University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 151

Semiparametric Quantitative-Trait-Locus Mapping: I. on Functional Growth Curves

Ying Qing Chen* Rongling Wu†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, yqchen@stat.berkeley.edu

†Dept. of Statistics, Institute of Food & Agriculture Sciences, University of Florida, Gainesville, FL, rwu@hes.hmc.psu.edu

Semiparametric Quantitative-Trait-Locus Mapping: I. on Functional Growth Curves

Ying Qing Chen and Rongling Wu

Abstract

The genetic study of certain quantitative traits in growth curves as a function of time has recently been of major scientific interest to explore the developmental evolution processes of biological subjects. Various parametric approaches in the statistical literature have been proposed to study the quantitative-trait-loci (QTL) mapping of the growth curves as multivariate outcomes. In this article, we view the growth curves as functional quantitative traits and propose some semiparametric models to relax the strong parametric assumptions which may not be always practical in reality. Appropriate inference procedures are developed to estimate the parameters of interest which characterise the possible QTLs of the growth curves in the models. Recently developed multiple comparison testing procedures are applied to locate the statistically meaningful QTLs. Numerical examples are presented with simulation studies and analysis of real data.

1 Introduction

To study the developmental process of biological subjects, the repeated measurements of certain quantitative trait are often collected over time. For instance, in Kenward (1987), the repeated measures on the weight of cattle are collected at ten two-week intervals till the final measurement at one week after the tenth interval. These repeated measurements are often considered as some underlying random functional growth curves observed at a set of the point processes over time. Similar to the single measurement of quantitative trait, the entire growth curves as functional-valued traits may be inherently controlled by genetic factors (Kirkpatrick and Heckman, 1989; Pletcher and Geyer, 1999).

There has been growing interest in scientific research to map the quantitative-trait-loci (QTL) for the functional-valued growth curves, $\{Y(t); t \geq 0\}$, say, to study the potential genetic association with the developmental processes. To analyse these growth curves, the heterogeneity of the repeated measurements of the same subject has to be considered. That is, the repeated measurements of one subject are usually considered to be correlated with each other. Parametric approaches have been proposed and studied in literature (Ma, et al., 2002; Wu, et al., 2002).

In general, these approaches impose certain parametric assumptions on both the mean and the covariance structures of the functional quantitative traits. For example, the models may assume the mean of the growth curves, $\mu(t)$, say, are of special shapes, such as the sigmoidal logistic functions:

$$\mu(t) = EY(t) = \frac{\alpha_1}{1 + \alpha_2 \exp(-\alpha_3 t)},$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$ are the parameters governing the logistic functions, as in Wu, et al. (2002). Here, ^{T} denotes the transpose of a vector or matrix. If different parameter estimates are obtained from the observed data for different genotypes at a putative QTL, then this QTL is identified as a potential genetic location to moderate the growth curves. For the covariance structure, some error distributions, such as the zero-mean Gaussian or autoregressive processes, have to be assumed as well. All these assumptions enable the usual

maximum-likelihood-based approaches, such as the EM-algorithms, to be straightforwardly implemented, although the inferences on the estimates are often lacked.

There are, however, challenges embedded with these parametric approaches. These parametric assumptions determine the ultimate estimates, and different selections of the parametric functions may thus lead to different conclusions on the QTL's. For example, instead of using the logistic functions to model the mean structure, there are alternative shapes such as exponential or saturating functions (Niklas, 1994). In some situations, there is even no prior information on a clear choice of $\mu(t)$. For another example, although the usual Gaussian and the autoregressive processes yield explicit likelihood functions to be maximized, their stationary assumptions may not be appropriate for the growth curves.

To avoid such difficulties with the stringent parametric assumptions, one approach is by way of the semiparametric modeling. For example, the semiparametric model proposed by Zeger and Diggle (1994) may be used to model $Y(t)$:

$$Y(t) = \mu_0(t) + \beta^{\mathrm{T}} Z(t) + \varepsilon(t), \tag{1}$$

where $\mu_0(\cdot)$ is some unknown baseline function, $Z(t)$ be the $p-$vector covariates of potential genotypes and other phenotypic variables, and $\beta \in \mathcal{B} \subset \mathbf{R}^p$ are the $p-$vector parameters. Here, $\varepsilon(\cdot)$ are assumed to be the zero-mean Gaussian processes. The model (1) generalizes the linear regression models with time-specific intercepts at each distinct time $t > 0$ to the ones with the baseline function of $\mu_0(\cdot)$ in continuous time. With the unknown $\mu_0(\cdot)$, it gains more flexibility to model the mean functional trait. In addition, the covariates $Z(t)$ are not limited to the subject's genotypes but include all possibly observed phenotypic or environmental factors that may potentially confound the genetic association. In fact, a more general semiparametric model by Lin and Ying (2001) is proposed to model the marginal mean of $\mu(\cdot)$ only:

$$E\{Y(t) \mid Z(s), 0 \leq s \leq t\} = \mu_0(t) + \beta^{\mathrm{T}} Z(t). \tag{2}$$

In addition to maintaining the flexibility in the mean structure, this model allows the covariance structure and the distributional form of errors to be unspecified, and can be easily extended to the functional qualitative traits of binary or categorical types.

In this article, we will use some semiparametric mean response models in the QTL analysis of the functional growth curves and further develop appropriate inference procedures in identifying the potential QTLs. The new methodologies may overcome the disadvantage of the current parametric models by allowing more flexibility in both mean and covariance structure, and hence the final estimates may be more robust to the potential model misspecification. Some recently developed multiple comparison procedures will be adapted to evaluate the parameter estimates for detecting the statistically meaningful QTLs. In the rest of the article, the methods will be presented in §2. In §3, the method will be applied to the QTL mapping of the diameters of the forest trees. Some issues related to the proposed methodologies are discussed in §4. Technical proofs are collected in the Appendix.

2 Methods

2.1 Genetic design

In this article, we use a standard backcross design to illustrate the proposed statistical methodologies. As discussed later, the methodologies can be further extended to more complex designs, such as an F_2 or full-sib design. The backcross design is initiated with two contrasting homozygous inbred lines. There are assumed two genotypes at a specific locus on the genome. A marker-based genetic linkage map is constructed and aims to the QTL identification affecting the time-dependent functional trait. Suppose there are n progeny subjects in the data set. The functional traits are considered as the random curves denoted by $\{Y_i(t); i = 1, 2, \ldots, n, t \geq 0\}$. In reality, the whole curves are usually not observed at every single t, but the repeated measurements are observed for the ith subject, $(Y_{i1}, Y_{i2}, \ldots, Y_{im_i})$, $i = 1, 2, \ldots, n$, say. They can be considered as the random growth curve observed at a set of time points of $(T_{i1}, T_{i2}, \ldots, T_{i,m_i})$, i.e., $Y_{ij} = Y_i(T_{ij})$, $j = 1, 2, \ldots, m_i$.

