Fine-Scale Mapping of Quantitative Trait Loci Using Historical Recombinations

Momiao Xiong and Sun-Wei Guo

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029

Manuscript received August 26, 1996
Accepted for publication December 2, 1996

ABSTRACT

With increasing popularity of QTL mapping in economically important animals and experimental species, the need for statistical methodology for fine-scale QTL mapping becomes increasingly urgent. The ability to disentangle several linked QTL depends on the number of recombination events. An obvious approach to increase the recombination events is to increase sample size, but this approach is often constrained by resources. Moreover, increasing the sample size beyond a certain point will not further reduce the length of confidence interval for QTL map locations. The alternative approach is to use *historical* recombinations. We use analytical methods to examine the properties of fine QTL mapping using historical recombinations that are accumulated through repeated intercrossing from an F_2 population. We demonstrate that, using the historical recombinations, both simple and multiple regression models can reduce significantly the lengths of support intervals for estimated QTL map locations and the variances of estimated QTL *map locations*. We also demonstrate that, while the simple regression model using historical recombinations does not reduce the variances of the estimated additive and dominant effects, the multiple regression model does. We further determine the power and threshold values for both the simple and multiple regression models. In addition, we calculate the Kullback-Leibler distance and Fisher information for the simple regression model, in the hope to further understand the advantages and disadvantages of using historical recombinations relative to F_2 data.

VIRTUALLY every organ and function of any species in nature exhibits continuous variations. It has been well documented that many traits that vary continuously are determined by a number of loci (called quantitative trait loci, or QTL), each with small effect, and working in concert with environmental factors.

The mapping of QTL is important not only in identifying genes' underlying traits of interest in economically important species but also in gaining new insight into gene mapping and identification, and the structure and function of the human genome. Indeed, the progress of gene mapping, identification, and characterization in humans often depends on the development and study of suitable animal models. For example, the successful mapping of new susceptibility loci for type I diabetes in mouse certainly shed new light into the study of human type I diabetes (TODD 1989; RISCH *et al.* 1993; DAVIES *et al.* 1994; HASHIMOTO *et al.* 1994).

The basic idea of mapping QTL has been known for over 30 years since THODAY's seminal paper (THODAY 1961). The idea was simple enough: if genetic markers are scattered throughout the genome of an organism of interest, the segregation of these markers can be used to detect and estimate the effects of linked QTL, making possible the mapping and characterization of underlying QTL (TANKSLEY 1993).

Simple as it may be, however, putting the idea into practice had been difficult, primarily due to the lack of appropriate genetic maps. With the rapid development of genetic maps based on DNA markers in the last decade, coupled with the explosive development of statistical methods (WELLER 1986; JENSEN 1989; LANDER and BOTSTEIN 1989; KNAPP *et al.* 1990; HALEY and KNOTT 1992; DARVASI *et al.* 1993; ZENG 1993, 1994; HALEY *et al.* 1994), our ability to map QTL has been greatly enhanced.

In general, the mapping and characterization of QTL consists of two different but closely related problems: the localization of QTL and the estimation of their effects on the trait value. Clearly, the larger effect a QTL has on the trait value, the easier it can be mapped. Conversely, if all QTL locations were known, it would be relatively easier to estimate their individual and joint effects. In practice, however, neither location nor effect of individual QTL is known. In fact, one does not even know how many contributing QTL there are underlying the trait of interest. Relatively speaking, the estimation of individual QTL effect is much more difficult than localization, since a precise estimate of the genetic effect for a specific locus depends not only on the mode of genetic interaction between the locus of interest and other QTL, but also on the specific environment in which the organism lives and on possible gene-environment interactions.

A point that does not seem to be well appreciated is that the localization, or mapping, or positioning, of QTL has two levels: one is low-resolution localization,

Corresponding author: Sun-Wei Guo, Division of Epidemiology, University of Minnesota, 1300 S. Second St., Suite 300, Minneapolis, MN 55454-1015. E-mail: guo_s@epivax.epi.umn.edu

or coarse-scale mapping, with a resolution of ~1–5 cM or over (depending on species, of course), and the other high-resolution localization, or fine-scale mapping, with a resolution of <1–5 cM. Low-resolution localization of QTL may only have limited usefulness in identification and characterization of QTL, for two reasons. First, if several QTL are clustered in a small chromosomal region, coarse-scale mapping will not be able to distinguish them and to identify each QTL individually, even if these QTL are correctly mapped to a region. This would be of little use in selective breeding based on flanking markers and would cause enormous difficulties in estimating individual QTL effect. Second, despite rapid advances in DNA sequencing technology, the cloning of QTL will still take years of hard work if these QTL are not further zeroed-in to narrow chromosomal regions.

As more and more empirical results have demonstrated (see below), and as we will show later in this paper, these two different levels of mapping require entirely different mapping strategies, experimental designs, and analytical methods. Most statistical methods developed in the past 7 or 8 years for mapping QTL in experimental species are designed for coarse-scale mapping.

The interval mapping method (LANDER and BOTSTEIN 1989) has been shown to be a powerful tool for mapping QTL. The method uses flanking markers to detect any QTL lying in the interval flanked by the two markers. Compared with methods using only single markers, interval mapping is more powerful and can provide much more accurate estimates of QTL effect and position when QTL are unlinked, and is relatively robust (KNOTT and HALEY 1992). However, it is still difficult for the interval mapping method to allow simultaneous analysis of several linked QTL, and to distinguish multiple linked QTL effects. When two or more QTL are located on the same chromosome region, they may be mapped to wrong positions by interval mapping (KNOTT and HALEY 1992; MARTINEZ and CURNOW 1992; WRIGHT 1994).

To circumvent these problems, several authors proposed to map QTL by linear regression models (JANSEN 1993; RODOLPHE and LEFORT 1993; ZENG 1993, 1994; HALEY *et al.* 1994). These authors demonstrated that, using multiple markers, one can detect effects of QTL and distinguish multiple QTL using both the flanking markers and the markers in other regions. This approach is sensible, because quantitative traits are unlikely to be controlled by a single QTL, and because use of multiple markers in different regions of chromosomes would help one detect multiple QTL. While one can still use the interval mapping method to search multiple QTL simultaneously, the heavy computational burden makes this approach impractical. It is also difficult to establish proper threshold values for declaring the existence of QTL. In contrast, the multiple regression method is computationally feasible, although an optimal mapping strategy does not exist yet.

One notion is that, with a dense genetic map, one can finely map QTL. Unfortunately, however, this is not quite true. A dense map is necessary for fine-mapping of QTL, but it is not sufficient. A key, limiting factor is the number of recombination events. One obvious way to ensure enough number of recombinations is to increase the sample size. However, besides practical constraints on resource and time, this approach has several drawbacks. RODOLPHE and LEFORT (1993) pointed out that the variances of the estimated additive and dominant QTL effects by the multiple regression model increases with the density of the markers typed. This will increase the chance of error in statistical inference. Therefore, the density of the genetic map cannot be too high. Furthermore, even if one has infinite number of markers, DARVASI *et al.* (1993) showed that a QTL with a moderate effect can only be assigned to a map location in a rather broad chromosome region due to the lack of sufficient recombinant events. HYNE *et al.* (1995) also reported that the estimates of QTL locations are unreliable even with a large sample from an F_2 population. Thus, unless one has enough number of recombinations, an overly dense map would be a waste.

What, then, can we increase the number of recombinations without typing huge number of subjects? As high-density genetic maps with highly polymorphic markers are increasingly becoming available for experimental species (see, for example, DIETRICH *et al.* 1996), and as more and more genes are mapped at a coarse scale, this issue is becoming increasingly urgent in mapping and identifying QTL. Without addressing this issue, it would be hard to imagine that one can take full advantage of a dense map.

One alternative approach, which has not been appreciated very much until recently, is to use *historical* recombinations. The haplotypes of any individual in the current population is a result of recombinations of different genotypes from different ancestors, if we trace his lineage far enough. In other words, given enough time, and barring strong selection, an ancestor's haplotype will be eventually break up, no matter how close the two loci are. Therefore, historical recombination events, if accumulated enough, will provide ample opportunities to observe recombinations between any two linked loci, and thus can be used for fine-scale mapping purpose. This phenomenon, the decay of linkage disequilibrium, was first noted by JENNINGS (1917) and ROBBINS (1918), and later studied extensively by LEWONTIN and KOJIMA (1960). It was the basis for BODMER (1986), apparently the first person, to argue for the use of linkage disequilibrium for fine-scale mapping in humans.

PATERSON *et al.* (1990) proposed a fine-mapping method in which recombinant individuals are identified in primary generations and selectively multipled in

subsequent generations so that the recombinant classes occur at near equal frequency with the nonrecombinant ones. In the ideal situation, a series of nearly isogenic lines, differing in recombination in the QTL regions, would be compared for the quantitative trait of interest, allowing, potentially, the high-resolution localization of QTL. CHURCHILL *et al.* (1993) discussed the use of DNA pooling in high-resolution mapping. Recently, DARVASI and SOLLER (1995) proposed a fine-mapping method based on what they called "advanced intercross line", or AIL. An AIL is produced from an F_2 population resulting from crossing two inbred lines assumed homozygous for different alleles at all loci. The subsequent generations, F_3, F_4, ..., are sequentially produced by randomly intercrossing the previous generation. For mapping purposes, only individuals in the last generation (F_t) are phenotyped and genotyped. As long as sizes of breeding individuals in F_s ($s \leq t$) generations are >100, and as long as there is no strong selection, DARVASI and SOLLER (1995) convincingly demonstrated, by simulation, that a simple regression model using data on an F_t population can significantly reduce the length of the support interval for estimated QTL map location.

Interesting as they are, the results of DARVASI and SOLLER (1995) actually prompt many more questions than they have answered regarding the use of AIL for fine-scale mapping. Because of limitations in simulation studies, is it possible to demonstrate *analytically* the advantages and disadvantages of using F_t data? How to determine the threshold for the corresponding test statistics? How to determine the power of the test? What is the relationship between the generation t and power? What is the relationship between the generation t and the threshold value? Is there any difference between data from F_2 and F_t ($t > 2$)? Since DARVASI and SOLLER (1995) only considered a *simple* regression model, can one generalize their results to a *multiple* regression model? And how can one determine the power and the threshold value? These questions are not only of theoretical importance but also of practical importance.

In this paper, we will address these issues. Using an asymptotical analysis, we demonstrate that both simple and multiple regression models can reduce significantly the lengths of support intervals for estimated QTL map locations and the variances of estimated QTL *map locations* using F_t data. We also demonstrate that, while the simple regression model using data from an F_t population does not reduce the variances of the estimated additive and dominant effects, the multiple regression model does. We further determine the power and threshold values for both the simple and multiple regression models. In addition, we calculate the Kullback-Leibler distance and Fisher information for the simple regression model, in the hope to further understand the advantages and disadvantages of using F_t data relative to F_2 data.

Due to the technical nature of this paper, our treatment is unavoidably very mathematical. Less mathematically inclined readers can skip derivations and proofs and read the part on the statement of the problem and our conclusions. For excellent discussions on the use of AIL design, the readers should consult DARVASI and SOLLER (1995).

A GENETIC MODEL FOR AN F_t POPULATION

The haplotype frequencies of an experimental population change over time due to various evolutionary forces. Barring mutations and selections, the change, on average, is a function of two variables: the recombination fraction between two linked loci and the number of generations. DARVASI and SOLLER (1995) derived a formula for calculating the expectation of the frequency of the recombinant haplotype for an F_t population. Since in calculating the Fisher information and the Kullback-Leibler distance, not only the expectation but also the higher moments of the frequencies of the recombinant haplotypes are needed, we derive the generator of a diffusion process that approximates a stochastic process that describes the evolution of change in haplotype frequencies.

