Joint Analysis of Strain and Parent-of-Origin Effects for Recombinant Inbred Intercrosses Generated from Multiparent Populations with the Collaborative Cross as an Example

Yanyan Liu,* Sican Xiong,* Wei Sun,† and Fei Zou‡

*School of Mathematics and Statistics, Wuhan University, Hubei 430072, China, †Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, and ‡Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina 27599

ORCID IDs: 0000-0002-8304-5663 (S.X.); 0000-0002-6350-1107 (W.S.)

ABSTRACT Multiparent populations (MPP) have become popular resources for complex trait mapping because of their wider allelic diversity and larger population size compared with traditional two-way recombinant inbred (RI) strains. In mice, the collaborative cross (CC) is one of the most popular MPP and is derived from eight genetically diverse inbred founder strains. The strategy of generating RI intercrosses (RIX) from MPP in general and from the CC in particular can produce a large number of completely reproducible heterozygote genomes that better represent the (outbred) human population. Since both maternal and paternal haplotypes of each RIX are readily available, RIX is a powerful resource for studying both standing genetic and epigenetic variations of complex traits, in particular, the parent-of-origin (PoO) effects, which are important contributors to many complex traits. Furthermore, most complex traits are affected by >1 genes, where multiple quantitative trait locus mapping could be more advantageous. In this paper, for MPP-RIX data but taking CC-RIX as a working example, we propose a general Bayesian variable selection procedure to simultaneously search for multiple genes with founder allelic effects and PoO effects. The proposed model respects the complex relationship among RIX samples, and the performance of the proposed method is examined by extensive simulations.

KEYWORDS imprinting Bayesian variable selection parameter-expanded Gaussian priors Multiparental Populations MPP

Most human genes have functional mouse counterparts, and genomes of both organisms are organized similarly. Thus, mouse serves as a good model organism for complex human diseases. Recombinant inbred (RI) mice are among the major mouse resources in biomedical and genetic research. However, traditional mouse RI lines are derived from only two inbred parental strains, with limited a number of lines available, resulting in a low percentage (15%) of genetic variation across all mouse inbred strains (Yuan et al. 2011) and extensive blind spots where a sizable proportion of the genome is identical by descent. These limitations make traditional RI insufficient for studying complex traits. With the need for more powerful resources, multiparent populations (MPP) (de Koning and McIntyre 2017), a set of inbred lines using multiple lines as founders, can overcome the limitations of traditional RI lines and have become an innovative tool for fine quantitative trait locus (QTL) mapping. In the past 15 yr, different kinds of MPP have been established in plants and animals, such as nested association mapping design (Yu et al. 2008) and multiparent advanced generation intercross (MAGIC) populations (Cavanagh et al. 2008; Kover et al. 2009; Huang et al. 2015; Ladejobi et al. 2016) in plants, and the Drosophila Synthetic Population Resource (King et al. 2012) in animals. The mouse MPP include the collaborative cross (CC) (Complex Trait Consortium 2004), in which a genetically diverse set of eight inbred strains were selected as breeding founders (Iraqi et al. 2008) to generate a large number of RI lines. The eight founder strains were predicted to represent ~90% of the genetic variation presented in laboratory mice (Threadgill and Churchill 2012).
The CC thus greatly overcomes the limitations of the traditional mouse RI lines, and is the only mammalian resource with high genome-wide genetic variation that is uniformly distributed across a large, heterogeneous, and infinitely reproducible population (Threadgill et al. 2011).

Through the generation of RI intercrosses (RIX) of RI lines, a large number of potential “outbred” RIX samples can be generated. That is, given $L$ RI lines, $L(L-1)$ reciprocal RIX and $L(L-1)/2$ nonreciprocal $F_1$ individuals, or RIX, can be produced. An example of using CC-RIX to illustrate a distinct response to Ebola virus infection, by Rasmussen et al. (2014), was summarized by the editor: “the CC-RIX mice could prove valuable for preliminary screens of candidate therapeutics and vaccines.” Since all parental RI lines are isogenic at each locus, genotypes of RIX can be imputed in advance from those of their parental RI lines. Additional advantages of RIX can be found in Threadgill and Churchill (2012), in particular, its power in support of analysis of parent-of-origin (PoO) effects, where effects of certain alleles are different depending on whether those alleles are inherited maternally or paternally. PoO makes a significant contribution to the heritability of most complex traits (Mott et al. 2014). In addition, genomic PoO effects provide a great model to study epigenetic regulation of gene expression (Barlow 2011).

