulation studies did not lead to noticeable numerical difference compared to $a_0 = b_0 = 0.1$.

The likelihood function is defined by $f(Y|\boldsymbol{\theta}, \sigma^2, M_k) \sim N(X_k\boldsymbol{\theta}, \sigma^2 I_n)$, which yields a normal-inverse-gamma posterior distribution for $(\boldsymbol{\theta}, \sigma^2)$, and the corresponding marginal distributions of $\boldsymbol{\theta}$ and σ^2 can be derived. Specifically, $\pi(\boldsymbol{\theta}, |Y, M_k)$, the posterior distributions for $\boldsymbol{\theta}$ under each model M_k, is a multivariate t distribution with $2a$ degrees of freedom (df henceforth), location parameter $\boldsymbol{\mu_k}$ and scale parameter $\frac{b_k}{a}\Lambda_k^{-1}$, i.e. density function

$$\pi(\boldsymbol{\theta}|Y, M_k) \propto \left[1 + \frac{(\boldsymbol{\theta} - \boldsymbol{\mu}_k)'\Lambda_k(\boldsymbol{\theta} - \boldsymbol{\mu}_k)}{2b_k}\right]^{-\frac{2a+2}{2}},$$

and the posterior of σ^2 is $\pi(\sigma^2|Y, M_k) = IG(a, b_k)$, where

$$\Lambda_k = X_k'X_k + \Lambda_{0k} \quad (\Lambda_{0k} = \frac{\lambda}{n}X_k'X_k),$$
$$\boldsymbol{\mu}_k = \Lambda_k^{-1}(\Lambda_{0k}\boldsymbol{\mu_0} + X_k'Y),$$
$$a = a_0 + \frac{n}{2}, \text{ and } b_k = b_0 + \frac{1}{2}(Y'Y + \boldsymbol{\mu_0}'\Lambda_{0k}\boldsymbol{\mu_0} - \boldsymbol{\mu}_k'\Lambda_k\boldsymbol{\mu}_k).$$

Focusing on the primary parameter of interest here, we obtain the coefficient β from the posterior of $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$ under each model M_k. If we let μ_{k2} be the second element of $\boldsymbol{\mu}_k$, and $(\Lambda_k^{-1})_{22}$ be the $(2,2)_{th}$ entry in Λ_k^{-1}, we obtain that β has univariate t distribution with $2a$ df and μ_{k2} and $\frac{b_k}{a}(\Lambda_k^{-1})_{22}$, respectively,

as the location and scale parameters, i.e.

$$\pi(\beta|Y, M_k) = \mu_{k2} + t_{2a}\sqrt{\frac{b_k}{a}(\Lambda_k^{-1})_{22}}, \tag{2.1}$$

where t_{2a} is the standard t distribution with $2a$ df. The normalizing constant for the posterior under model M_k is then

$$P(Y|M_k) = \frac{f(Y|\boldsymbol{\theta}, \sigma^2, M_k)\pi(\boldsymbol{\theta}|\sigma^2, M_k)\pi(\sigma^2|M_k)}{\pi(\boldsymbol{\theta}, \sigma^2|Y, M_k)} = \frac{1}{(2\pi)^{n/2}}\sqrt{\frac{|\Lambda_{0k}|}{|\Lambda_k|}}\frac{b_0^{a_0}\Gamma(a)}{b_k^a\Gamma(a_0)},$$

which leads to the Bayes factor between M_1 and M_2 as

$$BF_{12} = \sqrt{\frac{|\Lambda_2|}{|\Lambda_1|} \times \frac{|\Lambda_{01}|}{|\Lambda_{02}|}}\left(\frac{b_2}{b_1}\right)^a. \tag{2.2}$$

The BMA (Hoeting *and others*, 1999) of the two models takes the form

$$\pi(\boldsymbol{\theta}, \sigma^2|Y) = P(M_1|Y)\pi(\boldsymbol{\theta}, \sigma^2|Y, M_1) + P(M_2|Y)\pi(\boldsymbol{\theta}, \sigma^2|Y, M_2).$$

Let $P(Y)$ be the marginal probability of the data obtained after averaging over both models,

$$P(Y) = P(Y|M_1)P(M_1) + P(Y|M_2)P(M_2). \tag{2.3}$$

In the absence of prior information, it is customary to assume equal prior probabilities for the two models, i.e. $P(M_1) = P(M_2) = 0.5$. Therefore, we have

$$\begin{aligned}\pi(\boldsymbol{\theta}, \sigma^2|Y) &= \frac{P(Y|M_1)P(M_1)}{P(Y|M_1)P(M_1) + P(Y|M_2)P(M_2)}\pi(\boldsymbol{\theta}, \sigma^2|Y, M_1) \\ &\quad + \frac{P(Y|M_2)P(M_2)}{P(Y|M_1)P(M_1) + P(Y|M_2)P(M_2)}\pi(\boldsymbol{\theta}, \sigma^2|Y, M_2) \\ &= \frac{BF_{12}}{1 + BF_{12}}\pi(\boldsymbol{\theta}, \sigma^2|Y, M_1) + \frac{1}{1 + BF_{12}}\pi(\boldsymbol{\theta}, \sigma^2|Y, M_2).\end{aligned} \tag{2.4}$$

Note that the posterior distribution $\pi(\boldsymbol{\theta}, \sigma^2|Y)$, which we call *BMA posterior*, is a mixture of the two posterior distributions resulting from models M_1 and M_2. Because, it is not obtained from a given sampling distribution and a particular prior, it may not be a canonical posterior.

The BMA posterior relies on the Bayes factor as the weighting factor, favoring one model over using weights based on BF_{12}. Given an established association, we expect the Bayes factor provide evidence supporting one of the two models. Intuitively, if $BF_{12} > 1$ then we have more support for M_1 from the data and vice versa when $BF_{12} < 1$. For the priors considered here, we show that when data was generated from M_1, $Y = X_1\boldsymbol{\theta}_1 + \epsilon_1$, $BF_{12} \xrightarrow{p} \infty$ as $n \to \infty$ for any values of the hyperparameters, and similarly when $Y = X_2\boldsymbol{\theta}_2 + \epsilon_2$, $BF_{12} \xrightarrow{p} 0$ (Supplementary Materials available at *Biostatistics* online).

