Score Statistics for Mapping Quantitative Trait Loci

Myron N. Chang, Rongling Wu, Samuel S. Wu, and George Casella*
Department of Statistics, University of Florida
Gainesville, Florida 32611-8545

December 15, 2003

Abstract

We propose a method to detect the existence of a quantitative trait loci (QTL) in a backcross population using a score test. Under the null hypothesis of no QTL, all phenotype random variables are independent and identically distributed, and the maximum likelihood estimates (MLEs) of parameters in the model are usually easy to obtain. Since the score test only uses the MLEs of parameters under the null hypothesis, it is computationally simpler than the likelihood ratio test (LRT). Moreover, because the location parameter of the QTL is unidentifiable under the null hypothesis, the distribution of the maximum of the LRT statistics, typically the statistic of choice for testing $H_0 :$ no QTL, does not have the standard chi-square distribution asymptotically under the null hypothesis. From the simple structure of the score test statistics, the asymptotic null distribution can be derived for the maximum of the square of score test statistics. Numerical methods are proposed to compute the asymptotic null distribution and the critical thresholds can be obtained accordingly. A simple backcross design is used to demonstrate the application of the score test to QTL mapping. The proposed method can be readily extended to more complex situations.

*Professor and Chair, Department of Statistics, University of Florida, Gainesville, FL 32611. Supported by National Science Foundation Grant DMS-9971586. Email: casella@stat.ufl.edu. This work was completed while Casella was on sabbatical at the University of Granada.
1 Introduction

Since the publication of the seminal mapping paper by Lander and Botstein (1989), there has been a large amount of literature concerning the development of statistical methods for mapping complex traits (reviewed in Jansen 2000; Hoechle 2001). Although the idea of associating a continuously varying phenotype with a discrete trait (marker) dates back to the work of Sax (1923), it was Lander and Botstein (1989) who first established an explicit principle for linkage analysis. They also provided a tractable statistical algorithm for dissecting a quantitative trait into their individual genetic locus components, referred to as quantitative trait loci (QTL).

The success of Lander and Botstein in developing a powerful method for linkage analysis of a complex trait has roots in two different developments. First, the rapid development of molecular technologies in the middle 1980s led to the generation of a virtually unlimited number of markers that specify the genome structure and organization of any organism (Drayna et al. 1984). Second, almost simultaneously, improved statistical and computational techniques, such as the EM algorithm (Dempster et al. 1977), made it possible to tackle complex genetic and genomic problems.

Lander and Botstein’s (1989) model for interval mapping of QTL is regarded as appropriate for an ideal (simplified) situation, in which the segregation patterns of all markers can be predicted on the basis of the Mendelian laws of inheritance and a trait under study is strictly controlled by one QTL on a chromosome. This work was extended and improved by many (Jansen and Stam 1994; Zeng 1994; Haley and Knott 1994; Xu 1996), with successful identification of so-called “outcrossing” QTL in real-life data sets from pigs (Andersson et al. 1994) and pine (Knott et al. 1997). A general framework for QTL analysis was recently established by Wu et al. (2002) and Lin et al. (2003).

Despite many substantial extensions of statistical mapping methods, the characterization of critical thresholds used to declare the statistical significance of a QTL has been considered one of the thorniest issues in the genetic analysis of complex traits (Lander and Schork 1994) and has not well been addressed yet in the current literature. Three approaches can be used to calculate the threshold value throughout a genome, (1) analytical methods, (2) simulation studies, and (3) permutation tests. The analytical approach critically depends on the distribution of an underlying test statistic. Typically the profile of likelihood-ratio (LR) test statistics is constructed over the grid of possible QTL locations in a linkage group or an entire genome and the maximum of the LR (MLR) is used as a global test statistic. At a given position of the QTL, the LR test statistic is asymptotically \( \chi^2 \)-distributed under the null hypothesis with degrees of freedom equal to the number of associated QTL effects. However, under the null hypothesis \( H_0 \): no QTL, the QTL position is unidentified and, therefore, the MLR test statistic does not follow the standard \( \chi^2 \)-distribution asymptotically. Based on the results of Davies (1977, 1987), several authors have derived approximate formulas to determine critical thresholds for a particular design, where closed form thresholds are not available (Rebai et al 1994; Doerge
To overcome the limitations due to the failure of the test statistic to follow a standard statistical distribution, a distribution-free simulation approach has been proposed to calculate critical values for different experimental settings with intermediate marker densities (Lander and Botstein 1989; van Ooijen 1992; Darvasi et al 1993). A more empirical approach for determining critical thresholds is based on a permutation test procedure (see Churchill and Doerge 1994 and Doerge and Churchill 1996). The simulation- or permutation-based approach requires a very high computational workload, and this makes its application impractical in many situations. For example, to obtain a reasonably accurate estimate of a critical threshold at a genome-wide type I error rate of 0.01, one need perform at least 10,000 permutations for the same data set (Doerge and Rebai 1996).

In this article, we propose a general statistical framework to detect the existence of a QTL in a backcross population using a score test. Schaid et al. (2002) and Wang and Huang (2002) proposed a score-statistic approach for mapping QTL based on sibships or a natural population in humans. But these authors did not provide a general framework for implementing score statistics in QTL mapping. Under the $H_0$: no QTL, all phenotype random variables in the sample are independent and identically distributed and the maximum likelihood estimates (MLEs) of parameters in the model are usually easy to obtain. Since the score test only uses MLEs of parameters under the null hypothesis, the score test is much less demanding computationally than the LR test. We form score test statistics for given locations in the permissible range and then to use the maximum of the square of these score test statistics as a global test statistic. 

We show that under $H_0$: no QTL, the maximum of the square of score test statistics asymptotically follows the distribution of $\sup Z^2(d)$, where $Z(d)$ is a Gaussian stochastic process with mean zero and well-formulated covariances, and the supremum is over the permissible range of $d$, the Haldane distance between one end of the genome and the location of the QTL. The distribution of $\sup Z^2(d)$ only depends on Haldane distances or recombination fractions between markers, and the critical thresholds can be obtained by simulation. Since the LR test and the score test are asymptotically equivalent (Cox and Hinkley 1974), the critical thresholds obtained for the maximum of the square of score test statistics can be used for MLR test also.

In Section 3, we derive the score test statistic and the asymptotic null distribution of the maximum of the square of score test statistics for interval mapping. The score test and the null distribution are extended to composite interval mapping in Section 4, where we show that, in the dense map case, the distribution of $\sup Z^2(d)$ is the same as that derived by Lander and Botstein (1989). Numerical methods to compute the null distribution are addressed in Section 5. In Section 6, examples are presented and the thresholds obtained by our method are compared with that from permutation procedures in Churchill and Doerge (1994) and that from the approximate formula in Lander and Botstein (1989) and in Piepho (2001). Section 7 is a discussion section. Some technical derivations are deferred to Appendices A.1 and A.2.
Since our focus in this article is the asymptotic distribution under the null hypothesis, the phrase “under the null hypothesis” will be omitted hereafter we refer to asymptotic distribution, asymptotic convergence, or asymptotic equivalence.

2 QTL Mapping in Backcross Populations

This section provides a short introduction to QTL mapping in backcross populations, and then introduces the model that we will use in the remainder of the paper. A more thorough introduction to QTL mapping can be found in Doerge et al. (1997).

2.1 Basics

A chromosome can be thought of as a string of material, some of which are genes (and can be linked to certain traits) and some of which is inter-genic material, and carries no information. The aim of QTL mapping is to associate genes with quantitative phenotypic traits. For example, we might be interested in QTL that affect the growth rate of pine trees, so we might be looking for regions on a chromosome that are associated with growth rate in pines.

In the simplest case we assume that there are two parents that have alleles $QQ$ and $qq$ at a certain (unknown) place, or locus, on the chromosome. If the parents mate, the offspring (called the $F_1$ generation) will have the allele $Qq$. In a backcross, the offspring is mated back to a parent, say $QQ$, producing a new generation with alleles $QQ, Qq$. We want to investigate the association between the new generation (sometimes called $BC_1$) and the quantitative trait.

Now the problems arise. We typically cannot see if the locus is $Q$ or $q$, but rather we see a “marker”, with alleles $M$ and $m$. If for example, the parent has alleles $MQ$ and from the $F_1$ we get $mq$, then we will see $Mm$ and the genotype is $Qq$. But if a recombination occurs (so the $M$ switches chromosomes), we
then get $Mq$ from the $F_1$, and we will see $MM$, but the genotype is $Qq$. If $r$ denotes the recombination probability, then the distribution of the phenotype in the backcross population (Doerge et al. 1997) is

$$Y | MM \sim \begin{cases} N(\mu_{QQ}, \sigma^2) & \text{with probability } 1 - r \\ N(\mu_{Qq}, \sigma^2) & \text{with probability } r \end{cases},$$

and

$$Y | Mm \sim \begin{cases} N(\mu_{QQ}, \sigma^2) & \text{with probability } r \\ N(\mu_{Qq}, \sigma^2) & \text{with probability } 1 - r \end{cases}.$$ 

When categorized by the markers, the difference in means of the populations is

$$\mu_{MM} - \mu_{Mm} = (1 - 2r)(\mu_{QQ} - \mu_{Qq}).$$

Assuming that $\mu_{QQ} - \mu_{Qq} \neq 0$, a test of $H_0$: no linkage between $M$ and $Q$ can be carried out by testing $H_0: \mu_{MM} - \mu_{Mm}$ (which is also equivalent to the null hypothesis $H_0: r = 1/2$). This test can be carried out with something as simple as a $t$-test, but now likelihood methods are more popular. Using this mixture model in a likelihood analysis, we can not only test for linkage but also estimate $r$. (The backcross is the simplest design in which we get enough information to estimate $r$.)

### 2.2 Interval Mapping

In practice, we typically use a somewhat more complex model, where we assume that the QTL is bracketed by two flanking markers $M^{l-1}$ (with alleles $M^{l-1}$ and $m^{l-1}$) and $M^l$ (with alleles $M^l$ and $m^l$) from a genetic linkage map. These two markers form four distinct genotype groups

$$\{M^{l-1}m^{l-1}M^lm^l\}, \{M^{l-1}m^{l-1}m^lm^l\}, \{m^{l-1}m^{l-1}M^lm^l\}, \{m^{l-1}m^{l-1}m^lm^l\}. \tag{1}$$

The recombination fraction between the two markers $M^{l-1}$ and $M^l$ is denoted by $r$, and that between $M^l$ and the QTL by $r_1$, and that between the QTL and $M^l$ by $r_2$. The corresponding genetic distances between these three loci are denoted by $D$, $x$ and $D - x$, respectively. Genetic distances and recombination fractions are related through a map function (Ott 1991). Here we use the Haldane map function, given by $r = (1 - e^{-2D})/2$, to convert the genetic distance to the corresponding recombination fraction.

To express the conditional probability of a QTL genotype, say $Qq$, given each of the four two-marker genotypes, we introduce two ratio parameters,

$$\theta_1 = \frac{r_1r_2}{1 - r} = \frac{r_1r_2}{(1 - r_1)(1 - r_2) + r_1r_2} = \frac{1 + e^{-2D} - e^{-2x} - e^{-2(D-x)}}{2(1 + e^{-2D})}, \tag{2}$$

and

$$\theta_2 = \frac{r_1(1 - r_2)}{r} = \frac{r_1(1 - r_2)}{r_1(1 - r_2) + (1 - r_1)r_2} = \frac{1 - e^{-2D} - e^{-2x} + e^{-2(D-x)}}{2(1 - e^{-2D})}.$$
Furthermore, assume that the putative QTL has an additive effect on the trait, and define the expected genotypic values as $\mu + \Delta$ and $\mu - \Delta$ for the two QTL genotypes Qq and qq, respectively, where $\mu$ is the overall mean of the backcross population. The corresponding distributions of the phenotypic values for the four marker genotypes can be modeled, respectively, as

$$(1 - \theta_1)N(\mu - \Delta, \sigma^2) + \theta_1 N(\mu + \Delta, \sigma^2),$$

$$(1 - \theta_2)N(\mu - \Delta, \sigma^2) + \theta_2 N(\mu + \Delta, \sigma^2),$$

and

$$\theta_2 N(\mu - \Delta, \sigma^2) + (1 - \theta_2)N(\mu + \Delta, \sigma^2),$$

where $\mu$, $\Delta$, $\sigma^2$, and $x$, $0 \leq x \leq D$, are the four unknown parameters contained in the above mixture models.