Consider two loci, A and B, each with two alleles (A_i and B_j, $i, j = 1, 2$). The recombination fraction between the two loci is assumed to be θ. Let $P_{ij}(t)$ denote the frequency of the gamete A_iB_j ($i, j = 1, 2$) and $N(t)$ denote the size of the F_t population. Denote $r_t = P_{12}(t) + P_{21}(t)$. We can show that the population process {$X(t)$ = number of recombinant haplotypes in the t generation} evolves as a Markov chain that can be approximated by a diffusion process with the following generator:

$$L = \frac{1}{2}\frac{r_t(1 - r_t)}{2N(t)}\frac{\partial^2}{\partial r_t^2} + \left(\frac{\theta}{2} - \theta r_t\right)\frac{\partial}{\partial r_t}.$$

Now, let $f(r_t) = r_t$. Then, by the Hille-Yosida theorem (ETHIER and KURTZ 1986), we obtain (see APPENDIX A)

$$\frac{dE[r_t]}{dt} = E[L(r_t)] = -\theta E[r_t] + \frac{\theta}{2}.$$

Solving it for $E[r_t]$ yields

$$E[r_t] = \tfrac{1}{2}(1 - e^{-\theta t}). \quad (1)$$

If θ is small, $E[r_t] \approx \theta t/2$, which agrees with DARVASI and SOLLER (1995).

THE CASE OF SIMPLE REGRESSION MODELS

Assume that there is no epistasis and no interaction between the environment and QTL. The simple linear regression model for QTL mapping is

$$y_i = \mu + \alpha x_i + \delta z_i + e_i, \quad i = 1, 2, \ldots, n, \quad (2)$$

where n is the size of the sample taken randomly from the F_t population, y_i is the trait value of the ith individual

in the sample, μ is the overall population mean, α and δ are additive and dominance effects, respectively, e_i's are independently and identically distributed random variables with $E[e_i] = 0$ and $\text{Var}(e_i) = \sigma^2$, and $x_i(d)$ and $z_i(d)$ are dummy variables for the ith individual with the following values

$$x_i(d) = \begin{cases} 1 & M_i = AA \\ -1 & M_i = aa \\ 0 & M_i = Aa \end{cases}$$

$$z_i(d) = \begin{cases} 1 & M_i = Aa \\ -1 & \text{otherwise,} \end{cases}$$

where A and a are two alleles of the marker M_i at d.

Since the vectors $x_i(d)$ and $z_i(d)$ are asymptotically orthogonal, we can estimate the additive and the dominance effects by (DUPUIS 1994)

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$\hat{\delta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})z_i}{\sum_{i=1}^{n}(z_i - \bar{z})^2},$$

respectively.

The estimated additive and dominance effects of QTL: The use of historical recombinations increases the recombinant events, which, in turn, effectively increase the genetic distance between the markers and QTL. As a result, the estimated additive and dominance effects associated with markers will be reduced. Here, we evaluate the amount of reduction, as a function of t, in the estimated additive and dominance effects associated with markers.

Assume that there are k QTL with kth QTL having additive effect α_k and dominance effect δ_k along the genome. Denote the genetic distance between the marker M and the kth QTL by Δ_k. Then, we can show (APPENDIX B) that, asymptotically,

$$\hat{\alpha}(t) \xrightarrow{a.s} \sum_{k=1}^{K} \alpha_k e^{-t\Delta_k} \tag{3}$$

and

$$\hat{\delta}(t) \xrightarrow{a.s} \sum_{k=1}^{K} \delta_k e^{-2t\Delta_k}. \tag{4}$$

To evaluate how much amount of additive and dominant effects at particular markers are reduced, we consider, for simplicity, the case of $k = 1$. Equations 3 and 4 can then be written as

$$\hat{\alpha} \xrightarrow{a.s} \alpha e^{-t\Delta}$$

and

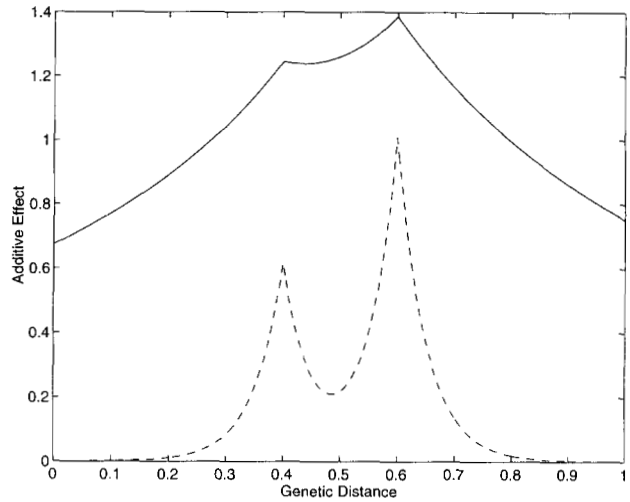

FIGURE 1.—The effect of t on the asymptotic additive effects at markers with the simple linear regression model. Two QTL, with the additive effects 0.6 and 1.0, respectively ($\sigma_r^2 = 1$), are located at 0.4 and 0.6 cM from the left end of the chromosome. —, the asymptotic additive effects using F_2 data; ---, the asymptotic additive effects using F_t data.

$$\hat{\delta} \xrightarrow{a.s} \delta e^{-2t\Delta},$$

where Δ is the genetic distance between the marker M and the QTL. Clearly, the additive and dominance effects associated with the markers decrease exponentially with the generation t.

When several QTL with comparable effects are closely linked, it may be difficult to separate them and may even map them to wrong positions. The above results indicate that at any given marker locus linked with the QTL, the estimated genetic (additive and dominant) effects decrease with t. Furthermore, the rate of decrease in the estimated genetic effects is exponential. This implies that, as one moves away from the QTL locus, the estimated genetic effects decrease exponentially, which, in turn, implies that the estimated genetic effects due to linked QTL can be separated/disentangled provided t is large enough. Figure 1 illustrates this point graphically. It can be seen that two linked QTL are hardly separated if F_2 data are used. However, the two loci can be distinguished very well if F_{10} data are used.

The thresholds of the test: To implement the fine-scale mapping of QTL using F_t data, it is critical to determine the threshold for a given significance level, so that one can reject or accept the null hypothesis H_0:$\alpha \neq 0$ or H_0:$\delta \neq 0$ depending on whether or not the statistic exceeds the threshold. Note that we simply cannot use the thresholds of the test for simple regression models based on F_2 data because some changes of the threshold are needed. In this subsection, we give procedures for computing the thresholds.

Under the assumption that e_i's are normally distributed with mean 0 and known variance σ_r^2, the log likeli-

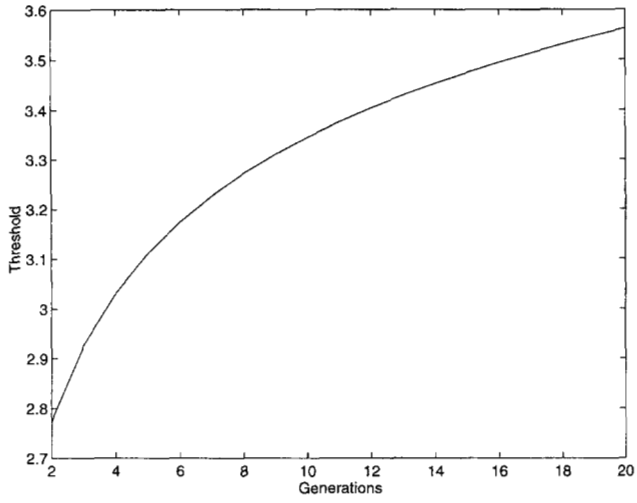

FIGURE 2.—Thresholds of test statistic X_d at a significance level of 5% as a function of generation t.

hood ratio for testing H_0:$\alpha = 0$ *vs.* H_1:$\alpha \neq 0$ for the presence of a QTL at d and known $x_i(d)$ is

$$\left[\frac{\sqrt{n}\hat{\alpha}(d)}{\sqrt{2}\sigma_e}\right]^2.$$

When d is unknown, the log-likelihood ratio statistics becomes

$$\max_d\left\{\frac{\sqrt{n}\hat{\alpha}(d)}{\sqrt{2}\sigma_e}\right\}^2,$$

where the maximum is taken over all loci d where $x_i(d)$ is known. The variable $x_i(d)$ is known when a marker is available at d.

Now let

$$X_d = \sqrt{n}\,\frac{\hat{\alpha}(d)}{\sqrt{2}\sigma_e}.$$

Along the same line as that of LANDER and BOTSTEIN (1989), we can show that, under the null hypothesis of H_0:$\alpha = 0$, X_d is a Gaussian process with mean 0 and covariance function $R(u) = e^{-t|u|}$ as $n \to \infty$. Note that for F_2 data, $R(u) = e^{-2|u|}$. This limiting distribution holds even when the e_i's are not normally distributed (DUPUIS 1994). Therefore, X_d is still an Ornstein-Uhlenbeck process.

Using the results of FEINGOLD *et al.* (1993), we have

$$P_0(\max_d X_d > b) \approx 1 - \Phi(b) + tlb\phi(b), \qquad (5)$$

where l is the length of the chromosome, $\phi(x)$ and $\Phi(x)$ are the density and cumulative functions of the standard normal distribution, respectively. As an example, we calculated the threshold as a function of t for the test statistic X_d at 5% level with $l = 100$ cM. The results are shown in Figure 2.

Thus, we can see that with a fixed significance level α, increasing t will decrease $1 - \Phi(b)$ and hence increase the threshold b. This can be explained intuitively. The recombination is a measure of genetic distance between the two linked loci. Increasing t corresponds effectively to increasing the number of recombination events, which in turn, is equivalent to increasing the genetic distances, or the length of the genome. This implies that we would search a QTL in a "longer" genome. Therefore, to maintain the same significance level α of the test as in the F_2 case, we need to increase the threshold of the test for F_t data.

When the genetic map is not dense, *i.e.*, markers are not available at some locations, it is usually assumed that $x_i(d)$ is known at equispaced distances of Δ cM. For the case where the $x_i(d)$ are only known at equispaced distance of Δ cM, (5) becomes

$$P_0(\max_k X_{k\Delta} > b) \approx 1 - \Phi(b) + tlb\phi(b)\nu(b\sqrt{2t\Delta}), \qquad (6)$$

where $\nu(x) \approx e^{-0.583x}$ (FEINGOLD *et al.* 1993). Note that this equation is equivalent to (5) when $\Delta = 0$. Here, the function $\nu(x)$ is a discreteness correction factor to account for the fact that we are computing the likelihood ratio statistic at discrete points on the chromosome instead of continuously as is the case for a dense map.

Similarly, when d is unknown, the log-likelihood ratio statistic for testing the dominant effect is

$$\max_d Z_d^2, \qquad (7)$$

where

$$Z_d = \frac{\sqrt{n}\hat{\delta}(d)}{2\sigma_e}.$$

The threshold of the test is determined by

$$P_0\{\max_d Z_d > b\} \approx 1 - \Phi(b) + 2tlb\phi(b), \qquad (8)$$

if the genetic map is dense, or

$$P_0\{\max_d Z_{k\Delta} > b\} \approx 1 - \Phi(b) + 2tlb\phi(b)\nu(b\sqrt{4t\Delta}), \qquad (9)$$

if the map is equispaced map with the distances of Δ cM.

Next we address the issue of testing for the presence of either additive or dominance effect, which amounts to a general hypothesis H_0: $\alpha = \delta = 0$ *vs.* H_1:$\alpha \neq 0$ or $\delta \neq 0$. The corresponding log-likelihood ratio is

$$\max_d\left[\left(\frac{\sqrt{n}\hat{\alpha}(d)}{\sqrt{2}\sigma_e}\right)^2 + \left(\frac{\sqrt{n}\hat{\delta}(d)}{2\sigma_e}\right)^2\right].$$

Thus, as $n \to \infty$, X_d and Z_d are Gaussian processes with mean 0 and covariance function $e^{-t|u|}$ and $e^{-2t|u|}$, respectively. Moreover, X_d and Z_d are asymptotically independent (DUPUIS 1994).