2.3. *BMA-based HPD interval for the genetic effect of a SNP*

To assess the genetic effect of a SNP based on the posterior distribution of β, the simplest approach is to use the posterior mode or mean of β as a point estimate. The HPD region, however, provides more information with an interval estimate. To calculate BMA-based HPD, we note that the posterior density of β from each of the M_1 and M_2 models is a univariate t with location and scale parameters as specified in equation (2.1). The BMA posterior of β is therefore a mixture of two known t distributions with the mixture proportion depending on BF_{12}. It is thus possible to calculate the exact HPD region for β.

A $(1-\alpha)$% HPD is defined as $R(c_\alpha) = \{\beta : \pi(\beta|Y) \geq c_\alpha\}$, where $\pi(\beta|Y)$ is the BMA posterior density of β and c_α is the threshold such that the area under the posterior density is $1-\alpha$. Depending on the similarity between the two posterior distributions corresponding to M_1 and M_2 for a given credible level α, a BMA HPD region can be either one single interval or made up of two disconnected intervals. In all examples, we have studied the HPD region is a single interval at $\alpha = 0.05$, due to the correlation between the two models (Supplementary Materials available at *Biostatistics* online). Specifically, let β_l and β_u to be the two solutions of $\pi^{-1}(c_\alpha)$. The $1-\alpha$ HPD region is then (β_l, β_u), where

$$\int_{\beta_l}^{\beta_u} \pi(\beta|Y)d\beta = 1-\alpha,$$
$$\pi(\beta_l|Y) = \pi(\beta_u|Y) = c_\alpha. \tag{2.5}$$

The closed form of $\pi(\beta|Y)$ is in fact available, thus we can solve the equations defined in (2.5) numerically to find c_α as well as β_l and β_u, using function `multiroot` in R package `rootSolve`. Note that for notation simplicity, we use α here to denote the desired credible level; its distinction from the intercept parameter, also denoted by α, should be clear from the context.

In practice, besides assessing association evidence for a single SNP, scientists are often interested in ranking multiple SNPs from a whole-genome scan and selecting the top ones for follow-up studies. The lower bounds of the HPD intervals can be used for this purpose. We will demonstrate the performance of this method in Section 3 using simulations, as well as in Section 4 where we rank over 14 000 X-chromosome SNPs studying their association evidence with meconium ileus in cystic fibrosis patients. In each setting, we compare the proposed ranking method with the frequentist method of Wang *and others* (2014) and the more conventional Bayes factor-based approach.

2.4. *Assessing genetic effect by Bayes factor*

In Bayesian framework, Bayes factor (Kass and Raftery, 1995; Stephens and Balding, 2009) is another important measure of evidence. In the presence of model uncertainty, we propose using the Bayes factor calculated by comparing the averaging model between M_1 and M_2 with the null model of no effect, M_N. Under the null model of $\beta = 0$, let $X_N = (\mathbf{1_n}, S)$ be the corresponding design matrix. Using the same prior distributions and hyperparameter values for the remaining parameters, σ^2, α, and γ, the calculation of $P(Y|M_N)$ is then similar to that of $P(Y|M_1)$ and $P(Y|M_2)$ as described in Section 2.2. Let

$$BF_{1N} = \frac{P(Y|M_1)}{P(Y|M_N)}, \quad BF_{2N} = \frac{P(Y|M_2)}{P(Y|M_N)}$$

be the Bayes factors comparing, respectively, the XCI M_1 and no XCI M_2 with the null model M_N, the Bayes factor for comparing the averaging model with the null model is defined as

$$BF_{AN} = \frac{P(Y|M_1)P(M_1) + P(Y|M_2)P(M_2)}{P(Y|M_N)}.$$

Because $P(M_1) = P(M_2) = 0.5$ in our setting, we thus have

$$BF_{AN} = \frac{1}{2}(BF_{1N} + BF_{2N}). \tag{2.6}$$

The Bayes factor BF_{AN} has similar asymptotic properties as BF_{12}. We show in the Supplementary Materials available at *Biostatistics* online that in our setting if $\lambda > 0$ (the precision parameter for β), then BF_{AN} converges in probability to either 0 or ∞, depending on whether $\beta = 0$ or not.

For finite sample, we find BF_{AN} computed following Section 2.2 is highly sensitive to the choice of λ. The sensitivity was noted by Raftery (1996, 1999). Because λ is often unknown in practice, we recommend approximating Bayes factors by Bayesian information criterion (BIC). Raftery (1996, 1999) has also noted that BIC provides a close approximation to the Bayes factor when $\lambda = 1$, which he called "unit information prior". If λ is known and not close to 1, BIC approximation may have an error of $O(1)$ (Kass and Raftery, 1995). Therefore, in practice, if there is sufficient evidence that λ should equal to any specific value, we recommend following Section 2.2 to compute BF_{12} and BF_{AN}. If there is little information about λ, we recommend using BIC approximation to avoid the complexity of choosing λ. In simulation and application studies below, we use BIC approximation for the more general case when λ is unknown.

2.5. *Binary outcomes*

When we measure binary responses, M_1 and M_2 are logistic regression models. Assuming the prior $\boldsymbol{\theta}_k \sim N(\boldsymbol{\mu_0}, \Lambda_{0k}^{-1})$, the BMA framework described above can still be used although computational complexities arise due to the lack of conjugacy. We use the R package `MCMCpack` to draw samples from the posterior distributions under M_1 and M_2. To obtain samples from the averaged model, we draw samples from M_1 with probability $BF_{12}/(1 + BF_{12})$ and from M_2 with probability $1/(1 + BF_{12})$ based on equation (2.4). And we use these samples to construct the $1 - \alpha$ HPD interval via the function `HPDinterval` in the R package `coda`.