Lastly, suppose that we take a sample of size $N$ from the backcross populations, with respective sample sizes of $n_1$, $n_2$, $n_3$ and $n_4$ in the four marker groups of (1). We then have

$$\lim(n_1/N) = \lim(n_4/N) = \frac{1}{2}(1 - r)$$

and

$$\lim(n_2/N) = \lim(n_3/N) = \frac{1}{2}r,$$

as $N \to \infty$.

### 3 Single Interval Mapping

Since the existence of the QTL in the interval is indicated by $\Delta \neq 0$, its statistical significance can be tested by the hypotheses:

$$H_0 : \quad \Delta = 0 \quad \text{vs.} \quad H_a : \quad \Delta \neq 0.$$  \hspace{1cm} (4)

Note that under the null hypothesis, all observations are independent and identically distributed from $N(\mu, \sigma^2)$. Thus, the QTL position parameter $\theta = (\theta_1, \theta_2) = \theta(x, D)$ is unidentifiable because there is no information on $\theta$ if $H_0$ is true. In this case, the MLR test statistic does not follow a standard $\chi^2$-distribution with two degrees of freedom. To see this, consider a case where the two markers $M_{\ell-1}$ and $M_{\ell}$ are close, and the sample sizes $n_2$ and $n_3$ can be small. The MLR test will be approximately equivalent to a two-sample comparison test and asymptotically distributed as the $\chi^2$-distribution with one degree of freedom rather than two degrees of freedom.

We derive a score statistic to overcome this problem. Recall that $n_1$, $n_2$, $n_3$ and $n_4$ are the sample sizes of the four marker groups. If we define $N_1 = n_1$,  

6
\[ N_2 = N_1 + n_2, \ N_3 = N_2 + n_3 \] and \[ N_4 = N = N_3 + n_4, \] the log likelihood of the phenotype data conditional on the marker information can be written

\[
\ell(\Delta, \mu, \sigma^2, x) = -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \sigma^2 \\
+ \sum_{i=1}^{N_1} \log[(1 - \theta_1)e^{-\frac{x}{2\sigma^2}(y_i - \mu + \Delta)^2} + \theta_1 e^{-\frac{x}{2\sigma^2}(y_i - \mu - \Delta)^2}] \\
+ \sum_{i=N_1+1}^{N_2} \log[(1 - \theta_2)e^{-\frac{x}{2\sigma^2}(y_i - \mu + \Delta)^2} + \theta_2 e^{-\frac{x}{2\sigma^2}(y_i - \mu - \Delta)^2}] \\
+ \sum_{i=N_2+1}^{N_3} \log[\theta_2 e^{-\frac{x}{2\sigma^2}(y_i - \mu + \Delta)^2} + (1 - \theta_2)e^{\frac{x}{2\sigma^2}(y_i - \mu - \Delta)^2}] \\
+ \sum_{i=N_3+1}^{N} \log[\theta_1 e^{-\frac{x}{2\sigma^2}(y_i - \mu + \Delta)^2} + (1 - \theta_1)e^{-\frac{x}{2\sigma^2}(y_i - \mu - \Delta)^2}]. \tag{5}
\]

The score function is the derivative of (5) with respect to \( \Delta \) evaluated at \( \Delta = 0 \):

\[
u(\mu, \sigma^2, x) = \frac{2\ell}{2\Delta} \bigg|_{\Delta=0} = \frac{1}{\sigma^2} \left\{ -(1 - 2\theta_1) \sum_{i=1}^{N_1} (y_i - \mu) \\
-(1 - 2\theta_2) \sum_{i=N_1+1}^{N_2} (y_i - \mu) + (1 - 2\theta_2) \sum_{i=N_2+1}^{N_3} (y_i - \mu) \\
+(1 - 2\theta_1) \sum_{i=N_3+1}^{N} (y_i - \mu) \right\}. \tag{6}
\]

Under the null hypothesis, the MLEs of \( \mu \) and \( \sigma^2 \) are

\[
\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} y_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\mu})^2,
\]

respectively. For a given \( x \in [0, D] \), to form the score test statistic we need to replace \( \mu \) and \( \sigma^2 \) in \( u(\mu, \sigma^2, x) \) by \( \hat{\mu} \) and \( \hat{\sigma}^2 \).

For ease of notation denote

\[
\bar{y}_t = \frac{1}{n_t} \sum_{i=N_{t-1}+1}^{N_t} y_i, \quad t = 1, 2, 3, 4.
\]

\[
a(x, N) = (\theta_1 - \theta_2)n_2 - (1 - \theta_1 - \theta_2)n_3 - (1 - 2\theta_1)n_4,
\]

\[
b(x, N) = (\theta_2 - \theta_1)n_1 - (1 - 2\theta_2)n_3 - (1 - \theta_1 - \theta_2)n_4,
\]

\[
c(x, N) = (1 - \theta_1 - \theta_2)n_1 + (1 - 2\theta_2)n_2 + (\theta_1 - \theta_2)n_4
\]

\[
d(x, N) = (1 - 2\theta_1)n_1 + (1 - \theta_1 - \theta_2)n_2 + (\theta_2 - \theta_1)n_3,
\]

7
where $\mathbf{N} = (n_1, n_2, n_3, n_4)$. We can then write

$$u(\hat{\mu}, \hat{\sigma}^2, x) = \frac{2}{N\hat{\sigma}^2} [a(x, \mathbf{N})n_1\overline{y}_1 + b(x, \mathbf{N})n_2\overline{y}_2 + c(x, \mathbf{N})n_3\overline{y}_3 + d(x, \mathbf{N})n_4\overline{y}_4],$$

where, under the null hypothesis, the variance of $u(\hat{\mu}, \hat{\sigma}^2, x)$ is

$$\text{Var}(u(\hat{\mu}, \hat{\sigma}^2, x)) = \left(\frac{2}{N\hat{\sigma}^2}\right)^2 [a^2(x, \mathbf{N})n_1 + b^2(x, \mathbf{N})n_2 + c^2(x, \mathbf{N})n_3 + d^2(x, \mathbf{N})n_4].$$

The score test statistic, $U(x)$, is $u(\hat{\mu}, \hat{\sigma}^2, x)$ divided by its standard deviation, that is,

$$U(x) = \frac{u(\hat{\mu}, \hat{\sigma}^2, x)}{\sqrt{\text{Var}(u(\hat{\mu}, \hat{\sigma}^2, x))}} \quad (7)$$

We see that the calculation of the score test statistic $U(x)$ is only based on sample sizes, sample means, sample standard deviation, and Haldane distances (or recombination fractions between the markers and the QTL). This makes the score test computationally very simple (in contrast to the computation of the MLR test statistic). Because of (3), we can replace $\hat{\sigma}^2$ asymptotically equivalent form of the score test statistic. To do this we derive a simplified and computationally very simple (in contrast to the computation of the MLR test statistic). This makes the score test statistic are asymptotically equivalent to the maximum likelihood ratio statistic, as we now show.

For a given $x \in [0, D]$, let $\hat{\Delta}_x$, $\hat{\mu}_x$, and $\hat{\sigma}^2_x$ be the MLEs of $\Delta, \mu$, and $\sigma^2$. By Cox and Hinkley (1974, pages 323-324), the LR test statistic and the square of the score test statistic are asymptotically equivalent at this $x$, that is,

$$2[\ell(\hat{\Delta}_x, \hat{\mu}_x, \hat{\sigma}^2_x, x) - \ell(0, \hat{\mu}, \hat{\sigma}^2, x)] \approx U^2(x). \quad (8)$$

The equivalence of the maximum over both sides of (8), over the range of $x \in [0, D]$, is not trivial. This result is proved in Appendix A.3, where we establish the following theorem.

**Theorem 1** Under the conditions in Section 4.1

$$\lim_{N \to \infty} \sup_{x \in [0, D]} \left| 2 \{\ell_N(\hat{\eta}^{(N)}(x), x) - \ell_N(\hat{\eta}^{(N)}_0)\} - \sup_{x \in [0, D]} U^2(x) \right| = 0 \text{ a.s.}$$

where the log likelihood $\ell$ and the score statistics $U(x)^2$ are described in Section 3, and the almost sure convergence is with respect to the distribution of the phenotypic trait under $H_0$.

Thus, the difference between the two sides in (8) converges to zero with probability one uniformly for $x \in [0, D]$ as $N \to \infty$.

To test (4) we need the null distribution of the maximum of $U(x)$ (which is equivalent to the MLR test statistic). To do this we derive a simplified and asymptotically equivalent form of the score test statistic. Because of (3), we can replace $n_1, n_2, n_3,$ and $n_4$ in $U(x)$ by $\frac{1}{2}(1-r)N, \frac{1}{2}rN, \frac{1}{2}rN$, and $\frac{1}{2}(1-r)N$, respectively. We then have

$$U(x) \approx U^*(x) = \left(\frac{\sqrt{N}}{2\hat{\sigma}}\right) \frac{(1-r)(1-2\theta_1)(\overline{y}_2-\overline{y}_1) + r(1-2\theta_2)(\overline{y}_4-\overline{y}_3)}{(1-r)(1-2\theta_1)^2 + r(1-2\theta_2)^2}. \quad (9)$$
Under the null hypothesis,
\[
\frac{1}{2} \sqrt{(1-r)N} \frac{(y_1 - y_2)}{\hat{\sigma}} \approx W_1 \quad \text{and} \quad \frac{1}{2} \sqrt{rN} \frac{(y_1 - y_2)}{\hat{\sigma}} \approx W_2, \tag{10}
\]
where \(W_1\) and \(W_2\) are two independent standard normal random variables. Thus, \(U^*(x)\) is asymptotically equivalent to
\[
Z(x) = \frac{(1-2\theta_1)\sqrt{1-r}W_1 + (1-2\theta_2)\sqrt{r}W_2}{\sqrt{(1-r)(1-2\theta_1)^2 + r(1-2\theta_2)^2}}. \tag{11}
\]

Next, consider two QTL positions \(x'\) and \(x''\) \(\in [0, D]\), whose corresponding probabilities in (2) can be written as \((\theta'_1, \theta'_2) = \Theta(x', D)\) and \((\theta''_1, \theta''_2) = \Theta(x'', D)\). The covariance between \(Z(x')\) and \(Z(x'')\) is
\[
\text{cov}(Z(x'), Z(x'')) = \frac{(1-r)(1-2\theta'_1)(1-2\theta''_1) + r(1-2\theta'_2)(1-2\theta''_2)}{\sqrt{(1-r)(1-2\theta'_1)^2 + r(1-2\theta'_2)^2}\sqrt{(1-r)(1-2\theta''_1)^2 + r(1-2\theta''_2)^2}}. \tag{12}
\]
Therefore, the MLR test statistic or the maximum of the square of score test statistics asymptotically distributed as
\[
\sup_{0 \leq x \leq D} Z^2(x), \tag{13}
\]
where \(Z(x)\) is a Gaussian stochastic process with mean 0 and covariances (12). Note that the distribution of (13) only depends on the Haldane distance \(D\) or the recombination fraction \(r\) between two markers \(M^{\ell-1}\) and \(M^{\ell}\). In Section 5, we will compute the distribution of (13) through simulation.