Using the results of DUPUIS (1994), we can determine the threshold of the test by solving the following inequality for b:

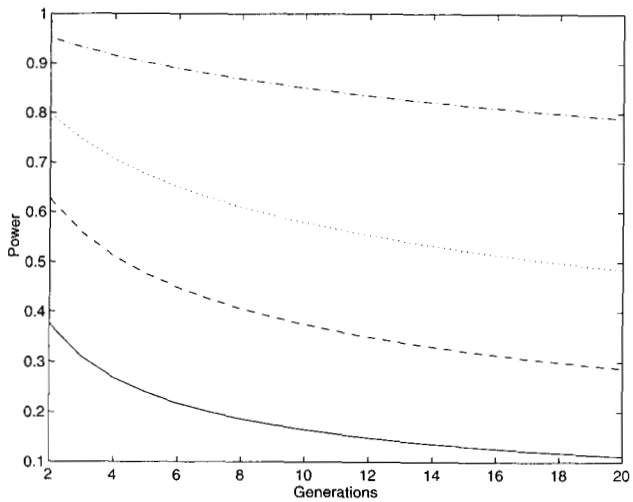

FIGURE 3.—Power of test with a significance level of 5% and an additive effect $a = 0.25$ (σ_e^2 is set to be 1). —, ---, · · ·, – · – · represent $n = 100$, 200, 300, and 500, respectively.

$$P_0\{\max_k [X^2(k\Delta) + Z^2(k\Delta)] \geq b^2\} \approx e^{-1/2b^2} + \tfrac{3}{2}b^2 t l \nu(b\sqrt{3t\Delta})e^{-1/2b^2} \leq \alpha. \quad (10)$$

The power of the test: The power of a test is the probability of detecting the effect of the QTL when it exists. To give an approximation to the power of the test, we need to calculate $E[X_d]$ under H_1:$\alpha \neq 0$. Assume that there is a QTL, we can show in Appendix C that

$$E[X_d] \approx \xi e^{-t|u|}, \quad (11)$$

where $\xi = \sqrt{(n/2)}\alpha/\sigma_e$ and $|u|$ is the distance between the marker and the QTL.

Again, using the results of FEINGOLD *et al.* (1993), we obtain the following approximations:

(1) for a dense map

$$P_{d,\xi}(\max_d X_d > b) \approx 1 - \Phi(b - \xi) + \phi(b - \xi)[2\xi^{-1} - (b + \xi)^{-1}], \quad (12)$$

(2) for an equispaced map

$$P_{d,\xi}(\max_k X_{k\Delta} > b) \approx 1 - \Phi(b - \xi) + \phi(b - \xi)[2\xi^{-1}\nu - (b + \xi)^{-1}\nu^2], \quad (13)$$

where $\nu = \nu(b\sqrt{2t\Delta})$.

Although the formula for calculating the power of the test is the same in form for any F_t population, the power of the test in F_t population decreases with t because increasing t means increasing the threshold as we have shown in the previous section.

Figure 3 shows the power of the test statistic X_d for a significance level of 0.05 and an additive effect $\alpha = 0.25$. Note that the power of the test in general decreases with increasing t, but the decreases in power for different sample sizes are different. The powers of the test for $n = 100$, 200, 300, and 400 based on F_{20} are 30%, 46%, 82%, and 98% as those based on F_2, respectively. Thus, the reduction in power with increasing t is smaller for larger sample size.

Support intervals for QTL locations: The construction of a support interval for the QTL location on the chromosome is a useful way to assess the uncertainty in QTL localization. It also helps to narrow down the search of QTL to a small chromosomal region. Compared with that in F_2 population, the length of support interval of mapping QTL in F_t population will be dramatically reduced.

To illustrate this, we use a lod-support interval. An α-lod support interval includes all the loci s such that

$$lod(s) \geq \max_d lod(d) - \alpha, \quad (14)$$

where $lod(s)$ is the base-10 logarithm of the likelihood ratio statistic at the locus s.

From the previous section, we know that the asymptotic log-likelihood ratio for testing H_0: $\alpha = \delta = 0$ *vs.* H_1:$\alpha \neq 0$ or $\delta \neq 0$ in the presence of a QTL at d is

$$LR(d) = X_d^2 + Z_d^2.$$

The lod score can be written as

$$lod(d) = \tfrac{1}{2}(\log 10e)LR(d). \quad (15)$$

Assuming there are k QTL, we know from the previous section that, provided n is larger enough,

$$lod(d) \approx \tfrac{1}{4}(\log 10e)\frac{n}{\sigma_e^2}\left\{\left[\sum_{k=1}^{K} \alpha_k e^{-t|d-d_k|}\right]^2 + \tfrac{1}{2}\left[\sum_{k=1}^{K} \delta_k e^{-2t|d-d_k|}\right]^2\right\}. \quad (16)$$

As we have shown in the previous section, the estimated genetic effects $\hat{\alpha}(d)$ and $\hat{\delta}(d)$ decrease exponentially with t as well as the distance between the marker and the QTL. As a result, the length of support interval of the QTL position will be reduced exponentially as t increases. Figure 4 illustrates the expected length of the support interval using F_t data. It can be seen that the expected lengths of the support interval decreases exponentially. The lengths of support interval using F_2 data are reduced by fivefold if F_{10} data are used. However, there is a diminished return: only a further 1.8-fold reduction is achieved after another 10 generations. This agrees with what DARVASI and SOLLER (1995) found in their simulations.

Kullback-Leibler (KL) distance and Fisher information: To further characterize the gain in fine-scale mapping of QTL using the historical recombinations, we compute the KL distance between the probability distributions with true and estimated QTL locations. We also compute the Fisher information for the likelihood ratio.

The KL distance measures the mean information for discrimination in favor of one hypothesis, say, A, against

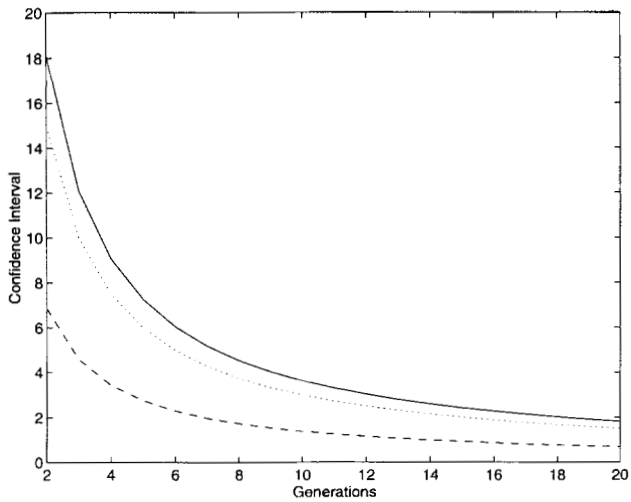

FIGURE 4.—Expected lengths of the support intervals of the estimated QTL locations using F_t data. —, the case of $n = 100$, $\alpha = 0.5$, $\delta = 0.25$; ---, the case of $n = 1000$, $\alpha = 0.25$, $\delta = 0.1$; · · ·, the case of $n = 500$, $\alpha = 0.25$, $\delta = 0.1$. In all cases, σ_e^2 is set to be 1.

another when A is true (KULLBACK 1983a). Let γ_n be an estimated QTL map location and γ^* the true QTL map location. We can show (APPENDIX D) that, for a dense map, the KL distance between the likelihood functions $L(\gamma_n)$ and $L(\gamma^*)$, defined as $K_t(L(\gamma^*), L(\gamma_n)) = E_{\gamma^*}(\log[L(\gamma^*)/L(\gamma_n)])$, is given by

$$K_t(L(\gamma^*), L(\gamma_n)) = \frac{n}{2\sigma^2}\{2E[\theta_{\gamma_n\gamma^*}](4\delta^2 + \alpha^2) - 8E[\theta^2_{\gamma_n\gamma^*}]\delta^2\}, \quad (17)$$

where $\theta_{\gamma_n\gamma^*}$ is the frequency of the recombinant haplotypes between the locations γ_n and γ^*.

If $\delta = 0$, *i.e.*, there is no dominant effect,

$$K_t(L(\gamma^*), L(\gamma_n)) = \frac{n}{\sigma^2}E[\theta_{\gamma_n\gamma^*}]\alpha^2 = \frac{n}{2\sigma^2}(1 - e^{-\theta_n t})\alpha^2 \approx \frac{t}{2}\frac{n}{\sigma^2}\theta_n\alpha^2 \approx \frac{t}{2}K_2(L(\gamma^*), L(\gamma_n)), \quad (18)$$

where θ_n is the recombination fraction between γ_n and γ^*.

Thus, the *KL* distance $K_t(L(\gamma^*), L(\gamma_n))$ for F_t data is approximately $t/2$ times greater than that for F_2 data. In other words, the information for discriminating γ^* against γ_n is increased by $t/2$-fold.

KONG and WRIGHT (1994) showed that above result implies that

$$\frac{L(\gamma_n)}{L(\gamma^*)} \xrightarrow{p} 0$$

as $n|\gamma_n - \gamma^*| \to \infty$ with the rate proportional to the KL distance. Hence, with a large enough sample size and a dense map, the likelihood function will be concentrated in intervals encompassing the true locations QTL γ^*, with widths of the intervals in the order of $2/nt$ in F_t population, which is $t/2$ times narrower than that in F_2 population. This further confirms that the widths of support intervals of the QTL location in F_t population would be reduced by $t/2$ times.

If $\delta \neq 0$, we can show that (APPENDIX E)

$$E[\theta^2_{\gamma_n\gamma^*}] = be^{-\int_0^t \lambda(\tau)\,d\tau}\int_0^t h(\eta)\,d\eta, \quad (19)$$

where

$$\lambda(t) = 2\theta_n + \frac{1}{2N(t)},$$

$$b = {}^1/_2\left(\theta_n + \frac{1}{2N(t)}\right),$$

$$h(\eta) = e^{\int_0^\eta \lambda(\tau)\,d\tau}(1 - e^{-\theta_n\eta}).$$

Thus, $K_t(L(\gamma^*), L(\gamma_n))$ can be evaluated in principle by substituting $E[\theta_{\gamma_n\gamma^*}] = {}^1/_2[1 - e^{-\theta_n t}]$ and (19) into (17). Of course, the expression for $K_t(L(\gamma^*), L(\gamma_n))$ will become more complicated.

The generalized Fisher information of the likelihood ratio measures the amount of information supplied by the data about the unknown parameter, *e.g.*, the QTL location γ^* (KULLBACK 1983b). However, the typical formula of Fisher information requires that the log-likelihood function be differentiable. Here, however, the log-likelihood ratio function is a function of the distance between γ_n and γ^*. Therefore, this log-likelihood ratio function is not differentiable at the true QTL map location γ^*. Thus, the formula for calculating the Fisher information cannot simply be applied to our case. KONG (A. KONG, personal communication) modified the formula of Fisher information to allow a nondifferentiable log-likelihood function at the true parameter as follows.

