Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 4

Nonparametric Functional Mapping of Quantitative Trait Loci Underlying Programmed Cell Death

Yuehua Cui* Rongling Wu†

George Casella‡ Jun Zhu**

*Michigan State University, cui@stt.msu.edu

†University of Florida, rwu@ufl.edu

‡University of Florida, casella@ufl.edu

**Zhejiang University, jzhu@zju.edu.cn

Copyright ©2008 The Berkeley Electronic Press. All rights reserved.

Nonparametric Functional Mapping of Quantitative Trait Loci Underlying Programmed Cell Death*

Yuehua Cui, Rongling Wu, George Casella, and Jun Zhu

Abstract

The development of an organism represents a complex dynamic process, which is controlled by a network of genes and multiple environmental factors. Programmed cell death (PCD), a physiological cell suicide process, occurs during the development of most organisms and is, typically, a complex dynamic trait. Understanding how genes control this complex developmental process has been a long-standing topic in PCD studies. In this article, we propose a nonparametric model, based on orthogonal Legendre polynomials, to map genes or quantitative trait loci (QTLs) that govern the dynamic features of the PCD process. The model is built under the maximum likelihood-based functional mapping framework and is implemented with the EM algorithm. A general information criterion is proposed for selecting the optimal Legendre order that best fits the dynamic pattern of the PCD process. The consistency of the order selection criterion is established. A nonstationary structured antedependence model (SAD) is applied to model the covariance structure among the phenotypes measured at different time points. The developed model generates a number of hypothesis tests regarding the genetic control mechanism of the PCD process. Extensive simulation studies are conducted to investigate the statistical behavior of the model. Finally, we apply the model to a rice tiller number data set in which several QTLs are identified. The developed model provides a quantitative and testable framework for assessing the interplay between genes and the developmental PCD process, and will have great implications for elucidating the genetic architecture of the PCD process.

KEYWORDS: Legendre polynomial, maximum likelihood, order selection, programmed cell death, quantitative trait loci

---

*The authors thank two anonymous referees for their constructive comments on this manuscript. This work was partially supported by grants from Michigan State University (IRGP 91-4533) and the National Science Foundation (DMS 0707031) to Y. Cui, and by a grant from the National Science Foundation (DMS 0540745) to R. Wu.

1 Introduction

As a physiological cell suicide process that occurs during the development of most organisms, programmed cell death (PCD) is typically a complex dynamic longitudinal trait (Ameisen, 2002). PCD functions as a defense mechanism against disease or virus attacks, to balance an organism's metabolism (Elis et al., 1991). It is involved in multiple developmental stages of an organism: cell differentiation, proliferation, aging and dying, and can be mathematically described by five distinguishable developmental phases (Cui et al., 2006). PCD has been universally observed in a wide range of phyla spanning from plants and animals to humans (Jacobson et al., 1997; Pennell and Lamb, 1997; Vaux and Korsmeyer, 1999). While its role in shaping an organism's development has been unanimously recognized, the mechanisms underlying this complex developmental process are poorly understood (Vaux and Korsmeyer, 1999). It is highly expected that identifying genes that contribute to this unique process would greatly enhance our understanding of the pathogenesis of important diseases such as cancer (Hanahan and Weinberg, 2000; Yuan and Horvitz, 2004) and, ultimately, help us to explore opportunities to therapeutically prevent, control and treat diseases (Martin, 2006).

PCD-related phenomena have been studied experimentally using simple structure model systems such as the nematode *Caenorhabditis elegans* and the fruitfly *Drosophila*. Several PCD-associated genes have been identified under experimental conditions (Ellis and Horvitz, 1986; Horvitz, 1999, 2003; Yuan and Horvitz, 2004). The identification of PCD genes in more complicated organisms, such as humans or flowering plants, however, has not been quite successful. This is partly due to the higher-order complex structure of these organisms. Before the positional cloning of a candidate gene, a natural way to target those genetic regions harboring PCD-related genes is to implement a QTL mapping strategy. Considering the longitudinal feature of the PCD trait (Ameisen, 2002), traditional QTL mapping that only considers the phenotypic trait measured at a particular time point will be less powerful for detecting PCD QTLs. More recently, a series of statistical models, called functional mapping models, have been developed to map genes responsible for the dynamic features of a trait (Ma et al., 2002; Wu et al., 2004; reviewed in Wu and Lin, 2006). By incorporating various well-established mathematical functions into the mapping framework, functional mapping has flexibility for mapping genes that underlie complex longitudinal traits. It has been applied to the study of many genetic mechanisms of biological or biomedical processes, such as allometric scaling (Wu et al., 2002), tumor progression (Liu et al., 2005), HIV dynamics (Wang et al., 2004), and drug response (Lin et al., 2005). Spe-

cific parametric functions have been adopted in these studies to fit specific developmental patterns. Simulation and real data analysis from these studies indicate that functional mapping has high power to detect dynamic QTL effects.

In real life, however, many biological characteristics, such as PCD, show unique developmental patterns, which cannot be explained by currently available parametric mathematical functions or equations (Cui et al., 2006). Traditional parametric approaches, such as parametric regression, require specific quantitative information about regression forms. Hence, they do not have enough flexibility to capture the nature of this process and have certain limitations in fitting a developmental PCD curve. To better understand the dynamic gene effects that govern complex longitudinal PCD traits, two major statistical issues need to be addressed: (1) How does one model the mean curve to better capture the dynamic developmental pattern of the PCD process? (2) How does one model the variation at different time points and the intra-individual correlation structure to better explain the variations caused by natural gene and environmental perturbations? These two issues form two inter-related processes and should be clearly addressed to dissect the genetic effects of major QTLs.

Recently, Cui et al. (2006) proposed a semi-parametric approach for mapping PCD genes by dividing the overall development process into two separate stages and fitting them with different functions. This unique modelling strategy brings advantages in fitting PCD trajectories, since it incorporates the growth law in the early developmental stage. However, the growth function needs at least three time points to estimate growth parameters, which could be easily violated in practice. Moreover, the asymptote might not be reached to allow it to be able to fit a logistic growth function (Lin and Wu, 2006). These limitations certainly restrict the utility of a semiparametric approach. Relaxing the assumption of parametric and semi-parametric approaches, a more natural way to fit the developmental process would be to use a nonparametric approach, such as nonparametric regression (Hart and Wehrly, 1986; Müller, 1988; Altman, 1990; Fraiman and Meloche, 1994; Altman and Casella, 1995; Boularan et al., 1995; Ferreira et al., 1997). These methods treat time as the only explanatory variable and estimate the mean response curve by smoothing the raw data. No study has been reported to apply these methods in genetic mapping.

In genetic studies, the random regression (RR) model has been used commonly for modelling dynamic additive effects in which the Legendre polynomial (LP) is applied to fit the developmental curves, partly due to the nature of the flexibility of the orthogonal polynomial (Pool et al., 2000a, 2000b; Kirk-

patrick and Heckman, 1989). Lin and Wu (2006) recently applied the Legendre polynomial to jointly model longitudinal traits and time-to-event data. The authors applied regular AIC or BIC to select the Legendre order. However, the consistency of order selection is not established, especially under the mixture model-based multivariate functional mapping framework. Moreover, the authors proposed to select the Legendre order only under the null hypothesis (i.e., one mean trajectory). The same order was then assigned to different QTL genotypes under the alternative (i.e., more than one mean trajectory) across the whole linkage group. In general, the PCD structure is much more complicated under the alternative than that under the null. This restriction could lead to a potential under-fit of the data under the alternative at different testing positions, and may consequently lead to information loss, such as missing QTLs, or to wrong inferences such as false QTL detection and biased QTL location estimation.

To overcome the limitations of the parametric and semiparametric approaches, in this article we will develop a nonparametric approach for mapping QTL underlying the complex developmental PCD process. The Legendre function is applied to fit a dynamic PCD curve and the structured antedependence (SAD) covariance structure is used to model the intra-individual correlation. We further propose a general information selection criterion and study its consistency property under the current mapping framework. The consistency of the selection criteria is established and its small sample performance is evaluated through simulation studies. Detailed parameter estimation procedures are given. Monte Carlo simulation studies and a real example using rice tiller number data, are given to demonstrate the utility of the developed model. Comparisons with current approaches are discussed.

2 Statistical Method

2.1 The Finite Mixture Model and Likelihood Function

For simplicity, we consider a standard backcross design. The model can easily be extended to other genetic designs, such as an F_2 design. Consider a backcross design, initiated with two contrasting homozygous inbred lines, in which there are two genotypes at each locus. A genetic linkage map is constructed with molecular markers, aimed at identifying QTL responsible for the PCD process. The longitudinal PCD trait, such as cell count or other measurement at tissue or organ level, is observed at a finite set of time points for each individual. The observed PCD trait is quantitative in nature and its mean

function can be modelled by a parametric or nonparametric function.

Suppose there is a putative segregating QTL, with alleles Q and q, that affects the PCD trait, but with different degrees. In QTL mapping studies, QTL genotype is generally considered as missing. Statistically, this is a missing data problem where the phenotypic trait at time t can be described by a mixture of two mean functions for different genotypes. Assuming independence and multivariate normality distribution, the joint likelihood function conditional on the observed phenotype ($\mathbf{y}$) and marker data ($\mathcal{M}$) is given by

$$L(\mathbf{\Omega}|\mathbf{y},\mathcal{M}) = \prod_{i=1}^{n}[\pi_{1|i}f_1(\mathbf{y}_i|\mathbf{\Omega},\mathcal{M}) + \pi_{0|i}f_0(\mathbf{y}_i|\mathbf{\Omega},\mathcal{M})] \quad (1)$$

where $\mathbf{y}_i = [y_i(t_1),\cdots,y_i(t_\tau)]$ is the observed trait vector for individual i ($i = 1,\cdots,n$) over τ time points; $\pi_{j|i}$ ($j = 0, 1$) is the mixture proportion for individual i with genotype j, which can be obtained based on the Mendelian segregation theory (Lynch and Walsh, 1998); the unknown parameters in $\mathbf{\Omega}$ contain three sets of parameters, one defining the co-segregation between the QTL and markers and, thereby, the location of the QTL relative to the markers, denoted by $\mathbf{\Omega}_q$, and the other two defining the distribution of the PCD trait for each QTL genotype, denoted by $(\mathbf{\Omega}_m, \mathbf{\Omega}_v)$, where $\mathbf{\Omega}_m = (\mathbf{\Omega}_{m_1}, \mathbf{\Omega}_{m_0})$ defines the mean vector for different genotypes and $\mathbf{\Omega}_v$ defines the covariance matrix among different time points.

