Statistics and Its Interface Volume 10 (2017) 387–397

A nonparametric approach for functional mapping of complex traits

Han Hao†, Lidan Sun†, Xuli Zhu, and Rongling Wu*

Functional mapping is a statistical tool for mapping quantitative trait loci (QTLs) involved with a function-valued phenotypic trait. The utility of functional mapping is often displayed when the phenotypic trait represent a developmental process and can be modeled by a parametric approach. However, there are many practical situations in which no explicit parametric forms are feasible to capture the dynamic change of phenotypic traits across a time or space scale. We address this issue to expand the applying scope of functional mapping by utilizing a nonparametric adaptive high-dimensional ANOVA (HANOVA) method. A discrete Fourier transformation was implemented to eliminate the dependence structure of errors that are assumed to be stationary along the measurement process, followed by the choice of the first several Fourier coefficients that can explain a majority of phenotypic variation for QTL mapping. From simulation tests, HANOVA-based functional mapping was observed to display high statistical power for detecting subtle variation. By analyzing the real dataset of a mapping population for mei, a woody ornamental plant naturally distributed in China, the new model has successfully identified many significant QTLs that control leaf shape. The model should find its immediate implications for mapping any high-dimensional phenotypic measurements with no explicit form.

AMS 2000 subject classifications: Primary 62H15, 62G05; secondary 62P10.
Keywords and phrases: Functional mapping, Functional ANOVA, QTL, Discrete Fourier transformation, Woody plant.

1. INTRODUCTION

One of the most important tasks in modern biology is to construct the precise genotype-phenotype map by detecting key genes, known as quantitative trait loci (QTLs), and their interactions with phenotypic traits [1]. Genetic mapping based on genetic linkage maps constructed from molecular markers has been considered as the most popular approach for QTL detection and mapping since its emergence in the 1980s [1]. This approach dissects complex phenotypes into their underlying genetic components expressed at the DNA level, allowing the genetic architecture of complex traits to be elucidated. With the advent of high-throughput genotyping and phenotyping techniques, genetic mapping has developed toward two different directions: genome-wide association studies (GWAS) by collecting genetic polymorphisms throughout the entire genome to detect and capture all possible genes for a complex trait [3, 4, 5, 6] and functional mapping by collecting multiple trait values measured along a time or space scale and integrating biological principles and processes into genetic mapping to elucidate the developmental genetic mechanisms of a complex phenotype [7, 8, 9]. A growing body of analytical approaches for GWAS has been developed, both to meet challenges in statistical modeling and analysis of genetic data and account for the complexity of phenotypic information [10, 11, 12, 13, 14, 15]. For example, renovated model selection techniques have been implemented in a GWAS setting to analyze a large number of SNPs simultaneously based on a relatively small sample size [10, 11, 16, 18].

In a parallel with GWAS that attempt to analyze many markers, functional mapping focuses on the complexity of trait values and models a series of phenotypic measures made for the same trait across a time and space scale [7, 8, 9, 19, 10, 21, 22]. These so-called function-valued traits are widely seen in growth analysis, shape analysis, network analysis, and clinical trials. Functional mapping and its extension, functional GWAS (fGWAS) were derived to accommodate to whole-genome analysis [23, 24], integrating the functional feature of traits into genetic mapping to increase the statistical power and the biological relevance of the results obtained. Functional mapping is particularly powerful for mapping QTLs controlling a biological trait which can be described by parametric curves with a fixed set of parameters, such as generalized logistic growth equations [19, 20, 21] and pharmacological Emax models [22, 25] among others. A set of nonparametric or semi-parametric approaches have also been developed to model the dynamic pattern of phenotypic traits that cannot be specified by any explicit form of mathematical equations [23, 24].

While current functional mapping models have been instrumental for mapping many traits, such as body mass growth in mice [26], human body shape [27] and plant height growth in rice [28], soybean [20] and scot pine [21], they may have not been well used in the situation where a series of measures of a complex trait are such high-dimensional

*Corresponding author.
†These two authors contributed to this work equally.

and arbitrary that no parametric models can approximate its dynamic pattern. For example, by high-throughput phenotyping, one can record the daily or even hourly change of physiological parameters, such as photosynthesis rate or respiration rate, leading to an ultrahigh-dimensional data of phenotypic measurements [29]. In shape analysis, it is crucial to describe a detailed structure of an organ by using a high-dimensional set of data [20]. Modeling and analysis of such high-dimensional phenotypic data have been proven of great challenge, whose integration with functional mapping has not been explored thus far.

In this article, we present an adaptive nonparametric method for detecting significant genes for function-valued traits based on the high dimensional ANOVA (HANOVA). Compared with previous work of functional mapping [19, 20, 21], this method requires no parametric pattern of trait values, allowing time- or space-dependent changes of the trait in an arbitrary form. We assume stationary but possibly dependent errors for each individual along the measurement process. A discrete Fourier transformation is first performed on the trait values. As a result, the error terms become approximately independent, and large absolute values intend to locate in the first several Fourier coefficients [30, 31]. Then we perform an adaptive HANOVA test at each marker, where the number of Fourier coefficients involved in the test is adaptively chosen in order to achieve optimal statistical power [30]. The method was applied to analyze a real data set of leaf shape in mei, a woody ornamental plant naturally distributed in China. The method was compared to several previous multivariate analysis of variance (MANOVA) approaches designed for high dimensional data, including: 1) MANOVA where the Moore-Penrose inverse is used as a substitute of the inverse of singular covariance matrices [32], 2) a testing procedure using Dempster's trace of within and between group covariance matrices [33], and 3) MANOVA on a transformed dataset using principle component analysis (PCA) [34]. The methods were compared in both real data analysis and simulation to illustrate the utility and usefulness of HANOVA in practice.

2. METHOD

2.1 A general testing procedure

Consider a mapping population of N individuals, such as the backcross, F_2, or full-sib family, or a random set of samples from a natural population, which is genome-wide genotyped by a large set of molecular markers from which to construct a high-density linkage map. Let $\boldsymbol{y}_n = (y_n(1), \cdots, y_n(M))$ denote a series of trait values for individual n $(n = 1, \cdots, N)$ measured at M points. Since the linkage map is highly dense, we perform genetic mapping by directly associating markers with the trait value. Consider a single marker with G genotypes and record the genotypes as $1, \cdots, G$, respectively. Let g_n denote the genotype of individual n $(g_n = 1, \cdots, G)$. Then, the trait value of this individual is expressed as

$$y_n(j) = f_{g_n}(j) + \epsilon_n(j), \tag{1}$$

where $n = 1, \cdots, N$, $j = 1, \cdots, M$, $y_n(j)$ is the jth element of the trait vector of individual n, $f_{g_n}(j)$ is the jth element of the mean (genotypic) trait value for individuals with genotypic value g_n, and $\epsilon_n(j) \sim N(0, \sigma^2)$ is the error term. We assume the error terms are independent across different individuals, and for any fixed n, $\{\epsilon_n(j), j = 1, 2, \cdots, M\}$ are stationary but possibly dependent.