During the past 20 yr, QTL mapping methods, including analysis of variance (ANOVA), interval mapping (Lander and Botstein 1989), composite interval mapping (Jansen and Stam 1994; Zeng 1994), and multiple interval mapping (Kao and Zeng 1997; Kao et al. 1999), have been well developed using experimental crosses, such as backcross, $F_2$, and RI (see Broman (2001) for reviews), and many excellent open software packages, such as QTLCart (Basten et al. 1999), MapManager (Manly and Olson 1999), and R/qtl (Broman et al. 2003), are freely available online. For diallel data obtained from the CC founder strains, Lenarcic et al. (2014) employed a general Bayesian model for decomposing phenotypic variance into biologically intuitive components. Crowley et al. (2014) also applied a Bayesian method to a diallel cross of the eight founder strains to estimate genome-wide genetic and PoO effects (not QT LI mapping) on responses to haloperidol, an antipsychotic drug. Zhang et al. (2014) proposed a more flexible, nonparametric single QTL mapping method for detecting QTL with time-varying coefficients. Hallin et al. (2016) applied a random-effects model to map single QTL by partitioning the variance of growth traits across differing environments of yeast strains into additive, dominant, and pairwise epistatic components. However, many complex traits are affected by more than one gene, and multiple QTL mapping may be more powerful than locus-by-locus analysis. For CC-RIX data, Yuan et al. (2011) constructed a mixed-effects model to simultaneously map multiple QTL, but treated the eight CC founder alleles as standard biallelic single-nucleotide polymorphisms (SNPs). For many complex traits, it is arguable that modeling the effects of the eight founder alleles could lead to improved mapping power (Collaborative Cross Consortium 2012; Vered et al. 2014). In addition, for complex traits where PoO effects are suspected, modeling such effects may further improve QTL mapping power (Lawson et al. 2013) and understanding of etiologies of complex traits (Threadgill and Churchill 2012).

In this paper, for MPP-RIX data, we develop a Bayesian variable selection procedure to simultaneously map multiple genes with founder allelic effects and PoO effects. We demonstrate our method with CC-RIX data, but the method is general enough for other MPP-RIX populations. We place parameter-expanded Gaussian (PeG) priors on both the random founder allelic effects and PoO effects for variable selection.

The paper is organized as follows. In the Statistical Method section, we first introduce the CC-RIX experiment, then propose a random-effects model and describe a Bayesian variable selection procedure. In the Simulation Study section, we perform extensive simulation to examine the proposed method. Summary comments are given in the Discussion section.

### Statistical Method

The CC-RIX panel is the RI intercross of CC lines. For $L$ CC RI lines, a total of $L(L-1)$ reciprocal CC-RIX and $L(L-1)/2$ nonreciprocal CC-RIX can be generated [see Figure 1 of Zou et al. (2005), Yuan et al. (2011)]. Let $n$ be the total number of CC-RIX samples and $p$ be the total number of genetic markers. Further, let the phenotype of individual $i$ (the dependent variable) be $y_i$ ($i = 1, \cdots, n$).

#### Model

In order to account for the unbalanced relatedness between the CC-RIX mice and model the eight founder alleles and PoO effects, we extend the mixed additive random-effects model of Gong and Zou (2012) and Yuan et al. (2011) as follows:

$$y_i = \mu + \sum_{j=1}^{p} \gamma_{jQ}^T \beta_j + \sum_{j=1}^{p} \gamma_{jP}^T \xi_j + \sum_{l=1}^{L} a_l \alpha_l + e_i$$