4 Composite Interval Mapping

In the preceding section, we discussed the asymptotic theory of the score test statistic for QTL mapping from a single marker interval. Such an interval mapping approach can not adequately use information from all possible markers on the genome. In this section, we use the full marker information from the entire genome to derive the score test statistics for mapping a QTL. Zeng (1994) proposed a so-called composite interval mapping technique to increase the precision of QTL detection by controlling the chromosomal region outside the marker interval under consideration. Statistically, composite interval mapping is a combination of interval mapping based on two given flanking markers and a partial regression analysis on all markers except for the two ones bracketing the QTL. We consider two different situations associated with a sparse genetic map and a dense genetic map, respectively.
4.1 QTL mapping on a sparse map

Suppose there are $k+1$ diallelic markers, $M^0, M^1, \ldots, M^k$, on a sparse linkage map. These markers encompass $k$ marker intervals. In a backcross population, four marker genotypes at a given interval $[M^{\ell-1}, M^{\ell}]$ ($\ell = 1, \ldots, k$) are indicated by

$$G^{(\ell)} = \begin{cases} 
1 & \text{if a marker genotype is } M^{\ell-1}m^{\ell-1}M^{\ell}m^{\ell} \\
2 & \text{if a marker genotype is } M^{\ell-1}m^{\ell-1}m^{\ell}m^{\ell} \\
3 & \text{if a marker genotype is } m^{\ell-1}m^{\ell-1}M^{\ell}m^{\ell} \\
4 & \text{if a marker genotype is } m^{\ell-1}m^{\ell-1}m^{\ell}m^{\ell}.
\end{cases}$$

The number of observations within each of these four marker genotypes at a given marker interval can be expressed as

$$n_{i}^{(\ell)} = \sum_{i=1}^{N} I(G_{i}^{(\ell)} = t), \quad t = 1, 2, 3, 4$$

where $I(A)$ is the indicator function of event $A$. We then use

$$N^{(\ell)} = (n_{1}^{(\ell)}, n_{2}^{(\ell)}, n_{3}^{(\ell)}, n_{4}^{(\ell)})$$

to array the observations of all the four marker genotypes at the interval. We can then express the phenotypic mean of a quantitative trait ($y$) for a marker genotype $t$ at the interval $[M^{\ell-1}, M^{\ell}]$ as

$$\bar{y}_{i}^{(\ell)} = \frac{1}{n_{i}^{(\ell)}} \sum_{i=1}^{N} y_{i} I(G_{i}^{(\ell)} = t).$$

Let $D_{\ell}$ and $r_{\ell}$ ($\ell = 1, \ldots, k$) denote the Haldane distance and recombination fraction between $M^{\ell-1}$ and $M^{\ell}$, respectively, and denote the Haldane distance between the two extreme markers $M^{0}$ and $M^{k}$ by $D = \sum_{\ell=1}^{k} D_{\ell}$. A putative QTL must be located somewhere on the map, at unknown position $d \in [0, D]$. For each $d \in [0, D]$ define

$$\ell_{d} = \min\{\ell : \sum_{i=1}^{\ell-1} D_{i} \leq d < \sum_{i=1}^{\ell} D_{i}\} \quad \text{and} \quad x_{d} = D - \sum_{i=1}^{\ell-1} D_{i},$$

(14)

to identify the marker $M_{\ell_{d}-1}$ and possible QTL within the interval $[M^{\ell_{d}-1}, M^{\ell_{d}}]$ that is associated with the distance $d$. As derived in Section 3, the score test statistic (7) for detecting the QTL at $d \in [0, D]$, the asymptotically equivalent forms (9) and (11), and the relationship (10) are all valid with $N_{i}, y_{i}$, and $r$ replaced by $N^{(\ell)}, \bar{y}_{i}^{(\ell)}$, and $r_{\ell}$, respectively. In (7), (9), and (11), $(\theta_{1}, \theta_{2}) = \theta(x, D_{\ell})$ is the function of $x$ and $D_{\ell}$ defined in (2).

For example, we express (11) as

$$Z(d) = \frac{(1 - 2\theta_{1}) \sqrt{1 - r_{\ell}} W_{1}^{(\ell)} + (1 - 2\theta_{2}) \sqrt{r_{\ell}} W_{2}^{(\ell)}}{\sqrt{(1 - r_{\ell})(1 - 2\theta_{1})^{2} + r_{\ell}(1 - 2\theta_{2})^{2}}},$$

(15)
where $W_1^{(\ell)}$ and $W_2^{(\ell)}$ are two independent standard normal random variables,
\[
W_1^{(\ell)} \approx \frac{1}{2} \sqrt{1 - r_\ell} N \left( \frac{y_4^{(\ell)} - \bar{y}_1^{(\ell)}}{\hat{\sigma}} \right) \quad \text{and} \quad W_2^{(\ell)} \approx \frac{1}{2} \sqrt{r_\ell} N \left( \frac{y_3^{(\ell)} - \bar{y}_2^{(\ell)}}{\hat{\sigma}} \right). \tag{16}
\]

In order to search for a QTL throughout the entire genome, we need to compute the covariance between $Z(d')$ and $Z(d'')$, $0 \leq d' \leq d'' \leq D$. Assume $(\ell', x')$ and $(\ell'', x'')$ are associated with $d'$ and $d''$ through (14). Let $(\theta_1^1, \theta_2^1) = \theta(x', D_{\ell'})$ and $(\theta_1^2, \theta_2^2) = \theta(x'', D_{\ell''})$ as defined in (2). If $\ell' = \ell'' = \ell$, then the covariance between $Z(d')$ and $Z(d'')$ is given in (12) with $r$ replaced by $r_\ell$ (see Appendix A.1). It now follows that the MLR test statistic or the maximum of the square of score test statistics is asymptotically distributed as
\[
\sup_{0 \leq d \leq D} Z^2(d), \tag{17}
\]
where $Z(d)$ is a Gaussian stochastic process with mean 0 and covariance in (12) when $\ell' = \ell''$ or in (31) when $\ell' < \ell''$. Note that the distribution (12) only depends on Haldane distances or recombination fractions between the markers. In Section 5, we will use a numerical approach to compute the distribution of (17).

### 4.2 QTL mapping on a dense map

For a dense genetic map, we can assume that markers are located at every point along the genome. This can be considered as the limit of the sparse map when $k$ (the number of diallelic markers) converges to infinity and the genonic distance between any two consecutive markers converges to zero.

For a dense map, two markers $\mathcal{M}^{\ell-1}$ and $\mathcal{M}^\ell$ can be assumed to have the identical distance $(d)$ to marker $\mathcal{M}^0$. Furthermore, for any interval $[\mathcal{M}^{\ell-1}, \mathcal{M}^\ell]$, we have $v_2^{(\ell)} \approx v_3^{(\ell)} \approx 0$, $n_1^{(\ell)} \approx n_4^{(\ell)} \approx \frac{1}{2} N$, $\theta_1 \approx 0$, $a(x, N^{(\ell)}) \approx -n_4^{(\ell)}$, and $d(x, N^{(\ell)}) \approx n_1^{(\ell)}$. If the Haldane distance between $\mathcal{M}^0$ and the QTL, which is located within interval $[\mathcal{M}^{\ell-1}, \mathcal{M}^\ell]$, is $d \in [0, D]$, the score statistic (7) for detecting the QTL at $d$ is reduced to
\[
U(d) = \frac{y_4^{(\ell)} - \bar{y}_1^{(\ell)}}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_4}}}.
\]

For two possible locations $d'$ and $d''$ of the QTL, $d' < d''$, where $d'$ and $d''$ belong to $[\mathcal{M}^{\ell-1}, \mathcal{M}^\ell]$ and $[\mathcal{M}^{\ell'-1}, \mathcal{M}^{\ell''}]$, the covariance (31) is reduced to
\[
\text{cov}(U(d'), U(d'')) \approx 1 - 2r_{\ell', \ell''} \approx 1 - 2r_{\ell' \ell''}. \tag{18}
\]
This suggests that the asymptotic distribution of MLR test statistic, or the maximum of the square of the score test statistics, is the same as the distribution of
\[
\sup_{0 \leq d \leq D} Z^2(d),
\]

11
where $Z(\cdot)$ is a Gaussian stochastic process with mean 0 and covariance (18). This result is consistent with the finding of Lander and Botstein (1989), which can be viewed as a special case of Section 4.1 where $k \to \infty$ and the genomic distance between any two consecutive markers converges to zero.

## 5 Simulation of the Null Distribution

Although it is difficult to derive an explicit formula for the asymptotic null distribution of the MLR test statistic or the maximum of the square of score test statistics derived in Sections 3 and 4.1, it is relatively straightforward to simulate it. Here, we show how to compute the maxima of the these statistics, and give a numerical method to simulate the asymptotic null distributions of (13) and (17) for a sparse genetic map.

### 5.1 Interval mapping

For any given values of two standard normal random variables $W_1$ and $W_2$, we can compute

$$\sup_{0 \leq x \leq D} Z^2(x)$$

where $Z(x)$ is given in (11). Although $x$ does not appear explicitly in $Z(x)$, recall from (2) that $(\theta_1, \theta_2) = \theta(x, D)$, so $Z^2(x)$ is a function of $x$. By straightforward computation, it can be shown that $\frac{dZ^2(x)}{dx} = 0$ is equivalent to

$$(e^{4x-2D} - e^{2D-4x})(W_2^2 - W_1^2) + \frac{2W_1W_2e^{-2D}}{\sqrt{1-e^{-4D}}}((e^{2D-4x} + e^{4x-2D} - 2e^{2D}) = 0. \quad (20)$$

Defining $y = e^{4x-2D}$, the solutions to (20) are

$$y_{1,2} = \frac{-2e^{2D}W_1W_2 \pm (W_1^2 + W_2^2)\sqrt{e^{4D} - 1}}{(W_1^2 - W_2^2)\sqrt{e^{4D} - 1} - 2W_1W_2}. \quad (21)$$

Therefore, in terms of $x$ the solutions of (20) are

$$x_i = \frac{1}{4}(\log y_i + 2D), \quad i = 1, 2, \quad (22)$$

provided that $y_i > 0$.

We can then simulate (19) by simulating $W_1$ and $W_2$ iid standard normal and taking

$$\sup_{0 \leq x \leq D} Z^2(x) = \begin{cases} \max\{Z^2(0), Z^2(x_i), Z^2(D)\} & \text{if } y_i > 0 \text{ and } 0 \leq x_i \leq D \\ \max\{Z^2(0), Z^2(D)\} & \text{otherwise} \end{cases}. \quad (23)$$

By repeating this many times (for example, 10,000 times), we obtain the approximate distribution of (19), i.e., the asymptotic null distribution of the MLR test or the maximum of the square of score test statistics. From such a distribution, we can obtain the critical thresholds at any significance level.
5.2 Composite interval mapping

The expression (15), which is asymptotically equivalent to the score test statistic, involves standard normal random variables $W_{i}^{(ℓ)}$, $i = 1, 2; ℓ = 1, ⋯ , k$. From the derivation in Appendix A.1, $W_{1}^{(1)}, W_{2}^{(1)}, W_{1}^{(ℓ)}$, $ℓ = 2, ⋯ , k$, follow a joint multivariate normal distribution with mean 0 and variance-covariance matrix Σ, where cov($W_{1}^{(1)}, W_{2}^{(1)}$) = 0 and the remaining covariances are given by (29) and (30) in Appendix A.1. It is also shown that

$$W_{2}^{(ℓ+1)} = \frac{1}{\sqrt{r_{ℓ+1}}} \left( \sqrt{1 - r_{ℓ}}W_{1}^{(ℓ)} - \sqrt{r_{ℓ}}W_{2}^{(ℓ)} - \sqrt{1 - r_{ℓ+1}}W_{1}^{(ℓ+1)} \right),$$

(24)

To simulate the null distribution of $\sup_{0 ≤ d ≤ D} Z^{2}(d)$, where $Z(d)$ is in (15), we can do the following:

(i) Generate $(W_{1}^{(1)}, W_{2}^{(1)}, W_{1}^{(ℓ)}, ℓ = 2, ⋯ , k) \sim N(0, Σ)$;

(ii) Compute $W_{2}^{(ℓ+1)}$, $ℓ = 1, 2, ⋯ , k - 1$ by (24);

(iii) For each $ℓ$, find

$$\sup_{\sum_{i=1}^{ℓ-1} D_{i} ≤ d ≤ \sum_{i=1}^{ℓ} D_{i}} Z^{2}(d)$$

(25)

using (23);

(iv) Take the maximum value of (25) over $ℓ = 1, 2, ⋯ , k$.