The calculation of BF_{12} is based on the Bridge sampling method proposed by Meng and Wong (1996) and further refined by Gelman and Meng (1998) which we delineate below. Suppose we have J posterior samples, $\boldsymbol{\theta_{kj}}$, from the two models, $k = 1$ and 2 and $j = 1, ..., J$. For each parameter sample $\boldsymbol{\theta_{kj}}$, we can calculate the corresponding unnormalized posterior density based on the logistic model under the M_1 XCI assumption,

$$\begin{aligned} q_1(\boldsymbol{\theta_{kj}}) &= \pi(\boldsymbol{\theta_{kj}}|M_1) f(Y|\boldsymbol{\theta_{kj}}, M_1) \\ &= \pi_1(\boldsymbol{\theta_{kj}}) \prod_{i=1}^{n} p_{1i}(\boldsymbol{\theta_{kj}})^{Y_i} (1 - p_{1i}(\boldsymbol{\theta_{kj}}))^{1-Y_i}, \end{aligned}$$

where $p_{1i}(\boldsymbol{\theta_{kj}}) = [1 + \exp(-X_{1i}\boldsymbol{\theta_{kj}})]^{-1}$, and X_{1i} is the i_{th} row of the design matrix X_1 that contains the genotype data coded under model M_1 for the i_{th} individual. π_1 is the density function of $N(\boldsymbol{\mu_0}, \Lambda_{01}^{-1})$. Similarly, we obtain

$$\begin{aligned} q_2(\boldsymbol{\theta_{kj}}) &= \pi(\boldsymbol{\theta_{kj}}|M_2) f(Y|\boldsymbol{\theta_{kj}}, M_2) \\ &= \pi_2(\boldsymbol{\theta_{kj}}) \prod_{i=1}^{n} p_{2i}(\boldsymbol{\theta_{kj}})^{Y_i} (1 - p_{2i}(\boldsymbol{\theta_{kj}}))^{1-Y_i}, \end{aligned}$$

where $p_{2i}(\boldsymbol{\theta_{kj}}) = [1 + \exp(-X_{2i}\boldsymbol{\theta_{kj}})]^{-1}$ under model M_2, and π_2 is the density function of $N(\boldsymbol{\mu_0}, \Lambda_{02}^{-1})$. We then define the ratio of unnormalized densities as $l_{kj} = q_1(\boldsymbol{\theta_{kj}})/q_2(\boldsymbol{\theta_{kj}})$ and compute the Bayes factor iteratively. Specifically, we set $BF_{12}^{(1)} = 1$ and compute at the $(t+1)_{th}$ iteration until convergence,

$$BF_{12}^{(t+1)} = \frac{\sum_{j=1}^{J} \frac{l_{2j}}{l_{2j}+BF_{12}^{(t)}}}{\sum_{j=1}^{J} \frac{1}{l_{1j}+BF_{12}^{(t)}}}. \tag{2.7}$$

When comparing the averaged model vs. null model, the above procedure cannot be directly implemented to calculate BF_{1N} and BF_{2N}, since the null model has different dimension of parameter $\boldsymbol{\theta}$. Instead of finding the ratio of normalizing constants by the numerical method above, we find $P(Y|M_1)$, $P(Y|M_2)$, and $P(Y|M_N)$ by calculating the ratio between them and known quantities. The latter will be the normalizing constants corresponding to Gaussian approximations of the posterior distributions of interest. More precisely, we use the following steps:

- To calculate $P(Y|M_1)$, we approximate the posterior under M_1 using a multivariate normal distribution with independent components. So we find the sample mean and sample variance of posterior sample $\boldsymbol{\theta_{1j}} = (\alpha_{1j}, \beta_{1j}, \gamma_{1j})$, which are $(\bar{\alpha}_1, \bar{\beta}_1, \bar{\gamma}_1)$ and $\begin{pmatrix} s_{\alpha_1}^2 & 0 & 0 \\ 0 & s_{\beta_1}^2 & 0 \\ 0 & 0 & s_{\gamma_1}^2 \end{pmatrix}$.
- We simulate α'_{1j}, β'_{1j} and γ'_{1j} from the above multivariate approximation to the posterior whose normalizing constant is $c_1 = (2\pi)^{3/2} s_{\alpha_1} s_{\beta_1} s_{\gamma_1}$ and set $\boldsymbol{\theta'_{1j}} = (\alpha'_{1j}, \beta'_{1j}, \gamma'_{1j})$.
- We use the iterative approach in equation (2.7) to compute the ratio of normalizing constants between the posterior under M_1 and the corresponding approximation, $BF_1 = P(Y|M_1)/c_1$. Since c_1 is known, we can easily derive the normalizing constant $P(Y|M_1)$.
- To calculate $P(Y|M_N)$, we repeat the procedure used for $P(Y|M_1)$ but this time the dimension of the parameter is two instead of three.
- The unnormalized posterior density for M_N is

$$\begin{aligned} q_N(\boldsymbol{\theta_{Nj}}) &= \pi(\boldsymbol{\theta_{Nj}}|M_N) f(Y|\boldsymbol{\theta_{Nj}}, M_N) \\ &= \pi_N(\boldsymbol{\theta_{Nj}}) \prod_{i=1}^{n} p_{Ni}(\boldsymbol{\theta_{Nj}})^{Y_i} (1 - p_{Ni}(\boldsymbol{\theta_{Nj}}))^{1-Y_i}, \end{aligned}$$

  where $p_{Ni}(\boldsymbol{\theta_{Nj}}) = [1 + \exp(-X_{Ni}\boldsymbol{\theta_{Nj}})]^{-1}$, and π_N is the prior density of $N(\mathbf{0}, \Lambda_{0N}^{-1})$.
- We then use equation (2.7) to compute $BF_N = P(Y|M_N)/c_N$, where $c_N = 2\pi s_{\alpha_1} s_{\gamma_1}$, and we obtain BF_{1N} as

$$BF_{1N} = \frac{BF_1}{BF_N} \times \frac{c_1}{c_N}.$$

- We repeat the above steps for M_2 to calculate BF_{2N}.
- Finally, we use equation (2.6) to calculate BF_{AN} by averaging BF_{1N} and BF_{2N}.

2.6. *Revisit the maximum likelihood approach*

Let Z_1 and Z_2 be the frequentist's test statistics for testing $\beta_k = 0$ derived from the two regression models, $Y = \alpha_k + \beta_k G_k + \gamma_k S + \epsilon_k, k = 1$ and 2, respectively under the XCI M_1 and no XCI M_2 assumptions. The maximum likelihood approach of Wang *and others* (2014), in essence, uses $Z_{max} = max(|Z_1|, |Z_2|)$ as the test statistic and calculates the p-value of Z_{max} empirically via a permutation-based procedure. We note that the significance of Z_{max} can be obtained more efficiently by recognizing that Z_1 and Z_2 have an approximate bivariate normal distribution under the null hypothesis of no association (Supplementary Materials available at *Biostatistics* online). This principle has been used in another setting where for an un-genotyped SNP, instead of imputing the missing genotype data, the association statistic is directly inferred based on the association statistic at a genotyped SNP and the correlation between the two SNPs estimated from a reference sample (Lee *and others*, 2013; Pasaniuc *and others*, 2014). In the simulation study below and in the application study of Section 4, for each simulated SNP and each of the 14 000 or so SNPs analyzed, we will obtain the corresponding p-value using this method because of the computational cost for assessing p-values less than 10^{-6}.