For a pleiotropic QTL that affects the functional trait $Y(t)$, it is assumed to be bracketed by two flanking genetic markers, P_l and P_{l+1}, with two genotypes of $(A_l a_l, A_l a_l)$ and $(A_{l+1} a_{l+1}, A_{l+1} a_{l+1})$ at each marker, respectively, $l = 1, 2, \ldots, L-1$, with L being the total number of markers on the genome. Therefore, there are four combinations of possible geno-

types for the flanking markers of the progeny at P_l and P_{l+1}: $(A_l a_l, A_{l+1} a_{l+1})$, $(A_l a_l, a_{l+1} a_{l+1})$, $(a_l a_l, A_{l+1} a_{l+1})$ or $(a_l a_l, a_{l+1} a_{l+1})$, which are denoted as $M_{l,i} = 1, 2, 3$ or 4, respectively. Suppose that there are two possible alleles for the genotype that determines the functional trait, Q and q. They segregate in the backcross population with two different genotypes of Qq and qq of complete penetrance in phenotypes. Denote G_i the genotype indicator of 1 being Qq and 0 being qq. Additional observed phenotypic covariates and environmental factors are denoted as $(p-2)-$vector $R_i(t)$. Let $Z_{l,i}(t) = (M_{l,i}, G_i, R_i(t))^{\mathrm{T}}$.

2.2 Statistical models

In a well-planned study, the observed data of $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{i,m_i})^{\mathrm{T}}$'s are often collected at a pre-determined set of observation times of $(T_{i1}, T_{i2}, \ldots, T_{i,m_i})$. In reality, however, the observations times may be irregular or even random. Consider $N_i(t) = \sum_{j=1}^{m_i} I(T_{ij} \leq t)$, which is the counting process for the ith subject's observation times. Here $I(\cdot)$ is the indicator function. Similar to the assumptions in Cheng and Wei (2000), $E\{N_i(t)\} = \Omega(t)$, where $\Omega(\cdot)$ is unspecified. In addition, there is usually a follow-up time or censoring time for $Y_i(\cdot)$, C_i, say. The C_i's are assumed to be noninformative such that

$$E\{Y_i(t) \mid Z_i(s), 0 \leq s \leq t, C_i\} = E\{Y_i(t) \mid Z_i(s), 0 \leq s \leq t\}.$$

Denote $\Delta_i(t) = I(C_i \geq t)$ as the "at-risk" indicator.

To model the functional trait $Y_i(\cdot)$, we consider the following models to relate $Y_i(t)$ and its covariates $Z_i(t)$:

$$E\{Y_i(t) \mid Z_i(s), 0 \leq s \leq t\} = \mu_0(t) + \beta_{\mathrm{G}} G_i + \beta_{\mathrm{R}}^{\mathrm{T}} R_i(t), \tag{3}$$

where $\mu_0(\cdot)$ is unspecified smooth function and $\beta = (\beta_{\mathrm{G}}, \beta_{\mathrm{R}}^{\mathrm{T}})^{\mathrm{T}}$ are parameters. This model assumes that the genetic effect at any putative QTL is additive, regardless of the flanking markers. This assumption can be relaxed to include the scenarios when the marker loci can be QTL as well, by introducing the interaction terms between G_i and $M_{l,i}$. In fact, when $R_i(\cdot)$ are not included in the model, $\mu(\cdot)$ itself becomes the mean function for $G_i = 0$, i.e., the mean function for the genotype qq. In this case, the magnitude of β_{G} characterises the

genetic effect on the functional trait due to the different genotypes of Qq and qq. If $\beta_{\mathrm{G}} > 0$, that means the genotype Qq is associated with a positive change in the functional trait from the genotype of qq, while otherwise it is associated with a negative change. Thus, β_{G} can be used to make inference on the QTL, if the magnitude of its estimate is unusually large. When the $R_i(t)$ are included in the model, the parameter β_{G} measures the genetic effect adjusted for the potentially heterogeneous environmental factors or the observed phenotypes other than the functional trait of interest.

To model the genetic effect as multiplicative, the functional growth curves can be first log-transformed, and then the same additive mean structure is applied:

$$E\{\log Y_i(t) \mid Z_i(s), 0 \leq s \leq t\} = \mu_0(t) + \beta_{\mathrm{G}} G_i + \beta_{\mathrm{R}}^{\mathrm{T}} R_i(t). \tag{4}$$

In model (4), $\exp(\beta_G)$ thus characterises the multiplicative effect due to the different genotypes at the QTL. These models, similar to the one by Lin and Ying (2001), only model the mean structure of the functional trait, while leaving the dependence structure completely unspecified. Its semiparametric feature of the unspecified baseline function would allow the models to embrace much broader classes of functions with different shapes.

In reality, the exact genotype of a progeny subject, G_i, is usually unknown. Its probability distribution, however, depends on the two-locus genotype of the flanking markers and the QTL position in the marker interval. Assume that the recombination fractions between the marker P_l and the potential QTL, the potential QTL and the marker P_{l+1} and the markers P_l and P_{l+1} are r_{l1}, r_{l2} and r_l, respectively. Then some straightforward calculations show that the joint probability distribution is determined for all the potential genotypes listed as in Table 4. Furthermore conditional on the genotypes of the bracketing markers, the probability distribution of G_i is listed in Table 2 as well.

Table 1: *Joint and conditional probabilities of genotype indicator G at a QTL bracketed by markers P_l and P_{l+1} in a backcross population. When r_{l1} or r_{l2} is relatively small, $r_{l1} + r_{l2}$ approximates r_l.*