The multivariate normal distribution for progeny i, which carries genotype j, can be expressed as

$$f_j(\mathbf{y}_i|\mathbf{\Omega},\mathcal{M}) = \frac{1}{(2\pi)^{\tau/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{y}_i-\mathbf{m}_j)\mathbf{\Sigma}^{-1}(\mathbf{y}_i-\mathbf{m}_j)^T\right], \quad (2)$$

where $\mathbf{m}_j = [m_j(t_1),\cdots,m_j(t_\tau)]$ is the mean vector common for all individuals with genotype j. At a particular time point (say t), the relationship between the response and mean can be written as a linear regression model,

$$y_i(t) = \delta_i m_1(t) + (1-\delta_i)m_0(t) + e_i(t), \quad (3)$$

where δ_i is an indicator variable with the value taken to be 1 or 0, depending on $j = 1$ or 0, respectively, and $e_i(t)$ is the residual error, which is normal, with mean zero and variance $\sigma^2(t)$. The errors for individual i at two different time points, t_1 and t_2, are correlated with covariance $\mathrm{cov}(y_i(t_1), y_i(t_2))$. The variance and covariance parameters comprise the covariance matrix $\mathbf{\Sigma}$, whose elements are the common parameters specified by $\mathbf{\Omega}_v$.

2.2 Modelling the Mean PCD Process

One of the statistical challenges in mapping QTLs governing the PCD process is how to model the dynamic mean function, m, given in Eq. (2). The challenge lies in the complexity of intra- and inter-individual variations, as well as in the unique developmental pattern that the PCD process possesses. A typical PCD developmental curve can be described by five stages (Cui et al., 2006). No appropriate mathematical function or equation has been developed to fit this unique developmental pattern. In fact, different organisms may carry different PCD patterns, partly due to their mass differences. This also makes it difficult to formulate a unified parametric function to describe the developmental trajectory. A natural and flexible way to model this process is in a nonparametric fashion that lets the data specify the best fit. Among a pool of choices, the orthogonal Legendre function has shown a number of merits for modelling genetics data (Pool et al., 2000a, 2000b; Kirkpatrick and Heckman, 1989), and can thus be applied to model the PCD process.

Denote the Legendre polynomial (LP) of order r at time t as $P_r(t)$. By choosing different orders of orthogonal polynomials, the Legendre function has the potential to approximate the functional relationships between trait values and different time points to any specified degree of precision. The measurement time can be adjusted to fit the orthogonal function range [-1, 1], by

$$t' = -1 + \frac{2(t - t_1)}{t_\tau - t_1}, \tag{4}$$

where t_1 and t_τ are the first and last measurement time points, respectively.

With an appropriate order r, the time-dependent genotypic means for different QTL genotypes at time t can be fitted by the orthogonal LP. A family of such polynomials is denoted by

$$\mathbf{P}(t') = [P_0(t'), P_1(t'), \cdots, P_r(t')]^T$$

and a vector of time-independent values, specific for genotype j with order r is denoted by

$$\mathbf{u}_j = (u_{j0}, u_{j1}, \cdots, u_{jr})^T.$$

This genotype-related vector is called the *base genotypic vector*, and the parameters within the vector are called the *base genotypic means* for QTL genotype j. A vector of polynomials $\mathbf{P}(t')$ is also called the base function. Then, the time-dependent genotypic values at time t, $m_j(t)$, can be described as a linear combination of $\mathbf{u}_j$ weighted by a series of the polynomials, i.e.

$$m_j(t) = \mathbf{u}_j^T \mathbf{P}(t') \tag{5}$$

Thus, for individual i, whose QTL genotype is j, its genotype means at different time points can be modelled by the following vector

$$\mathbf{m}_{j|i} = [m_{j|i}(t_1), \cdots, m_{j|i}(t_\tau)] \tag{6}$$

This modelling approach has great flexibility in modelling curves in which the logistic or other parametric functions do not fit. By choosing the appropriate order, the model can better capture the intrinsic developmental PCD trend.

2.3 LP Order Selection

One of the major advantages in using a nonparametric approach is that the best fit is specified by the data themselves, which offers a certain degree of flexibility and accuracy in terms of model fitting. If the degree of LP is k (the highest power of the polynomial), the order of LP is $k + 1$ (the number of the coefficients defining the polynomial). The LP order selection is similar to selecting variables in a regression study. Normally, a higher order always provides a better fit. However, if a model contains too many parameters, it will greatly reduce the model efficiency and increase computation burden. On the other hand, if the developmental curve is fit by low orders, a model that contains too few parameters will not be flexible enough to approximate important features in the data. Consequently, this will result in a bias contribution to the misfit, due to a lack of flexibility. In both cases, the use of poor or redundant orders can be harmful. There should be a tradeoff between the LP order and the model efficiency. It is essential to select the optimal LP order without over- or under-fitting the PCD curve. One choice is to conduct the order selection using an information-theoretic criterion.

A general form for the basic information criterion to select the LP order can be given by

$$IC_r(n) = -2 \ln L(\hat{\mathbf{\Omega}}|Y, r) + c(n)p_r(\tau) \tag{7}$$

where the first term on the right hand side is the negative maximum log-likelihood of the data Y, given the model parameter estimates; $\hat{\mathbf{\Omega}}$ contains the MLEs of mean and covariance parameters; $p_r(\tau)$ represents the number of free parameters, which only depends on the number of measurement time with an order of r, i.e., $p_r(\tau) = \text{dimension}(\mathbf{\Omega}|r)$; $c(n)$ is a penalty term. The model that minimizes the criterion is considered to be the optimal one. Clearly, both the AIC and BIC information criteria are two special cases of this general form, with AIC having $c(n) = 2$ and BIC having $c(n) = \tau log(n)$.

When the proposed information criterion is applied to the current mixture model-based likelihood framework, the consistency property of the selection criterion needs to be established. The consistency of a selection criterion is defined as the probability of choosing the correct model approaching one as sample size goes to infinity. Hence, a model selection criterion is consistent if

$$P[\text{choose model } M_2] = P[IC_{r_1}(n) < IC_{r_2}(n)] \rightarrow 1, \text{as n} \rightarrow \infty$$

where we assume that M_2 is the correct model; $IC_{r_1}(n)$ and $IC_{r_2}(n)$ are the information criteria for models M_1 and M_2, with LP order r_1 and r_2, respectively. For a backcross design assuming one QTL, there are two developmental trajectories corresponding to two different genotypes. The density function for each observation is modelled as a mixture of two distributions corresponding to two different genotypes. To apply the selection criterion, we first show the consistency property demonstrated by the following theorem.

Theorem *Under the current functional mapping framework, a model selection criterion $IC_k(n)$ defined in Equation (7) is consistent if*

$$\frac{c(n)[p_{r_0}(\tau) - p_r(\tau)]}{n} \rightarrow 0, \text{ as } n \rightarrow \infty$$

where r_0 and r are the optimal and selected LP order, respectively.

Proof See Appendix B for a rigorous proof.

Based on the theorem, the regular AIC and BIC information criteria are consistent under the current mixture model based functional mapping framework. The asymptotic consistency is based on infinite sample size. In practice, sample size is often limited. Simulation studies are designed to check the finite sample performance of the selection criteria, which is given in section 3.

2.4 Modelling the Covariance Structure

Covariance structure modelling is another important and challenging step in the functional mapping of PCD genes. Dissection of the intra-individual correlation will help us to understand how QTLs mediate the developmental pattern. The nonstationary nature of the covariance structure can be best described by the structured antedependence (SAD) model (Jaffrézic et al. 2003). The SAD model, with order p for modelling the error term in Eq. (3), is given by

$$e_i(t) = \phi_1 e_i(t-1) + \cdots + \phi_p e_i(t-r) + \epsilon_i(t) \tag{8}$$

where $\epsilon_i(t)$ is the "innovation" term assumed to be independent and distributed as $\mathcal{N}(0, \sigma_t^2)$. Therefore, the variance-covariance matrix of the PCD process can be expressed as

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{\Sigma}_\epsilon \mathbf{A}^{\mathrm{T}}, \tag{9}$$

where $\mathbf{\Sigma}_\epsilon$ is a diagonal matrix, with diagonal elements being the innovation variance. For the first-order SAD or SAD(1) model, the matrix $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & & 0 \\ \phi_1 & 1 & 0 & & 0 \\ \vdots & & & \ddots & \\ \phi_1^{\tau-1} & \phi_1^{\tau-2} & \cdots & \phi_1 & 1 \end{pmatrix}$$

In general, the SAD order (p) can be selected through an information criterion (Zhao et al., 2005). The closed forms for the inverse and determinant of matrix $\mathbf{\Sigma}$ are given in Appendix A.