where $\gamma_{jQ}$ and $\gamma_{jP}$ are binary $[0, 1]$ variables used to decide whether the $j$th QTL and PoO should be included in or excluded from the model and $\mu$ is the overall mean; $\beta_j = (\beta_{j1}, \cdots, \beta_{j8})^T$ and the $k$th element of $\beta_{j}$ equals 2, 1, or 0, depending on whether CC-RIX inherits 2, 1, or 0 copies of the $k$th founder allele ($k = 1, \cdots, 8$) at the $j$th candidate locus; $a_l$ are the proportions of CC-RIX, that are equal to CC-RII. Let $A$ be an $n \times L$ matrix whose $(i,j)$th element equals $a_{ij}$. Clearly, $\sum_{i=1}^{n} a_{ij} = 2$ for all $i = 1, 2, \cdots, n$, since each CC-RIX has two and only two parents. Therefore, $\beta_{jk}$ ($j = 1, \cdots, p$) represents the $k$th allelic effect of founder allele $j$, and $\alpha_l$ ($l = 1, \cdots, L$) represents the random polygenic effect of founder strain $l$. We let $\alpha_l$ follow $N(0, \sigma_{a_l}^2)$ ($l = 1, 2, \cdots, L$), and the random error $e$ follow $N(0, \sigma_e^2)$ ($i = 1, 2, \cdots, n$).

At a given locus, PoO effects can only be estimated from individuals with heterogeneous genotypes, since for those with homozygote genotypes, the parental contributions cannot be distinguished. Label the eight-founder alleles A, B, C, D, E, F, G, and H by numbers 1, 2,…, 8, and let the $k$th element of $z_{ij} = (z_{ij1}, \cdots, z_{ij8})^T$ be equal to 1 or -1 if the maternal or paternal allele of the $k$th subject at the $j$th locus equals the $k$th founder allele, and 0 otherwise. For those CC-RIX samples with a homozygous genotype at a given locus, such as AA or BB, we let $z_{ij} = 0_{8 \times 1}$. Therefore, $\xi_j = (\xi_{j1}, \cdots, \xi_{j8})^T$ represents the PoO effects of the $j$th locus.

Model (1) can be expressed by the following matrix form:

$$y = \mu + \sum_{j=1}^{p} \gamma_{jQ}^T x_j \beta_j + \sum_{j=1}^{p} \gamma_{jP}^T z_j \xi_j + A \alpha + e$$

where $y = (y_1, \cdots, y_n)^T$, $\mu = \mu 1_n$, $x_j = (x_{j1}, \cdots, x_{jn})^T$, $z_j = (z_{j1}, \cdots, z_{jn})^T$, $\alpha = (\alpha_{1}, \cdots, \alpha_{L})^T$, $\gamma_{jQ} = (\gamma_{jQ1}, \cdots, \gamma_{jQp})^T$, and $1_n = (1, \cdots, 1)^T$.

To enable selection of QTL and PoO effects, we reexpress the eight-dimensional vectors $\beta_j$ and $\xi_j$, as the expanded parameters $\tilde{\beta}_j$ and $\tilde{\xi}_j$, for $j = 1, \cdots, p$, respectively, such that $\tilde{\beta}_j = \gamma_{jQ} \beta_j$ and $\tilde{\xi}_j = \gamma_{jP} \xi_j$. This
parameter-expansion idea was first proposed by Kuo and Mallick (1998) and has an advantage in selecting important variables and shrinking coefficients.

In the above model (2), the observed data are $y_i$, marker genotypes $x_i = (x_{i1}, \ldots, x_{ip})$, PoO information matrix $z_i = (z_{i1}, \ldots, z_{ip})$, and parental CC-RI information matrix $A$. The unobserved variables include $u$, $b = (b_{j1}, \ldots, b_{j1})^T$, $f = (f_{j1}, \ldots, f_{j1})^T$, $g = (g_{j1}, \ldots, g_{j1})^T$, and the indicators $y_q = \{y_{q1}, \ldots, y_{qP}\}$ and $y_p = \{y_{p1}, \ldots, y_{pP}\}$. We assign the following conjugate Gaussian priors to the $j^{th}$ QTL and PoO effects as:

$$
\beta_j \sim \mathcal{N}_s\left(0, \sigma_{Qj}^2 I_8\right) \quad \text{and} \quad \xi_j \sim \mathcal{N}_s\left(0, \sigma_{Pj}^2 I_8\right) \quad \text{for} \quad j = 1, \ldots, p.
$$