Marker type	Marker genotypes		Joint probabilities		Marginal	Conditional probabilities	
M_l	P_l	P_{l+1}	$p_l^{\mathrm{G}} = \mathrm{pr}\{G=1\}$	$1-p_l^{\mathrm{G}} = \mathrm{pr}\{G=0\}$	probabilities	$p_l^{\mathrm{G}} = \mathrm{pr}\{G=1 \mid M_l\}$	$1-p_l^{\mathrm{G}} = \mathrm{pr}\{G=0 \mid M_l\}$
1	$A_l a_l$	$A_{l+1} a_{l+1}$	$\frac{1}{2}(1-r_{l1})(1-r_{l2})$	$\frac{1}{2} r_{l1} r_{l2}$	$\frac{1}{2}(1-r_l)$	$(1-r_{l1})(1-r_{l2})/(1-r_l)$	$r_{l1} r_{l2}/(1-r_l)$
2	$A_l a_l$	$a_{l+1} a_{l+1}$	$\frac{1}{2}(1-r_{l1}) r_{l2}$	$\frac{1}{2} r_{l1}(1-r_{l2})$	$\frac{1}{2} r_l$	$(1-r_{l1}) r_{l2}/r_l$	$r_{l1}(1-r_{l2})/r_l$
3	$a_l a_l$	$A_{l+1} a_{l+1}$	$\frac{1}{2} r_{l1}(1-r_{l2})$	$\frac{1}{2}(1-r_{l1}) r_{l2}$	$\frac{1}{2} r_l$	$r_{l1}(1-r_{l2})/r_l$	$(1-r_{l1}) r_{l2}/r_l$
4	$a_l a_l$	$a_{l+1} a_{l+1}$	$\frac{1}{2} r_{l1} r_{l2}$	$\frac{1}{2}(1-r_{l1})(1-r_{l2})$	$\frac{1}{2}(1-r_l)$	$r_{l1} r_{l2}/(1-r_l)$	$(1-r_{l1})(1-r_{l2})/(1-r_l)$

Table 2: *Conditional probabilities of genotype indicator G at a QTL bracketed by markers P_l and P_{l+1} in a backcross population, when $\rho_l = r_{l1}/r_l$ is relatively small.*

Marker type	Marker genotypes		Conditional probabilities	
M_l	P_l	P_{l+1}	$p_l^{\mathrm{G}} = \mathrm{pr}\{G=1\}$	$1-p_l^{\mathrm{G}} = \mathrm{pr}\{G=0\}$
1	$A_l a_l$	$A_{l+1} a_{l+1}$	1	0
2	$A_l a_l$	$a_{l+1} a_{l+1}$	$1-\rho_l$	ρ_l
3	$a_l a_l$	$A_{l+1} a_{l+1}$	ρ_l	$1-\rho_l$
4	$a_l a_l$	$a_{l+1} a_{l+1}$	0	1

2.3 Estimation procedures

As discussed previously, the exact genotypes of G is not known and neither does its location on the genome. However, given the configuration of the flanking markers, the marginalised model of (3) over G_i is,

$$E\{Y_i(t) \mid M_{l,i}, R_i(s), 0 \leq s \leq t\} = \mu_0(t) + \beta_{\mathrm{G}} p_{l,i}^{\mathrm{G}} + \beta_{\mathrm{R}}^{\mathrm{T}} R_i(t).$$

Let $\nu_i(t) = \beta_{\mathrm{G}} p_{l,i}^{\mathrm{G}} + \beta_{\mathrm{R}}^{\mathrm{T}} R_i(t)$. Furthermore, denote the true parameters in the aforementioned models as their respective counterparts with the subscript "$_*$." For instance, the true parameters of β_{G} and β_{R} are $\beta_{\mathrm{G}*}$ and $\beta_{\mathrm{R}*}$, respectively. Let $X_i(t) = \int_0^t [Y_i(s) - \nu_i(s)\}] dN_i(s)$, then

$$E\{dX_i(t) \mid M_{l,i}, R_i(s), 0 \leq s \leq t, C_i\} = \Delta_i(t) d\Omega_\mu(t),$$

where $d\Omega_\mu(t) = \mu_{0*}(t) d\Omega(t)$ and $\Omega(t) = E\{N_i(t)\}$ as defined previously. Let $M_i(t) = X_i(t) - \int_0^t \Delta_i(s) d\Omega_\mu(s)$. Then $M_i(\cdot; \beta_*)$ are the zero-mean stochastic processes. As pointed out in Lin and Ying (2001), the following estimating equations generalise the normal equations of the least-squares in the linear regression models, and can be used to estimate the parameters in the proposed model (3),

$$\sum_{i=1}^{n} \int_0^{\tau} \Delta_i(t) \Psi(t) \varphi_i(t) dM_i(t) = 0, \tag{5}$$

where $\Psi(\cdot)$ is the positive weight function which converges uniformly to a deterministic function $\psi(t) \in [0, \tau]$, τ is some upper limit of the observation times, and $\varphi_i(t)$ are the smooth functions of the same dimensions as β such that $\varphi_i(t)$ are measurable with respect to $\{Z_i(s), C_i; 0 \leq s \leq t, i = 1, 2, \ldots, n\}$. For instance, $\varphi_i(\cdot)$ can be chosen as $Z_i(\cdot)$ and some of its nonlinearly related functionals.

In addition to the unknown parameters of β in (5), the infinite-dimensional function of $\Omega_\mu(\cdot)$ is also unknown. An estimator of the Breslow-type, however, can be obtained for $\Omega_\mu(\cdot)$,

$$\widehat{\Omega}_\mu(t) = \int_0^t \frac{\sum_{i=1}^n dX_i(s)}{\sum_{i=1}^n \Delta_i(s)},$$

which is unbiased to $\Omega_\mu(t)$. Let $\widehat{M}_i(t) = X_i(t) - \int_0^t \Delta_i(s) d\widehat{\Omega}_\mu(s)$. Replace the $M_i(\cdot)$'s in (5) and thus result in $\sum_{i=1}^n \int_0^\tau \Delta_i(t) \Psi(t) \varphi_i(t) d\widehat{M}_i(t) = 0$. Straightforward algebra further leads

to

$$\mathcal{E}(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \Delta_i(t)\Psi(t)\left\{\varphi_i(t) - \bar{\varphi}(t)\right\} dX_i(t) = 0, \tag{6}$$

where $\bar{\varphi}(t) = \sum_{i=1}^{n} \Delta_i(t)\varphi_i(t) / \sum_{i=1}^{n} \Delta_i(t)$. Assume that $\widehat{\beta}$ is the solution in (6).

Let $\nu_i'(t)$ be the derivative of $\nu_i(t)$, $i = 1, 2, \ldots, n$. Then $-n^{-1}\mathcal{E}'(\beta_*)$ goes to

$$B = E\left[\int_0^{\tau} \Delta_1(t)\psi(t)\{\varphi_1(t) - \bar{\varphi}_*(t\}\nu_1'(t)^{\mathrm{T}} d\Omega(t)\right],$$

where $\bar{\varphi}_*(t)$ is the limit of $\bar{\varphi}(t)$ almost surely, as $n \to \infty$. Since the elements in $\varphi_i(\cdot)$ are not linearly related, B is nonsingular. Thus under mild conditions, the solutions to $\mathcal{E}(\beta) = 0$ are strongly consistent as $n \to \infty$ as shown in the Appendix. If the total variation of $\varphi_i(\cdot)$, $i = 1, 2, \ldots, n$, are bounded, it is true that

$$n^{-1/2}\mathcal{E}(\beta_*) \doteq n^{-1/2} \sum_{i=1}^{n} \int_0^{\tau} \Delta_i(t)\psi(t)\{\varphi_i(t) - \bar{\varphi}_*(t)\} dM_i(t; \beta_*).$$