2.5 Parameter Estimation

We implement the EM algorithm, originally proposed by Dempster et al. (1977), to obtain the maximum likelihood estimates (MLEs) of the unknown parameters. In general, we do not directly estimate the QTL-segregating parameters ($\mathbf{\Omega}_q$). Instead, we use a grid search approach to estimate the QTL location, by searching for a putative QTL at every 1 or 2 cM on a map interval bracketed by two markers throughout the entire linkage map. The log-likelihood ratio test statistic for a QTL at a testing position is displayed graphically, to generate a log-likelihood ratio plot called the LR profile plot. The genomic position corresponding to a peak of the profile is the MLE of the QTL location. The curve parameters contained in $\mathbf{\Omega}_{m_j}$ $(j = 1, 0)$ and the covariance parameters contained in $\mathbf{\Omega}_v$ can be estimated by the EM algorithm (see Appendix A for a detailed derivation).

2.6 Hypothesis Testing

2.6.1 Global test

The first step toward the understanding of the genetic architecture of the PCD process would be to test whether specific QTLs exist to affect the entire trajectory. After obtaining the MLEs of the parameters, the existence of

a QTL affecting the PCD curve can be tested by formulating the following hypotheses

$$\begin{cases} \mathrm{H}_0: \ \boldsymbol{\Omega}_{m_1} \equiv \boldsymbol{\Omega}_{m_0} \\ \mathrm{H}_1: \ \text{The equalities above do not hold,} \end{cases} \tag{10}$$

where H_0 corresponds to the reduced model, in which the data can be fit by a single curve, and H_1 corresponds to the full model, in which different curves exit that fit the data. The test statistic for testing the hypotheses is calculated as the log-likelihood (LR) ratio of the reduced model to the full model

$$\mathrm{LR} = -2[\log L(\widetilde{\boldsymbol{\Omega}}|\mathbf{y}, \mathcal{M}) - \log L(\widehat{\boldsymbol{\Omega}}|\mathbf{y}, \mathcal{M})]$$

where $\widetilde{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Omega}}$ denote the MLEs of the unknown parameters under H_0 and H_1, respectively. An empirical approach for determining the critical threshold is based on permutation tests (Churchill and Doerge, 1994). Specifically, we randomly reshuffle the individual phenotype data many times, while fixing the structure for the genotype data. The reshuffled data represent random samples from the null distribution, assuming no QTL effects, from which we can determine the threshold value. When reshuffling the data, the phenotype vector for each individual is maintained as a unit, to preserve the intra-individual correlation structure.

2.6.2 Regional test

Once we find the QTL, it would be interesting to test the difference in PCD trajectories over a certain time interval. The question of how a QTL exerts its effects on a period of PCD trajectories $[t_1, t_2]$ can be tested using a regional test approach based on the areas under the curve (AUC). The AUC for genotype j is calculated as

$$\mathrm{AUC}_j = \int_{t_1'}^{t_2'} \mathbf{u}_j^T \mathbf{P}(t') dt'$$

where t' is the adjusted time according to Eq. (4). If the AUC of the two genotypes for a testing period $[t_1, t_2]$ is the same, then we claim there is no QTL effect at that time interval. The hypothesis test for the genetic effect on a period of PCD process can be formulated as

$$\begin{cases} \mathrm{H}_0: \ \mathrm{AUC}_1 = \mathrm{AUC}_0 \\ \mathrm{H}_1: \ \mathrm{AUC}_1 \neq \mathrm{AUC}_0, \end{cases} \tag{11}$$

which is equivalent to testing the difference between the full model with no restriction and the reduced model with a restriction, where 0 and 1 corresponding to different genotypes in a backcross design. One of the applications of this test is to test whether a detected QTL affects the growth phase or the death phase. This can be achieved by comparing the AUC for different genotypes, calculated under the two phases as

$$\mathrm{AUC}_j = \int_{t_1'}^{t*'} \mathbf{u}_j^T \mathbf{P}(t')dt'$$

for testing growth, and

$$\mathrm{AUC}_j = \int_{t*'}^{\tau'} \mathbf{u}_j^T \mathbf{P}(t')dt'$$

for testing death, with $t^{*'}$ representing the transformed transition time point. Similarly, we can also test the QTL effect at different phases, such as at the lag phase or at the exponential phase, described in Cui et al. (2006).

3 Monte Carlo Simulation

Consider a backcross population with which a $100cM$ long linkage group, composed of 6 equidistant markers, is constructed. A putative QTL that affects the PCD process is located at 48 cM from the first marker on the linkage group. The Haldane map function is used to convert the map distance into the recombination fraction. We simulate data with different specifications, namely different heritability levels (H^2=0.1 vs 0.4) and different sample sizes (n=100 vs 200). For each backcross progeny, its phenotype is simulated with 9 equally spaced time points. The covariance is simulated assuming the first-order SAD structure.

We assume that individuals carrying different genotypes will follow the same LP order. For simplicity, we specify the true LP order as six. Data are then simulated assuming that the true order is known in order to demonstrate the selection power under different conditions. We only report the performance of two commonly used criteria, AIC and BIC. The selection power is evaluated as the percentage of the simulated data sets in which the correct models are selected.

Comparisons of the two model selection criteria are summarized in Table 1. The table gives the percentage of those simulations in which the correct models are chosen, for each combination of sample size, heritability level, and

selection criteria. Despite differences in performance between the two selection criteria, trends that hold across different criteria are evident. As might be expected, the overall power of the two selection criteria to select the true model generally increases as sample size and heritability increase. For example, for fixed heritability ($H^2 = 0.1$), the power of BIC increases from 0.82 to 0.93 when sample size increases from 100 to 200. For a fixed sample size ($n = 100$), the power of BIC increases from 0.82 to 1 when heritability level increases from 0.1 to 0.4. However, relative to the effect of sample size, the increase of H^2 from 0.1 to 0.4 can lead to a more significant power improvement than the increase of sample size n from 100 to 200. For example, the power would increase by 22% when H^2 is increased from 0.1

Table 1: *Power comparison for LP order selection with different information criteria, from 100 simulation replicates.*

	$H^2 = 0.1$		$H^2 = 0.4$	
Information criteria	n=100	n=200	n=100	n=200
AIC	82%	85%	90%	91%
BIC	82%	93%	100%	100%

The true order is specified as six. Power is calculated as the percentage of the number of those simulations in which the correct order is selected.

to 0.4, but only by 13% when n is increased from 100 to 200.

Even though both AIC and BIC are asymptotically consistent, their finite sample performances are quite different under a number of conditions. We observe a distinct pattern in which BIC outperforms AIC under different sample sizes and heritability levels. For example, with a heritability level of 0.4 and a sample size of 200, BIC has 100% power to pick the right model while the AIC criterion has only 91% power to choose the correct model. In theory, the BIC criterion can select more parsimonious models than AIC since BIC puts more penalty term ($log(n)$) than AIC (2) does to the likelihood function. The simulation results also confirm this conclusion. Therefore, one can apply BIC to select the optimal order in practice.

With the appropriate LP order selected, we would like to check how well the parameters are estimated. Simulation results are summarized in Table 2. The precision of parameter estimation is evaluated in terms of the square root of the mean squared errors (RMSE) of the MLEs. In general, the model can provide reasonable estimates of the QTL positions (λ) and effects of various kinds, with estimation precision dependent on heritability, sample size and sampling strategy. As might be expected, the precision of the QTL parameter

estimation increases with increased sample size. Histogram plots for all parameter estimates out of the 100 simulation runs under different sample sizes indicate good convergence of the parameter estimates (data not shown). It has also been noted that precision is greatly improved with increased heritability levels, rather than with increased sample sizes. For example, the RMSE of the mean parameter u_{20} for genotype QQ decreases from 0.59 to 0.45 when sample size n increases from 100 to 200. However, for a fixed sample size ($n = 100$), we observe a decrease of RMSE from 0.59 to 0.22 when H^2 increases from 0.1 to 0.4. Thus, the relative precision increase with increased sample size is less attractive, when compared with the precision increase with increased heritability. This information also suggests that, in practice, well-managed experiments, through which residual errors are reduced and therefore H^2 is increased, are more important than simply increasing sample size.

4 A Case Study

We apply the developed model to a real data set to show its utility. Two inbred lines, semi-dwarf IR64 and tall Azucena, were crossed to generate an F_1 progeny population. By doubling haploid chromosomes of the gametes derived from the heterozygous F_1, a doubled haploid (DH) population of 123 lines were founded (Huang et al., 1997). With 123 DH lines, Huang et al. (1997) and Yan et al. (1998) genotyped 175 genetic markers to construct a genetic linkage map of length 2005 cM, representing good coverage of 12 rice chromosomes. The 123 DH lines were planted in two blocks, with each block divided into different plots, each containing eight plants per line. Starting from 10 days after transplanting, tiller numbers were measured every 10 days for five central plants in each plot until all lines had headed. The mean number of tiller numbers for the two blocks were used in QTL analysis.

