Published in final edited form as:
J Stat Plan Inference. 2009 March 1; 139(3): 978–989. doi:10.1016/j.jspi.2008.06.009.

A Robust QTL Mapping Procedure

Fei Zou[*,1,2], **Lei Nie**[**], **Fred A. Wright**[*], and **Pranab K. Sen**[*]

[*] Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

[**] Georgetown University Medical Center, Washington District of Columbia 20057

Abstract

In quantitative-trait linkage studies using experimental crosses, the conventional normal location-shift model or other parameterizations may be unnecessarily restrictive. We generalize the mapping problem to a genuine nonparametric setup and provide a robust estimation procedure for the situation where the underlying phenotype distributions are completely unspecified. Classical Wilcoxon-Mann-Whitney statistics are employed for point and interval estimation of QTL positions and effects.

Keywords

GEE; Generalized least squares estimate; Quantitative trait; Weighted least squares estimate; Wilcoxon- Mann-Whitney statistic

1 Introduction

Genetic mapping of quantitative trait loci (QTL) has fundamental importance in revealing the genetic basis of phenotypic differences (Belknap *et al.*, 1997; Haston *et al.*, 2002; Wang *et al.*, 2003). In plants and laboratory animals, backcross or F2 intercross populations are widely used for mapping quantitative traits (see Lynch and Walsh 1998 for details). In QTL mapping, the basic problems are to test the existence of one or more QTLs, and to estimate the QTL map position and effect if there is evidence of linkage to a chromosomal region. QTL mapping methodologies, including the single marker t-tests (Sax 1923) and likelihood interval mapping (Lander and Botstein 1989; Haley and Knott 1992; Kruglyak and Lander 1995), have traditionally relied on parametric assumptions. In Kruglyak and Lander (1995), a nonparametric approach has been explored for testing linkage, but cannot produce QTL confidence intervals or specify effect sizes. Zou *et al.* (2002) proposed a semiparametric model that specifies an exponential tilt relationship between phenotype densities for different genotypes at the QTL.

In standard parametric linkage scans, the (profile) likelihood ratio test statistic is calculated for each position, and the maximum likelihood estimate (MLE) used as a point estimate for the QTL position. A difficulty in the use of the MLE in this setting is that it may exhibit non-standard asymptotic behavior, depending on the asymptotic regime used (Kong and Wright 1994). For realistic sample sizes and marker densities, the consequences are that the MLE of

1*Correspoding author. Fax: 1-919-966-3804 and e-mail address: fzou@bios.unc.edu.
2This work was supported by National Institutes of Health (NIH) grant MH070504 to F.Z..

the QTL position might not be efficient and accurate confidence intervals are not readily available from the profile likelihood in the vicinity of the MLE. However, the reporting of plausible intervals is important (Flaherty *et al.* 2003). A number of approximate methods have been described, including LOD-drop intervals (Lander and Botstein 1989), which may have unreliable coverage (Dupuis and Siegmund, 1999), and formulae in Darvasi and Soller (1997) for 95% confidence intervals based on their extensive simulations. Other computation-intensive approaches include bootstrapping (Visscher *et al.* 1996) and the method of Mangin *et al.* (1994), which requires simulation to obtain an asymptotic distribution of a test statistic.

For backcross population, Kearsey and Hyne (1994), Wu and Li (1994, 1996) proposed a multipoint mapping by modeling the mean phenotype difference between two genotype groups at a marker as a function of the recombination frequency between that locus and a putative QTL. Their approach jointly uses the information of every marker on a chromosome. Instead of working on the profile likelihood across genomic positions, they proposed several least squares methods to estimate the QTL position and its effect simultaneously. Therefore, both the detection of the QTL and its position (with correct confidence intervals) are done simultaneously. Liang *et al.* (2001A, B) proposed a similar multipoint mapping of complex diseases for affected sib pair studies. The method carries out a parametric inference procedure to locate a susceptibility gene, using generalized estimating equations (GEE) to model the expected identical by descent (IBD) allele sharing on all genotyped markers at once with the ultimate goal of locating the susceptible gene more robustly.

The objectives of the current study are to extend the procedure of Kearsey and Hyne (1994), Wu and Li (1994, 1996) to relax stringent model assumptions on the underlying phenotype distributions. Our proposed method differs from the approach of Kearsey and Hyne (1994), Wu and Li (1994, 1996) in several ways. First, they considered mean phenotype differences at each marker while we calculate the rank difference of phenotype at each marker, which as shown later, increases mapping efficiency dramatically. Second, we directly express the covariance matrix analytically in terms of several meaningful parameters, while Kearsey and Hyne (1994), Wu and Li (1994, 1996) did not. To simplify the illustration, we describe the method for backcross populations as done in Kearsey and Hyne (1994), Wu and Li (1994, 1996).

The paper is organized as described below. Section 2 formulates the estimation procedures. Simulation studies in Section 3 demonstrate the properties of the proposed method and its utility. The discussion section describes extensions and suggestions for future work.

2 Methodology

Consider a backcross experiment with n genotyped individuals. For the inbred parental lines P1 and P2, we label an allele from P1 as m and that from P2 as M. The hybrid F1 individuals are completely heterozygous, with genotype Mm at each locus. Crossing F1 with one of the parental lines (say P2) generates a backcross population in which a subject's genotype has an equal probability ½ of being either MM or Mm at every locus. For each individual i, $i = 1, \cdots, n$ where n is the total number of observations, the observed dataconsists of a quantitative trait value y_i and genotypes at K molecular markers $\{M_{ik}\}_{k=1}^{K}$. Details of the QTL experiments can be found in Lynch and Walsh (1998).

Suppose there exists a putative QTL at position μ on the genome. Further assume that the quantitative traits for individuals with QTL genotypes Qq and QQ follow distribution functions F and G, respectively. F and G will differ, for otherwise locus μ would not be considered a QTL. The quantity $\int F\, dG$ is often used to measure the difference between F and G, and is interpretable as the probability that a random value from G exceeds a random value from F. It

is also the area under the receiver-operator characteristic curve (AUC) comparing the two distributions, and is invariant to increasing monotone transformations. It is conceptually helpful to use the rescaled parameter $\delta = 2\int F\,dG - 1$. Note that $|\delta|$ ranges from 0 (when $F = G$) to 1 (where F and G are completely non-overlapping with each other).

For the QTL mapping problem, we note that the QTL position μ is unknown and the only genetic information consists of the marker genotypes, from which the genetic distances of the markers are estimated. If the recombination frequency between a particular marker locus $k \in \{1, \cdots, K\}$ and the QTL is θ_k, then given its kth marker genotype M_{ik}, the conditional phenotype distributions of individual i, will be $y_i \mid (M_{ik} = Mm) \sim \tilde{F}_k(y) = (1-\theta_k)F(y) + \theta_k G(y)$ and $y_i \mid (M_{ik} = MM) \sim \tilde{G}_k(y) = \theta_k F(y) + (1-\theta_k)G(y)$. Here θ_k is a function of μ, and by definition of the conditional distributions we have

$$\tilde{F}_k - \tilde{G}_k = (1 - 2\theta_k)(F - G).$$

This equation drives our ability to detect linkage nonparametrically, as $\tilde{F}_k$ and $\tilde{G}_k$ will exhibit their greatest difference for the marker closest to the QTL, and will show no difference at markers unlinked to the QTL (where $\theta_k = 0.5$). That is, the phenotypic differences between the two marker genotype groups will decrease as the marker and QTL distance increases. Specifically, when marker k is the QTL itself, $\theta_k = 0$ and $\tilde{F}_k = F$, $\tilde{G}_k = G$ (although the QTL need not be at a marker location). At the other extreme of no linkage, $\theta_k = ½$ and $\tilde{F}_k = \tilde{G}_k = \frac{1}{2}(F+G)$.

For testing the existence of a QTL, we have the following two hypotheses:

H_0: There exist no QTLs, that is, $F = G$ for all positions on the chromosome vs.

H_A: There exists a QTL, that is, $F \neq G$ for μ somewhere on the chromosome.

