

# Varying Coefficient Models for Mapping Quantitative Trait Loci Using Recombinant Inbred Intercrosses

Yi Gong and Fei Zou<sup>1</sup>

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

**ABSTRACT** There has been a great deal of interest in the development of methodologies to map quantitative trait loci (QTL) using experimental crosses in the last 2 decades. Experimental crosses in animal and plant sciences provide important data sources for mapping QTL through linkage analysis. The Collaborative Cross (CC) is a renewable mouse resource that is generated from eight genetically diverse founder strains to mimic the genetic diversity in humans. The recombinant inbred intercrosses (RIX) generated from CC recombinant inbred (RI) lines share similar genetic structures of  $F_2$  individuals but with up to eight alleles segregating at any one locus. In contrast to  $F_2$  mice, genotypes of RIX can be inferred from the genotypes of their RI parents and can be produced repeatedly. Also, RIX mice typically do not share the same degree of relatedness. This unbalanced genetic relatedness requires careful statistical modeling to avoid false-positive findings. Many quantitative traits are inherently complex with genetic effects varying with other covariates, such as age. For such complex traits, if phenotype data can be collected over a wide range of ages across study subjects, their dynamic genetic patterns can be investigated. Parametric functions, such as sigmoidal or logistic functions, have been used for such purpose. In this article, we propose a flexible nonparametric time-varying coefficient QTL mapping method for RIX data. Our method allows the QTL effects to evolve with time and naturally extends classical parametric QTL mapping methods. We model the varying genetic effects nonparametrically with the B-spline bases. Our model investigates gene-by-time interactions for RIX data in a very flexible nonparametric fashion. Simulation results indicate that the varying coefficient QTL mapping has higher power and mapping precision compared to parametric models when the assumption of constant genetic effects fails. We also apply a modified permutation procedure to control overall significance level.

**D**URING the past 2 decades, there has been considerable development in statistical methodologies for mapping quantitative trait loci (QTL), since Lander and Botstein (1989) implemented a maximum-likelihood approach to the interval-mapping technique (Goldgar 1990; Amos 1994; Jansen and Stam 1994; Zeng 1994; Almasy and Blangero 1998; Kao *et al.* 1999; Zou *et al.* 2001; Xu *et al.* 2005). In addition to the interval-mapping approach, many other statistical approaches have been used in QTL mapping, such as regression analyses (Haley and Knott 1992)

and Bayesian approaches (Satagopan *et al.* 1996; Sillanpaa and Arjas 1998; Yi and Xu 2000; Yi 2004; Hoeschele 2007).

While these methods have been instrumental for QTL identification, they are not able to capture the temporal pattern of QTL effect. Many quantitative traits, such as body size, are inherently too complex to be described by a single value, because their phenotypes, for example, change with age. Instead of being measured at one fixed time point, each subject's phenotype may be measured at different time points across samples, which allows us to study genetic effects that vary with the change of time. For example, genetic correlations among age-specific weights in a laboratory population of rats were shown to involve variable gene action at different ages (Cheverud *et al.* 1983). Vaughn *et al.* (1999) located QTL responsible for age-specific weights in mice, and they found that some QTL affect the early growth patterns and some affect the late growth patterns. To study

Copyright © 2012 by the Genetics Society of America  
doi: 10.1534/genetics.111.132522

Manuscript received July 10, 2011; accepted for publication November 17, 2011

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.111.132522/-/DC1>

<sup>1</sup>Corresponding author: Department of Biostatistics, University of North Carolina, 4115D McGavran-Greenberg Hall, CB# 7420, Chapel Hill, NC 27599. E-mail: fzou@bios.unc.edu

genetic determination of such functional traits, Wu and colleagues (Ma *et al.* 2002; Wu *et al.* 2002, 2004; Lin and Wu 2006) developed the functional mapping approach. They used growth curve data as an example of functional traits, and the genetic effect was modeled by a parametric function such as sigmoidal or logistic function (Ma *et al.* 2002). While the parametric nature of functional mapping offers tremendous biological and statistical advantages, a reliance on the availability of mathematical functions limits its applicability (Yang *et al.* 2009).

Varying coefficient models are very useful statistical tools for exploring dynamic effects. The varying coefficient models were introduced by Cleveland *et al.* (1991), and discussed by Hastie and Tibshirani (1993) in more detail, to extend the applications of local regression techniques from one-dimensional to multidimensional settings. In varying coefficient models, there are many ways to model the function of the varying effect, such as polynomials, Fourier series, piecewise polynomials, and more general nonparametric functions (Hastie and Tibshirani 1993). For nonparametric varying-coefficient models, various basis systems can be used, and the most common choice is the B-spline basis (He and Shi 1998; Pittman 2002; Huang *et al.* 2004; Wang *et al.* 2007, 2008). One advantage of B-splines over some other nonparametric approaches, like smoothing splines, is that the smoother matrix is independent of the responses. Yang *et al.* (2009) proposed a nonparametric functional mapping framework for genetic mapping of QTL controlling for a dynamic trait, implemented with B-splines.

Although important, QTL mapping in humans is difficult, time consuming, expensive, and hampered by ethical problems and uncontrollable environments. These obstacles are nearly all overcome in laboratory mice. Furthermore, most human genes have functional mouse counterparts and both genomes are organized similarly. Hence, the laboratory mouse has become an important model organism in mapping QTL related to human disease. Recombinant inbred (RI) lines have contributed greatly to genetic dissection of simple and complex traits. A major advantage of RI panels over other commonly used mapping approaches is their ability to support genetic mapping and correlations among many traits, even under different environmental conditions (Plomin *et al.* 1991). However, the traditional inbred mice have a limited amount of variation (Darvasi 1998). Typical mouse RI panels have only 15–35 strains from a single pair of parental inbred lines (Zou *et al.* 2005; Tsaih *et al.* 2005). This is a particularly acute problem when one wants to examine numerous gene–environment interactions or study disease progression at many stages and ages (Zou *et al.* 2005). Mouse RI panels generally have low power and precision compared to other resources because of their small size.

The Collaborative Cross (CC) project (Threadgill *et al.* 2002; Churchill *et al.* 2004; Collaborative Cross Consortium 2012) has been carried out to create a large panel

of new RI mouse strains. The CC RI lines are generated from an eight-way cross using eight genetically diverse founder strains, which makes the CC RI lines closer to natural populations than regular RI lines with more genetic variation. A novel derivative of RI lines, called recombinant inbred intercrosses (RIX), has been designed that permits repeated interrogations of a fixed genotype to reduce non-genetic variance while increasing the power of the original RI panel (Threadgill *et al.* 2002; Collaborative Cross Consortium 2012). The RIX panel is created as F<sub>1</sub> hybrids of RI lines. Linkage analyses can be performed, using these resources, to fine map genetic loci that are responsible for most inheritable complex traits. Since all RI mice are homozygous at each locus, the genotypes of the derivative RIX mice will be known in advance by imputing from the genotypes of the parental RI lines. RIX mice with identical genotypes can be regenerated whenever needed. Compared to RI, the RIX design has several advantages that include twice the number of recombination sites in a single individual since each is derived from two parental RI, dominance effects can be estimated, there is a large expansion of different RIX genomes over the parental RI, and, because of the buffering capacity of their heterogeneous genome structure, RIX genomes should provide more reliable trait means than the parental RIs. The RIX approach also has advantages over classical crosses like the F<sub>2</sub> design since each RIX has a higher recombination density than a single F<sub>2</sub> individual when performing interval mapping (Broman 2005; Broman 2012), RIX are especially useful for long-term collaborative research because their genotypes are renewable making the phenotypic data cumulative within the research community, and since RIX genomes are easily replicated, experiments with different environmental variables or temporal relationships can be performed on the same genotypes. At the individual level, although the genome of each RIX mouse has similar genetic structures of F<sub>2</sub> individuals, statistical analyses for F<sub>2</sub> data cannot be directly applied to RIX data. This is because some RIX individuals share a common parental RI line, making them genetically more related to each other than those that do not share any parental lines. Several QTL mapping methods have been introduced (Zou *et al.* 2005; Tsaih *et al.* 2005; Yuan *et al.* 2011) for dealing with the special genetic structure of RIX data. However, none has considered the situation in which the QTL effect varies with other covariates. In this study, we propose a new method to properly model both the (time) varying genetic effects and the genetic complexity of RIX data. The proposed model investigates gene-by-time interactions in a flexible nonparametric fashion for RIX data.

