Biometrics 65, 30–39
March 2009
DOI: 10.1111/j.1541-0420.2008.01063.x

Nonparametric Functional Mapping of Quantitative Trait Loci

Jie Yang,[1,*] Rongling Wu,[2,] and George Casella[1,2,***]**

[1]Genetics Institute, University of Florida, Gainesville, Florida 32611, U.S.A.
[2]Department of Statistics, University of Florida, Gainesville, Florida 32611, U.S.A.
[*]*email:* jyang81@.ufl.edu
[**]*email:* rwu@stat.ufl.edu
[***]*email:* casella@stat.ufl.edu

Summary. Functional mapping is a useful tool for mapping quantitative trait loci (QTL) that control dynamic traits. It incorporates mathematical aspects of biological processes into the mixture model-based likelihood setting for QTL mapping, thus increasing the power of QTL detection and the precision of parameter estimation. However, in many situations there is no obvious functional form and, in such cases, this strategy will not be optimal. Here we propose to use nonparametric function estimation, typically implemented with B-splines, to estimate the underlying functional form of phenotypic trajectories, and then construct a nonparametric test to find evidence of existing QTL. Using the representation of a nonparametric regression as a mixed model, the final test statistic is a likelihood ratio test. We consider two types of genetic maps: dense maps and general maps, and the power of nonparametric functional mapping is investigated through simulation studies and demonstrated by examples.

Key words: B-splines; Maximum likelihood; Mixed model.

1. Introduction

The past two decades, since Lander and Botstein's (1989) pioneering work, have witnessed the considerable development of statistical methodologies for mapping quantitative trait loci (QTL) with genetic linkage maps (Jansen and Stam, 1994; Zeng, 1994; Sen and Churchill, 2001; Kao and Zeng, 2002; Wu, Ma, and Casella, 2007). These methods have been instrumental for the identification of significant QTL that contribute to phenotypic variation in a variety of quantitative traits in plants (Paterson, 2006), animals (Anholt and Mackay, 2004), and humans (Weiss et al., 2005, 2006).

Ma, Casella, and Wu (2002) proposed a statistical framework, called functional mapping, for mapping QTL that regulate developmental trajectories of a trait, allowing, for example, to test the timing of a QTL to turn on or turn off, and the duration of QTL expression (Wu et al., 2004). Because the parametric functions chosen are usually derived from a universal biological law, functional mapping facilitates the testing of numerous biologically meaningful hypotheses by testing separately or jointly the parameters that define the curves. Functional mapping uses fewer parameters to model biological processes, and can increase the power of QTL mapping. While the parametric nature of functional mapping offers tremendous biological and statistical advantages, a reliance on the availability of mathematical functions limits its applicability. In some cases there are many different functions that describe the same phenotypic trajectory. For example, there are functions in three different categories to describe a growth trajectory: exponential, saturating, and sigmoidal (Von Bertalanffy, 1957; Niklas, 1994). Thus, it may not be clear which one should be used, especially when there are not enough observations for each subject to show obvious characteristics. Moreover, in many situations, there are no obvious functional forms.

These issues have started to draw the attention of several statistical geneticists. For example, Yang, Tian, and Xu (2006) and Yang and Xu (2007) attempted to use Legendre orthogonal polynomials to fit various shapes of curves and further test the dynamic genetic effects of QTL in a time course. The motivation of Yang et al.'s model stemmed from the strong application of the normalized Legendre polynomial to the prediction of breeding values and time-dependent covariance in dairy milk production (Meyer, 2005a, 2005b). Lin and Wu (2006) incorporated Legendre orthogonal polynomials in joint modeling of longitudinal and time-to-event traits to test whether a pleiotropic QTL exists affecting vegetative growth and reproductive behavior in plants.

In all aforementioned models multiple tests have to be performed on many putative QTL positions across the whole genome. Then some form of adjustment of the critical threshold value of the likelihood ratio test (LRT) statistic is necessary to control the genome-wise type I error rate. Permutation test procedures advocated by Churchill and Doerge (1994) and Doerge and Churchill (1996) are a popular approach to find critical values because of their conceptual simplicity, distribution-free nature, and generality in different population structures. But this method has a serious drawback in its computational intensity. For a genome-wise type I error at 0.01, at least 10,000 permutations are required (Doerge and Rebaï, 1996). An alternative is to use approximate threshold

values for maps of different intensity and different population structures. Lander and Botstein (1989, 1994) derived an approximation formula for an infinitely dense map and backcross population. Rebaï, Goffinet, and Mangin (1994) and Piepho (2001) also gave formulas to approximate critical values in the intermediate density case. These methods only considered a univariate phenotypic trait.

In this article, we introduce a nonparametric functional mapping framework for genetic mapping of QTL controlling for a dynamic trait based on a more general and flexible nonparametric smoothing approach, implemented with B-splines. In our setting an exact *p*-value calculation formula for genome-wise inference is provided for a dense map. For a map with intermediate or sparse density we propose a computationally less intensive method to find critical values that control the genome-wise type I error rate for the LRT statistic.

We incorporate B-splines (He and Shi, 1998; Pittman, 2002) into the procedure for modeling biological processes and formulate a nonparametric test for the existence of a QTL and its effects on various biological events and processes within the the maximum likelihood context. (An excellent introduction to B-splines is given by Eilers and Marx, 1996.) In Section 2, we develop the model and derive the test statistics; the subject-specific model we propose can incorporate both within- and among-subject variation due to individual environmental adjustments. Section 3 contains applications to real data, as well as simulation studies, and we show that nonparametric functional mapping works well in a variety of situations. Section 4 contains some conclusions.

2. Statistical Modeling

2.1 *Genetic Design*

For simplicity, we illustrate our method using a backcross population. Consider a backcross with N individuals, each genotyped with polymorphic markers to construct a genetic linkage map. This map is used to identify the genome-wide distribution of QTL that control a dynamic trait of interest. All the backcross individuals are measured repeatedly for the trait at a multitude of time points. Let y_{ij} be the observed value of individual i for the trait at observation point $t_j (i = 1, \ldots, N$ and $j = 1, \ldots, T)$.

Now suppose there is a putative QTL segregating with two different genotypes Qq (coded by 1) and qq (coded by 2) in the assumed backcross that affects the shape of the dynamic trait. For dense maps, we can assume that the QTL is one of the markers and the genotypes of each putative QTL are known. For general maps, the unknown QTL can be detected by the linkage map. In this case, assume that the QTL resides between a pair of flanking markers $\mathcal{M}_1$ (with two alleles M_1 and m_1) and $\mathcal{M}_2$ (with two alleles M_2 and m_2). For each backcross individual, it may carry one (and only one) QTL genotype, 1 or 2. The probability of a particular individual (i) to carry QTL genotype 1 or 2 depends on the marker genotype of this individual at the two flanking markers ($\mathcal{M}_1$ and $\mathcal{M}_2$) that bracket the QTL. Let r_1, r_2, and r be the recombination fractions between $\mathcal{M}_1$ and QTL, between QTL and $\mathcal{M}_2$, and between the two markers, respectively. Under the assumption of no double crossovers, we approximate the conditional probability (p_{iq}) of a QTL genotype (q) given the marker genotype of individual i as a function of $\theta = r_1/r$ (Web Table 1), $q = 1, 2$.