At marker k, we divide the n individuals into $n_{1,k}$ individuals with genotype MM and $n_{2,k} = n - n_{1,k}$ individuals with genotype Mm. Let $y_{(1,1)}, \cdots, y_{(1,n^{1,k})}$ and $y_{(2,1)}, \cdots, y_{(2,n^{2,k})}$ be the corresponding trait values of those $n_{1,k}$ and $n_{2,k}$ individuals. We propose the following approach for estimation and testing. Define the Wilcoxon-Mann-Whitney (WMW) statistic at the kth marker as

$$U_{k,n} = \frac{1}{n_{1,k}n_{2,k}} \sum_{i=1}^{n_{1,k}} \sum_{j=1}^{n_{2,k}} \varphi(y_{(1,i)}; y_{(2,j)}), \quad k = 1, 2, \ldots, K,$$

where $\varphi(x; y) = sign(x - y)$. Then

$$E(U_{k,n}) = (1 - 2\theta_k)\delta. \tag{1}$$

Let $\mathbf{U}_n = (U_{1,n}, \ldots, U_{K,n})^T$ so that $E(\mathbf{U}_n) = [\mathbf{1} - 2\boldsymbol{\theta}]\delta$ where $\mathbf{1} = (1, 1, \ldots, 1)^T$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$. For simplicity, we consider the Euclidean norm:

$$L_1 = (\mathbf{U}_n - E(\mathbf{U}_n))^T (\mathbf{U}_n - E(\mathbf{U}_n)). \tag{2}$$

L_1 is a function of μ and δ, and we define the *Ordinary Least Squares Estimates* (OLSE) of μ and δ as $\hat{\mu}_{ols}$ and $\hat{\delta}_{ols}$, which minimize L_1. Recall that the elements of $\mathbf{U}_n$ are not generally independent. Thus these least squares estimates may not be very efficient, since the dependence structure of $U_{k,n}(k = 1, \ldots, K)$ has not been considered. When the distance between two markers is small, the correlation between the U statistics at the two markers tends toward 1. In linkage studies performed at realistic marker densities, the correlation structure may be nontrivial, and important to take into consideration. As shown below, the conditional covariance matrix Var ($\mathbf{U}_n$) of $\mathbf{U}_n$ given the marker genotypes can be expressed in terms of several unknown but meaningful quantities. Denote $\Delta_1 = \int F^2(y)dG(y)$ and $\Delta_2 = \int G^2(y)dF(y)$. Also define $\zeta_k = ½ + ½ (1 - 2\theta_k)\delta$, $k = 1, 2, \ldots, K$.

Lemma 1

Let Var($\mathbf{U}_n$) *be the covariance matrix of* $\mathbf{U}_n$ *and* θ_{kl} *be the recombination frequency between markers k and l. For the matrix* $\mathbf{v} = \lim_{n\to\infty} nVar(\mathbf{U}_n)$, *the k, lth element is*

$$v_{k,l}=\begin{cases} 8[\int \tilde{F}^2(y)d\tilde{G}(y)+\int(1-\tilde{G}(y))^2 d\tilde{F}(y) - 2\xi_k^2] & k=l \\ 8[(1-\theta_{kl})^2\theta_{kl}\delta_{kl,1}+\theta_{kl}^2(1-\theta_{kl})\delta_{kl,2}+(1-\theta_{kl})^3\delta_{kl,3}] & k\neq l \\ +\theta_{kl}^3\delta_{kl,4} - (6(1-\theta_{kl})\theta_{kl}+2(1-\theta_{kl})^3+2\theta_{kl}^3)\xi_k\xi_l \end{cases} \quad (3)$$

For additional notation and detailed derivations, the reader is referred to the Appendix. Note that the $\mathbf{v}_{k,l}$ are functions of the QTL location μ, δ, Δ_1, and Δ_2, where the last two parameters can be viewed as the second moments of F and G. Let

$$L_2=(\mathbf{U}_n - E(\mathbf{U}_n))^T \sum^{-1} (\mathbf{U}_n - E(\mathbf{U}_n)), \quad (4)$$

$$L_3=(\mathbf{U}_n - E(\mathbf{U}_n))^T \mathrm{Var}^{-1}(\mathbf{U}_n)(\mathbf{U}_n - E(\mathbf{U}_n)), \quad (5)$$

where Σ is the diagonal matrix consisting of the diagonal elements of $Var(\mathbf{U}_n)$. Define the *Weighted Least Squares estimates* (WLSE) of μ and δ as $\hat{\mu}_{wls}$ and $\hat{\delta}_{wls}$, which minimize L_2. Similarly, define the *Generalized Least Squares estimates* (GLSE) of μ and δ as $\hat{\mu}_{gls}$ and $\hat{\delta}_{gls}$, which minimize L_3.

Given Σ or $Var(\mathbf{U}_n)$, L_2 or L_3 can be easily minimized. In the Appendix, a simple procedure is proposed to estimate Δ_1 and Δ_2, which can then be used to estimate Σ and $Var(\mathbf{U}_n)$. After obtaining the estimates of Σ and $Var(\mathbf{U}_n)$, we minimize L_2 or L_3, where Σ and $Var(\mathbf{U}_n)$ are substituted by their estimates. Replacing an unknown covariance matrix by its estimate is generally accompanied by a small increase in the variability of the estimates derived from the resulting estimating equation. This increase becomes negligible for large n. The simulation results in the next section show that for realistic sample sizes, this increase can be safely ignored. Specifically, we adopt the following scheme:

1. Calculate $\hat{\mu}_{ols}$ and $\hat{\delta}_{ols}$ from L_1.
2. Compute the ordinary least squares (OLS) of Δ_1 and Δ_2, $\widehat{\Delta_{1,ols}}$ and $\widehat{\Delta_{2,ols}}$, as described in Appendix.

3. Estimate $\hat{\Sigma}$ and $\widehat{Var}(\mathbf{U}_n)$ by substituting (μ, δ, Δ_1, Δ_2) in Σ and $Var(\mathbf{U}_n)$ with ($\hat{\mu}_{ols}$ $\hat{\delta}_{ols}$, $\widehat{\Delta_{1,ols}}$, $\widehat{\Delta_{2,ols}}$).

4. Minimize L_2 and L_3, where Σ and $Var(\mathbf{U}_n)$ are substituted by $\hat{\Sigma}$ and $\widehat{Var}(\mathbf{U}_n)$.

The method has been implemented in a SAS macro and the source codes are available from the authors upon request.

We now outline the general (asymptotic) properties of the estimator $\hat{\mu}_n$ of μ as well as $\hat{\delta}_n$ of δ. We assume that μ lies in the chromosomal range considered, therefore δ is strictly positive.

Theorem 1

Conditional on the marker genotypes, as n increases, $\sqrt{n}(\widehat{\delta_n} - \delta, \widehat{\mu_n} - \mu)$ asymptotically has a bivariate normal distribution with null mean vector and variance-covariance matrix $\tilde{\mathbf{W}}$, provided μ belongs to the range considered, where

$$\tilde{\boldsymbol{W}} = [\boldsymbol{D}^T \tilde{\boldsymbol{V}}^{-1} \boldsymbol{D}]^{-1} [\boldsymbol{D}^T \tilde{\boldsymbol{V}}^{-1} \boldsymbol{v} \tilde{\boldsymbol{V}}^{-1} \boldsymbol{D}][\boldsymbol{D}^T \tilde{\boldsymbol{V}}^{-1} \boldsymbol{D}]^{-1}.$$

Note that in the definition of $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{V}} = \boldsymbol{I}/n$ for the OLSE and $\hat{\boldsymbol{V}} = \hat{\boldsymbol{\Sigma}}$ or $\widehat{Var}(\mathbf{U}_n)$ for the WLSE and GLSE procedures, respectively with $\boldsymbol{I}$ denoting the $n \times n$ identity matrix. For definitions of $\boldsymbol{D}$, $\boldsymbol{v}$, we refer readers to Appendix.

The quantity $E(U_{k;n})$ is a function of $|\mu_k - \mu|$, where μ_k and μ are the locations of the *kth* marker and the QTL. Thus L_1, L_2 and L_3 are not differentiable if the QTL is located exactly at a marker. A minor modification is necessary. We can fix the problem by replacing

$$|x-\mu| \text{ with } \begin{cases} |x-\mu| & \text{if } |x-\mu| > \varepsilon \\ 1/(2\varepsilon)(x-\mu)^2 + 1/2\varepsilon & \text{if } |x-\mu| < \varepsilon \end{cases}$$ where ε is a prespecified small positive number (see Liang *et al.* 2001A for more discussions). This strategy is commonly used in robust regression (Huber 1964).

Theorem 1 requires that $\delta \neq 0$ and μ is in the chromosomal range considered. To test for H_0 : $\delta = 0$ vs. H_A : $\delta \neq 0$, a simple test statistic can be proposed:

$$L_{3,0} = \mathbf{U}_n^T \mathrm{Var}_0^{-1}(\mathbf{U}_n)\mathbf{U}_n.$$

Under H_0 : $\delta = 0$, $\mathrm{Var}_0(\mathbf{U}_n)$, the variance of $\mathbf{U}_n$, is only a function of the relative distances of the markers. Further, under H_0, $L_{3;0}$ will have an asymptotically chi-squared distribution with K degrees of freedom.

