Mapping quantitative trait loci underlying function-valued traits using functional principal component analysis and multi-trait mapping

Il-Youp Kwak[*,1], Candace R. Moore[†], Edgar P. Spalding[†], Karl W. Broman[‡,2]

Departments of [*]Statistics, [†]Botany, and [‡]Biostatistics and Medical Informatics,

University of Wisconsin–Madison, Madison, Wisconsin 53706

25 Aug 2015

Running head: QTL for function-valued traits

Key words: QTL, function-valued traits, model selection, growth curves, multivariate analysis

[1]**Present address:** Division of Biostatistics, University of Minnesota, Minneapolis, MN, 55455

[2]**Corresponding author:**

Karl W Broman

Department of Biostatistics and Medical Informatics

University of Wisconsin–Madison

2126 Genetics-Biotechnology Center

425 Henry Mall

Madison, WI 53706

Phone: 608–262–4633

Email: kbroman@biostat.wisc.edu

Abstract

We previously proposed a simple regression-based method to map quantitative trait loci underlying function-valued phenotypes. In order to better handle the case of noisy phenotype measurements and accommodate the correlation structure among time points, we propose an alternative approach that maintains much of the simplicity and speed of the regression-based method. We overcome noisy measurements by replacing the observed data with a smooth approximation. We then apply functional principal component analysis, replacing the smoothed phenotype data with a small number of principal components. Quantitative trait locus mapping is applied to these dimension-reduced data, either with a multi-trait method or by considering the traits individually and then taking the average or maximum LOD score across traits. We apply these approaches to root gravitropism data on Arabidopsis recombinant inbred lines and further investigate their performance in computer simulations. Our methods have been implemented in the R package, funqtl.

Introduction

Technology developments have enabled the automated acquisition of numerous phenotypes, included function-valued traits, such as phenotypes measured over time. High-dimensional phenotype data are increasingly considered as part of efforts to map the genetic loci (quantitative trait loci, QTL) that influence quantitative traits.

Numerous methods are available for QTL mapping with function-valued traits. Ma *et al.* (2002) considered parametric models such as the logistic growth model, $g(t) = \frac{a}{1+be^{-rt}}$. The high-dimensional phenotype is reduced to a few parameters. This works well if the parametric model is approximately correct but in many cases the correction functional form is not clear. Yang *et al.* (2009) proposed a non-parametric functional QTL mapping method, with a selected number of basis functions to fit the function-valued phenotype. Min *et al.* (2011) extended this method to multiple-QTL models using Markov chain Monte Carlo (MCMC). Sillanpaa *et al.* (2012) proposed another Bayesian multiple-QTL mapping method based on hierarchical modeling. Xiong *et al.* (2011) proposed an additional non-parametric functional mapping method based on estimating equations. An important barrier to these methods is the long computation time required for the analysis.

In Kwak *et al.* (2014), we proposed two simple regression-based methods to map quantitative trait loci underlying function-valued phenotypes, based on the results from the individual analysis of the phenotypes at each time point. With the SLOD score, we take the average of the LOD scores across time points, and with the MLOD score, we take the maximum. These approaches are fast to compute, work well when the trait data are smooth, and provide results that are easily interpreted. Another important advantage is the ability to consider multiple QTL, which can

improve power and enable the separation of linked QTL. However, the approaches do not work as well when the trait data are not smooth, and they do not take account of the correlation among time points.

In the present paper, we describe methods to overcome these weaknesses. First, we replace the observed trait data with a smooth approximation. Second, we apply functional principal component analysis (PCA) as a dimension-reduction technique, and replace the smoothed phenotype data with a small number of principal components.

QTL analysis is then performed on these dimension-reduced data. We consider either the multivariate QTL mapping method of Knott and Haley (2000), or use the SLOD or MLOD scores, as in Kwak *et al.* (2014). These methods all have analogs for multiple-QTL models, by extending the penalized LOD scores of Broman and Speed (2002) and Manichaikul *et al.* (2009).

We illustrate these methods by application to the root gravitropism data of Moore *et al.* (2013), measured by automated image analysis over a time course of 8 hr across a population of *Arabidopsis thaliana* recombinant inbred lines (RIL). We further investigate the performance of these approaches in computer simulations.

Methods

We will focus on the case of recombinant inbred lines, with genotypes AA or BB. Consider n lines, with function-valued phenotypes measured at T discrete time points $(t_1, \cdots, t_T)$.

Smoothing

We first smooth the phenotype data for each individual. Let $y_i(t_j)$ denote the observed phenotype for individual i at time t_j. We assume underlying smooth curves $x_i(t)$, with $y_i(t_j) = x_i(t_j) + e_i(t_j)$.

We approximate the functional form of $x_i(t)$ as $x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$ for a set of basis functions $\phi_1(t), \cdots, \phi_K(t)$, where the number of basis functions, K, is generally much smaller than the number of time points, T. There are many possible choices of basis functions. We are using B-splines (Ramsay and Silverman 2005, pp. 49–53).

Define the K by T matrix Φ, with $\Phi_{kj} = \phi_k(t_j)$, for $k = 1, \ldots, K$ and $j = 1, \ldots, T$. We estimate the coefficients, c_{ik} by minimizing the least squares criterion (Ramsay and Silverman 2005):

$$\mathrm{SMSSE}(\mathbf{y}|\mathbf{c}) = \sum_{i=1}^{n}\sum_{j=1}^{T}\left[y_i(t_j) - \sum_{k=i}^{K} c_{ik}\phi_k(t_j)\right]^2 = (\mathbf{y} - \Phi\mathbf{c})'(\mathbf{y} - \Phi\mathbf{c}).$$

This gives the solution $\hat{\mathbf{c}} = (\Phi'\Phi)^{-1}\Phi'\mathbf{y}$. We then obtain the smoothed phenotypes $\hat{\mathbf{y}} = \Phi\hat{\mathbf{c}}$, which are used in all subsequent analyses.

A key issue is the choice of the number of basis functions. We use 10-fold cross-validation and choose the number of basis functions that minimizes the estimated sum of squared errors.

Functional principal component analysis

Having replaced the phenotype data with a smooth approximation, $\hat{y}_i(t)$, we use functional

principal component analysis (Ramsay and Silverman 2005) to reduce the dimensionality with little loss of information.