Suppose that a random variable X is distributed with density $p(x, \theta)$ and δ is an estimator of $g(\theta)$. Furthermore, assume that

$$\left|\frac{\partial}{\partial\theta}\log p(x, \theta)\Big|_{\theta_-}\right| = \left|\frac{\partial}{\partial\theta}\log p(x, \theta)\Big|_{\theta_+}\right|.$$

Here, we can show that (APPENDIX F), in our case,

$$\mathrm{Var}(\delta) \geq \frac{[g'(\theta)]^2}{E_\theta\left[\frac{\partial}{\partial\theta}\log p(x, \theta)\,|_{\theta_-}\right]^2}. \quad (20)$$

The Fisher information $I(\theta)$ is defined as

$$I(\theta) = E_\theta\left[\frac{\partial}{\partial\theta}\log p(x, \theta)\Big|_{\theta_-}\right]^2. \quad (21)$$

Now, applying the definition (21) to our problem, we can define the Fisher information $I_t(\gamma^*)$ for measur-

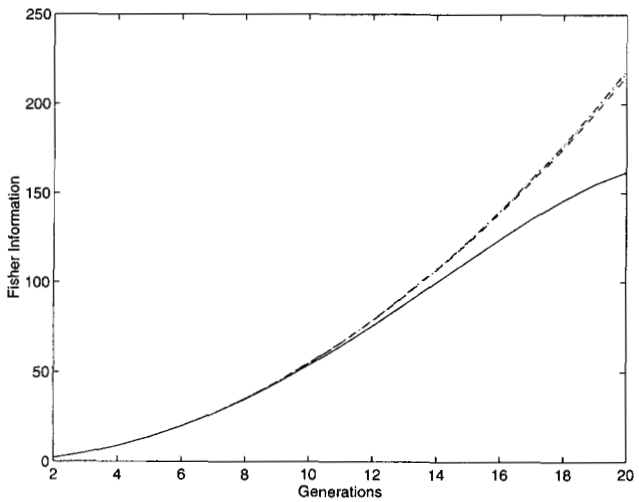

FIGURE 5.—The Fisher information as a function of population size. —, ---, · · ·, –·–· are Fisher informations for n = 100, 500, 1000 and 10000, respectively.

ing the difficulty of estimating the true QTL location γ^* in F_t population as

$$I_t(\gamma^*) = E_{\gamma^*}\left[\frac{\partial}{\partial\gamma}\log\frac{L(\gamma^*)}{L(\gamma)}\bigg|_{\gamma^*}\right]^2.$$

For the ease of exposition, we use the first-order of approximate to the likelihood ratio. In APPENDIX G, we show that

$$\begin{aligned} I_t(\gamma^*) \approx{}& \frac{(a-b)^2}{8\sigma^4}\left[2(2\delta-\alpha)^4 + 2(2\delta+\alpha)^4\right. \\ &\left. + 4\delta(2\delta-\alpha)^2(2\delta+\alpha) + 2(4\delta^2-\alpha^2)^2\right] \\ &+ \frac{b(a-b)\alpha^2}{2\sigma^4}\left[(2\delta-\alpha)^2 + 3(2\delta+\alpha)^2\right] \\ &+ \frac{b^2\alpha^4}{2\sigma^4} + \frac{(a-b)^2}{2\sigma^2}(8\delta^2 + 5\alpha^2 - 8\alpha\delta) \\ &+ \frac{b(a-b)}{\sigma^2}\alpha(\alpha+2\delta), \end{aligned}$$

where

$$a = -\frac{t}{2},$$

$$b = -\frac{t}{2}\frac{1}{4N(t)}\, e^{-\int_0^t d\tau/2N(\tau)} \int_0^t \eta e^{1/2\int_0^\eta d\tau/N(\tau)}\, d\eta.$$

If we assume $N(t) = N$, then $b \approx -(t^3/16N)$. To see how the Fisher information increases as generation t increases, we calculated the Fisher information for $\alpha = \delta = 0.5$, and $\sigma^2 = 1$ (Figure 5). We can see that the Fisher information increases as the generation t increases, which implies that the observed data sampled from an F_t population contain more information about the true QTL location than that in F_2 population. It also can be seen from the figure that, for $t \leq 12$, the Fisher information for data sampled from populations with different sizes (100, 500, 1000, and 10000) are practically identical. After 12 generations, the Fisher information for data sampled from populations of >500 are indistinguishable. However, the difference in Fisher information increases as t increases for $N = 100$ and $N = 500$. These observations suggest that for $t \leq 12$, an effective population size of 100 should be enough. This seems to agree with DARVASI and SOLLER (1995). For $t > 12$, however, a population size of >100 is recommended. This is because that, while a population of 100 individuals may be large enough to avoid genetic fixation for $t \leq 12$, it may not be large enough if further intercrossing is required.

EXTENSION TO THE MULTIPLE REGRESSION MODEL

The previous analysis can be extended to the multiple regression model for fine-scale mapping of QTL using F_t data. Assume n individuals are sampled at random from an F_t population. For the ith individual ($i = 1, \ldots, n$), denote the corresponding quantitative trait value as Y_i, and marker genotype (codominant) at jth locus as $M_i(j)$ ($j = 1, \ldots, m$). Assuming no epistasis and no interaction between environment and QTL, the multiple regression can be written as follows:

$$Y_i = \mu + \sum_{j=1}^{m}\begin{Bmatrix} +\alpha_j - \delta_j & M_i(j) = AA \\ \delta_j & M_i(j) = AB \\ -\alpha_j - \delta_j & M_i(j) = BB \end{Bmatrix} + \epsilon_i \quad \text{for an } F_t, \qquad (22)$$

where α_j and δ_j are additive and dominant effects associated with the jth marker, ϵ_i's are independent and identically distributed random variables with $E[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.

Let X_i be a row vector containing the coefficient of the parameters μ, α and δ in (22) and $X = [X_1^T, \ldots, X_N^T]^T$. In matrix notations, (22) can be rewritten as

$$Y = X\beta + \varepsilon, \quad E[\varepsilon] = 0, \quad V(\varepsilon) = \sigma^2 I, \qquad (23)$$

where $\beta = [\mu, \alpha_1, \ldots, \alpha_m, \delta_1, \ldots, \delta_m]$. It is a classical result that the least square estimate of β:

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

By the strong law of large numbers (RODOLPHE and LEFORT 1993),

$$\frac{1}{n}X^TX = \frac{1}{n}\sum_{i=1}^{n} X_i^TX_i \xrightarrow{a.s} E[X_1^TX_1] = U_t$$

where

$$U_t = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & A_1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & A_2 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & A_\nu & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & B_1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & B_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & B_t \end{pmatrix},$$

where A_i and B_i denote the matrices of additive and dominance effects in the ith chromosome respectively. We can show (APPENDIX H) that for F_t data

$$A_i = [\tfrac{1}{2}e^{-t\Delta_{jj'}}] \quad \text{and} \quad B_i = [e^{-2t\Delta_{jj'}}],$$

where $\Delta_{jj'}$ is the genetic distance between the markers j and j'.

It can be seen that the matrix U_t for the F_t data has the same structure as that for the F_2 data. If all genetic distances $\Delta_{jj'}$ increase by $t/2$ fold, the matrix U_2 for the F_2 data become identical to U_t.

Since

$$n\,\mathrm{Var}(\hat{\beta}) = n\,\sigma^2(X^TX)^{-1} \overset{a.s}{\to} \sigma^2U^{-1}$$

by the Central Limit Theorem, we have

$$\sqrt{n}(\hat{\beta}_n - \beta) \sim N(0, \sigma^2U^{-1}).$$

Hence, $\hat{\beta}_n$ is an unbiased and consistent estimator of β.

Asymptotic variance of the estimated genetic effects: As we mentioned before, the matrix U_t has the same structure of the matrix U in F_2 population. Therefore, using similar arguments as that of RODOLPHE and LEFORT (1993), we have

$$\begin{aligned} V(\hat{\alpha}_j(t)) &= \frac{2\sigma^2}{n}\frac{1 - e^{-2t\Delta_{j-1,j+1}}}{(1 - e^{-2t\Delta_{j-1,j}})(1 - e^{-2t\Delta_{j,j+1}})} \\ &\approx \frac{\sigma^2}{n}\frac{1}{t}\frac{\Delta_{j-1,j+1}}{\Delta_{j-1,j}\Delta_{j,j+1}} \\ &= \frac{2}{t}V(\hat{\alpha}_j) \end{aligned}$$

and

$$\begin{aligned} V(\hat{\delta}_j(t)) &= \frac{\sigma^2}{n}\frac{1 - e^{-4t\Delta_{j-1,j+1}}}{(1 - e^{-4t\Delta_{j-1,j}})(1 - e^{-4t\Delta_{j,j+1}})} \\ &\approx \frac{\sigma^2}{4n}\frac{1}{t}\frac{\Delta_{j-1,j+1}}{\Delta_{j-1,j}\Delta_{j,j+1}} \\ &= \frac{2}{t}V(\hat{\delta}_j), \end{aligned}$$

where $V(\hat{\alpha}_j)$ and $V(\hat{\delta}_j)$ are the variances of $\hat{\alpha}_j$ and $\hat{\delta}_j$, estimated by multiple regression using F_2 data (RODOLPHE and LEFORT 1993). It can be seen from the above equation that, as the distance between adjacent markers get smaller, *i.e.*, the marker density increases, both $V(\hat{\alpha}_j)$ and $V(\hat{\delta}_j)$ increase, which makes it difficult to increase the resolution of mapping QTL.

We can see from above that one way to effectively use dense map and retain small variances of estimated genetic effects is to map QTL by F_t data. Indeed, the variances of the estimated additive and dominance effects by multiple regression for F_t data are reduced approximately by $t/2$ times as compared with F_2 data. Thus, the use of F_t data will be more effective in separating individual QTL when they are linked, and provide greater precision of the estimated genetic effects.

Note that for the simple regression, since

$$U_t = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{2}I & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{2}I & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \frac{1}{2}I & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & I & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & I & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & I \end{pmatrix}$$

we have

$$V(\hat{\alpha}_j(t)) = V(\hat{\alpha}_j) \quad \text{and} \quad V(\hat{\delta}_j(t)) = V(\hat{\delta}_j).$$

This suggests that for the simple regression model, increasing t has no effect on variances of the estimated genetic effects.

Effects of QTL: By the same argument as that of RODOLPHE and LEFORT (1993), we can show that using F_t data to map QTL does not destroy the most important property of multiple regression model for mapping QTL: the genetic effects associated with the markers depends on only those QTL that are located on the interval flanked by the two neighboring markers, and is independent of the effects of QTL located on other intervals. Therefore, to calculate the effects of QTL associated with the marker, we need to consider only those QTL located in the interval flanked by neighboring markers.

The increase of recombination events, as a result of the use of historical recombinations, effectively increases genetic distances between the markers and QTL. Therefore, as in the case of the simple regression, the estimated additive and dominance effects due to QTL will be reduced quickly when moving away from QTL. In this subsection, we evaluate the amount of the reduction as a function of t.

Using the same argument as that of RODOLPHE and LEFORT (1993), we can show that (APPENDIX I)

$$\alpha_j(t) \overset{a.s}{\to} \alpha_j(t) = \sum_{k=1}^{K} b_k e^{-t\Delta_{kj}} \frac{1 - e^{-2t\Delta_{kj-1}}}{1 - e^{-2t\Delta_{j-1j}}}$$

and

$$\hat{\delta}_j(t) \xrightarrow{a.s} \delta_j(t) = \sum_{k=1}^{K} c_k e^{-2t\Delta_{kj}} \frac{1 - e^{-4t\Delta_{kj-1}}}{1 - e^{-4t\Delta_{j-1j}}},$$

where K is the number of QTL located between marker $j-1$ and marker j, b_k and c_k are additive and dominance effects of the kth QTL, respectively.

To quantify how much the additive and dominance effects, associated with the markers, in F_t population are reduced, we assume, for simplicity, that only a single QTL is located between markers $j-1$ and j. This assumption holds when the genetic map is dense enough. Then, after some algebra, we have

$$\frac{\alpha_j(t)}{\alpha_j(2)} \approx e^{-(t-2)\Delta_{jq}}$$

and

$$\frac{\delta_j(t)}{\delta_j(2)} \approx e^{-2(t-2)\Delta_{jq}},$$

where Δ_{jq} is the genetic distance between the marker j and the QTL and $\alpha_j(2)$ and $\delta_j(2)$ are the additive and dominance effects estimated from F_2 data. Thus, the estimated genetic effects decrease exponentially as one moves away from QTL, with the dominant effect decreasing faster than the additive effect.

Recall that $\sqrt{n}\hat{\beta} \sim N(\beta, \sigma^2 U^{-1})$, and that variances of the additive and dominance effects based on F_t data decrease linearly with t as compared with those based on F_2 data. Taken together, it is easy to visualize that $\hat{\alpha}_j(t)$ and $\hat{\delta}_j(t)$ will be concentrated on the increasingly narrow region surrounding the QTL. Consequently, the "ghost locus" and other problems associated with using F_2 data in mapping QTL will be alleviated significantly using F_t data.

