Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Wang et al., J Biom Biostat 2017, 8:1
DOI: 10.4172/2155-6180.1000329

Research Article Open Access

Fast Computation of Significance Threshold in QTL Mapping of Dynamic Quantitative Traits

Nating Wang[1], Hongxiao Tian[1], Yongci Li[1], Rongling Wu[2], Jiangtao Luo[3] and Zhong Wang[2]*

[1]School of Science, Beijing Forestry University, Beijing 100083, China
[2]College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, PR China
[3]Department of Biostatistics, College of Public Health, University of Nebraska Medical Center, Omaha, NE 68198-4375, USA

Abstract

Functional Mapping is a popular statistical method in QTL mapping studies for longitudinal data. The threshold for declaring statistical significance of a QTL is commonly obtained through permutation tests, which can be time consuming. To improve the computational efficiency of a permutation test of mixture models used in Functional Mapping, we first quantified the correlation between QTL and longitudinal data, using a curve clustering method. Then, the QTLs which are highly correlated with the outcome were computed in the improved permutation tests. As a result, it reduces the amount of computation in permutation tests and speeds up the computation for Functional Mapping analysis. Simulation studies and real data analysis were conducted to demonstrate that the proposed approach can greatly improve the computational efficiency of QTL mapping without loss of accuracy.

Keywords: Quantitative trait locus; QTL mapping; Functional mapping; Permutation test; Log-likelihood ratio test

Introduction

High-throughput sequencing of genomes in studies of human and model organism has made rapid progress in recent years. Gene mapping based on QTL identification plays an important role in the traditional agriculture and forestry fields. Research on QTL mapping has made much advancement, such as detecting QTL related to rice flower development in recombinant inbred lines (RIL) [1], barley height [2], maize genome controlling roots [3], disease resistance and fiber quality of cotton [4], and anti-thrips ability of pepper [5]. Moreover, the researches on crop by QTL mapping methods has developed from phenotypic effects to gene expression level, and from static gene mapping in a number of stages during development to dynamic QTL mapping in the whole growth process [6,7].

Researchers have constructed a series of different QTL mapping methods with a combination of biological and statistical knowledge, which allows us to conduct various hypothesis tests to understand biological mechanisms of traits controlled by genes. Gene mapping of quantitative traits has also been extended to longitudinal data measured at multiple time points [8-10]. Function mapping, a framework of gene mapping of dynamic quantitative traits [11-16], has been applied to various contexts related to biological growth curves and allometric growth [15]. It has been used to study pleiotropy [15], gene-gene interactions [16] and gene-environment interactions [16]. Functional Mapping has been thoroughly formulated as a system, and made great advancements in genetics research on agriculture and medicine.

Functional Mapping is based on mixture models, and likelihood ratio (LR) statistic is used to test different hypotheses on genetic models. Because it is difficult to determine the distribution under null hypothesis and satisfy the conditions for approximation of chi-squared distribution, statistical literature recommends using permutation tests [17] to determine a threshold under a certain significance level. Permutation tests are time consuming and cannot use for genome-wide study in practice. We propose a new computation method for permutation tests in Functional Mapping. It generates a threshold very close to that from permutation tests, but its computational time is 1/10 of the original time, sometimes even cuts to 1/20. We first partition samples into several groups basing on the trajectory of the phenotypical values using a curve clustering method. The probability that an individual belongs to each group can be calculated. The clustering probability matrix is constructed according to each individual's probability. Then, we compute the correlation between the clustering probability matrix and QTL genetic probability matrix defined by the genetic recombination rate. Note that QTLs with stronger correlation are more significant in QTL mapping. Lastly, we pick the QTLs with stronger correlation to calculate their LR values in the final permutation test, which will reduce the computation burden in Functional Mapping.

The paper is organized as following. We first briefly introduce the principle of Functional Mapping. Then we discuss our algorithm, which includes three steps. The first step is to obtain the clustering probability matrix by using curve clustering method, the second is to define different correlation calculation methods to quantify the correlation between clustering probability matrix and genetic probability matrix, and the third one is to apply traditional permutation test to the selected QTLs with high correlation. Finally, we present simulation studies and real data analysis to demonstrate our method.