By the Central Limit Theorem, it is shown in the Appendix that $n^{-1/2}\mathcal{E}(\beta_*)$ is asymptotically normal with mean zero and the variance-covariance matrix,

$$\Sigma = E\left[\int_0^{\tau} \Delta_1(t)\psi(t)\{\varphi_1(t) - \bar{\varphi}_*(t)\} dM_1(t)\right]^{\otimes 2},$$

where $a^{\otimes 2}$ denotes aa^{T}. In addition, a Taylor's expansion of $\mathcal{E}(\widehat{\beta})$ at β_* yields that $n^{1/2}(\widehat{\beta} - \beta_*)$ is asymptotically equivalent to $\{-\mathcal{E}'(\beta_*)/n\}^{-1} \cdot n^{-1/2}\mathcal{E}(\beta_*)$. As a result of the Appendix, it is true that $\widehat{\beta}$ are consistent, and

$$n^{1/2}(\widehat{\beta} - \beta_*) \to N(0, B^{-1}\Sigma B^{-1})$$

in distribution in a neighbourhood of (β_*), where B and Σ can be approximated by their empirical counterparts,

$$\begin{aligned} \widehat{B} &= n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \Delta_i(t)\Psi(t)\{\varphi_i(t) - \bar{\varphi}(t)\}\nu_i'(t)^{\mathrm{T}} dN_i(t), \text{ and} \\ \widehat{\Sigma} &= n^{-1} \sum_{i=1}^{n} \left[\int_0^{\tau} \Delta_i(t)\Psi(t)\{\varphi_i(t) - \bar{\varphi}(t; \widehat{\beta})\} d\widehat{M}_i(t; \widehat{\beta})\right]^{\otimes 2}, \end{aligned}$$

respectively.

The estimating equations used in the weighted estimating equations of (6) are somewhat *ad hoc*, although the estimators defined in the equations carry the appealing statistical properties such as consistency and asymptotic normality. It is desirable to choose an optimal weight function to minimize the variance among the estimators. With an application of Cauchy-Schwarz inequality, it is straightforward to see that such choice is $1/\text{var}\{Y(t)-\nu(t)\}$, which is the essentially the diagonal elements in the variance-covariance matrix of $Y(\cdot)$. Hence the optimal choice of $\psi(\cdot)$ would improve the efficiency. In addition, as pointed out in Wang and Wang (2001), the efficiency should be further improved if the weight function can be selected among the bivariate functions of $\Phi(s,t)$ to account for the covariance of $(Y(s), Y(t))$ for different $s > 0$ and $t > 0$.

To estimate the baseline $\mu(\cdot)$, it is natural to consider the estimator of

$$\widetilde{\mu}(t) = \bar{Y}(t) - \bar{\nu}(t; \widehat{\beta}),$$

where $\bar{Y}(t) = \sum_{i=1}^n \Delta_i(t) Y_i(t) / \sum_{i=1}^n \Delta_i(t)$ and $\bar{\nu}(t;\beta) = \sum_{i=1}^n \Delta_i(t)\nu_i(t;\beta)/\sum_{i=1}^n \Delta_i(t)$, respectively. This is the pointwise average of $Y_i(t) - \nu_i(t)$ when $\Delta_i(t) = 1$, i.e., the subjects are still "at risk." When the observation times are observed in a continuous time scale, some smoothing technique has to be implemented to obtain a reasonable estimate. In Lin and Ying (2001), a simple singleton nearest neighbour smoother was used. This approach may not be the most efficient. But it has advantage "in non-linear, non-Gaussian situations" without constructing explicit smoothers (Rice, 2003). To improve efficiency, however, more sophisticated smoothing techniques such as the one by Capra and Müller (1997) can be can be adapted to estimate $\mu(\cdot)$. Specifically, consider the time interval $[0, \tau]$ is partitioned into L consecutive equidistant intervals: (t_{l-1}, t_l), with $l = 1, 2 \ldots, L \to \infty$ and $t_0 = 0$. Assume the smoothing parameter h such that $h \to 0$ and $n_* h \to 0$, as $n_* \to 0$, where n_* is the total number of observation time points. Then a smoothed estimate of $\widetilde{\mu}(\cdot)$ is

$$\widehat{\mu}(t) = \arg\min_{a_0, a_1} \left[\sum_{l=1}^{L} K\left(\frac{t - t_l}{h}\right) \{\widetilde{\mu}(t_l) - a_0 - a_1(t_l - t)\}^2 \right].$$

Here $K(s) = 1 - s^2$, if $|s| \leq 1$, and 0 otherwise. Other smoothers including higher-order kernel smoothers or local fitting with high-order polynomials can be also used under the necessary

conditions of linearity, consistency and consistency with needed rate in Capra and Müller (1997).

2.4 Multiple comparison procedures in QTL detection

The regression models and their estimation are proposed mainly to evaluate the association between the genotypes and the functional quantitative trait at a putative locus bracketed by one specific pair of markers. To detect the QTLs, the following null hypotheses would be used: $\mathrm{H}_{l,0} : \beta_{\mathrm{G}} = 0$, for $l = 1, 2, \ldots, L$. Specifically for the lth pair of markers, two statistics can be used: one is the difference statistic of $D_{n,l} = n^{1/2}(\widehat{\beta}_{\mathrm{G},l} - 0)$, and the other is its standardized version of $T_{n,l} = n^{1/2}(\widehat{\beta}_{\mathrm{G},l} - 0)/\sigma_{l,n}$, where $\sigma_{n,l}/\sqrt{n}$ is the estimated standard error of $\widehat{\beta}_{\mathrm{G},l}$. When the testing procedure is repeated at every 1 or 2 cM on a map bracketed by two consecutive markers throughout the entire linkage map, L multiple-comparison procedures are thus conducted.

A common approach to identify the amount of support for a QTL at a particular map position is often by graphically displaying the likelihood ratio test statistics as a function of the map position of a putative QTL (Lander and Bostein, 1989). However, given the semiparametric framework of our models, the underlying distributional form of the errors are usually not assumed, and it is thus almost impossible to obtain the usual likelihood maps or profiles to construct the linkage map. In fact, when a large number of hypothesis testing are performed, the rate of false QTL claims usually needs to be controlled. Conventional approaches, such as the ones discussed in Hochberg and Tamhane (1987), are mainly aimed to controlling the so-called family-wise error rate (FWER), i.e., the probability of at least one false QTL claim when there is no QTL bracketed by any pair of markers in the entire linkage map. When certain proportion of markers to be tested actually depart from their corresponding null hypotheses, these procedures are often conservative and less powerful, as discussed extensively in literature. An important alternative has been developed to focus on the control of the so-called false discovery rate (FDR), which is the expected false positive rate of the rejected hypotheses, since the work by Benjamini and Hochberg (1995). There are both Frequentist and Bayesian FDR-based approaches. Yet most of them rely on the

Table 3: *Error types in QTL multiple comparisons*

	QTL not claimed	QTL claimed	Total hypotheses tested
No QTL existed	U	V	L_0
QTL existed	T	S	$L - L_0$
Total claims	$L - R$	R	L

assumptions of the independence among the test statistics, although certain specific form of dependence may be allowed.