3 Sizmulation and application

Simulations were conducted to study the properties of the proposed method in a backcross population. For simplicity, one chromosome is simulated under different marker densities. The total genetic length of the chromosome is 100 *cM*. The inter-marker distances are either 20, 10 or 5 *cM*, with the first marker at 0 *cM*. The true QTL is located at 45 *cM* (for marker distance 5 *cM*, the marker genotypes at 45*cM* have been removed). Two sets of error distributions of F and G are simulated, which are (1) $F \sim N(-1; 1)$ and $G \sim N(0; 1)$ and (2) $F \sim \exp(2)$ and $G \sim \exp(1)$, respectively. The total sample sizes are either 100 or 200. 1000 simulations were performed for each combination of distribution and sample size. The sample means and

standard deviations from the 1000 simulated data sets were calculated and the results are presented in Tables 1 to 3.

The averages of the estimated standard deviations based on Theorem 1 are also listed in the tables and they are close to the sample standard deviations, indicating that the asymptotically-derived estimates work quite well at sample sizes of 100 or larger. Further, the coverage of the constructed 95% confidence interval of the QTL position based on L_3 are close to the nominal 95%. The improved efficiency of the GLSE procedure over weighted or ordinary least squares is apparent in reduced standard deviations of the estimates, especially for estimating the QTL position μ.

For comparison, the normal likelihood-based interval mapping results are also presented in Tables 1–3. It is very interesting to see that even for truly normally distributed data, the proposed method (based on L_3) is more efficient in estimating the QTL position than the traditional MLE interval mapping procedure. To illustrate that this result does not reflect mere bias in the proposed procedures (e.g., a tendency to favor the middle of a chromosome), we also performed simulations with the first (normal) error distribution as described above and with 6 markers (20 *cM* spacing), but with the QTL at μ =25 *cM*. For n = 100, we obtained estimates with mean ± SD of 24.93 ± 8.78 (interval mapping) and 25.21 *cM* ± 5.99 (GLS procedure). For n = 200, the results were 24.13 *cM* ± 5.69 (interval mapping) and 24.63 *cM* ± 4.27 (GLS procedure). The results of Kong and Wright (1994) indicate that nonstandard asymptotic results apply if the number of markers is large in comparison to the sample size, while standard maximum likelihood efficiency results apply if the markers are held fixed while n increases. However, for the (realistic) sample sizes and marker densities considered, there is no guarantee that the maximum likelihood estimate has the lowest mean-squared error among point estimates. Accordingly, the difference in efficiency between interval mapping and the GLS procedure (as measured by ratios of standard errors) should become less extreme as (i) sample size increases; (ii) the marker density decreases. Both of these observations are borne out in Tables 1–3, noting also that Table 3 has the same marker density surrounding the QTL (5*cM* to each side) as Table 2.

To see how our conclusions vary with the QTL positions and effect sizes, we run two additional sets of simulations and the results are reported in Tables 4 and 5. The simulation set ups of Table 4 are similar to those of Table 1 except that the new QTL position is located at 25 cM instead of 45 cM. For normal traits, in Table 5, we further reduce the heritability of the QTL in Table 4 down to 5%. These simulations assure us that the performance of the proposed method is consistent regardless of the QTL positions and effects.

We next apply the proposed method to the breast cancer study described in Lan *et al.* (2001). In this study, pure inbred female rats from the Wistar-Kyoto (WKY) strain resistant to mammary carcinogenesis were crossed with pure inbred male rats from the Wistar-Furth (WF) strain susceptible to cancer (Lan *et al.*, 2001). The F1 progeny were then mated to WF animals, producing 383 female rats which were either WF/WF or WKY/WF at each locus. These backcross rats were scored for the number of mammary carcinomas and were genotyped at 58 markers on chromosome 5. Using several mapping strategies, Lan *et al.* (2001) found marker D5Rat22 on chromosome 5 was strongly associated with low tumor counts. That is, female rats with the WKY allele at DFRat22 had fewer carcinomas than rats without the WKY allele. We assign the distributions F and G to tumor counts from the rats with WKY/WF and WF/WF genotypes, respectively. In contrast to Lan *et al.* (2001), we test the hypotheses $H_0 : F = G$ vs $H_1 : F \neq G$. The dataset provides a nice example of how nonparametric approaches provide practical and conceptual simplicity to a mapping problem. Lan *et al.* (2001) applied normal parametric interval mapping as one analysis approach to the data, although tumor counts of zero were not uncommon in the dataset.

The data were reanalyzed with our proposed method. Note that datasets with truncated observations, or involving mixtures of discrete and continuous phenotype values, present no difficulty for our proposed approach. Approximately 5% of the genotypes were missing, and because the rate was relatively low, we simply discarded those individuals whose genotypes were missing at the *k*th marker in the calculation of U_k. However, the same individuals were not discarded at markers for which their genotypes were not missing. More sophisticated imputation of missing genotypes would use all observations (Little and Rubin 1987). The tumor counts in the dataset are discrete and ties occur. Conceptually we may use tied ranks (or $\varphi(x - y)=0$), but the application of the analytic variance calculations becomes more complicated. For these data we randomly broke ties with equal probability, which produces valid results but may entail a small loss in power.

For these data, we computed $L_{3;0} = 135$, which is compared to a χ^2 distribution with 58 degrees of freedom. The p-value is 4×10^{-8}, a highly significant result supporting the existence of QTL (s) on chromosome 5. The estimate of the QTL location was 39.6 *cM*, with a 95% confidence interval (37.6,41.6) *cM*. The estimate of δ was 0.539 (corresponding to $\int F\, dG=0.77$), with standard error 0.018 .

In Figure 1, the profile LRT scores based on the normal-likelihood interval mapping procedure of Lander and Botstein (1989) has been plotted. The LRT score peaks near 45 *cM*. This value is not included in the 95% confidence interval of the QTL locus from the proposed method, illustrating that the traditional approach and our approach can exhibit meaningful differences. Samuelson *et al.* (2003) have since provided further evidence for the mammary carcinoma QTL on chromosome 5 and narrowed the region modestly, but the involved gene has not yet been characterized. For easy comparison, we have reversed and rescaled the L_3 curve to the same range of the LRT and plotted it in Figure 1.

4 Discussion

We have proposed a new nonparametric QTL mapping method without assuming the forms of two underlying phenotype distributions. Our approach is more general than the conventional normal location-shift model and relaxes the model assumptions of traditional interval mapping by working with the Wilcoxon-Mann-Whitney (WMW) statistic using genotypes of all markers simultaneously. Once the variance approximation is derived, the minimization of the objective functions can be implemented fairly easily in an iterative procedure. Appropriate thresholds for the test statistic and a confidence interval for the QTL locus are readily available. One situation where our proposed methods lack power is when a QTL influences the *variability* of a trait rather than its *mean*, and it is possible that $\delta = 0$ even under H_A. However, our methods are general enough to cover many applications.

Instead of modeling $U_{k;n}$, Kearsey and Hyne (1994), Wu and Li (1994, 1996) used the trait mean difference $T_{k,n}=\frac{1}{n_{1,k}}\sum_{i=1}^{n_{1,k}} y_{1,i} - \frac{1}{n_{2,k}}\sum_{j=1}^{n_{2,k}} y_{2,j}$ to construct the objective functions L_1, L_2 and L_3, from which the parameter estimates are derived. Only Kearsey and Hyne (1994) compared their procedure with the standard interval mapping and showed that standard interval mapping is slightly more efficient. However, their method does not take the variance-covariance structure of $T_{k;n}$ into consideration. For more fair comparison, we have also done extensive simulations on minimizing L_2 and L_3 that derived from $T_{k;n}$. Our simulation results (not shown) indicate that the efficiency of the QTL position estimate from $T_{k;n}$ is comparable to that from standard interval mapping and therefore is lower than that from $U_{k;n}$ even when F and G are normal. The reason is that, though F and G are normally distributed, $y_{1;i}$ and $y_{2,j}$ are not normally distributed and instead follow the mixture of normals. Therefore, using rank-based $U_{k,n}$ is more efficient than using $T_{k,n}$. For the sample sizes and marker densities

considered, our GLS procedure gives more efficient location estimates than the traditional interval mapping in all our simulations.

The method can be readily applied to other mapping populations, such as Double Haploid, Recombinant Inbred Lines, etc, which are widely used in plant and animal genetic studies. Extending the method to more complicated designs, such as F2 intercrosses, requires careful consideration. Rank-based nonparametric inference methods for stochastically ordered distributions with more than two groups are available and can be directly applied to F2 population (Shanubhogue 1988). However, for F2 population, the covariance structure is more complicated than that of the backcross, and approximations or empirical estimations using for example, the jackknife procedure, may be explored to simplify the problem. We will address this problem in a follow up paper.