For convenience, we name the above hierarchical priors for QTL and PoO coefficients as PeG priors in the sequel. We assign conjugate noninformative hyper-priors to the variance parameters

$$
P\left(\sigma_{Qj}^2\right) \propto \frac{1}{\sigma_{Qj}^2}, \quad P\left(\sigma_{Pj}^2\right) \propto \frac{1}{\sigma_{Pj}^2}, \quad P\left(\alpha_i^2\right) \propto (\alpha_i^2)^{\delta - 1} \quad \text{and} \quad P\left(\gamma_i\right) \propto \frac{1}{\gamma_i^2},
$$

where $0 < \delta \leq 1/2$ is used to ensure that the posterior distribution is proper (ter Braak et al. 2005). The above PeG priors are equivalent to "spike and slab" point mass mixture Gaussian priors (Mitchell and Beauchamp 1988; George and McCulloch 1993) for the expanded parameters

$$
\beta_j \sim \gamma_{Qj}\delta_0 + \left(1 - \gamma_{Qj}\right)\mathcal{N}_s\left(0, \sigma_{Qj}^2 I_8\right) \quad \text{for} \quad j = 1, \ldots, p.
$$

It is well known that these priors can achieve variable selection. However, compared with mixture priors, the PeG priors can employ a block Gibbs sampler to update the two blocks of parameters, one for $\beta^s$ and $\xi^s$ (corresponding to the selected predictors) and one for $\beta^u$ and $\xi^u$ (corresponding to the unselected predictors) in turn, and therefore can dramatically reduce computational time, especially for high-dimensional data with large $p$.

For indicator variables $y_q$ and $y_p$, we specify Bernoulli priors with inclusion probability $0 < \eta_{qj} < 1$ and $0 < \eta_{pj} < 1$ for $j = 1, \ldots, p$, respectively. To be more flexible, we further apply hierarchical uniform priors to $\eta_{qj}$ and $\eta_{pj}$:

$$
\eta_{qj} \sim U[0, 1], \quad \text{and} \quad \eta_{pj} \sim U[0, 1], \quad \text{for} \quad j = 1, \ldots, p.
$$

Lastly, we specify a flat prior on $\mu$ as $P(\mu) \propto 1$.

**Block Gibbs sampling algorithm for posterior computation**

The specific priors above result in known marginal conditional distributions for all variables. The blockwise Gibbs sampling algorithm that we employ can be summarized as follows:

First, we initiate $\sigma_{Q1}^2, \sigma_{Q2}^2, \sigma_{Q3}^2 = \{\sigma_{Q1}^2, \ldots, \sigma_{Qp}^2\}$, $\sigma_{P1}^2 = \{\sigma_{P1}^2, \ldots, \sigma_{Pp}^2\}$, $\eta_{Q1} = \{\eta_{Q1}, \ldots, \eta_{QP}\}$, and $\eta_{P1} = \{\eta_{P1}, \ldots, \eta_{PP}\}$ from uniform distribution $U(0, 1)$, then sample other parameters $\beta$, $\xi$, $\alpha$, and indicators $y_q$ and $y_p$ from their priors. We then perform the following block Gibbs sampling procedures. Superscripts $(t)$ and $(t+1)$ signify the Markov chain Monte Carlo (MCMC) iterations, and $t = 0$ refers to the initial iteration.
Step 1. Updating $\mu$: $\mu^{(t+1)}$ is drawn from the normal distribution.

$$N\left(\frac{1}{n} \left( y - \sum_{j=1}^{p} \gamma_{Q,j} x_j - \sum_{j=1}^{p} \gamma_{P,j} z_j - \mathbf{A} \mathbf{e}^{(t)} \right), \frac{\sigma^2_{e}(t)}{n} \right)$$