Statistical Model

Functional mapping for dynamic quantitative traits

Functional Mapping can operate with any biological growth curve and covariance matrix describing correlation among the observed phenotypic values. We here use logistic growth curve as an example to describe its fundamental concept. Under the control of QTL (or gene),

***Corresponding author:** Wang Z, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, PR China; Tel: +86 10 6233 1279; E-mail: zhongwang@bjfu.edu.cn

Received December 05, 2016; **Accepted** January 03, 2017; **Published** January 09, 2017

Citation: Wang N, Tian H, Li Y, Wu R, Luo J, et al. (2017) Fast Computation of Significance Threshold in QTL Mapping of Dynamic Quantitative Traits. J Biom Biostat 8: 329. doi:10.4172/2155-6180.1000329

growth trajectories are different across QTL genotypes. For example, the growth curve can be described in the following form:

$$g_j(t) = \frac{a_j}{1 + b_j * e^{-r_j * t}} \quad (1)$$

where g_j is the phenotypic values of genotype j. In F2 breeding, we have QQ, Qq, and qq three types, which are denoted by 1, 2, and 3, respectively. is growth limit for genetype j, b_j is the initial balance parameter, and r_j is the relative growth rate [11].

Suppose that the growth of subject i follows the trajectory of genotype j, and the observed growth vector at y_i multiple time points follows multivariate normal distribution then its probability density function is:

$$f_j(y_i) = \frac{1}{(2\pi)^{m/2}|\sum|^{1/2}} \exp\left[-(y_i - g_j)^T \sum^{-1} (y_i - g_j)/2\right] \quad (2)$$

Assuming that y_i is a vector of the phenotypic values at m time points; g_j denotes the overall mean vector for genotype j; Σ is the variance-covariance matrix. We here use the first-order autoregressive model [AR(1)] as our example. In the AR(1) model, the variance-covariance matrix is:

$$\sum = \begin{vmatrix} 1 & \rho & \dots & \rho^{m-1} \\ \rho & 1 & \dots & \rho^{m-2} \\ \dots & \dots & \dots & \dots \\ \rho^{m-1} & \rho^{m-2} & \dots & 1 \end{vmatrix} \quad (3)$$

Where σ^2 is the variance and ρ is the correlation. In order to test whether a QTL affects the growth trajectories, likelihood function is based on the following mixture model,

$$L(\Omega) = \prod_{i=1}^{N}\left[\sum_{j=1}^{3} p_{ij} f_j(y_i)\right]. \quad (4)$$

Where Ω is the set of the parameters for the growth curves of different genes and for variance-covariance matrix (Σ). The genotype at the QTL of subject i is not directly observed, but we can calculate a probability matrix by using its distances to the markers at both ends, where each element p_{ij} in the matrix is the genetic probability of subject i and gene j. Please see [13] for details of calculations.

The purpose of Functional Mapping is to utilize the curves from biological profiles (e.g., growth curves) to compute log likelihood ratio statistic for each QTL, and then choose the significant QTL according to the values of LR, which is defined by (6). The null hypothesis H_0 is that there is no gene that controls the growth process and the alternative hypothesis is the growth processes are different across the QTL genotypes, specifically,

$$\begin{cases} H_0: a_1 = a_2 = a_3, b_1 = b_2 = b_3, r_1 = r_2 = r_3 \\ H_1: \textit{ at least one of the equalities above does not hold} \end{cases} \quad (5)$$

The test statistic for testing the hypothesis (5) is the log-likelihood ratio (LR) test.

$$LR = -2\log\left[\frac{L(\tilde{\Omega})}{L(\hat{\Omega})}\right] \quad (6)$$

Where $\tilde{\Omega}$ and $\hat{\Omega}$ are the maximum likelihood estimates of parameters under the hypothesis H_0 and H_1 respectively. However, it is difficult to determine the distribution of the log-likelihood ratio in statistics; therefore, permutation test is a feasible approach to access the threshold at different significance levels. At least 100 permutation tests have to be done for the rough threshold at the 5% confidence level, generally 1000 tests are recommended for more accurate threshold. That means 1000 QTL scanning process in the whole molecular markers scope, which requires a large number of CPU computing.