Tiller growth is thought to be an excellent example of PCD in plants (Greenberg, 1996), since it experiences several developmental stages during rice ontogeny. Different development stages, such as vegetative, reproductive and ripening phases, represent a dynamic PCD process that can be summarized as two distinct growth and death phases (Cui et al., 2006). A genome-wide scan is conducted at every 2 cM distance on each of the 12 chromosomes. At each test position, an LP order is selected using the BIC information criterion and an LR test is conducted. Figure 1 shows the log-likelihood profile plot between the full (there is a QTL) and reduced (no QTL) models for tiller number trajectories across the 12 rice chromosomes. The solid curve represents the LR value at each test position. The 5% significant threshold value

Table 2: *The MLEs of the model parameters and the QTL position derived from 100 simulation replicates, with the SAD(1) covariance structure. The squared root of the mean square errors (RMSEs) of the MLEs are given in parentheses.*

True	$H^2 = 0.1$		$H^2 = 0.4$	
parameters	n=100	n=200	n=100	n=200
QTL position				
$\lambda = 48$	47.98(4.62)	47.02(3.47)	46.38(3.49)	46.26(2.96)
Mean parameters for QQ				
$u_{20} = 9.049$	9.329(0.59)	9.323(0.45)	9.080(0.22)	9.091(0.15)
$u_{21} = 1.151$	1.337(0.45)	1.311(0.32)	1.165(0.18)	1.168(0.13)
$u_{22} = -6.019$	-5.684(0.49)	-5.677(0.44)	-6.006(0.13)	-6.001(0.09)
$u_{23} = 2.651$	2.592(0.29)	2.580(0.23)	2.658(0.11)	2.663(0.09)
$u_{24} = 0.652$	0.774(0.27)	0.753(0.20)	0.672(0.10)	0.658(0.07)
$u_{25} = -0.797$	-0.975(0.27)	-0.949(0.23)	-0.820(0.08)	-0.826(0.07)
$u_{26} = 0.621$	0.619(0.25)	0.636(0.19)	0.621(0.10)	0.626(0.06)
Mean parameters for Qq				
$u_{00} = 7.148$	7.345(0.49)	7.393(0.43)	7.139(0.19)	7.145(0.17)
$u_{01} = 1.379$	1.520(0.40)	1.559(0.32)	1.370(0.17)	1.374(0.13)
$u_{02} = -4.489$	-3.929(0.67)	-3.981(0.59)	-4.467(0.15)	-4.482(0.09)
$u_{03} = 2.004$	1.862(0.34)	1.843(0.25)	2.006(0.12)	1.987(0.09)
$u_{04} = 0.662$	0.695(0.23)	0.699(0.17)	0.679(0.10)	0.682(0.08)
$u_{05} = -0.836$	-0.904(0.24)	-0.928(0.16)	-0.862(0.09)	-0.852(0.07)
$u_{06} = 0.432$	0.410(0.24)	0.403(0.16)	0.448(0.10)	0.445(0.07)
Covariance parameters				
$\sigma^2_{0.1} = 5.065$	4.732(0.41)	4.755(0.35)		
$\sigma^2_{0.4} = 0.844$			0.825(0.04)	0.829(0.03)
$\phi = 0.95$	0.915(0.04)	0.915(0.04)	0.951(0.02)	0.948(0.01)

The location of the simulated QTL is described by the map distance (in cM) from the first marker of the linkage group (100 cM long).

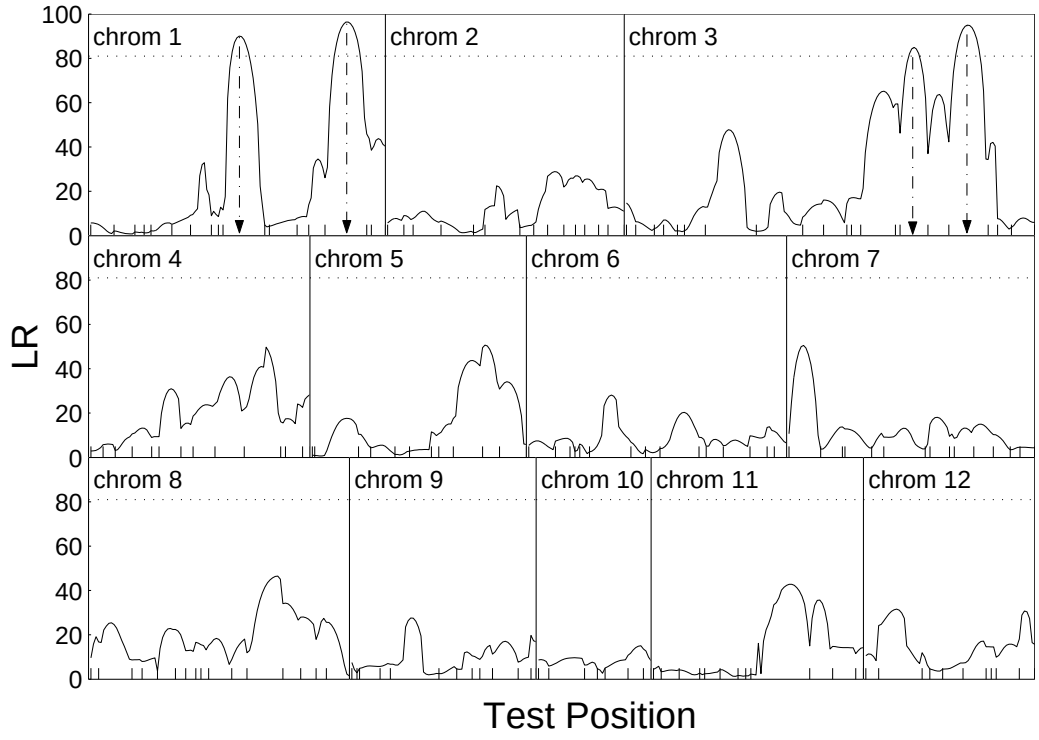

Figure 1: The profile plot of the log-likelihood ratios between the full (there is a QTL) and reduced model (there is no QTL) for tiller number trajectories across the 12 rice chromosomes. The genomic positions corresponding to the peak of the curve are the MLEs of the QTL location (indicated by the arrows). The threshold value for claiming the existence of QTLs is given as the horizonal dotted line for the 5% genome-wide level. The positions of markers on the linkage groups (Huang et al. 1997) are indicated at ticks.

for claiming the existence of QTLs at the genome-wide level is marked with the horizonal solid line, based on permutation tests.

As clearly shown by the genome-wide LR profile plot in Figure 1, the model detected 4 major QTLs that are all significant at the 5% genome-wide significance level, based on permutations. Table 3 tabulates the estimated QTL positions on the chromosome, the marker intervals of the QTLs, the MLEs of curve parameters that specify the developmental pattern, as well as the asymptotic standard errors of the estimators (in the parenthesis). Most parameters can be reasonably estimated with a small sampling error. As clearly indicated in the table, the LP orders at different test positions are not

the same. Two QTLs have order 7 (Q_{11} and Q_{31}) and two QTLs have order 8 (Q_{12} and Q_{32}).

The developmental trajectories of the identified QTLs are shown in Figure 2, with tiller number trajectories for all individuals indicated in the background. The four QTLs detected are found between marker RZ276 and RG146, denoted as CH1-1(Q_{11}), between markers RZ730 and RZ801, denoted as CH1-2 (Q_{12}) (both located on chromosome 1), between marker RZ337A and RZ448, denoted as Ch3-1 (Q_{31}) and between marker RZ519 and Pgi_1, denoted as Ch3-2 (Q_{32}) (both located on chromosome 3). Among these four QTLs detected, three of them show similar development patterns (Q_{11}, Q_{31} and Q_{32}), while QTL Q_{12} shows a different pattern. A statistical test based on hypotheses (11) shows that QTL Q_{12} only controls the growth phase ($P < 0.05$) and all the other three QTLs control the entire development process ($P < 0.05$ for both growth and death tests). Also, there is a genetic effect switch for QTL Q_{12} in which two genotypes switch their genetic effects, roughly at the end of the exponential growth phase. The goodness of fit of the polynomial to the data is assessed through one of the criteria summarized in Zheng (Zheng, 2000). The criterion is defined as

$$R^2 = 1 - \frac{\sum_{t=1}^{\tau}\sum_{n=1}^{n}(y_{it} - \hat{y}_{it})^2}{\sum_{t=1}^{\tau}\sum_{n=1}^{n}(y_{it} - \bar{y})^2} \tag{12}$$

where $\hat{y}_{it} = \pi_{1|i}m_1(t) + \pi_{0|i}m_0(t)$ and $\bar{y} = \frac{1}{n\tau}\sum_{t=1}^{\tau}\sum_{n=1}^{n} y_{it}$. The estimated R^2 values for the four detected QTLs are $0.69(Q_{11})$, $0.71(Q_{12})$, $0.73(Q_{31})$ and $0.79(Q_{32})$, which indicate adequate fit. At a particular time point, the goodness of fit is assessed through a similar measure defined as

$$R_t^2 = 1 - \frac{\sum_{n=1}^{n}(y_{it} - \hat{y}_{it})^2}{\sum_{n=1}^{n}(y_{it} - \bar{y})^2} \tag{13}$$

where t refers to the tth time point. We observe adequate fit of the polynomial to the data at early stages for the four QTLs. For example, the R_1^2 values are close to one and the R_6^2 values are greater than 0.6 for the four QTLs. However, we observe small R_t^2 value at the right tail, where $R_t^2 < 0.4$ for $t \geq 7$.

Since the genetic distance between the two QTLs detected on chromosome 3 are only $42cM$ apart, these two QTLs show signs of a weak linkage. If one QTL is the true one, the other one could be false positive caused by its linkage with the true positive. Based on one referee's suggestion, we did a linkage scan, using the parametric bootstrapped samples, to see whether the model can really separate the two linked QTLs. We first fixed the marker and linkage map information and then simulated bootstrapped samples assuming

Table 3: *The QTL location, MLEs of the estimated parameters and their asymptotic standard errors in the parentheses, with the SAD(1) covariance structure.*