In functional PCA, we seek a sequence of orthonormal functions, $\psi_j(t)$, that take the role of principal components but for functional data. By orthonormal, we mean $\langle \psi_j(t), \psi_k(t) \rangle = \int \psi_j(t)\psi_k(t)\,dt = 0$, for all $j \neq k$, and $\|\psi_j(t)\|^2 = \langle \psi_j(t), \psi_j(t) \rangle = 1$ for all j.

With our smoothed functional data, $\hat{y}_i(t)$, we first find the function $\psi_1(t)$ that maximizes $\sum_i (\langle \hat{y}_i(t), \psi_1(t) \rangle)^2 = \sum_i (\int \hat{y}_i(t)\psi_1(t)\,dt)^2$, subject to the constraint $\|\psi_1(t)\| = 1$. The computations make use of the B-spline basis representation of the $\hat{y}_i(t)$.

Conceptually, the procedure then proceeds inductively. Having identified $\psi_1, \ldots, \psi_j$, we choose ψ_{j+1} that maximizes $\sum_i (\langle \hat{y}_i(t), \psi_1(t) \rangle)^2$ subject to the constraint that $\psi_1, \ldots, \psi_{j+1}$ are orthonormal.

We focus on a small number, p, of functional PCs that explain 99% of the data variation and consider the coefficients $\langle \hat{y}_i(t), \psi_j(t) \rangle$, as derived traits.

Single-QTL analysis

Having smoothed and dimension-reduced the phenotype data, we then use the p principal components as derived traits for QTL analysis, using one of three methods. First, we apply the multivariate QTL mapping method of Knott and Haley (2000). Our second and third approaches are to analyze the p derived traits individually and take the average or maximum LOD score, respectively, at each putative QTL position, as in Kwak *et al.* (2014).

HKLOD score: In the method of Knott and Haley (2000), we take $n \times p$ matrix of derived phenotypes and scan the genome, and at each position, λ, we fit a multivariate regression model with a single QTL. The basic model is $Y = XB + E$, where Y is the $n \times p$ matrix of phenotypes, X

is an $n \times 2$ matrix of QTL genotype probabilities, and B is a $2 \times p$ matrix of QTL effects. The rows of the $n \times p$ matrix of errors, E, are assumed to be independent and identically distributed draws from a multivariate normal distribution.

The maximum likelihood estimate of the coefficients is $\hat{B} = (X'X)^{-1}X'Y$, the same as if the traits were analyzed separately. A key component of the likelihood is the matrix of sums of squares and cross-products of residuals, $\text{RSS} = (Y - XB)'(Y - XB)$, and particularly its determinant, $|\text{RSS}|$. The $\log_{10}$ likelihood ratio comparing the model with a single QTL at position λ, to the null model of no QTL, is

$$\text{HKLOD} = \frac{n}{2} \log_{10} \left\{ \frac{|\text{RSS}_0|}{|\text{RSS}(\lambda)|} \right\},$$

where $|\text{RSS}_0|$ is for the null model with no QTL and $|\text{RSS}(\lambda)$ is for the model with a single QTL at λ.

SL and ML scores: As further approaches, we apply the method of Kwak *et al.* (2014) to the p derived traits. We perform a genome scan by Haley-Knott regression (Haley and Knott 1992) with each trait separately, giving the $LOD_j(\lambda)$ for trait j at position λ.

Our second criterion, is to take the average LOD score, across traits: $\text{SL}(\lambda) = \frac{1}{p} \sum_{j=1}^{p} \text{LOD}_j(\lambda)$. We call this the SL score, to distinguish it from SLOD of Kwak *et al.* (2014), calculated using the original phenotype data.

Our third criterion is to take the maximum LOD score, $\text{ML}(\lambda) = \max_j \text{LOD}_j(\lambda)$. We call this the ML score, to distinguish it from the MLOD score of Kwak *et al.* (2014).

Multiple-QTL analysis

As in Kwak *et al.* (2014), we use the penalized LOD score criterion on Broman and Speed (2002) to extend each of the LOD-type statistics defined above, for use with multiple-QTL models. The penalized LOD score is $\mathrm{pLOD}(\gamma) = \mathrm{LOD}(\gamma) - \mathrm{T}|\gamma|$, where γ denotes a multiple-QTL model with strictly additive QTL, and $|\gamma|$ is the number of QTL in the model γ. T is a penalty on model size, chosen as the $1-\alpha$ quantile of the genome-wide maximum LOD score under the null hypothesis of no QTL, derived from a permutation test (Churchill and Doerge 1994). We may replace the LOD score in the above equation with any of the HKLOD, SL and ML scores.

To search the space of models, we use the stepwise model search algorithm of Broman and Speed (2002): we use forward selection up to a model of fixed size (e.g., 10 QTL), followed by backward elimination to the null model. The selected model $\hat{\gamma}$ is that which maximizes the penalized LOD score criterion, among all models visited.

The selected model is of the form $Z = Q\beta + \varepsilon$, where Z contains the derived traits (the coefficients from the functional principal component analysis) and Q contains an intercept column and genotype probabilities at the inferred QTL. The derived phenotypes, Z, are linearly related to the smoothed phenotypes, $\hat{Y}$, by the equation $\hat{Y} = Z\Psi$, where Ψ is a matrix with (i,j)th element $\psi_i(t_j)$. Thus we have $\hat{Y} = Z\Psi = Q(\beta\Psi) + \varepsilon'$, and so $\hat{\beta}\Psi$ are the estimated QTL effects, translated back to the time domain (see Figure 3, below).

Application

As an illustration of our approaches, we considered data from Moore *et al.* (2013) on gravitropism in Arabidopsis recombinant inbred lines (RIL), Cape Verde Islands (Cvi) $\times$ Landsberg erecta (Ler). For each of 162 RIL, 8–20 replicate seeds per line were germinated and then rotated 90 degrees, to change the orientation of gravity. The growth of the seedlings was captured on video, over the course of eight hours, and a number of phenotypes were derived by automated image analysis.

We focus on the angle of the root tip, in degrees, over time (averaged across replicates within an RIL), and consider only the first of two replicate data sets examined in Moore *et al.* (2013). There is genotype data at 234 markers on five chromosomes; the function-valued root tip angle trait was measured at 241 time points (every two minutes for eight hours).