3. Simulation study

We conduct simulation studies to evaluate the performance of the proposed BMA methods and the frequentist method of Wang *and others* (2014), for studying both normally distributed traits and binary outcomes.

3.1. *Simulation settings*

In our simulations, we vary the sample size n, proportion of males and frequencies of allele D for males and females (p_m and p_f, respectively). In each case, we first generate data for G, where we simulate female genotypes using a multinomial distribution with probabilities of $(1-p_f)^2$, $2p_f(1-p_f)$ and p_f^2, respectively, for dd, dD, and DD, and we simulate male genotypes using a binomial distribution with probabilities of $(1-p_m)$ and p_m, respectively, for d and D.

We then generate outcome data for Y based on the simulated G coded under the XCI M_1 or no XCI M_2 assumption, and various parameter values of the regression models. For linear models we fix $\alpha = 0$ and $\gamma = 0$; the intercept parameter has negligible effects on result interpretation (e.g. $\alpha = 1$ lead to similar conclusion). Since the effect of sex is not of primary interest here, we set $\gamma = 0$ without loss of generality. We also fix $\sigma^2 = 1$. Under the null model, $\beta = 0$ and Y does not depend on the XCI and no XCI assumptions, i.e. $Y \sim N(0, \sigma^2 I_n)$. Under alternatives and for each M_k, method performance depends on both genetic effect size β and allele frequencies p_m and p_f, via the quantity EV, the variation of Y explained by genotype, where $EV = Var(E(Y|G))/Var(Y)$. Although allele frequency affects method performance as we will see in the application study below, fixing EV instead of β has the benefit of not requiring specification of the relationship between β and allele frequency (e.g. variants with lower frequencies tend to have bigger effects or smaller effects, vs. β and allele frequency are independent of each other); Derkach *and others* (2014) explored this in a frequentist setting for jointly analyzing multiple autosome SNPs. For linear models, it is easy to show that $EV = \beta^2\sigma_G^2/(\beta^2\sigma_G^2 + \sigma^2)$, where σ_G^2 is the variance of G depending on p_m and p_f. Thus, for a given EV value we obtain $\beta = \sigma/\sigma_G \cdot \sqrt{EV/(1-EV)}$ for different values of p_m, p_f and codings of G based on the M_1 XCI or M_2 no XCI assumption. We then simulate Y for continuous outcomes from $N(X_k\boldsymbol{\theta}, \sigma^2 I_n)$ based on $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$ and $X_k = (\mathbf{1}_n, G_k, S)$.

For studying binary outcomes using logistic regression, we assume the typical study design of equal numbers of cases and controls. Under the null of $\beta = 0$, we randomly assign $Y = 0$ to half of the sample and $Y = 1$ to the other half. Under alternatives, the derivation of β given EV and allele frequencies is a bit more involved, and we outline the details in the Supplementary Materials available at *Biostatistics*

online. We then simulate Y from $Bino(n^*, (1 + \exp(-X_k\boldsymbol{\theta}))^{-1})$, $n^* > n$, until $n/2$ numbers of cases and controls are generated.

To summarize, the parameters involved in the simulation studies include the sample size (n and the proportion of males), allele frequencies in males and females (p_m and p_f), the variation of Y explained by genotype (EV and in turn β; without loss of generality, $\alpha = 0$, $\gamma = 0$, $\sigma^2 = 1$), as well as equal numbers of cases and controls for studying binary traits. The number of MCMC samples for analyzing each binary dataset is $J = 1000$.

3.2. *Results*

Figure 1 shows representative results when $n = 1000$ (and assuming the proportion of males is half), and the minor allele frequency (MAF) of the associated SNP ranges from 0.01 to 0.5. The top panel shows the results when β, the regression coefficient in the linear model (also known as the genetic effect size) is set to be 0.3. In that case, the ability of a method (frequentist or the proposed BMA) to identify an associated SNP depends on the MAF of the SNP. Indeed, results in the top panel of Figure 1 show that as the MAF increases, $-log_{10}$ p-value (top left graph) and $log_{10}BF_{AN}$ (top middle graph) increase, and the BMA-based 95% HPD intervals (top right graph) become narrower and the corresponding lower bounds are further away from zero. Note that for easy of presentation, the results here are the averages across 100 independently simulated datasets; results of each of the 100 simulation replicates are provided in the Supplementary Materials available at *Biostatistics* online. Results for no XCI, binary traits and other parameter values (i.e. different β and MAF values) are also provided in the Supplementary Materials available at *Biostatistics* online.

When β is fixed, we observed that rankings of SNPs based on frequentist p-values or the proposed lower bounds of BMA HPD intervals are quite similar (top panel of Figure 1). However, this is not the case when EV, the phenotypic variation explained by SNP genotype, is fixed (bottom panel of Figure 1). When EV is fixed, SNPs with lower allele frequencies have stronger effects (larger β) and intuitively they should be ranked higher. However, the p-values (bottom left graph) are quite similar across the allele frequencies. This is also the case for BF_{AN} (bottom middle graph) if we use the Bayes facto to rank SNPs. On the other hand, method based on the BMA HPD intervals (bottom right graph) exhibits superior performance, where the lower bound is further away from zero for larger effect size β while a smaller MAF is reflected by a wider interval. A frequentist confidence interval can be easily constructed under one given model, but an weighted average CI is inherently difficult to derive under the frequentist paradigm.

An astute reader may notice that the true value of β is not in the center of each BMA-based HPD interval. When data are simulated from X-inactivated models, β is to the right of the center (Figure 1); when the simulation model is X not inactivated, β is to the left (Supplementary Materials available at *Biostatistics* online). This is because these HPD intervals are computed under the averaged model rather than the true simulation model. When $n = 1000$, BF_{12} is not close to 0 or ∞ and both XCI and no XCI models have non-zero weights. As we show in the Supplementary Materials available at *Biostatistics* online, BF_{12} converges to either 0 or ∞ as $n \to \infty$, which implies the averaged HPD intervals also converges to the HPD intervals under the true model. This theoretical justification further supports the use of BMA-based HPD intervals for inference of β and ranking of SNPs.