Genotype-oriented curve clustering

To improve the efficiency of permutation test, we design a method for selecting QTL and we only choose highly correlated QTLs to compute LR at each permutation. Our method filters out the points with lower correlations between QTL and phenotypic values. We call it the filtering method. The filtering method of the QTL is constructed such that only QTLs highly relevant to the character remain to compute LR. Therefore we use genotype-oriented curve clustering to partition longitudinal phenotypic data and obtain a grouping probability matrix for different genotypes. Curve clustering was originally used for differentiating the genetic expression pattern under time sequence [18]. We use this method to divide biological traits into three groups corresponding to 3 genotypes in F2 population, and assume each group is controlled by a specific genotype. The phenotypic curve y_i of subject i can be expressed as

$$y_i(t) = \sum_{j=1}^{J} \xi_{ij} g_j(t) + \sum_{c=1}^{C} \beta_c u_{ic} + e_i(t) \quad (7)$$

Where ξ_{ij} is the indicator for genotype j if the genotype of subject i belongs to the j class, the value is 1; otherwise, it is 0. u_{ic} is covariate for the subject i (c=1,…,C); β_c is the coefficient of covariate c, which is the effect value. For simplicity, all covariates are assumed zeros in the follow up context. $e_i(t)$ represents the residue, which follows the multivariate Gaussian distribution with mean 0, and variance-covariance matrix Σ.

Therefore, the mixture model with three genotypes is

$$L(\Theta) = \prod_{i=1}^{n}\sum_{j=1}^{3}\left[\omega_j f_j(y_i; g_j, \sum)\right] \quad (8)$$

Where $\omega=(\omega_1,\omega_2,\omega_3)$ is the probability weight vector, namely every element of the vector is non-negative and the sum of them is 1. Θ is the vector consisting of all unknown parameters including weight vector, parameters for biological curves of each clustering group, and variance-covariance, e.g. σ^2 and ρ in the AR(1) model.

To get the best estimation for clustering, we need to maximize the log-likelihood $\log(L(\Theta))$. In practical, we use the Expectation–maximization (EM) algorithm. In the estimation step, the mean vector of three genetic curves can be estimated by the method of Least squares, and then in the maximization step, all parameters except the biological curves can be solved by the global optimizer solution (e.g. Simplex) based on the fixed biological curves in the estimate step. Through the above calculation, we obtain the probability of each subject's phenotypic curve in each genotype group, and the probabilities of all subjects form a curve clustering matrix denoted by Q. Each element of Q is the probability q_{ij} of the subject i belongs to the j-class, i.e.

$$q_{ij} = f_j(y_i; g_j, \sum) / \sum_{j=1}^{3}\left[f_j(y_i; g_j, \sum)\right] \quad (9)$$

Filtering method of QTL

Theoretically QTLs that are strongly associated with the trait values tend to have large LR statistic. Thus we set a threshold based the correlation between the genetic probability matrix P of QTL and the

curve clustering matrix *Q* from curve clustering. As each permutation test only retains QTLs with a large correlation, we can filter out many loci to substantially reduce the computational burden. We propose three approaches to evaluate the correlation between the two probability matrixes, and then select QTLs based on the correlation.

Method 1: Calculate the maximum likelihood estimate L in Functional Mapping with the genetic probability matrix *P* and the curve clustering matrix *Q*. The formula of the statistics T_1 is

$$T_1 = \prod_{i=1}^{n}\left(\sum_{j=1}^{3} p_{ij} * q_{ij}\right) \quad (10)$$

The larger the statistics T_1, the larger the maximum likelihood and the higher correlation between the probability matrix *P* and *Q*.

Method 2: Calculate the average of the correlation coefficients between the genetic probability matrix *P* and the curve clustering matrix *Q*. The formula of the statistics T_2 is

$$T_2 = \frac{\sum_{i=1}^{n} cor(p_i, q_i)}{n} \quad (11)$$

Method 3: Treat matrix *P* and *Q* as a vector respectively and calculate the correlation coefficient between *P* and *Q*. The formula of the statistics T_3 is

$$T_3 = cor(P, Q) \quad (12)$$

We use the three methods to evaluate the correlation between the probability matrix *P* and *Q*. The higher is the correlation, the larger are the three statistics. Our method only performs on QTLs with high correlation to fulfill the maximum likelihood ratio LR in permutation test. Simulation results suggest the ability to search out the LR is a bit insufficient with just a single one of them. The accuracy of the solutions will considerably be improved when combing these three methods together. Therefore, we integrate all three methods to locate the most significant QTL in practice.

New method for permutation test

In traditional permutation tests used for calculating critical threshold of LR, we first randomly shuffle phenotypic data, calculate the LR test statistic for the shuffled data, and then obtain the largest one as one permutation result. After repeating the process N times and ranking the corresponding test statistics, we find the thresholds for the LR test statistic at significant levels of 1% or 5%. This permutation method will provide reliable cutoff points only if *N* is large enough. Ideal situation is to perform *N*=n! times, which is a burden for computing.