## Methods

For an RI panel with  $L$  lines, there are at most  $L(L - 1)/2$  nonreciprocal RIXs that can be generated, which is a huge number when  $L$  is large. A useful sampling and mating

scheme is the loop design as described by Zou *et al.* (2005) and Yuan *et al.* (2011). With the loop design,  $L$  RI lines were randomly ordered to form a circle. Then each RI line is mated with the next  $J$  RI lines after it, resulting in total of  $LJ$  samples. That is, we mate  $RI_1$  with  $RI_2, RI_3, \dots,$  and  $RI_{J+1}; \dots; RI_i$  with  $RI_{m(i+1,L)}, RI_{m(i+2,L)}, \dots,$  and  $RI_{m(i+j,L)}; \dots;$  and  $RI_L$  with  $RI_1, RI_2, \dots,$  and  $RI_J$ , where

$$m(x, L) = \begin{cases} x, & \text{if } x \leq L; \\ x - L, & \text{if } x > L. \end{cases}$$

Not only in the loop design, but in many RIX populations, pairs of RIX sharing one parent are more closely related than those RIX that do not share a parent. For example, RIX produced by crossing  $RI_1$  and  $RI_2$  ( $RIX_{12}$ ) is expected to be more similar to RIX produced by crossing  $RI_1$  and  $RI_3$  ( $RIX_{13}$ ) than to RIX from crosses between  $RI_3$  by  $RI_4$  ( $RIX_{34}$ ) since ( $RIX_{12}$ ) and ( $RIX_{13}$ ) share a parental RI ( $RI_1$ ) while ( $RIX_{12}$ ) and ( $RIX_{34}$ ) do not share any parental RI lines.

To study the RIX data, we fit a mixed-effect model by applying a random effect to model the polygenic effect. For simplicity, a model with only additive effect is considered. Also, we assume that all putative QTL are located on markers. For individual  $i$ , define the observed data as  $\{y_i, t_i, x_{i1}, \dots, x_{im}\}$ , where  $y_i$  is the phenotype,  $t_i$  is the measure of the covariate and is nonconstant across subjects,  $M$  is the total number of markers, and  $x_{im}$  is the genotype at the  $m$ th marker, coded as  $-1, 0,$  or  $1$  for genotypes  $aa, Aa,$  and  $AA$ , respectively. We consider one putative QTL at a time and therefore suppress the subindex  $m$  in the sequel. The model can be expressed as

$$y_i = \mu(t_i) + x_i\beta(t_i) + \sum_{l=1}^L a_{il}\alpha_l + \epsilon_i,$$

where  $\mu(t)$  and  $\beta(t)$  are the overall population mean and QTL effect that vary with time  $t$ , respectively. The random polygenic effect  $\alpha_l$  follows  $N(0, \sigma_a^2)$  for  $l = 1, 2, \dots, L$ ; the random error  $\epsilon_i$  follows  $N(0, \sigma_0^2)$ ; and

$$a_{il} = \begin{cases} 1, & \text{if one of } i\text{th individual's parents is } RI_l; \\ 0, & \text{otherwise.} \end{cases}$$

This model can be applied to any RIX population in addition to the loop design as described above. The hypotheses for whether there exists any major QTL at a given locus are  $H_0: \beta(t) = 0$  vs.  $H_a: \beta(t) \neq 0$ .

We incorporate B-spline bases to model the varying coefficient functions  $\beta(t)$  and  $\mu(t)$ . The smoothness of the function modeled by B-splines is controlled by the parameter  $K = n_j + d + 1$ , where  $n_j$  is the number of interior knots and  $d$  is the degree of spline. The interior knots of the splines can be either equally spaced or placed on the sample quantiles of the data, so that there are about the same number of observations between any two adjacent knots. We use

equally spaced knots for all numerical examples for this study, and hence  $B_k(t)$  is determined for any given  $t$ .

The mixed-effects model becomes

$$y_i = \sum_{k=1}^K \gamma_{0k} B_k(t_i) + \sum_{k=1}^K \gamma_k B_k(t_i) x_i + \sum_{l=1}^L a_{il} \alpha_l + \epsilon_i,$$

where  $B_k(t_i)$ 's are basis functions of B-splines of order  $K$ , and  $\gamma_{0k}$ 's and  $\gamma_k$ 's are coefficients for B-spline basis. Here  $\mu(t)$  is approximated by  $\sum_{k=1}^K \gamma_{0k} B_k(t)$  and  $\beta(t)$  is approximated by  $\sum_{k=1}^K \gamma_k B_k(t)$ . To test for genetic effect of QTL, the hypotheses  $H_0: \beta(t) = 0$  vs.  $H_a: \beta(t) \neq 0$  are equivalent to  $H_0: \gamma_1 = \dots = \gamma_K = 0$  vs.  $H_a: \text{not all the } \gamma_k\text{'s are 0}$ .

We can rewrite the model above in the matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{A}\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

where  $\mathbf{y} = (y_1, \dots, y_n)^T$ ;  $\boldsymbol{\gamma} = (\gamma_{01}, \dots, \gamma_{0K}, \gamma_1, \dots, \gamma_K)^T$ ;  $\mathbf{X}$  is the corresponding  $n \times 2K$  design matrix for the time-varying fixed effect;  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_L)^T$ ;  $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ ; and  $\mathbf{A}$  is an  $n \times L$  design matrix for the random polygenic effect. The design matrix  $\mathbf{X}$  can be expressed as

$$\mathbf{X} = \begin{pmatrix} B_1(t_1) & \dots & B_K(t_1) & x_1 B_1(t_1) & \dots & x_1 B_K(t_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_1(t_n) & \dots & B_K(t_n) & x_n B_1(t_n) & \dots & x_n B_K(t_n) \end{pmatrix}.$$

We therefore observe  $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\gamma}, \boldsymbol{\Sigma})$  with  $\boldsymbol{\Sigma} = \sigma_a^2 \mathbf{A}\mathbf{A}^T + \sigma_0^2 \mathbf{I}$ , which can be reparameterized as  $\boldsymbol{\Sigma} = \sigma_0^2(\boldsymbol{\theta}\mathbf{D} + \mathbf{I}) = \sigma_0^2 \mathbf{V}$ , with  $\boldsymbol{\theta} = \sigma_a^2/\sigma_0^2$ ,  $\mathbf{D} = \mathbf{A}\mathbf{A}^T$ , and  $\mathbf{V} = \boldsymbol{\theta}\mathbf{D} + \mathbf{I}$ . Regardless of the form of the covariance matrix  $\boldsymbol{\Sigma}$ , the generalized least squares (GLS) is an appropriate estimate for parameter  $\boldsymbol{\gamma}$  as

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}.$$

The profile log-likelihood functions with only unknown parameters in  $\boldsymbol{\Sigma}$ , based on the maximum likelihood (ML) and restricted/residual maximum likelihood (REML), can be written as

$$-2l(\sigma_0^2, \boldsymbol{\theta} | \mathbf{y}) = \log |\mathbf{V}| + n \log(\sigma_0^2) + \sigma_0^{-2} \mathbf{r}^T \mathbf{V}^{-1} \mathbf{r} + n \log(2\pi),$$

for ML and

$$-2l_r(\sigma_0^2, \boldsymbol{\theta} | \mathbf{y}) = \log |\mathbf{V}| + (n-p) \log(\sigma_0^2) + \log |\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}| + \sigma_0^{-2} \mathbf{r}^T \mathbf{V}^{-1} \mathbf{r} + (n-p) \log(2\pi)$$