In the QTL detection, the independence assumption does not always hold, given the same set of observations of the functional quantitative traits being repeatedly used in the semiparametric models. In this section, we adapt the framework recently constructed by Pollard and van der Laan (2003), Dudoit, et al. (2003), and van der Laan, et al. (2003a, 2003b) to the test statistics on the QTL parameter. In this framework, two kinds of Type I error rate, θ_n, are considered: the generalized family-wise error rate (gFWER) and the proportion of false QTL claims of the rejected hypotheses (PFP). A $gFWER(k)$ is the probability of allowing at least k false claims for some $k+1 \geq 0$, while a $PFP(\kappa)$ is the probability of false claims larger than some κ in (0,1) among the total rejections. Consider the notations used in Benjamini and Hochberg (1995), as seen in Table 3. Then the $gFWER(k)$ and $PFP(\kappa)$ are actually $\text{pr}\{V \geq k+1\}$ and $\text{pr}\{V/R > \kappa\}$, respectively. Compared with the definitions of the FWER and FDR, it is not difficult to find that

$$FWER = gFWER(0), \text{ and } FDR = E(V/R) = \int_0^1 PFP(\kappa)d\kappa,$$

respectively. For a prespecified α-value, it is said to be of finite sample control if $\theta_n \leq \alpha$, whereas it is of asymptotic control if $\overline{\lim}_{n\to\infty} \theta_n \leq \alpha$. Usually α is chosen to be 0.05.

Let $D_n = (D_{n,1}, D_{n,2}, \ldots, D_{n,L})^{\mathrm{T}}$ and $T_n = (T_{n,1}, T_{n,2}, \ldots, T_{n,L})^{\mathrm{T}}$, respectively. Assume that P is the underlying data generating distribution. Denote $Q_{n,\mathrm{D}}(P)$ and $Q_{n,\mathrm{T}}(P)$ the joint distributions of D_n and T_n with limiting distributions of $Q_{\mathrm{D}}(P)$ and $Q_{\mathrm{T}}(P)$, respectively. Then the distributions of V is determined by the corresponding $Q_{n,\mathrm{D}}(P)$ and $Q_{n,\mathrm{T}}(P)$. Since P is usually unknown, it needs to be estimated to ensure appropriate control of $gFWER(k)$ and $PFP(\kappa)$ in the QTL detection under the null distributions of $Q_{0,\mathrm{D}}(P)$ and $Q_{0,\mathrm{T}}(P)$,

respectively. Since

$$D_{n,l} = n^{1/2}(\widehat{\beta}_{\mathrm{G},l} - \beta_{\mathrm{G}*,l}) + n^{1/2}\beta_{l,\mathrm{G}} = D^*_{n,l} + n^{1/2}\beta_{\mathrm{G}*,l}, \text{ and}$$
$$T_{n,l} = \frac{n^{1/2}(\widehat{\beta}_{\mathrm{G},l} - \beta_{\mathrm{G}*,l})}{\sigma_{l,n}} + \frac{n^{1/2}\beta_{\mathrm{G}*,l}}{\sigma_l} \cdot \frac{\sigma_l}{\sigma_{n,l}} = T^*_{n,l} + \frac{n^{1/2}\beta_{\mathrm{G}*,l}}{\sigma_l} \cdot \frac{\sigma_l}{\sigma_{n,l}},$$

it is therefore true that

$$D^*_{n,l} \xrightarrow{\mathcal{L}} N(0, V_{\mathrm{D}}(P)) \text{ and } T^*_{n,l} \xrightarrow{\mathcal{L}} N(0, \rho_{\mathrm{T}}(P)),$$

where $V_{\mathrm{D}}(P)$ is the covariance matrix and $\rho_{\mathrm{T}}(P)$ is the correlation matrix. Thus according to the Theorem 2 in Dudoit, et al. (2003), the bootstrapping algorithm such as the following can be used to estimate the null distribution:

Algorithm 1.

1. Obtain a bootstrapping set of samples as $\{(Y^b_i, Z^b_i), i = 1, 2, \ldots, n\}$;
2. Compute D^b_n and T^b_n, respectively;
3. Repeat Step 1 and 2 for a total of B times;
4. Compute the sample mean and the sample variance for each element in D^b_n and T^b_n;
5. Compute

$$D^{*,b}_{n,l} = \sqrt{\min\{1, 1/\widehat{\mathrm{var}}(D^b_{n,l})\}}\{D^b_{n,l} - \widehat{E}(D^b_{n,l})\}, \text{ and}$$
$$T^{*,b}_{n,l} = \sqrt{\min\{1, 1/\widehat{\mathrm{var}}(T^b_{n,l})\}}\{T^b_{n,l} - \widehat{E}(T^b_{n,l})\},$$

   respectively.

6. Compute the empirical distributions of $D^{*,b}_{n,l}$ and $T^{*,b}_{n,l}$ for $b = 1, 2, \ldots, B$.

After the null distribution Q_0 is estimated, there are two procedures to choose actual cutoffs, $\beta^c_{\mathrm{G}} = (\beta^c_{\mathrm{G},1}, \beta^c_{\mathrm{G},2}, \ldots, \beta^c_{\mathrm{G},L})^{\mathrm{T}}$, say, to decide the rejection regions for $D_{n,l}$ and $T_{n,l}$, $l = 1, 2, \ldots, L$, namely, single-step common-quantile and single-step common-cutoff, to control the FWER. For the single-step common-quantile procedure, the cutoffs can be selected

as the common quantile of the marginal distributions of the estimated Q_0. For the single-step common-cutoff, the common cutoff can be selected as $\inf\{c : \theta_n(R \mid Q_0) \leq \alpha\}$. Furthermore, their adjusted p-values can be computed as $\widetilde{p}_{n,l} = \inf\{\alpha : l \in S_n(\alpha)\}$, $l = 1, 2, \ldots, L$, where $S_n = \{l : T_{n,l} > c_l(\alpha)\}$ (Pollard and van der Laan, 2003).