We are interested in testing whether the genotypes differ significantly in the trait value from each other. The hypothesis testing problem is then expressed as

$$\mathrm{H}_0: \ f_1(j) = f_2(j) = \cdots = f_G(j) \ \forall j = 1, \cdots, M \text{ vs.} \tag{2}$$

$$\mathrm{H}_a: \text{ otherwise} \tag{3}$$

The test statistic is calculated from the hypotheses above. By comparing it with a critical threshold determined from simulation studies with Bonferroni multiple test correction, we can judge whether we have detected a significant SNP.

2.2 Discrete Fourier transformation

The discrete Fourier transformation is first applied to each vector of trait values,

$$\tilde{y}_n(k) = \sum_{j=1}^{M} y_n(j) e^{-i\frac{2\pi kj}{M}} \tag{4}$$

where $k = 0, 1, \cdots, M-1$, i is the imaginary unit.

The real and imaginary parts of the first S Fourier coefficients are then compiled to form the new trait data, and the rest of Fourier coefficients are discarded as white noise. S can be any fixed number such that the length of the resulting transformed vector is no larger than the length of the original vector. For simplicity, we set $S = M$. So the resulting transformed vector should be:

$$\boldsymbol{y}_n^* = (y_n^*(1), y_n^*(2), y_n^*(3), \cdots, y_n^*(M)) \tag{5}$$

where $y_n^*(2k+1) = \mathrm{RE}((\tilde{y}_n(k)), \ y_n^*(2k) = \mathrm{IM}(\tilde{y}_n(k)), k = 1, \cdots, \left[\frac{M}{2}\right]$.

After the discrete Fourier transformation, the new trait values can be expressed as

$$y_n^*(j) = f_{g_n}^*(j) + \epsilon_n^*(j) \tag{6}$$

where $n = 1, \cdots, N$, $j = 1, \cdots, M$.

The Fourier bases are chosen because the leaf shape is usually symmetric, plus the two following purposes:

1. After transformation, the stationary errors are converted into approximately independent Gaussian errors [30, 31]; i.e., the new error terms $\epsilon_n^*(j)$ are approximately independent for all n and j.
2. After transformation, large absolute values of the trait vector tend to locate in the first several elements. This feature allows us to apply the HANOVA method [30] in the following hypothesis testing procedure.

2.3 HANOVA method for testing the genotype effect

After the discrete Fourier transformation, the hypothesis testing problem 2, is transformed into

(7) $\mathrm{H}_0: \ f_1^*(j) = f_2^*(j) = \cdots = f_G^*(j) \ \forall j = 1, \cdots, M$ vs.

(8) $\mathrm{H}_a:$ otherwise

Then, an adaptive analysis of variance procedure for comparing high dimensional data through HANOVA is used to perform such a hypothesis testing problem. Let

$$N_g = \sum_{n=1}^{N} I(g_i = g),$$
$$\bar{y}_g^*(j) = \frac{1}{N_g} \sum_{g_n=g} y_i^*(j),$$
$$s_g^{2^*}(j) = \frac{1}{N_i - 1} \sum_{g_i=g} [y_i^*(j) - \bar{y}_g^*(j)]^2,$$
$$\bar{y}^*(j) = \frac{\sum_{g=1}^{G} N_g \bar{y}_g^*(j)/s_g^{2^*}(j)}{\sum_{g=1}^{G} N_g/s_g^{2^*}(j)}$$

with $n = 1, \cdots, N$, $g = 1, \cdots, G$, $j = 1, \cdots, M$. The test statistic is

(9)
$$F_M = \max_K \frac{1}{\sqrt{2M}} \sum_{j=1}^{K} \sum_{g=1}^{G} \left\{ \frac{N_g}{s_g^{2^*}(j)} [\bar{y}_g^*(j) - \bar{y}^*(j)]^2 - 2K \right\}.$$

The idea behind the test statistic is similar to the F-statistic in the ordinary analysis of variance (ANOVA). Here the term $[\bar{y}_g^*(j) - \bar{y}^*(j)]^2$ is a generalization of the between-group sum of squares, and $s_g^{2^*}(j)$ is a generalization of the within-group sum of squares. Thus the test statistic still represents the proportion of between-group sum of squares out of the total sum of squares.

When $M \to \infty$, the asymptotic distribution [30] of the test statistic is

(10) $$P(\tilde{F}_M < x) \to \exp\{-\exp(-x)\}$$

where

$$\tilde{F}_M = \sqrt{2 \log\log M} F_M - [2 \log\log M + 0.5 \log\log\log M - 0.5 \log(4\pi)].$$

Thus, the asymptotic p-value for the hypothesis testing problem with a calculated statistic F is

(11) $$P(\tilde{F}_M > x) \to 1 - \exp\{-\exp(-x)\}$$

The asymptotic p-value can be directly used to approximate the true p-value when the number of measurements is large enough. When the number of measurements is smaller, the true p-value should be determined using a random simulation.

3. APPLICATION

3.1 Real data analysis

We tested and validated the usefulness of the HANOVA-based mapping model by analyzing a real dataset of leaf shape in a woody plant, mei (Prunus mume Sieb. et Zucc.). Two commercially used mei cultivars differing in morphological traits, selected from the Qingdao Mei Garden, Qingdao, China, were crossed to generate a full-sib family composed of 228 hybrids, grown in the Xiao Tangshan Horticultural Trial, Beijing, China [35, 36]. A total of 1484 SNPs were generated from this full-sib family, including 261 intercross markers and 1223 testcross markers segregating in the hybrids. Sun et al. [35] provided a procedure for testing genetic effects of these two types of markers.

We collected three representative leaves on the main stem for each hybrid and took their photos. Then we delimited the contour of a shape into 360 equally-angled semilandmarks and recorded their coordinates, from which the radiuses from the centroid to semilandmarks were measured [37, 38]. In order to minimize variance caused by scale and pose, we performed shape alignment on the original radius measures [37, 38]. We first normalized the radius vectors for each leaf by setting the Euclidean norm as a fixed value. Then we used the average of all normalized contours as the reference shape for shape alignment. The idea of shape alignment is to rotate each contour and find the pose most similar to the reference shape. The result of shape alignment of each leaf contour is a rotation of the normalized shape contour that achieves the minimum Euclidean distance between the reference shape and any rotations. After shape normalization and alignment, we obtained a vector of 360 values for each leaf and use this vector for further data analysis.

The normalized radius values were first processed using a discrete Fourier transformation. Fourier bases are suitable for approximating the radius values because the radius values are symmetric about the center. To verify that the Fourier bases provide good approximations for the radius values, we used both Fourier bases and the 3rd order B-spline bases to approximate the mean radius vector of all leaves. Table 1 gives the mean squared error of the approximation result under various degrees of freedom. The result shows that Fourier basis provides a good approximation even at very small degrees of freedom. When the degree of freedom is 2, the mean squared error using Fourier bases is 7.69×10^{-7}, while the mean squared error using B-spline bases is 1.22×10^{-6}.