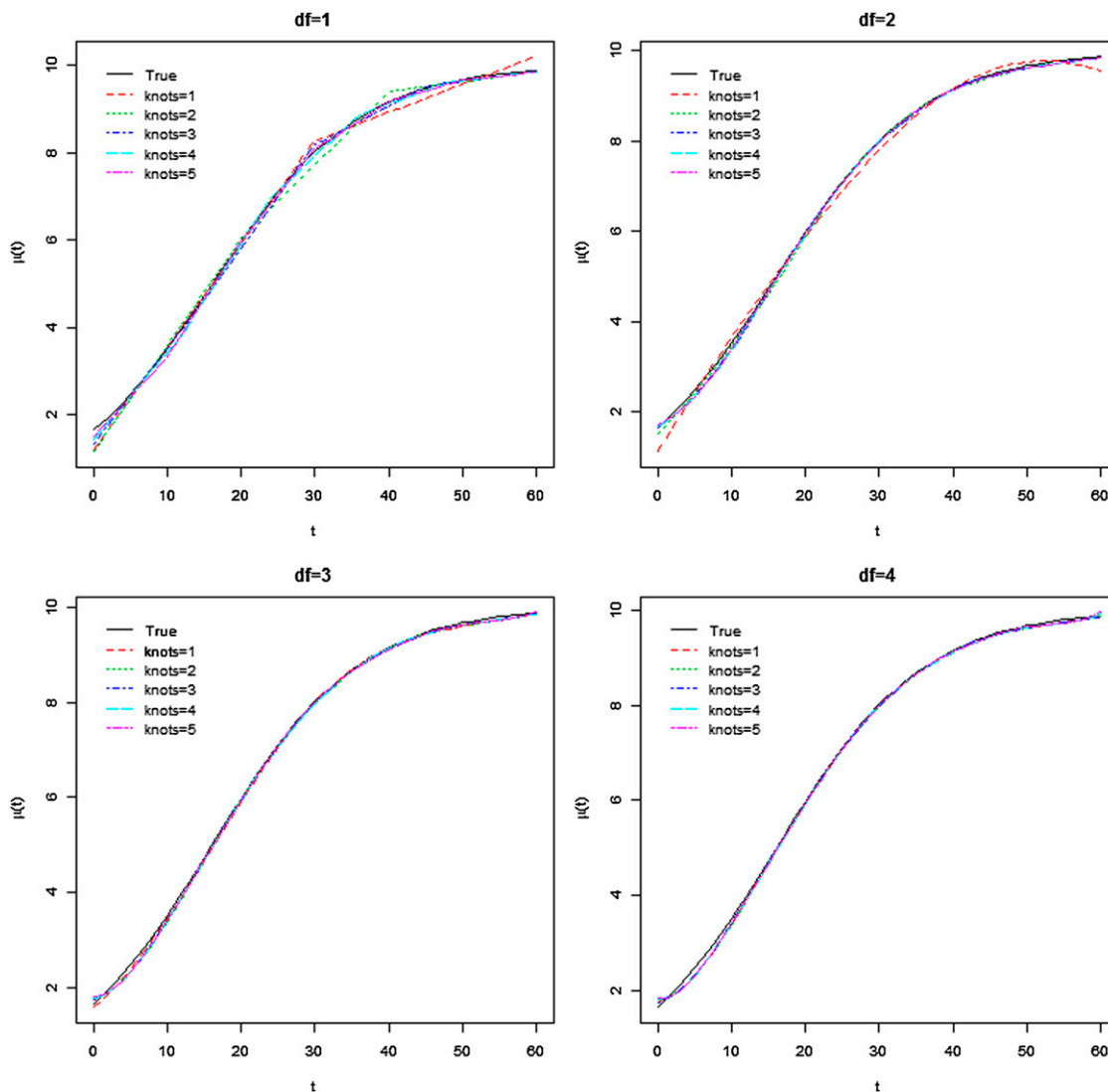
for REML, where  $p$  is the rank of  $\mathbf{X}$  and  $\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}$  is a function of  $\boldsymbol{\theta}$ .

To simplify the computation, we further solve for the ML or REML estimate of  $\sigma_0^2$  as a function of  $\boldsymbol{\theta}$ ,

$$\hat{\sigma}_0^2 = \frac{1}{n} \mathbf{r}^T \mathbf{V}^{-1} \mathbf{r},$$

for ML and

$$\hat{\sigma}_0^2 = \frac{1}{n-p} \mathbf{r}^T \mathbf{V}^{-1} \mathbf{r}$$



**Figure 1** The varying coefficient  $\mu(t) = 10/(1+5e^{-0.1t})$  (solid curve). Dotted curves are the mean estimates of  $\mu(t)$  for different combinations of  $n_j$  (number of internal knots) and  $d$  (degree of spline).

for REML. Substitute the expressions above, we obtain the final profile log-likelihoods for  $\theta$  as

$$-2l(\theta | y) = \log |V| + n \log(r^T V^{-1} r) + n \log(2\pi),$$

and

$$-2l_R(\theta | y) = \log |V| + \log |X^T V^{-1} X| + (n-p) \log(r^T V^{-1} r) + (n-p) \log(2\pi).$$

Note that the profile log-likelihood above involves only the nuisance parameter  $\theta$ . Hence its MLE can be easily computed by the Newton–Raphson algorithm. Once  $\theta$  is estimated,  $\hat{\gamma}$  and  $\hat{\sigma}_0^2$  can be subsequently estimated by

$$\hat{\gamma} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y,$$

and

$$\hat{\sigma}_0^2 = \frac{1}{n} (y - X\hat{\gamma})^T \hat{V}^{-1} (y - X\hat{\gamma}),$$

for ML and

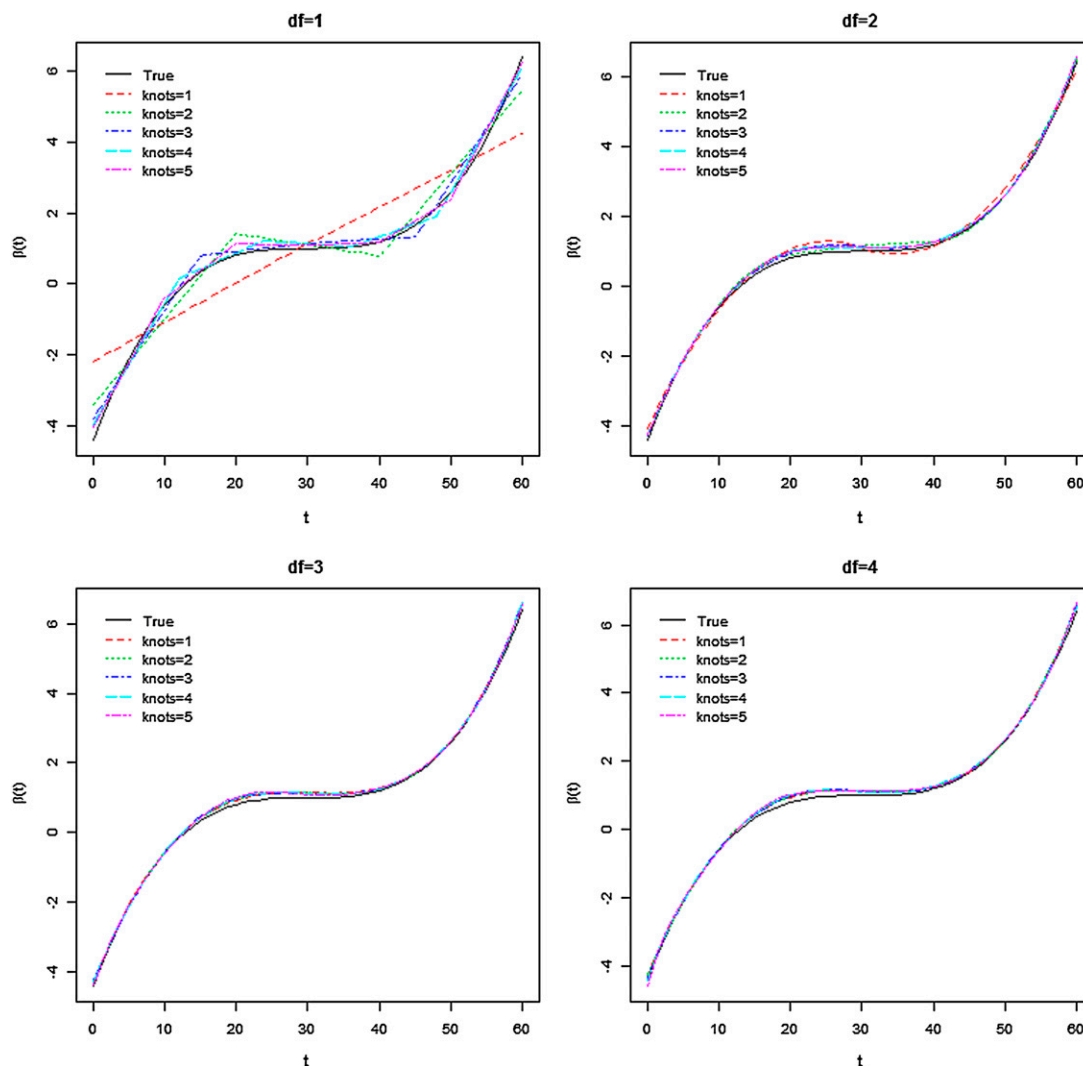
$$\hat{\sigma}_0^2 = \frac{1}{n-p} (y - X\hat{\gamma})^T \hat{V}^{-1} (y - X\hat{\gamma}),$$

for REML. We use REML in the following simulation studies, since it has some advantages over ML, such as taking into account the degrees of freedom for fixed effects (McCulloch and Searle 2001).