Repeat the above steps many times (for example, 10,000 times), to obtain the approximate distribution of $\sup_{0 ≤ d ≤ D} Z^{2}(d)$, the asymptotic null distribution of the MLR test or the maximum of the square of score test statistics. The critical thresholds can be obtained accordingly.

Note that the random variables $W_{1}^{(1)}, W_{2}^{(1)}$ and $W_{1}^{(ℓ)}$, $ℓ = 2, 3, ⋯ , k$, can be calculated as $A'Z$, where $A$ is the Cholesky factorization of $Σ$ (that is, $A'A = Σ$) and $Z$ follows multivariate standard normal distribution. Note that we only need to factor the matrix once for the entire simulation.

6 Examples

To examine the statistical properties of our score test statistic proposed, we perform a simulation study to determine the asymptotic null distribution and the critical threshold for declaring the existence of a significant QTL. Our results here are compared to those obtained by Lander and Botstein (1989), Doerge and Churchill (1996) and Piepho (2001). The score test statistic is further validated using a case study from a forest tree genome project.
6.1 Simulation

We use the same simulation design as used by Lander and Botstein (1989) to simulate a backcross mapping population of size 250. Thus, our results can be directly compared with those of Lander and Botstein. In our simulation study, equal spacing of the six markers is assumed throughout 12 chromosomes of 100 cM each. Five QTL are hypothesized to be located at genetic positions 70, 49, 27, 8 and 3 cM from the left end on the first five chromosomes with effects of 1.5, 1.25, 1.0, 0.75 and 0.50, respectively. For each individual, genotypes at the markers were generated assuming no interference. The corresponding quantitative trait was simulated by summing the five QTL effects and random standard normal noise.

Both the likelihood approach of Lander and Botstein (1989) and the score-statistic approach are used to analyze the simulated marker and phenotype data. The QTL likelihood profiles are drawn from the likelihood ratio test statistics and the score test statistics. As shown in Figure A.3.5, these two statistics gave similar profiles across each of the simulated linkage groups. Based on 10,000 simulations under the null model without a QTL, the 95% percentile of score test statistics is 12.83, which is used as the threshold for declaring a QTL at the significance level \( \alpha = 0.05 \).

We also look at other approaches to estimate the threshold for the same data set. All these approaches obtain a threshold value similar to ours. For example, Lander and Botstein (1989) suggested using the Bonferroni method for the sparse-map case, which yields a threshold of 11.16 at the 0.05 level, given a total of 60 intervals tested. For a dense map, we would solve \( \alpha = (C + Gt)Pr(\chi^2_1 > t) \), where \( C \) is the number of chromosomes and \( G \) is the total genetic length, which gives a cut-off point of 14.33. The critical threshold from the quick method of Piepho (2001) is 12.96 at the 0.05 level. His method solves equation \( \alpha = CP\chi^2_1 > t) + V \exp(-t/2) = \sqrt{2}\pi, \) where \( V = 75.83 \) is the sum of absolute differences in the square roots of score test statistics between every two adjacent hypothesized QTL positions across the genome. Churchill and Doerge (1994) proposed the permutation test approach to estimate an empirical threshold. This approach gives 11.44 for our simulated data based on 10,000 simulation replicates.

6.2 A case study

We use an example to further examine the performance of our statistical method. The study we consider was derived from an interspecific hybridization of *Populus* (poplar). A *P. deltoides* clone (designated I-69) was used as a female parent to mate with a *P. euramericana* clone (designated I-45) as a male parent (Wu et al. 1992). Both *P. deltoides* I-69 and *P. euramericana* I-45 were selected at the Research Institute for Poplars in Italy in the 1950s and were introduced to China in 1972. A genetic linkage map has been constructed using 90 genotypes randomly selected from the 450 hybrids with random amplified polymorphic DNAs (RAPDs), amplified fraction length polymorphisms (AFLPs) and inter-
simple sequence repeats (ISSRs) (Yin et al. 2002). This map comprises the 19 largest linkage groups for each parental map, which roughly represent 19 pairs of chromosomes. The 90 hybrid genotypes used for map construction were measured for wood density with wood samples collected from 11-year-old stems in a field trial in a completely randomized design. The measurement for each genotype was repeated 4 – 6 times to reduce measurement errors. The means of these genotypes were calculated and used for QTL mapping here.

The 19 linkage groups constructed for each parent were scanned for possible existence of QTL affecting wood density using both our score statistic and the likelihood approach. We successfully detected a wood density QTL on linkage group, D17, composed of 16 normally segregated markers. The QTL profiles obtained from these two approaches, as shown in Figure 2, consistently exhibit a marked peak within a narrow marker interval AG/CGA-480 – AG/CGA-330. Both the score statistic value and log-likelihood ratio at the peak are greater than the threshold of 10.10 ($\alpha = 0.05$) obtained from the simulated asymptotic null distribution. Other approaches are also used to estimate the $\alpha = 0.05$ level threshold, which give 7.87 for permutation tests, 8.87 for Piepho’s quick method, 8.73 for the Bonferroni method and 10.20 for Lander and Botstein’s formula for dense markers. All these suggest that both approaches have consistently detected a significant QTL, located at the peak of the profiles, affecting wood density in hybrid poplars.

The similarity of the QTL profiles from our score statistic and the likelihood approach, in conjunction with consistent thresholds calculated from these two different approaches, suggest that the score statistic has the power to detect the hypothesized QTL as well as the likelihood approach. However, in the wood density example, the calculation of the score statistic threshold based on 10,000 simulations from the asymptotic null distribution only took 5 minutes, where the calculation of the threshold based on 10,000 permutation likelihood ratio tests took 4,100 minutes. (All calculations were performed using Matlab software on a Dell Inspiron 7000 with a 300 MHz CPU.) Thus, the score statistic approach has a considerable advantage in reducing computing times.

We also used an EM algorithm to estimate the additive effect of this significant QTL detected on wood density. This results in two QTL genotypes to differ in wood density by 0.033, or equivalently 7% relative to the overall mean. This QTL is found to explain about 30% of the phenotypic variance for wood density in hybrid poplars.

7 Discussion

The idea of associating a continuous phenotype with a discrete trait (marker) dates back to Sax (1923), but it was Lander and Botstein (1989) who established an explicit analytical principle for such an association study and provided a tractable statistical algorithm for dissecting a quantitative trait into its individual QTL components. They embedded a traditional quantitative genetic model within a likelihood based statistical framework, implemented with the
EM algorithm developed by Dempster et al. (1977). This statistical strategy has now been used as a standard approach for QTL mapping in a variety of organisms (Mackay 2001; Barton and Keightley 2002; Doerge 2002). Although the maximum likelihood method has many desirable properties from a statistical perspective, its implementation requires intensive computation in maximizing the likelihood function. For many complex genetic models, this method can be computationally prohibitive.

In this article, we devise a score-statistic method for QTL mapping based on a genetic linkage map. A score test statistic is much easier to compute. Because of the simple structure of the score test statistic, the null asymptotic distribution of the maximum of the square of score test statistics can be readily derived, and is asymptotically equivalent to the likelihood ratio test statistic. Moreover, the asymptotic null distribution is the distribution of the maximum of the square of a well-defined Gaussian stochastic process. The critical thresholds obtained in Section 5 can be used for both the MLR test and the maximum of the square of score test statistics.

Many authors advocated using score tests. Davies (1977, 1987) proposed the utility of the maximum of score test statistics when a nuisance parameter is present only under the alternative. Some authors pointed out that LR tests are often not most powerful (see, for example, Berger 1997). More recently, score test approaches were developed for QTL mapping in humans to facilitate computation (Schaid et al. 2002; Wang and Huang 2002). A study of the association of haplotype with humoral immune response to measles vaccination was performed to illustrate the utility of score tests (Schaid et al. 2002).

The computational method in Section 5 is feasible even for a large $k$, the number of molecular markers in a genetic linkage map. For a dense map, we can assume that a QTL is located at the same position as a marker. In this case, the numbers of recombinant groups $n_2^{(r)}$ and $n_3^{(r)}$ approaches zero for all intervals and our results on the asymptotic distribution of MLR test statistic in Section 4 reduce to that of Lander and Botstein (1989).

Here we used a simple backcross design to demonstrate the idea of the application of score test statistics to QTL mapping. The proposed method can be readily extended to more complex designs, such as the F2 design, full-sib family mapping (Wu et al. 2002; Lin et al. 2003) and natural populations, and to model multiple QTL with epistasis. Our approach was devised on the basis of an ideal set of molecular and phenotypic data that have no missing values, no genotypic errors and whose markers segregate in the expected Mendelian ratio. In practice, however, mapping data is complicated by all of these factors (Shields et al. 1991). One of the most significant effects of these factors in a QTL mapping project is that they always lead to considerably more computational times. Our approach would be particularly helpful in increasing computational efficiency arising from data imperfection. With thorough investigations on these experimental complexities, the score test statistic may be more broadly applied to unravel the genetic basis of quantitative variation in agriculture, biology and biomedicine.
Acknowledgment

This work is partially supported by grants from the National Science Foundation to G. C. (DMS9971586) and the University of Florida Research Opportunity Fund to R. W. (02050259). The publication of this manuscript is approved as journal series R-xxxx by the Florida Agricultural Experiment Station.
A  Some Technical Details

A.1 The Covariance of \( Z(d') \) and \( Z(d'') \) in (31)

Assume \((\ell', x')\) and \((\ell'', x'')\) are associated with \(d'\) and \(d''\) through (14), respectively. Since the covariance of \(Z(d')\) and \(Z(d'')\) is available in (12) for \(\ell' = \ell''\), we derive the covariance for \(\ell' < \ell''\).