We compare our work to that of Liang et al. (2001A), in that we similarly do not advocate our approach as a replacement to existing parametric interval mapping approaches. Our procedure is intended as a *supplement* to the traditional interval mapping, with the main goal of providing robust estimates of QTL locations. As with Liang et al. (2001A), our work implicitly assumes that there is preliminary evidence of linkage to a chromosomal region. This evidence can be assessed via interval mapping or the test statistic proposed in this paper, followed by location and confidence interval estimation as proposed here. The robustness of our method gives the researcher an additional chance to discover linkages that might have been missed due to unrealistic parametric assumptions, and any observed conflict between the two approaches will spur the researcher to carefully investigate the data further.Chen *et al.* (2004) have used GEE approach for mapping quantitative traits. However, their approach uses one marker at a time instead of all available markers simultaneously. Further our approach assumes that there exists only one QTL in the region considered. Liang *el al.* (2001B) have extended their approach to two loci, as have Biernacka *et al.* (2005). Extensions of our approach to handle multiple QTLs are currently under investigation.

References

Barlow, RE.; Bartholomew, DJ.; Bremner, JM.; Brunk, HD. Statistical inference under order restrictions. John Wiley and Sons; New York: 1972.

Belknap JK, Richards SP, O'Toole LA, Helms ML, Phillips TJ. Short-term selective breeding as a tool for QTL mapping: ethanol preference drinking in mice. Behavior Genetics 1997;27:55–66. [PubMed: 9145544]

Biernacka JM, Sun L, Bull SB. Simultaneous localization of two linked disease susceptibility genes. Genetic Epidemiology 2005;28:33–47. [PubMed: 15481103]

Chen WM, Broman KW, Liang KY. Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. Genetic Epidemiology 2004;26:265–272. [PubMed: 15095386]

Darvasi A, Soller M. A simple method to calculate resolving power and confidence interval of QTL map location. Behavior Genetics 1997;27:125–132. [PubMed: 9145551]

Dupuis J, Siegmund D. Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 1999;151:373–386. [PubMed: 9872974]

Flaherty L, et al. The nature and identification of quantitative trait loci: a community's view. Nature Genetics Review 2003;4:911–916.

Haley CS, Knott SA. A simple regression method for mapping quantitative trait in line crosses using flanking markers. Heredity 1992;69:315–324. [PubMed: 16718932]

Haston CK, Zhou X, Gumbiner-Russo L, Irani R, Dejournett R, Gu X, Weil M, Amos CI, Travis EL. Universal and radiation-specific loci influence murine susceptibility to radiation-induced pulmonary fibrosis. Cancer Research 2002;62:3782–3788. [PubMed: 12097289]

Huber PJ. Robust estimation of a location parameter. Annal of Mathematical Statistics 1964;144:1214–1223.

Kearsey MJ, Hyne V. QTL analysis: a simple 'marker regression' approach. Theoretical and Applied Genetics 1994;89:698–702.

Kong A, Wright F. Asymptotic theory for gene mapping. Proceedings of the National Academy of Sciences, USA 1994;91:9705–9709.

Kruglyak L, Lander ES. A nonparametric approach for mapping quantitative trait loci. Genetics 1995;139:1421–1428. [PubMed: 7768449]

Lan H, Kendziorski CM, Haag JD, Shepel LA, Newton MA, Gould MN. Genetic loci controlling breast cancer susceptibility in the Wistar-Kyoto rat. Genetics 2001;157:331–339. [PubMed: 11139513]

Lander ES, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 1989;121:185–199. [PubMed: 2563713]

Liang KY, Chiu YF, Beaty TH. A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases. Human Heredity 2001A;51:64–78. [PubMed: 11096273]

Liang KY, Chiu YF, Beaty TH, Wjst M. Multipoint analysis using affected sib pairs: incorporating linkage evidence from unlinked regions. Genetic Epidemiology 2001B;21:105–122. [PubMed: 11507720]

Little, RJ.; Rubin, DB. Statistical analysis with missing data. John Wiley and Sons; New York: 1987.

Lynch M, Walsh JB. Genetics and analysis of quantitative traits. Sinauer Associations. 1998

Mangin B, Goffinet B, Rebai A. Constructing confidence intervals for QTL location. Genetics 1994;138:1301–1308. [PubMed: 7896108]

Robertson, T.; Wright, FT.; Dykstra, RL. Ordered restricted inference. Wiley; New York: 1988.

Samuelson DJ, Haag JD, Lan H, Monson DM, Shultz MA, Kolman BD, Gould MN. Physical evidence of Mcs5, a QTL controlling mammary carcinoma susceptibility, in congenic rats. Carcinogenesis 2003;24:1455–1460. [PubMed: 12844486]

Sax K. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus Vulgaris*. Genetics 1923;8:552–560. [PubMed: 17246026]

Shanubhogue A. Distribution-free test for homogeneity against stochastic ordering. Metrika 1988;35:109–119.

Visscher PM, Thompson R, Haley CS. Confidence intervals in QTL mapping by bootstrapping. Genetics 1996;143:1013–1020. [PubMed: 8725246]

Wang X, Le Roy I, Nicodeme E, Li R, Wagner R, Petros C, Churchill GA, Harris S, Darvasi A, Kirilovsky J, Roubertoux PL, Paige B. Using advanced intercross lines for high-resolution mapping of HDL Cholesterol quantitative trait loci. Genome Research 2003;13:1654–1664. [PubMed: 12805272]

Wu WR, Li WM. A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. Theoretical and Applied Genetics 1994;89:535–539.

Wu WR, Li WM. Model fitting and model testing in the method of joint mapping of quantitative trait loci. Theoretical and Applied Genetics 1996;92:477–482.

Zou F, Fine JP, Yandell BS. On empirical likelihood for a semiparametric mixture model. Biometrika 2002;89:61–75.

5 Appendix: proofs and derivations

Derivations of first and second moments of U_n

$$
\begin{aligned}
&E(U_{k,n}) = P(Y_{(1,i)} > Y_{(2,j)}) - P(Y_{(1,i)} \le Y_{(2,j)}) \\
&= 2P(Y_{(1,i)} > Y_{(2,j)}) - 1 \\
&= 2\int \tilde{F}_k(y)\, d\tilde{G}_k(y) - 1 \\
&= 2\int \{(1-\theta_k)F(y) + \theta_k G(y)\}\, d\{\theta_k F(y) + (1-\theta_k)G(y)\} - 1 \\
&= (1-2\theta_k)\delta,
\end{aligned}
$$

which is (1).

$$\begin{aligned}
\mathrm{Var}(U_{k,n}) &= \frac{1}{n_{1,k}^2 n_{2,k}^2}\mathrm{Var}\left[\sum_{i=1}^{n_{1,k}n_{2,k}}\varphi(y_{(1,i)}, y_{(2,j)})\right] \\
&= \frac{1}{n_{1,k}^2 n_{2,k}^2}\sum_{i=1}^{n_{1,k}n_{2,k}}\sum_{j=1}^{n_{2,k}}\sum_{l\neq j}(E(\varphi(y_{(1,i)}, y_{(2,j)})\varphi(y_{(1,i)}, y_{(2,l)})) - E^2(U_{k,n})) \\
&+ \frac{1}{n_{1,k}^2 n_{2,k}^2}\sum_{i=1}^{n_{1,k}n_{2,k}}\sum_{j=1}^{n_{1,k}}\sum_{l\neq i}(E(\varphi(y_{(1,i)}, y_{(2,l)})\varphi(y_{(1,j)}, y_{(2,l)})) - E^2(U_{k,n})) \\
&+ \frac{1}{n_{1,k}^2 n_{2,k}^2}\sum_{i=1}^{n_{1,k}}\sum_{j=1}^{n_{2,k}}(E(\varphi^2(y_{(1,i)}, y_{(2,j)})) - E^2(U_{k,n})) \\
&= \frac{4(n_{2,k}-1)}{n_{1,k}n_{2,k}}[\int\tilde{F}_k^2(y)d\tilde{G}_k(y) - \xi_k^2] + \frac{4(n_{1,k}-1)}{n_{1,k}n_{2,k}}[\int(1-\tilde{G}_k(y))^2 d\tilde{F}_k(y) - \xi_k^2] \\
&+ \frac{4}{n_{1,k}n_{2,k}}[\xi_k - \xi_k^2],
\end{aligned}$$

where

$$\begin{aligned}
\int\tilde{F}_k^2(y)d\tilde{G}_k(y) &= \frac{4\theta_k(1-\theta_k)}{3} + (2\theta_k - 1)\{\theta_k\Delta_2 - (1-\theta_k)\Delta_1\}, \\
\int(1-\tilde{G}_k(y))^2 d\tilde{F}_k(y) &= (1-2\theta_k)\delta + \frac{4\theta_k(1-\theta_k)}{3} + (1-2\theta_k)\{(1-\theta_k)\Delta_2 - \theta_k\Delta_1\},
\end{aligned}$$

all are functions of μ, δ, Δ_1, Δ_2. Note $\lim_{n\to\infty}\frac{n_{1,k}}{n} = \lim_{n\to\infty}\frac{n_{2,k}}{n} = \frac{1}{2}$, thus

$$\lim_{n\to\infty} n\mathrm{Var}(U_{k,n}) = 8[\int\tilde{F}^2(y)d\tilde{G}(y) + \int(1-\tilde{G}(y))^2 d\tilde{F}(y) - 2\xi_k^2].$$

This leads to the first equation of (3).

