Asymptotic theory for gene mapping

AUGUSTINE KONG AND FRED WRIGHT

Department of Statistics, The University of Chicago, Chicago, IL 60637

Communicated by Herman Chernoff, October 12, 1993

ABSTRACT In genetic linkage mapping, the precision of the location estimate depends on both the number of individuals and the number of markers. Convergence rates and limit laws are established for several situations in which the number of markers depends on the sample size. Practical implications of the results are discussed.

Section 1. Introduction and Summary

This paper explores some asymptotic properties of likelihood-based location estimates of trait susceptibility genes. With the increased resolution of genetic marker maps, mathematical theory which addresses the impact of marker density on the convergence rates of location estimates becomes increasingly relevant. Due to its simplicity and its importance to animal and plant studies, the backcross breeding design is used to develop the theory in this paper. However, the key conclusions established can be extended qualitatively to much more general settings, including the study of human pedigree data, where controlled breeding is not possible. Moreover, it is emphasized that our results apply to genes that may control complex traits—i.e., incomplete penetrance, phenocopy, quantitative phenotype (QTL), etc. Some asymptotic results regarding the log-likelihood function can be found in ref. 1. The focus there was on hypothesis testing, and the results were developed under the null assumption that no gene is present. Our results are different because we assume that a gene is present, and at issue is the resolution of location estimates.

Section 2 establishes terminology and notation. Section 3, divided into four subsections, contains results on convergence rates and provides limit laws for the log-likelihood function. Depending on the density of the markers, the rate of convergence to the true location of the gene can be anywhere between order $n^{-1/2}$ and n^{-1} , where n is the number of bred individuals. Section 4 discusses issues of nuisance parameters, model misspecifications, and multiple trait genes.

Section 2. The Backcross Design and Likelihood Estimation

For simplicity it is assumed here that a single gene affects the trait phenotype. To create a backcross population, the breeder starts by creating two different populations, P0 and P1, of genetically identical individuals. Furthermore, these individuals are assumed to be homozygous at all loci. For each locus, the two identical alleles of a P0 individual are identified as 0 and the two identical alleles of a P1 individual are identified as 1. For the trait gene locus, the alleles 0 and 1 are assumed to have different effects on the trait phenotype. Any locus where the alleles 0 and 1 can be distinguished experimentally is a potential marker locus. Hybrids are formed by mating a P0 and a P1 individual. Hence, for a hybrid individual, each chromosome pair consists of one chromosome which has all 0 alleles and another which has all 1 alleles.

To produce a backcross individual, a hybrid individual is mated with a P1 individual. For each chromosome pair of the offspring, the P1 parent contributes a chromosome with alleles 1 for all loci. The hybrid parent, however, contributes a random chromosome which usually consists of both 0 and 1 alleles due to crossovers. Backcross individuals are assumed to be independent of each other.

The focus of this paper is on the localization of the trait gene to a small region. Because it is intuitively clear that with enough data one can quickly determine on which chromosome pair the trait gene lies, the notations and mathematical results to be established focus on the chromosome pair that contains the trait gene of interest. For simplicity, it is assumed that there is no genetic interference—i.e., the crossover process is Poisson. The genetic length (expected number of crossovers for an individual) of the chromosome pair of interest is Γ . The genetic map will be used—i.e., a position on the chromosome is denoted by a number γ which lies in the interval $[0, \Gamma]$. The true location of the gene is denoted by γ^* . The genotype of backcross individual i at a position γ is coded by

$$x_i(\gamma) = \text{genotype at location } \gamma = \begin{cases} 0 & \text{alleles are } \{0, 1\} \text{ at } \gamma \\ 1 & \text{alleles are } \{1, 1\} \text{ at } \gamma. \end{cases}$$

A position γ is a marker locus if $x_i(\gamma)$ is observed. For an individual i , the two conditional phenotype distributions corresponding to the two possible genotypes at γ^* are defined as

$$f(y_i | x_i(\gamma^*) = 0) = f_0(y_i)$$

and

$$f(y_i | x_i(\gamma^*) = 1) = f_1(y_i).$$

For simplicity, $f_0(y)$ and $f_1(y)$ are assumed to be known here. For example, for a quantitative trait, $f_0(y)$ and $f_1(y)$ may be two normal densities with different means.