Let \(A_{s,t', t''}\) denote the subset of individuals who belong to Group \(s\) in interval \(\ell'\) and belong to Group \(t\) in interval \(\ell''\), i.e.,

\[
A_{s,t', t''} = \{i: 1 \leq i \leq N, \ G_{i\ell'} = s, \ G_{i\ell''} = t\}, \ 1 \leq s, t \leq 4, \ 1 \leq \ell' \leq \ell'' \leq k.
\]

Denote the size of \(A_{s,t', t''}\) by \(n_{s,t', t''}\), i.e.,

\[
n_{s,t', t''} = \sum_{i=1}^{N} I(G_{i\ell'} = s)I(G_{i\ell''} = t), \ 1 \leq s, t \leq 4, \ 1 \leq \ell' \leq \ell'' \leq k.
\]

Let \(N^{(\ell', \ell'')} = (n_{s,t', t''})_{s,t=1,2,3,4}\) be the \(4 \times 4\) matrix with \(s, t\) as the row index and column index, respectively. It is seen that

\[
\lim_{N \to \infty} \frac{1}{N} N^{(\ell', \ell'')} = \frac{1}{2} \begin{pmatrix}
1 - r_{\ell'} & 0 & 0 & 0 \\
0 & r_{\ell'} & 0 & 0 \\
0 & 0 & r_{\ell'} & 0 \\
0 & 0 & 0 & 1 - r_{\ell'}
\end{pmatrix}
\]

\[
\times \begin{pmatrix}
1 - r_{\ell', \ell'' - 1} & 1 - r_{\ell', \ell''} & r_{\ell', \ell'' - 1} & r_{\ell', \ell'' - 1} \\
r_{\ell', \ell'' - 1} & 1 - r_{\ell', \ell''} & r_{\ell', \ell'' - 1} & r_{\ell', \ell'' - 1} \\
1 - r_{\ell', \ell'' - 1} & 1 - r_{\ell', \ell''} & r_{\ell', \ell'' - 1} & r_{\ell', \ell'' - 1} \\
r_{\ell', \ell'' - 1} & r_{\ell', \ell'' - 1} & 1 - r_{\ell', \ell''} & 1 - r_{\ell', \ell'' - 1}
\end{pmatrix}
\]

\[
\times \begin{pmatrix}
1 - r_{\ell''} & 0 & 0 & 0 \\
0 & r_{\ell''} & 0 & 0 \\
0 & 0 & r_{\ell''} & 0 \\
0 & 0 & 0 & 1 - r_{\ell''}
\end{pmatrix}.
\]

In addition, we have results similar to (3):

\[
\lim_{N \to \infty} \frac{n_{1}^{(\ell')}}{N} = \lim_{N \to \infty} \frac{n_{4}^{(\ell')}}{N} = \frac{1}{2} (1 - r_{\ell'}),
\]

\[
\lim_{N \to \infty} \frac{n_{2}^{(\ell')}}{N} = \lim_{N \to \infty} \frac{n_{3}^{(\ell')}}{N} = \frac{1}{2} r_{\ell'},
\]

\[
\lim_{N \to \infty} \frac{n_{1}^{(\ell'')}}{N} = \lim_{N \to \infty} \frac{n_{4}^{(\ell'')}}{N} = \frac{1}{2} (1 - r_{\ell''}) \ \ (27)
\]

\[
\lim_{N \to \infty} \frac{n_{2}^{(\ell'')}}{N} = \lim_{N \to \infty} \frac{n_{3}^{(\ell'')}}{N} = \frac{1}{2} r_{\ell''}. \ \ (28)
\]
Using (7), (26), and (27), we have

\[
\text{cov}(W_1^{(\ell')}, W_1^{(\ell'\prime)}) \approx \frac{N}{4\sigma^2} \sqrt{(1 - r_{\ell'})(1 - r_{\ell'\prime})} \\
\times \text{cov}(\bar{y}_1^{(\ell')}, \bar{y}_1^{(\ell'\prime)} - \bar{y}_1)
\]

\[
= \frac{N}{4} \sqrt{(1 - r_{\ell'})(1 - r_{\ell'\prime})} \\
\times \left( \frac{n_{4,1,\ell',\ell'\prime}}{n_4^{(\ell')}} - \frac{n_{4,1,1,\ell'\prime}}{n_4^{(\ell')}} - \frac{n_{4,1,\ell',1}}{n_4^{(\ell')}} + \frac{n_{4,1,1,1}}{n_4^{(\ell')}} \right)
\approx \sqrt{(1 - r_{\ell'})(1 - r_{\ell'\prime})}(1 - 2r_{\ell',\ell'\prime})
\]

Similarly, we have

\[
\text{cov}(W_1^{(\ell')}, W_2^{(\ell'\prime)}) \approx \sqrt{(1 - r_{\ell'})(1 - 2r_{\ell',\ell'\prime})},
\]

\[
\text{cov}(W_2^{(\ell')}, W_1^{(\ell'\prime)}) \approx -\sqrt{r_{\ell'}(1 - r_{\ell'\prime})}(1 - 2r_{\ell,\ell'\prime}),
\]

\[
\text{cov}(W_2^{(\ell')}, W_2^{(\ell'\prime)}) \approx -\sqrt{r_{\ell'\prime}(1 - 2r_{\ell'\prime})}(1 - 2r_{\ell,\ell'\prime}).
\]

By (29) and (30), for \( \ell' < \ell'\prime \) the covariance of \( Z(d') \) and \( Z(d'\prime) \) is

\[
\text{cov}(Z(d'), Z(d'\prime)) = (1 - 2r_{\ell',\ell'\prime}) \]

\[
 \times \left[ (1 - 2\ell')(1 - 2\ell'\prime)(1 - r_{\ell'})(1 - r_{\ell'\prime}) \\
+ (1 - 2\ell')(1 - 2\ell'\prime)(1 - r_{\ell'})r_{\ell'\prime} - (1 - 2\ell')(1 - 2\ell'\prime)(1 - r_{\ell'})r_{\ell'} \\
- (1 - 2\ell')(1 - 2\ell'\prime)r_{\ell'}r_{\ell'\prime} \right] \\
\div \left[ \sqrt{(1 - r_{\ell'})(1 - 2\ell')^2 + r_{\ell'}(1 - 2\ell')^2} \\
\sqrt{(1 - r_{\ell'\prime})(1 - 2\ell'\prime)^2 + r_{\ell'\prime}(1 - 2\ell'\prime)^2} \right],
\]

where \( r_{\ell',\ell'\prime} \) denotes the recombination fraction between two markers \( \mathcal{M}_{\ell'} \) and \( \mathcal{M}_{\ell'\prime} \) (\( \ell' \leq \ell'\prime \)), where \( r_{\ell-1,\ell} = r_{\ell} \) and \( r_{\ell,\ell} = 0 \).

### A.2 The Linear Relationship Between \( W_1^{(\ell)}, W_2^{(\ell)}, W_1^{(\ell+1)} \), and \( W_2^{(\ell+1)} \) in (24)

From (29) and (30), the variance-covariance matrix of \( W_1^{(\ell)}, W_2^{(\ell)}, W_1^{(\ell+1)}, \) and \( W_2^{(\ell+1)} \) has the form

\[
\begin{pmatrix}
1 & 0 & \sqrt{(1 - r_{\ell})(1 - r_{\ell+1})} & \sqrt{(1 - r_{\ell})r_{\ell+1}} \\
0 & 1 & -\sqrt{r_{\ell}(1 - r_{\ell+1})} & -\sqrt{r_{\ell}r_{\ell+1}} \\
\sqrt{(1 - r_{\ell})(1 - r_{\ell+1})} & -\sqrt{r_{\ell}(1 - r_{\ell+1})} & 1 & 0 \\
\sqrt{(1 - r_{\ell})r_{\ell+1}} & -\sqrt{r_{\ell}r_{\ell+1}} & 0 & 1
\end{pmatrix}
\]
It can be seen that the above matrix is singular and hence (24) holds. The relationship (24) has an intuitive interpretation. Denote by $B_{s,\ell}$ the subset of individuals who belong to Group $s$ in interval $\ell$, i.e.,
\[
B_{s,\ell} = \{ i : 1 \leq i \leq N, G_{i\ell} = s \}, \quad 1 \leq s \leq 4; \quad 1 \leq \ell \leq k.
\]
Consider two consecutive intervals $[M_{\ell-1}, M_{\ell}]$ and $[M_{\ell}, M_{\ell+1}]$. For any individual in Group 1 or Group 3 (Group 2 or Group 4), the marker condition is $(M^\ell, m^\ell)$ $(m^\ell, m^\ell)$ at marker $M_{\ell}$. Thus this individual must belong to Group 1 or Group 2 (Group 3 or Group 4) in interval $[M_{\ell}, M_{\ell+1}]$. Therefore,
\[
B_{1,\ell} \cup B_{3,\ell} = B_{1,\ell+1} \cup B_{2,\ell+1}
\]
and similarly,
\[
B_{2,\ell} \cup B_{4,\ell} = B_{3,\ell+1} \cup B_{4,\ell+1}.
\]
Consequently,
\[
n_1^{(\ell)} y_1^{(\ell)} + n_3^{(\ell)} y_3^{(\ell)} = n_1^{(\ell+1)} y_1^{(\ell+1)} + n_2^{(\ell+1)} y_2^{(\ell+1)}
\]
and
\[
n_2^{(\ell)} y_2^{(\ell)} + n_4^{(\ell)} y_4^{(\ell)} = n_3^{(\ell+1)} y_3^{(\ell+1)} + n_4^{(\ell+1)} y_4^{(\ell+1)}.
\]
(32) Combining (10), (27), and (32) we see that
\[
\sqrt{1 - r_\ell W_1^{(\ell)}} - \sqrt{1 - r_{\ell+1} W_1^{(\ell+1)}} = \sqrt{1 - r_\ell W_2^{(\ell)}} - \sqrt{1 - r_{\ell+1} W_2^{(\ell+1)}}
\]
\[
\approx \frac{\sqrt{N}}{2} \left[ \left( 1 - r_\ell \right) \left( y_4^{(\ell)} - y_1^{(\ell)} \right) - \left( 1 - r_{\ell+1} \right) \left( y_4^{(\ell+1)} - y_1^{(\ell+1)} \right) \right]
\]
\[
= \frac{1}{\sqrt{N}} \left[ \left( n_4^{(\ell)} y_4^{(\ell)} - n_1^{(\ell)} y_1^{(\ell)} \right) - \left( n_3^{(\ell+1)} y_3^{(\ell+1)} - n_2^{(\ell+1)} y_2^{(\ell+1)} \right) \right]
\]
\[
= 0,
\]
which provides a justification of (24).
A.3 Equivalence of the Suprema of the MLR and Score Statistic

In this section we demonstrate the equivalence of the suprema of the MLR statistic and the score test statistic over one marker interval \( [M^0, M^1] \). The equivalence over multiple intervals follows immediately. Specifically, we prove Theorem 1 of Section 3:

**Theorem 1** For the situation described in Section 4.1

\[
\lim_{N \to \infty} \left| \sup_{x \in [0, D]} 2\{\ell_N(\hat{\eta}^{(N)}(x), x) - \ell_N(\hat{\eta}_0^{(N)})\} - \sup_{x \in [0, D]} U^2(x) \right| = 0 \text{ a.s.}
\]

where the log likelihood \( \ell \) and the score statistics \( U(x)^2 \) are described in Section 3, and the almost sure convergence is with respect to the distribution of the phenotypic trait under \( H_0 \).

The proof of Theorem 1 is based on two lemmas. We will state the two lemmas and then, using them, will prove the theorem. In separate sections we will prove the two lemmas, and two other needed results. Before stating the lemma we review some notation.

The genotype variable \( G^\ell \) is defined at the beginning of Section 4.1. The superscript \( "^\ell" \) will be omitted since we focus on only one interval \( [M^0, M^1] \).

The probability function of \( G \) is defined by

\[
p(g) = P(G = g) = \begin{cases} 
\frac{1}{2}(1 - r), & g = 1 \\
\frac{1}{2}r, & g = 2 \\
\frac{1}{2}r, & g = 3 \\
\frac{1}{2}(1 - r), & g = 4.
\end{cases}
\]

To express the likelihood function in a compact form, we define

\[
\varphi(y; \lambda, \theta) = (1 - \theta)e^{-\frac{1}{2\pi\sigma}(y - \mu_1)^2} + \theta e^{-\frac{1}{2\pi\sigma}(y - \mu_2)^2}.
\]

where \( \lambda = (\mu_1, \mu_2, \sigma)^T \). We will also use the alternative parameterization \( \mu_1 = \mu - \Delta, \mu_2 = \mu + \Delta \) and \( \eta = (\Delta, \mu, \sigma)^T \).

The density function of the phenotype variable \( Y \) conditional on \( G = j \) is

\[
f_{Y|G=j}(y; \lambda, x) = \frac{1}{\sqrt{2\pi}\sigma}\varphi(y; \lambda, \theta^*_j),
\]

where \( \theta^*_j = \theta_1, \theta_2, 1 - \theta_2, 1 - \theta_1 \), for \( j = 1, 2, 3, 4 \), respectively, and \( \theta_1, \theta_2 \) are functions of \( x \) as defined in (2).