Support intervals for QTL locations: Suppose that we have estimated the location of QTL, which is flanked by the markers M_i and M_{i+1}. Now we want to consider the confidence interval for the estimated location of QTL q. From classical regression theory (Graybil 1976), it is well-known that the likelihood ratio statistic for testing H_0:$\alpha = \delta = 0$ *vs.* H_1:$\alpha \neq 0$ or $\delta \neq 0$ is given by

$$LR = \frac{Y^T(XX^- - X_0X_0^-)Y}{\sigma^2}, \tag{24}$$

where $XX^- = X(X^TX)^{-1}X^T$ and $X_0 = [11\cdots1]^T$.

The lod score can be written as

$$lod = \tfrac{1}{2}\log_{10} eLR,$$

which can be approximated asymptotically by

$$\frac{n}{2}\sum_{j=1}^{m} u_j\alpha_j + n\sum_{j=1}^{m} v_j\delta_j, \tag{25}$$

where α_j and δ_j are the statistic additive and dominance effects at the marker j defined in the previous section. If Δ_{jk} is the genetic distance between the marker M_j and the kth QTL in the interval flanked by the markers M_j and M_{j+1}, and b_k and c_k are the corresponding additive and dominance effects of the kth QTL, respectively,

$$u_j = \sum_{k=1}^{K_j} b_k e^{-t\Delta_{jk}},$$

$$v_j = \sum_{k=1}^{K_j} c_k e^{-2t\Delta_{jk}},$$

where K_j is the number of QTL in the jth interval.

For simplicity, suppose that we have a dense map and the marker M_d, located at locus d in the i interval $[M_i, M_{i+1}]$, is used to detect the location of QTL. Adding this marker to the set of the original markers, and regressing the phenotypic value Y on this augmented set of the markers yields an approximation to the asymptotic likelihood ratio statistic:

$$LR(d) = \frac{n}{2} u_d\alpha_d + nv_d\delta_d + a,$$

where

$$u_d = \sum_{k=1}^{k_d} b_k e^{-t\Delta_{dk}},$$

$$v_d = \sum_{k=1}^{k_d} c_k e^{-2t\Delta_{dk}},$$

$$a = \frac{n}{2}\sum_{j=1}^{m} u_j\alpha_j + n\sum_{j=1}^{m} v_j\delta_j,$$

and k_d is the number of QTL in the interval flanked by the markers $[M_i, M_d]$.

Recall that an x-lod support interval is an interval including all the loci s such that

$$lod(s) \geq \max_d lod(d) - x. \tag{26}$$

To determine the support interval, we need to calculate $\max_d lod(d)$. For simplicity, assuming that there is only a single QTL, indexed as q, in that interval. Then,

$$\max_d lod(d) = \tfrac{1}{2}\log_{10} e\left(\frac{n}{2} b_q^2 + nc_q^2 + a\right).$$

Substitute this into (26), and find a marker located at d_t such that

$$\frac{1}{2}\log_{10} e\left[\frac{n}{2} b_q^2(1 - e^{-2t\Delta_{d_tq}}g_t) + nc_q^2(1 - e^{-4t\Delta_{d_tq}}h_t)\right] = \alpha,$$

where Δ_{d_tq} denotes the genetic distance between the marker located at d_t and the QTL q or half α-lod support interval in F_t population, Δ_{d_ti} denotes the genetic distance between the marker located at d_t and the marker M_i,

$$g_t = \frac{1 - e^{-2t\Delta_{iq}}}{1 - e^{-2t\Delta_{d_i i}}}$$

and

$$h_t = \frac{1 - e^{-4t\Delta_{iq}}}{1 - e^{-4t\Delta_{d_i i}}}.$$

To see how much the length of the support interval for a QTL map location is reduced by multiple regression model using F_t data, we assume $c_q = 0$, *i.e.*, there is no dominant effect. After some algebra, we have

$$e^{-2t\Delta_{d_1 q} + 4t\Delta_{d_2 q}} = \frac{g_2}{g_t},$$

where $\Delta_{d_1 q}$ and $\Delta_{d_2 q}$ are half of the length of the α-lod support interval for a QTL map location by multiple regression using F_t and F_2 data, respectively. It is easy to see that, for $t > 2$,

$$\frac{g_2}{g_t} > 1.$$

Hence,

$$-2t\Delta_{d_1 q} + 4t\Delta_{d_2 q} > 0.$$

That is

$$\Delta_{d_1 q} < \frac{\Delta_{d_2 q}}{\frac{t}{2}},$$

which means that the length of an α-lod support interval for a QTL map location by multiple regression using F_t data will be reduced by more than $t/2$ fold.

Threshold and power: For a particular chromosomal interval flanked by two markers, the statistic genetic effects estimated through the multiple regression depend only on those QTL located within the interval. It is thus natural to test whether or not there exists a QTL in a given interval. Suppose that we want to test the interval $[M_{j-1}, M_j]$. Similar to the simple regression model, we let

$$X_d = \sqrt{n}\,\frac{\hat{\alpha}(d)}{\sqrt{2}\sigma_e}$$

and

$$Z_d = \sqrt{n}\,\frac{\hat{\delta}(d)}{2\sigma_e},$$

where $\hat{\alpha}(d)$ and $\hat{\delta}(d)$ are associated with a marker located at locus d in the interval $[M_{j-1}, M_j]$ and are estimated by the multiple regression method.

It can be shown that under the null hypothesis of $H_0: b_q = 0$ and $c_q = 0$, X_d and Z_d are asymptotically Gaussian processes with mean 0 and a complicated covariance function, which can be approximated by the function $R(u) = e^{-t|u|}$. Therefore, X_d and Z_d can be approximated by an Ornstein-Uhlenbeck process.

To determine the thresholds of the test, we can use formulae similar to (5), (6), (8), (9), and (10), but with the length of genome replaced by the length of the interval. Consequently, for fixed t, the threshold of the test becomes lower.

Assume that in the interval $[M_{j-1}, M_j]$ there exists only one QTL with additive and dominance effects b_k and c_k, respectively. Under $H_1: b_k \neq 0$ or $c_k \neq 0$, we have

$$E[\hat{\alpha}(d)] \approx b_k e^{-t\Delta_{dk}} \frac{1 - e^{-2t\Delta_{k,j-1}}}{1 - e^{-2t\Delta_{j-1,d}}} \approx b_k \frac{\Delta_{k,j-1}}{\Delta_{j-1,d}} e^{-t\Delta_{dk}}. \quad (27)$$

The coefficient of $e^{-t\Delta_{dk}}$ depends in general on the genetic distance between the marker and the QTL and do not have the form of (11). However, if Δ_{dk} is small, $E[\hat{\alpha}(d)]$ can be approximated by

$$E[\hat{\alpha}(d)] \approx b_k e^{-t\Delta_{dk}}.$$

In this case, letting $\xi = \sqrt{(n/2)}(b_k/\sigma_e)$, (12) and (13) can still be used for calculating the power. Note that ξ is independent of the time and the length of the interval.

As we discussed before, increasing the number of recombinant events is effectively equivalent to increasing the genetic distance between loci, and hence to decreasing the genetic effects of QTL. Therefore, to maintain a prespecified type I error, we need to increase the threshold of the test, which, in turn, will decrease the power of detecting QTL. To increase the power of detecting QTL in the particular interval, we can reduce the length of the interval. If the product tl is kept constant, then we know from (5) that the threshold of the test will remain constant, as will the power.

DISCUSSION

With increasing popularity of QTL mapping in economically important animals and in experimental species, the need for statistical methodology for fine-scale QTL mapping becomes increasingly urgent. An obvious approach is to increase sample size to increase the recombination events. However, this approach is often constrained by resources. Moreover, as shown by HYNE *et al.* (1995), increasing the sample size beyond a certain point will not further reduce the length of confidence interval for QTL map locations.

Intrigued by the work of DAVASI and SOLLER (1995), we have carried out a theoretical analysis of QTL mapping using historical recombinations. We have demonstrated that both simple and multiple regression models can reduce significantly the lengths of support intervals for estimated QTL map locations and the variances of estimated QTL *map locations* using F_t data. We also have demonstrated that, while the simple regression model using data from an F_t population does not reduce the variances of the estimated additive and dominant ef-

fects, the multiple regression model does. To further understand the advantages and disadvantages of using F_t data relative to F_2 data, we have calculated the Kullback-Leibler distance and Fisher information for the simple regression model. In addition, to help implement fine-mapping methods, we have derived the formula to compute the power and threshold values for both the simple and multiple regression models.

The idea behind the work of DAVASI and SOLLER (1995) is to use historical recombinations for fine-scale mapping. This idea has been known in human genetics for about 10 years, although it has recently been shown to be quite successful in locating disease genes in fine scale (HASTBACKA *et al.* 1994). The idea is simple indeed: in the absence of selection, the linkage disequilibrium between any QTL and marker loci, however closely linked they are, will gradually dissipate as the population continues intercrossing. Hence, the evolution of haplotype frequencies in the population reflects the action of recombination through past generations, making it possible to disentangle and finely map closely linked QTL.

It should be pointed out that, although the population model that we considered is for the F_t population, or, in DAVASI and SOLLER's terms, the AIL design, the model and subsequent analysis can be modified easily to allow for other features such as selection or particular design.

The analytical techniques that we used in this paper enable us to discover much more interesting features of fine-scale mapping via historical recombinations than found by simulation studies. We have demonstrated that, since the estimated additive and dominance effects based on F_t data decrease exponentially as one moves away from QTL, the length of the support interval for the estimated QTL map location also decreases exponentially with t and by approximately $t/2$ fold. We also have demonstrated that, for the simple regression model that was investigated by DAVASI and SOLLER (1995), the variance of estimated genetic effects using F_t data is the same as that in F_2 population. In other words, there is no gain in precision in estimation of the genetic effects using the simple regression even F_t data are used. If, however, a multiple regression model is used, that variance will be reduced approximately by $t/2$ times as compared with F_2 data.

By extending the definition of the Fisher information to the case of nondifferentiable likelihood functions, we have showed that, for the simple regression model, the Fisher information evaluated at the true QTL location with F_t data is much larger than that for F_2 data. We also have used Kullback-Leibler distance to measure the rate of convergence of the estimated QTL map location to the true QTL location. We found that that Kullback-Leibler distance between the likelihood function evaluated at estimated QTL map location and at the true QTL location using F_t data is approximately $t/2$ times as large as that using F_2 data. Thus, for the simple regression model, the information on QTL locations using F_t data is much higher than that for F_2 data.

To help implement fine-scale mapping of QTL using F_t data, we have evaluated the thresholds of the test for both simple and multiple regression models. Intuitively, increasing t means more accumulated recombinant events, which is effectively equivalent to increasing the genetic distances, or the length of the genome on which QTL are being searched. Consequently, the thresholds of the test for F_t data should be higher. We have showed that the thresholds for the test of QTL for a genome segment of length l with F_t data are roughly equal to those for a genome segment of length tl with F_2 data. As a result, the power of detecting QTL for a *given* marker will be reduced. This seems to contradict with the common perception that experimental parameter values that increase/decrease the power also tend to decrease/increase the lengths of support (or confidence) intervals for QTL map locations. This apparent contradiction can be easily resolved by noting that here the power refers to the power of the test for whether or not there exists a QTL for a given marker, not the power of testing the QTL locations. In addition, because the genetic effects decrease exponentially as one moves away from the QTL, the length of support interval for QTL location also decreases exponentially, thus increasing the precision of estimation of QTL map location. This point has been made by DAVASI and SOLLER (1995) using a simple argument. Based on our analysis, the point can be illuminated more clearly.

We point out, however, that the use of historical recombinations in fine-mapping QTLs is not without limit. In fact, as shown in Figure 4, there is a diminished return with increasing t in terms of increasing the precision of the estimation of QTL map location. The greatest gain in precision seems to be achieved in the first eight to 10 generations after F_2. This suggests that for $t \geq 10$, the use of F_t data may not be very cost-efficient.