Parameters	Q_{11}	Q_{12}	Q_{31}	Q_{32}
QTL position (λ)	$112cM$	$200cM$	$220cM$	$262cM$
Marker interval	$RZ146$-$RG345$	$RZ730$-$RZ801$	$RZ337A$-$RZ448$	$RZ519$-Pgi-1
Parameters for QQ				
u_{00}	10.108(0.25)	8.848(0.26)	8.933(0.19)	9.136(0.18)
u_{01}	1.426(0.25)	1.920(0.19)	1.542(0.16)	1.624(0.16)
u_{02}	-6.231(0.33)	-6.119(0.20)	-5.594(0.20)	-6.213(0.15)
u_{03}	2.472(0.22)	1.332(0.15)	2.039(0.13)	2.076(0.15)
u_{04}	0.719(0.27)	1.069(0.13)	0.855(0.16)	0.533(0.12)
u_{05}	-1.425(0.16)	-1.154(0.11)	-1.137(0.09)	-1.181(0.10)
u_{06}	1.341(0.17)	0.627(0.11)	0.919(0.10)	1.054(0.11)
u_{07}	0.973(0.21)	1.144(0.17)	0.695(0.14)	0.715(0.16)
u_{08}	-1.099(0.38)	-	-0.729(0.25)	-
Parameters for qq				
u_{20}	7.462(0.14)	7.517(0.19)	7.126(0.20)	6.982(0.19)
u_{21}	1.377(0.12)	1.055(0.16)	1.329(0.17)	1.289(0.17)
u_{22}	-4.431(0.17)	-4.690(0.14)	-3.927(0.22)	-4.343(0.16)
u_{23}	2.027(0.10)	2.718(0.13)	2.117(0.14)	2.118(0.17)
u_{24}	0.761(0.13)	0.054(0.10)	0.807(0.18)	0.421(0.13)
u_{25}	-0.717(0.07)	-0.774(0.09)	-0.548(0.10)	-0.572(0.11)
u_{26}	0.537(0.09)	0.929(0.10)	0.397(0.12)	0.547(0.12)
u_{27}	-0.061(0.09)	-0.444(0.14)	-0.425(0.15)	-0.372(0.17)
u_{28}	-0.543(0.20)	-	-0.768(0.28)	-
Covariance parameters				
σ^2	0.725(0.04)	0.748(0.04)	0.787(0.04)	0.789(0.04)
ϕ	0.848(0.02)	0.901(0.02)	0.884(0.02)	0.842(0.02)

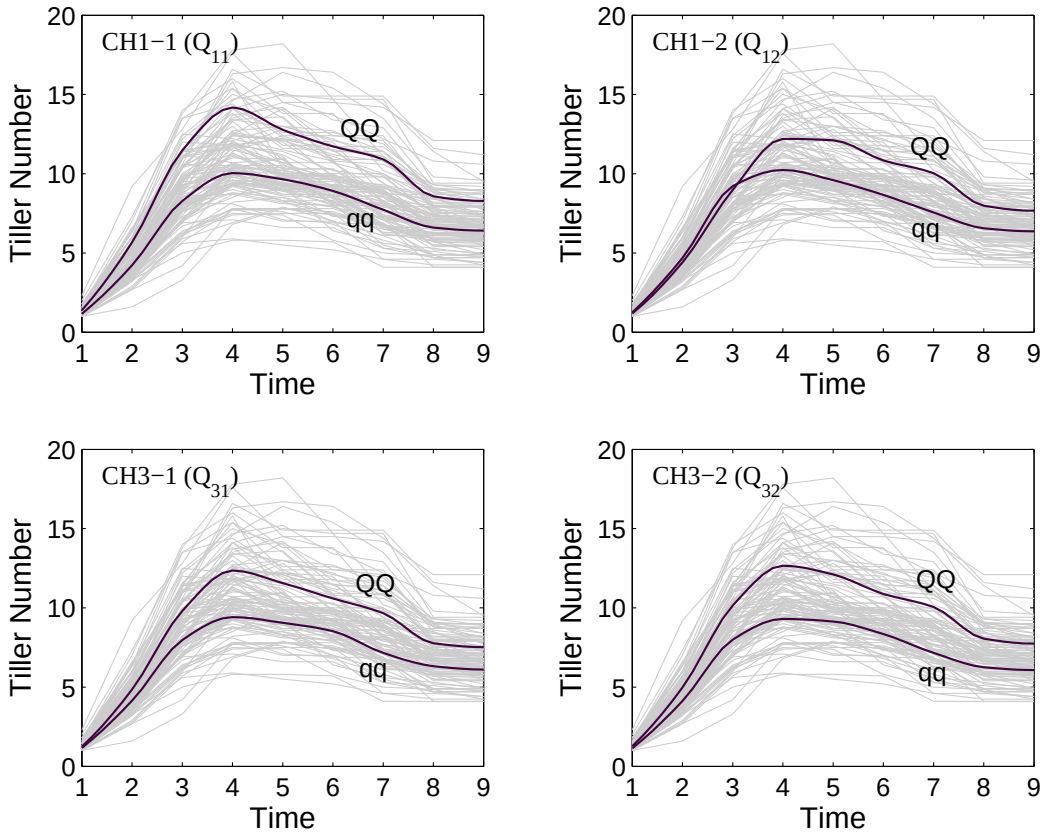

Figure 2: Two curves for the dynamic changes of tiller numbers, each representing one of the two groups of genotypes, QQ and qq, at each of the four significant QTLs. Tiller number trajectories for all observed individuals are indicated in the gray background.

there were two QTLs located at the positions indicated in Table 3. The bootstrapped samples were drawn from a multivariate normal distribution, with the mean and covariance variables tabulated in Table 3. Fig. 3A shows the averaged LR profile plot of 100 bootstrapped samples for chromosome 3. We observed consistent peaks for the two QTLs in the 100 samples, and the model has 100% power to detect these two QTLs. The highest LR peak, at the interval $RZ337A$-$RZ448$, coincides with the original QTL position, based on real data analysis. The highest peak at the interval $RZ519$-Pgi-1 is shifted a little to the left of the original QTL position (Q_{32}) partially due to the effect of the QTL located at the interval $RZ337A$-$RZ448$. Also seen in the figure, the linkage signals between the two intervals harboring the two QTLs are inflated,

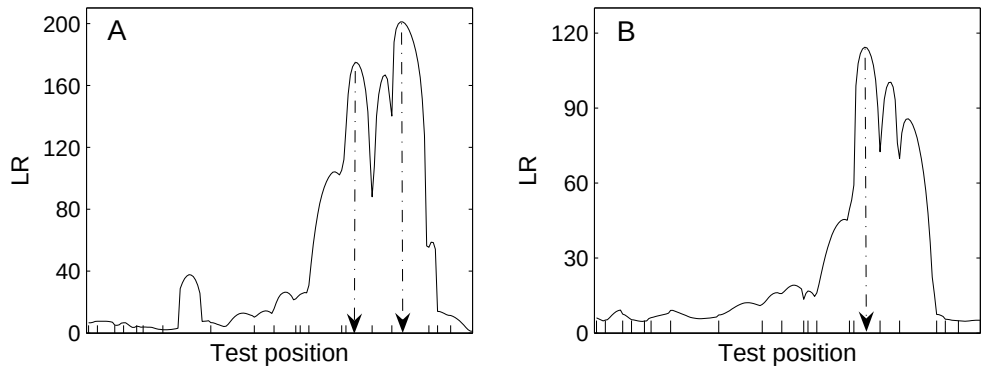

Figure 3: The averaged LR profile plot of 100 bootstrapped samples for chromosome 3. For each parametric bootstrapped sample, the marker information remains fixed, and the phenotype data are simulated assuming that there are two QTLs located at the intervals given in Table 3 (**A**) and that there is only one QTL located at the interval $RZ337A - RZ448$ (**B**). The genetic parameters for the two QTLs are tabulated in Table 3. The arrows indicate the estimated QTL positions.

due to linkage.

We also ran another parametric bootstrapping for a situation in which there is only one true positive; how likely are we to detect another one? Without a loss of generality, we assume the true one is located at the interval $RZ337A$-$RZ448$. Data were simulated assuming there is only one QTL; and the genetic and QTL location parameters are given in Table 3. Fig. 3B plots the averaged LR profile plot of 100 bootstrapped samples for chromosome 3. We can clearly see that the highest LR peak is always located at the interval where the true QTL is assumed to be. The linkage signal at the interval $RZ519$-Pgi-1 does not reach the significant level. The averaged QTL location estimate is $4cM$ left of the initial estimate listed in Table 3. This information indicates that the presence of other linked QTLs might potentially affect the tested QTL location estimation when the effect of a QTL located outside of the tested interval is not adjusted. In summary, information from the two bootstrap scenarios confirms that the two QTLs detected have a high probability of being two separate QTLs.

5 Discussion

The proper development of any organism is systematically maintained by the functional balance of cell growth and death. Proper cell death is an important requirement for the normal development of an organism. Meanwhile, inappropriate cell death may often lead to a variety of disorders or cancers (Vaux and Korsmeyer, 1999). Using genetic approaches, scientists have identified at least 10 genes associated with PCD in model organisms. However, a complete understanding of how this unique process is triggered by specific genes is still in its infancy. By integrating modern statistics and molecular techniques, the recently developed functional mapping approach provides an alternative quantitative platform for testing the interplay between gene actions and complex PCD processes (Ma et al., 2002; Cui et al., 2006; reviewed in Wu and Lin, 2006).

To fully enhance the flexibility of functional mapping for the PCD trait, we have extended the mapping approach for mapping QTLs responsible for longitudinal traits, by using the orthogonal Legendre function. We propose selecting the LP order under the alternative hypothesis and further propose a general information criterion to select the LP order and show its consistency property. The small sample properties, when applying AIC or BIC information criteria to select the optimal order under the functional mapping framework, are assessed through simulation studies. Since BIC has a stronger penalty term to the likelihood function, it likely favors more parsimonious models when compared with AIC, and this fits our modelling objective for a PCD curve. Simulation also confirms that BIC has a higher power than AIC when sample size is small. It might be expected that different genotypes could follow different development patterns, which may result in different LP orders. Order selection that assumes different orders for different genotypes, will need to be developed in the future.