2.2 *Subject-Specific Model*

As pointed out above, each backcross individual should carry one and only one of the two possible QTL genotypes. The phenotypic value of individual i can be described by a general linear model,

$$\mathbf{y}_i(\vec{\mathbf{t}}) = \delta_{iq}\mu_q(\vec{\mathbf{t}}) + \alpha_i \mathbf{1}_T + \boldsymbol{\varepsilon}_i(\vec{\mathbf{t}}), \tag{1}$$

where $\vec{\mathbf{t}} = (t_1, t_2, \ldots, t_T)'$, $\mu_1(\vec{\mathbf{t}})$ and $\mu_2(\vec{\mathbf{t}})$ are the genotypic mean vectors of Qq and qq, respectively, δ_{iq} is an indicator variable for individual i, defined as 1 if a particular QTL is indicated and 0 otherwise (for a backcross population, $q = 1, 2$), $\alpha_i \mathbf{1}_T$ models the covariance structure of observations among individuals, and ε_i is a parameter that accounts for the within-individual covariance structure of the observations on individual i. The variables α_i and ε_i are independently distributed with normal distribution $N(0, \sigma^2)$ and multivariate normal distribution $MVN(0, V_T)$, respectively. The probability that $\delta_{iq} = 1$ depends on the genotype of the flanking markers and the position of the QTL on the marker interval.

Let $\mathbf{B} = \{\beta_\ell(t_i)\}_{T\times L}$ be a smoothing matrix composed of L ($L \leq T$) B-spline basis functions at T time points. Then, we have $\mu_q(\vec{\mathbf{t}}) = \mathbf{B}\xi_q$, where ξ_q is the coefficient vector for the matrix $\mathbf{B}$. The comparison of these coefficient vectors between two different QTL genotypes can determine whether this putative QTL affects phenotypic trajectories.

2.3 *Estimation and Tests*

Depending on the density of the map we have different models and testing strategies.

2.3.1 *Dense map.* If the markers on the linkage map are dense enough so that we can assume that the QTL are located on the marker locus or very close to it, then we know the value of δ_{iq} without requiring Web Table 1. In this situation, the genome-wise search for existing QTL amounts to testing the null hypothesis $H_0 : \mu_1 = \mu_2$ *at every marker* versus $H_1 : \mu_1 \neq \mu_2$ *at some marker*. This is a multiple testing problem with each test deciding whether two underlying functions are different at a particular marker location.

We start by deriving LRTs at each particular marker, $H_0' : \mu_1 = \mu_2$ *at a particular marker k*, and then extend our result to the entire genome.

Let $n_{k,1}$ and $n_{k,2}$ be the total number of subjects in each group, with $n_{k,1} + n_{k,2} = N$. Then,$\mathbf{y}_i(\vec{\mathbf{t}}) \sim I(i \leq n_{k,1}) \times MVN\,(\mathbf{B}\xi_1, \Sigma) + I(n_{k,1} < i \leq N)\; MVN\,(\mathbf{B}\xi_2, \Sigma)$. The likelihood function for$\{\xi_1, \xi_2\}$ given a fixed Σ is

$$L \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_{k,1}} \|\mathbf{y_i} - \mathbf{B}\xi_1\|_\Sigma - \frac{1}{2}\sum_{i=n_{k,1}+1}^{N} \|\mathbf{y_i} - \mathbf{B}\xi_2\|_\Sigma\right\}.$$

From such likelihood function it is straightforward to calculate the likelihood test statistic $\lambda_k = \max_{H_0'} L/\max L$, which can be written as (Web Appendix A)

$$\begin{aligned} G_k &= \frac{n_{k,1}n_{k,2}}{N}\|\bar{\mathbf{y}}_{k,1} - \bar{\mathbf{y}}_{k,2}\|_{A_0} \\ &\sim \chi^2_L\left(\frac{1}{2}(\mu_1 - \mu_2)' A_0 (\mu_1 - \mu_2)\right), \end{aligned} \tag{2}$$

where $A_0 = \Sigma^{-1}B(B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}$. Under H_0', each LRT statistic G has a central chi-squared distribution with degrees of freedom L, the number of B-spline basis functions used in the estimation of the phenotypic curves. H_0' is rejected when G is large. For the genome-wise hypothesis test, the test statistic is the maximum of all test statistics at each putative QTL position, which, in this case, is each marker. When this statistic is large enough we can conclude that there is significant evidence of existing QTL and the positions where the test statistics exceed the critical value indicate the possible QTL locations.

Under this dense map assumption we can derive a formula to calculate threshold values for genome-wise testing directly. If we write $\mathbf{Y}_k = (y_{11}, \ldots, y_{1T}, y_{21}, \ldots, y_{NT})'$, where the first $n_1 * Ty_{ij}$s are from the group with genotype Qq, then the test statistic could be written as a quadratic form $G_k = \mathbf{Y}_k' A_k \mathbf{Y}_k$, where

$$A_k = \frac{n_{k,1}n_{k,2}}{N}\begin{pmatrix} \frac{1}{n_{k,1}^2}J_{n_{k,1}} & -\frac{1}{n_{k,1}n_{k,2}}\mathbf{1}_{n_{k,1}} \otimes \mathbf{1}'_{n_{k,2}} \\ -\frac{1}{n_{k,1}n_{k,2}}\mathbf{1}_{n_2} \otimes \mathbf{1}'_{n_{k,1}} & \frac{1}{n_{k,2}^2}J_{n_{k,2}} \end{pmatrix} \otimes A_0 \equiv U \otimes A_0, \tag{3}$$

and $\mathrm{var}(\mathbf{Y}) = I_N \otimes \Sigma \equiv \tilde{\Sigma}$. (When there are three or more possible genotypes at each marker [e.g., an F_2 population or single-nucleotide polymorphism (SNP); genotypes] the expression for G_k can be found in Web Appendix B.) Using a permutation matrix P, we can write the LRT statistic G_k corresponding to each marker k, $k = 1, \ldots, m$, as $G_k = \mathbf{Y}_1' P_k' \times A_k P_k \mathbf{Y}_1$, where Y_1 represents the Y vector corresponding to the first marker. Under H_0, these test statistics G_ks have the same chi-squared distribution with degrees of freedom L but are correlated with each other.

Rewrite $G_k = \mathbf{Z}_k' \mathbf{Z}_k$, where $\mathbf{Z}_k = W_k' \tilde{\Sigma}^{-\frac{1}{2}} P_k \mathbf{Y}_1 \sim MVN \times (W_k' \tilde{\Sigma}^{-\frac{1}{2}} P_k \tilde{\mu}, W_k' W_k)$ and $W_k W_k'$ are the spectral decompositions of $\tilde{\Sigma}^{\frac{1}{2}} A_k \tilde{\Sigma}^{\frac{1}{2}}$. The entire vector $\mathbf{Z} = (\mathbf{Z}_1', \mathbf{Z}_2', \ldots, \mathbf{Z}_m')'$ has distribution $\mathbf{Z} \sim MVN(\mu_Z, \Delta)$ where $\mu_Z = (\mu_{Z_1}, \ldots, \mu_{Z_m}), \mu_{Z_i} = \mathbf{W}_i' \tilde{\Sigma}^{-\frac{1}{2}} \tilde{\mu}$ and $\Delta = \tilde{\mathbf{W}}' \tilde{\Sigma} \tilde{\mathbf{W}}$ with $\tilde{\mathbf{W}} = (\tilde{\Sigma}^{-\frac{1}{2}} \mathbf{W}_1, P_2' \tilde{\Sigma}^{-\frac{1}{2}} \mathbf{W}_2, \ldots, P_m' \tilde{\Sigma}^{-\frac{1}{2}} \mathbf{W}_m)$. It is straightforward to show that under H_0, $\mu_Z = 0$.