A discrete Fourier transformation turns the data to be approximately independent along different measurement points (degrees 1 to 360). To verify the approximate independence, we performed hypothesis tests on the Pearson correlations between measurement points of simulated data sets with AR(1) covariance structure and different auto-correlations. The hypothesis tests are performed using Fisher's Z-transformation. Similar to the real data, each

Table 1. Summary of the approximation results using Fourier bases and 3rd order B-spline bases. DOF is the degree of freedom, i.e. the total number of bases used for the approximation. MSE is the mean squared error of the approximation. The range of the original data is 0.015 – 0.021

DOF	MSE: Fourier	MSE: B-spline
2	7.69×10^{-7}	1.22×10^{-6}
3	6.59×10^{-7}	1.10×10^{-6}
4	5.62×10^{-7}	8.11×10^{-7}
5	4.99×10^{-7}	5.09×10^{-7}
10	3.55×10^{-7}	1.24×10^{-7}
20	3.10×10^{-7}	1.94×10^{-8}
50	3.07×10^{-7}	1.11×10^{-9}
100	3.06×10^{-7}	7.11×10^{-11}

Table 2. Summary of the test of pairwise correlation. Auto-correlation is the correlation value used to generate the AR(1) random error for simulation. Max Cor. Est. and Min Cor. Est. are the maximum and minimum correlation value estimations of all pairwise correlations of the transformed data. Min P-value is the minimum p-value for the hypothesis test of no correlation. The total number of tests performed is 64620 and Min Adjusted P-value is the minimum p-value of the hypothesis test adjusted using Bonferroni correction

Auto -cor	Max Cor.	Min. Cor.	Min P-value	Min Adjusted P-value
0.9	0.256	-0.268	4.000×10^{-5}	1.000
0.7	0.267	-0.264	4.580×10^{-5}	1.000
0.5	0.283	-0.275	1.438×10^{-5}	0.919
0.3	0.265	-0.284	1.285×10^{-5}	0.821
0.1	0.269	-0.270	3.654×10^{-5}	1.000
−0.1	0.252	-0.269	3.804×10^{-5}	1.000

simulated data sets contains 228 vectors of dimension 360. Each simulated vector is the average of all leaf contour vectors in the real data, plus an AR(1) error vector with standard deviation estimated from the real data, and different auto-correlations including the auto-correlation value estimated from the real data. Table 2 and Table 3 gives the testing result, where the null hypothesis is independence, i.e. correlation equals 0. Table 2 gives the result of pairwise tests on the correlations between any two of the measurement points (degree 1 to 360), where the total number of tests is $360 \times 359/2 = 64620$. Table 3 gives the result of tests on adjacent measurement points, where the total number of tests is 360-1=359. Adjusted p-values were calculated using Bonferroni correction. Table 2 and Table 3 show that all adjusted p-values are greater than 0.05, indicating that the transformed data has no significant deviation from independence.

After the discrete Fourier transformation, a HANOVA test was performed for each marker. First, we calculate an original p-value for each single test using the simulation re-

Table 3. Summary of the test of adjacent correlation (time lag=1). Auto-correlation is the correlation value used to generate the AR(1) random error for simulation. Max Cor. Est. and Min Cor. Est. are the maximum and minimum correlation value estimations of all pairwise correlations of the transformed data. Min P-value is the minimum p-value for the hypothesis test of no correlation. The total number of tests performed is 359 and Min Adjusted P-value is the minimum p-value of the hypothesis test adjusted using Bonferroni correction

Auto -cor	Max Cor.	Min. Cor.	Min P-value	Min Adjusted P-value
0.9	0.190	-0.219	8.482×10^{-4}	0.301
0.7	0.185	-0.217	9.707×10^{-4}	0.345
0.5	0.207	-0.195	1.693×10^{-3}	0.601
0.3	0.197	-0.215	1.071×10^{-3}	0.380
0.1	0.179	-0.174	6.858×10^{-4}	1.000
−0.1	0.186	-0.208	1.625×10^{-3}	0.577

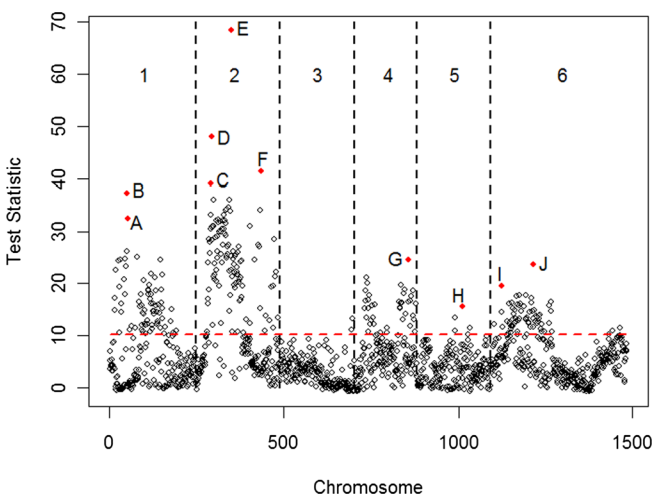

Figure 1. The Manhattan plot of test statistics for individual loci across the genome of mei, Prunus mume Sieb. et Zucc (containing six chromosomes). Ten representatives of significant SNPs (labeled by A – J) are highlighted in red. The broken horizontal line is the critical threshold at the 5% significance level after Bonferroni correction.

sult of 100,000,000 simulation runs described in the section Simulation 1, with the covariance and autocorrelation parameters estimated from the real data. Then we use a Bonferroni correction to adjust for multiple testing. Figure 1 illustrates a Manhattan plot of test statistic values for all loci across six different chromosomes. It was found that a total of 281 loci each display a significant effect on the leaf shape (adjusted p-value < 0.05 after Bonferroni correction). Furthermore, 54 of the significant markers have (simulated) original p-values less than 10^{-8}, indicating that the testing method is very powerful. The detected significant loci tend

Table 4. Summary of ten representative significant QTLs (labeled by A – J) detected from Figure 3.1. The adjusted p-values are calculated using Bonferroni correction. The simulated p-values are computed using results from Simulation 1, where we had 100,000,000 runs. Adjusted p-values $< 1.48 \times 10^{-5}$ means the corresponding test statistic is larger than the maximum test statistic in the simulations

QTL	LG	Type	Name	P-value	R^2
A	1	2	fheter 106496	1.65×10^{-13}	0.00798
B	1	2	fheter 105087	1.10×10^{-11}	0.00678
C	2	2	mheter 117411	$< 10^{-16}$	0.07075
D	2	2	mheter 110968	$< 10^{-16}$	0.06230
E	2	3	hk hk865	$< 10^{-16}$	0.06763
F	2	2	fheter 102985	$< 10^{-16}$	0.00772
G	4	2	mheter 111568	3.19×10^{-8}	0.01417
H	5	2	mheter 116168	$2.52 \times 10^{-4}6$	0.00458
I	6	3	hk hk2360	4.77×10^{-6}	0.02698
J	6	3	hk hk1140	7.50×10^{-8}	0.07064

Table 5. Estimated recombination rate of representative QTLs that belong to the same linkage group

Linkage group	QTL 1	QTL 2	Recombination rate
1	A	B	0.0219
2	C	D	0.0263
2	C	E	0.0833
2	C	F	0.4517
2	D	E	0.0965
2	D	F	0.4430
2	E	F	0.3465
6	I	J	0.3684

to congregate, which probably results from the genotype dependence of adjacent linked genes. Most of the significant SNPs locate on chromosomes 1, 2, 4 and 6 (Figure 1). Table 4 gives the location, genotype number, name, adjusted p-value and R^2 for the 10 representatives, where R^2 is generalized from ordinary ANOVA and calculated by

$$R^2 = 1 - \frac{\sum_{j=1}^{M}\sum_{g=1}^{G}\sum_{g_n=g}[y_n(j) - \bar{y}_g(j)]^2}{\sum_{j=1}^{M}\sum_{n=1}^{N}[y_n(j) - \bar{y}(j)]^2} \tag{12}$$

These loci are highly significant in impacting on the overall difference of leaf shape. They each explain 0.68 – 7.08% of the phenotypic variation in leaf shape. Table 5 gives the estimated recombination rate of the representative loci that belong to the same linkage group. Representative loci A, B in linkage group 1, and C, D, E in linkage group 2 are highly related indicated by a small recombination rate.