The data are available at the QTL Archive, which is now part of the Mouse Phenome Database, as the *Moore1b* data set:

`http://phenome.jax.org/db/q?rtn=projects/projdet&reqprojid=282`

Single-QTL analysis

We first performed genome scans with a single-QTL model by the multiple methods: SLOD and MLOD from Kwak *et al.* (2014), EE(Wald) and EE(Residual) from Xiong *et al.* (2011), and the HKLOD, SL, and ML methods described above (and using four principal components). We used a permutation test (Churchill and Doerge 1994) with 1000 permutation replicates to estimate 5% significance thresholds, which are shown in Supporting Information, Table S1.

The results are shown in Figure 1. The SLOD method (Figure 1A) gave similar results to the EE(Residual) method (Figure 1D), with significant evidence for QTL on chromosomes 1, 4, and

5. The MLOD method (Figure 1B) also showed evidence for a QTL on chromosome 3. The EE(Wald), HKLOD, and SL methods (Figure 1C, 1E, and 1F) all gave similar results, with significant evidence for QTL on each chromosome. The ML method (Figure 1G) is different, with significant evidence for QTL on chromosomes 1, 2, and 4.

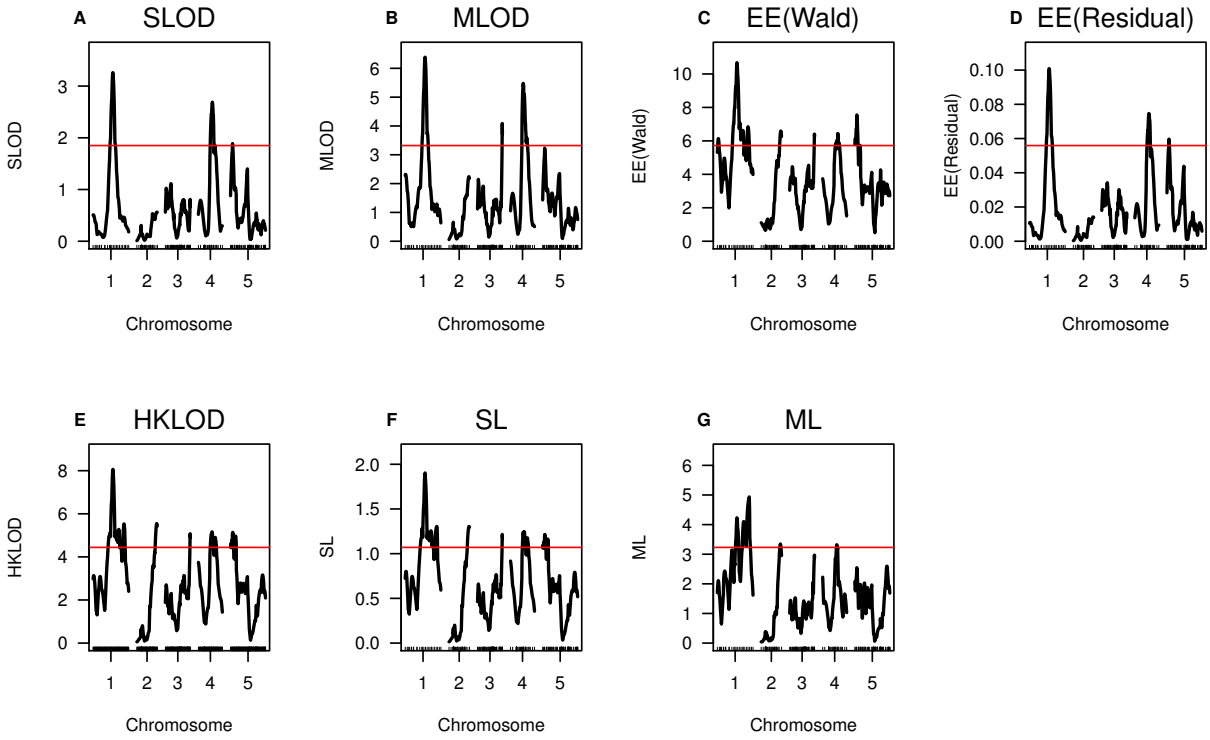

Figure 1: The SLOD, MLOD, EE(Wald) and EE(Residual), HKLOD, SL and ML curves for the root tip angle data. A red horizontal line indicates the calculated 5% permutation-based threshold.

Multiple-QTL analysis

We further applied multiple-QTL analysis, extending the HKLOD, SL, and ML methods to use the penalized LOD score criterion of Broman and Speed (2002) for function-valued traits. We focused on additive QTL models, and we used the 5% permutation-based significance thresholds (Table S1) as the penalties.

The penalized-HKLOD and penalized-SL criteria each indicated a five-QTL model, with a QTL on each chromosome. The inferred positions of the QTL showed only slight differences. The penalized-SL criterion indicated a two-QTL model, with QTL on chromosomes 1 and 4.

LOD profiles for these models are displayed in Figure 2. These curves, which visualize both the evidence and localization of each QTL in the context of a multiple-QTL model, are calculated following an approach developed by Zeng *et al.* (2000): The position of each QTL was varied one at a time, and at each location for a given QTL, we derived a LOD-type score comparing the multiple-QTL model with the QTL under consideration at a particular position and the locations of all other QTL fixed, to the model with the given QTL omitted. For the SL (or ML) method, the profile is calculated for the four derived traits, individually, and then the SL (or ML) profiles are obtained by averaging (or maximizing) across traits. For the HKLOD method, the profiles are calculated using the multivariate LOD test statistic.

In Kwak *et al.* (2014), we applied, to these same data, the analogous multiple-QTL modeling approach, with the penalized-SLOD and penalized-MLOD criteria (that is, using the original phenotype data, without the smoothing and dimension-reduction steps). The result (see Figure 3 in Kwak *et al.* 2014) with the penalized-SLOD criterion was the same two-QTL model identified by ML (Figure 2C), while the penalized-MLOD criterion gave a three-QTL model, with a QTL on chromosome 3, similar to that inferred by single-QTL analysis with MLOD (Figure 1B).