If we let B_x denote the L-dimensional ball with radius equal to x, then

$$\begin{aligned} P_{H_0}\Big(\max_{1\le k\le m} \mathbf{Y}_k' A_k \mathbf{Y}_k \le x\Big) &= P_{H_0}\Big(\max_{1\le k\le m} \mathbf{Z}_k' \mathbf{Z}_k \le x\Big) \\ &= P_{H_0}(\mathbf{Z}_1'\mathbf{Z}_1 \le x, \ldots, \mathbf{Z}_m'\mathbf{Z}_m \le x) \\ &= \int \cdots \int_{\{\mathbf{Z}_i \in B_x\}} \frac{\exp\left\{-\frac{1}{2}\mathbf{Z}'\Delta^{-1}\mathbf{Z}\right\}}{\sqrt{2\pi}^{mL}|\Delta|^{\frac{1}{2}}} d\mathbf{Z}_1 \cdots d\mathbf{Z}_m . \end{aligned} \tag{4}$$

This probability, which is one minus the p-value for H_0 when $x = \max_k G_k$, can be directly calculated by simulating $\mathbf{Z} \sim MVN(0, \Delta)$, or with importance sampling. In the other direction, once setting this probability equal to $1 - \alpha$, it is easy to numerically search for the threshold value that controls the genome-wise type I error α.

2.3.2 *General map.* When the assumption that possible QTL are located on or very near to marker positions cannot be satisfied, if the above testing procedure for a dense map is still used, obviously the power of detecting existing QTL decreases. In this case, we can use a mixture model following the idea of interval mapping first proposed by Lander and Botstein (1989). The probability that $\delta_{iq} = 1$ is listed in Web Table 1. The ratio θ is unknown and needs to be estimated, but in practical computations, the QTL position parameter θ can be viewed as a fixed parameter because we put a putative QTL at every 1 or 2 cM on a map interval bracketed by two markers throughout the entire genome (Lander and Botstein, 1989).

The likelihood function of these backcross progenies with a general marker map can be represented as a multivariate mixture model

$$L(\Omega) = \prod_{i=1}^{N}\left[\sum_{q=1}^{2} p_{iq} f_q(\mathbf{y}_i)\right], \tag{5}$$

where $f_q(\mathbf{y}_i) = \frac{1}{(2\pi)^{T/2}|\Sigma|^{1/2}} \exp[-\|\mathbf{y}_i - \mathbf{B}\xi_q\|_\Sigma/2]$ with $\|a\|_\Sigma = a'\Sigma a$ and $\Sigma = \sigma^2 J_T + V$ containing the matrix of all ones in J_T, and $\mathbf{\Omega}$ contains unknown parameters that model the QTL effect (ξ_1 and ξ_2) and residual (co)variances. The maximum likelihood estimates (MLEs) of the unknown parameters for a pleiotropic QTL (Ω_q) can be computed by implementing the expectation–maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977).

Let $\Phi(y; \mathbf{B}\xi, \Sigma) \equiv \exp\{-\frac{\|y - \mathbf{B}\xi\|_\Sigma}{2}\}$ and

$$P(y;t) = \frac{(1-t)\Phi(y; \mathbf{B}\xi_1, \Sigma)}{(1-t)\Phi(y; \mathbf{B}\xi_1, \Sigma) + t\Phi(y; \mathbf{B}\xi_2, \Sigma)}. \tag{6}$$

Suppose the kth marker interval is considered and $\mathbf{y}_i$ has been sorted with respect to the four possible combination of the two bracketing markers. The corresponding EM algorithm for a backcross population design is as follows (the subscript k is omitted):

EM Algorithm: For fixed θ and known Σ, iterate until convergence:

Step t: Calculate $P(\mathbf{y}_i; 1-\theta)^{(t)}$ and $P(\mathbf{y}_i; \theta)^{(t)}$ using $\hat{\xi}_1^{(t)}$ and $\hat{\xi}_2^{(t)}$.

Step $t+1$: Calculate

$$\hat{\xi}_1^{(t+1)} = (\mathbf{B}'\Sigma^{-1}\mathbf{B})^{-1}\mathbf{B}'\Sigma^{-1} \times \left(\frac{\sum_{i=1}^{n_1}\mathbf{y}_i + \sum_{i=n_1+1}^{n_2} P(\mathbf{y}_i;\ \theta)^{(t)}\mathbf{y}_i + \sum_{i=n_2+1}^{n_3} P(\mathbf{y}_i; 1-\theta)^{(t)}\mathbf{y}_i}{n_1 + \sum_{i=n_1+1}^{n_2} P(\mathbf{y}_i;\ \theta)^{(t)} + \sum_{i=n_2+1}^{n_3} P(\mathbf{y}_i; 1-\theta)^{(t)}}\right)$$

and

$$\hat{\xi}_2^{(t+1)} = (\mathbf{B}'\Sigma^{-1}\mathbf{B})^{-1}\mathbf{B}'\Sigma^{-1} \times \left(\frac{\sum_{i=n_3+1}^{N}\mathbf{y}_i + \sum_{i=n_1+1}^{n_2}(1 - P(\mathbf{y}_i;\theta)^{(t)})\mathbf{y}_i + \sum_{i=n_2+1}^{n_3}\left(1 - P(\mathbf{y}_i; 1-\theta)^{(t)}\right)\mathbf{y}_i}{N - n_3 + \sum_{i=n_1+1}^{n_2}\left(1 - P(\mathbf{y}_i;\theta)^{(t)}\right) + \sum_{i=n_2+1}^{n_3}\left(1 - P(\mathbf{y}_i; 1-\theta)^{(t)}\right)}\right).$$

The details of deriving this EM algorithm are given in the Web Appendix C.

In classical interval mapping, a profile of LRT statistics across the entire linkage map is constructed through calculating the likelihood ratio at each putative $\mathcal{Q}$ along the map, where the position can be characterized by the outside markers $\mathcal{M}_1, \mathcal{M}_2$ and the position θ. Unlike the dense map case, we do not have an explicit formula for the p-value. However, under our nonparametric setting we can simulate the critical threshold value to control genome-wise type I error in a less computationally intense way than permutation.