Step 2. Updating $\beta$: we divide the long vector $\beta^{(t)}$ into two blocks $\beta_1^{(t)} = \left( (\beta^{(t)}_1, \ldots, \beta^{(t)}_m) \right)$ and $\beta_2^{(t)} = \left( (\beta^{(t)}_{m+1}, \ldots, \beta^{(t)}_{m+n}) \right)$, corresponding to the selected $(y^{(t)}_j = 1)$ and unselected $(y^{(t)}_j = 0)$ predictors, respectively. Then we sample $\beta^{(t+1)}_m$ from their priors, i.e., the eight-dimensional multivariate normal distribution with zero mean and covariance matrix $\Sigma^{(t)}_\Omega$, and sample $\beta^{(t+1)}_n(h = 1, \ldots, m_1)$ from the eight-dimensional multivariate normal distribution with mean

$$1/\sigma^2_e^{(t)} \sum_{j=1}^{p} (x_j - \mu^{(t+1)})_n = -\sum_{j>h} \gamma_{Q,j} x_j - \sum_{j=1}^{p} \gamma_{P,j} z_j - \mathbf{A} \mathbf{e}^{(t)}$$

and covariance matrix $\Sigma^{(t)}_\Omega = \sigma^2_{e}^{(t)} (\sigma^2_{e}^{(t)} \mathbf{A}^2 + \sigma^2_{e}^{(t)} \mathbf{I})^{-1}$.

Step 3. Updating $\xi$: similar to Step 2, we divide $\xi^{(t)}$ into two parts, $\xi_1^{(t)} = \left( (\xi^{(t)}_1, \ldots, \xi^{(t)}_m) \right)$ and $\xi_2^{(t)} = \left( (\xi^{(t)}_{m+1}, \ldots, \xi^{(t)}_{m+n}) \right)$, corresponding to the selected $(y^{(t)}_j = 1)$ and unselected $(y^{(t)}_j = 0)$ predictors, respectively. Then $\xi^{(t+1)}_m(h = 1, \ldots, m_0)$ is sampled from their priors, i.e., the eight-dimensional multivariate normal distribution with zero mean and covariance matrix $\sigma^2_{e}^{(t)} \mathbf{I}$, and $\xi^{(t+1)}_n(h = 1, \ldots, m_1)$ is sampled from the eight-dimensional multivariate normal distribution with mean

$$1/\sigma^2_e^{(t)} \sum_{j=1}^{p} (x_j - \mu^{(t+1)})_n = -\sum_{j>h} \gamma_{Q,j} x_j - \sum_{j=1}^{p} \gamma_{P,j} z_j - \mathbf{A} \mathbf{e}^{(t)}$$

and covariance matrix $\Sigma^{(t)}_\Omega = \sigma^2_{e}^{(t)} (\sigma^2_{e}^{(t)} \mathbf{A}^2 + \sigma^2_{e}^{(t)} \mathbf{I})^{-1}$.

Step 4. Updating $\alpha$: $\alpha^{(t+1)}$ is drawn from the multivariate normal distribution with mean

$$1/\sigma^2_a^{(t)} \sum_{j=1}^{p} (x_j - \mu^{(t+1)})_n = -\sum_{j>h} \gamma_{Q,j} x_j - \sum_{j=1}^{p} \gamma_{P,j} z_j - \mathbf{A} \mathbf{e}^{(t)}$$

and covariance matrix $\Sigma^{(t)}_\Omega = \sigma^2_{e}^{(t)} (\sigma^2_{a}^{(t)} \mathbf{A}^2 + \sigma^2_{e}^{(t)} \mathbf{I})^{-1}$.

Step 5. Updating $\sigma_{Q,j}^2$: $\sigma_{Q,j}^{(t+1)}$ is sampled from the scale-inverted $\chi^2$ distribution, $\left| \beta_j^{(t+1)} \right|^2 / \chi^2_{\nu_{Q,j}}$.

Step 6. Updating $\sigma_{P,j}^2$: $\sigma_{P,j}^{(t+1)}$ is sampled from the scale-inverted $\chi^2$ distribution, $\left| \xi_j^{(t+1)} \right|^2 / \chi^2_{\nu_{P,j}}$.

Step 7. Updating $\sigma_a^2$: the random-effects variance $\sigma_a^{2(t+1)}$ is sampled from the scale-inverted $\chi^2$ distribution, $\left| \alpha^{(t+1)} \right|^2 / \chi^2_{\nu_{a}}$.