The conditional covariance at two marker loci k and l can be calculated similarly. Based on the genotypes at markers k and l, observations may be grouped. Namely, there are 4 possible marker genotypes $MM{=}MM$, $MM{=}Mm$, $Mm{=}MM$ and $Mm{=}Mm$ with corresponding number of observations as $m_{1,kl}$, $m_{2,kl}$, $m_{3,kl}$ and $m_{4,kl}$, respectively. Let $\tilde{F}_{type}$, $type = MM{=}MM$, $MM{=}Mm$, $Mm{=}MM$ and $Mm{=}Mm$, be the corresponding distributions of phenotypes of the above four possible marker genotype groups. In other words, $\tilde{\boldsymbol{F}}_{\boldsymbol{type}} = P(\mathrm{Qq}|\mathrm{type})F + P(QQ|type)G$, where the conditional probabilities, $P(QQ|type)$ and $P(Qq|type)$, depend on the relative orders of marker M_k, M_l, putative QTL and their relative distances (Chapter 15 of Lynch and Walsh 1998). For ease of notation, we denote, $\tilde{F}_{MM=MM} = \tilde{F}_{1,kl}$, $\tilde{F}_{MM=Mm} = \tilde{F}_{2,kl}$, $\tilde{\boldsymbol{F}}_{\boldsymbol{Mm=MM}} = \tilde{\boldsymbol{F}}_{\boldsymbol{3,kl}}$ **and** $\tilde{\boldsymbol{F}}_{\boldsymbol{Mm=Mm}} = \tilde{\boldsymbol{F}}_{\boldsymbol{4,kl}}$**. The covariance between** $\boldsymbol{U_{k,n}}$ **and** $\boldsymbol{U_{l,n}}$**, Cov** $(\boldsymbol{U_{k,n}}, \boldsymbol{U_{l,n}})$**, equals**

$$\frac{4}{(m_{1,kl}+m_{2,kl})(m_{3,kl}+m_{4,kl})(m_{1,kl}+m_{3,kl})(m_{2,kl}+m_{4,kl})}\{$$
$$m_{1,kl}m_{2,kl}m_{3,kl}[\int\tilde{F}_{2,kl}\tilde{F}_{3,kl}d\tilde{F}_{1,kl}-\xi_k\xi_l]+m_{1,kl}m_{3,kl}m_{4,kl}[\int\tilde{F}_{3,kl}\tilde{F}_{4,kl}d\tilde{F}_{1,kl}-\xi_k\xi_l]$$
$$+m_{1,kl}m_{2,kl}m_{3,kl}[\int(1-\tilde{F}_{1,kl})\tilde{F}_{2,kl}d\tilde{F}_{3,kl}-\xi_k\xi_l]+m_{1,kl}m_{3,kl}m_{4,kl}[\int(1-\tilde{F}_{1,kl})\tilde{F}_{4,kl}d\tilde{F}_{3,kl}-\xi_k\xi_l]$$
$$+m_{1,kl}m_{2,kl}m_{3,kl}[\int(1-\tilde{F}_{1,kl})\tilde{F}_{3,kl}d\tilde{F}_{2,kl}-\xi_k\xi_l]+m_{2,kl}m_{3,kl}m_{4,kl}[\int(1-\tilde{F}_{2,kl})\tilde{F}_{4,kl}d\tilde{F}_{3,kl}-\xi_k\xi_l]$$
$$+m_{1,kl}m_{2,kl}m_{4,kl}[\int\tilde{F}_{2,kl}\tilde{F}_{4,kl}d\tilde{F}_{1,kl}-\xi_k\xi_l]+m_{1,kl}m_{3,kl}m_{4,kl}[\int(1-\tilde{F}_{1,kl})(1-\tilde{F}_{3,kl})d\tilde{F}_{4,kl}-\xi_k\xi_l]$$
$$+m_{1,kl}m_{2,kl}m_{4,kl}[\int(1-\tilde{F}_{1,kl})\tilde{F}_{4,kl}d\tilde{F}_{2,kl}-\xi_k\xi_l]+m_{1,kl}m_{2,kl}m_{4,kl}[\int(1-\tilde{F}_{1,kl})(1-\tilde{F}_{2,kl})d\tilde{F}_{4,kl}-\xi_k\xi_l]$$
$$+m_{2,kl}m_{3,kl}m_{4,kl}[\int(1-\tilde{F}_{3,kl})\tilde{F}_{4,kl}d\tilde{F}_{2,kl}-\xi_k\xi_l]+m_{2,kl}m_{3,kl}m_{4,kl}[\int(1-\tilde{F}_{2,kl})(1-\tilde{F}_{3,kl})d\tilde{F}_{4,kl}-\xi_k\xi_l]$$
$$+m_{1,kl}m_{4,kl}(m_{4,kl}-1)[\int\tilde{F}_{4,kl}^2d\tilde{F}_{1,kl}-\xi_k\xi_l]+m_{2,kl}m_{3,kl}(m_{3,kl}-1)[\int\tilde{F}_{3,kl}(1-\tilde{F}_{3,kl})d\tilde{F}_{2,kl}-\xi_k\xi_l]$$
$$+m_{1,kl}m_{4,kl}(m_{1,kl}-1)[\int(1-\tilde{F}_{1,kl})^2d\tilde{F}_{4,kl}-\xi_k\xi_l]+m_{2,kl}m_{3,kl}(m_{2,kl}-1)[\int(1-\tilde{F}_{2,kl})\tilde{F}_{2,kl}d\tilde{F}_{3,kl}-\xi_k\xi_l]$$
$$+m_{1,kl}m_{4,kl}[\int\tilde{F}_{4,kl}d\tilde{F}_{1,kl}-\xi_k\xi_l]-m_{2,kl}m_{3,kl}\xi_k\xi_l\Big\}$$

Further, $\lim_{n\to\infty}\frac{m_{1,kl}}{n}=\lim_{n\to\infty}\frac{m_{4,kl}}{n}$ and $\lim_{n\to\infty}\frac{m_{2,kl}}{n}=\lim_{n\to\infty}\frac{m_{3,kl}}{n}$, therefore the above covariance can be simplified as

$$n\mathrm{Cov}(U_{k,n},U_{l,n})\approx\frac{64\{n_{a,kl}^2n_{b,kl}\delta_{kl,1}+n_{b,kl}^2n_{a,kl}\delta_{kl,2}+n_{a,kl}^3\delta_{kl,3}+n_{b,kl}^3\delta_{kl,4}-(3nn_{a,kl}n_{b,kl}+2n_{a,kl}^3+2n_{b,kl}^3)\xi_k\xi_l\}}{n^3}$$

where

$$\delta_{kl,1}=\int\tilde{F}_{3,kl}\tilde{F}_{4,kl}d\tilde{F}_{1,kl}+\int(1-\tilde{F}_{1,kl})\tilde{F}_{4,kl}d\tilde{F}_{3,kl}+\int\tilde{F}_{2,kl}\tilde{F}_{4,kl}d\tilde{F}_{1,kl}$$
$$+\int(1-\tilde{F}_{1,kl})(1-\tilde{F}_{3,kl})d\tilde{F}_{4,kl}+\int(1-\tilde{F}_{1,kl})\tilde{F}_{4,kl}d\tilde{F}_{2,kl}+\int(1-\tilde{F}_{1,kl})(1-\tilde{F}_{2,kl})d\tilde{F}_{4,kl}$$
$$\delta_{kl,2}=\int\tilde{F}_{2,kl}\tilde{F}_{3,kl}d\tilde{F}_{1,kl}+\int(1-\tilde{F}_{1,kl})\tilde{F}_{2,kl}d\tilde{F}_{3,kl}+\int(1-\tilde{F}_{1,kl})\tilde{F}_{3,kl}d\tilde{F}_{2,kl}$$
$$+\int(1-\tilde{F}_{2,kl})\tilde{F}_{4,kl}d\tilde{F}_{3,kl}+\int(1-\tilde{F}_{3,kl})\tilde{F}_{4,kl}d\tilde{F}_{2,kl}+\int(1-\tilde{F}_{2,kl})(1-\tilde{F}_{3,kl})d\tilde{F}_{4,kl}$$
$$\delta_{kl,3}=\int\tilde{F}_{4,kl}^2d\tilde{F}_{1,kl}+\int(1-\tilde{F}_{1,kl})^2d\tilde{F}_{4,kl}$$
$$\delta_{kl,4}=\int\tilde{F}_{3,kl}(1-\tilde{F}_{3,kl})d\tilde{F}_{2,kl}+\int(1-\tilde{F}_{2,kl})\tilde{F}_{2,kl}d\tilde{F}_{3,kl}$$