Note that when there is no QTL, that is, $\xi_1 = \xi_2$, equation (6) is actually free of the phenotypic value y. So we can directly calculate $\hat{\xi}_1$ and $\hat{\xi}_2$ for each fixed θ without using the EM algorithm. Then the LRT statistic at each putative QTL position can be easily determined using $\hat{\xi}_0, \hat{\xi}_1, \hat{\xi}_2$, and the likelihood map follows. The $(1 - \alpha) * 100$ percentile of the highest peaks of the likelihood maps from 1000 sets of simulated phenotypic values y under H_0 is taken as the cut-off point for significance level α. Simulation studies confirm that our simulation procedure gives threshold values similar to those of the permutation test (Yang, 2006).

All of the above derivations are made under the assumption that we know the (co)variance matrix Σ, which is typically untrue in practice. We suggest substituting a restricted maximum likelihood estimation (REML) estimate of the variance–covariance matrix $\hat{\Sigma}$ from a saturating model with some dependence structure selected using a model selector such as Bayesian information criterion. For example, for a backcross population with a general map we can get a REML estimate of the (co)variance matrix based on the four possible marker genotype combinations of two adjacent markers. Such REML estimates are used when a putative QTL position is placed in this marker interval. The simulation studies in Section 3.2.2 demonstrate that the REML estimate performs similar to the true variance matrix even for moderate sample sizes. For the dense map case, if the chosen structure is correct, then the calculated p-value is correct asymptotically.

3. Examples

We illustrate our procedure on two data sets, one with a dense map, and one with a general map. We also give the results of simulations to compare our procedure with a parametric approach, and evaluate the performance of the REML variance estimate.

3.1 *SNPs Influencing 5-Fluorouracil Cytotoxicity*

In this example, we illustrate how our method for the dense map case can be used to discover SNPs influencing the sensitivity of lymphoblastoid cells to death caused by incubation with 5-fluorouracil, a uracil analog widely used to treat colorectal and breast tumors. The cellular viability data of lymphoblastoid cell lines from 38 Center d'Etude du Polymorphisme Humain (CEPH) families were collected by Dr McLeod's group (Watters et al., 2004) and downloaded from PharmGKB (`www.pharmgkb.org`). Two drugs were considered in their paper: docetaxel and 5-fluorouracil. Here we used 5-fluorouracil as our example. Studied dosages of 5-fluorouracil drug were set at 0 (vehicle only), 0.76, 1.92, 3.84, 5.77, 7.68, 19.2, 38.4, 76.8 μM. Watters et al. (2004) performed genome-wide linkage analysis for QTL influencing 5-fluorouracil cytotoxicity at each dose separately based on microsatellite markers and found a region on chromosome 9q13-q22 with supportive evidence of linkage for 5-fluorouracil response (Figure 1 in Watters et al., 2004).

Because high-resolution SNP genotype data for a subset of CEPH individuals is available (produced by the International Haplotype Map Project, `www.hapmap.org`), we applied our proposed model for the dense map case to perform a fine mapping using SNPs in a 1 log of odds (LoD) interval (about 70–122 cM) that they found on chromosome 9. There were 57 cell lines in Watters et al. (2004) with both SNP data and complete cytotoxicity profiles. Web Figure 1 plots the cell viabilities of 57 cell lines at each dose.

Five order 3 B-splines were used to estimate the cell viability curves with inner knots at 4 μM and 7 μM and an autoregressive(AR)(1) structure was chosen to model the (co)variance matrix. The sex-average genetic length of chromosome 9 is about 164 cM with 176,336 nonredundant SNPs (HapMap project release 22). The selected 6634 SNPs with three genotypes are located within the 1 LOD confidence interval of linkage region, and have minor allele frequency at least 10%. Figure 1a illustrates the LRT results from using nonparametric functional mapping. Based on equation (4), the cut-off value of LRT statistic with family-wise type I error at 0.05 is 39.23 with standard error 0.04, which leads to SNP *rs*7039978 (LRT value = 39.58) being significantly associated different drug-response dynamic curves. If a Bonferroni multiplicity adjustment or a controlling false discovery rate FDR at 0.05 were used, there were no significant SNPs discovered. Figure 1b shows the observed and estimated mean drug-response curve for each genotype group corresponding to SNP *rs*7039978, where the largest LRT value occurred.

3.2 *Genome-Wide Mapping for Poplar Growth Curves*

Here we illustrate our proposed model for a general map through genome-wise mapping for genes controlling stem growth of poplar. This data set comes from an experiment of the triple hybridization of Populus (poplar). The study materials used were described in Ma et al. (2002). Autoregressive regression with AR(1) measurement errors were assumed to model the within-subject correlation and a log transformation was applied to the raw data of stem diameters to stabilize the age-dependent variance heteroscedasticity (Wu et al., 2004). The REML estimate of Σ was calculated from `Proc Mixed` (Littell, Pendergast, and Natarajan, 2000). The empirical estimate of the critical value is obtained from 1000 simulations and we find the threshold value for declaring the genome-wise existence of a QTL is 32.01 at the significance level $p = 0.01$. The QTL candidate positions are the positions corresponding to the peaks of curves higher than the critical value. There is significant evidence showing that several QTL candidates exist in linkage group 1, 2, 4, 7, 10, 14, and 18 to control the growth trajectory of stem diameter in the interspecific hybrids of poplar (Figure 2).

The poplar data were also used by Ma et al. (2002) to illustrate their functional mapping method, where they found a QTL controlling for diameter growth trajectory across linkage group 10 in the Populus deltoides parent map, which was not

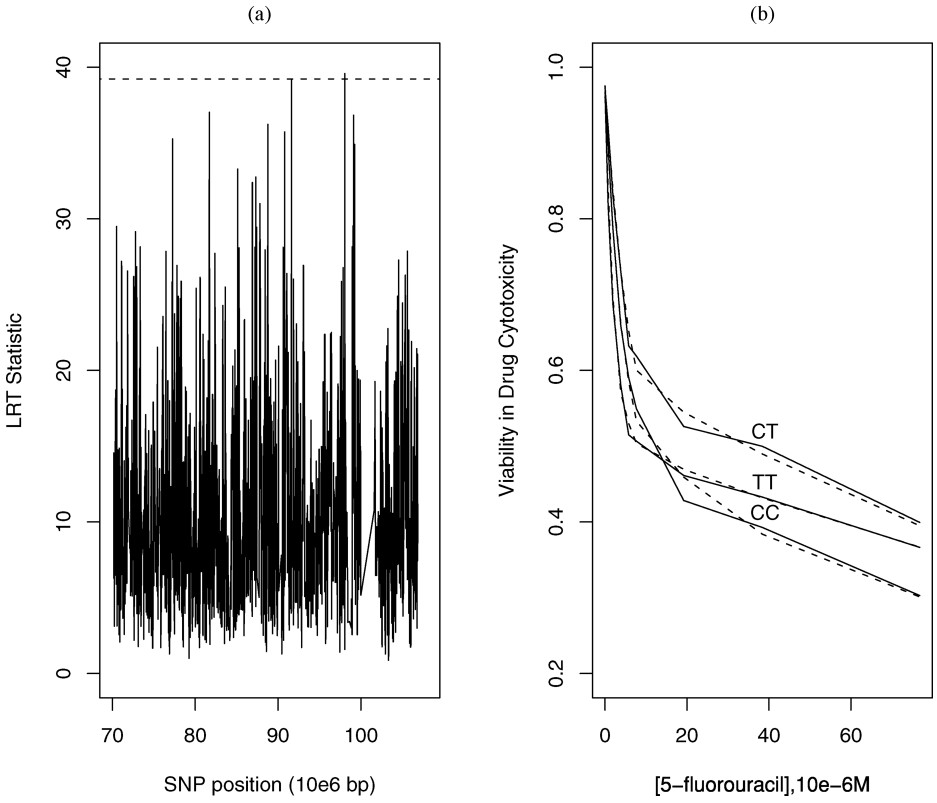

Figure 1. (a) The LRT statistics for all 6635 SNPs using nonparametric functional mapping. The dashed line shows the cut-off point corresponding to significance level 0.05 calculated from equation (4). (b) Mean drug response curves (solid lines) with their nonparametric estimates (dashed lines) of three genotype groups classified by SNP rs7039978.