Once the parameters are estimated, likelihood-ratio (LR) tests can be performed to evaluate the evidence of QTL effect, and LOD scores can be calculated at the locations of all genetic markers

$$\text{LOD} = \log_{10} L_R(\hat{\gamma}, \hat{\theta}, \hat{\sigma}_0^2) - \log_{10} L_R(0, \tilde{\theta}, \tilde{\sigma}_0^2),$$

where  $(\tilde{\theta}, \tilde{\sigma}_0^2)$  is the MLE under  $H_0: \gamma_1 = \dots = \gamma_K = 0$ .



**Figure 2** The varying coefficient  $\beta(t) = 1 + (30-t)^3/5000$  (solid curve). Dotted curves are the mean estimates of  $\beta(t)$  for different combinations of  $n_j$  (number of internal knots) and  $d$  (degree of spline).

Since the hypothesis testing is performed on a number of markers, it is necessary to adjust the significance level for multiple testing. The threshold, in practice, is usually obtained by permutation procedures (Churchill and Doerge 1994). However, it is quite complicated to obtain appropriate threshold values for RIX data, because direct permutation will not only destroy the relationship between QTL and the trait, but also destroy the relationship between polygenes and the trait, which will result in incorrect thresholds (Anderson and Ter Braak 2003; Zou *et al.* 2005; Churchill and Doerge 2008). To overcome this difficulty, Zou *et al.* (2005) extended the permutation method of Churchill and Doerge (1994) to a novel permutation method for the RIX data. The modified permutation method starts with permuting parental RI strain numbers  $1, 2, \dots, L$  into  $\phi(1), \phi(2), \dots, \phi(L)$ . Then the permuted marker genotypes of  $\text{RIX}_{ij}$  will be the corresponding marker genotypes of  $\text{RIX}_{\min(\phi(i), \phi(j))\max(\phi(i), \phi(j))}$  in the original data. The permuted samples are analyzed with the same

model as the original data to generate an empirical distribution of maximum LOD scores, where the threshold value can be obtained.

## Results

All analysis and simulation code used below are included in supporting information, [File S1](#). In simulation studies, we set the number of parental RI lines  $L = 100$ , and applied the loop design with  $J = 3$  to generate a total of 300 RIX samples (Zou *et al.* 2005). A single chromosome with 101 evenly spaced markers was simulated with either a 2-cM interval or 5-cM interval between nearby markers (resulting in a total length of 2 M or 5 M, respectively). The QTL is located at the 41th marker, which is at either 80 or 200 cM, for the two marker densities, respectively. The marker genotypes were simulated using R/qtl (Broman *et al.* 2003). We set  $\mu(t)$ , the mean temporal growth function for QTL genotype Aa to  $10/(1+5e^{-0.1t})$ , which is a logistic growth curve (Ma *et al.*

**Table 1** Counts based on the smallest SQD or AIC

	$n_k =$	SQD					AIC				
		1	2	3	4	5	1	2	3	4	5
Case 1	$d = 1$	0	73	39	10	0	40	126	36	21	3
	$d = 2$	326	19	4	1	1	182	14	9	4	3
	$d = 3$	17	2	0	0	0	28	4	2	2	1
	$d = 4$	6	1	0	1	0	7	8	5	2	3
Case 2	$d = 1$	47	73	32	7	0	122	85	28	18	6
	$d = 2$	257	34	5	1	1	155	17	12	4	2
	$d = 3$	36	2	0	0	0	19	2	2	2	3
	$d = 4$	3	1	0	1	0	7	7	4	2	3
Case 3	$d = 1$	68	259	13	18	1	136	180	12	24	6
	$d = 2$	114	4	6	1	1	78	6	10	2	4
	$d = 3$	9	1	0	0	0	13	2	1	2	3
	$d = 4$	4	0	0	1	0	8	4	5	1	3

2002; Yang *et al.* 2009). We randomly generated  $t_i$  from (0, 60) for each subject.

We considered the three different functions for  $\beta(t)$ :

Case 1:

$$\beta(t) = 1 + 3 \sin\left(\frac{\pi t}{30}\right);$$

Case 2:

$$\beta(t) = 1 + \frac{(30-t)^3}{5000};$$

Case 3:

$$\beta(t) = \frac{3}{2} \left( \arctan\left(\frac{t-30}{4}\right) + \frac{\pi}{2} \right).$$

Cases 1 and 2 are nonlinearly increasing functional effects, used in simulation studies by Wang *et al.* (2008). Case 3 mimics the situation in which the genetic effect is hardly perceptible until after certain age, such as some breast cancer-susceptibility genes (Foulkes *et al.* 2004). To test the performance of the model under various signal/noise ratios, two different sets of variances for random effect and random error were considered for each case:  $\sigma_a^2 = 10$ ,  $\sigma_0^2 = 20$  and

$\sigma_a^2 = 30$ ,  $\sigma_0^2 = 30$ . In all cases, the average heritability is between 0.02 and 0.18.

To choose a good combination of the interior knot number  $n_j$  and the degree of spline  $d$  for the genetic effect, 500 runs of simulation were performed. In those simulations, we set  $\sigma_a^2 = 30$ ,  $\sigma_0^2 = 30$ , and the interval length to 5 cM. Figure 1 and Figure 2 plotted the mean  $\hat{\mu}(t)$  and mean  $\hat{\beta}(t)$  for case 2 with different combinations of  $n_j$  and  $d$ . The figures showed that relatively small  $n_j$  and  $d$  in general fit the curves well, and the same is true for the other two cases. We calculated the squared differences (SQD) between  $\hat{\mu}(t)$  and  $\mu(t)$ , and between  $\hat{\beta}(t)$  and  $\beta(t)$  as  $SQD = \int_0^{60} \{(\hat{\mu}(t) - \mu(t))^2 + (\hat{\beta}(t) - \beta(t))^2\} dt$  for each of the  $n_j$  and  $d$  combination. We recorded the number of combinations of  $n_j$  and  $d$  with the smallest SQD in Table 1, left. The results suggest that the combination of  $n_j = 1$  and  $d = 2$  is the best for cases 1 and 2, while for case 3, it is  $n_j = 2$  and  $d = 1$ .

In practice, the true  $\beta(t)$  is unknown, so the choice of  $n_j$  and  $d$  needs to be estimated. We propose the following approach to choose  $n_j$  and  $d$  using the AIC (Akaike 1970, 1974) as the selection criterion. First, we set  $n_j = 1$  and  $d = 1$  and identify the marker with the highest LOD score. Then at the selected marker, we calculate the AIC values for a set of  $n_j$  and  $d$ , and choose the one with the smallest AIC. In the simulation study, we computed the AIC values for the 500

**Table 2** Mean estimated locations of QTL (in centimorgans) and standard errors

Variance	Distance	Method	Case 1	Case 2	Case 3	
$\sigma_a^2 = 10$ , $\sigma_0^2 = 20$	2 cM	B-spline	79.59 (0.71)	81.60 (1.17)	80.80 (0.74)	
		Linear	79.46 (1.55)	83.62 (1.19)	81.01 (0.74)	
	QTL at 80 cM	Quadratic	80.23 (1.41)	83.13 (1.55)	82.66 (1.10)	
		5 cM	B-spline	198.15 (2.11)	203.28 (3.26)	201.75 (1.43)
			Linear	203.15 (3.74)	209.00 (3.29)	201.05 (1.25)
		QTL at 200 cM	Quadratic	209.80 (5.08)	211.98 (4.43)	204.95 (2.37)
$\sigma_a^2 = 30$ , $\sigma_0^2 = 30$	2 cM		B-spline	80.02 (1.93)	81.92 (2.15)	81.78 (1.32)
		Linear	79.81 (2.65)	84.65 (1.99)	81.88 (1.20)	
	QTL at 80 cM	Quadratic	79.88 (2.56)	83.63 (2.21)	82.42 (1.38)	
		5 cM	B-spline	204.03 (5.73)	207.20 (5.95)	206.10 (4.40)
			Linear	210.23 (6.74)	209.68 (5.47)	203.80 (3.19)
		QTL at 200 cM	Quadratic	201.08 (7.07)	207.70 (5.76)	208.33 (4.25)