Because of random crossovers, $x_i(\gamma)$ is a random 0–1 step function on $[0, \Gamma]$. We use γ^* as the reference point for this random process. In particular, $x_i(\gamma^*)$ is equally likely to be 0 or 1. For any other location γ , $x_i(\gamma) = x_i(\gamma^*)$ if there are zero or an even number of crossovers (no recombination) between γ and γ^* , and $x_i(\gamma) = 1 - x_i(\gamma^*)$ if there are an odd number of crossovers (recombination) between γ and γ^* . Assuming that the crossover process is Poisson, we have Haldane's map function

$$d = -\frac{1}{2} \log(1 - 2\theta), \quad \theta = \frac{1}{2} (1 - e^{-2d}),$$

where d and θ denote, respectively, the genetic distance (expected number of crossovers) and recombination probability between two loci. For small d , θ is very close to d . Finally, $\mathbf{a}_i$ denotes the (observable) marker genotypes for the i th individual—i.e., $\mathbf{a}_i = \{x_i(\gamma); \gamma \text{ is a marker locus}\}$. For n backcross individuals, the likelihood function of the trait gene location, conditional on the markers, is

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: KL, Kullback–Leibler.

$$L(\gamma) = \prod_{i=1}^n L_i(\gamma) \\ = \prod_{i=1}^n [f_0(y_i)P(x_i(\gamma) = 0|\mathbf{a}_i) + f_1(y_i)P(x_i(\gamma) = 1|\mathbf{a}_i)],$$

and the log-likelihood is

$$\ell(\gamma) = \sum_{i=1}^n \ell_i(\gamma) \\ = \sum_{i=1}^n \log[f_0(y_i)P(x_i(\gamma) = 0|\mathbf{a}_i) + f_1(y_i)P(x_i(\gamma) = 1|\mathbf{a}_i)].$$

The number of individuals n is referred to as the “sample size” for the backcross experiment.

Consider two cases for the density of the markers: (i) Dense markers—following ref. 1, it is assumed here that markers are everywhere on the chromosome. In other words, $\mathbf{a}_i = \{x_i(\gamma); \gamma \in [0, \Gamma]\}$. (ii) Nondense markers—there are m markers which, for simplicity, are assumed to be equally spaced across $[0, \Gamma]$. The genetic distance between adjacent markers is denoted by $\delta = \Gamma/m$.

With dense markers, $x_i(\gamma)$ is observed at each γ . In the nondense case, let γ_1 and $\gamma_1 + \delta$ be the two markers immediately to the left and right of $\gamma = \gamma_1 + u$, $0 \leq u \leq \delta$. Then

$$P[x_i(\gamma) = 1|x_i(\gamma_1) = 0, x_i(\gamma_1 + \delta) = 1] \\ = P[x_i(\gamma) = 0|x_i(\gamma_1) = 1, x_i(\gamma_1 + \delta) = 0] \\ = \frac{\theta_{\gamma_1, \gamma}(1 - \theta_{\gamma, \gamma_1 + \delta})}{\theta_{\gamma_1, \gamma_1 + \delta}} = \frac{1}{2} \frac{(1 - e^{-2u})(1 + e^{-2(\delta-u)})}{1 - e^{-2\delta}} \approx \frac{u}{\delta}. \quad [1]$$

The approximation applies for small δ . Similarly

$$P[x_i(\gamma) = 0|x_i(\gamma_1) = x_i(\gamma_1 + \delta) = 1] \\ = P[x_i(\gamma) = 1|x_i(\gamma_1) = x_i(\gamma_1 + \delta) = 0] \\ = \frac{\theta_{\gamma_1, \gamma}\theta_{\gamma, \gamma_1 + \delta}}{1 - \theta_{\gamma_1, \gamma_1 + \delta}} = \frac{1}{2} \frac{(1 - e^{-2u})(1 - e^{-2(\delta-u)})}{1 + e^{-2\delta}} \approx u(\delta - u). \quad [2]$$

Example: Monogenic disease with incomplete penetrance and phenocopy. Consider a qualitative trait such that an individual i is either affected ($y_i = 1$) or normal ($y_i = 0$). Let

$$f_0(y) = \begin{cases} 1 - \tau_0, & y = 0 \\ \tau_0, & y = 1, \end{cases} \quad f_1(y) = \begin{cases} 1 - \tau_1, & y = 0 \\ \tau_1, & y = 1. \end{cases}$$