detected by other traditional models/tools. Using our method we can also find the same QTL, but we do not need to specify the functional form and needed much less computation time. We also applied our proposed model for the dense map case to linkage group 10, where the smallest genetic distance between two neighbor markers is about 10 cM. The small p-value ($p = 0.037$ with standard error of $4.6e-4$) suggests that a QTL exists. The biggest LRT statistic G appears at the first marker, Marker *CA/CCC-640R*, which is also consistent with Ma et al. (2002), where they found a QTL located about 13 cM away from the first marker and 20 cM away from the second marker.

3.3 *Simulation Studies*

We present simulation studies to compare our method to the parametric procedure of Ma et al. (2002), and to evaluate the performance of the REML variance estimators.

3.3.1 *Comparison of parametric and nonparametric functional mapping.* We assume a backcross in which 10 equidistant markers are simulated to generate a genome with length 180. A QTL was located between markers 5 and 6, 88 cM from marker 1. The dependence structure for log-transformed observations on the same individual was set to be autoregressive with order 1. The phenotypic values, $\mathbf{y}_i(\vec{\mathbf{t}})$, at different time points were simulated by assuming $\mathbf{y}_i(\vec{\mathbf{t}})$ to be distributed as $MVN(\delta_{i1}\mu_1(\vec{\mathbf{t}}) + (1-\delta_{i1})\mu_2(\vec{\mathbf{t}}), \sigma^2 J_T + V)$. Two different simulation scenarios were performed, each assuming different sample sizes ($N = 100$ in study $S1$ and 400 in study $S2$), and different heritabilities for longitudinal traits in a middle of time course ($H^2 = 0.1$ and 0.4).

In the first simulation scenario, we assumed that the genotypic means are logistic growth curves $\frac{20}{1+20e^{-0.6t}}$ for QTL genotype Qq and $\frac{30}{1+27e^{-0.9t}}$ for QTL genotype qq ($t = 1, \ldots, 11$). These two curves were chosen to mimic the estimated curves from the Poplar stem growth example in the previous section. The true logistic function is used in the parametric functional mapping procedure. In the second simulation scenario, we used the following biexponential functions with time-varying coefficient equations to model two genotypic mean vectors: $e^{13.5-0.35t} + e^{9.0-\alpha_1(t)t}$ and $e^{12.0-0.35t} + e^{8.0-\alpha_2(t)t}$, with $t = 0$, 2, 7, 10, 14, 21, 28, 56, 84, 115, 145, 175, 205, 235, 265, 295, 336, where $\alpha_1(t)$ and $\alpha_2(t)$ were assumed to change over time according to $\alpha_1(t_k) = 0.05 - \frac{0.06k}{17}$ and $\alpha_2(t_k) = 0.05 - \frac{0.055k}{17} (k = 1, \ldots, 17)$. These two mean vectors mimic the estimated long-term HIV dynamics in the AIDS Clinical Trial Group Protocol 315 data from Wu and Zhang (2002). Parametric functional mapping uses the existing regular biexponential form to estimate the underlying phenotypic curves.

The simulation data sets in each scenario were analyzed by parametric and nonparametric functional mapping. The cut-off values were determined by the proposed simulation procedure (1000 times) for nonparametric functional mapping and permutation tests (500 times) for parametric functional

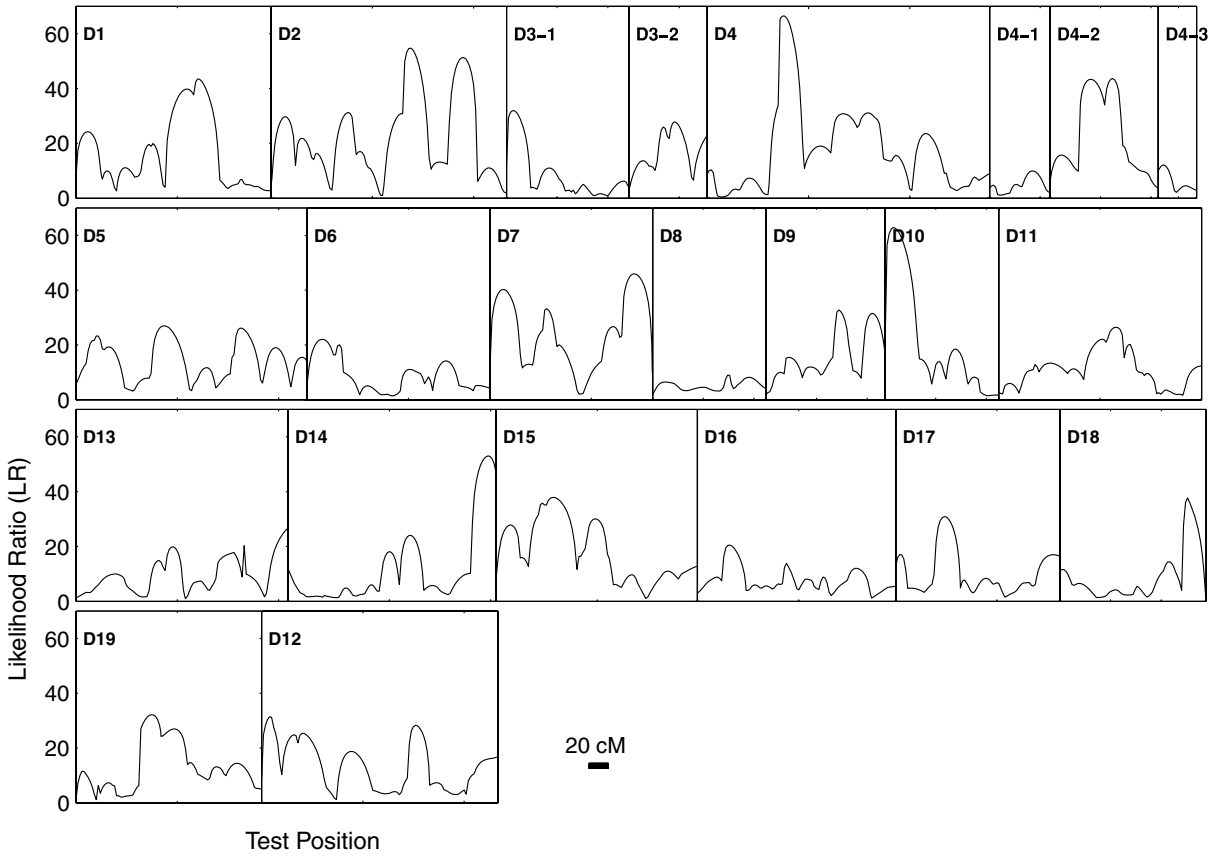

Figure 2. The profile of the LRT statistics between the full and reduced (no QTL) subject-specific model for the diameter growth trajectories across the whole *Populus deltoides* parent map. The genomic positions corresponding to the peak of the curve are the MLEs of the QTL localization.