**Table 3 Power of likelihood ratio test**

Variance	Distance (cM)	Method	Case 1	Case 2	Case 3
$\sigma_a^2 = 10, \sigma_0^2 = 20$	2	B-spline	0.855	0.765	0.925
		Linear	0.620	0.735	0.920
		Polynomial	0.595	0.685	0.895
	5	B-spline	0.845	0.735	0.930
		Linear	0.610	0.730	0.910
		Polynomial	0.590	0.695	0.875
$\sigma_a^2 = 30, \sigma_0^2 = 30$	2	B-spline	0.470	0.425	0.630
		Linear	0.220	0.415	0.640
		Polynomial	0.230	0.370	0.605
	5	B-spline	0.450	0.430	0.685
		Linear	0.220	0.400	0.685
		Polynomial	0.195	0.360	0.645

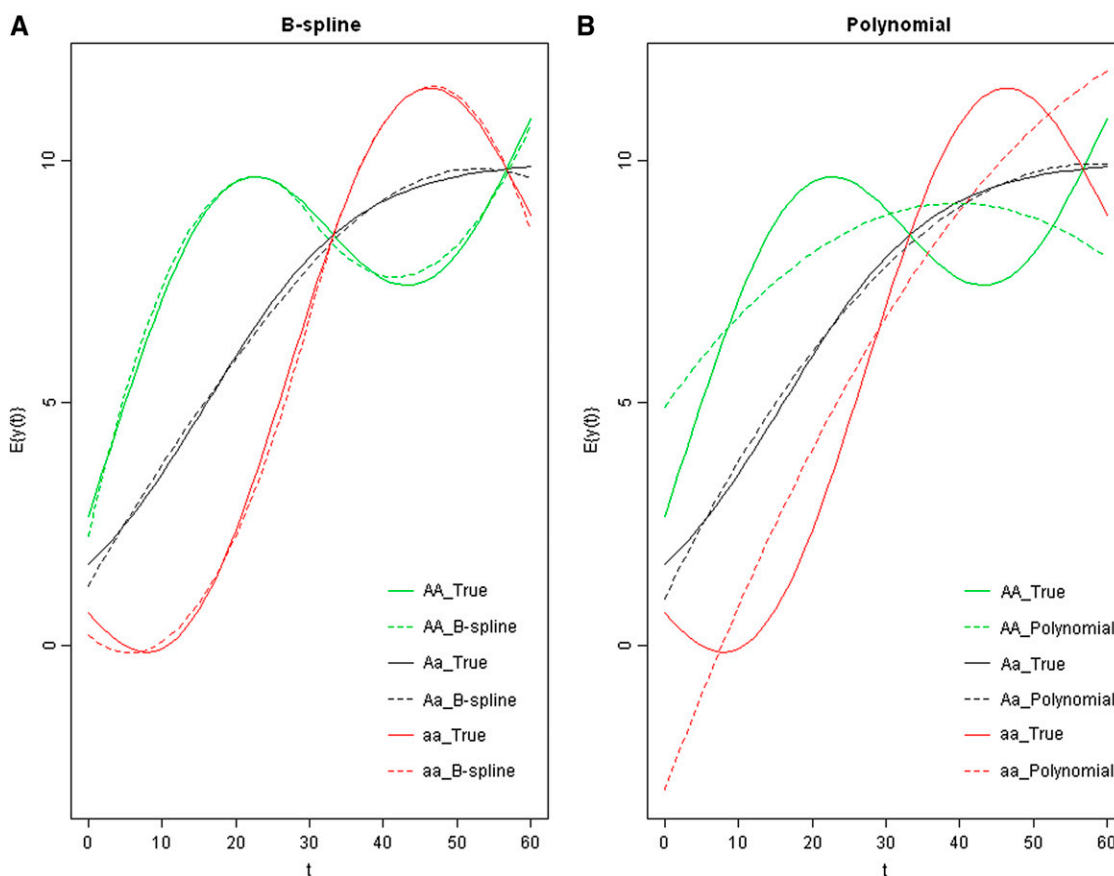
simulations. Table 1, right, shows the number of combinations of  $n_j$  and  $d$  with the smallest AIC. The results are consistent with the SQD results presented in Table 1, left.

For model comparison, we also fitted  $\beta(t)$  parametrically. Specifically, we used polynomial functions

$$\beta(t) = \sum_{k=0}^s \gamma_k t^k.$$

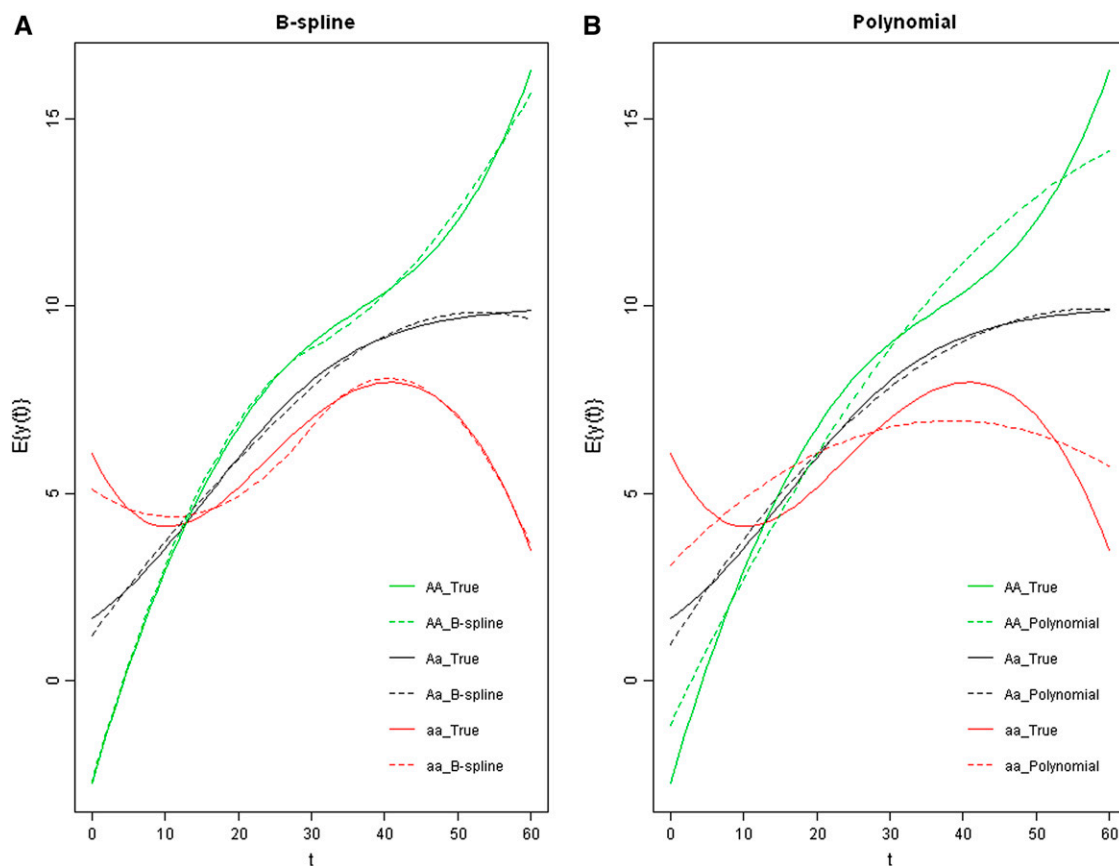
We set  $s = 1$  and  $s = 2$ , for linear and quadratic polynomial functions, in the simulation studies.

Under each case, 200 runs of simulation were conducted for all models mentioned above. For each case, we estimated  $\beta(t)$  using both B-splines and the polynomial functions. Hypothesis testings were performed on  $H_0: \beta(t) = 0$  vs.  $H_a: \beta(t) \neq 0$ , and LOD scores were calculated. For accessing the significance of the hypothesis testing,



**Figure 3** The varying coefficient  $\beta(t) = 1 + 3 \sin(\pi t/30)$ . (A) The estimated phenotypic mean curves by the B-spline method (in dotted lines) with the true genetic curves (in solid lines). (B) The estimated phenotypic mean curves by the polynomial method (in dotted lines) with the true genetic curves (in solid lines).