4. APPLICATION STUDY

Sun *and others* (2012) performed a whole-genome association scan on meconium ileus, a binary intestinal disease occurring in about 20% of the individuals with cystic fibrosis. Their GWAS included X-chromosome but assumed the inactivation M_1 model. They identified a gene called *SLCA14* to be associated with meconium ileus, and in their Table 2 they reported p-values in the range of 10^{-12}, 10^{-8},

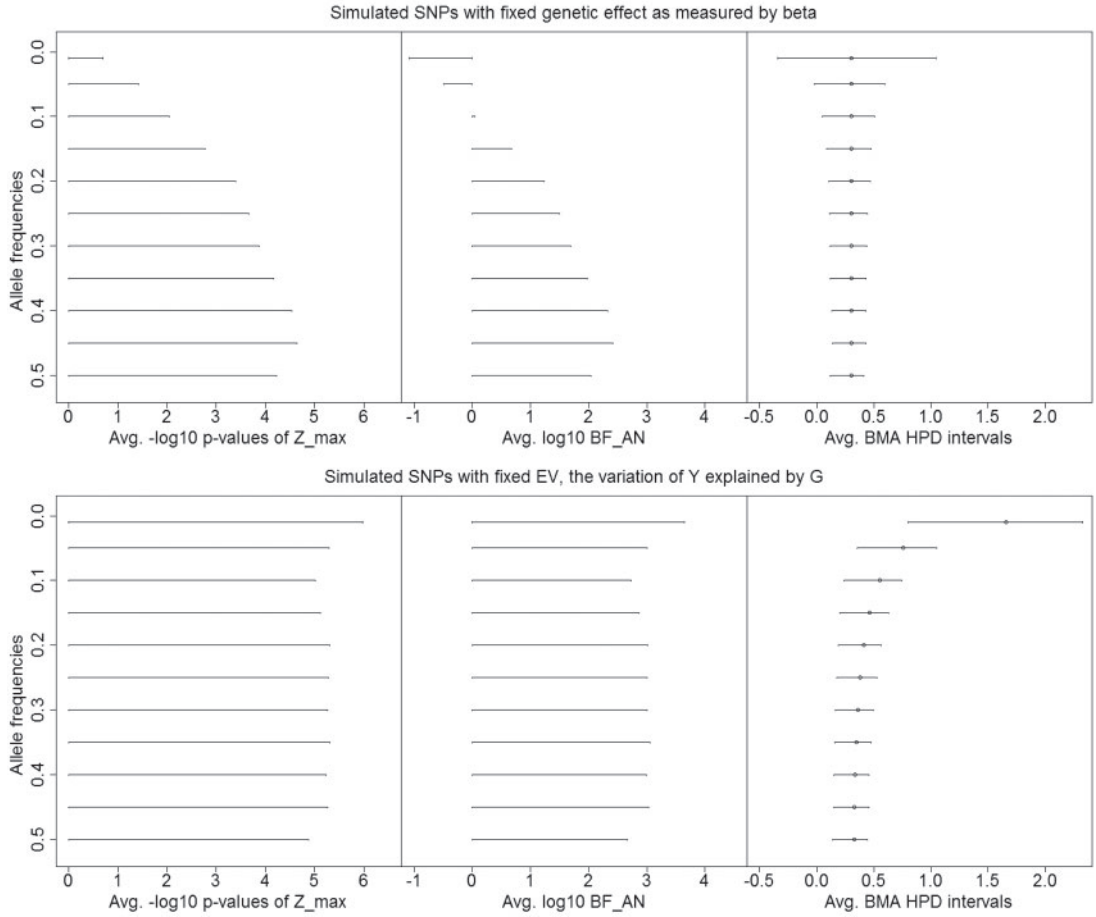

Fig. 1. Simulation results averaged across 100 independently simulated datasets. Left panel: frequentist $-log_{10}$ p-values of Z_{max} (based on the approximate asymptotic distribution in Section 2.6 instead of the original permutation method of Wang *and others* (2014) because of the prohibitive computational cost in this setting). Middle panel: the $log_{10}BF_{AN}$, comparing the Bayesian averaged model with the null model of no association. Right panel: BMA-based 95% HPD intervals for β, the genetic effect size of the associated SNP. The allele frequency of the SNP ranges from 0.01 to 0.5, shown in the Y-axis. Top row: the effect size is fixed at $\beta = 0.3$. Bottom row: the explained variance by the SNP is fixed at $EV = 0.02$, and thus the corresponding effect size varies depending on the allele frequency. The circles mark the true values of the β in each setting. The outcome here is a normally distributed trait simulated under the true model of X-chromosome inactivation (XCI). Results for no XCI, binary traits and other parameter values are provided in the Supplementary Materials available at *Biostatistics* online.

and 10^{-6}, respectively, for *rs*3788766, *rs*5905283, and *rs*12839137 from the region. We revisited this data by applying the maximum likelihood approach and the proposed Bayesian model average method.

The data consists of $n = 3199$ independent CF patients, and there are slightly more males ($n_m = 1722$, 53.8%) than females ($n_f = 1477$, 46.2%). Among the study subjects, 574 are cases with meconium ileus and 2625 are controls, and the rates of meconium ileus do not appear to differ between the male and female groups (17.7% vs. 18.3%). Genotypes are available for 14 280 X-chromosome SNPs, but 60 are monomorphic (no variation in the genotypes within the sample). Thus, the association analyses were performed between 14 220 X-chromosome SNPs and the binary outcome of interest. By convention, for