mapping. The simulation and estimation procedure was repeated 100 times to estimate the power of QTL detection and the means of estimated genotypic curves for each approach. Both approaches obtained a similar power of QTL detection and similar estimates of QTL location under two different simulation scenarios (Table 1). As expected, more power and more precise estimates can be obtained for bigger sample sizes and heritabilities. The estimates of the genotypic mean curves were similar between the two approaches under the first simulation scenario, but were very different in the second. The nonparametric approach provided better estimates of genotypic curves than the parametric approach under the second scenario, even though the heritability and sample size were rather high. Figure 3 illustrates that the estimated genotypic mean curves deviate from the true mean curves dramatically with the parametric approach but they are well estimated by the nonparametric approach for sample size $N = 400$ and heritability $H^2 = 0.4$. Also, the

Table 1

Comparison between nonparametric functional mapping and parametric functional mapping. The first line in each cell is the p-value for existence of QTL while the second line is the location of the highest peak in the likelihood map, which is the QTL candidate position when there is evidence that a QTL exists. The true QTL was set at 88 *cM. The number of subjects is* 100 *for* $S1$ *and* 400 *for* $S2$*. "NPFM" stands for nonparametric functional mapping and "PFM" is parametric functional mapping. The symbol "-" means that all the p-values are the same, hence there is no variation.*

Mapping method	$S1$		$S2$	
	$H^2 = 0.1$	$H^2 = 0.4$	$H^2 = 0.1$	$H^2 = 0.4$
NPFM	0.0644 (0.0136) 88.02 (2.044)	<0.001 (-) 86.28 (0.3000)	<0.001 (-) 83.22 (0.2467)	<0.002 (-) 86.02 (0.0200)
PFM	0.0423 (0.0044) 88.06 (1.6000)	<0.002 (-) 86.28 (0.2594)	<0.001 (-) 84.68 (0.4322)	<0.002 (-) 86.02 (0.0200)

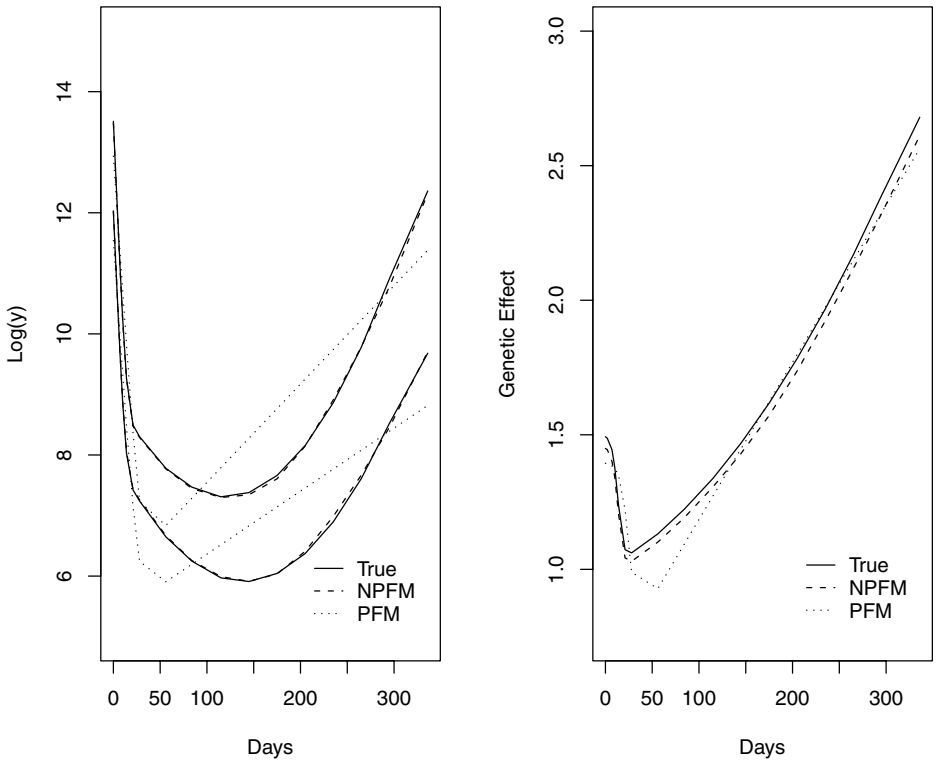

Figure 3. The estimated underlying phenotypic mean curves (left panel) and genetic effect along time (right panel) from true model (True), parametric functional mapping (PFM) and nonparametric functional mapping (NPFM), respectively. In this study, the number of subjects was 400 and the heritability value at the middle time point was set at 0.4.

nonparametric functional mapping, on average, took only about 5% of the computation time used by parametric functional mapping.

3.3.2 *Evaluating the REML variance estimate.* A second set of simulation studies was done to check the statistical behavior of using REML to estimate the unknown (co)variance matrix and then substituting the fixed (co)variance matrix in the EM algorithm for the MLE of the unknown coefficient vectors for the smoother matrix. We used the REML estimate of the (co)variance matrix from a saturating model, which is a consistent estimate of Σ only when the covariance structure is correctly specified. We also include the consistent empirical Bayes' (EB) estimate of Daniels and Kass (2001).

The first simulation study used the 61 subjects' marker information from linkage group 10 in the poplar data set. The underlying functions were two logistic growth curves: $\frac{20}{1+20e^{-0.6t}}$ and $\frac{30}{1+27e^{-0.9t}}$, where $t = 1, \ldots, 11$. Autoregressive correlation was assumed for any two observations. The covariance matrix was determined by letting the heritability on year 4 (the year with the largest genetic variance) equal (0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6), respectively. The heritability curves across all 11 years are shown in the left part of Figure 4. One hundred data sets for each heritability value were generated to perform the nonparametric functional mapping procedure for the general map. The average p-values are shown in the right part of Figure 4 corresponding to each heritability value. This figure tells us that EB performs better when heritability is bigger.