As revealed by real data analysis, the current nonparametric approach and our previous semiparametric approach (Cui et al., 2006) only agree with two significant QTLs, namely Q_{11} and Q_{12}. Both QTLs reach the genome-wide significant level by using the current nonparametric approach, but only one (Q_{11}) reaches the genome-wide significant level by using the previous semiparametric approach. One of the QTLs detected on chromosome 9 by our previous semiparametric approach is not identified by the current approach. This QTL, however, also did not show significance in an analysis by Yan et al. (1998), which is consistent with the current finding. Moreover, the current model detects two QTLs located on chromosome 3 which were not identified by our previous semiparametric approach. Even though the two QTLs are weakly

linked, bootstrapping results confirm that they are likely two separate QTLs. These evidences demonstrate the power of the current approach, compared to the semiparametric approach, for this particular data set. Possible reasons for these differences might be directly related to different ways of modelling mean function. Further experimental evidence needs to be collected to reach a conclusion as to which model is more robust. Besides the effects of mean function on the power of functional mapping, covariance structure also plays a pivotal role in the precision and power of QTL detection. Modelling the covariance structure can be accomplished either parametrically or nonparametrically (Zimmerman and Núñez-Antón, 2001; Diggle and Verbyla, 1998; Kirkpatrick et al., 1994). However, it is difficult to compare which structure is optimal, given that simulating the true biological mechanisms is unrealistic due to our limited knowledge of the underlying true residual covariance structure. More studies on covariance structure modelling are desired.

For the rapidly changing values of a PCD curve to be approximated, an appropriate LP order needs to be selected. Intuitively, a higher order indicates a higher complexity to the developmental PCD process. Significant QTLs at different genomic locations may trigger different genetic effects on the PCD trait, which is revealed by the complex structure of the PCD curve. The curve complexity can then be described by the varying degrees of LP orders. Thus, the LP order reflects the functional complexity of the underlying QTL effects. Real data analysis indicates that the LP order does show different orders at different genomic regions (Table 3). For example, the two QTLs detected in chromosome 1 (Q_{11} and chromosome 3 Q_{31}) have an LP order of 7. If the order is selected under the null hypothesis as described by Lin and Wu (2006), the optimal order would be 6, using the BIC criterion. When the 6th order LP is applied to fit data under the alternative hypothesis, we observe a substantial likelihood reduction for the QTL in chromosome 3 (Q_{31}) due to potential mis-fitting. The likelihood ratio test statistic does not reach the genome-wide significant level. Thus, simply selecting the LP order under the null could lead to potential information loss and wrong inferences. Selecting the order under the alternative should be more informative. In general, the Legendre polynomial fits the data adequately as revealed by the trajectory plot in Fig. 2. The overall measure of the goodness of fit (R^2) defined in Eq. (12) shows that the polynomial fits the data well for the four QTLs. The poor fit of the polynomial to the right tail of the data, indicated by the measure defined in Eq. (13), may be due to the poor performance of the criterion itself. Since the mixture density does not directly lead to the mixture mean function, more study is needed to assess the goodness of fit in a nonlinear mixture model-based longitudinal regression setting.

It should be noted that the current nonparametric functional mapping approach does not consider the effects of background markers. This can be done by applying the composite interval mapping idea (Zeng, 1994) and integrating a simple multivariate multiple regression approach to select background markers as cofactors. Since the power of the functional mapping is to incorporate mathematical functions into a mapping framework by reducing the number of parameters to be estimated, using this marker selection approach without considering the functional curve information for a longitudinal or dynamic trait would reduce the power of the functional mapping. Moreover, preserving biologically relevant information for the responses during background marker selection presents great challenges to statistical modelling. Little research has been done for such an analysis; and modelling through multiple QTL by composite or multiple interval mapping should make the current mapping approach more useful in practice.

Our nonparametric approach using the Legendre function is built under the maximum likelihood-based functional mapping framework, and provides a testable quantitative platform for understanding the genetic basis of genes that account for quantitative variations of the PCD trait. The proposed framework is not restricted to the PCD trait. Any dynamic developmental process, whether or not it follows a particular parametric function, can be modelled and tested under the current framework.

APPENDIX A: DERIVATION OF EM ALGORITHM

The MLEs of the parameters contained in $\mathbf{\Omega} = (\mathbf{\Omega}_q, \mathbf{\Omega}_m, \mathbf{\Omega}_v)$ are derived as follows. Since we use a grid search algorithm by assuming known QTL positions, $\mathbf{\Omega}$ only contains two sets of parameters, i.e., $(\mathbf{\Omega}_m, \mathbf{\Omega}_v)$. The first derivative of the log-likelihood function, with respect to specific parameter φ contained in $\mathbf{\Omega}$, is given by

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{\Omega}_\varphi} \log \ell(\mathbf{\Omega}|\mathbf{y}, \mathcal{M}) &= \sum_{i=1}^{n} \sum_{j=0}^{1} \frac{\pi_{j|i} \frac{\partial}{\partial \mathbf{\Omega}_\varphi} f_j(\mathbf{y_i}|\mathbf{\Omega}, \mathcal{M})}{\sum_{j=0}^{1} \pi_{j|i} f_j(\mathbf{y_i}|\mathbf{\Omega}, \mathcal{M})} \\
&= \sum_{i=1}^{n} \sum_{j=0}^{1} \frac{\pi_{j|i} f_j(\mathbf{y_i}|\mathbf{\Omega}, \mathcal{M})}{\sum_{j'=0}^{1} \pi_{j'|i} f_{j'}(\mathbf{y_i}|\mathbf{\Omega}, \mathcal{M})} \frac{\partial}{\partial \mathbf{\Omega}_\psi} \log f_j(\mathbf{y_i}|\mathbf{\Omega}, \mathcal{M}) \\
&= \sum_{i=1}^{n} \sum_{j=0}^{1} \Pi_{j|i} \frac{\partial}{\partial \mathbf{\Omega}_\varphi} \log f_j(\mathbf{y_i}|\mathbf{\Omega}, \mathcal{M})
\end{aligned}$$

where we define

$$\Pi_{j|i} = \frac{\pi_{j|i} f_j(\mathbf{y_i}|\mathbf{\Omega}, \mathcal{M})}{\sum_{j'=0}^{1} \pi_{j'|i} f_{j'}(\mathbf{y_i}|\mathbf{\Omega}, \mathcal{M})} \tag{A1}$$

The MLEs of the parameters contained in $(\mathbf{\Omega}_m, \mathbf{\Omega}_v)$ are obtained by solving

$$\frac{\partial}{\partial \mathbf{\Omega}_\varphi} \log \ell(\mathbf{\Omega}|\mathbf{y}, \mathcal{M}) = 0 \tag{A2}$$

Direct estimation is unavailable since there is no closed form for the MLEs of parameters. The EM algorithm is applied to solve these unknowns iteratively.

Define

$$\mathbf{X} = \begin{bmatrix} P_0(t'_1) & P_1(t'_1) & \cdots & P_r(t'_1) \\ P_0(t'_2) & P_1(t'_2) & \cdots & P_r(t'_2) \\ \vdots & \cdots & \cdots & \vdots \\ P_0(t'_\tau) & P_1(t'_\tau) & \cdots & P_r(t'_\tau) \end{bmatrix} \quad \text{and} \quad \boldsymbol{u}_j = \begin{bmatrix} u_{0j} \\ u_{1j} \\ \vdots \\ u_{rj} \end{bmatrix} \tag{A3}$$

where $\mathbf{X}$ is a matrix of the LP base function with order r and $\boldsymbol{u}_j$ is the *base genotypic vector* for genotype j, which contains the mean parameters to be estimated. We then have the mean vector $\mathbf{m}_j = \mathbf{X}\boldsymbol{u}_j$.

For the SAD(1) covariance structure given in Eq. (9) with constant innovation variance $\sigma_t^2 = \sigma^2$, we have the following properties

(1) $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2} \mathrm{L}^T \mathrm{L}$

(2) $|\mathbf{\Sigma}| = (\sigma^2)^\tau$

(3) $(\mathbf{y}_i - \mathbf{m}_j)^T \mathbf{\Sigma}^{-1} (\mathbf{y}_i - \mathbf{m}_j)$

$$\begin{aligned} &= \frac{1}{\sigma^2} (\mathbf{y}_i - \mathbf{m}_j)^T \mathbf{\Gamma}(\phi) (\mathbf{y}_i - \mathbf{m}_j) \\ &= \frac{1}{\sigma^2} \Big\{ -2\phi \sum_{s=2}^{\tau-1} \Big[y_i(t_s) - m_j(t_s) \Big] \Big[y_i(t_{s+1}) - m_j(t_{s+1}) \Big] \\ &\quad + (\phi^2 + 1) \sum_{s=1}^{\tau-1} \Big[y_i(t_s) - m_j(t_s) \Big]^2 - \big[y_i(t_\tau) - m_j(t_\tau) \big]^2 \Big\} \end{aligned}$$

where L is a lower triangular matrix with 1s on the diagonal and with the

negative of the antedependence coefficient ϕ as below diagonal entries.

$$\mathrm{L} = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ -\phi & 1 & 0 & 0 & \cdots & 0 \\ 0 & -\phi & 1 & 0 & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & -\phi & 1 & 0 \\ 0 & \cdots & \cdots & 0 & -\phi & 1 \end{bmatrix}$$

and

$$\mathbf{\Gamma}(\phi) = \begin{bmatrix} \phi^2+1 & -\phi & 0 & \cdots & 0 \\ -\phi & \phi^2+1 & -\phi & \cdots & 0 \\ 0 & -\phi & \phi^2+1 & -\phi & \vdots \\ \vdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & -\phi & \phi^2+1 & -\phi \\ 0 & \cdots & 0 & -\phi & 1 \end{bmatrix}$$

In solving Eq. A2 with respect to the unknowns to get the MLEs of the unknown parameters, we have

$$\hat{\boldsymbol{u}}_j = \frac{\sum_{i=1}^{n} \Pi_{j|i} \mathbf{X}^T \mathbf{\Sigma}^{-1} \mathbf{y}_i}{\sum_{i=1}^{n} \Pi_{j|i} \mathbf{X}^T \Sigma^{-1} \mathbf{X}} = \frac{\sum_{i=1}^{n} \Pi_{j|i} \mathbf{X}^T \mathrm{L}^T \mathrm{L} \mathbf{y}_i}{\sum_{i=1}^{n} \Pi_{j|i} \mathbf{X}^T \mathrm{L}^T \mathrm{L} \mathbf{X}} \tag{A4}$$

and the MLEs of the covariance parameters are given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \sum_{j=0}^{1} \Pi_{j|i} (\mathbf{y}_i - \mathbf{X}\hat{\mathbf{u}}_j)^T \Gamma(\hat{\phi}) (\mathbf{y}_i - \mathbf{X}\hat{\boldsymbol{u}}_j)}{n\tau} \tag{A5}$$

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \sum_{j=0}^{1} \Pi_{j|i} \sum_{s=1}^{\tau-1} [y_i(t_s) - \mathbf{X}_s^T \hat{\boldsymbol{u}}_j][y_i(t_{s+1}) - \mathbf{X}_{s+1}^T \hat{\boldsymbol{u}}_j]}{\sum_{i=1}^{n} \sum_{j=0}^{1} \Pi_{j|i} \sum_{s=1}^{\tau} [(y_i(t_s) - \mathbf{X}_s^T \hat{\boldsymbol{u}}_j]^2} \tag{A6}$$

where $\mathbf{X}_s$ and $\mathbf{X}_{s+1}$ are the (s)th and $(s+1)$th row of the design matrix $\mathbf{X}$.
E-step: Given initial values for $(\mathbf{\Omega}_m, \mathbf{\Omega}_v)$, calculate the posterior probability matrix $\mathbf{\Pi} = \{\Pi_{j|i}\}$ in Eq. (A1).