To explain the results by our nonparametric approach, we chose 10 representative significant loci, labelled by A – J, from the chromosomes that are likely to harbor important loci (Figure 1). Some of these chosen SNPs belong to testcross markers each with two genotypes, whereas the rest

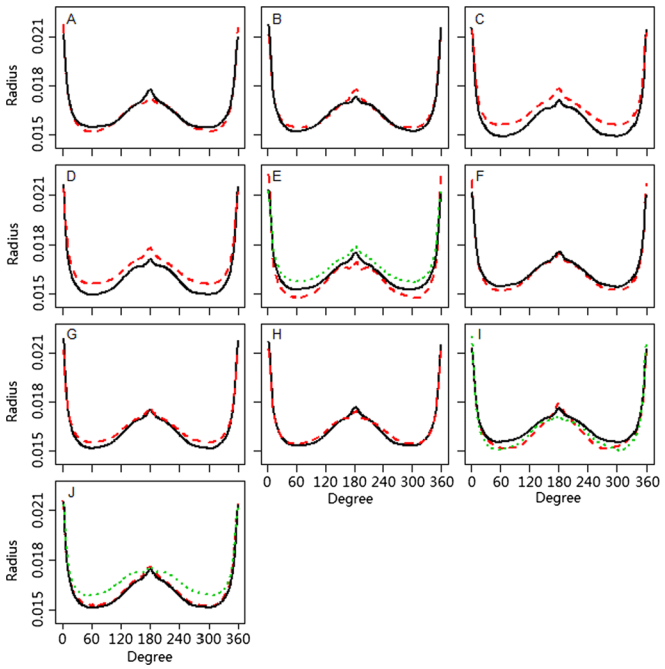

Figure 2. The profiles of the radius values from the centroid to semilandmarks on the contour of leaf shape for different genotypes (solid, broken or dotted) at ten representatives of significant SNPs chosen from Figure 1.

are intercross markers each with three genotypes. At each locus, we drew the profile of genotypic means for radius values the centroid to semilandmarks over a range of 0 – 360° (Figure 2). It can be seen that some SNPs possess larger across-genotype differences in the overall radius profile than the other, suggesting the stronger genetic control by the former than latter. It is also interesting to see that different genotypes at the same loci vary, depending on the region of the radius profile. This provides detailed information about which part of the leaf shape a genetic locus determines.

To visualize the detail of genetic control over leaf shape, we transformed the radius profiles back to leaf shape for each genotype at each chosen loci (Figure 3). Although different genotypes for many significant loci appear to be visually similar in leaf shape, they do differ from each other significantly. This shows a powerful capacity of our approach to discern subtle differences controlled by individual genes.

3.2 Simulation

To further study the statistical properties of the proposed method in real application, we have conducted several simulations based on the mei dataset.

3.2.1 Simulation 1

To study the null distribution and false discovery rate in real application, we have conducted the following simulation based on the real data:

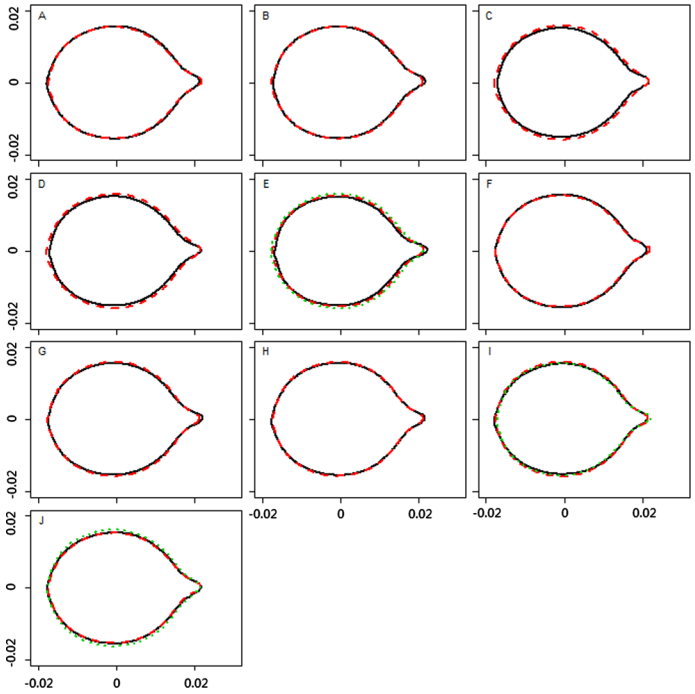

Figure 3. Average leaf shape of each genotype at ten representatives of significant loci chosen from Figure 1, transformed back from the radius profiles in Figure 2.

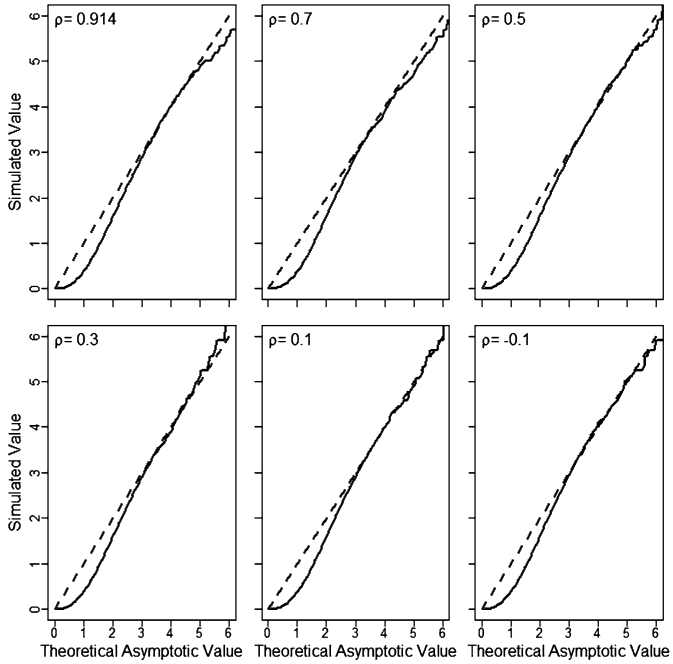

Figure 4. Plots of simulated p-value versus the theoretical asymptotic p-value in the $-log_{10}$ scale based on 1,000,000 simulation runs (that is, a value of 2 on the graph means p-value=10^{-2}). The solid lines are the simulated p-values versus theoretical asymptotic p-values, and the dashed lines are the reference line $y = x$. Simulations are performed using different auto-correlation values $\rho = 0.914, 0.7, 0.5, 0.3, 0.1, -0.1$. The auto-correlation value is marked at the top-left corner of each plot.