Assume the observations are \((Y_i, G_i) = (y_i, g_i), i = 1, \ldots, N\), and are arranged so that the first \( n_1 \) observations belong to group 1; the next \( n_2 \) belong to group 2; and so on. The likelihood functions conditional on the genotype variable are

\[
f(y^{(N)}; \lambda, x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \prod_{i=1}^{N_1} \varphi(y_i; \lambda, \theta_1) \prod_{i=N_1+1}^{N_2} \varphi(y_i; \lambda, \theta_2)
\]
\[ \prod_{i=N_2+1}^{N_3} \varphi(y_i; \lambda, 1 - \theta_2) \prod_{i=N_3+1}^{N} \varphi(y_i; \lambda, 1 - \theta_1), \]

where \( y^{(N)} = (y_1, \ldots, y_N) \). The full likelihood function for observations \((Y, G) = (y_i, g_i), i = 1, \ldots, N\), is

\[ L_N(\lambda; x) = f(y^{(N)}; \lambda, x) \prod_{j=1}^{4} [p(j)]^{n_j}. \quad (33) \]

Since the second factor does not contain parameters of interest, we consider the log likelihood function for \( y^{(N)} \) conditional on the genotype variable \( \ell_N(\lambda; x) = \log f(y^{(N)}; \lambda, x) \), which is given in (5).

For a fixed \( x \in [0, D] \), the mle’s of \( \mu_1, \mu_2, \sigma, \Delta, \mu, \lambda \eta \) are denoted, respectively, by \( \hat{\mu}_1^{(N)}(x), \hat{\mu}_2^{(N)}(x), \hat{\sigma}^{(N)}(x), \hat{\Delta}^{(N)}(x), \hat{\mu}^{(N)}(x) \),

\[ \hat{\lambda}^{(N)}(x) = \left( \hat{\mu}_1^{(N)}(x), \hat{\mu}_2^{(N)}(x), \hat{\sigma}^{(N)}(x) \right)^T \quad \text{and} \quad \hat{\eta}^{(N)}(x) = \left( \hat{\Delta}^{(N)}(x), \hat{\mu}^{(N)}(x), \hat{\sigma}^{(N)}(x) \right)^T, \]

with \( \hat{\mu}_0 = \mu_0(x) - \hat{\Delta}^{(N)}(x) \) and \( \hat{\mu}_2^{(N)}(x) = \hat{\mu}^{(N)}(x) + \Delta^{(N)}(x) \).

Let \( \mu_0 \) be the true common value of \( \mu_1 \) and \( \mu_2 \) and \( \sigma_0 \) be the true value of \( \sigma \) under \( H_0 \), and denote the normal density function with mean \( \mu_0 \) and variance \( \sigma_0^2 \) by \( n(y; \mu_0, \sigma_0) \). Under \( \lambda = \lambda_0 = (\mu_0, \mu_0, \sigma) \), the densities of \((Y, G)\) and \( Y \) are

\[ f_{Y,G}(y; g; \lambda_0, x) = n(y; \mu_0, \sigma_0)p(g) \quad \text{and} \quad f_Y(y; \lambda_0, x) = n(y; \mu_0, \sigma_0), \]

where neither depend on \( x \) under \( H_0 \).

From now on, all probabilities \( P_0(\cdot) \), expectations \( E_0(\cdot) \), variances \( Var_0(\cdot) \), and all probability statements such as “converges in probability”, “converges almost surely (a.s.)”, and “with probability one” are under \( \lambda = \lambda_0 \).

**Lemma 1 (Information Matrix)** The information matrix \( I(x) \) is given by

\[ I(x) = -E_0 \left( \frac{\partial^2}{\partial \gamma \partial x} \right) \begin{pmatrix} \frac{\partial^2}{\partial \gamma \partial \mu} & \frac{\partial^2}{\partial \gamma \partial \sigma} \\ \frac{\partial^2}{\partial \gamma \partial \Delta} & \frac{\partial^2}{\partial \gamma \partial \sigma} \end{pmatrix} \right) \begin{pmatrix} \gamma(x) & 0 \\ 0 & 1 \end{pmatrix}, \quad (34) \]

where \( \gamma(x) = (1 - 2\theta_1)^2(1 - r) + (1 - 2\theta_2)^2r \) (recall that \( \theta_1 \) and \( \theta_2 \) depend on \( x \); see (2)).

A sequence of random variables \( X_N(\eta, x) \) is almost surely (a.s.) uniformly bounded with respect to \( \eta \in \Gamma = \{(\Delta, \mu, \sigma) : |\mu - \Delta| \leq \xi, |\mu + \Delta| < \xi, 0 < \sigma_1 \leq \sigma \leq \sigma_2 < \infty \} \) and \( x \in [0, D] \) if for any \( \epsilon > 0 \), there exists a constant \( K \) such that

\[ P_0 \left( \sup_{\eta \in \Gamma} \right) \left( \sup_{\eta \in \Gamma} \left( \sup_{x \in [0, D]} N \right) \sup_{N \geq 1, 2, \ldots} \left( X_N(\eta, x) \right) \right) \leq K > 1 - \epsilon. \]
A sequence of random variables $X_N(x)$ converges to a random variable $Y(x)$ almost surely and uniformly for $x \in [0, D]$ if

$$\lim_{N \to \infty} \sup_{x \in [0, D]} |X_N(x) - Y(x)| = 0 \text{ a.s.}$$

Two sequences of random variables $X_N(s)$ and $Y_N(x)$ are called asymptotically equivalent, denoted by $X_N(x) \equiv Y_N(x)$, if $X_N(x) - Y_N(x)$ converges to zero a.s. and uniformly for $x \in [0, D]$.

**Lemma 2 (Higher Order Terms)**

(A) All elements of the $3 \times 3 \times 3$ matrix

$$\frac{1}{N} \frac{\partial^3}{\partial \eta^3} \ell_N(\eta, x)$$

are a.s. bounded uniformly for $\eta \in \Gamma$ and $x \in [0, D]$.

(B)

$$\sqrt{N} \left[ I(x) \right]^{\frac{1}{2}} (\hat{\eta}(N)(x) - \eta_0) \equiv \frac{1}{\sqrt{N}} \left[ I(x) \right]^{-\frac{1}{2}} \frac{\partial}{\partial \eta} \ell_N(\eta, x) \bigg|_{\eta=\eta_0}.$$  

**A.3.1 Proof of Theorem 1:**

It is sufficient to show that

$$2\{\ell_N(\hat{\eta}(N)(x), x) - \ell_N(\hat{\eta}_0(N))\} \equiv U^2(x).$$

Recall from Section 3 that, under $H_0$, we can write $U(x)$ as

$$U(x) = \frac{u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)}{\sqrt{\text{Var}_0[u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)]}}.$$

From (6) it follows that,

$$\frac{u(\mu_0, \sigma_0, x) - u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)}{\sqrt{\text{Var}_0[u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)]}} = \frac{1}{\sigma_0} \left( \frac{\hat{\mu}_0(N) - \mu_0}{\sqrt{\text{Var}_0[u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)]}} + \left( \frac{1}{\sigma_0^2} - \frac{1}{(\hat{\sigma}_0(N))^2} \right) \frac{u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)}{\sqrt{\text{Var}_0[u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)]}} \right) \equiv 0,$$

and hence

$$U(x) \equiv \frac{u(\mu_0, \sigma_0, x)}{\sqrt{\text{Var}_0[u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)]}}.$$
Next, using the facts (following from (3)) that
\[
\frac{1}{N} \text{Var}_0[u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)] = \left( \frac{1}{\sigma_0} \right)^2 \gamma(x) \text{ and } \frac{\partial}{\partial \Delta} \ell_N(\eta, x) \bigg|_{\eta=\eta_0} = u(\mu_0, \sigma_0, x),
\]
we can write
\[
U(x) = \frac{u(\mu_0, \sigma_0, x)}{\sqrt{\text{Var}_0[u(\hat{\mu}_0(N), \hat{\sigma}_0(N), x)]}} \frac{\partial}{\partial \Delta} \ell_N(\eta, x) \bigg|_{\eta=\eta_0} = \frac{\sqrt{N}}{\sigma_0} \Delta^{(N)}(x) \sqrt{\gamma(x)},
\]
where the last equality follows from Lemma 2(B).

Next we consider the likelihood ratio statistic. Using a Taylor expansion, together with Lemma 2 (A), we can write
\[
2[\ell_N(\eta^{(N)}(x), x) - \ell_N(\eta_0)] = N(\eta^{(N)}(x) - \eta_0)^T I(x)(\eta^{(N)}(x) - \eta_0),
\]
and, under \(H_0\), since there is no dependence on \(x\), we can use classical results to write
\[
2[\ell_N(\eta_0^{(N)}) - \ell_N(\eta_0)] = \frac{N}{\sigma_0^2} (\hat{\nu}_0^{(N)} - \nu_0)^T \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} (\hat{\nu}_0^{(N)} - \nu_0),
\]
where \(\nu = (\mu, \sigma)^T\), \(\nu_0 = (\mu_0, \sigma_0)^T\), and \(\hat{\nu}_0^{(N)} = (\hat{\mu}_0^{(N)}, \hat{\sigma}_0^{(N)})^T\).

Using the derivatives of the log likelihood (see the proof of Lemma 1 we have
\[
\sqrt{N}(\hat{\nu}_0^{(N)} - \nu_0) \approx \frac{\sigma_0^2}{\sqrt{N}} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \frac{\partial}{\partial \nu} \ell_N(\eta, x) \bigg|_{\eta=\eta_0} \approx \sqrt{N} \begin{pmatrix} \hat{\mu}^{(N)}(x) \\ \hat{\sigma}^{(N)}(x) \end{pmatrix} - \nu_0,
\]
where the last equality follows from Lemma 2 (B). The likelihood ratio statistic is obtained by subtracting (39) from (38), which results in
\[
2[\ell(\mathbf{y}^{(N)}; \eta^{(N)}(x), x) - \ell(\mathbf{y}^{(N)}; \eta_0^{(N)})] = \frac{N}{\sigma_0^2} (\Delta^{(N)}(x))^2 \gamma(x)
\]
\[\approx U^2(x),\]
where the last equality follows from (37). The proof of Theorem 1 is now complete.