The estimated QTL effects, translated back to the time domain, are displayed in Figure 3. The red curves are for the five-QTL model identified with the penalized-HKLOD and penalized-SL criteria. The blue dashed curves are for the two-QTL identified with the penalized-ML criterion. The estimated QTL effects in panels B–F are for the difference between the Cvi allele and the Ler

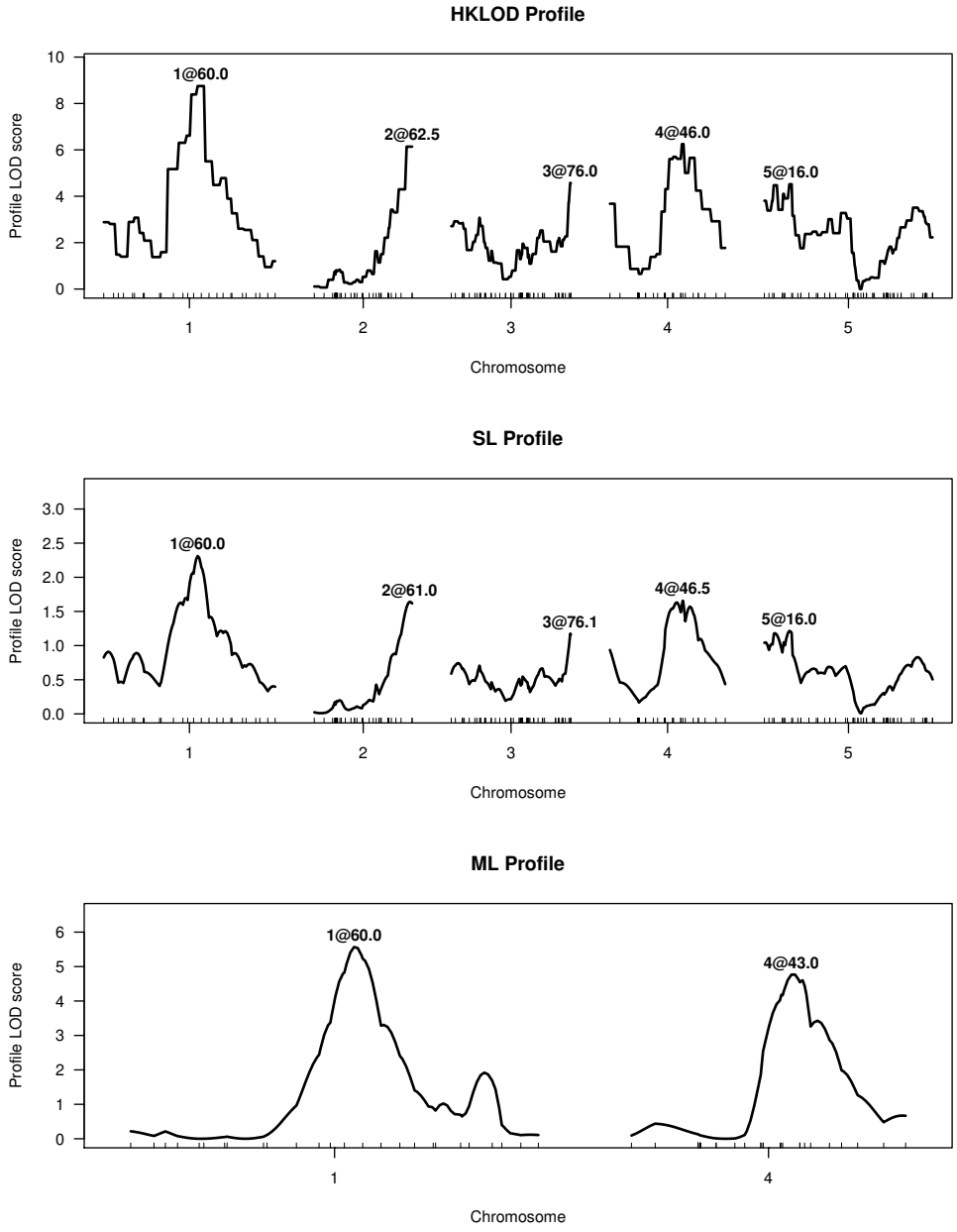

Figure 2: HKLOD, SL, and ML profiles for a multiple-QTL model with the root tip angle data set.

allele.

The effects of the QTL on chromosomes 1 and 4 are approximately the same, whether or not the chromosome 2, 3 and 5 QTL are included in the model. The chromosome 1 QTL has greatest

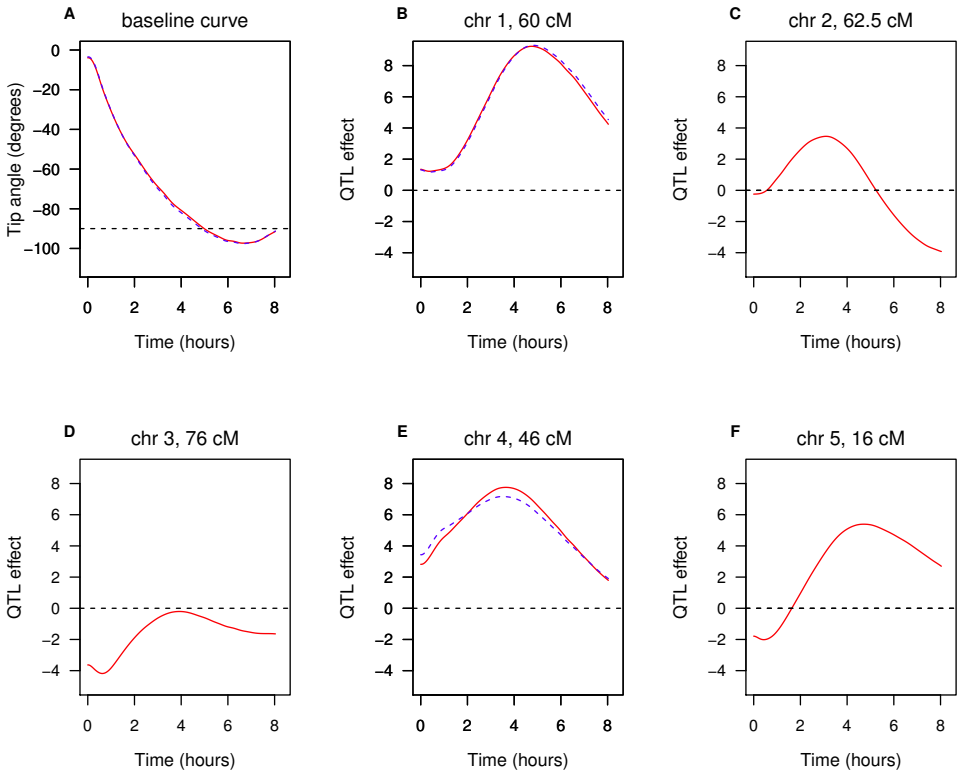

Figure 3: The estimated QTL effects for the root tip angle data set. The red curves are for the five-QTL model (from the penalized-HKLOD and penalized-SL criterion) and the blue dashed curves are for the two-QTL model (from the penalized-ML criterion). Positive values for the QTL effects indicate that the Cvi allele increases the tip angle phenotype.