Assuming the null hypothesis, i.e. the leaf contours come from a homogeneous population with a fixed probability distribution. Let $y_n = (y_n(1), \cdots, y_n(M))$ denote leaf contour radius values for individual n $(n = 1, \cdots, N)$ measured at degree $1, \cdots, M$, and let $\boldsymbol{f} = (f(1), \cdots, f(M))$ be the population mean vector of the leaf contour, $M = 360$. $\boldsymbol{y}_n(j) \sim N(\boldsymbol{f}, \Sigma)$ where Σ is the covariance matrix for the AR(1) model with error variance σ^2 and autocorrelation parameter ρ.

We first calculate the maximum likelihood estimator (MLE) of parameters $\boldsymbol{f}$, σ^2 and ρ using the real data containing 228 mei leaf samples, and denote the MLE as $\hat{\boldsymbol{f}}$, $\hat{\sigma}^2$ and $\hat{\rho}$.

For the first part of the simulation, we generate 228 random samples of genotype and contour radius vector with length 360, same as in the original data. The genotype is generated as a random vector of AA, Aa, and aa with probabilities 0.25,0.5,0.25, respectively. The simulated contour radius vector is the estimated mean vector $\hat{\boldsymbol{f}}$ plus a random error vector generated from an AR(1) process:

$$\epsilon = (\epsilon(1), \cdots, \epsilon(M)),\ \epsilon(t) = \hat{\rho}\epsilon(t-1) + \eta(t) \tag{13}$$

Here $\eta(t) \sim i.i.d\ N(0, \hat{\sigma}^2)$, $\hat{\rho} = 0.914$, $\hat{\sigma} = 0.00113$. We have also included simulation results for different auto-correlation values $\rho = 0.7, 0.5, 0.3, 0.1, -0.1$.

For the second part of the simulation, we generate 228 random samples of genotype and contour radius vector with lengths 60, 90, 120, 180, 360, 720, respectively. The genotype is still generated as a random vector of AA, Aa, and aa with probabilities 0.25,0.5,0.25, respectively. The simulated mean contour radius vector is generated using the original data. When generating a simulated contour radius vector with a length smaller than 360, this simulated mean vector is a sparsification of the estimated mean vector $\hat{\text{f}}$ from the original data. For example, a simulated mean vector of length 60 is $\boldsymbol{f}_{\text{sim,length60}} = (\hat{f}(6), \hat{f}(12), \hat{f}18), \cdots, \hat{f}(360))$. When generating a simulated contour of length 720, we have $\boldsymbol{f}_{\text{sim,length720}} = (h(1), h(2), h(3), \cdots, h(720))$, where $h(2t) = \hat{f}(t)$, $h(2t+1) = (h(2t) + h(2t+2))/2$, $t = 1, 2, 3, \cdots, 360$. The simulated contour radius vector is the simulated mean vector plus a random error vector generated from an AR(1) process:

$$\epsilon = (\epsilon(1), \cdots, \epsilon(M)),\ \epsilon(t) = \hat{\rho}\epsilon(t-1) + \eta(t) \tag{14}$$

Here $\eta(t) \sim i.i.d\ N(0, \hat{\sigma}^2)$, $\hat{\rho} = 0.914$, $\hat{\sigma} = 0.00113$.

Figure 4 is the graph of simulated p-value versus the theoretical asymptotic p-value in the $\log_{10}$ scale based on 100,000,000 runs for different auto-correlation values marked at the top-left corner of each plot. This result shows that when the sample size is 228, the simulated p-value and the theoretical asymptotic agrees with each other quite well

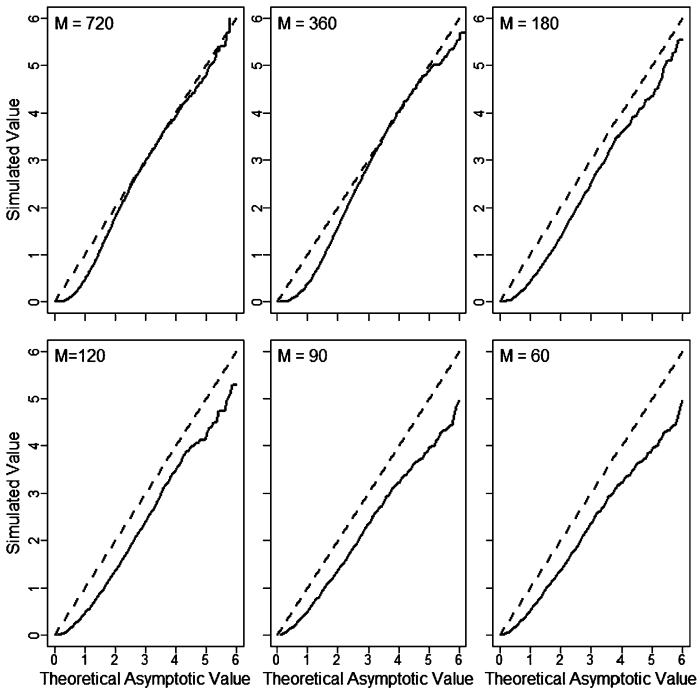

Figure 5. Plots of simulated p-value versus the theoretical asymptotic p-value in the $-log_{10}$ scale based on 1,000,000 simulation runs (that is, a value of 2 on the graph means p-value=10^{-2}). The solid lines are the simulated p-values versus theoretical asymptotic p-values, and the dashed lines are the reference line $y = x$. Simulations are performed using different contour radius vector length values M=720,360,180,120,60,30. The contour radius vector length value is marked at the top-left corner of each plot.

in the range 0.0001 to 0.01 for all the auto-correlation values. However, when the autocorrelation value is large and the theoretical asymptotic p-value goes below 0.0001, the simulated p-values can be different from the theoretical ones. Thus we need to be careful about small p-values in real application when the number of measurement points is limited. In the real data analysis all p-values are calculated using the simulation result.

Figure 5 is the graph of simulated p-value versus the theoretical asymptotic p-value in the $\log_{10}$ scale based on 100,000,000 runs for different contour radius vector lengths marked at the top-left corner of each plot. The result shows that with a larger sample size, the simulated p-value becomes closer to the theoretical p-value. When the vector length increases to 720, the simulated p-value agrees with the theoretical p-value in the whole range from 10^{-6} to 0.01. When the vector length decreases to 180, the simulated p-value deviates from the theoretical p-value, and such difference becomes larger as the vector length continue to decrease. This result verifies the theoretical conclusion that the distribution of the test statistic has an asymptotic theoretical distribution when the vector length approaches infinity. Also, together with Figure 4, the difference between the simulated p-value and the theoretical p-value also depend on the covariance matrix. Thus, whether the sample size is enough for the theoretical p-value to be reliable should still be verified by simulation with each real data set.