and $2n_{a,kl}$ = number of observations with marker genotypes *MM=MM* or *Mm=Mm* and $2n_{b,kl}$= number of observations with marker genotypes *MM=Mm* or *Mm=MM*. Let the recombination rate between the two markers be θ_{kl}, then $\lim_{n\to\infty}\frac{2n_{a,kl}}{n}=1-\theta_{kl}$ and $\lim_{n\to\infty}\frac{2n_{b,kl}}{n}=\theta_{kl}$, thus the above formula can be further simplified as $n\mathrm{Cov}(U_{k,n},U_{l,n})\approx 8[(1-\theta_{kl})^2\theta_{kl}\delta_{kl,1}+\theta_{kl}^2(1-\theta_{kl})\delta_{kl,2}+(1-\theta_{kl})^3\delta_{kl,3}+\theta_{kl}^3\delta_{kl,4}-(6(1-\theta_{kl})\theta_{kl}+2(1-\theta_{kl})^3+2\theta_{kl}^3)\xi_k\xi_l]$, which leads to the second equation of (3).

Estimation of Δ

Define the Z-statistics $Z_{k,1}$ and $Z_{k,2}$ at the *k*th marker as

$$Z_{k,1}=\frac{2}{n_{1,k}n_{2,k}(n_{2,k}-1)}\sum_{i=1}^{n_{1,k}}\sum_{1\le j<l\le n_{2,k}}\varphi_1(y_{(1,i)},y_{(2,j)},y_{(2,l)})\text{ and}$$

$$Z_{k,2}=\frac{2}{n_{1,k}(n_{1,k}-1)n_{2,k}}\sum_{1\le i<l\le n_{1,k}}\sum_{j=1}^{n_{2,k}}\varphi_2(y_{(1,i)},y_{(2,j)},y_{(1,l)}),\text{ where}$$

$$\varphi_1(x,y,z)=\begin{cases}1 & x>\max\{y,z\}\\ 0 & x\le\max(y,z)\end{cases}\quad\text{and}\quad\varphi_2(x,y,z)=\begin{cases}1 & y>\max\{x,z\}\\ 0 & y\le\max(x,z)\end{cases};$$

$y_{(1,i)}, y_{(2,j)}$, $i = 1, 2 \ldots, n_{1,k}$; $j = 1, 2, \ldots, n_{2,k}$ are defined as in section 2. The mean of $Z_{k,1}$ given the kth marker genotypes is:

$$\begin{aligned}E(Z_{k,1})&=\int\tilde{F}_k^2 d\tilde{G}_k\\ &=\int\{(1-\theta_k)F(y)+\theta_k G(y)\}^2 d\{\theta_k F(y)+(1-\theta_k)G(y)\}\\ &=\tfrac{\theta_k(1-\theta_k)}{3}+\theta_k^3\Delta_2+(1-\theta_k)^3\Delta_1+2\theta_k^2(1-\theta_k)\int F(x)G(x)dF(x)\\ &+2\theta_k(1-\theta_k)^2\int F(x)G(x)dG(x)\end{aligned}$$

The following relations $\int F(x)G(x)dF(x)=\frac{1-\Delta_1}{2}$ and $\int F(x)G(x)dG(x)=\frac{1-\Delta_2}{2}$ lead to

$$E(Z_{k,1})=\frac{4\theta_k(1-\theta_k)}{3}+(2\theta_k-1)\{\theta_k\Delta_2-(1-\theta_k)\Delta_1\},$$

and similarly, the expected value of $Z_{k,2}$ is

$$E(Z_{k,2})=\frac{4\theta_k(1-\theta_k)}{3}+(2\theta_k-1)\{\theta_k\Delta_1-(1-\theta_k)\Delta_2\}.$$

Thus,

$$\begin{aligned}E(Z_{k,1}+Z_{k,2})&=\tfrac{8\theta_k(1-\theta_k)}{3}+(2\theta_k-1)^2(\Delta_1+\Delta_2),\\ E(Z_{k,1}-Z_{k,2})&=(2\theta_k-1)(\Delta_2-\Delta_1).\end{aligned}$$

A least squares estimator of Δ_1 and Δ_2 can be obtained in minimizing the following two objective functions:

$$(\boldsymbol{Z}_1+\boldsymbol{Z}_2-E(\boldsymbol{Z}_1+\boldsymbol{Z}_2))^T(\boldsymbol{Z}_1+\boldsymbol{Z}_2-E(\boldsymbol{Z}_1+\boldsymbol{Z}_2)),\tag{6}$$

$$(\mathbf{Z}_1-\boldsymbol{Z}_2-E(\boldsymbol{Z}_1-\boldsymbol{Z}_2))^T(\boldsymbol{Z}_1-\boldsymbol{Z}_2-E(\boldsymbol{Z}_1-\boldsymbol{Z}_2)).\tag{7}$$

where $\mathbf{Z}_1 = (Z_{1,1}, Z_{2,1}, \ldots, Z_{K,1})^T$ and $\mathbf{Z}_2 = (Z_{1,2}, Z_{2,2}, \ldots, Z_{K,2})^T$. By minimizing (6) and (7), we can get the estimates of $\Delta_1 + \Delta_2$ and $\Delta_2 - \Delta_1$ and therefore, the estimates of Δ_1 and Δ_2.

For more complicated designs, it may not be easy to find appropriate statistics to estimate parameters for the variance and covariance calculations, such as Δ_1 and Δ_2. We thus propose the following alternative empirical method to estimate conditional variance of $\mathbf{U}_n$, $Var(\mathbf{U}_n)$. According to the definition of $U_{k,n}$, given the kth marker genotypes, we have

$Var(U_{k,n})=\frac{\sigma_{10}^2}{n_{(1,k)}}+\frac{\sigma_{01}^2}{n_{(2,k)}}+\frac{\sigma_{11}^2}{n_{(1,k)}n_{(2,k)}}$, where

$\sigma_{10}^2=Var[\,E(\varphi(y_{(1,1)},y_{(2,1)})|y_{(1,1)})]$, $\sigma_{01}^2=Var[\,E(\varphi(y_{(1,1)},y_{(2,1)})|y_{(2,1)})]$ and

$\sigma_{11}^2=Var[\,E(\varphi(y_{(1,1)},y_{(2,1)}))]$. Consider the set

$\{\varphi(y_{(1,i)},y_{(2,j)})\varphi(y_{(1,i)},y_{(2,l)}),\ j\neq l=1,2,\ldots,n_{2,k}\}$, we get

$E(\varphi(y_{(1,i)},y_{(2,j)})\varphi(y_{(1,i)},y_{(2,l)}))=\sigma_{10}^2+E^2(U_{k,n})$. Thus,

$\frac{1}{n_{1,k}n_{2,k}(n_{2,k}-1)}\sum_{i=1}^{n_{1,k}}\sum_{j=1}^{n_{2,k}}\sum_{l\neq j}^{n_{2,k}}\varphi(y_{(1,i)},y_{(2,j)})\varphi(y_{(1,i)},y_{(2,l)})-E^2(\widehat{U}_{k,n})$ estimates σ_{10}^2. Similarly, σ_{01}^2 and σ_{11}^2 can be estimated by

$\frac{1}{n_{1,k}(n_{1,k}-1)n_{2,k}}\sum_{i=1}^{n_{1,k}}\sum_{l=1}^{n_{2,k}}\sum_{j=1}^{n_{2,k}}\varphi(y_{(1,i)},y_{(2,j)})\varphi(y_{(1,l)},y_{(2,j)})-E^2(\widehat{U}_{k,n})$ and

$\frac{1}{n_{1,k}n_{2,k}}\sum_{i=1}^{n_{1,k}}\sum_{j=1}^{n_{2,k}}\varphi^2(y_{(1,i)},y_{(2,j)})-E^2(\widehat{U}_{k,n})$, respectively, **where** $E^2(\widehat{U}_{k,n})=U_{k,n}^2$.

When n is large,

$$Var(U_{k,n})=\frac{\sigma_{10}^2}{n_{(1,k)}}+\frac{\sigma_{01}^2}{n_{(2,k)}}+O(n^{-2}).$$

The estimate for conditional covariance can be derived similarly and omitted for brevity of presentation.