effect at later time points, while the chromosome 4 QTL has greatest effect earlier and over a wider interval of time. For both QTL, the Cvi allele increases the root tip angle phenotype.

In summary, the HKLOD and SL methods gave similar results. For these data, the HKLOD and SL methods indicate evidence for a QTL on each chromosome. The results suggest that these approaches, with the initial dimension-reduction via functional PCA, may have higher power to detect QTL, as the multiple-QTL analysis with SLOD and MLOD, without dimension-reduction, indicated fewer QTL.

Simulations

In order to investigate the performance of our proposed approaches and compare them to existing methods, we performed computer simulation studies with both single-QTL models and models with multiple QTL. For the simulations with a single-QTL model, we compared the HKLOD, SL, and ML methods to the SLOD and MLOD methods of Kwak *et al.* (2014) and the estimating equation approaches of Xiong *et al.* (2011). For the simulations with multiple QTL, we omitted the methods of Xiong *et al.* (2011), as they have not yet been implemented for multiple QTL.

Single-QTL models

To compare approaches in the context of a single QTL, we considered the simulation setting described in Yap *et al.* (2009), though exploring a range of QTL effects.

We simulated an intercross with sample sizes of 100, 200, or 400, and a single chromosome of length 100 cM, with 6 equally spaced markers and with a QTL at 32 cM. The associated phenotypes was sampled from a multivariate normal distribution with the mean curve following a logistic function, $g(t) = \frac{a}{1+be^{-rt}}$. The AA genotype had $a = 29, b = 7, r = 0.7$; the AB genotype had $a = 28.5, b = 6.5, r = 0.73$; and the BB genotype had $a = 27.5, b = 5, r = 0.75$. The shape of the growth curve with this parameter was shown in Figure S3 of Kwak *et al.* (2014). The phenotype data were simulated at 10 time points.

The residual error was assumed to following multivariate normal distributions with a covariance structure $c\Sigma$. The constant c controls the overall error variance, and Σ was chosen to have one of three forms: (1) auto-regressive with $\sigma^2 = 3, \rho = 0.6$, (2) equicorrelated with $\sigma^2 = 3, \rho = 0.5$, or (3) an "unstructured" covariance matrix, as given in Yap *et al.* (2009) (also

shown in Table S2 of Kwak *et al.* 2014)).

The parameter c was given a range of values, which define the percent phenotypic variance explained by the QTL (the heritability). The effect of the QTL varies with time; we took the mean heritability across time as an overall summary. For the auto-regressive and equicorrelated covariance structures, we used $c = 1, 2, 3, 6$; for the unstructured covariance matrix, we took $c = 0.5, 1, 2, 3$.

For each of 10,000 simulation replicates, we applied the previous SLOD and MLOD methods of Kwak *et al.* (2014), the EE(Wald) and EE(Residual) methods of Xiong *et al.* (2011), and the HKLOD, SL, and ML methods proposed here. For all seven approaches, we fit a three-parameter QTL model (that is, allowing for dominance).

The estimated power to detect the QTL as a function of heritability due to the QTL, for $n = 100, 200, 400$ and for the three different covariance structures, is shown in Figure 4. With the autocorrelated variance structure, all methods worked well, though HKLOD had noticeably lower power. With the equicorrelated variance structure, EE(Wald), SL and ML methods had higher power than the other four methods. The HKLOD method also worked reasonably well, but the EE(Residual), SLOD and MLOD methods had low power. In the unstructured variance setting, EE(Wald), MLOD, SL and ML methods worked better than the other four methods. EE(Residual) did not work well in this setting.

The precision of QTL mapping, measured by the root mean square error in the estimated QTL position, is displayed in Figure S1. Performance, in terms of precision, corresponds quite closely to the performance in terms of power: when power is high, the RMS error of the estimated QTL position is low, and vice versa.