**Figure 4** The varying coefficient  $\beta(t) = 5/(1+e^{-0.1t})$ . (A) The estimated phenotypic mean curves by the B-spline method (dotted) with the true genetic curves (solid). (B) The estimated phenotypic mean curves by the polynomial method (dotted) with the true genetic curves (solid).

simulations were carried out from the following null model:

$$y_i = \mu(t_i) + \sum_{l=1}^L a_{il}\alpha_l + \epsilon_i.$$

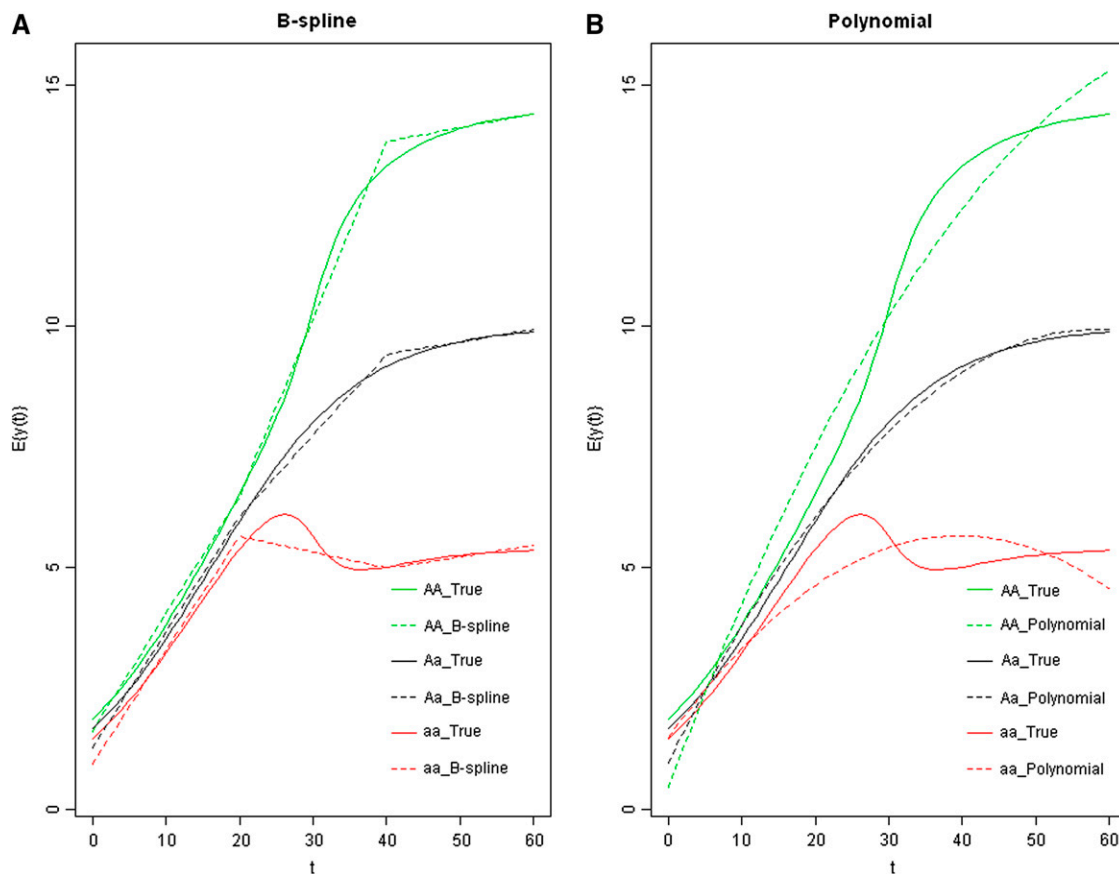
Total 1000 runs of simulations were performed and the 95% percentile of the maximum LOD score was calculated.

The QTL position was estimated as the location where the maximum LOD is reached. The mean and standard error of the estimated QTL position, by the three approaches, are listed in Table 2. Power is listed in Table 3. All the three methods have similar performance on estimating the QTL position and power for mapping QTL under cases 2 and 3. However, the B-spline approach has substantially higher power than the other two approaches under case 1, as well as higher precision in estimating the QTL location. The mean of the estimated phenotypic curves  $\hat{y}(t) = \hat{\mu}(t) + x\hat{\beta}(t)$  are plotted along time in Figure 3, Figure 4, and Figure 5, for all cases with 5-cM intervals,  $\sigma_a^2 = 10$  and  $\sigma_0^2 = 20$ . The nonparametric approach provides better fit to the true underlying phenotypic curves than the parametric approach in all three cases. Overall speaking, the B-spline method outperforms the parametric method.

To evaluate the performance of the modified permutation, we further carried out the following simulation studies. From 100 RI lines, 300 RIX subjects were simulated. A single 100-cM chromosome with evenly spaced markers was simulated with the QTL located at 40 cM. There were either 51 markers separated by 2-cM intervals or 21 markers separated by 5-cM intervals, on the chromosome. Two different  $\beta(t)$  functions, as described in cases 1 and 2 above, were simulated. We set  $\mu(t) = 10/(1+5e^{-0.1t})$ ,  $\sigma_a^2 = 30$ , and  $\sigma_0^2 = 30$ . A total of 100 simulations were conducted. Within each simulation, 1000 permutations were performed and the permutation threshold was calculated. To obtain the empirical thresholds, we ran additional 5000 simulations under  $H_0: \beta(t) = 0$ . We compared the permutation thresholds with the empirical ones. The results listed in Table 4 indicate that the modified permutation performs reasonably well. The permutation thresholds were close to the empirical ones. Type I errors were slightly inflated. This is probably due to the small number of RI lines used in the simulations, as well as the sampling variation due to the small number (100) of simulations conducted.

Besides the biallelic marker data, our method can be extended to model founder effects by fitting individual functional curve for each of the eight CC founder alleles. Assuming additive model, we fit





**Figure 5** The varying coefficient  $\beta(t) = \frac{3}{2}(\arctan((t-30)/4) + \pi/2)$ . (A) The estimated phenotypic mean curves by the B-spline method (dotted) with the true genetic curves (solid). (B) The estimated phenotypic mean curves by the polynomial method (dotted) with the true genetic curves (solid).

$$y_i = \sum_{j=1}^8 \beta_j(t_i) x_{ij} + \sum_{l=1}^L a_{il} \alpha_l + \epsilon_i,$$

where  $x_{ij}$  is the number of the  $j$ th founder alleles in the  $i$ th RIX sample and  $\beta_j(t)$  is the functional effect of the  $j$ th founder allele ( $j = 1, \dots, 8$ ). To model  $\beta_j(t)$  nonparametrically, we use  $\sum_{k=1}^K \gamma_{jk} B_k(t)$  to approximate it.

We carried out a simple simulation study to demonstrate the performance of this model. We again simulated 300 RIX subjects from 100 RI lines by the loop design. Every parental RI line has the same probability,  $1/8$ , to carry one of the eight founder alleles. We set  $\beta_j(t) = 10/(1+5e^{-0.1t})$  for  $j = 1, 2, 3, 4$  and  $\beta_j(t) = 1+3 \sin(\pi t/30)$  for  $j = 5, 6, 7, 8$ , respectively. We further set  $\sigma_a^2 = 10$  and  $\sigma_0^2 = 20$ . The simulations were conducted for 100 runs. In the simulations, we