Proof of Theorem 1

Let $\boldsymbol{\beta}=(\delta,\mu)^T$ and $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. L_i, i = 1, 2, 3 can be generally expressed as special cases of the following objective functions

$$S(\boldsymbol{\beta})=\{\mathbf{U}_n-g(\boldsymbol{\beta})\}^T\boldsymbol{V}^{-1}\{\mathbf{U}_n-g(\boldsymbol{\beta})\},$$

where $g(\boldsymbol{\beta})=E(\mathbf{U}_n)$ and $\boldsymbol{V}=\boldsymbol{I}/n$, $\boldsymbol{\Sigma}$, $Var(\mathbf{U}_n)$ for L_1, L_2, L_3, respectively. To emphasize the dependence of the parameter estimate of $\boldsymbol{\beta}$ on the sample size n, let $\{\hat{\boldsymbol{\beta}}\}$be a sequence of solutions which minimizes the objective function $S(\boldsymbol{\beta})$.

Further, denote $g'(\boldsymbol{\beta})=[\delta g(\boldsymbol{\beta})/\delta\boldsymbol{\beta}^T]_{k\times 2}$, $D=g'(\boldsymbol{\beta}_0)$, and for k = 1, 2,...,K, $g_k''(\boldsymbol{\beta})=\left[\partial^2 g_k(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\right]_{2\times 2}$, where $g_k(\boldsymbol{\beta})$ is the kth element of the vector $g(\boldsymbol{\beta})$.

For general $S(\boldsymbol{\beta})$, we can show that under the regularity conditions

(A)There exists a positive definite matrix $\boldsymbol{\Lambda}(\boldsymbol{\beta})$ in a neighborhood of $\boldsymbol{\beta}_0$ such that

$$\lim_{n\to\infty}n^{-1}g'(\boldsymbol{\beta})^T\boldsymbol{V}^{-1}g'(\boldsymbol{\beta})=\boldsymbol{\Lambda}(\boldsymbol{\beta}).$$

Let $\boldsymbol{\Lambda}_0=\boldsymbol{\Lambda}(\boldsymbol{\beta}_0)=\lim_{n\to\infty}n^{-1}\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D}$ and $\boldsymbol{\Lambda}(\boldsymbol{\beta})\to\boldsymbol{\Lambda}_0$ as $\boldsymbol{\beta}\to\boldsymbol{\beta}_0$

(B) There exists a positive definite matrix such that

$$\lim_{m\to\infty} n^{-1}\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{V}_0\boldsymbol{V}^{-1}\boldsymbol{D}=\boldsymbol{\Lambda}^*,$$

where $\mathbf{V}_0 = Var(\mathbf{U}_n)$ is the true variance-covariance of $\mathbf{U}_n$ given $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

(C) $\lim_{n\to\infty} n^{-1}, \boldsymbol{V}^{-1}$, $\lim_{n\to\infty} n^{-1} g'(\boldsymbol{\beta})^T \mathbf{V}^{-1}\mathbf{V}_0 \mathbf{V}^{-1} g'(\boldsymbol{\beta})$, $g'(\boldsymbol{\beta})^T g'(\boldsymbol{\beta})$, and $g_k''(\boldsymbol{\beta})^T g_k''(\boldsymbol{\beta})$, $k = 1,\ldots,K$, are all bounded in a neighborhood of $\boldsymbol{\beta}_0$

then, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ will be asymptotically normal.

Due to the special set up of our proposed model, we can show that $g(\boldsymbol{\beta})$, $g'(\boldsymbol{\beta})$ and $g_k''(\boldsymbol{\beta})$s are actually bounded for all $\boldsymbol{\beta}$. Further, under the alternative hypothesis where there exists a QTL, it can be shown that $(n\boldsymbol{V})^{-1}$ is bounded and conditions A), B) and C) hold. Thus the asymptotical normality of our estimate $\hat{\boldsymbol{\beta}}_n$ follows.

Proof

By the regularity assumptions, it follows by standard methods (Sen and Singer 1993, pp 206 – 207) that for every (3xed) $C(< \infty)$,

$$\sup_{\|t\|\le C}\left\{n^{-1/2}\,\|\,S'(\boldsymbol{\beta}_0+\tfrac{1}{\sqrt{n}}t) - S'(\boldsymbol{\beta}_0) - n^{1/2}[\,\tfrac{1}{n}S''(\boldsymbol{\beta}_0)]t\,\|\right\} \xrightarrow{p} 0 \text{ as } n\to\infty.$$

This implies that $\hat{\boldsymbol{\beta}}$ lies in the $O(n^{-1/2})$-ball around $\boldsymbol{\beta}_0$ with probability going to 1 (as $n \to \infty$), and hence,

$$0=n^{-1/2}S'(\widehat{\boldsymbol{\beta}}_n)=n^{-1/2}S'(\boldsymbol{\beta}_0)+\left[S''(\boldsymbol{\beta}_0)/n\right]\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)+o_p(1)$$

Thus $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ is asymptotically equivalent to $[S''(\boldsymbol{\beta}_0)/n]^{-1}\, n^{-1/2}S'(\boldsymbol{\beta}_0)$. Further it can be shown that $\frac{1}{2}n^{-1/2}S'(\boldsymbol{\beta}_0)=n^{-1/2}\boldsymbol{D}^T\boldsymbol{V}^{-1}\{\mathbf{U}_n - g(\boldsymbol{\beta}_0)\} \xrightarrow{\mathcal{D}} N(0,\boldsymbol{\Lambda}^*)$ and $\lim_{n\to\infty} S''(\boldsymbol{\beta}_0)/n = 2\boldsymbol{\Lambda}_0$ *a.e.* Using the Slutsky theorem, it follows that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\Lambda}^*\boldsymbol{\Lambda}_0^{-1}).$$

By replacing unknown $\boldsymbol{\Lambda}_0$ or $\boldsymbol{\Lambda}^*$ with their consistent estimates yields Theorem 1 using Slutsky theorem again.

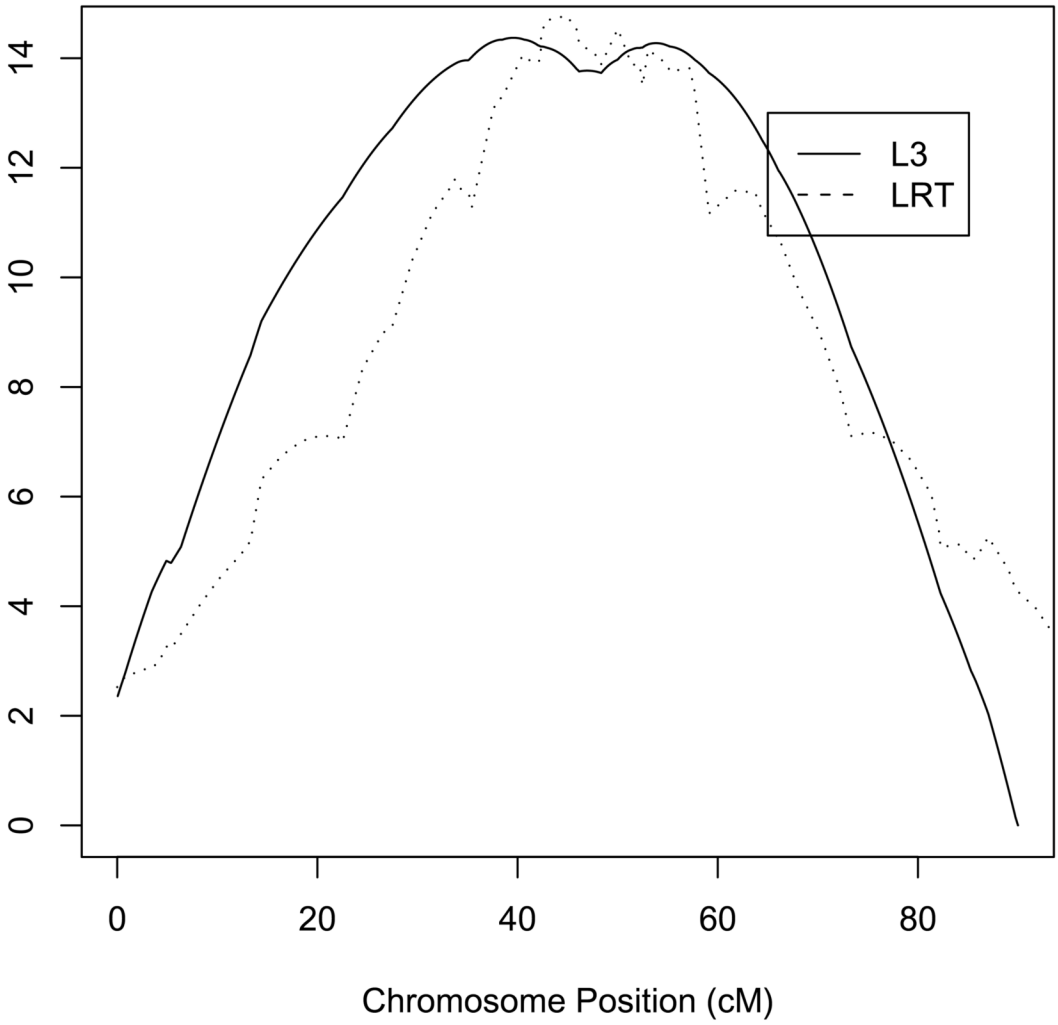

Figure 1.
Comparison of the profile likelihood ratio test *LRT* and the objective function L_3.