**A.3.2 Proof of Lemma 1**

The proof of Lemma 1 is quite straightforward, and we only sketch it here. Evaluation of the derivatives of the log likelihood is long and tedious, and we
omit the details. However, for completeness, we give the derivatives evaluated
at the null parameter.

\[
\begin{align*}
\frac{\partial \ell_N}{\partial \Delta} \bigg|_{\eta = \eta_0} &= \frac{1}{\sigma_0^2} \left[ - (1 - 2\theta_1) \sum_{i=1}^{N_1} (y_i - \mu_0) - (1 - 2\theta_2) \sum_{N_1+1}^{N_2} (y_i - \mu_0) 
+ (1 - 2\theta_1) \sum_{N_2+1}^{N_3} (y_i - \mu_0) + (1 - 2\theta_1) \sum_{N_3+1}^{N} (y_i - \mu_0) \right], \\
\frac{\partial \ell_N}{\partial \mu} \bigg|_{\eta = \eta_0} &= \frac{1}{\sigma_0} \sum_{i=1}^{N} (y_i - \mu_0), \\
\frac{\partial \ell_N}{\partial \sigma} \bigg|_{\eta = \eta_0} &= -\frac{N}{\sigma_0} + \frac{1}{\sigma_0^3} \sum_{i=1}^{N} (y_i - \mu_0)^2, \\
\frac{\partial^2 \ell_N}{\partial \Delta^2} \bigg|_{\eta = \eta_0} &= -\frac{N}{\sigma_0^2} \sum_{i=1}^{N} (y_i - \mu_0)^2 
- (1 - 2\theta_2)^2 \sum_{i=N_1+1}^{N_2} (y_i - \mu_0)^2 
- (1 - 2\theta_2)^2 \sum_{i=N_2+1}^{N_3} (y_i - \mu_0)^2 
+ (1 - 2\theta_1)^2 \sum_{i=N_3+1}^{N} (y_i - \mu_0)^2, \\
\frac{\partial^2 \ell_N}{\partial \mu^2} \bigg|_{\eta = \eta_0} &= -\frac{N}{\sigma_0^2}, \\
\frac{\partial^2 \ell_N}{\partial \sigma^2} \bigg|_{\eta = \eta_0} &= -\frac{N}{\sigma_0^2} \sum_{i=1}^{N} (y_i - \mu_0)^2, \\
\frac{\partial^2 \ell_N}{\partial \Delta \partial \mu} \bigg|_{\eta = \eta_0} &= \frac{1}{\sigma_0^2} \left[ (1 - 2\theta_1)(n_1 - n_4) + (1 - 2\theta_2)(n_2 - n_3) \right], \\
\frac{\partial^2 \ell_N}{\partial \Delta \partial \sigma} \bigg|_{\eta = \eta_0} &= \frac{2}{\sigma_0^3} \left[ (1 - 2\theta_1) \sum_{i=1}^{N_1} (y_i - \mu_0) + (1 - 2\theta_2) \sum_{N_1+1}^{N_2} (y_i - \mu_0) 
- (1 - 2\theta_2) \sum_{N_2+1}^{N_3} (y_i - \mu_0) - (1 - 2\theta_1) \sum_{N_3+1}^{N} (y_i - \mu_0) \right], \\
\frac{\partial^2 \ell_N}{\partial \mu \partial \sigma} \bigg|_{\eta = \eta_0} &= \frac{2}{\sigma_0^3} \sum_{i=1}^{N} (y_i - \mu_0). 
\end{align*}
\]
A.3.3 Proof of Lemma 2

The proof of Lemma 2 relies on two other results, which we state here and prove in the next sections.

Lemma 3 (Boundedness of MLE) (A) For any \( x \in [0, D] \), the mle \( \hat{\lambda}^{(N)}(x) \) exists with probability one.

(B) For any \( \delta > 0 \), there exist a measurable subset \( A \) of the sample space with \( P(A) \geq 1 - \delta \), real numbers \( \sigma_1, \sigma_2 \) satisfying \( 0 < \sigma_1 < \sigma_0 < \sigma_2 < \infty \), a real number \( \xi > 0 \) and an integer \( M > 0 \), such that for any sample point in \( A \), \( x \in [0, D] \), and \( N > M \), we have

\[
\left| \hat{\mu}^{(N)}_{j}(x) \right| \leq \xi, \quad j = 1, 2, \quad \text{and} \quad \sigma_1 \leq \left| \hat{\sigma}^{(N)}(x) \right| \leq \sigma_2,
\]

where \( \delta, A, \sigma_1, \sigma_2, \xi \), and \( M \) are independent of \( x \).

Lemma 4 (Uniform convergence of the MLE) The mles \( \hat{\mu}^{(N)}_{1}(x), \hat{\mu}^{(N)}_{2}(x), \) and \( \hat{\sigma}^{(N)}(x) \) converge uniformly to \( \mu_0, \mu_0 \), and \( \sigma_0 \), respectively, a.s. and uniformly for \( x \in [0, D] \).

Since the second order moments of the elements in (35) are bounded uniformly for \( \eta \in \Gamma \) and \( x \in [0, D] \), the conclusion in (A) follows from Chebyshev’s inequality, which also implies that

\[
\frac{1}{\sqrt{N}} \left. \frac{\partial}{\partial \eta} \ell_N(\eta, x) \right|_{\eta=\eta_0}
\]

is a.s. bounded uniformly for \( x \in [0, D] \), since the elements in (40) have mean zero and variances bounded uniformly for \( x \in [0, D] \). Applying Chebyshev’s inequality once more, we can show that

\[
\frac{1}{N} \left. \frac{\partial^2}{\partial \eta^2} \ell_N(\eta, x) \right|_{\eta=\eta_0} \equiv I(x),
\]

since the right-hand side of (41) is the mean of the left-hand side and the variances of the elements of the left-hand side converge to zero uniformly for \( x \in [0, D] \).

From the Taylor expansion

\[
\frac{\partial}{\partial \eta} \ell_N(\hat{\eta}^{(N)}(x), x) - \left. \frac{\partial}{\partial \eta} \ell_N(\eta, x) \right|_{\eta=\eta_0}
\]

\[
= \left. \frac{\partial^2}{\partial \eta^2} \ell_N(\eta, x) \right|_{\eta=\eta_0} (\hat{\eta}^{(N)}(x) - \eta_0) + \frac{1}{2} \left. \frac{\partial^3}{\partial \eta^3} \ell_N(\eta^{*}(x), x) \right|_{\eta=\eta_0} (\hat{\eta}^{(N)}(x) - \eta_0)
\]

\[-\eta_0, \eta^{(N)}(x) - \eta_0],
\]

26
where \( y^* = \eta_0 + \zeta(y^{(N)}(x) - \eta_0) \) with \( 0 < \zeta < 1 \), we have
\[
\sqrt{N} \left[ I(x) \right]^{\frac{1}{2}} \left( \hat{\eta}(N)(x) - \eta_0 \right) - \frac{1}{\sqrt{N}} \left[ I(x) \right]^{-\frac{1}{2}} \frac{\partial}{\partial \eta} \ell_N(\eta, x) \bigg|_{\eta=\eta_0} \\
= \left\{ I(x) \left[ \frac{1}{\sqrt{N}} \frac{\partial}{\partial \eta} \ell_N(\eta, x) \bigg|_{\eta=\eta_0} \right] + \frac{1}{2} \frac{\partial}{\partial \eta} \ell_N \left( y^*(x), x \right)(\eta^*(x)(x) - \eta_0) \right\}^{-1} - I \\
\cdot \frac{1}{\sqrt{N}} I^{-\frac{1}{2}}(x) \frac{\partial}{\partial \eta} \ell_N(\eta, x) \bigg|_{\eta=\eta_0},
\]
(42)
where \( I \) is the \( 3 \times 3 \) identity matrix. The argument in the first paragraph of this proof and Lemmas 3 and 4 imply that the right-hand side of (42) is asymptotically equivalent to 0.

A.3.4 Proof of Lemma 3

Under \( H_0 \), the mle’s of \( \mu_1 = \mu_2 = \mu_0 \) and \( \sigma^2 = \sigma_0^2 \) are the sample mean and sample variance, respectively, and let \( \lambda_0^{(N)} = (\hat{\mu}_0^{(N)}, \hat{\sigma}_0^{(N)}, \hat{\sigma}_0^{(N)}) \). Clearly,
\[
\sup_{\lambda} f(y^{(N)}; \lambda, x) \geq f(y^{(N)}; \lambda_0^{(N)}, x) = \left( \frac{1}{\sqrt{2\pi\hat{\sigma}_0^{(N)}}} \right)^N e^{-\frac{1}{2} \frac{y^{(N)} - \lambda_0^{(N)}}{\hat{\sigma}_0^{(N)}}^2}.
\]
(43)
For any positive number \( b \), the variance of \( Y \) conditional on \(|Y| \leq b\) is denoted by
\[
\sigma_b^2 = \text{Var}_0(Y \mid |Y| \leq b),
\]
which converges to \( \sigma_0^2 = \text{Var}_0(Y) \) as \( b \to \infty \). For any \( \epsilon > 0 \), select a positive number \( b \) such that
\[
P_0(|Y| \leq b) > 1 - \epsilon \text{ and } \sigma_b > (1 - \epsilon)\sigma_0.
\]
The sample mean and sample variance using \( y_i \)’s with \(|y_i| \leq b\) are denoted by \( \bar{y}_b^{(N)} \) and \((\hat{\sigma}_b^{(N)})^2\), respectively. We have
\[
\lim_{N \to \infty} \hat{\sigma}_b^{(N)} = \sigma_b > (1 - \epsilon)\sigma_0.
\]
(44)
Now choose intervals \( I_1 = [-2, -1], I_2 = [0, 1], I_3 = [2, 3], \) and \( I_4 = [-b, b] \). The choice of these intervals is somewhat arbitrary, we are only using \( I_1 - I_3 \) to get a lower bound on \( \sigma^2 \), and \( I_4 \) to get a bound on \( \mu \).

Let \( W_j^{(N)} \) be the number of \( Y_1, \ldots, Y_N \) in \( I_j \), \( j = 1, 2, 3, 4 \). Denote by \( n^{(b)}_j \) the number of \( Y_1, \ldots, Y_N \) in \([-b, b] \) and in group \( j, j = 1, 2, 3, 4 \). Clearly,
\[
W_4^{(N)} = \sum_{j=1}^{4} n^{(b)}_j.
\]
Following from the fact that
\[
\lim_{N \to \infty} \frac{W_j^{(N)}}{N} = P_0(Y \in I_j), \quad j = 1, 2, 3, 4,
\]
\[
\lim_{N \to \infty} \hat{\sigma}_0^{(N)} = \sigma_0,
\]
and the limits in (3) and (44), there exist a measurable subset \(A\) of the sample space with \(P_0(A) > 1 - \delta\), a positive number \(\alpha\), and an integer \(M\) such that for any sample point in \(A\) and \(N > M\),
\[
W_j^{(N)} > \alpha N, \quad j = 1, 2, 3, 4,
\]
\[
\hat{\sigma}_0^{(N)} < \frac{3}{2} \sigma_0,
\]
\[
n_j > (1 - \epsilon) \frac{1}{2} (1 - r) N, \quad j = 1, 4, \quad (45)
\]
\[
n_j > (1 - \epsilon) \frac{1}{2} r N, \quad j = 2, 3,
\]
\[
\hat{\sigma}_b^{(N)} > (1 - \epsilon) \sigma_0.
\]

Hereafter in the proof of Lemma 3, we assume that the sample point is in \(A\) and \(N > M\) and then, (45) hold. We then have,
\[
\sup_{\lambda} f(y^{(N)}; \lambda, x) \geq \left( \frac{1}{\sqrt{2\pi} \hat{\sigma}_0^{(N)}} \right)^N e^{-\frac{N}{2}} > \left( \frac{2e^{-\frac{1}{2}}}{3\sqrt{2\pi} \sigma_0} \right)^N. \quad (46)
\]

Now, because of the gaps between \(I_1 - I_3\), for any real values of \(\mu_1\) and \(\mu_2\), there exists an interval, \(I^*\), among \(I_1, I_2\) and \(I_3\) such that \(\inf \left\{ |y - \mu_j| : j = 1, 2, y \in I^* \right\} \geq \frac{1}{2}\). This fact, together with the inequality
\[
f(y^{(N)}; \lambda, x) \leq \left( \frac{1}{\sqrt{2\pi} \sigma} \right)^N \prod_{y_i \in I_1 \cup I_2 \cup I_3} \exp \left\{ -\frac{1}{2\sigma^2} \min \left[(y_i - \mu_1)^2, (y_i - \mu_2)^2\right]\right\} \quad (47)
\]
and (45), imply that
\[
f(y^{(N)}; \lambda, x) \leq \left( \frac{1}{\sqrt{2\pi} \sigma} \right)^N \left[ \exp \left( -\frac{1}{2\sigma^2} \cdot \frac{1}{4} \right) \right]^\alpha N = \left( \frac{e^{-\frac{\alpha}{8\sigma^2}}}{\sqrt{2\pi} \sigma} \right)^N. \quad (48)
\]

As \(\sigma \to 0\) or \(\sigma \to \infty\), the quantity in parentheses converges to zero and then (48) < right-hand side of (46). As a result, the supremum of \(f(\cdot)\) can be
achieved in the area of \(0 < \sigma_1 \leq \sigma \leq \sigma_2 < \infty\), where \(\sigma_1\) and \(\sigma_2\) are two constants independent of \(x\). Furthermore, for any \(y_i \in I_1 \cup I_2 \cup I_3\), the limit
\[
\lim_{|\mu_1|, |\mu_2| \to \infty} \min \left( (y_i - \mu_1)^2, (y_i - \mu_2)^2 \right) = \infty,
\]
implies that the right-hand side of (47) < right-hand side of (46) for a bounded \(\sigma\) as both \(|\mu_1|\) and \(|\mu_2|\) converge to \(\infty\). Therefore, we have shown that the supremum of \(f(\cdot)\) can be achieved in the area where \(\sigma_1 \leq \sigma \leq \sigma_2\) and at least one of \(\mu_1\) and \(\mu_2\) is bounded.