The second simulation study also used the 61 subjects' marker information from linkage group 10 in the poplar data set as genotypic data. The two underlying biological trajectories were from the HIV dynamics mechanism, which have double exponential forms $e^{12-0.7t} + e^{7.5-0.05t}$ and $e^{11-0.4t} + e^{5-0.03t}$ and the growth curves with logistic forms $\frac{20}{1+20e^{-0.6t}}$ and $\frac{30}{1+27e^{-0.9t}}$. Assume there were 20 observation points for the HIV curves and the covariance matrix was randomly generated without a known structure. There were 11 time points for the growth curves and the covariance matrix was set to be $\Sigma_3 = 0.3J_{11} + \text{Autoregressive}$ ($\tau^2 = 0.1$, $\rho = 0.8$), where J_{11} is a dimension 11 square matrix of all ones, and Autoregressive(τ^2, ρ) is the autoregressive covariance matrix of order one. One hundred data sets with 200 subjects were analyzed using nonparametric functional mapping for a general map. This analysis was also conducted for a sub data set containing 61 subjects randomly selected from each data set. The best structure picked by `SAS Proc Mixed` for the HIV data set was an autoregressive moving average structure, ARMA(1, 1) while for the growth data set the true dependence structure was selected. The results are in Table 2. From this table we can conclude that when the sample size increases, the p-value gets smaller and so does the standard deviation regardless of the covariance matrix estimate. If the true covariance matrix has some structure such as autoregressive, the REML estimate usually outperforms the EB estimate, as suggested by the results from the growth data set. If the true covariance matrix is actually unstructured, the EB

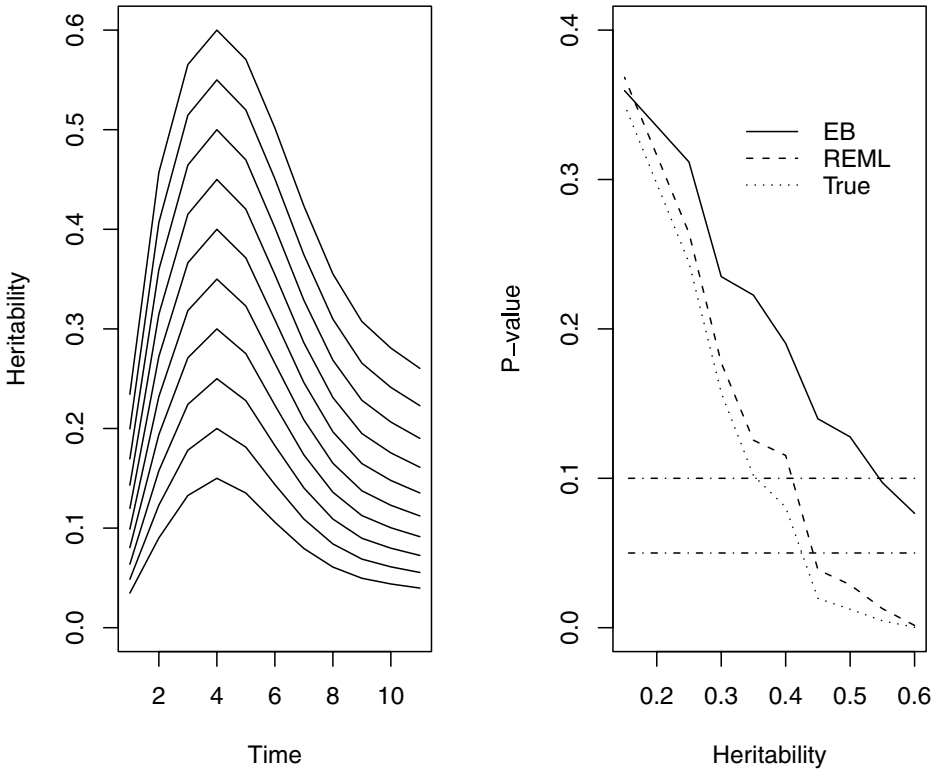

Figure 4. The left plot shows the 10 heritability curves of the simulated growth data sets. The right plot shows the trend of p-values when heritability on year 4 increases. "EB" and "REML" denote the empirical Bayes' estimate and REML estimate of covariance matrix, respectively. "True" represents the covariance matrix used to generate data.

Table 2

P-values and standard deviations of nonparametric functional interval mapping (NPFIM) from the HIV dynamics data and the growth data for different combinations of variance–covariance estimators and sample size. "EB" means the shrinkage estimator, which is guaranteed to be a consistent estimator. The "REML" estimator is obtained from `SAS Proc Mixed`, *assuming each subject has a different underlying mean curve. The "True" estimator is, of course, the matrix we actually used to generate the data.*

Variance estimate	HIV dynamics data		Growth data	
	$N = 61$	$N = 200$	$N = 61$	$N = 200$
EB	0.4018 (0.0300)	0.00126 (<0.0001)	0.384 (0.0279)	0.0239 (0.00787)
REML	0.1644 (0.0169)	0.03346 (0.0042)	0.209 (0.0238)	0.00105 (0.00064)
True	0.0376 (0.0167)	<0.0001(<0.0001)	0.197 (0.0240)	0.00095 (0.00058)

The numbers in parentheses are the sampling errors of the p-values.

estimate is better than the REML estimate. However, when analyzing a real data set, there is typically some pattern in the correlations among repeated measurements/longitudinal data. So the REML estimate is still recommended even though the EB estimate performs well when the sample size is large.

4. Discussion

As a direct extension of functional mapping (Ma et al., 2002), nonparametric functional mapping inherits significant advantages over other traditional mapping tools and models. For example, the results are closer to biological reality because of the simultaneous analysis of repeated measurements for a quantitative trait. By treating the process as a smooth curve, a small sample size could also achieve adequate power for QTL mapping because multiple measurements for each subject are analyzed simultaneously. Moreover, because of its nonparametric regression nature, this approach for general maps also has favorable computational advantages in model fitting over parametric functional mapping, especially when the parametric form is rather complicated and therefore a numerical optimization algorithm has to be used to search for MLEs of unknown parameters.

Our proposed nonparametric functional mapping with a dense map is essentially an association analysis with an exact multiple testing adjustment. It is a widely applicable strategy

as shown in our example of pharmacogenomic discovery using linkage-directed association studies with dense SNP markers. The multiple testing adjustment part for backcross populations involves computation of a matrix with dimension mL *(*$2mL$ *for* F_2 *or SNPs)*, the product of marker numbers m and the number of B-splines L, which may require computers with enough memory to generate $\mathbf{Z}$ from $MVN(0, \Delta)$ directly when the number of dense markers considered is huge. The key to avoid this problem is that the structure of the covariance matrix of $\mathbf{Z}$ allows one to simulate random variables with $MVN(0, \Sigma)$ and then get random variables with $MVN(0, \Delta)$ through multiplying by the matrix $\tilde{\mathbf{W}}$.

Instead of using B-splines, as in this article, many other basis functions can be used (e.g., penalized splines, regression splines, or wavelets). One advantage of using B-splines here is that we can use a simulation procedure to determine the critical value for nonparametric functional interval mapping because the smoother matrix is independent of observed phenotypic values. For smoothing with B-splines, the number of knots one should use and where to put them is an open question. In the poplar stem growth example, we used evenly distributed inner knots because the observation time points were all equidistant. We used about $\lceil T/2 \rceil$ inner knots to estimate the mean curves in those examples, and estimation of the underlying mean curves seems acceptable, as seen in the simulation. Because in reality the dynamic traits through time or other observation units are usually smooth, a small number of splines with order three may be enough for estimation purposes. From our experience, the results are not sensitive to different sets of basis functions or different numbers of knots used, given that the underlying curves are reasonably estimated.