Since increasing the mapping resolution by using historical recombinations will result in decrease in power of detecting QTL, one might want to use a denser genetic map to detect QTL. However, increasing the density of the map will increase the variances of the estimated genetic effects at the markers, which will cause difficulty in increasing mapping resolution. To maintain the power of detecting QTL while improving the accuracy of estimated QTL map location, it may be appropriate to use a two-stage mapping strategy. First, use a sample of moderate size taken from the F_2 population and a map with moderate density to detect QTL. Second, once the existence of QTL in some regions is detected, one can proceeding by fine-scale mapping using F_t data.

Fine-scale mapping of QTL using historical recombinations is particularly applicable to species with a short generation cycle that can be easily reproduced by inter-

crossing and for which inbred lines exist. As DARVASI and SOLLER (1995) pointed out, inbred lines of mice, chicken, corn and cultivars as well as accession lines of most annual selfing plant species are good candidates for F_t population. Furthermore, the time scale required for production of a F_t population applies only to the first time that such a population is produced. In plant species, an F_t, once produced, can be stored in the form of seeds, and constitute a permanent mapping population. In mice, or other animals, once an F_{10}, say, has been produced, it can be maintained by reproducing at a much slower rate. Hence, for a relative small investment, such produced population can be used as resource population for fine mapping.

This research was supported by the National Institutes of Health grants R29-GM-52205 and R01-GM-56515.

LITERATURE CITED

BODMER, W. F., 1986 Human genetics: the molecular challenge. Cold Spring Harbor Symp. Quant. Biol. **LI:** 1–13.

CHURCHILL, G. A., J. J. GIOVANNONI and S. D. TANKSLEY, 1993 Pooled-sampling makes high-resolution mapping practical with DNA markers. Proc. Natl. Acad. Sci. USA **90:** 16–20.

DARVASI, A., and M. SOLLER, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141:** 1199–1207.

DARVASI, A., V. WEINREB, J. MINKE, I. WELLER and M. SOLLER, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134:** 943–951.

DAVIES, J. L., Y. KAWAGUCHI, S. T. BENNETT, J. B. COPEMAN, H. J. CORDELL *et al.*, 1994 A genome-wide search for human type 1 diabetes susceptibility genes. Nature **371:** 130–136.

DIETRICH, W. F., J. MILLER, R. STEEN, M. A. MERCHANT, D. DAMRON-BOLES *et al.*, 1996 A comprehensive genetic map of the mouse genome. Nature **380:** 149–152.

DUPUIS, J. 1994 Statistical problems associated with mapping complex and quantitative traits from genomic mismatch scanning data. Technical Report No. 2. Department of Statistics, Stanford University, Stanford, CA.

ETHIER, S., and T. G. KURTZ, 1986 *Markov Processes: Characterization and Convergence.* Wiley, New York.

FEINGOLD, E., O. P. BROWN and D. SIEGMUND, 1993 Guassian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am. J. Hum. Genet. **53:** 234–251.

GRAYBILL, F. A., 1976 *Theory and Application of the Linear Model.* Wordsworth & Books/Cole Advanced Books & Software, Pacific Grove, CA.

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

HALEY, C. S., S. A. KNOTT and J. M. ELSEN, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics **136:** 1195–1207.

HASHIMOTO, L., C. HABITA, J. P. BERESSL, M. DELEPINE, C. BESSE *et al.*, 1994 Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q. Nature **371:** 161–164.

HÄSTBACKA, J., A. DE LA CHAPELLE, M. M. MAHTANI, G. CLINES, M. P. REEVE-DALY *et al.*, 1994 The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloining by fine-structure linkage disequilibrium mapping. Cell **78:** 1078–1087.

HYNE, V., M. J. KEARSEY, D. J. PIKE and J. W. SNAPE, 1995 QTL analysis-unrealiability and bias in estimation procedures. Mol. Breed. **1:** 273–282.

JANSEN, R. C., 1989 Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker loci. Theor. Appl. Genet. **78:** 613–618.

JANSEN, R. C., 1993 Interval mapping of multiple quantitative traits. Genetics **136:** 205–214.

JENNINGS, H. S., 1917 The numerical results of diverse systems of breedingh special relation to the effects of linkage. Genetics **2:** 97–154.

KARLIN, S., and H. M. TAYLOR, 1981 *A Second Course in Stochastic Processes.* Academic Press, Inc. New York.

KNAPP, S. J., W. C. BRIDGES and D. BIRKES, 1990 Mapping quantitative trait loci using molecular marker linkage map. Theor. Appl. Genet. **79:** 583–592.

KNOTT, S. A., and C. S. HALEY, 1992 Maximum likelihood mapping of quantitative trait loci using full-sib families. Genetics **132:** 1211–1222.

KONG, A., and F. WRIGHT, 1994 Asymptotic theory for gene mapping. Proc. Natl. Acad. Sci. USA. **91:** 9705–9709.

KULLBACK, S., 1983a Fisher information, pp. 115–118 in *Encyclopedia of Statistics,* Vol. 3, edited by S. KOTZ and N. L. JOHNSON. John Wiley, New York.

KULLBACK, S., 1983b Kullback information, pp. 421–425 In *Encyclopedia of Statistics,* Vol. 4, edited by S. KOTZ and N. L. JOHNSON. John Wiley, New York.

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

LANGE, K., L. KUNKEL, J. ALDRIDGE and S. A. LATT, 1985 Accurate and superaccurate gene mapping. Am. J. Hum. Genet. **37:** 853–867.

LEWONTIN, R. C., and K. KOJIMA, 1960 The evolutionary dynamics of complex polymorphisms. Evolution **14:** 450–472.

MARTINEZ, O., and R. N. CURNOW, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85:** 480–488.

PATERSON, A. H., J. W. DEVERNA, B. LANINI and S. D. TANKSLEY, 1990 Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes in an interspecies cross of tomato. Genetics **124:** 735–742.

RISCH, N., S. GHOSH and J. A. TODD, 1993 Statistical evaluation of multiple-locus linkage data in experimental species and its relevance to human studies: application to nonobese diabetic (NOD) mouse and human insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. **53:** 702–714.

ROBBINS, R. B., 1918 Some applications of mathematics to breeding problems. III. Genetics **3:** 375–389.

RODOLPHE, F., and M. LEFORT, 1993 A multi-marker model for detecting chromosomal segments displaying QTL activity. Genetics **134:** 1277–1288.

TANKSLEY, S. D., 1993 Mapping polygenes. Annu. Rev. Genet. **27:** 205–233.

THODAY, J. M., 1961 Location of polygenes. Nature **191:** 368–370.

TODD, J. A., C. MIJOVIC, J. FLETCHER, D. JENKINS, A. R. BRADWELL *et al.*, 1989 Identification of susceptibility loci for insulin-dependent diabetes mellitus by trans-racial gene mapping. Nature **338:** 587–589.

WELLER, J. I., 1986 Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics **42:** 627–640.

WRIGHT, F., 1994 *Asymptotics and Robustness for Genetic Linkage Mapping.* Ph.D. thesis, University of Chicago.

ZENG, Z. B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972–10976.

ZENG, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: W. J. EWENS

APPENDIX A

Assume that the recombination between loci 1 and 2, if any, occurs before reproduction. Denote the frequency of the gamete $A_iB_j (i, j = 1, 2)$ after recombination by $P^*_{ij}(t)$. Then, in general, we have

$$P^*_{ij}(t) = \sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{m=1}^{2}\sum_{n=1}^{2}\theta_{ij,kl,mn}P_{kl}P_{mn}, \quad (28)$$

where

$$\theta_{ij,kl,mn} = \tfrac{1}{2}\delta_{ik}[(1-\theta)\delta_{jl} + \theta\delta_{jn}] + \tfrac{1}{2}\delta_{im}[(1-\theta)\delta_{jn} + \theta\delta_{jl}]$$

and

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

In particular, applying (28) for P^*_{12} and $P^*_{21}(t)$ yields

$$\begin{aligned} P^*_{12}(t) &= \tfrac{1}{2}P^2_{12}(t-1) + \tfrac{1}{2}P_{12}(t-1) + \tfrac{1}{2}P_{11}(t-1)P_{12} \\ &\quad \times (t-1) + \tfrac{1}{2}P_{22}(t-1)P_{12}(t-1) \\ &\quad - \frac{\theta}{2}P_{12}(t-1)P_{21}(t-1) + \frac{\theta}{2}P_{11}(t-1) \\ &\quad \times P_{22}(t-1) + \frac{\theta}{2}P_{11}(t-1)P_{22}(t-1) \\ &\quad + \frac{1-\theta}{2}P_{12}(t-1)P_{21}(t-1), \\ P^*_{21}(t) &= \tfrac{1}{2}P^2_{21}(t-1) + \tfrac{1}{2}P_{21}(t-1) + \tfrac{1}{2}P_{11}(t-1) \\ &\quad \times P_{21}(t-1) + \tfrac{1}{2}P_{22}(t-1)P_{21}(t-1) \\ &\quad - \frac{\theta}{2}P_{21}(t-1)P_{12}(t-1) + \frac{\theta}{2}P_{11}(t-1) \\ &\quad \times P_{22}(t-1) + \frac{\theta}{2}P_{22}(t-1)P_{11}(t-1) \\ &\quad + \frac{1-\theta}{2}P_{12}(t-1)P_{21}(t-1). \quad (29) \end{aligned}$$

Since we are mainly concerned with recombinant haplotypes $r_t = P_{12}(t) + P_{21}(t)$, we obtain, by summarizing (28) and (29),

$$\begin{aligned} r^*_t &= P^*_{12}(t) + P_{*21}(t) = r_{t-1} + 2\theta[P_{11}(t-1) \\ &\quad \times P_{22}(t-1) - P_{12}(t-1)P_{21}(t-1)]. \quad (30) \end{aligned}$$

Since the population F_t are produced by randomly mating from an F_2 population, the expectation of frequencies $P_{11}(t)$ and $P_{22}(t)$ will be identical. The same is true for $P_{12}(t)$ and $P_{21}(t)$. Thus,

$$\begin{aligned} P_{11}(t-1) &= P_{22}(t-1) = \tfrac{1}{2}(1 - r_{t-1}), \\ P_{12}(t-1) &= P_{21}(t-1) = \tfrac{1}{2}r_{t-1}. \quad (31) \end{aligned}$$

It follows from (30) and (31) that

$$r^*_t = \frac{\theta}{2} + (1-\theta)r_{t-1} \quad (32)$$

or

$$\Delta r_t = r^*_t - r_t = -\theta r_{t-1} + \frac{\theta}{2}.$$

Thus, the population process $\{X(t) = \text{number of recombination haplotypes at the } t\text{th generation}\}$ evolves as a Markov chain with the transition probability

$$\binom{2N}{j}(\Delta r_t)^j(1 - \Delta r_t)^{2N-j},$$

which can be approximated by the diffusion process with a generator given by (KARLIN and TAYLOR 1981)

$$L = \tfrac{1}{2}\frac{r_t(1-r_t)}{2N(t)}\frac{\partial^2}{\partial r_t^2} + \left(\frac{\theta}{2} - \theta r_t\right)\frac{\partial}{\partial r_t}. \quad (33)$$

APPENDIX B

Suppose that there are K QTL with the kth QTL, located at d_k, having additive effect α_k and dominance effect δ_k. Then, the true model is

$$y_i = \mu + \sum_{k=1}^{K}\alpha_k x_i(d_k) + \sum_{k=1}^{K}\delta_k z_i(d_k) + e_i, \quad i = 1, 2, \ldots, n,$$

where $x_i(d_k)$ and $z_i(d_k)$ are indicator variables with the values

$$x_i(d_k) = \begin{cases} 1 & Q_k = QQ \\ -1 & Q_k = qq \\ 0 & Q_k = Qq \end{cases}$$

and

$$z_i(d_k) = \begin{cases} 1 & Q_k = Qq \\ -1 & \text{otherwise,} \end{cases}$$

where Q and q are two alleles of the QTL. Recall that in the text we assumed the following simple linear regression model:

$$y_i = \mu + \alpha x_i + \delta z_i + e_i, \quad i = 1, 2, \ldots, n, \quad (34)$$

where x_i and z_i are dummy variables representing the genotype at the marker M_i.