Step 8. Updating $\sigma_e^2$: the residual variance $\sigma_e^{2(t+1)}$ is sampled from the scale-inverted $\chi^2$ distribution,

$$\left| y - \mu^{(t+1)} \right|^2 / \chi^2_{\nu_{e}}$$

Step 9. Updating $\gamma_{Q,j}(1 \leq j \leq p)$: $\gamma_{Q,j}^{(t+1)}$ is sampled from the Bernoulli distribution with success probability

$$p_j^{(t+1)} = \eta_j^{(t)} / (\eta_j^{(t)} + (1 - \eta_j^{(t)}) \rho_j^{(t+1)})$$

where

$$
\rho_j^{(t+1)} = \exp \left[ -\frac{1}{2\sigma_e^{2(t)}} \left( 2\left( K_j^{(t)} \right) x_j z_j \right) \right],
$$

and

$$K_j^{(t+1)} = y - \mu^{(t+1)} - \sum_{j \neq h} \gamma_{Q,j} x_j - \sum_{j \neq h} \gamma_{P,j} z_j - \mathbf{A} \mathbf{e}^{(t)}.$$
genes only have QTL effects, and the last two only have PoO effects. However, the middle-selected marker has both QTL and PoO effects. The corresponding variances of QTL or PoO effects of the selected markers are all set to 1.

To better assess the performance of our method in situations where multiple nearby SNPs jointly affect the outcome, for cases 1–3, we now let the number of causal SNPs in each of the first two QTL be 2 instead of 1. Specifically, we randomly select two nearby SNPs for each of the first two simulated QTL and denote the alleles of the two SNPs $A$ and $a$. Based on the haplotype frequencies $f_{A\alpha}, f_{A\alpha}, f_{A\alpha},$ and $f_{A\alpha}$ (without loss of generality, we assume that $f_{A\alpha} \geq f_{A\alpha} \geq f_{A\alpha} \geq f_{A\alpha}$), we create three haplotype allelic groups, where haplotype $AA$ is group 1, $Aa$ is group 2, and the other two are group 3, and set the genetic effects of the three groups to 1, 2, and 3, respectively. The three new simulation setups are labeled as cases 1’ to 3’.

For each simulated data point, we generate a single long chain with 20,000 cycles, of which the first 10,000 cycles are discarded as burn-in, resulting in a total of 10,000 samples for post-MCMC analysis. All the analysis is done in MATLAB, and the MATLAB source code is submitted as supplemental material (File S2).

For comparison, we fit each simulated data point with the following three models.

The same model as (2), but without the PoO terms

$$y = \mu + \sum_{j=1}^{p} \gamma_{j}x_{ji} + A + e$$  

(3)

All the priors are set to be the same as their counterparts in model (2). For convenience, we subsequently call model (3) the “mixed model”.

Linear mixed-effects model (LMM) for single gene mapping

$$y = x_{ji}B_{j} + z_{ij}B_{j} + A + e, 1 \leq j \leq p.$$  

(4)

Here, random effects $A \sim N(0, \sigma^2_{A})$ and random errors $e \sim N(0, \sigma^2_{e})$, and the random effects $A$ are also assumed to be independent of the random errors $e$ as before. Moreover, $B_{j} = (\beta_{j1}, \ldots, \beta_{j8})'$ and $\xi_{j} = (\xi_{j1}, \ldots, \xi_{j8})'$ are the fixed QTL and PoO effects, respectively, of the $j$th tested locus. Given the constraints that $\sum_{k=1}^{8}z_{ijk} = 2$ and $\sum_{k=1}^{8}z_{ijk} = 0$ for $i = 1, \ldots, n; j = 1, \ldots, p$, we force the intercept $\mu$ in model (4) to be 0 to overcome the identifiability problem, and jointly test the effects of the $j$th locus as:

$$H_{0j}: \beta_{j1} = \beta_{j2} = \cdots = \beta_{j8}, \text{ and } \xi_{j1} = \xi_{j2} = \cdots = \xi_{j8}.$$  

In contrast to $H_{0j}$, some of the equations in $H_{0j}$ are not satisfied. After obtaining the maximum likelihood estimates of the parameters in model (4), we perform a test with the following log-odds ratio (LOD):

$$LOD_{j} = \log_{10} \frac{L_{ij}}{L_{0j}}$$

which is equivalent to the log-likelihood ratio test, where $L_{0j}$ and $L_{ij}$ are the likelihood of model (4) under null hypothesis $H_{0j}$ and alternative hypothesis $H_{ij}$, respectively. Since the above hypothesis test is performed $p$ times, it is necessary to find an appropriate significance threshold to control the multiple testing, which can be obtained, for example, by modified permutation procedures (Zou et al. 2005).