Based on the aforementioned control of FWER, there are augmentation procedures to select additional rejections to control the gFWER and PFP (van der Laan, et al., 2003b). Specifically, the augmentations are done in the following algorithm:

Algorithm 2:

1. Sort the adjusted FWER p-values as

$$\widetilde{p}_{n,(1)} \leq \widetilde{p}_{n,(2)} \leq \ldots \leq \widetilde{p}_{n,(L)},$$

   where $(\cdot)$ defines a permutation of $\{1, 2, \ldots, L\}$. Then the rejected null hypotheses of S_n consist of $\{l : \widetilde{p}_{n,l} \leq \alpha\}$ or $\{(l) : l = 1, 2, \ldots, R\}$;

2. Additional rejections are selected as $\{(l) : l = R+1, \ldots, R+k\}$, for $k = k_0$ of a given $0 \leq k_0 \leq L-R$ in the $gFWER$-control, and for $k = \max\{0 \leq l \leq L-R : l/(l+R) \leq \kappa\}$ of a given κ in FPF-control, respectively.

Thus the adjusted p-value for controlling the $gFWER(k)$ is calculated as $\widetilde{p}_{n,(l-k)} I(l > k)$, and the adjusted p-value for controlling the $PFP(\kappa)$ is calculated as $\inf\{\alpha : \{l - R(\alpha)\}/l \leq \kappa\}$.

3 An example

A study of forest tree growth was conducted at a forest farm in Xuzhou City of Jiangsu Province in China since the Spring of 1988. The study materials used in the study were derived from the triple hybridization of Populus (poplar). As described in Wu, et al. (1992), a *Populus deltoides* clone (designated I-69) was used as a female parent to mate with an interspecific *P. deltoides* × *P. nigra* clone (designated I-45) as a male parent to produce the hybrids Euramerica poplar, *P. euramericana*. A total of 450 one-year-old rooted three-way

hybrid seedlings were planted at a spacing four by five meters in the forest farm. The total stem heights and diameters are measured and collected at the end of each of the 11 growing seasons.

The genetic linkage maps based on the pseudo-test backcross design were constructed using 90 randomly selected genotypes of the 450 hybrids with random amplified polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs), and intersimple sequence repeats (ISSRs), see Yin, et al. (2002). These parent-specific maps consist of the 19 largest linkage groups for each parent parental map. They amount to 19 pairs of chromosomes. To contrast with previously reported results in Wu, et al. (2002), we also choose the linkage group 10 of the *P. deltoides* parental map to detect statistically meaningful QTLs that potentially affect the diameter growth of the forest trees with the proposed methodologies.

In Wu, et al. (2002), it was observed that most of the growth curves might display sigmoidal shape for the phenotypes, such as the diameter, as function of time, i.e., the years. The plot of the observed curves is reproduced in Figure 1(a). Two logistic functions of different set of parameters were chosen to model the functional quantitative traits linked to the possible genotypes. When the growth curves are log-transformed, as shown in Figure 1(b), they mostly share similar shape and are also parallel, which may suggest that the assumption of common baseline function and the additive differences among the curves are reasonable in the semiparametric model (3). Thus we use $E[\log\{Y_i(t)\} \mid Z_i] = \mu_0(t) + \beta_{\mathrm{G}} G_i$ to estimate the QTL parameter β_{G} for each of the two consecutive markers. By applying the multiple comparison procedures, it was found that the null hypotheses of no QTL was rejected at the first pair of markers, with an adjusted p-value of 0.01, which is highly significant for a potential QTL located between the markers of CA/CCC-640R and CG/CCC-825. This is consistent with the finding reported in Wu, et al. (2002) using both fixed and random effects model.