Our improved method does not calculate test statistics for all QTLs, it only calculates LR for the QTLs whose correlations between *P* and *Q* pass a predefined threshold. The 3 methods in 1.3 are used to select the QTLs. In practice, we use 5% or 10% as the threshold to select the strong QTLs by each method, and then make a union set for these 3 QTL sets. Because of small amount of QTLs are selected to do permutation, the computation burden is largely reduced.

Simulation Studies

To verify the effectiveness of our method, we conducted two numerical experiments in an F2 population. The genotype data are generated from a chromosome with length of 250 cM, which contains 15 markers with genetic distances between two adjacent markers 23, 9, 29, 14, 12, 10, 20, 20, 26, 6, 22, 21, 24, and 14 respectively, and the significant QTL position is defined at 119 cM. We also assume that the growth follows logistic growth curves with parameters *a,b,r* for three genotypes *QQ* (18.18, 9.98, 0.99), *Qq* (17.08, 9.78, 0.97), and *qq* (15.95, 9.88, 0.98), respectively. In our simulated data, we utilize AR(1) matrix with ρ=0.5 and σ^2=1.2 and 4.8, respectively.

We employed two different variance-covariance matrices to illustrate the effectiveness of our method. When the parameter σ^2 is 1.2, the results of curve clustering for the simulation data are shown in Figure 1. As we can see, three curve groups has been clustered clearly, therefore, the curve clustering method is effective. The results for σ^2=4.8 are quite similar and not displayed here.

After clustering the curves, we combined the three methods to choose the QTLs with high correlation. Through 1000 permutation tests, we obtain the threshold curve of significant level, shown in Figures 2 and 3.

In Figures 2 and 3, the black line denotes the traditional permutation test results, the blue, green, and red ones denote the threshold curves of QTL after 20%, 10%, and 5% of QTLs were selected, respectively. It can be seen that curves are almost identical when the *p*-value are between 0.001~0.01. Even if *p*-value is larger, these curves are very close, which means significant QTLs are mostly been chosen by the filtering methods.

The most important one that we care about is whether the maximum

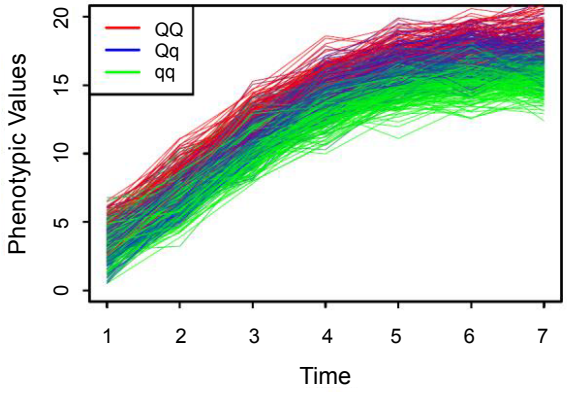

Figure 1: Clustering result of growth curves.

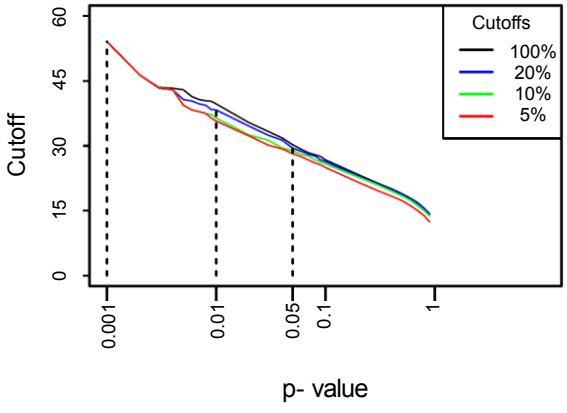

Figure 2: Cutoff curves obtained from 1000 permutation tests for the simulation (σ^2=1.2).

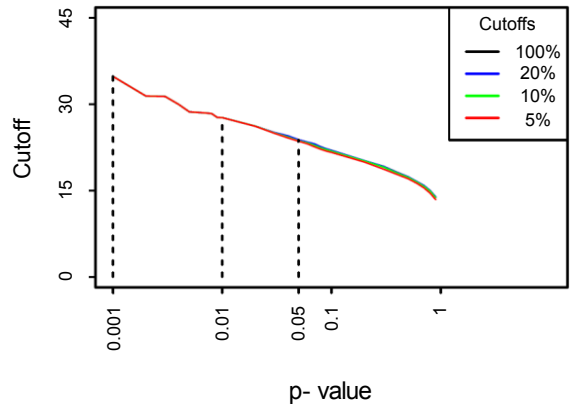

Figure 3: Cutoff curves obtained from 1000 permutation tests for the simulation (σ^2=4.8).