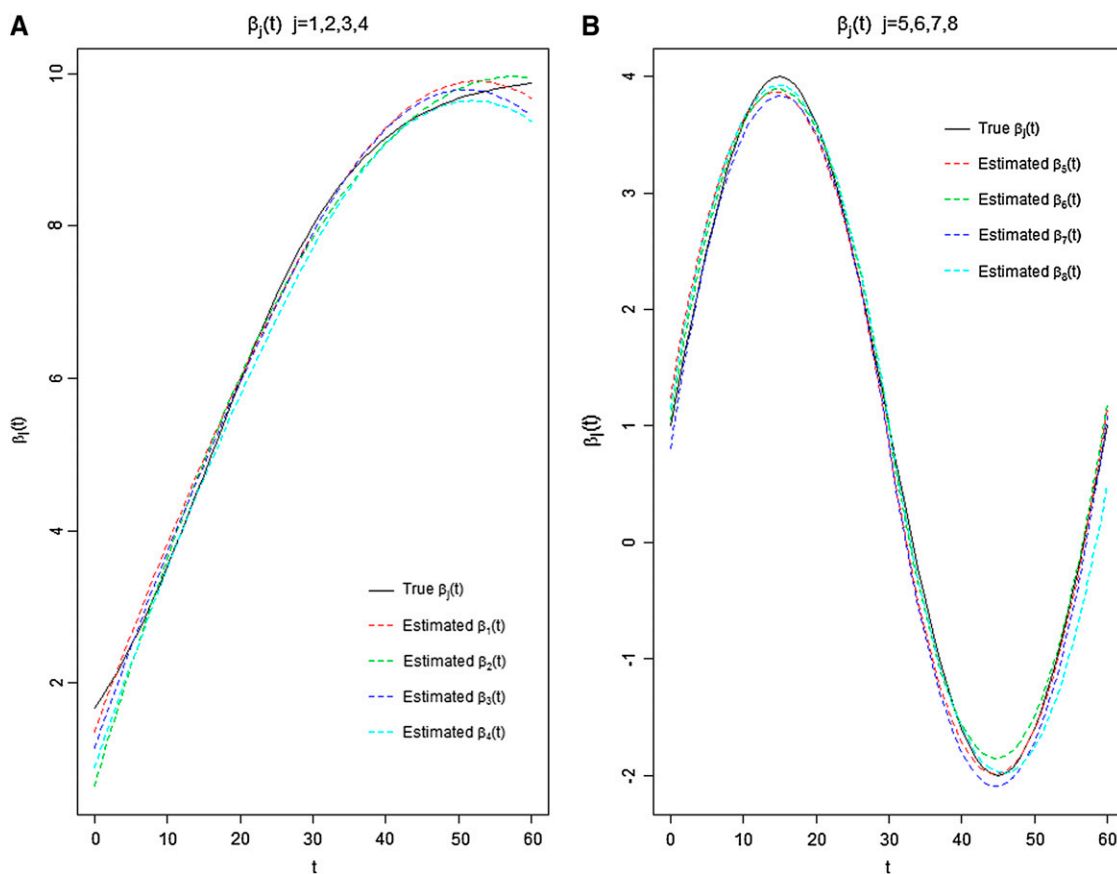
assumed that the founder alleles were known for all RI lines. The means of the estimated functions of the eight founder alleles were plotted in Figure 6. All estimate the true functions well. With no prior knowledge that the genetic effects of the first four founder alleles were the same, our model obtained four very similar estimated curves, which allows us to group founder alleles with similar genetic effects.

## Discussion

This study is largely motivated by the availability of the CC lines (Collaborative Cross Consortium 2012; Kelada *et al.* 2012). The CC project aims to generate and maintain >300 multiparental CC RI lines, and our ability to map complex traits will be greatly increased by making use of

**Table 4** Threshold and power estimates with LOD scores

	Interval (cM)	Empirical	Permutation		
			$\beta(t) = 0$	Case 1	Case 2
LOD score	2	3.92	3.83	3.82	3.80
	5	3.70	3.56	3.54	3.67
Power	2	0.05	0.07	0.60	0.66
	5	0.05	0.07	0.65	0.59



**Figure 6** (A) The true varying genetic effect,  $\beta_j(t)$ , for founder alleles 1, 2, 3, and 4 equals  $10/(1+5e^{-0.1t})$  (solid curve). Dotted curves are the mean estimates of  $\beta_j(t)$  for  $j = 1, 2, 3, 4$ . (B) The true varying genetic effect,  $\beta_j(t)$ , for founder alleles 5, 6, 7, and 8 equals  $1+3 \sin(\pi t/30)$  (solid curve). Dotted curves are the mean estimates of  $\beta_j(t)$  for  $j = 5, 6, 7, 8$ .

these resources. RIX samples derived from CC RI lines process some good properties from both RI lines and  $F_2$  populations. Genotypes of RIX can be directly inferred from those of their parental RI lines. Unlike the parental RIs whose genotypes are homozygous, the genetic structure of an RIX resembles  $F_2$  animals, reducing the phenotypic anomalies associated with inbred genomes. However, RIX animals typically do not share the same degree of relatedness. This unbalanced genetic relatedness requires careful statistical modeling to avoid a large number of false-positive findings. The functional mapping idea is not new in statistical genetics community (Ma *et al.* 2002; Wu *et al.* 2002, 2004; Lin and Wu 2006; Yang *et al.* 2009). However, this article is the first one that develops the functional mapping method for the RIX data and specifically models the unique genetic structure of RIX samples. In addition to B-spline approximation, other nonparametric approaches can be used to model the varying coefficients, such as the local polynomial regression (Fan and Gijbels 1996), the smoothing splines (Hastie and Tibshirani 1993; Hoover *et al.* 1998), and wavelet-based approaches (Donoho and Johnstone 1994). One advantage of using B-splines is that the smoother matrix  $\{B_k(t_i)\}$  is independent of the responses. Unlike other nonparametric approaches, how to determine the smoothness is

still an open question, although the choice of the number of knots is generally not critical (Yang *et al.* 2009). Our simulation results (for example, Figures 1 and 2) show that the estimated functional effects are not very sensitive to the choices of  $d$  and  $n_j$ .

In our simulations, we applied single marker analysis because the high marker density of the parental RI (Aylor *et al.* 2011; Durrant *et al.* 2011; Collaborative Cross Consortium 2012; Kelada *et al.* 2012), and thus RIX, makes results similar to those that would be obtained using more complicated mapping methods, such as traditional interval mapping (Lander and Botstein 1989) or regression interval mapping (Haley and Knott 1992). We also assume no parent-of-origin QTL and polygenic effects. The model can be extended to include additional effects. For example, with two random effects—one for the maternal effect and another for the paternal effect—we can model the parent-of-origin polygenic effects. Our method mainly considers quantitative trait nucleotide (QTN) mapping, we have shown by simulation studies that it can be extended to model founder allelic effects by fitting one functional curve for each of the eight CC founder alleles. Although our model considers only the additive genetic effects, the dominant effects can be easily included in the model by adding additional functional effects.

This is not a concern for QTN models but for models with founder allelic effects, the subsequent increase in the number of parameters can be very large. We may need to consider grouping certain founder alleles with some prior knowledge or genetic similarity as was done in haplotype analysis (Schaid *et al.* 2002; Park *et al.* 2003; Wang *et al.* 2004; Lin *et al.* 2005) to maximize mapping power.

When more than one QTL is on a chromosome, the test statistic at one position will be affected by all the other QTL, the genetic estimates are likely to be biased, and QTL can be mapped to wrong positions (Knott and Haley 1992; Martinez and Curnow 1992). Our model can be extended to multiple regression for multiple QTL mapping, and some model selection approaches can be modified for QTL selection.

Our model investigates gene-by-time interactions for RIX data in a flexible nonparametric fashion. In this model, correlations among subjects are modeled as a function of their relatedness, which dramatically simplifies the covariance matrix of the data. The final result is a framework for mapping in complex genetic designs, which is computationally tractable.

## Acknowledgments

The authors are grateful for constructive comments and suggestions from the reviewers and the associate editor. Support was provided in part by National Institute of General Medical Sciences R01GM074175 and NIMH/NHGRI Center of Excellence in Genomic Sciences (P50MH090338).