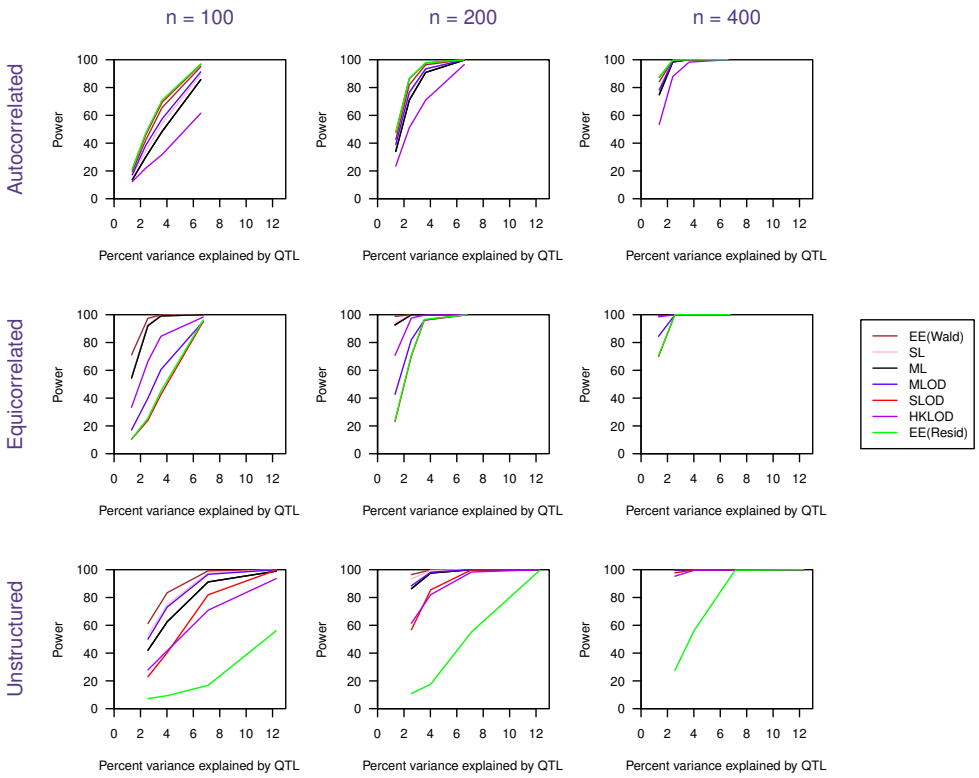

Figure 4: Power as a function of the percent phenotypic variance explained by a single QTL. The first column is for $n = 100$, the second column is for $n = 200$ and the third column is for $n = 400$. The three rows correspond to the covariance structure (autocorrelated, equicorrelated, and unstructured). In each panel, SLOD is in red, MLOD is in blue, EE(Wald) is in brown, EE(Residual) is in green, HKLOD is in purple, SL is in pink, and ML is in black.

A possible weakness of SLOD and MLOD approaches was that they do not make use of the function-valued form of the phenotypes. The methods may further suffer lower power in the case of noisy phenotypes. The methods proposed in this paper use smoothing to handle the measurement error. We repeated the same simulations in Kwak *et al.* (2014) with $n = 200$, adding independent, normally distributed errors (with standard deviation 1 or 2) at each time point.

The estimated power to detect the QTL as a function of heritability due to the QTL, for added noise with SD $= 0, 1, 2$ and the three different covariance structures, is shown in Figure 5. The power of the SLOD and MLOD were greatly affected by the introduction of noise. However, the

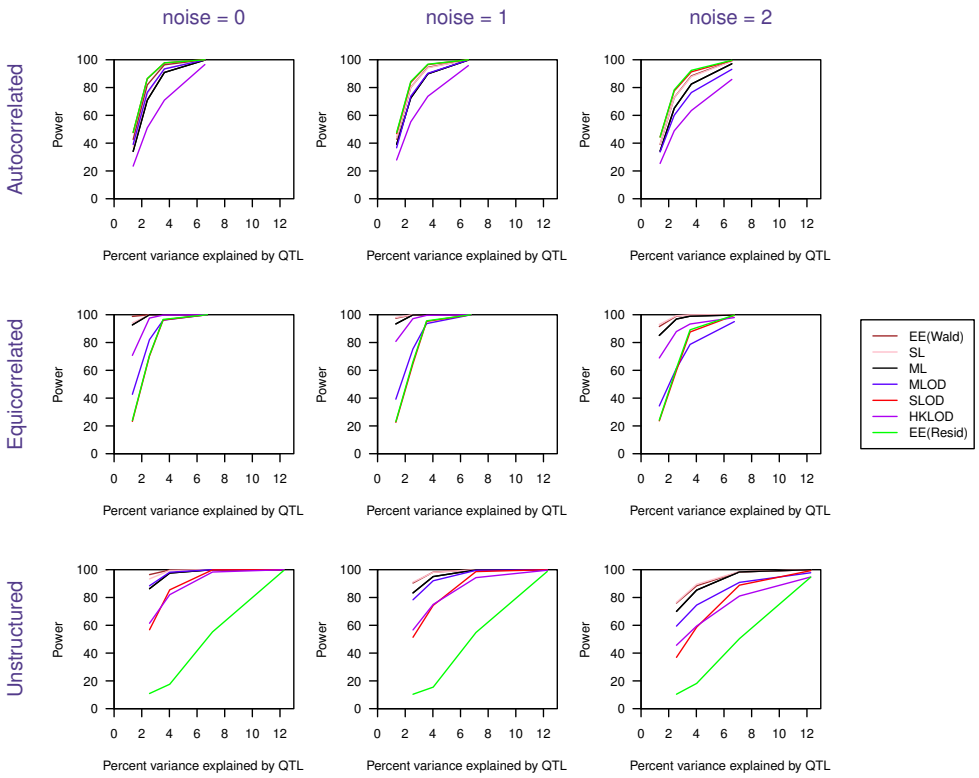

Figure 5: Power as a function of the percent phenotypic variance explained by a single QTL, with additional noise added to the phenotypes. The first column has no additional noise; the second and third columns have independent normally distributed noise added at each time point, with standard deviation 1 and 2, respectively. The three rows correspond to the covariance structure (autocorrelated, equicorrelated, and unstructured). In each panel, SLOD is in red, MLOD is in blue, EE(Wald) is in brown, EE(Residual) is in green, HKLOD is in purple, SL is in pink, and ML is in black. The percent variance explained by the QTL on the x-axis refers, in each case, to the variance explained in the case of no added noise.

SL and ML methods, which used both smoothing and dimension-reduction, work well in the presence of noise. The EE(Wald) method performed well in this case, as well. The EE(Residual) method did not work well compared to the other six methods. Overall, the EE(Wald) method continued to perform best.

Table 1 shows the average computation time for SLOD/MLOD, EE(Wald), EE(Residual), HKLOD and SL/ML methods. For the single-QTL simulations, the computation time for the

Table 1: Average computation time for each method, in the single-QTL simulations and in the application.

	Ave. computation time (in sec.)	
Method	Simulations	Application
SLOD/MLOD	0.011	0.96
EE(Wald)	0.853	224.4
EE(Residual)	0.030	3.97
HKLOD	0.013	0.21
SL/ML	0.114	0.21

SLOD/MLOD and HKLOD methods are similar, while the SL/ML methods are somewhat slower. This is because, in the simulation data set, the phenotype was measured at 10 time points. In the application, with 241 time points, the functional PCA based methods (HKLOD and SL/ML) are faster than the SLOD/MLOD methods. The EE(Wald) method requires considerably longer computation time; on the other hand, it provided the highest power to detect QTL.