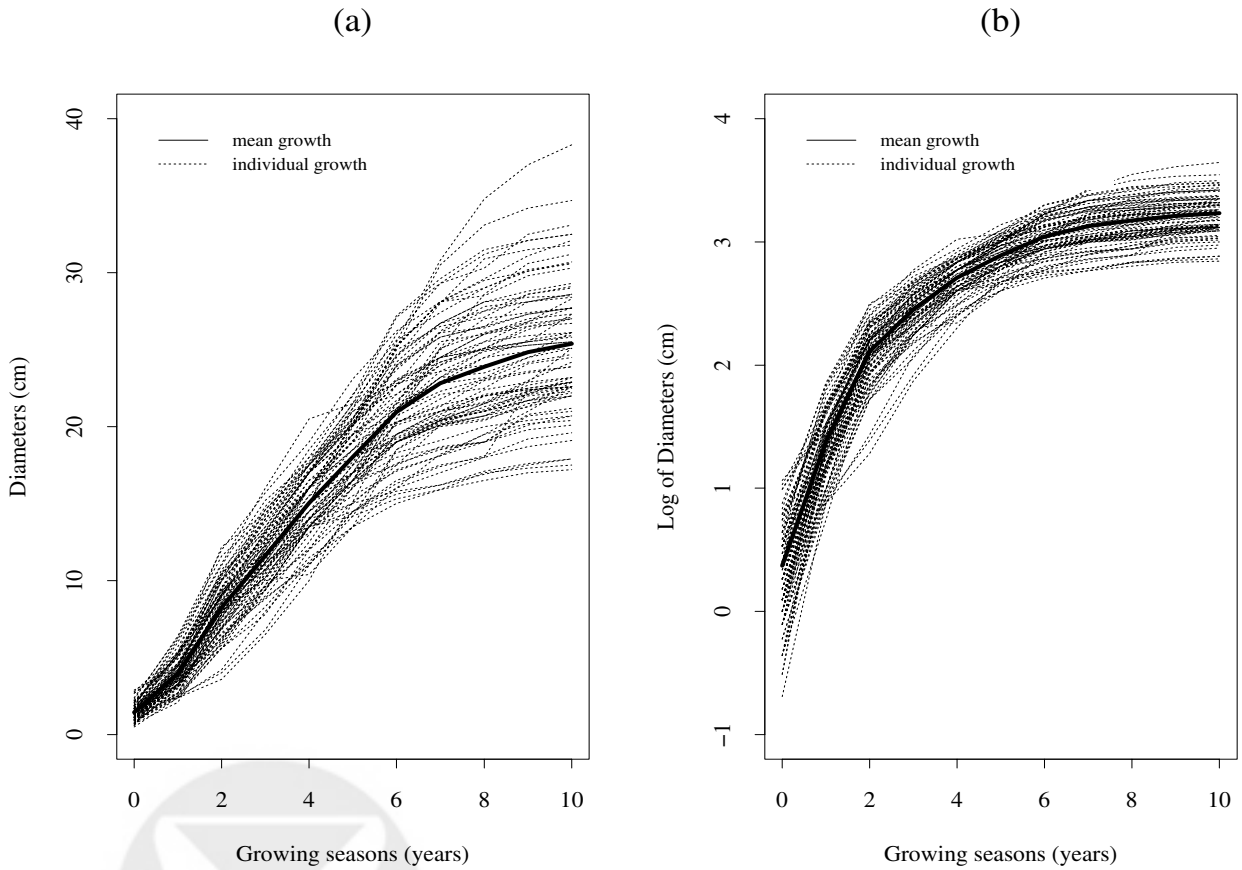

Figure 1: Growth curves of the diameters of forest trees

4 Discussion

The methodology proposed in this article is a type of regression interval mapping (Lander and Bostein, 1989; Haley and Knott, 1992). In general, there are two advantages of regression interval mapping: the first, the interval mapping uses more information from the consecutive markers, which can be more precise to determine the QTL, and the other advantage of regression mapping is that the statistical analysis is straightforward and involves less computing burden than the conventional maximum likelihood mapping. Our approach has two additional features. The first one is that the quantitative trait is no longer viewed as one-dimensional outcome but a functional curve, such as the growth curves discussed before. In fact, there are other examples of functional traits may be of important scientific interest as well, such as the CD4 dynamics of HIV-infected patients, or the blood pressure of hypertensive patients. The second feature is the semiparametric framework of our methods, which does not assume the explicit structure of the errors. In fact, there has been effort by Zou, et al. (2003) to develop rank-based regression approaches to deal with the unknown symmetric error distributions, although their methods are still limited to the one-dimensional outcomes.

Our methodologies can be easily extended to several other occasions. For examples, one such occasion is the so-called Composite Interval Mapping. In the interval mapping, although it has greater advantage than the single marker mapping, it may be still biased if multiple QTLs are linked to the marker or the interval between the markers. To extend the proposed methods to the Composite Interval Mapping (Zeng, 1994) for the functional traits, consider

$$E\{Y(t) \mid Z(t)\} = \mu_0(t) + \beta_{\mathrm{G}} G + \sum_{k=1}^{K} \beta_k G_k,$$

where G_k are the markers selected for genetic background control. This would adjust for the effect of other potential QTLs outside the interval containing the putative QTL of interest. Another occasion is to extend the methodologies to more complex design such as the F_2-design. Specifically, the conditional probabilities of three genotypes at a QTL bracketed by two markers are determined in Table 3. Using this table, we can derive similar marginalised model and hence to estimate the parameter β_{G} to determine the potential existence of a

Table 4: *Conditional probabilities of genotype indicator G at a QTL bracketed by markers P_l and P_{l+1} in an F_2 population. When r_{l1} or r_{l2} is relatively small, $r_{l1}+r_{l2}$ approximates r_l.*

Marker type	Marker genotypes		Conditional probabilities		
M_l	P_l	P_{l+1}	$p_l^{\mathrm{G}} = \mathrm{pr}\{G=-1\}$	$p_l^{\mathrm{G}} = \mathrm{pr}\{G=0\}$	$\mathrm{p}_l^{\mathrm{G}} = \mathrm{pr}\{G=1\}$
1	A_lA_l	$A_{l+1}A_{l+1}$	$\frac{(1-r_{l1})^2(1-r_{l2})^2}{(1-r_l)^2}$	$\frac{2r_{l1}r_{l2}(1-r_{l1})(1-r_{l2})}{(1-r_l)^2}$	$\frac{r_{l1}^2r_{l2}^2}{(1-r_l)^2}$
2	A_lA_l	$A_{l+1}a_{l+1}$	$\frac{(1-r_{l1})^2(1-r_{l2})r_{l2}}{(1-r_l)r_l}$	$\frac{r_{l1}(1-r_l1)\{r_{l2}^2+(1-r_{l2})\}^2}{(1-r_l)r_l}$	$\frac{r_{l1}^2r_{l2}(1-r_{l2})}{(1-r_l)r_l}$
3	A_lA_l	$a_{l+1}a_{l+1}$	$\frac{(1-r_{l1})^2r_{l2}^2}{r_l^2}$	$\frac{2r_{l1}r_{l2}(1-r_l1)(1-r_{l2})}{r_l^2}$	$\frac{r_{l1}^2(1-r_{l2})^2}{r_l^2}$
4	A_la_l	$A_{l+1}A_{l+1}$	$\frac{r_{l1}(1-r_{l1})(1-r_{l2})^2}{r_l(1-r_l)}$	$\frac{\{r_{l1}^2+(1-r_{l1})^2\}r_{l2}(1-r_{l2})}{r_l(1-r_l)}$	$\frac{r_{l1}(1-r_{l1})r_{l2}^2}{r_l(1-r_l)}$
5	A_la_l	$A_{l+1}a_{l+1}$	$\frac{2r_{l1}r_{l2}(1-r_{l1})(1-r_{l2})}{r_l^2+(1-r_l)^2}$	$\frac{\{(1-r_{l1})^2+r_{l1}^2\}\{(1-r_{l2})^2+r_{l2}^2\}}{r_l^2+(1-r_l)^2}$	$\frac{2r_{l1}r_{l2}(1-r_{l1})(1-r_{l2})}{r_l^2+(1-r_l)^2}$
6	A_la_l	$a_{l+1}a_{l+1}$	$\frac{r_{l1}(1-r_{l1})r_{l2}^2}{r_l(1-r_l)}$	$\frac{\{(1-r_{l1})^2+r_{l1}^2\}r_{l2}(1-r_{l2})}{r_l(1-r_l)}$	$\frac{r_{l1}(1-r_{l1})(1-r_{l2})^2}{r_l(1-r_l)}$
7	a_la_l	$A_{l+1}A_{l+1}$	$\frac{r_{l1}^2(1-r_{l2})^2}{r_l^2}$	$\frac{2r_{l1}(1-r_{l1})(1-r_{l2})}{r_l^2}$	$\frac{(1-r_{l1})^2r_{l2}^2}{r_l^2}$
8	a_la_l	$A_{l+1}a_{l+1}$	$\frac{r_{l1}^2r_{l2}(1-r_{l2})}{r_l(1-r_l)}$	$\frac{r_{l1}(1-r_{l1})\{r_{l2}^2+(1-r_{l2})^2\}}{r_l(1-r_l)}$	$\frac{(1-r_{l1})^2r_{l2}(1-r_{l2})}{r_l(1-r_l)}$
9	a_la_l	$a_{l+1}a_{l+1}$	$\frac{r_{l1}^2r_{l2}^2}{(1-r_l)^2}$	$\frac{2r_{l1}(1-r_{l1})r_{l2}(1-r_{l2})\}}{(1-r_l)^2}$	$\frac{(1-r_{l1})^2(1-r_{l2})^2}{(1-r_l)^2}$

putative QTL.