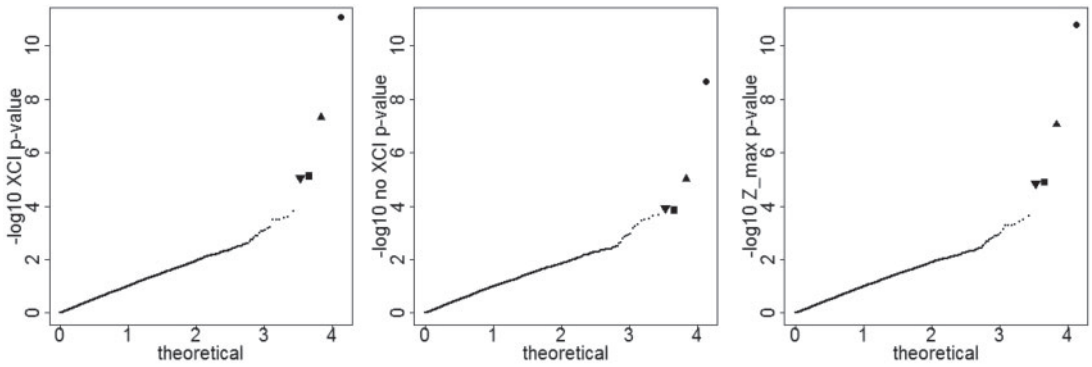

Fig. 2. QQplots of $-log_{10}$ p-values of analyzing association evidence between 14 220 X-chromosome SNPs and meconium ileus in 3199 cystic fibrosis patients, under the XCI M_1 assumption (left), the no-XCI M_2 assumption (middle), and using Z_{max} (right). Circle (●) for *rs*3788766, up-pointing triangle (▲) for *rs*5905283, square (■) for *rs*12839137, and down-pointing triangle (▼) for *rs*5905284.

each SNP we assumed the minor allele as the risk allele D and we used the two coding schemes of G_1 and G_2 under the XCI M_1 and no XCI M_2 models.

Figure 2 shows the QQplots of p-values obtained using the frequentist framework. The left graph is under the XCI M_1 assumption as in the original analysis of Sun *and others* (2012). The middle graph is under the no XCI M_2 assumption, and the right one is based on the adjusted minimal p-value of the maximum likelihood approach (Wang *and others*, 2014). It needs to be noted that the original permutation-based is computationally prohibitive for estimating p-values as small as 10^{-12} as in our case. Alternatively, we used the approximate asymptotic distribution for Z_{max} (Section 2.6 and Supplementary Materials available at *Biostatistics* online). As expected, most of the SNPs are from the null, but there are four clear outliers/signals with evidence for association with meconium ileus regardless of the methods used. Contrasting the left graph with the middle one in Figure 2 shows that the XCI M_1 assumption lead to smaller p-values for these four SNPs than the no XCI M_2 assumption.

Figure 3 presents the Bayesian results for the top 50 ranked SNPs, as well as the corresponding p-values. Similarly to the presentation of the simulation results in Section 3, the left graph shows the $-log_{10}$ p-values of Z_{max}, while the middle one is for $log_{10}BF_{AN}$, and right one is for the BMA-based 95% HPD intervals. Note that for ease of presentation and without loss of generality, we mirrored all negative intervals to positive ones. Table 1 provides results for the first 15 of the top 50 ranked SNPs.

Several important remarks can be made here. First, the proposed Bayesian method clearly identifies the four SNPs suggested by the p-value approach. Second, the Bayesian framework in this setting provides more feature-rich quantities such as the BMA-based HPD intervals, and it pinpoints additional SNPs that merit follow-up studies. Note that although p-values lead to similar rankings between the two models themselves, they could miss potentially important SNPs. Taking *rs*12689325 as an example, this SNP is ranked 331 based on the p-value of 0.0268, the p-value of the maximum test statistic calculated under M_1 and M_2. However, this SNP is ranked second based on the lower bounds of the BMA-based HPD intervals averaged over M_1 and M_2 (the first set of red thick lines in Figure 3). The wide BMA HPD interval is a result of small MAF (1.3%) coupled with a moderate effect size. Similar results are obtained for *rs*12845594, the fourth ranked SNP based on the BMA-based HPD intervals. This result is consistent with that of simulations in Section 3, where we demonstrated that the HPD intervals may have stronger ability to identify truly associated SNPs with large effect sizes but small MAFs. Also consistent is the observation that the conventional Bayes factor is one single measure of evidence that can be complemented by an interval measure. Given a trait of interest in practice, if genetic etiology implies

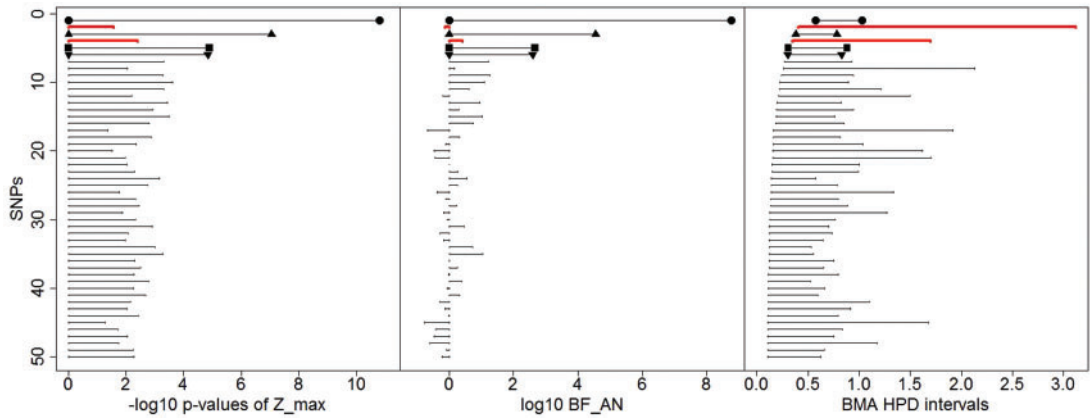

Fig. 3. Application results for 50 top ranked SNPs, selected from analyzing association evidence between 14 220 X-chromosome SNPs and meconium ileus in 3199 cystic fibrosis patients. SNPs are ordered by their lower bounds of the BMA-based HPD intervals. Left panel: frequentist $-log_{10}$ p-values of Z_{max} (based on the approximate asymptotic distribution in Section 2.6 instead of the original permutation method of Wang *and others* (2014) because of the prohibitive computational cost in this setting). Middle panel: the $log_{10}BF_{AN}$, comparing the Bayesian averaged model with the null model of no association. Right panel: BMA-based 95% HPD intervals for β, the genetic effect size of the associated SNP. The four top SNPs identified by p-values are marked here using the same symbol: black circle (●) for *rs*3788766, up-pointing triangle (▲) for *rs*5905283, square (■) for *rs*12839137, and down-pointing triangle (▼) for *rs*5905284. SNP *rs*12689325 (the second ranked SNP) and *rs*12845594 (the fourth ranked SNP) discussed in Section 4 are marked in thick lines.