Multiple-QTL models

To investigate the performance of the penalized-HKLOD, penalized-SL and penalized-ML criteria in the context of multiple QTL, we used the same setting as Kwak *et al.* (2014). We simulated data from a three-QTL model modeled after that estimated, with the penalized-MLOD criterion, for the root tip angle data of Moore *et al.* (2013) considered in the Application section.

We assumed that the mean curve for the root tip angle phenotype followed a cubic polynomial, $y = a + bt + ct^2 + dt^3$, and assumed that the effect of each QTL also followed such a

cubic polynomial. The four parameters for a given individual were drawn from a multivariate normal distribution with mean defined by the QTL genotypes and variance matrix estimated from the root tip angle data. Details on the parameter values used in the simulations appear in Kwak *et al.* (2014).

Normally distributed measurement error (with mean 0 and variance 1) was added to the phenotype at each time point for each individual. Phenotypes are taken at 241 equally spaced time points in the interval of 0 to 1. We considered two sample sizes: $n = 162$ (as in the Moore *et al.* (2013) data) and twice that, $n = 324$.

We performed 100 simulation replicates. For each replicate, we applied a stepwise model selection approach with each of the penalized-HKLOD, penalized-SL, penalized-ML, penalized-SLOD, and penalized-MLOD criteria. The simulation results are shown in Table 2.

The penalized-SL and penalized-HKLOD criteria had higher power to detect all three QTL. In particular, the power to detect the QTL on chromosome 3 was greatly increased, in comparison to the penalized-SLOD and penalized-MLOD criteria.

Table 2: Power to detect QTL, for a three-QTL model modeled after the Moore *et al.* (2013) data.

	QTL position		Power				
	chr	cM location	HKLOD	SL	ML	SLOD	MLOD
n=162	1	61	82	86	65	89	54
	3	76	24	27	2	12	15
	4	40	88	94	60	82	77
n=324	1	61	100	100	94	100	59
	3	76	78	77	13	31	43
	4	40	100	100	93	100	91

Discussion

We have described two techniques for improving the regression-based methods of Kwak *et al.* (2014) for QTL mapping with function-valued phenotypes: smoothing and dimensional-reduction. Smoothing leads to better performance in the case of noisy phenotype measurements, and dimension-reduction improves power. The particular methods we used (smoothing via B-splines, and dimensional reduction via functional principal component analysis) are not the only possibilities, but they are natural choices widely used in functional data analysis (Ramsay and Silverman 2005).

Following smoothing and dimensional-reduction, we applied QTL analysis to the small number of derived traits, either by analyzing the traits individually and then combining the log likelihoods, as in Kwak *et al.* (2014), or by applying the multivariate QTL mapping method of Knott and Haley (2000). The latter approach cannot be applied directly to the original phenotypes, due to the large number of time points at which the traits were measured, but it can work well with the dimension-reduced derived traits.

Key advantages of our proposed methods include speed of computation and the ability to consider multiple-QTL models. The EE(Wald) method of Xiong *et al.* (2011), based on estimating equations, was seen to be most powerful for QTL detection in our simulation study, but it is orders of magnitude slower and has not yet been implemented for multiple-QTL models.

Many other methods have been developed for QTL mapping with function-valued traits. However, most focus on single-QTL models (e.g., Ma *et al.* 2002; Yang *et al.* 2009; Yap *et al.* 2009). Bayesian methods for multiple-QTL mapping with function-valued traits have been proposed (Min *et al.* 2011; Sillanpaa *et al.* 2012), but these methods are computationally

intensive, and software is not available.

In considering multiple-QTL models, we have focused on strictly additive QTL. Manichaikul *et al.* (2009) extended the work of Broman and Speed (2002) by considering pairwise interactions among QTL. Our approaches may be similarly extended to handle interactions.

The enormous recent growth in capabilities for high-throughput phenotyping, including images and time series, particularly in plants (see, for example, Cabrera-Bosquet *et al.* 2012; Araus and Cairns 2014; Ghanem *et al.* 2015), is accompanied by a growth in interest in the genetic analysis such phenotype data. Speed of computation will be particularly important in the analysis of such high-dimensional data, as will the joint consideration of multiple loci. The methods we have proposed can meet many of these challenges.

We implemented our methods as a package, funqtl, for the general statistical software R (R Core Team 2015). Our package makes use of the fda package (Ramsay *et al.* 2014) for smoothing and functional PCA. It is available at `https://github.com/ikwak2/funqtl`.

Acknowledgments

The authors thank Nathan Miller for suggestions to improve the manuscript. This work was supported in part by grant IOS-1031416 from the National Science Foundation Plant Genome Research Program to E.P.S. and by National Institutes of Health grant R01GM074244 to K.W.B.

Literature Cited

Araus, J. L., and J. E. Cairns, 2014 Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci. **19**: 52–61.

Broman, K. W., and T. P. Speed, 2002 A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). J. R. Statist. Soc. B **64**: 641–656, 737–775.

Cabrera-Bosquet, L., J. Crossa, J. von Zitzewitz, M. D. Serret, and J. Luis Araus, 2012 High-throughput phenotyping and genomic selection: The frontiers of crop breeding converge. J. Integr. Plant Biol. **54**: 312–320.

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

Ghanem, M. E., H. Marrou, and T. R. Sinclair, 2015 Physiological phenotyping of plants for crop improvement. Trends Plant Sci. **20**: 139–144.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**: 315–324.

Knott, S. A., and C. S. Haley, 2000 Multitrait least squares for quantitative trait loci detection. Genetics **156**: 899–911.

Kwak, I.-Y., C. R. Moore, E. P. Spalding, and K. W. Broman, 2014 A simple regression-based method to map quantitative trait loci underlying function-valued phenotypes. Genetics **0**: 0.

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

Manichaikul, A., J. Y. Moon, S. Sen, B. S. Yandell, and K. W. Broman, 2009 A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. Genetics **181**: 1077–1086.

Min, L., R. Yang, X. Wang, and B. Wang, 2011 Bayesian analysis for genetic architecture of dynamic traits. Herdity **106**: 124–133.

Moore, C. R., L. S. Johnson, I.-Y. Kwak, M. Livny, K. W. Broman *et al.*, 2013 High-throughput computer vision introduces the time axis to a quantitative trait map of a plant growth response. Genetics **195**: 1077–1086.