Table 1

Results for simulation with 6 markers (spacing 20cM)

Model Set Up[a]		OLSE			WLSE			GLSE		
para	*n*	Mean[b]	SD[c]	ESD(cp)[d]	Mean[b]	SD[c]	ESD(cp)[d]	Mean[b]	SD[c]	ESD(cp)[d]
Normal										
δ=.52, μ=45					Estimate of δ					
	100	.53	.12	.12(.94)	.528	.12	.12(.94)	.53	.11	.11(.94)
	200	.524	.082	.086(.95)	.522	.08	.084(.95)	.524	.076	.08(.96)
					Estimate of μ					
	100	44.7	9.0	7.7(.91)	44.7	8.7	7.5(.91)	44.7 [44.38]	5.9 [9.45]	5.9(.95)
	200	44.8	5.7	5.3(.93)	44.8	5.4	5.1(.93)	44.9 [44.10]	3.8 [5.60]	4.1(.96)
Exponential										
δ=.333, μ=45					Estimate of δ					
	100	.34	.138	.132(.93)	.34	.136	.132(.94)	.34	.124	.124(.94)
	200	.334	.094	.094(.95)	.334	.094	.094(.94)	.336	.088	.088 (.95)
					Estimate of μ					
	100	44.9	15.3	15.4(.89)	44.9	15.3	15.5(.89)	45.5 [45.18]	10.6 [17.20]	12.3(.93)
	200	44.7	9.9	9.1(.91)	44.6	9.8	9.0(.91)	45.1 [44.40]	6.4 [10.39]	7.0(.94)

[a] Parameters of simulated data and sample size.

[b] Average of the parameter estimate over 1000 simulated data sets. Numbers in [ ] are based on traditional interval mapping.

[c] Empirical standard deviation of parameter estimate over 1000 simulated data sets. Numbers in [ ] are based on traditional interval mapping.

[d] Average of estimated standard deviation of parameter estimate over 1000 simulated data sets.

cp: 95% coverage probability over 1000 simulated data sets.

Table 2

Results for simulation with 11 markers (spacing 10cM)

Model Set Up		OLSE			WLSE			GLSE		
para	*n*	Mean	SD	ESD(cp)	Mean	SD	ESD(cp)	Mean	SD	ESD(cp)
Normal										
δ=.52, μ=45					Estimate of δ					
	100	.524	.114	.118(.95)	.52	.12	.114(.95)	.525	.118	.10(.93)
	200	.518	.082	.08(.94)	.512	.080	.080(.94)	.52	.075	.073(.95)
					Estimate of μ					
	100	44.6	8.7	6.8(.90)	44.6	8.8	6.5(.89)	44.9 [44.86]	3.8 [8.72]	3.9(.94)
	200	44.9	5.4	4.7(.91)	44.9	5.2	4.5(.90)	44.9 [44.95]	2.8 [4.12]	2.70(.96)
Exponential										
δ=.333, μ=45					Estimate of δ					
	100	.343	.126	.127(.96)	.342	.125	.127(.95)	.341	.116	.116(.95)
	200	.337	.089	.089(.96)	.336	.088	.09(.96)	.336	.08	.082(.96)
					Estimate of μ					
	100	45.3	14.8	12.8(.85)	44.1	14.8	12.5(.84)	44.9 [45.14]	6.8 [15.75]	7.0(.93)
	200	45.0	9.9	7.7(.88)	45.0	9.6	7.6(.88)	45.0 [45.06]	4.6 [9.32]	4.4(.94)

All columns are the same as in Table 1.

Table 3

Results for simulation with 20 markers (spacing 5cM, no marker at 45cM)

Model Set Up		OLSE			WLSE			GLSE		
para	***n***	**Mean**	**SD**	**ESD(cp)**	**Mean**	**SD**	**ESD(cp)**	**Mean**	**SD**	**ESD(cp)**
Normal										
δ=.52, μ=45					Estimate of δ					
	100	.53	.116	.118(.95)	.53	.112	.112(.95)	.53	.11	.102(.93)
	200	.52	.081	.082(.96)	.52	.08	.081(.95)	.52	.076	.072(.93)
					Estimate of μ					
	100	45.1	8.5	7.0(.88)	45.0	8.3	6.7(.88)	45.1 [44.87]	4.1 [8.79]	3.6(.90)
	200	45.3	5.7	4.9(.90)	45.2	5.6	4.7(.90)	45.1 [45.23]	2.7 [5.16]	2.6(.93)
Exponential										
δ=.333, μ=45					Estimate of δ					
	100	.34	.12	.12(.96)	.34	.12	.12(.95)	.34	.12	.12(.94)
	200	.34	.09	.09(.96)	.34	.09	.09(.96)	.33	.081	.80(.94)
					Estimate of μ					
	100	44.9	15.8	15.2(.86)	44.9	16.0	16.4(.84)	44.7 [45.52]	6.3 [15.95]	6.0(.88)
	200	45.2	10.8	8.3(.88)	45.3	11.1	8.2(.88)	45.0 [45.07]	4.6 [9.32]	4.1(.88)

All columns are the same as in Table 1.

Table 4

Results for simulation with 6 markers (spacing 20cM)

Model Set Up[a]		OLSE			WLSE			GLSE		
para	*n*	Mean[b]	SD[c]	ESD(cp)[d]	Mean[b]	SD[c]	ESD(cp)[d]	Mean[b]	SD[c]	ESD(cp)[d]
Normal										
δ=.52, μ=25				Estimate of δ						
	100	.528	.117	.122(.94)	.526	.115	.120(.94)	.533	.110	.113(.93)
	200	.520	.081	.086(.96)	.520	.080	.085(.96)	.523	.075	.080(.95)
				Estimate of μ						
	100	25.13	8.07	7.62(.93)	25.04	7.82	7.38(.93)	25.21 [25.05]	5.99 [8.80]	6.00(.93)
	200	24.5	5.57	5.3(.91)	24.48	5.40	5.11(.92)	24.63 [24.21]	4.21 [5.80]	4.22 (.94)
Exponential										
δ=.333, μ=25				Estimate of δ						
	100	.340	.128	.133(.95)	.339	.127	.132(.96)	.347	.117	.124(.94)
	200	.330	.090	.093(.97)	.330	.089	.093(.96)	.334	.089	.088 (.96)
		Estimate of μ								
	100	25.87	14.18	15.61(.90)	25.99	14.62	14.36(.90)	26.1 [28.30]	9.7 [17.51]	10.53(.93)
	200	24.80	9.98	8.90(.91)	24.81	10.09	8.76(.91)	24.99 [24.84]	7.08 [10.47]	7.00(.93)

Simulation set ups are similar to Table 1 except that the simulated QTL position is 25 *cM*; All columns are the same as in Table 1.

Table 5

Results for simulation with 6 markers (spacing 20cM)

Model Set Up[a]		OLSE			WLSE			GLSE		
para	***n***	**Mean[b]**	**SD[c]**	**ESD(cp)[d]**	**Mean[b]**	**SD[c]**	**ESD(cp)[d]**	**Mean[b]**	**SD[c]**	**ESD(cp)[d]**
Normal										
δ=.28, μ=25		Estimate of δ								
	100	.291	.130	.135(.88)	.290	.129	.134(.88)	.298	.123	.126(.92)
	200	.280	.088	.095(.90)	.280	.088	.094(.93)	.284	.082	.080(.97)
		Estimate of μ								
	100	26.65	16.50	38.6(.97)	26.62	16.45	28.71(.96)	25.90 [31.47]	11.17 [22.83]	14.23(.97)
	200	25.1	12.31	10.82(.90)	24.92	11.90	10.75(.90)	24.77 [26.87]	8.00 [16.19]	8.35 (.93)

Simulation set ups are similar to those of the normally distributed traits in Table 4 but with significantly reduced QTL effects;

All columns are the same as in Table 1.