### Table 1 Simulation settings and AUC values

| Case | IS | $p$ | AUC$_{p}^{b}$ | AUC$_{M}^{c}$ | AUC$_{p}^{d}$ | AUC$_{M}^{e}$ |
|------|----|----|--------------|--------------|--------------|--------------|
| 1    | 10 | 133| 0.9642       | 0.8499       | 0.5155       | 0.9436       |
| 2    | 5  | 266| 0.9722       | 0.9022       | 0.5029       | 0.9463       |
| 3    | 133| 0.9915| 0.9489       | 0.5000       | 0.9525       |
| 1+  | 10 | 133| 0.9499       | 0.6885       | 0.5085       | 0.9264       |
| 2+  | 5  | 266| 0.9624       | 0.9056       | 0.4998       | 0.9238       |
| 3+  | 133| 0.9908| 0.9393       | 0.5040       | 0.9482       |

$\text{Interval space (IS; cm) between nearby markers.}$  
$^{b} \text{AUC values of our proposed model (2).}$  
$^{c} \text{AUC values of the mixed model (3).}$  
$^{d} \text{AUC values of Yuan’s model (5).}$  
$^{e} \text{AUC values of the LMM model (4).}$

However, in our comparison, we evaluate the receiver operating characteristic (ROC) curve, which only requires the use of LOD scores.

**Yuan’s Bayesian method**

$$y_{i} = \mu + \sum_{j=1}^{p} z_{ij} \beta_{j} + \sum_{l=1}^{p} a_{jl} \alpha_{l} + e_{i}$$  

(5)

Here, $\beta_{j}$ is the effect of the $j$th putative QTL with $z_{ij} = \sqrt{2m_{ij}}$, where $m_{ij} = 1, \ldots, n; j = 1, \ldots, p$ equals $-1, 0, 1$, or $1$ if the putative QTL genotype is $aa$, $Aa$, or $AA$, respectively. The other parameters are set the same as those in the mixed model. The prior of $\beta_{j}$ is set to $N(0, \sigma_{\beta j}^{2})$, $1 \leq j \leq p$ in Yuan et al. (2011). To compare the methods, ROC curves (Fawcett 2006), where true positive rates (also known as sensitivity) are plotted against false positive rates (also known as 1-specificity) evaluated at various threshold cut-offs. To estimate the sensitivity (the proportion of positives that are correctly identified as such) and specificity (the proportion of negatives that are correctly identified as such), we define false and true positive findings as follows. If a detected locus falls no more than 10 cm apart from any simulated genes, we call it a true positive finding, otherwise a false positive finding. Then, combining the outputs of each model for the 100 data sets, we can calculate the corresponding sensitivity and specificity. Specifically, for the $j$th marker, we record the LOD scores, $LOD_{j}$, for the LMM model; the maximum posterior frequency between $\gamma_{ij}$ and $\gamma_{pj}$ for our proposed method; the posterior frequency $\gamma_{ij}$ for the mixed model; and the posterior mean of $\sigma_{\beta j}^{2}$ for Yuan’s model. The corresponding area under the ROC curve (AUC) is also calculated for each of the four models, and the results are presented in Table 1. Generally speaking, a model with a higher AUC value indicates on average a better performance compared with those with lower AUC scores (Fawcett 2006).