It is easy to see that

$$E[x_1 - \bar{x}]^2 = \left(1 - \frac{1}{n}\right)E[x_1^2] = \tfrac{1}{2}\left(1 - \frac{1}{n}\right).$$

By the strong law of large numbers, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}[x_i - \hat{x}]^2 \xrightarrow{a.s} \tfrac{1}{2}. \quad (35)$$

Now we calculate $E[x_i(d_k)x_i)]$. Let $\rho_k(t)$ be the expectation of the recombinant haplotype at the marker M and the kth QTL. It can be shown that

$$\rho_k(t) = \tfrac{1}{2}(1 - e^{-t\Delta_k}),$$

where Δ_k is the genetic distance between the marker M and the kth QTL. Thus,

$$E[P\{x_i = 1,\ x_i(d_k) = 1\}] = E[P\{M = AA,\ Q_k = QQ\}] \approx \tfrac{1}{4}(1 - \rho_k(t))^2,$$

$$E[P\{x_i = -1,\ x_i(d_k) = -1\}] = E[P\{M = aa,\ Q_k = qq\}] \approx \tfrac{1}{4}(1 - \rho_k(t))^2,$$

$$E[P\{x_i = 1,\ x_i(d_k) = -1\}] = E[P\{M = AA,\ Q_k = qq\}] \approx \tfrac{1}{4}\rho_k^2(t),$$

$$E[P\{x_i = -1,\ x_i(d_k) = 1\}] = E[P\{M = aa,\ Q_k = QQ\}] \approx \tfrac{1}{4}\rho_k^2(t).$$

It follows that

$$E[x_i x_i(d_k)] \approx \tfrac{1}{2}[1 - 2\rho_k(t)] = \tfrac{1}{2}e^{-t\Delta_k}.$$

Since

$$P\{x_i = 1,\ z_i(d_k) = 1\} = P\{x_i = -1,\ z_i(d_k) = 1\}$$

and

$$P\{x_i = -1,\ z_i(d_k) = -1\} = P\{x_i = 1,\ z_i(d_k) = -1\},$$

we have

$$E[x_i z_i(d_k)] = 0.$$

Let

$$\bar{x}(d_k) = \frac{1}{n}\sum_{i=1}^{n} x_i(d_k) \quad \text{and} \quad \bar{z}(d_k) = \frac{1}{n}\sum_{i=1}^{n} z_i(d_k).$$

Then,

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y}]x_i = \frac{1}{n}\sum_{k=1}^{K}\alpha_k\sum_{i=1}^{n}[x_i(d_k) - \bar{x}(d_k)]x_i + \frac{1}{n}\sum_{k=1}^{K}\delta_k\sum_{i=1}^{n}[z_i(d_k) - \bar{z}(d_k)]x_i + \frac{1}{n}\sum_{i=1}^{n}x_i e_i.$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y}]x_i \xrightarrow{a.s.} \sum_{k=1}^{K}\frac{1}{2}\alpha_k e^{-t\Delta_k}. \tag{36}$$

Therefore, combining (35) and (36) yields

$$\hat{\alpha}(t) = \frac{\frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y}]x_i}{\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2} \xrightarrow{a.s} \sum_{k=1}^{K}\alpha_k e^{-t\Delta_k}.$$

Similarly, by noting that

$$E[z_i(d_k)z_i] = \tfrac{1}{2}e^{-2t\Delta_k},$$

$$E[x_i(d_k)z_i] = 0,$$

we can show that

$$\hat{\delta}(t) \xrightarrow{a.s} \sum_{k=1}^{K}\delta_k e^{-2t\Delta_k}.$$

APPENDIX C

Suppose a QTL is located at s. Then, from standard regression analysis, we have

$$\hat{\alpha}(d) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i(d)}{\sum_{i=1}^{n}[x_i(d) - \bar{x}(d)]^2}.$$

Recall that in APPENDIX B, we showed that

$$\frac{1}{n}\sum_{i=1}^{n}[x_i(d) - \bar{x}(d)]^2 \xrightarrow{a.s} \tfrac{1}{2}.$$

Clearly,

$$E\left\{x_1(d)\left[x_1(s) - \frac{1}{n}\sum_{j=1}^{n}x_j(s)\right]\right\} = \left(1 - \frac{1}{n}\right)E[x_1(d)x_1(s)].$$

By the strong law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})x_i(d) = \frac{1}{n}\alpha\sum_{i=1}^{n}\left\{x_i(d)\left[x_i(s) - \frac{1}{n}\sum_{j=1}^{n}x_j(s)\right]\right\} + \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})x_i(d) \xrightarrow{a.s} \frac{\alpha}{2}e^{-t|d-s|}.$$

Therefore,

$$\hat{\alpha}(d) \xrightarrow{a.s} \alpha e^{-t|d-s|},$$

which implies,

$$\lim_{n\to\infty} E[\hat{\alpha}(d)] = \alpha e^{-t|d-s|},$$

$$E[X_d] \sim \sqrt{\frac{n}{2}}\frac{\alpha}{\sigma_e}e^{-t|d-s|}.$$

APPENDIX D

Assume that the map is dense. Given observed phenotypes $y = [y_1, \ldots, y_n]^T$ and the genotypes of the markers $M(\gamma^*) = [M_1(\gamma^*), \ldots, M_n(\gamma^*)]^T$ under the true QTL map location γ^* as well as $M(\gamma_n) = [M_1(\gamma_n), \ldots, M_n(\gamma_n)]^T$ under an alternative γ_n, the log likelihood ratio $\log(L(y, \gamma^*)/L(y, \gamma_n))$ can be expressed as

$$\log\left(\frac{L(\gamma^*)}{L(\gamma_n)}\right) = \sum_{i=1}^{n}\log\left(\frac{L(y_i, \gamma^*)}{L(y_i, \gamma_n)}\right), \tag{37}$$

where

$$L(y_i, \gamma) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(y_i - \mu - \alpha x_i - \delta z_i)^2}{2\sigma^2}\right\}$$

and γ is the assumed QTL map location. After some calculations, we have

$$2\sigma^2\log\left(\frac{L(y_i, \gamma^*)}{L(y_i, \gamma_n)}\Big|_{M(\gamma^*),M(\gamma)}\right)$$

$$= \begin{cases} -2(y_i - \mu)(2\delta - \alpha) + (2\delta - \alpha)\delta, & M_i(\gamma^*) = AA, M_i(\gamma_n) = Aa \\ (2\delta - \alpha)[2(y_i - \mu) - \delta], & M_i(\gamma^*) = Aa, M_i(\gamma_n) = AA \\ 2\alpha[2(y_i - \mu) + 2\delta], & M_i(\gamma^*) = AA, M_i(\gamma_n) = aa \\ -2\alpha[2(y_i - \mu) + 2\delta], & M_i(\gamma^*) = aa, M_i(\gamma_n) = AA \\ (\alpha + 2\delta)[2(y_i - \mu) + \alpha], & M_i(\gamma^*) = Aa, M_i(\gamma_n) = aa \\ -(\alpha + 2\delta)[2(y_i - \mu) + \alpha], & M_i(\gamma^*) = aa, M_i(\gamma_n) = Aa \\ 0 & \text{otherwise.} \end{cases} \quad (38)$$

Thus,

$$E_{\gamma^*}\left[\log\left(\frac{L(y_i, \gamma^*)}{L(y_i, \gamma_n)}\right)\right] = \frac{1}{2\sigma^2}\{(2\delta - \alpha)^2 \times E[\theta_{\gamma^*\gamma_n}(1 - \theta_{\gamma^*\gamma_n})] + 4\alpha^2 1/2 E[\theta^2_{\gamma^*\gamma_n}] + (2\delta + \alpha)^2 E[\theta_{\gamma^*\gamma_n}(1 - \theta_{\gamma^*\gamma_n})]\}$$

$$= \frac{1}{2\sigma^2}[2E[\theta_{\gamma^*\gamma_n}](4\delta^2 + \alpha^2) - 8E[\theta^2_{\gamma^*\gamma_n}]\delta^2], \quad (39)$$

where $\theta_{\gamma^*\gamma_n}$ denotes the frequency of the recombinant haplotype at the markers $M(\gamma^*)$ and $M(\gamma_n)$. Combining (37) and (39) yields

$$E_{\gamma^*}\left[\log\frac{L(Y, \gamma^*)}{L(Y, \gamma_n)}\right] = nE_{\gamma^*}\left[\log\frac{L(y_1, \gamma^*)}{L(y_1, \gamma_n)}\right]$$

$$= \frac{n}{2\sigma^2}\{2E[\theta_{\gamma^*\gamma_n}](4\delta^2 + \alpha^2) - 8E[|\theta^2_{\gamma^*\gamma_n}]\delta^2\}. \quad (40)$$

By definition, the KL distance between two distributions $L(Y, \gamma^*)$ and $L(Y, \gamma_n)$ is defined as (KONG and WRIGHT 1995)

$$K_t(L(Y, \gamma^*), L(Y, \gamma_n)) = E_{\gamma^*}\left[\log\frac{L(Y, \gamma^*)}{L(Y, \gamma_n)}\right]$$

$$= \frac{n}{2\sigma^2}\{2E[\theta_{\gamma^*\gamma_n}](4\delta^2 + \alpha^2) - 8E[\theta^2_{\gamma^*\gamma_n}]\delta^2\},$$

where the subscript t indicates that KL distance is measured for the likelihood functions $L(Y, \gamma^*)$ and $L(Y, \gamma_n)$ in F_t population.

APPENDIX E

Let $f = \theta^2_{\gamma^*\gamma_n}(t)$. Then by the Hille-Yosida theorem, we obtain

$$\frac{dE[\theta^2_{\gamma^*\gamma_n}(t)]}{dt} = E[L(\theta^2_{\gamma^*\gamma})]$$

$$= -\left(2\theta_n + \frac{1}{2N(t)}\right)E[\theta^2_{\gamma^*\gamma_n}] + \left(\theta_n + \frac{1}{2N(t)}\right)E[\theta_{\gamma^*\gamma_n}]$$

$$= -\left(2\theta_n + \frac{1}{2N(t)}\right)E[\theta^2_{\gamma^*\gamma}] + 1/2\left(\theta_n + \frac{1}{2N(t)}\right)[1 - e^{-\theta_n t}]$$

$$= -\lambda(t)E[\theta^2_{\gamma^*\gamma_n}] + b[1 - e^{-\theta_n t}], \quad (41)$$

where

$$\lambda(t) = 2\theta_n + \frac{1}{2N(t)},$$

$$b = \frac{1}{2}\left(\theta_n + \frac{1}{2N(t)}\right),$$

and θ_n is the recombination fraction between γ^* and γ_n.

Note that

$$\frac{d\{e^{\int_0^t\lambda(\tau)d\tau}E[\theta^2_{\gamma^*\gamma_n}(t)]\}}{dt} = \lambda(t)e^{\int_0^t\lambda(\tau)d\tau} \times E[\theta^2_{\gamma^*\gamma_n}] + e^{\int_0^t\lambda(\tau)d\tau}\frac{d\{E[\theta^2_{\gamma^*\gamma_n}]\}}{dt} = be^{\int_0^t\lambda(\tau)d\tau}$$

$$[1 - e^{-\theta_n t}] \text{ by (41)} = bh(t), \quad (42)$$

where $h(t)$ is defined as in the text. Integrating both sides of (42) yields

$$e^{\int_0^t\lambda(\tau)d\tau}E[\theta^2_{\gamma^*\gamma}(t)] = b\int_0^t h(\eta)d\eta.$$

Thus,

$$E[\theta^2_{\gamma^*\gamma_n}(t)] = be^{-\int_0^t\lambda(\tau)d\tau}\int_0^t h(\eta)d\eta.$$

The proof is complete.