Fundamental assumptions for our proposed method are normality of the errors and homoscedasticity of the analyzed phenotypes, and the question of robustness is natural. As in Coppieters et al. (1998), because the significance levels are deduced from a simulation procedure or phenotype permutation, our proposed method for a general map is relatively insensitive to the nonnormality of the residual variation. For the dense map case we calculate a family-wise error rate directly from the theoretical joint distribution of the LRT statistics. Based on simulation studies, we found that again our method is relatively insensitive to model misspecification because the distribution of p-values obtained from data with nonnormal residuals is not significantly different from those obtained from data with normal residuals. Simulation study details can be found in Web Appendix D.

Functional mapping can address a number of biologically meaningful questions (see Wu et al., 2004). Nonparametric functional mapping can also generate hypotheses to shed light on biological questions. For example, whether the detected QTL affects the rate of the change of longitudinal trajectories at a particular time point t_0 can be tested by formulating the hypotheses

$$H_0 : \frac{\partial \mu_1(t)}{\partial t} = \frac{\partial \mu_2(t)}{\partial t}\bigg|_{t=t_0} \text{ versus } H_1 : \frac{\partial \mu_1(t)}{\partial t} \neq \frac{\partial \mu_2(t)}{\partial t}\bigg|_{t=t_0}. \tag{7}$$

The time at which the rate of the change of the longitudinal trait is maximum can be estimated by restricting the derivatives of $m_1(t)$ and $m_2(t)$ to equal zero and solving for t_0. Therefore, our model can be used to test how the QTL detected controls the timing of maximum change rate in a time course.

When we illustrated our proposed method for a general map, we used a backcross design in both the simulation studies and a real data example for clarity of description. However, our method for a general map can easily be extended for application in more complex designs, such as an F_2 or full-sib family. For example, for an F_2 population with a general map, the likelihood function is a mixture of three multivariate normal density functions instead of a mixture of two (as in a backcross population). We used SNPs to demonstrate our method for a dense map, and this can be easily extended to multiallelic dense marker cases (details are in Web Appendix B). Also, an extension to model the association or interaction of two or more QTL can make this nonparametric functional mapping methodology more powerful.

5. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2.1, 2.3.1, 2.3.2, 3.1, and 4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

We thank Dr Howard L. McLeod for providing us the 5-fluorouracil cytotoxicity data and detailed explanation of data set description. We appreciate the valuable comments made by the editor, associate editor, and referee, which have greatly improved the contents and the presentation of this article. This study was supported by National Science Foundation Grants DMS-04-05543, DMS-0631632, and SES-0631588

.

References

Anholt, R. R. and Mackay, T. F. C. (2004). Quantitative genetic analysis of complex behaviors in Drosophila. *Nature Review: Genetics* **5,** 838–849.

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–971.

Coppieters, W., Kvasz, A., Farnir, F., Arranz, J., Grisart, B., Mackinnon, M., and Georges, M. (1998). A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: Application to milk production in a granddaughter design. *Genetics* **149,** 1547–1555.

Daniels, M. J. and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57,** 1173–1184.

Dempster, A. P., Laird, N .M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Doerge, R. W. and Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142,** 285–294.

Doerge, R. W. and Rebaï, A. (1996). Significance thresholds for QTL interval mapping tests. *Heredity* **76,** 459–464.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11,** 89–121.

He, X. M. and Shi, P. D. (1998). Monotone B-spline smoothing. *Journal of the American Statistical Association* **93,** 643–650.

Jansen, R. C. and Stam, P. (1994). High resolution mapping of quantitative traits into multiple loci via interval mapping. *Genetics* **136,** 1447–1455.

Kao, C.-H. and Zeng, Z.-B. (2002). Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* **160,** 1243–1261.

Lander, E. S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121,** 185–199.

Lander, E. S. and Botstein, D. (1994). Corrigendum. *Genetics* **136,** 705.

Lin, M. and Wu, R. L. (2006). A joint model for nonparametric functional mapping of longitudinal trajectories and time-to-events. *BMC Bioinformatics* **7,** 138.

Littell, R. C., Pendergast, J., and Natarajan R. (2000). Tutorial in biostatistics: Modeling covariance structure of the analysis of repeated measures data. *Statistics in Medicine* **19,** 1793–1819.

Ma, C., Casella, G., and Wu, R. L. (2002). Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics* **161,** 1751–1762.

Meyer, K. (2005a). Estimates of genetic covariance functions for growth of Angus cattle. *Journal of Animal Breeding Genetics* **122,** 73–85.

Meyer, K. (2005b). Random regression analyses using B-splines to model growth of Australian Angus cattle. *Genetics Selection Evolution* **37,** 473–500.

Niklas, K. L. (1994). *Plant Allometry: The Scaling of Form and Process.* Chicago: University of Chicago.

Paterson, A. H. (2006). Leafing through the genomes of our major crop plants: Strategies for capturing unique information. *Nature Reviews: Genetics* **7,** 174–184.

Piepho, H. P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics* **157,** 425–432.

Pittman, J. (2002). Adaptive splines and genetic algorithms. *Journal of Computational and Graphical Statistics* **11,** 615–638.

Rebaï, A., Goffinet, B., and Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics* **153,** 235–240.

Sen, S. and Churchill, G. A. (2001). A statistical framework for quantitative trait mapping. *Genetics* **159,** 371–387.

Von Bertalanffy, L. (1957). Quantitative laws in metabolism and growth. *Quarterly Review of Biology* **32,** 217–231.

Watters, J. W., Kraja, A., Meucci, M. A., Province, M. A., and McLeod, H. L. (2004). Genome-wide discovery of loci influencing chemotherapy cytotoxicity. *The Proceedings of National Academy of Sciences* **101,** 11809–11814.

Weiss, L. A., Abney, M., Cook, E. H., and Ober, C. (2005). Sex-specific genetic architecture of whole blood serotonin levels. *American Journal of Human Genetics* **76,** 33–41.

Weiss, L. A., Pan, L., Abney, M., and Ober, C. (2006). The sex-specific genetic architecture of quantitative traits in humans. *Nature Genetics* **38,** 218–222.

Wu, H. and Zhang, J. T. (2002). The study of long-term HIV dynamics using semiparametric nonlinear mixed-effects models. *Statistics in Medicine* **21,** 3655–3675.

Wu, R. L., Ma, C.-X., Lin, M., Wang, Z. H., and Casella, G. (2004). Functional mapping of growth QTL using a transform-both-sides logistic model. *Biometrics* **60,** 729–738.

Wu, R. L., Ma, C.-X., and Casella, G. (2007). *Statistical Genetics of Quantitative Traits: Linkage, Maps and QTL.* New York: Springer-Verlag.

Yang, J. (2006). Nonparametric functional mapping of quantitative trait loci. Ph.D. Thesis, Department of Statistics, University of Florida, Gainesville, Florida.

Yang, R. Q. and Xu, S. Z. (2007). Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. *Genetics* **2007**, 1169–1185.

Yang, R. Q., Tian, Q., and Xu, S. Z. (2006). Mapping quantitative trait loci for longitudinal traits in line crosses. *Genetics* **173,** 2339–2356.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136,** 1457–1468.

Received September 2007. *Revised February* 2008.
Accepted February 2008.