M-step: With the updated posterior probability $\mathbf{\Pi}$, we can update the parameters $(\boldsymbol{u}_1, \boldsymbol{u}_0, \sigma^2, \phi)$ The above procedures are iteratively repeated between (A1) and (A4) - (A6), until a certain convergence criterion is met. The converged values are the MLEs of the parameters.

APPENDIX B: CONSISTENCY OF THE MODEL SELECTION CRITERION

Background

Under the alternative given in hypotheses (10), there are two genotypes in a backcross design. The mean vector for genotype j can be expressed as $\mathbf{m}_j = \mathbf{X}\boldsymbol{u}_j, j = 0, 1$. The log-likelihood function can be expressed as

$$\ell(\mathbf{\Omega}|\mathbf{y}, \mathcal{M}) = \sum_{i=1}^{n} log\{\pi_{1|i} f_1(\mathbf{y}_i|\mathbf{u_1}, \mathbf{\Sigma}, \mathcal{M}) + \pi_{0|i} f_0(\mathbf{y}_i|\mathbf{u_0}, \mathbf{\Sigma}, \mathcal{M})\} \qquad \text{(B1)}$$

where $\pi_{1|i}$ and $\pi_{0|i}$ are the mixture proportions; $f_j(y_i)$ is the density function for genotype j, which has the form given in Eq. (2)

The EM algorithm can be applied to estimate the parameters $\mathbf{\Omega} = (\mathbf{u}_1, \mathbf{u}_0, \sigma^2, \phi)$, as shown in Apendix A in details. The MLEs of the mean and covariance parameters are consistent under the assumption that the covariance matrix is the same for both mixture components (Redner 1981; Pourahmadi 2000).

Based on the segregation principle, there are two QTL genotypes at a particular test position for a backcross progeny. Different combinations of flanking markers form 4 groups, say $M_\eta M_\eta M_{\eta+1} M_{\eta+1}(\mathcal{M}_1)$, $M_\eta M_\eta M_{\eta+1} m_{\eta+1}(\mathcal{M}_2)$, $M_\eta m_\eta M_{\eta+1} M_{\eta+1}(\mathcal{M}_3)$, $M_\eta m_\eta M_{\eta+1} m_{\eta+1}(\mathcal{M}_4)$. The longitudinal PCD traits can be grouped as $\mathbf{y}_k$ corresponding to marker group $\mathcal{M}_k$, with each group containing n_k $(k = 1, \cdots, 4)$ observations. Now define $f(\mathbf{y}_{k,i}|\mathbf{\Omega}, \mathcal{M}_\mathbf{k}) = \pi^k_{1|i} f_1(\mathbf{y}_{k,i}|\mathbf{\Omega}, \mathcal{M}_\mathbf{k}) + \pi^k_{0|i} f_0(\mathbf{y}_i|\mathbf{\Omega}, \mathcal{M}_\mathbf{k})$ and $\pi^k_{1|i} + \pi^k_{0|i} = 1$, for $k = 1, \cdots, 4$ and $i = 1, \cdots, n$, where $\pi^k_{j|i}$ is the mixture proportion for individual i in group k with genotype j, and f_j is the multivariate density function for genotype j with parameters contained in $\mathbf{\Omega}$. A further partition of the log-likelihood function given in Eq. (B1) leads to

$$\begin{aligned} \ell(\mathbf{\Omega}|\mathbf{y}, \mathcal{M}) &= \sum_{i=1}^{n} log f(\mathbf{y}_i|\mathbf{\Omega}, \mathcal{M}) \\ &= \sum_{k=1}^{4} \sum_{i=1}^{n_k} log f(\mathbf{y}_{k,i}|\mathbf{\Omega}, \mathcal{M}_k) \end{aligned}$$

Order Selection Criterion

Under the current study's design, we aim at fitting the developmental PCD curve using the LP function with an optimal order r that best explains the variations of the dynamic PCD process. Our selection goal is to choose an optimal order that has great flexibility so as to capture the developmental PCD pattern, yet is parsimonious enough for modelling purposes. We propose using the following information criterion

$$IC_r(n) = -2\ell(\hat{\theta}_r|y) + c(n)p_r(\tau)$$

where $IC_r(n)$ is the information for the model, with an LP order r; $\hat{\theta}_r = (\widehat{\mathbf{\Omega}}_m, \widehat{\mathbf{\Omega}}_v) = (\hat{\boldsymbol{u}}_{1|r}, \hat{\boldsymbol{u}}_{0|r}, \hat{\sigma}_r^2, \hat{\phi}_r)$; $c(n)$ is a penalty term; $p_r(\tau)$ is the number of free parameters for the selected model.

Consistency of the Selection Criterion

Theorem: *Under the current mapping framework, a model selection criterion $IC_r(n)$ is consistent if*

$$\frac{c(n)[p_{r_0}(\tau) - p_r(\tau)]}{n} \to 0, \text{ as } n \to \infty$$

where r_0 and r represents the optimal and selected LP order, respectively.

The following arguments are based on the assumption that MLEs converge and are consistent.

Proof: To show the consistency of the selection criterion, we need to show that

$$\lim_{n\to\infty} p[IC_{r_0}(n) < IC_r(n)] = 1$$

Let θ^* be the parameters of the true model; $\hat{\theta}_r$ and $\hat{\theta}_{r_0}$ be the MLEs under model m_r and m_{r_0}, respectively. Then

$$\begin{aligned} P[IC_{r_0}(n) < IC_r(n)] &= P[2\ell(\hat{\theta}_r) - c(n)p_r(\tau) < 2\ell(\hat{\theta}_{r_0}) - c(n)p_{r_0}(\tau)] \\ &= P[2\ell(\hat{\theta}_r) - 2\ell(\hat{\theta}_{r_0}) - c(n)(p_r(\tau) - p_{r_0}(\tau)) < 0] \\ &= P[2\frac{\ell(\hat{\theta}_r) - \ell(\hat{\theta}_{r_0})}{n} + \frac{c(n)[p_{r_0}(\tau) - p_r(\tau)]}{n} < 0] \end{aligned}$$

By the consistency of the MLEs, we have $\hat{\theta}_{r_0} \to \theta^*_{r_0}$ and $\hat{\theta}_r \to \theta^*_r$, where $\theta^*_{r_0}$ and θ^*_r are the closest points in the parameter space of model m_{r_0} and m_r, to

the true parameter θ^*. Then we have

$$\begin{aligned}\frac{\ell(\hat{\theta}_r)-\ell(\hat{\theta}_{r_0})}{n} &= \frac{1}{n}\sum_{k=1}^{4}\sum_{i=1}^{n_k}\{logf(\mathbf{y}_{k,i}|\hat{\theta}_r,\mathcal{M}_k)-logf(\mathbf{y}_{k,i}|\hat{\theta}_{r_0},\mathcal{M}_k)\}\\ &\to \frac{1}{n}\sum_{k=1}^{n_k}\sum_{i=1}^{n_k}\{logf(\mathbf{y}_{k,i}|\theta^*_r,\mathcal{M}_k)-logf(\mathbf{y}_{k,i}|\hat{\theta}_{r_0},\mathcal{M}_k)\}\\ &= \frac{\ell(\theta^*_r)-\ell(\theta^*_{r_0})}{n}\end{aligned}$$

By the weak law of large numbers (WLLN), when $P_{\theta_{r_0}}$ is the true probability measure

$$\begin{aligned}&\frac{1}{n_k}\{\sum_{i=1}^{n_k}\{logf(\mathbf{y}_{k,i}|\hat{\theta}_r,\mathcal{M}_k)-logf(\mathbf{y}_{k,i}|\hat{\theta}_{r_0},\mathcal{M}_k)\}\\ &\to E_{\theta_{r_0}}[logf(\mathbf{y}_{k,1}|\theta^*_r,\mathcal{M}_k)-logf(\mathbf{y}_{k,1}|\theta^*_{r_0},\mathcal{M}_k)]\\ &= E_{\theta_{r_0}}[log\frac{f(\mathbf{y}_{k,1}|\theta^*_r,\mathcal{M}_k)}{f(\mathbf{y}_{k,1}|\theta^*_{r_0},\mathcal{M}_k)}],\ \text{as } n_k\to\infty\end{aligned}$$

with $P_{\theta_{r_0}}[f(\mathbf{y}_{k,1}|\theta^*_r,\mathcal{M}_k)=f(\mathbf{y}_{k,1}|\theta^*_{r_0},\mathcal{M}_k)]<1,\ \forall\ r_0\neq r$, and $0<\frac{n_k}{n}=\mathcal{O}(1)<1,\ \forall\ k=1,\cdots,4$