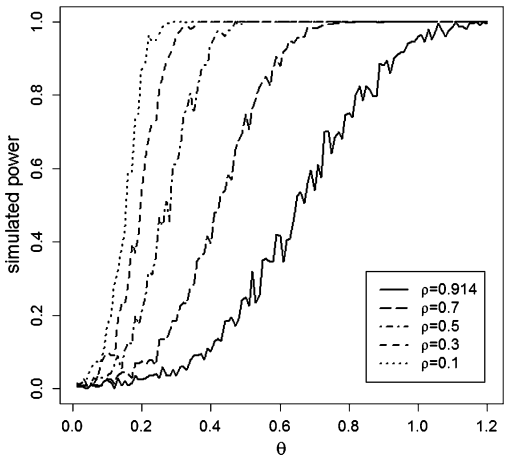

Figure 6. Plot of the simulated power (true positive rate) versus θ based on 1,000 runs. θ goes from 0 to 1.2 with a step of 0.01. The five curves are corresponding to auto-correlation values $\rho = 0.914, 0.7, 0.5, 0.3, 0.1$, respectively.

3.2.2 Simulation 2

To study the power the proposed method, we have conducted the following simulation:

Take QTL location E in the real data analysis as an example. Let $\boldsymbol{f}_g = (f_g(1), \cdots, f_g(M))$, $g = 1, 2, 3$ be the mean vectors of samples with QTL type hh, hk, kk, respectively. We then generate 228 random samples of genotype and measurement vector. The genotype is generated as a random vector of hh, hk, and kk with probabilities 0.25,0.5,0.25, respectively.

The simulated measurement vector is a mean vector plus a random error vector. The mean vector of samples with genotype hk is set to be $\boldsymbol{f}_2$. For several choices of $0 < \theta < 1.2$, the mean vector of samples with genotype hh is set to be $\boldsymbol{f}_2 + \theta(\boldsymbol{f}_1 - \boldsymbol{f}_2)$, the mean vector of samples with genotype kk is set to be $\boldsymbol{f}_2 + \theta(\boldsymbol{f}_3 - \boldsymbol{f}_2)$. The random error vector is again generated from an AR(1) process:

$$\epsilon = (\epsilon(1), \cdots, \epsilon(M)),\ \epsilon(t) = \hat{\rho}\epsilon(t-1) + \eta(t), \tag{15}$$

Here $\eta(t) \sim i.i.dN(0, \hat{\sigma}^2)$.

Same as in Simulation 1, the auto-correlation and variance are set as the maximum likelihood estimators using the original data, where $\hat{\rho} = 0.914$, $\hat{\sigma} = 0.00113$. We have also included results with different auto-correlation values $\rho = 0.7, 0.5, 0.3, 0.1$.

Figure 6 is the graph of the simulated power (true positive rate) versus θ based on 1,000 runs on θ changing from 0 to 1.2 with a step of 0.01 with five different

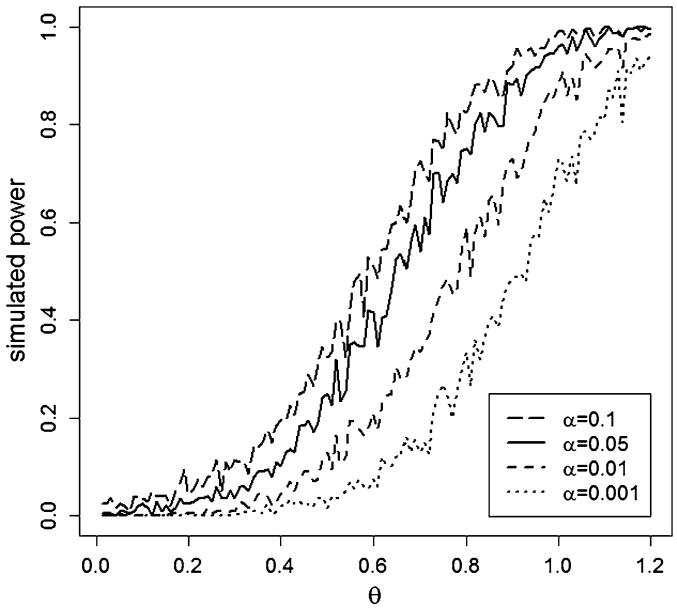

Figure 7. Plot of the simulated power (true positive rate) versus θ based on 1,000 runs. θ goes from 0 to 1.2 with a step of 0.01 using $\rho = 0.914$. The four curves are corresponding to significance levels $\alpha = 0.1, 0.05, 0.01, 0.001$, respectively.

auto-correlation values, with auto-correlaion values $\rho = 0.914, 0.7, 0.5, 0.3, 0.1$, respectively. Figure 6 shows that for this particular marker and at significance level 0.05, the power with auto-correlation value 0.914 (estimated from the real data) reaches 0.8 when $\theta = 0.82$, thus the marker can be detected with an adequate power even if the difference across different marker types slightly decreases. If the auto-correlation value decreases (e.g. to 0.7), the power can reach 0.8 even if the difference across different marker types decreases by half. These results indicate that HANOVA is quite powerful in detecting significant SNPs that affect the leaf shape.

Figure 7 is the graph of the simulated power (true positive rate) versus θ based on 1,000 runs. θ goes from 0 to 1.2 with a step of 0.01 using $\rho = 0.914$. The four curves are corresponding to significance levels $\alpha = 0.1, 0.05, 0.01, 0.001$, respectively. Figure 7 shows that for this particular marker, the power with auto-correlation value 0.914 (estimated from the real data) reaches 0.7 when $\theta = 1$ and $\alpha = 0.001$, thus using the original data, the marker can be detected with an adequate power even with significance level as small as 0.001.

3.3 Method comparison

At each single marker, hypothesis testing problem 2 is a multivariate analysis of variance (MANOVA) problem of testing the difference in mean between several groups of vectors. When the dimension of the vector is larger than the sample size, the MANOVA problem becomes high-dimensional and cannot be solved using classical MANOVA approaches, because the sample covariance matrix of the measured vectors, as the estimation of the covariance matrix, is singular.

Table 6. Comparison of the detected significant SNPs using HANOVA, Moore-Penrose inverse (MP), Dempster's trace and PCA

	# of significant SNP	# overlap with HANOVA
HANOVA	280	280
M-P	0	0
Dempster	129	80
PCA	55	54

Apart from the HANOVA approach we have used in this paper, other approaches have been developed to treat the high-dimensional MANOVA problem. These approaches mainly fall into two types: 1) constructing test statistics that avoid the usage of sample covariance matrix, and 2) performing dimension-reduction on the original data before a classical MANOVA testing procedure.

The HANOVA approach introduced in this paper is similar as the first type of approach. The data is first transformed in order to build up the test statistic on the transformed data, and dimension reduction procedure is carried out through the adaptive testing procedure. To further illustrate the power of the HANOVA approach, we compare this approach with three different previous approaches developed for the high-dimensional MANOVA problem. The first one is a testing procedure developed by Srivastava [32] using the Moore-Penrose inverse of the sample covariance matrix. The second one is a testing procedure developed by Fujikoshi et al. [33] using the Dempster's trace of the sample covariance matrix. The third one is a testing procedure developed by Kong et al. [34], where a principle component analysis (PCA) is first conducted on the original data for dimension reduction, then a classical MANOVA procedure is carried out after the dimension reduction.