In general, it is complicated to use multiple comparisons to determine the threshold of the test statistics. The Wald's test statistics with the usual pointwise significance levels are not adequate due to the genome-wise scanning of the makers. Our choice of multiple comparison approach are mainly for the dense map (Lander and Bostein, 1989). For sparse map in which the markers are sparse and widely separated, the marker intervals can be considered as independent approximately. Then the usual Bonferroni correction may be suffice to explore the potential QTLs. Unlike the permutation tests proposed in Churchill and Doerge (1994), the multiple comparison procedures used in this article do not involve specific choice of the null distribution and more protected from any misspecifications of the underlying distributions.

Appendix A: Asymptotics

A.1. Weak of Convergence of $n^{-1/2}\mathcal{E}(\cdot;\beta_*)$

Our proof follows an extension of the Appendix 2 in Cheng and Wei (2000). Denote $\mathcal{B}(t) = \sum_{i=1}^n \int_0^t \Delta_i(s)\Psi(s)dM_i(s)$ and $\mathcal{B}_\varphi(t) = \sum_{i=1}^n \int_0^t \Delta_i(s)\Psi(s)\varphi_i(s)dM_i(s)$. Then $\mathcal{E}(\beta_*) = \mathcal{B}_\varphi(\tau) - \int_0^\tau \bar{\varphi}(t)d\mathcal{B}(t)$. For any $t > 0$, $\mathcal{B}(t)$ and $\mathcal{B}_\varphi(t)$ are sums of independently and identically distributed zero-mean terms. By the Central Limit Theorem, $n^{-1/2}(\mathcal{B}(t), \mathcal{B}_\varphi(t))$ converges in distribution to a zero-mean Gaussian process, $(\mathcal{W}(t), \mathcal{W}_\varphi(t))$, say.

Assume that $\varphi_i(\cdot)$, $i = 1, \ldots, n$, are of bounded variation. Moreover, without loss of generality, $\varphi_i(\cdot)$ are assumed to be non-negative. Then the individual terms of $\mathcal{B}(\cdot)$ and $\mathcal{B}_\varphi(\cdot)$ can be written as sums of monotone functions in t and hence "manageable." Thus $n^{-1/2}(\mathcal{B}(t), \mathcal{B}_\varphi(t))$ converges weakly to $(\mathcal{W}(t), \mathcal{W}_\varphi(t))$, as $n \to \infty$ (Pollard, 1990, p.38 and p.53). By the strong embedding theorem in Shorack and Wellner (1986, p.47), there exists an induced probability space such that $(n^{-1/2}\mathcal{B}(t), n^{-1/2}\mathcal{B}_\varphi(t), n^{-1}\sum_{i=1}^n \Delta_i(t), n^{-1}\sum_{i=1}^n \Delta_i(t)\varphi_i(t))$ converges almost surely. By the Lemma 8.2.3 in Chow and Teicher (1988, p.265) coupled with the Helly's theorem in Serfling (1980, p.352), it is true that

$$n^{-1/2}\int_0^t \frac{n}{\sum_{i=1}^n \Delta_i(s)} d\mathcal{B}(s) \to \int_0^t \frac{1}{E\Delta_1(s)} d\mathcal{W}(s) \text{ and } n^{-1/2}\int_0^t \bar{\varphi}(s)d\mathcal{B}(s) \to \int_0^t \bar{\varphi}_*(s)d\mathcal{W}(s)$$

almost surely and uniformly in t. The weak convergence of $n^{-1/2}\mathcal{E}(\beta_*, \theta_*)$ thus follows in the original probability space, due to their convergence almost surely to $\mathcal{W}_\varphi(\tau) - \int_0^\tau \bar{\varphi}_*(s)d\mathcal{W}(s)$ in the induced probability. The calculation of the variance-covariance matrix is thus straightforward.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approached to multiple testing. *Journal of Royal Statistical Society* **57**, 289-300.

Capra B. and Muller H. G. (1997). An accelerated-time model for response curves. *Journal of American Statistical Association* **92**, 72-83.7

Cheng, S. C. and L. J. Wei (2000). Inferences for a semiparametric model with panel data. *Biometrika* **87**, 89-97.

Chow, Y. S. and Teicher, H. (1988). *Probability Theory: Independence, Interchangeability, Martingales*, 2nd Ed. New York: Springer.

Dudoit, S., van der Laan and Pollard, K. S. (2003). Multiple testing. Part I: Single-step procedures for control of general type I error rates. *UC Berkeley Biostatistics Technical Report #138*.

Haley, C. S. and Knott, S. A. (1992). A simple regression method for mapping quantitative traits in line crosses using flanking markers. *Heredity* **69**, 315-324.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley: New York.

Kenward, M. C. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics* **36**, 296-308.

Kirkpatrick, M. and Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology* **27**, 429-450.

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

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of American Statistical Association* **96**, 103-126.

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

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

Pletcher, S. D. and Geyer, C. J. (1999). The genetic analysis of age-dependent traits: modeling the character process. *Genetics* **153**, 815-835.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Hayward: Institute of Mathematical Sciences.

Pollard, K. S. and van der Laan, M. J. (2003). Resampling-based multiple testing: asymptotic control of type I error and application to gene expression data. *Journal of Statistical Planning and Inference*, in press.

Rice, J. A. (2003). Functional and longitudinal data analysis: perspectives on smoothing. *UC Berkeley Department of Statistics Technical Report*.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Process with Applications to Statistics*. New York: Wiley.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2003a). Multiple testing. Part II: Step-down procedures for control of the family-wise error rate. *UC Berkeley Biostatistics Technical Report #139*.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2003b). Multiple testing. Part III: Procedures of control of the generalized family wise error rate and proportion of false positives. *UC Berkeley Biostatistics Technical Report #141*.

Wang, J.-L. and Wang, W. (2001). Comment on "Semiparametric and nonparametric regression analysis of longitudinal data" by Lin and Ying. *Journal of American Statistical Association* **96**, 119-123.

Wu, R. L., Ma, C. X., Chang, M., Little, R. C., Wu, S. S., Yin, T. M., Huang, M. R., Wang, M. X. and Casella, G. (2002). A logistic mixture model for characterizing genetic determinants causing differentiation in growth trajectories. *Genetic Research* **19**, 235-245.

Yin, T. M., Zhang, X. Y., Huang, M. R., Wang, M. X., Zhuge, Q., Zhu, L. H., Zeng, Z.-B. and Wu, R. L. (2002). Molecular linkage maps of the Populus genome. *Genome* **45**, 541-555.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689-699.

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

Zou, F., Yandell, B. S. and Fine, J. P. (2003). Rank-based statistical methodologies for Quantitative Trait Locus Mapping. *Genetics* **165**, 1599-1605.