R Core Team, 2015 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J., and B. W. Silverman, 2005 *Functional Data Analysis Ed. 2*. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker, 2014 *fda: Functional Data Analysis*. R package version 2.4.4, `http://cran.r-project.org/package=fda`.

Sillanpaa, M. J., P. Pikkuhookana, S. Abrahamsson, T. Knurr, A. Fries *et al.*, 2012 Simultaneous estimation of multiple quantitative trait loci and growth curve parameters through hierarchical bayesian modeling. Herdity **108**: 134–146.

Xiong, H., E. H. Goulding, E. J. Carlson, L. H. Tecott, C. E. Mcculloch *et al.*, 2011 A flexible estimating equations approach for mapping function-valued traits. Genetics **189**: 305–316.

Yang, J., R. L. Wu, and G. Casella, 2009 Nonparametric functional mapping of quantitative trait loci. Biometrics **65**: 30–39.

Yap, J. S., J. Fan, and R. Wu, 2009 Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. Biometrics **65**: 1068–1077.

Zeng, Z. B., J. J. Liu, L. F. Stam, C. H. Kao, J. M. Mercer *et al.*, 2000 Genetic architecture of a morphological shape difference between two Drosophila species. Genetics **154**: 299–310.

Mapping quantitative trait loci underlying function-valued traits using functional principal component analysis and multi-trait mapping

SUPPLEMENT

Il-Youp Kwak*, Candace R. Moore†, Edgar P. Spalding†, Karl W. Broman‡

Departments of *Statistics, †Botany, and ‡Biostatistics and Medical Informatics,
University of Wisconsin--Madison, Madison, Wisconsin 53706

25 Aug 2015

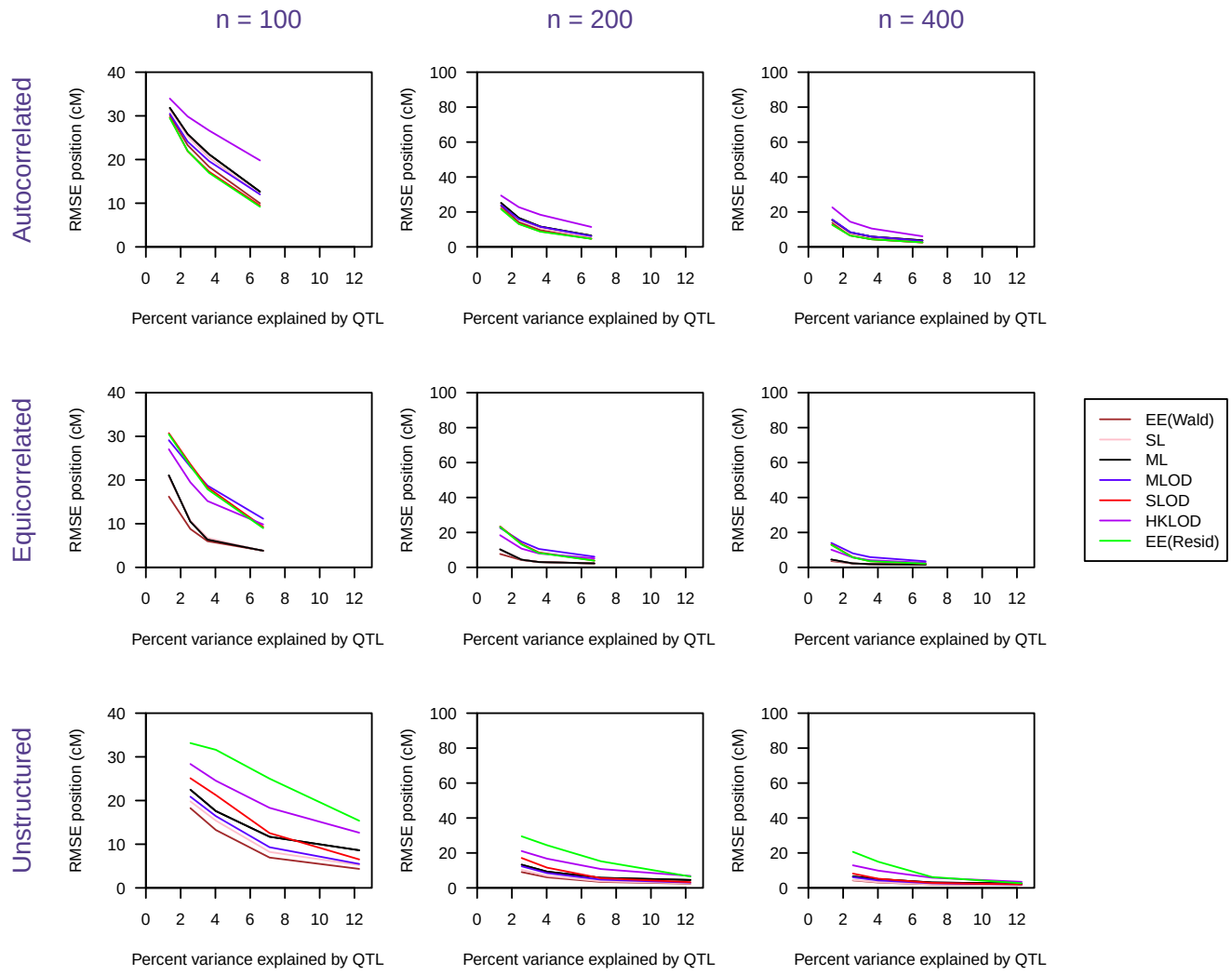

Figure S1 Root Mean Square Error (RMSE) of the estimated QTL position as a function of the percent variance explained by a single QTL. The first column is for $n = 100$, the second column is for $n = 200$ and the third column is for $n = 400$. The three rows correspond to the covariance structure (autocorrelated, equicorrelated, and unstructured). In each panel, SLOD is in red, MLOD is in blue, EE(Wald) is in brown, EE(Residual) is in green, HKLOD is in purple, FLOD is in gray, SL is in pink, and ML is in black.

Table S1 5% significance thresholds for the data from Moore *et al.* 2013, based on a permutation test with 1000 replicates.

Method	Threshold
HKLOD	4.44
SL	1.07
ML	3.23
SLOD	1.85
MLOD	3.32
EE(Wald)	5.72
EE(Residual)	0.0559