## Literature Cited

- Akaike, H., 1970 Statistical predictor identification. *Ann. Inst. Stat. Math.* 22: 203–217.
- Akaike, H., 1974 A new look at the Statistical model identification. *IEEE Trans. Automat. Contr.* 19: 716–723.
- Almasy, L., and J. Blangero, 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* 62: 1198–1211.
- Amos, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* 54: 535–543.
- Anderson, M. J., and C. J. F. ter Braak, 2003 Permutation tests for multi-factorial analysis of variance. *J. Statist. Comput. Simulation* 73: 85–113.
- Aylor, D. L., W. Valdar, W. Foulds-Mathes, R. J. Buus, R. A. Verdugo *et al.*, 2011 Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Res.* 21: 1213–1222.
- Broman, K. W., 2005 The genomes of recombinant inbred lines. *Genetics* 169: 1133–1146.
- Broman, K. W., 2012 Genotype probabilities at intermediate generations in the construction of multiple-strain recombinant inbred lines. *Genetics* 190: 403–412.
- Broman, K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19: 889–890.
- Cheverud, J. M., J. J. Rutledge, and W. R. Atchley, 1983 Quantitative genetics of development: genetic correlations among age-specific trait values and the evolution of ontogeny. *Evolution* 37: 895–905.
- Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* 138: 963–971.
- Churchill, G. A., and R. W. Doerge, 2008 Naive application of permutation testing leads to inflated type I error rates. *Genetics* 178: 609–610.
- Churchill, G. A., D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie *et al.*, 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.* 36: 1133–1137.
- Cleveland, W. S., E. Grosse, and W. M. Shyu, 1991 Local regression models, pp. 309–376 in *Statistical Models in S*, edited by J. M. Chambers, and T. J. Hastie. Wadsworth & Brooks, Pacific Grove, CA.
- Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics* 190: 389–401.
- Darvasi, A., 1998 Experimental strategies for the genetic dissection of complex traits in animal models. *Nat. Genet.* 18: 19–24.
- Donoho, D. L., and I. M. Johnstone, 1994 Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81: 425–455.
- Durrant, C., H. Tayem, B. Yalcin, J. Cleak, L. Goodstadt *et al.*, 2011 Collaborative Cross mice and their power to map host susceptibility to *Aspergillus fumigatus* infection. *Genome Res.* 21: 1239–1248.
- Fan, J., and I. Gijbels, 1996 *Local Polynomial Modeling and Its Applications*, Chapman & Hall, London.
- Foulkes, W. D., K. Metcalfe, P. Sun, W. M. Hanna, H. T. Lynch *et al.*, 2004 Estrogen receptor status in BRCA1- and BRCA2-related breast cancer: the influence of age, grade, and histological type. *Clin. Cancer Res.* 10: 2029–2034.
- Goldgar, D. E., 1990 Multipoint analysis of human quantitative genetic-variation. *Am. J. Hum. Genet.* 47: 957–967.
- Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69: 315–324.
- Hastie, T. J., and R. J. Tibshirani, 1993 Varying-coefficient models. *J. R. Stat. Soc., B* 55: 757–796.
- He, X. M., and P. D. Shi, 1998 Monotone B-spline smoothing. *J. Am. Stat. Assoc.* 93: 643–650.
- Hoeschele, I., 2007 Mapping quantitative trait loci in outbred populations, pp. 623–677 in *Handbook of Statistical Genetics*, Vol. 1, edited by D. J. Balding, M. Bishop, and C. Cannings. Wiley, New York.
- Hoover, D. R., J. A. Rice, C. O. Wu, and L. P. Yang, 1998 Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 85: 809–822.
- Huang, J. Z., C. O. Wu, and L. Zhou, 2004 Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* 14: 763–788.
- Jansen, R. C., and P. Stam, 1994 High-resolution mapping of quantitative traits into multiple loci via interval mapping. *Genetics* 136: 1447–1455.
- Kao, C. H., Z. B. Zeng, and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. *Genetics* 152: 1203–1216.
- Kelada, S. N. P., D. L. Aylor, B. C. E. Peck, J. F. Ryan, U. Tavarez *et al.*, 2012 Genetic analysis of hematological parameters in incipient lines of the Collaborative Cross. *G3: Genes, Genomes, Genetics* 2: 157–165.
- Knott, S. A., and C. S. Haley, 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genet. Res.* 60: 139–151.
- Lander, E. S., and D. Botstein, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage results. *Genetics* 121: 185–199.

- Lin, D. Y., D. Zeng, and R. Millikan, 2005 Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genet. Epidemiol.* 29: 299–312.
- Lin, M., and R. L. Wu, 2006 A joint model for nonparametric functional mapping of longitudinal trajectories and time-to-events. *BMC Bioinformatics* 7: 138.
- Ma, C., G. Casella, and R. L. Wu, 2002 Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics* 161: 1751–1762.
- Martinez, O., and R. N. Curnow, 1992 Estimation the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet.* 85: 480–488.
- McCulloch, C. E., and S. R. Searle, 2001 *Generalized, Linear and Mixed Models*, Wiley-Interscience, New York.
- Park, Y. G., R. Clifford, K. H. Buetow, and K. W. Hunter, 2003 Multiple cross and inbred strain haplotype mapping of complex-trait candidate genes. *Genome Res.* 13: 118–121.
- Pittman, J., 2002 Adaptive splines and genetic algorithms. *J. Comput. Graph. Statist.* 11: 615–638.
- Plomin, R., G. E. McClearn, G. Gora-Maslak, and J. M. Neiderhiser, 1991 An RI QTL cooperative data bank for recombinant inbred quantitative trait loci analyses. *Behav. Genet.* 21: 97–98.
- Satagopan, J. M., B. S. Yandell, M. A. Newton, and T. C. Osborn, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* 144: 805–816.
- Schaid, D. J., C. M. Rowland, D. E. Tines, R. M. Jacobson, and G. A. Poland, 2002 Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* 70: 425–434.
- Sillanpaa, M. J., and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* 148: 1373–1388.
- Threadgill, D. W., K. W. Hunter, and R. W. Williams, 2002 Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm. Genome* 13: 175–178.
- Tsaih, S. W., L. Lu, D. C. Airey, R. W. Williams, and G. A. Churchill, 2005 Quantitative trait mapping in a diallel cross of recombinant inbred lines. *Mamm. Genome* 16: 344–355.
- Vaughn, T. T., L. S. Pletscher, A. Peripato, K. King-Ellison, E. Adams *et al.*, 1999 Mapping quantitative trait loci for murine growth: a closer look at genetic architecture. *Genet. Res.* 74: 313–322.
- Wang, L., G. Chen, and H. Li, 2007 Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* 23: 1486–1494.
- Wang, L., H. Li, and J. Z. Huang, 2008 Variable selection in non-parametric varying-coefficient models for analysis of repeated measurements. *J. Am. Stat. Assoc.* 103: 1556–1569.
- Wang, X., R. Korstanje, D. Higgins, and B. Paigen, 2004 Haplotype analysis in multiple crosses to identify a QTL gene. *Genome Res.* 14: 1767–1772.
- Wu, R. L., C. Ma, R. Littell, and G. Casella, 2002 A statistical model for the genetic origin of allometric scaling laws in biology. *J. Theor. Biol.* 217: 275–287.
- Wu, R. L., C. Ma, M. Lin, and G. Casella, 2004 A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics* 166: 1541–1551.
- Xu, Z. L., F. Zou, and T. J. Vision, 2005 Improving QTL mapping resolution in experimental crosses by the use of genotypically selected samples. *Genetics* 170: 401–408.
- Yang, J., R. L. Wu, and G. Casella, 2009 Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. *Biometrics* 65: 30–39.
- Yi, N., 2004 A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* 167: 967–975.
- Yi, N., and S. Xu, 2000 Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics* 155: 1391–1403.
- Yuan, Z., F. Zou, and Y. Liu, 2011 Bayesian multiple quantitative trait loci mapping for recombinant inbred intercrosses. 188: 189–195.
- Zeng, Z. B., 1994 Precision mapping of quantitative trait loci. *Genetics* 136: 1457–1468.
- Zou, F., B. S. Yandell, and J. P. Fine, 2001 Statistical issues in the analysis of quantitative traits in combined crosses. *Genetics* 158: 1339–1346.
- Zou, F., J. L. Gelfond, D. C. Airey, L. Lu, K. F. Manly *et al.*, 2005 Quantitative trait locus analysis using recombinant inbred intercrosses (RIX): theoretical and empirical considerations. *Genetics* 170: 1299–1311.

*Edited by Lauren M. McIntyre, Dirk-Jan de Koning, and 4 dedicated Associate Editors*

# GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.111.132522/-/DC1>

## Varying Coefficient Models for Mapping Quantitative Trait Loci Using Recombinant Inbred Intercrosses

Yi Gong and Fei Zou

## File S1

### Compressed R Package

File S1 is available at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.111.132522/-/DC1> as a compressed folder.

The folder contains three programs written in R:

- sim1.r - for simulations for the biallelic marker model
- sim2.r - for simulations for the eight founder alleles model
- funs.r - basic functions used in sim1.r and sim2.r