Table 1. *Summary of frequentist and Bayesian analysis of the 15 top ranked SNPs, selected from analyzing association evidence between 14 220 X-chromosome SNPs and meconium ileus in 3199 cystic fibrosis patients*

		Log odds ratios		Frequentist P-values			BMA		
SNPs	MAF	M_1	M_2	M_1	M_2	Z_{max}	HPD interval	BF_{12}	BF_{AN}
rs3788766	0.388	−0.798	−0.484	8.50e−12	2.20e−09	1.61e−11	(0.572–1.033)	271	5.84e+08
rs12689325	0.013	−1.615	−1.386	4.02e−02	1.99e−02	2.68e−02	(0.405–3.118)	0.377	7.56e−01
rs5905283	0.487	−0.586	−0.326	4.79e−08	9.64e−06	8.88e−08	(0.379–0.784)	201	3.58e+04
rs12845594	0.047	−0.990	−0.546	2.73e−03	1.08e−02	3.93e−03	(0.344–1.700)	7.23	2.61e+00
rs12839137	0.237	−0.611	−0.360	7.55e−06	1.43e−04	1.25e−05	(0.307–0.884)	22.5	4.77e+02
rs5905284	0.249	−0.592	−0.358	8.61e−06	1.21e−04	1.43e−05	(0.302–0.830)	18.3	3.99e+02
rs579854	0.136	−0.642	−0.424	3.31e−04	5.66e−04	5.08e−04	(0.266–0.932)	1.88	1.70e+01
rs5955417	0.030	−1.229	−0.710	6.42e−03	1.25e−02	9.02e−03	(0.260–2.130)	3.28	1.50e+00
rs12720074	0.100	−0.715	−0.529	6.61e−04	3.49e−04	5.29e−04	(0.237–0.943)	0.533	1.84e+01
rs1921965	0.091	0.611	0.440	1.57e−04	2.30e−04	2.41e−04	(0.228–0.893)	1.34	1.22e+01
rs6623182	0.036	0.867	0.552	3.32e−04	1.28e−03	4.97e−04	(0.217–1.216)	2.89	4.17e+00
rs3027514	0.015	0.976	0.740	5.41e−03	4.58e−03	6.46e−03	(0.209–1.496)	0.834	6.23e−01
rs17338514	0.099	0.574	0.419	2.46e−04	3.07e−04	3.75e−04	(0.201–0.821)	1.19	9.00e+00
rs11797786	0.068	0.618	0.383	8.17e−04	5.29e−03	1.21e−03	(0.191–0.947)	4.68	1.96e+00
rs1921967	0.122	0.531	0.393	2.67e−04	2.12e−04	3.26e−04	(0.190–0.756)	0.750	1.07e+01

MAF is the pooled estimate of the frequency of the minor allele (frequencies do not differ between males and females); log odds ratio estimates under the XCI M_1 and no XCI M_2 assumptions. Frequentist results includes p-values corresponding to M_1 and M_2, and the bias-adjusted p-value of Z_{max} of the maximum likelihood approach (Wang *and others*, 2014). The adjusted p-values were obtained using the approximate asymptotic distribution (Section 2.6) instead of the original permutation-based because of the prohibitive computational cost in this setting. Results of the proposed approach include BMA-based 95% HPD intervals, Bayes factors BF_{12} comparing the XCI M_1 model with the no XCI M_2 model, as well as BF_{AN} comparing the average model with the null model.

the involvement of rare variants, the Bayesian results suggest that these two SNPs warrant additional investigation.

5. DISCUSSION

We propose a Bayesian approach to address the ambiguity involved in GWAS and NGS studies of SNPs situated on the X-chromosome. Depending on whether X-inactivation takes place or not, there are two regression models that can be used to explore the genetic effect of a given SNP on the phenotype of interest. The proposed method allows us to produce posterior-based inference that incorporates the uncertainty within and between genetic models. While the former is quantified by the posterior distribution under each model, the latter can be properly accounted for by considering a weighted average of the model-specific estimators. Following the Bayesian paradigm, the weights are proportional to the Bayes factor comparing the two competing models. The asymptotic properties of the Bayes factors considered in this article for linear models are included in the Supplementary Materials available at *Biostatistics* online. In the binary response case, the theoretical study is difficult due to the intractable posteriors, but the Monte Carlo estimators exhibit good properties in all the numerical studies performed.

The use of g-priors in this study setting is essential in that it allows us to avoid the effect of covariate rescaling on the Bayes factors, yet maintain results interpretation. In regression models, we know that the effect size β is inversely proportional to the size of the covariate value/genotype coding. Given a set of data, using $X/2$ or X should lead to identical inference. However, without g-priors, a model with smaller covariate value would be preferred based on BF. In our setting, the female component of the design matrix under the XCI M_1 coding is only half of that no XCI M_2 coding; male codings are the same for the two models. Consider the null case of $\beta = 0$ when the two competing models are identical. Using $\Lambda_{0k} = \lambda I$ for the precision of $\boldsymbol{\theta_k}$, we observed in our simulations that 80% of BF_{12} are greater than one, suggesting M_1 is preferred simply because of its smaller genotype coding. One statistical solution is to rescale the design matrix prior to the Bayesian inference. However, it is important to note that the coding difference for females is driven by a specific biological consideration, thus rescaling leads to difficulties in results interpretation. Instead, we use a g-prior in Section 2. Indeed, simulation results for the null case show that $BF_{12} > 1$ in about 50% of replicates, indicating proper calibration.

In our application, we did not observe a significant effect of sex. However, we note that the sex covariate S should be always included in association analysis of SNPs from the X-chromosome. Besides genetic epidemiological arguments, there is a strong statistical justification. For autosomes, the choice of the reference allele for coding of G only changes the sign of β but does not affect statistical significance. However, we note that this is not the case for analyzing X-chromosome SNPs under the no XCI M_2 model assumption; inference is identical under the XCI M_1 model. Interestingly, we can show that including S as a covariate resolves the issue. To see this, let G_2^* be the new coding of G_2 when the reference allele is switched. Because $G_2^* = 2 - G_2 - S$, switching reference allele in a regression model that includes S is then equivalent to changing the sign of β.