From Table 1, it is clear that our proposed method outperforms the other three methods for all cases, regardless of whether there is a single causal SNP or multiple causal SNPs in each QTL. Yuan’s method fails in all the simulated cases, which is expected as the method only models biallelic SNP effects instead of the founder allelic effects that we simulate. The LMM model outperforms the mixed model, in particular when the marker density is sparse. This phenomenon is further confirmed by Figure 1, where the ROC curves of the proposed method are always higher than the ROC curves of the other three methods; the ROC curves of Yuan’s model fall along the 45-degree line, indicating its low power for mapping genes with founder allelic effects. The ROC curves of the LMM model fall between those of our proposed model and the mixed model in cases 1 and 2, and cases 1’ and 2’, but cross
with the ROC curves of the mixed model in cases 3 and 3'. Figure 1 and Table 1 show that the mapping precisions of all the methods increase as the number of markers increases, as expected.

Figure 2 presents the Manhattan plots of the four models based on the average estimate of a total of 100 replications across the whole genome for case 1. Clearly, our proposed method is the winner in mapping genes with founder allelic effects (marked by an asterisk) and PoO effects (marked by a circle). The Manhattan plots for cases 2 and 3, and cases 1', 2', and 3', are included in Figures S1–S3 in File S1, respectively, and show similar patterns.

**DISCUSSION**

The CC (Threadgill and Churchill 2012), a panel of newly emerged multiparent RI mouse strains, was developed, similar to other MPPs, to provide greater genetic diversity than the traditional RI populations and thereby to improve our power of understanding of complex traits. CC-RIX, F1 crosses that are generated from parental CC RI lines (Zou et al. 2005) can serve as excellent mouse models for mapping genes with both traditional genetic effects and epigenetic effects, such as imprinting. This study extends the model of Gong and Zou (2012) and Yuan et al. (2011) and, to the best of our knowledge, it is the first use of Bayesian variable selection methods to jointly map multiple QTL with both founder allelic effects and PoO effects for MPP-RIX data, in particular, CC-RIX data.

In this article, it is assumed that the RI lines are equally distanced from each other in terms of genetic distance which is a sensible assumption given the funnel design used for generating CC RI lines. However, we do observe that some RI lines share more or fewer founder alleles than expected. Our limited simulation shows that such genetic unbalance has a negligible effect on mapping genes, but this issue deserves further investigation. Alternatively, as genotypes of the parental RI lines are available, we could modify the design matrices $\mathbf{x}_j$, $\mathbf{z}_j$, and $\mathbf{A}$ in model (1) and replace them with local and genome-wide similarity matrices as done in sequence kernel association tests (SKAT) (Ionita-Laza et al. 2013) and genome-wide complex trait analysis (GCTA) (Yang et al. 2011). In addition, in our model we assume additive founder allelic effects. This assumption can be easily relaxed to allow genes with both additive and dominance effects (Woods et al. 2012; Zhang et al. 2014). Because dominant genetic effects are orthogonal to PoO effects, missing the dominant genetic effects does not cause any bias in the PoO effects estimate.

One of the drawbacks of Bayesian methods is their computational cost, especially for data with large numbers of samples and markers. Our model employs a block Gibbs sampler, which dramatically reduces computational time. However, further improvements may be possible. For example, instead of modeling each marker separately, we could jointly model multiple nearby markers and reduce the magnitude of $p$. Similar ideas have been proposed for genome-wide association study data (Wu et al. 2010; Lu et al. 2015). This approach is also biologically meaningful, since for some complex traits multiple causal SNPs may be located in a single region, and our simulations (cases 1'–3') have shown that the proposed method is powerful in mapping regions with multiple causal SNPs. However, grouping nearby markers or SNPs may offer further help in improving mapping power, which deserves more thorough investigation.

**ACKNOWLEDGMENTS**

We thank two anonymous referees for their valuable comments on our manuscript. This research was partially supported by NIH grant R01 GM105785, and by the scholarship from China Scholarship Council (CSC) under the Grant CSC No.201506270042.

**LITERATURE CITED**

Barlow, D. P., 2011 Genomic imprinting: a mammalian epigenetic discovery model. Annu. Rev. Genet. 45: 359–403.

Basten, C. I., B. S. Weir, and Z.-B. Zeng, 1999 QTL Cartographer: A Reference Manual and Tutorial for QTL Mapping. Department of Statistics, North Carolina State University, Raleigh, NC.

Broman, K. W., 2001 Review of statistical methods for QTL mapping in experimental crosses. Lab Anim. (NY) 30: 44–52.