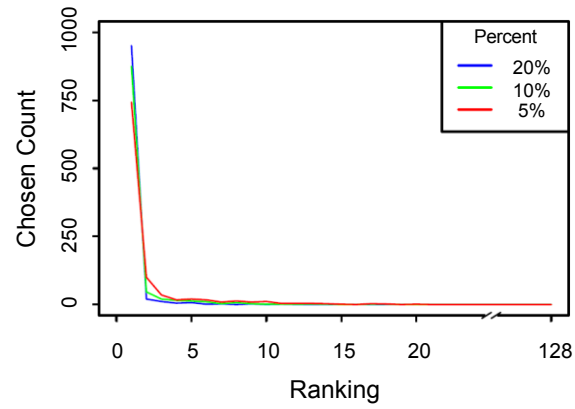

Figure 5: Times Chosen count at each rank of LR value in 1000 permutation tests (σ^2=4.8).

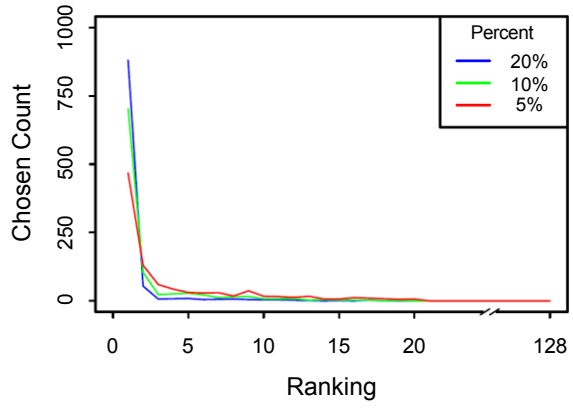

Figure 4: Chosen count at each rank of LR value in 1000 permutation tests (σ^2=1.2).

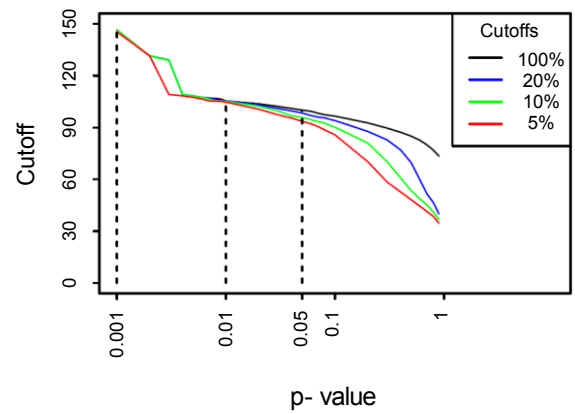

Figure 6: Cutoff curves obtained from 1000 permutation test for real data.

value of LR is selected in each permutation test. In other words, we expect the maximum LR can be selected in each permutation test by our proposed method. The rank level of LR chosen by our method among all QTLs could indicate the validation of our method. Figures 4 and 5 demonstrate the ranks of the chosen QTLs by our method in 1000 permutation tests, where the vertical axis lists the chosen count for each rank. For traditional permutation test, the top one will definitely be selected by all 1000 permutations, but our method requires 85%-90% less times. Some points from our method are not the maximum LR, but all of the QTLs are ranked top 20.

The original permutation algorithm takes 41.2 hours using 16 CPUs to do 1000 permutation simultaneously for these 15 markers and 125 QTLs, which is equivalent to 659.2 CPU hours. But it only takes 30.8, 65.9, and 131.8 CPU hours if we use 5%, 10%, and 20% of our selection filtering method, respectively. Therefore, our method has greatly reduced the computing time in Functional Mapping.

Real Data Analysis

We used a data set from poplar to verify our method. The data are from 450 backcross individuals with 19 linkage groups and 90 molecular markers. The phenotypic values were measures in 11 seasons. The results for permutation tests are displayed in Figure 6. The differences thresholds between traditional permutation and our method are very small for the significant levels 0.01~0.05, therefore, our method can be used in real data analysis.