The methods were compared on both the real data and simulated power. Table 6 is the summary of significant markers detected using the four different approaches. The HANOVA approach detected 280 out of 1484 significant markers. The approach using Dempster's trace detected 129 significant markers, where 80 of them were also detected by HANOVA. The approach using PCA detected 55 significant markers, where 54 of them were also detected by HANOVA. The approach using Moore-Penrose inverse failed to detect any significant marker. This result shows that for the shape analysis using the mei data, HANOVA is the most powerful one, detecting more than twice number of markers than any of the rest approaches. Also, a large percentage of the markers detected by the other approaches (62% of the Dempster's trace approach and 98% of the PCA approach) were also detected by HANOVA, indicating the method to be quite reliable.

Table 7 is the summary of power simulation. Same as in Simulation 2 in the previous part, the QTL E is used

Table 7. Comparison of simulated power using HANOVA, Moore-Penrose inverse (MP), Dempster's trace and PCA

ρ	0.9	0.7	0.5	0.3	0.1	-0.1
HANOVA	0.961	0.999	1.000	1.000	1.000	1.000
M-P	0.001	0.007	0.005	0.014	0.013	0.014
Dempster	0.425	0.389	0.540	0.561	0.664	0.658
PCA	0.134	0.170	0.152	0.148	0.136	0.142

to generate radius values corresponding to 3 different genotypes. Here we only consider mean curves corresponding to the original data, i.e. $\theta = 1$ in Simulation 2. The result shows the HANOVA method is quite powerful, with simulated powers greater than 0.961 with auto-correlation values ranging from 0.1 to 0.9. The power of the Dempster's trace approach varies around 0.5, with a higher power when the auto-correlation is smaller. The power values of the PCA and Moore-Penrose inverse approach are less than 0.2, indicating they are not so powerful for detecting SNPs in this situation.

4. DISCUSSION

An important challenge we face in genetic mapping is how to increase its biological relevance and statistical power to detect significant SNPs given that complex traits involve multiple stages of development. One way to increase the power of genetic mapping is to adopt a strategy of measuring the same trait repeatedly over time and space. When such repeated measurements span a developmental space, the simultaneous analysis and synthesis of multiple measurements can provide new insight into the biological mechanisms underlying trait formation. Thus, it has become crucial to integrate function-valued traits with a hypothesis testing procedure to unravel additional part of the genetic machineries which may be easily ignored by traditional approaches.

Wu and his group are among the first who have recognized the use of mathematical equations to model the temporal pattern of genetic effects by individual genes and their interactions [7, 8, 9, 19, 20]. In this article, we describe a more flexible approach based on HANOVA [30] to provide simple but powerful solution into testing differences between trait curves of different genotypes. The model-free testing procedure can be reasonably applied to any situation where the trait values are believed to be continuous curves along the measurements, although its power increases with the number of measurements. Before performing the test, we first perform a discrete Fourier transformation on the data to manipulate the possibly dependent error structure along measurements. The transformation will also help to locate large absolute values in the first few components of the transformed data, thus a lower dimension will be used in the adaptive testing procedure, leading to further increase of the power. Fourier bases are most suitable for approximating symmetric curves, but the application can be easily extended to asymmetric curves by adding a reflection of the original vector and making a new symmetric vector. The Fourier transformation can also be applied to unbalanced design situations, since it does not require the time or space points to be the same, and we can always perform the transformation on the union of all measurement points. The adaptive HANOVA testing procedure is very powerful in itself, such that even with a conservative Bonferroni correction, the statistical significance is still high enough to reveal a moderate number of significant SNPs.

The adaptive HANOVA test procedure is applied to analyze a real dataset of genetic association studies in a full-sib family of an ornamental woody plant, mei (Prunus mume Sieb. et Zucc). The new model is very powerful to detect significant loci for leaf shape variation from a pool of SNPs. Increased power of gene detection by the new model results from its capacity to detect cumulative subtle effects of phenotypic values over time. For example, for significant gene C (Table 3), the adjusted p-value is less than 1.48×10^{-5}, whereas a minimum p-value is 8.05×10^{-3} from a pointwise t-test on each single measurement. We also performed a simulation to study the power when the difference between mean trait values of different genotypes is further decreased. The result (Figure 6) shows that at significance level 0.05, even when the difference between mean trait values is diminished by half, the power of each single test can still reach 0.8. Also, the power of each single test can reach 0.7 with the original data, even at significance level 0.001 (Figure 7). It is found that the proportion of variance in the transformed trait values explained by a significant SNP (R^2) is generally large [39, 40, 41]. This may be because the variance structure of the original trait values is also transformed into the space of Fourier frequencies after the discrete Fourier transformation, thus revealing more inner structure that has been explained by the SNP. In other words, the significant SNP essentially explained a large part of variation hidden in the trait values.

The HANOVA method is compared with three other approaches designed for the high-dimensional MANOVA problem, including approaches using the Moore-Penrose inverse, the Dempster's trace and PCA. Both real data analysis and simulation results show that the HANOVA method is very powerful in detecting markers that significantly affect the leaf shape. Also, HANOVA and the other approaches share a large percentage of detected significant markers, indicating the testing result is quite consistent across different approaches, and the HANOVA approach is quite reliable.

This approach will find its immediate implications for genetic mapping and GWAS related to both biology and medicine. First, it is possible and sometimes essential in practice to take multiple measures for the same complex trait or disease. For example, blood pressure is an important predictor of human health, but it is also sensitive to many stochastic factors, such as emotion. Thus, any single measures of blood pressure may be imprecise, rather than repeated measures are crucial. By analyzing multiple measures at the same time using the new model, a more detailed

picture of the genetic control mechanisms for blood pressure can be illustrated. Second, many genetic mapping or GWAS are subject to missing heritability, since each significant SNP can only explain a tiny portion of the phenotypic variance. The new model can discern genetic variation in the accumulation of subtle differences over multiple measures, considerably enhancing the statistical power of gene detection. Third, statistical properties of the new model have been well studied by statisticians [30, 31], thus any application of it can assure the interpretability, generality and utility of the results obtained.

ACKNOWLEDGEMENTS

This work is supported by Special Fund for Forest Scientific Research in the Public Welfare (201404102), "One-thousand Person Plan" Award, grants from the Ministry of Science and Technology (2011AA100207, 2013AA102607) and the State Forestry Administration of China (201004012), and the Fundamental Research Funds for the Central Universities (BLX2013011).

Received 13 June 2015

REFERENCES

[1] Mackay, T. F., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10(8)** 565–577.

[2] Lander, E. S., & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121(1)** 185–199.

[3] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., ... & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **208(5720)** 385–389.

[4] Donnelly, P. (2008). Functional mapping—how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics* **456(7223)** 728–731.

[5] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9(5)** 728–731.