Therefore,

$$\begin{aligned}\frac{1}{n}\nabla\ell(n) &\stackrel{def}{=} \frac{1}{n}\sum_{k=1}^{4}\sum_{i=1}^{n_k}\{logf(\mathbf{y}_{k,i}|\hat{\theta}_r,\mathcal{M}_k)-logf(\mathbf{y}_{k,i}|\hat{\theta}_{r_0},\mathcal{M}_k)\}\\ &= \sum_{k=1}^{4}(\frac{n_k}{n})\frac{1}{n_k}\sum_{i=1}^{n_k}\{logf(\mathbf{y}_{k,i}|\hat{\theta}_r,\mathcal{M}_k)-logf(\mathbf{y}_{k,i}|\hat{\theta}_{r_0},\mathcal{M}_k)\}\\ &\to \sum_{k=1}^{4}\mathcal{O}(1)\{E_{\theta_{r_0}}[logf(\mathbf{y}_{k,1}|\theta^*_r,\mathcal{M}_k)-logf(\mathbf{y}_{k,1}|\theta^*_{r_0},\mathcal{M}_k)]\},\text{as } n_k\to\infty\\ &= \sum_{k=1}^{4}\mathcal{O}(1)E_{\theta_{r_0}}[log\frac{f(\mathbf{y}_{k,1}|\theta^*_r,\mathcal{M}_k)}{f(\mathbf{y}_{k,1}|\theta^*_{r_0},\mathcal{M}_k)}]\end{aligned}$$

Now using the fact that $log(\frac{1}{x})\leq\frac{1}{x}-1\ \forall\ x>0$ with equality iff $x=1$, putting $x=f(\mathbf{y}_{k,1}|\theta^*_{r_0},\mathcal{M}_k)/f(\mathbf{y}_{k,1}|\theta^*_r,\mathcal{M}_k)$ and then taking expectation, one gets

$$E_{\theta_{r_0}}[log\frac{f(\mathbf{y}_{k,1}|\theta^*_r,\mathcal{M}_k)}{f(\mathbf{y}_{k,1}|\theta^*_{r_0},\mathcal{M}_k)}]\leq E_{\theta_{r_0}}[\frac{f(\mathbf{y}_{k,1}|\theta^*_r,\mathcal{M}_k)}{f(\mathbf{y}_{k,1}|\theta^*_{r_0},\mathcal{M}_k)}]-1\leq 1-1=0,\ \forall\, k=1,\cdots,4$$

Hence,

$$\begin{aligned} & P[IC_{r_0}(n) < IC_r(n)] \\ = \; & P\{2\frac{\ell(\hat{\theta}_r) - \ell(\hat{\theta}_{r_0})}{n} + \frac{c(n)[p_{r_0}(\tau) - p_r(\tau)]}{n} < 0\} \\ = \; & P\{\frac{2}{n}\nabla\ell(n) + \frac{c(n)[p_{r_0}(\tau) - p_r(\tau)]}{n} < 0\} \\ \to \; & 1 \end{aligned}$$

if

$$\frac{c(n)[p_r(\tau) - p_{r_0}(\tau)]}{n} \to 0 \text{ as } n \to \infty.$$

Since $p_r(\tau) - p_{r_0}(\tau)$ is finite, $\frac{c(n)[p_r(\tau)-p_{r_0}(\tau)]}{n} \to 0$ as long as $\frac{c(n)}{n} \to 0$. So both AIC and BIC are consistent in this case.

References

Altman, N. S. (1990) Kernel smoothing of data with correlated errors. *J Am Stat Assoc*, **90**, 508-515.

Altman, N.S. and Casella, G. (1995) Nonparametric empirical Bayes growth curve analysis. *J Am Stat Assoc*, **90**, 508-515.

Ameisen, J. C. (2002) On the origin, evolution, and nature of programmed cell death: a timeline of four billion years. *Cell Death Differentiation*, **9**, 367-393.

Boularan, J., Ferre, L., and Vieu, P. (1994) Growth curves: a two-stage nonparametric approach. *J Stat Plan Infer*, **38**, 327-350.

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

Cui, Y., Zhu, J., and Wu, R. (2006) Functional Mapping for Genetic Control of Programmed Cell Death. *Physiol Genomics*, **25**, 458-469.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplet data via the EM algorithm. *J R Statist Soc* B, **39**, 1-38

Diggle, P. J. and Verbyla, A. P. (1998) Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, **54**, 401-415.

Ellis, H. M. and Horvitz, H. R. (1986) Genetic control of programmed cell death in the nematode *C. elegans*. *Cell*, **44**, 817-29.

Ellis, R. E., Yuan, J. Y., and Horvitz, H. R. (1991) Mechanisms and functions of cell death. *Ann Rev Cell Biol*, **7**, 663-698.

Ferreira, E., Núñez-Antón, V., and Rondriguez-Poo, J. (1997) Kernel regression estimates of growth curves using nonstationary correlated errors. *Stat Prob Let*, **34**, 413-423.

Fraiman R. and Meloche, J. (1994) Smoothing dependent observations. *Stat Prob Let*, **2**, 203-214.

Greenberg, J. T. (1996) Programmed cell death: A way of life for plants. *Proc Natl Acad Sci*, **93**, 12094-12097.

Hart, J. D. and Wehrly, T. E. (1986) Kernal regression estimation using repeated measurements data. *J Am Stat Assoc*, **81**, 1080-1088.

Hanahan, D. and Weinberg, R. A. (2000) The hallmarks of cancer. *cell*, **100**, 57-70.

Jacobson, M. D., Weil, M., and Raff, M. C. (1997) Programmed cell death in animal development. *Cell*, **88**, 347-354.

Horvitz, H.R. (1999) Genetic control of programmed cell death in the nematode *Caenorhabditis elegans*. *Cancer Res*, **59**, 1701-1706.

Horvitz, H.R. (2003) Worms, life, and death (Nobel Lecture). *Chembiochem*, **4**, 697-711.

Huang, N., Parco, A., Mew, T., Magpantay, G., McCouch, S., Guiderdoni, E., Xu, J., Subudhi, P., Angeles, E. R., Khush, G. S. (1997) RFLP mapping of isozymes, RAPD and QTL for grain shape, brown planthopper resistance in a doubled haploid rice population. *Mol Breeding*, **3**, 105-113.

Jacobson, M. D., Weil, M., and Raff, M. C. (1997) Programmed cell death in animal development. *Cell*, **88**, 347-354.

Jaffrézic, F., Thompson, R., and Hill, W. G. (2003) Structured antedependence models for genetic analysis of repeated measures on multiple quantitative traits. *Genet Res*, **82**, 55-65.

Kirkpatrick, M., Hill, W. G., and Thompson, R. (1994) Estimating the covariance structure of traits during growth and aging, illustrated with lactation in dairy cattle. *Genet Res*, **64**, 57-69.

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

Liu, T., Zhao, W., Tian, L. L., and Wu, R. L. (2004) An algorithm for molecular dissection of tumor progression. *J Math Biol*, **50**, 336-354.

Lin, M., Aquilante, C., Johnson, J. A., and Wu, R. (2005) Sequencing drug response with HapMap. *Pharm J*, **5**, 149-156.

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

Lynch, M. and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

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.

Martin, K.R. (2006) Targeting apoptosis with dietary biloactive agents. *Exp Biol Med*, **231**, 117-129.

Müller, H. G. (1988) *Nonparametric Analysis of Longitudinal Data*. Springer-Verlag, Berlin.

Pennell, R. I. and Lamb,C. (1997) Programmed cell death in plants. *Plant Cell*, **9**, 1157-1168.

Pool, M. H., Janss, L. L. G., and Meuwissen, T. H. E. (2000a) Genetic Parameters of Legendre Polynomials for First Parity Lactation Curves. *J Dairy Sci*, **83**, 2640-2649.

Pool, M. H. and Meuwissen, T. H. E. (2000b) Prediction of daily milk yields from a limited number of test days using test day models. *J Dairy Sci*, **82**, 1555-1564.

Pourahmadi, M. (2000) Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika*, **87**,425-435.

Redner, R. (1981) Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann Statist*, **9**, 225-228.

Vaux, D. L. and Stanley, J. K. (1999) Cell Death in Development. *Cell*, **96**, 245-254.

Wang, Z. H. and Wu, R. L. (2004) A statistical model for high-resolution mapping of quantitative trait loci determining human HIV-1 dynamics. *Stat Med*, **23**, 3033-3051.

Wu, R. L. and Lin, M. (2006) Functional mapping – How to map and study the genetic architecture of dynamic complex traits. *Nat Rev Genet*, **7**, 229-237.

Wu, R. L., Ma, C.-X., Littell, R. C., and Casella, G. (2002) A statistical model for the genetic origin of allometric scaling laws in biology. *J Theo Biol*, **219**, 121-135.

Wu, R. L., Ma, C. -X., Lin, M. and Casella, G. (2004) A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics*, **166**, 1541-1551.

Yan, J. Q., Zhu, J., He, C.X., Benmoussa, M., and Wu, P. (1998) Quantitative trait loci analysis for the developmental behavior of tiller number in rice. *Theor Appl Genet*, **97**, 267-274.

Yuan, J. and Horvitz, H. R. (2004) A first insight into the molecular mechanisms of apoptosis. *Cell*, **S116**, 53-56.

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

Zhao, W., Chen, Y.Q., Casella, G., Cheverud, J.M., and Wu, R.L. (2005) A non-stationary model for functional mapping of complex traits. *Bioinformatics*, **21**, 2469-2477.

Zheng, B. (2000) Summarizing the goodness of fit of generalized linear models for longitudinal data. *Stat Med*, **19**, 1265-1275.

Zimmerman, D. L. and Núñez-Antón, V. (2001) Parametric modelling of growth curve data: An overview. *Test*, **10**, 1-73.