Discussion

The flexibility and applicability of Functional Mapping are reduced due to a large computation time in real data analysis. In this paper we choose QTLs that are strongly correlated with phenotypic data to reduce computational burden. We quantify the correlations of QTL and phenotypic data using several methods in the QTL selection, so we can filter out some QTLs from our calculation. The simulation studies and real data analysis have shown that our method can greatly increase the computing speed without loss much of accuracy.

Since Functional Mapping utilizes mixture models to calculate LR, our method is designed for permutation tests of mixture models. Therefore our method can be used for the permutation tests of all other mixture models. Functional Mapping uses biological growth curves as the expected phenotypic vectors, so we group the growth curves using clustering methods. Different grouping methods should be used for different expected vectors.

Conclusion

The method we proposed in this paper can solve not only the efficiency of permutation test, but also QTL mapping, which is a process

of scanning each QTL. If we want to find the optimal LR of QTLs for a chromosome or linkage group, the method we have proposed here will provide a fast solution. Our method has been integrated into our software package *Funmap2* for Functional Mapping model (https://github.com/wzhy2000/Funmap2).

Acknowledgments

This study was supported by supported by National Training Program of Innovation and Entrepreneurship for Undergraduates (201510022069), the Fundamental Research Funds for the Central Universities (BLX2013026), NSFC Grant (31470675).

References

1. Wang J, Yu H, Weng X, Xie W, Xu C, et al. (2014) An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. Journal of experimental botany, ert464.
2. Wang J, Yang J, Jia Q, Zhu J, Shang Y, et al. (2014) A new QTL for plant height in barley (Hordeum vulgare L.) showing no negative effects on grain yield. PloS one 9: e90144.
3. Zurek PR, Topp CN, Benfey PN (2015) Quantitative trait locus mapping reveals regions of the maize genome controlling root system architecture. Plant physiology 167: 1487-1496.
4. Maharijaya A, Vosman B, Steenhuis-Broers G, Pelgrom K, Purwito A, et al. (2015) QTL mapping of thrips resistance in pepper. Theoretical and Applied Genetics 128: 1945-1956.
5. Chen W, Yao J.B, Chu L, Liu HJ (2016) QTL Mapping of Important Agronomic Characters in Cotton [J]. Fenzi Zhiwu Yuzhong 8: 6.
6. Shaohua Y (2011) Development of Plant QTL Analysis. Chinese Agricultural Science Bulletin 3: 046.
7. Zargar SM, Raatz B, Sonah H, Bhat JA, Dar ZA, et al. (2015) Recent advances in molecular marker techniques: Insight into QTL mapping, GWAS and genomic selection in plants. Journal of Crop Science and Biotechnology 18: 293-308.
8. Pletcher SD, Geyer CJ (1999) The genetic analysis of age-dependent traits: Modeling the character process. Genetics 153: 825-835.
9. Jaffrézic F, Pletcher SD (2000) Statistical models for estimating the genetic basis of repeated measures and other function-valued traits. Genetics 156: 913-922.
10. Kirkpatrick M, Heckman N (1989) A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. Journal of mathematical biology 27: 429-450.
11. Ma CX, Casella G, Wu R (2002) Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 161: 1751-1762.
12. Wu R, Lin M (2006) Functional Mapping-how to map and study the genetic architecture of dynamic complex traits. Nature Reviews Genetics 7: 229-237.
13. Wu R, Ma C, Casella G (2007) Statistical genetics of quantitative traits: linkage, maps and QTL. Springer Science & Business Media.
14. Sun L, Jiang L, Ye M, Zhu X, Wang J, et al. (2015) Functional Mapping: How to Map Genes for Phenotypic Plasticity of Development. In Evolutionary Biology: Bio diversification from Genotype to Phenotype (pp. 3-17). Springer International Publishing.
15. Li J, Wang Z, Li YC, Wu RL (2014) A novel QTL mapping model for allometric growth and pleiotropic extension[J]. Journal of Nanjing Forestry University (Natural Sciences Edition) 38: 35-39.
16. Wang Z, Pang X, Lv Y, Xu F, Zhou T, et al. (2012) A dynamic framework for quantifying the genetic architecture of phenotypic plasticity. Briefings in bioinformatics.
17. Doerge RW, Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142: 285-294.
18. Wang Y, Xu M, Wang Z (2012) How to cluster gene expression dynamics in response to environmental signals [J]. Briefings in Bioinformatics 13: 162-174.