[6] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., ... & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42(7)** 565–569.

[7] Ma, C. X., Casella, G., & Wu, R. (2002). Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics* **161(4)** 1751–1762.

[8] Wu, R., & Lin, M. (2006). Progress and challenges in genome-wide association studies in humans. *Nature Reviews Genetics* **7(3)** 229–237.

[9] Li, Y., & Wu, R. (2010). Functional mapping of growth and development. *Biological Reviews* **85(2)** 207–216.

[10] Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics* **4(7)** e1000130.

[11] Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (20098). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25(6)** 714–721.

[12] Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L., Peters, U., & Hsu, L. (2010). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *The American Journal of Human Genetics* **86(6)** 860–871.

[13] Logsdon, B. A., Hoffman, G. E., & Mezey, J. G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* **11(1)** 58.

[14] Guan, Y., & Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics* 1780–1815. MR2884922

[15] Dolejsi, E., Bodenstorfer, B., & Frommlet, F. (2014). Analyzing genome-wide association studies with an FDR controlling modification of the Bayesian Information Criterion. *PloS One* **9(7)** e103322.

[16] Li, J., Das, K., Fu, G., Li, R., & Wu, R. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatics* **27(4)** 516–523.

[17] Li, J., Li, R., & Wu, R. (2015). Bayesian group lasso for nonparametric varying coefficient models. *Annals of Applied Statistics* **9(2)** 640–664. MR3371329

[18] Li, J., Zhong, W., Li, R., & Wu, R. (2014). A fast algorithm for detecting gene–gene interactions in genome-wide association studies. *The Annals of Applied Statistics* **8(4)** 2292–2318. MR3292498

[19] He, Q., Berg, A., Li, Y., Vallejos, C. E., & Wu, R. (2010). Mapping genes for plant structure, development and evolution: functional mapping meets ontology. *Trends in Genetics* **26(1)** 39–46.

[20] Li, Q., Huang, Z., Xu, M., Wang, C., Gai, J., Huang, Y., ... & Wu, R. (2010). Functional mapping of genotype-environment interactions for soybean growth by a semiparametric approach. *Plant Methods* **6(1)** 1–11.

[21] Li, Z., Hallingbäck, H. R., Abrahamsson, S., Fries, A., Gull, B. A., Sillanpaa, M. J., & García-Gil, M. R. (2014). Functional multi-locus QTL mapping of temporal trends in scots pine wood traits. *G3: Genes — Genomes — Genetics* **4(12)** 2365–2379.

[22] Lin, M., Aquilante, C., Johnson, J. A., & Wu, R. (2005). Sequencing drug response with HapMap. *The Pharmacogenomics Journal* **5(3)** 149–156.

[23] Das, K., Li, J., Wang, Z., Tong, C., Fu, G., Li, Y., ... & Wu, R. (2011). A dynamic model for genome-wide association studies. *Human Genetics* **129(6)** 629–639.

[24] Das, K., Li, R., Sengupta, S., & Wu, R. (2013). A Bayesian semiparametric model for bivariate sparse longitudinal data. *Statistics in Medicine* **32(22)** 3899–3910. MR3102447

[25] Wang, Y., Tong, C., Wang, Z., Mauger, D., Tantisira, K. G., Israel, E., ... & Wu, R. (2015). Pharmacodynamic genome-wide association study identifies new responsive loci for glucocorticoid intervention in asthma. *The Pharmacogenomics Journal*.

[26] Zhao, W., Ma, C., Cheverud, J. M., & Wu, R. (2004). A unifying statistical model for QTL mapping of genotype sex interaction for developmental trajectories. *Physiological Genomics* **19(2)** 218–227.

[27] Wang, N., Wang, Y., Wang, Z., Hao, H., & Wu, R. (2012). Mapping Body Shape Genes through Shape Mapping. *J Biom Biostat* **3** e121.

[28] Zhao, W., Zhu, J., Gallo-Meagher, M., et al. (2004). A unified statistical model for functional mapping of genotype × environment interactions for ontogenetic development. *Genetics* **168** 1751–1762.

[29] Brown, T. B., Cheng, R., Sirault, X. R., Rungrat, T., Murray, K. D., Trtilek, M., ... & Borevitz, J. O. (2014). TraitCapture: genomic and environment modelling of plant phenomic data. *Current Opinion in Plant Biology* **18** 73–79.

[30] Fan, J., & Lin, S. K. (1998). Test of significance when data are curves. *Journal of the American Statistical Association* **93(443)** 1007–1021. MR1649196

[31] Brockwell, P. J., & Davis, R. A. (2013). *Time series: theory and methods* Springer Science & Business Media.

[32] Srivastava, M. S. (2007). Multivariate theory for analyzing high dimensional data. *Journal of the Japan Statistical Society* **37(1)** 53–86. MR2392485
[33] Fujikoshi, Y., Himeno, T., & Wakaki, H. (2004). Asymptotic results of a high dimensional MANOVA test and power comparison when the dimension is large compared to the sample size. *Journal of the Japan Statistical Society* **34(1)** 19–26. MR2084057
[34] Kong, S. W., Pu, W. T., & Park, P. J. (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics* **22(19)** 2373–2380.
[35] Sun, L., Yang, W., Zhang, Q., Cheng, T., Pan, H., Xu, Z., ... & Chen, C. (2013). Genome-wide characterization and linkage mapping of simple sequence repeats in mei (Prunus mume Sieb. et Zucc.). *PloS One* **8(3)** e59562.
[36] Sun, L., Wang, Y., Yan, X., Cheng, T., Ma, K., Yang, W., ... & Zhang, Q. (2014). Genetic control of juvenile growth and botanical architecture in an ornamental woody plant, Prunus mume Sieb. et Zucc. as revealed by a high-density linkage map. *BMC Genetics* **15(Suppl 1)** S1.
[37] Fu, G., Berg, A., Das, K., Li, J., Li, R., & Wu, R. (2010). A statistical model for mapping morphological shape. *Theor Biol Med Model* **7** 28.
[38] Fu, G., Bo, W., Pang, X., Wang, Z., Chen, L., Song, Y., ... & Wu, R. (2013). Mapping shape quantitative trait loci using a radius-centroid-contour model. *Heredity* **110(6)** 511–519.
[39] Goddard, M. E. & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* **10(6)** 381–391.
[40] Flint, J. & Mackay, T. F. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research* **10(5)** 723–733.
[41] Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics* **40(5)** 489–490.

Han Hao
Department of Mathematics
College of Arts and Sciences
1155 Union Circle #311430
University of North Texas
Denton, Texas 76203
USA
E-mail address: Han.Hao@unt.edu

Lidan Sun
Beijing Key Laboratory
of Ornamental Plants Germplasm Innovation
& Molecular Breeding
National Engineering Research Center for Floriculture
at Beijing Forestry University
Beijing 100083
China
E-mail address: sld656@126.com

Xuli Zhu
Center for Computational Biology
Beijing Forestry University
Beijing 100083
China
E-mail address: 349082942@qq.com

Rongling Wu
Center for Statistical Genetics
The Pennsylvania State University
Hershey, PA 17033
USA
E-mail address: rwu@phs.psu.edu