In our simulation and application studies, we focused on additive genetic models because of the earlier literature, most notably the work by Hill *and others* (2008). Both the frequentist and the proposed Bayes methods, however, can be readily applied to other genetic models such as the dominant, recessive and genotypic models. Consider the most general two degrees of freedom of the genotypic model, $Y \sim G_A + G_D + S$, where G_D represents the dominant effects, equal to 1 for genotype dD and 0 otherwise, and G_A has the same additive coding under XCI and no XCI assumptions as before. Supplementary Figures E.1 available at *Biostatistics* online show that results of the genotypical and additive genetic models are largely consistent in the application study, where *rs*3788766, *rs*5905283, *rs*12839137, and *rs*5905284 remain clearly associated. However, results also show differences for some of the lower ranked

SNPs, suggesting that the conventional choice of additive genetic model needs future investigations for both the X-chromosome and autosomes.

When the allele frequency is on the boundary, we have commented that the resulting HPD intervals can be quite wide as seen in the application above, e.g. *rs*12689325 with MAF of 1.3%, the second ranked SNP in Figure 3; ranked 331 by the minimal *p*-value approach. Among the 14,220 X-chromosome SNPs analyzed in Section 4, 829 SNPs have MAF less than 1%. In that case, there is little variation in the genotype variable thus limited information available for inference. The top ranked SNPs thus were chosen from the remaining 13 391 SNPs with MAF greater than 1%. In recent years, joint analyses of multiple rare (or common) variants (also known as the gene-based analyses) have received much attention but only for autosome SNPs (Derkach *and others*, 2014). Extension to X-chromosome SNPs remain an open question. Similarly, additional investigations are needed for X-chromosome SNPs in the areas of family-based association studies (Thornton *and others*, 2012), direct interaction studies (Cordell, 2009), as well as indirect interaction studies via scale-test for variance heterogeneity (Soave and Sun, 2017).

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgements

The authors would like to thank Dr Lisa J. Strug and her lab for providing the cystic fibrosis application data, and Prof Mike Evans for suggestions that have improved the presentation of the article. The author would also like to thank the Associate Editor for the constructive comments and suggestions that substantially improved the article. *Conflict of Interest:* None declared.

Funding

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-249547 and RGPIN-250053) to R.V.C. and L.S., and the Canadian Institutes of Health Research (CIHR 201309MOP-310732-G-CEAA-117978) to L.S.

References

Clayton, D. G. (2008). Testing for association on the X chromosome. *Biostatistics* **9**, 593–600.

Clayton, D. G. (2009). Sex chromosomes and genetic association studies. *Genome Medicine* **1**, 110.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics* **10**, 392–404.

Craiu, R. V. and Sun, L. (2014). Chapter 9: Bayesian methods in Fisher's statistical genetics world. In: Lawless, J. F. (editor), *Statistics in Action: A Canadian Outlook*. Boca Raton, FL: CRC Press, pp. 147–161.

Derkach, A., Lawless, J. F. and Sun, L. (2014). Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science* **29**, 302–321.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 45–97.

Gelman, A. and Meng, X. L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163–185.

Gendrel, A. V. and Heard, E. (2011). Fifty years of X-inactivation research. *Development* **138**, 5049–5055.

HEID, I. M., JACKSON, A. U., RANDALL, J. C., WINKLER, T. W., QI, L., STEINTHORSDOTTIR, V., THORLEIFSSON, G., ZILLIKENS, M. C., SPELIOTES, E. K., MAGI, R. *and others*. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genetics* **42**, 949–960.

HICKEY, P. F. AND BAHLO, M. (2011). X chromosome association testing in genome wide association studies. *Genetic Epidemiology* **35**, 664–670.

HILL, W. G., GODDARD, M. E. AND VISSCHER, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* **4**, e1000008.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. AND VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382–401.

KASS, R. E. AND RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

KONIG, I. R., LOLEY, C., ERDMANN, J. AND ZIEGLER, A. (2014). How to include chromosome X in your genome-wide association study. *Genetic Epidemiology* **38**, 97–103.

LEE, D., BIGDELI, T. B., RILEY, B. P., FANOUS, A. H. AND BACANU, S. A. (2013). DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* **29**, 2925–2927.

LOLEY, C., ZIEGLER, A. AND KONIG, I. R. (2011). Association tests for X-chromosomal markers—a comparison of different test statistics. *Human Heredity* **71**, 23–36.

MENG, X. L. AND WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica* **6**, 831–860.

PASANIUC, B., ZAITLEN, N., SHI, H., BHATIA, G., GUSEV, A. AND et al. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**(20), 2906–2914.

RAFTERY, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**, 251–266.

RAFTERY, A. E. (1999). Bayes factor and BIC: Comment on "A critique of the Bayesian Information Criterion for model selection". *Sociological Methods & Research* **27**, 411–427.

SASIENI, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

SOAVE, D. AND SUN, L. (2017). A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics* **73**, 960–971.

STEPHENS, M. AND BALDING, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics* **10**, 681–690.

SUN, L., ROMMENS, J. M., CORVOL, H., LI, W., LI, X. AND et al. (2012). Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nature Genetics* **44**, 562–569.

TESLOVICH, T. M., MUSUMURU, K., SMITH, A. V., EDMONDSON, A. C., STYLIANOU, I. M. AND et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713.

THORNTON, T., ZHANG, Q., CAI, X. C., OBER, C. AND MCPEEK, M. S. (2012). XM: Association testing on the X-chromosome in case-control samples with related individuals. *Genetic Epidemiology* **36**, 438–450.

WANG, J., YU, R. AND SHETE, S. (2014). X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. *Genetic Epidemiology* **38**, 483–493.

WELTER, D., MACARTHUR, J., MORALES, J., BURDETT, T., HALL, P., JUNKINS, H., KLEMM, A., FLICEK, P., MANOLIO, T., HINDORFF, L. AND PARKINSON H. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**, D1001–D1006.

WISE, A. L., GYI, L. AND MANOLIO, T. A. (2013). eXclusion: toward integrating the X chromosome in genome-wide association analyses. *American Journal of Human Genetics* **92**, 643–647.

WRIGHT, J. H. (2008). Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics* **146**, 329–341.

ZELLNER, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In: Goel, P. K. and Zellner, A. (editors), *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti.* Amsterdam: North-Holland, pp. 233–243.

ZHENG, G., JOO, J., ZHANG, C. AND GELLER, N. L. (2007). Testing association for markers on the X chromosome. *Genetic Epidemiology* **31**, 834–843.

[*Received June 24, 2017; revised June 24, 2018; accepted for publication June 30, 2018*]