APPENDIX F

From LEHMANN (1983, p. 116), we know that

$$\lim_{\Delta_-\to 0}\psi^2(x, \theta) = \lim_{\Delta_-\to 0}\left[\frac{P(x, \theta + \Delta_-)}{P(x, \theta)} - 1\right]^2$$

$$= \left[\frac{\partial p(x, \theta)}{\partial\theta}\Big|_{\theta_-}\right]^2 = \left[\frac{\partial P(x, \theta)}{\partial\theta}\Big|_{\theta_+}\right]^2$$

$$= \lim_{\Delta_+\to 0}\left[\frac{P(x, \theta + \Delta_+)}{P(x, \theta)} - 1\right]^2 = \lim_{\Delta_+\to 0}\psi^2(x, \theta). \quad (43)$$

Thus,

$$\mathrm{Var}(\delta) \geq \lim_{\Delta_- \to 0} \frac{[g(\theta + \Delta_-) - g(\theta)]^2}{E_\theta[\psi(x, \theta)^2]}$$

by (6) in LEHMANN (1983, p. 106)

$$= \frac{[g'(\theta)]^2}{E_\theta\left[\left.\frac{\partial P(x, \theta)}{\partial \theta}\right|_{\theta_-}\right]^2} = \frac{[g'(\theta)]^2}{E_\theta\left[\left.\frac{\partial P(x, \theta)}{\partial \theta}\right|_{\theta_+}\right]^2}$$

$$= \lim_{\Delta_+ \to 0} \frac{[g(\theta + \Delta_+) - g(\theta)]^2}{E_\theta[\psi(x, \theta)^2]}.$$

APPENDIX G

Note that

$$\left.\frac{\partial|\gamma - \gamma^*|}{\partial \gamma}\right|_{\gamma_-^*} = -1.$$

Since

$$E[\theta_{\gamma\gamma^*}] = \tfrac{1}{2}(1 - e^{-t|\gamma - \gamma^*|}),$$

it is easy to see that

$$\left.\frac{\partial E[\theta_{\gamma\gamma^*}]}{\partial \gamma}\right|_{\gamma_-^*} = -\frac{t}{2}. \tag{44}$$

Recall that

$$h(\eta) = e^{\int_0^\eta \lambda(\tau)\,d\tau}(1 - e^{-\theta_n \eta}).$$

Thus,

$$\left.\frac{\partial h(\eta)}{\partial \gamma}\right|_{\gamma_-^*} = -\frac{t}{2}\eta e^{\int_0^\eta d\tau/2N(\tau)}$$

and

$$h(\eta)\,|_{\gamma_-^*} = 0.$$

After some lengthy, but straightforward calculations, we obtain

$$\left.\frac{\partial E[\theta_{\gamma\gamma^*}]^2}{\partial \gamma}\right|_{\gamma_-^*} = -\frac{t}{2}\frac{1}{4N(t)} e^{-\int_0^t d\tau/2N(\tau)} \int_0^t \eta e^{1/2\int_0^\eta d\tau/N(\tau)}\, d\eta.$$

To calculate the Fisher information, we use the following notations:

$$a = \left.\frac{\partial E[\theta_{\gamma\gamma^*}]}{\partial \gamma}\right|_{\gamma_-^*} \tag{45}$$

and

$$b = \left.\frac{\partial E[\theta_{\gamma\gamma^*}]^2}{\partial \gamma}\right|_{\gamma_-^*}. \tag{46}$$

Then, the Fisher information can be calculated by conditioning on the markers $M(\gamma^*)$ and $M(\gamma)$:

$$I(t) = E\left[\left.\frac{\partial \log\left(\left.\frac{L(y_i, \gamma^*)}{L(y_i, \gamma)}\right|_{M(\gamma^*), M(\gamma)}\right) P(M(\gamma), M(\gamma^*))}{\partial \gamma}\right|_{\gamma = \gamma_-^*}\right]^2$$

$$= E\Bigg[\frac{(a-b)(2\delta - \alpha)}{4\sigma^2}(2\delta - \alpha - 2e_i) + \frac{(a-b)(2\delta - \alpha)}{4\sigma^2}(2\delta - \alpha + 2e_i) + \frac{b\alpha}{4\sigma^2} \times (2\alpha + 2e_i) + \frac{b\alpha}{4\sigma^2}(2\alpha - 2e_i) + \frac{(a-b)(2\delta + \alpha)}{4\sigma^2} \times (2\delta + \alpha + 2e_i) + \frac{(a-b)(2\delta + \alpha)}{4\sigma^2}(2\delta + \alpha - 2e_i)\Bigg]^2$$

$$= \frac{(a-b)^2}{8\sigma^4}[2(2\delta - \alpha)^4 + 2(2\delta + \alpha)^4 + 4\delta(2\delta - \alpha)^2 \times (2\delta + \alpha) + 2(4\delta^2 - \alpha^2)^2] + \frac{b(a-b)\alpha^2}{2\sigma^4} \times [(2\delta - \alpha)^2 + 3(2\delta + \alpha)^2] + \frac{b^2\alpha^4}{2\sigma^4} + \frac{(a-b)^2}{2\sigma^2} \times (8\delta^2 + 5\alpha^2 - 8\alpha\delta) + \frac{b(a-b)}{\sigma^2}\alpha(\alpha + 2\delta).$$

APPENDIX H

The proof is similar to that of RODOLPHE and LEFORT (1993) for the F_2 population.

In an F_t population, the allele frequencies of the markers remain unchanged, only the haplotype frequencies change with time t. Therefore, for additive effects,

$$E[x_{i,1}x_{i,j}] = 0, \quad E[x_{i,j}^2] = \tfrac{1}{2},$$

and for dominance effects

$$E[x_{i,1}x_{i,j}] = 0, \quad E[x_{i,j}^2] = 1,$$

which are the same as that in F_2 population.

Now, we modify the formulas for the F_2 population involving haplotype frequencies.

First, we consider j and j' as indexes for additive effects. Let $\rho_{jj'} = \tfrac{1}{2}(1 - e^{-t\Delta_{jj'}})$, the recombination fraction between two markers j and j'. Then,

$$P\{x_{ij} = 1, x_{ij'} = 1\} = P\{M_i(j) = AA, M_i(j') = BB\} = \tfrac{1}{4}(1 - \rho_{jj'}(t))^2,$$

$$P\{x_{ij} = -1, x_{ij'} = -1\} = P\{M_i(j) = aa, M_i(j') = bb\} = \tfrac{1}{4}(1 - \rho_{jj'}(t))^2.$$

Thus,

$$P\{x_{ij}x_{ij'} = 1\} = \tfrac{1}{2}(1 - \rho_{jj'}(t))^2.$$

Similarly,

$$P\{x_{ij}x_{ij'} = -1\} = \tfrac{1}{2}\rho_{jj'}^2(t).$$

Therefore,

$$E[x_{ij}x_{ij'}] = P\{x_{ij}x_{ij'} = 1\} - P\{x_{ij}x_{ij'} = -1\}$$
$$= \tfrac{1}{2}[1 - 2\rho_{jj'}(t)] = \tfrac{1}{2}e^{-t\Delta_{jj'}}. \quad (47)$$

Note that for $t = 2$, $\rho_{jj'}(t) = \tfrac{1}{2}(1 - e^{-2\Delta_{jj'}})$, which is the recombination fraction between the two markers j and j'. In this case, (47) is reduced to the formula of RODOLPHE and LEFORT (1993) for an F_2 population.

Now we consider j and j' as indexes for dominance effects. It is not hard to see that

$$P\{x_{ij} = 1, x_{ij'} = 1\} = P\{M_i(j) = Aa, M_i(j') = Bb\}$$
$$= \tfrac{1}{2}[(1 - \rho_{jj'}(t))^2 + \rho^2_{jj'}(t)],$$

and

$$P\{x_{ij} = -1, x_{ij'} = -1\} = P\{M_i(j) = AA, M_i(j') = BB\}$$
$$+ P\{M_i(j) = aa, M_i(j') = BB\} + P\{M_i(j)$$
$$= AA, M_i(j') = bb\} = \tfrac{1}{4}[1 - \rho_{jj'}(t)]^2$$
$$+ \tfrac{1}{4}\rho^2_{jj'}(t) + \tfrac{1}{4}\rho^2_{jj'}(t) + \tfrac{1}{4}[1 - \rho^2_{jj'}(t)]$$
$$= \tfrac{1}{2}[(1 - \rho_{jj'}(t))^2 + \rho^2_{jj'}(t)].$$

Thus,

$$P\{X_{ij}X_{ij'} = 1\} = 1 - 2\rho_{jj'}(t) + 2\rho^2_{jj'}(t).$$

Similarly,

$$P\{X_{ij}X_{ij'} = -1\} = 2\rho_{jj'}(t)[1 - \rho_{jj'}(t)].$$

Therefore,

$$E[X_{ij}X_{ij'}] = [1 - 2\rho_{jj'}(t)]^2 = e^{-2t\Delta_{jj'}}.$$

Considering j and j' as indexes for additive and dominance effects, we obtain exactly the same results as that of RODOLPHE and LEFORT (1993), *i.e.*, $E[X_{ij}X_{ij'}] = 0$.

APPENDIX I

We use j as an index for additive effects. First we show

$$E[X_{ij}Y_i] = \sum_{k=1}^{K} b_k e^{-t\Delta_{jk}},$$

where K is the number of QTL, b_k is the additive effects of the kth QTL and Δ_{jk} is the genetic distance between the marker M_j and the kth QTL. To see this, letting $G_i(k)$ denote the genotype of the kth QTL of the i individual and computing $E[X_{ij}Y_j]$ by conditioning on the genotype of the marker j, we obtain

$$E[X_{ij}Y_j] = P(M_i(j) = AA)E[X_{ij}Y_i|M_i(j) = AA]$$
$$+ P(M_i(j) = Aa)E[X_{ij}Y_j|M_i(j) = Aa]$$
$$+ P(M_i(j) = aa)E[X_{ij}Y_i|M_i(j) = aa]$$
$$= \tfrac{1}{4}\left\{\sum_{k=1}^{K} E[Y_i|G_i(k) = QQ]P(G_i(k)\right.$$
$$= QQ|M_i(j) = AA) + \sum_{k=1}^{K} E[Y_i|G_i(k)$$
$$\left. = qq]P(G_i(k) = qq|M_i(j) = AA)\right\}$$
$$- \tfrac{1}{4}\left\{\sum_{k=1}^{K} E[Y_i|G_i(k) = QQ]P(G_i(k)\right.$$
$$= QQ|M_i(j) = aa) + \sum_{k=1}^{K} E[Y_i|G_i(k)$$
$$\left. = qq]P(G_i(k) = qq|M_i(j) = aa)\right\}$$
$$= \tfrac{1}{4}\sum_{k=1}^{K} b_k[(1 - \rho_{ik}(t))^2 - \rho^2_{ik}(t)]$$
$$- \tfrac{1}{4}\sum_{k=1}^{K} b_k[\rho^2_{ij}(t) - (1 - \rho_{ik}(t))^2]$$
$$= \tfrac{1}{2}\sum_{k=1}^{K} b_k e^{-t\Delta_{jk}} = \tfrac{1}{2}Z_j,$$

where

$$Z_j = \sum_{k=1}^{K} b_k e^{-t\Delta_{jk}}.$$

By the strong law of large numbers, we have

$$\hat{\alpha}_j(t) \xrightarrow{a.s} \frac{-a_l}{1 - a_l^2} Z_{j-1} + \frac{1 - a_l^2 a_r^2}{(1 - a_l^2)(1 - a_r^2)} Z_j$$
$$- \frac{a_l}{(1 - a_r^2)} Z_{j+1} = \sum_k b_k e^{-t\Delta_{kj}} \frac{1 - e^{-2t\Delta_{j-1,k}}}{1 - e^{-2t\Delta_{j-1,j}}}.$$

where $a_l = e^{-t\Delta_{j-1,j}}$, $a_r = e^{-t\Delta_{jj+1}}$

Similarly, we can show

$$\hat{\delta}_j(t) \xrightarrow{a.s} \sum_k c_k e^{-2t\Delta_{kj}}(1 - e^{-4t\Delta_{j-1,k}})/(1 - e^{-4t\Delta_{j-1,j}}).$$