Now we show that for \(|\mu_1| \leq b\) and as \(|\mu_2| \to \infty\), the value of the likelihood function is smaller than the right-hand side of (46). Following from the inequalities (45), we have
\[
n_j^{(b)} > n_j - \epsilon N > (1 - \epsilon) \frac{1}{2} (1 - r)N - \epsilon N \geq (1 - k\epsilon) \frac{1}{2} (1 - r)N, \quad j = 1, 4,
\]
and
\[
n_j^{(b)} > n_j - \epsilon N > (1 - \epsilon) \frac{1}{2} rN - \epsilon N \geq (1 - k\epsilon) \frac{1}{2} rN, \quad j = 2, 3,
\]
where
\[
k = \max \left( \frac{3 - r}{1 - r}, \frac{2 + r}{r} \right).
\]
Consequently,
\[
\lim_{|\mu_2| \to \infty} f(y^{(N)}, \lambda, x) \
\leq \lim_{|\mu_2| \to \infty} \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^N \prod_{i=1}^{N_2} \varphi(y_i; \lambda, \theta_1) \prod_{i=N_2+1}^{N} \varphi(y_i; \lambda, \theta_2) \times \prod_{i=N_2+1}^{N_3} \varphi(y_i; \lambda, 1 - \theta_2) \prod_{i=N_3+1}^{N_4} \varphi(y_i; \lambda, 1 - \theta_1) \times \prod_{i=1}^{N} \exp \left[ -\frac{1}{2\sigma^2}(y_i - \mu_1)^2 \right] \
\leq \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^N \left( 1 - \theta_1 \right)^{n_1(b)} \left( 1 - \theta_2 \right)^{n_2(b)} \theta_2^{\eta_2(b)} \theta_1^{\eta_1(b)} \prod_{i=1}^{N} \exp \left[ -\frac{1}{2\sigma^2}(y_i - \mu_1)^2 \right] \
\leq \left( \frac{1}{\sqrt{2\pi \tilde{\sigma}^2}} \right)^N \left[ (1 - \theta_1) \theta_1^{(1 - k\epsilon)\frac{1}{2}(1 - r)N} [(1 - \theta_2) \theta_2^{(1 - k\epsilon)\frac{1}{2} rN} e^{-\epsilon N} \right] \
\leq \left( \frac{e^{-\epsilon \sigma_0}}{2^{1-k\epsilon} \sqrt{2\pi(1 - \epsilon)^2 \sigma_0}} \right)^N.
\]
Note that to obtain the inequality in (49), the mle’s \( \overline{y}^{(N)}_b \) and \( \hat{\sigma}^{(N)}_b \) have been used. For small \( \epsilon \), clearly (50) < the right-hand side of (46). Symmetrically, for \( |\mu_2| \leq b \) and as \( |\mu_1| \to \infty \), the value of the likelihood function is also smaller than the right-hand side of (46). Therefore, we have shown that the supremum of \( f(\cdot) \) is achieved in the area of \( \mu_1, \mu_2, \) and \( \sigma \) bounded uniformly with respect to \( x \in [0, D] \). The existence of mle’s of \( \mu_1, \mu_2, \) and \( \sigma \) for any \( x \in [0, D] \) follows from the continuity of \( f(\cdot) \) with respect to \( \lambda \).

### A.3.5 Proof of Lemma 4

Before we prove Lemma 4 we need a few preliminaries.

For any function \( h(\lambda, x) \) of \( \lambda \) and \( x \), any subset \( \Lambda \) of the parameter space of \( \lambda \), and subset \( \Phi \) of \([0, D]\), define

\[
h(\Lambda, \Phi) = \sup_{\lambda \in \Lambda, \ x \in \Phi} h(\lambda, x).
\]

For any \( \epsilon > 0 \), \( \lambda^* = (\mu_1^*, \mu_2^*, \sigma^*) \), and \( x^* \in [0, D] \), define the \( \epsilon \)-neighborhood of \( \lambda^* \) by

\[
S(\lambda^*, \epsilon) = \{ \lambda = (\mu_1, \mu_2, \sigma) : |\mu_1 - \mu_1^*| < \epsilon, |\mu_2 - \mu_2^*| < \epsilon, |\sigma - \sigma^*| < \epsilon \}
\]

and the \( \epsilon \)-neighborhood of \( x^* \) by

\[
S(x^*, \epsilon) = \{ x \in [0, D] : |x - x^*| < \epsilon \}.
\]

As \( \epsilon \downarrow 0 \), we have

\[
f_{Y|G=1}(y; S(\lambda, \epsilon), S(x, \epsilon)) \downarrow f_{Y|G=1}(y; \lambda, x)
\]

and

\[
f_{Y,G}(y, g; S(\lambda, \epsilon), S(x, \epsilon)) \downarrow f_{Y,G}(y, g; \lambda, x).
\]

From Rao (Section 6.6, page 59), or an application of Jensen’s inequality, for \( \lambda \neq \lambda_0 \),

\[
E_0 \left[ \log \left( \frac{f_{Y,G}(Y, G; \lambda, x)}{f_{Y,G}(Y, G; \lambda_0)} \right) \right] < 0.
\]

Now we prove Lemma 4.

Let \( \delta, M, \sigma_1, \sigma_2, \xi \) and \( A \) be as in Lemma 3. It is sufficient to show that for any \( \epsilon > 0 \), there exists a \( M_1 \) such that when \( N > M_1 \),

\[
\lambda^{(N)}(x) \in S(\lambda_0, \epsilon), \quad x \in [0, D],
\]

with probability \( 1 - 2\delta \). Define \( \Gamma = \{ \lambda : |\mu_1| \leq \xi, \sigma_1 \leq \sigma \leq \sigma_2 \} \) and observe that

\[
\left| \log \left( \frac{f_{Y,G}(y, g; \Gamma, [0, D])}{f_{Y,G}(y, g; \lambda_0)} \right) \right| \leq K|y| + Q,
\]

(52)
where \( K \) and \( Q \) are positive constants depending on \( \xi, \sigma_1, \) and \( \sigma_2 \). Let \( B = \Gamma - S(\lambda_0, \epsilon) \) and note that for any \( \lambda \in B \) and \( x \in [0, D] \), (51), (52) and the Lebesgue Dominated Convergence Theorem imply that

\[
\lim_{\epsilon \downarrow 0} E_0 \left[ \log \left( \frac{f_{Y,G}(Y,G;S(\lambda, \epsilon), S(x, \epsilon))}{f_{Y,G}(Y,G;\lambda_0)} \right) \right] = E_0 \left[ \log \left( \frac{f_{Y,G}(Y,G;\lambda, x)}{f_{Y,G}(Y,G;\lambda_0, x)} \right) \right] < C(\lambda, x) < 0.
\]

Therefore, there exists an \( \epsilon = \epsilon(\lambda, x) \) such that

\[
E_0 \left[ \log \left( \frac{f_{Y,G}(Y,G;S(\lambda, \epsilon), S(x, \epsilon))}{f_{Y,G}(Y,G;\lambda_0)} \right) \right] < C(\lambda, x) < 0.
\]

The preceding argument applies to each \((\lambda, x) \in B \times [0, D]\). Since \( B \times [0, D] \) is compact, there exist points \((\lambda_1, x_1), \ldots, (\lambda_k, x_k)\) and corresponding \( \epsilon_1, \ldots, \epsilon_k \) and \( C_1, \ldots, C_k \) such that

\[
B \times [0, D] \subset \bigcup_{n=1}^k S(\lambda_i, \epsilon) \times S(x_i, \epsilon)
\]

and for \( i = 1, \ldots, k, \)

\[
E_0 \left[ \log \left( \frac{f_{Y,G}(Y,G;S(\lambda_i, \epsilon_i), S(x_i, \epsilon_i))}{f_{Y,G}(Y,G;\lambda_0)} \right) \right] < C_i < 0. \quad (53)
\]

Let \( C = \max(C_1, \ldots, C_k) \). Because

\[
\frac{1}{N} \log \left( \frac{L_N(S(\lambda_i, \epsilon_i), S(x_i, \epsilon_i))}{L_N(\lambda_0)} \right) \quad (54)
\]

converges a.s. to the left-hand side of (53), there exist a measurable subset \( A_1 \subset A \) with \( P_0(A_1) > 1 - 2\delta \) and an integer \( M_1 > M \) such that when \( N > M_1, \)

\[
(54) < C < 0, \quad \text{for } i = 1, \ldots, k.
\]

Now, for any \( \lambda \in B \) and any \( x \in [0, D] \), there exists an \( i, 1 \leq i \leq k \), such that \((\lambda, x) \in S(x_i, \epsilon_i) \times S(x_i, \epsilon_i)\) and for any sample point in \( A_1 \) and \( N > M_1, \)

\[
\frac{1}{N} [\ell_N(\lambda, x) - \ell_N(\lambda_0)] = \frac{1}{N} \log \left( \frac{L_N(\lambda, x)}{L_N(\lambda_0)} \right) < \frac{1}{N} \log \left( \frac{L_N(S(\lambda_i, \epsilon_i), S(x_i, \epsilon_i))}{L_N(\lambda_0)} \right) < C < 0,
\]

which implies that \( \hat{\lambda}^{(N)}(x) \notin B \), for all \( x \in [0, D] \). By Lemma 3,

\[
\hat{\lambda}^{(N)}(x) \in \Gamma, \quad \text{for all } x \in [0, D].
\]

31
Thus, for any sample point in $A_1$ and $N > M_1$,

$$\hat{\lambda}^{(N)}(x) \in S(\lambda_0, \epsilon), \text{ for all } x \in [0, D],$$

which concludes the proof.
References


Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64** 247-254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74** 33-43.


Figure 1: Score test for a simulated data set with 250 backcross progeny. As in Lander and Botstein (1989), we assumed that markers are spaced 20 cM throughout 12 chromosomes of 1 Morgan each. The QTLs, located at genetic positions 70, 49, 27, 8 and 3 cM from the left end on the first five chromosomes, were assumed to have effects 1.5, 1.25, 1.0, 0.75 and 0.50, respectively. For each individual, genotypes at markers were generated assuming no interference. The corresponding quantitative trait was simulated by summing the five QTL effects and random standard normal noise. The solid line is for score test statistics and the dotted line is for likelihood ratio test statistics (-2*log(LR)). The dashed line indicates the threshold of 12.83 at the 0.05 level based on the simulated asymptotic null distribution. All five QTLs attained this threshold.
Figure 2: The value of the test statistics along chromosome 17, which has 16 markers. The solid line is for the score test, the dotted line is for the likelihood ratio test ($-2 \log \text{LR}$), and the dashed line indicates the threshold of 10.10 at the 0.05 level based on the simulated asymptotic null distribution.