In other words, allele 1 is recessive, τ_1 is the *penetrance*, and τ_0 is the probability of a phenocopy. An individual with $y_i = 1$ has $\ell_i(\gamma) = \log[\tau_0 P(x_i(\gamma) = 0|\mathbf{a}_i) + \tau_1 P(x_i(\gamma) = 1|\mathbf{a}_i)]$. Fig. 1A plots a realization of $\ell_i(\gamma)$ for such an individual on a chromosome pair of length 1, under three marker density scenarios with $\tau_0 = 0.1$ and $\tau_1 = 0.9$. With dense markers, $\ell_i(\gamma)$ jumps at exactly the two crossover points. With nondense markers, the log-likelihood is relatively flat between two adjacent markers that do not recombine. For the same model, Fig. 1B plots the overall log-likelihood curves for data simulated for 20 individuals and with $\gamma^* = 0.3$. Observe that for the nondense cases the log-likelihood at a (nondense) marker location is the same as the dense log-likelihood at that location. However, the dense marker log-likelihood is substantially higher in the neighborhood of the true location, illustrating its superior resolution. Data have also been simulated for a quantitative trait model, giving similar results. ■

FIG. 1. log-likelihoods for the monogenic disease model with $\tau_0 = 0.1$ and $\tau_1 = 0.9$. The log-likelihood is plotted vs. map location under three marker density scenarios for an individual (A) and for 20 individuals with true location $\gamma^* = 0.3$ (B). —, Dense markers; ---, markers at intervals of 0.2; ···, markers at $\gamma = 0$ and $\gamma = 1$ only.

Section 3. Rates of Convergence

We study the asymptotic behavior of the log-likelihood function by considering a number of situations where both n (the sample size) and m (the number of markers) are allowed to vary.

3.1. Dense Markers ($\delta = 0$). With dense markers, as n increases, the likelihood function becomes concentrated in intervals about the true location γ^* with widths of order n^{-1} . This can be seen most easily by inspecting the expected log-likelihood ratios between γ^* and other locations γ .

The observed data are the vectors of phenotypes and markers—i.e., **data** = $(\mathbf{y}, \mathbf{a})$ where $\mathbf{y} = \{y_1, \dots, y_n\}$ and $\mathbf{a} = \{\mathbf{a}_1, \dots, \mathbf{a}_n\}$. The distribution of the data under γ^* is $p(\mathbf{data}; \gamma^*)$, while the distribution of the data under an alternative γ is $p(\mathbf{data}; \gamma)$. The Kullback-Leibler (KL) distance between the two distributions, measured from $p(\mathbf{data}; \gamma^*)$, is

$$K(p_{\gamma^*}, p_{\gamma}) = E_{\gamma^*} \left\{ \log \left(\frac{p(\mathbf{y}, \mathbf{a}; \gamma^*)}{p(\mathbf{y}, \mathbf{a}; \gamma)} \right) \right\} \\ = E_{\gamma^*} \left\{ \log \left(\frac{p(\mathbf{a}; \gamma^*)p(\mathbf{y}|\mathbf{a}; \gamma^*)}{p(\mathbf{a}; \gamma)p(\mathbf{y}|\mathbf{a}; \gamma)} \right) \right\} \\ = \sum_{i=1}^n E \left[E_{\gamma^*} \left\{ \log \left(\frac{p(y_i|\mathbf{a}_i, \gamma^*)}{p(y_i|\mathbf{a}_i, \gamma)} \right) \middle| \mathbf{a}_i \right\} \right] \\ = n \times E \left[E_{\gamma^*} \left\{ \log \left(\frac{p(y_1|\mathbf{a}_1, \gamma^*)}{p(y_1|\mathbf{a}_1, \gamma)} \right) \middle| \mathbf{a}_1 \right\} \right].$$

Note that y and a are treated as random when evaluating the expectations. With dense markers,

$$\log\left(\frac{p(y_i|a_i, \gamma^*)}{p(y_i|a_i, \gamma)}\right) = \log\left(\frac{p(y_i|x_i(\gamma^*); \gamma^*)}{p(y_i|x_i(\gamma); \gamma)}\right)$$

$$= \begin{cases} 0, & x_i(\gamma^*) = x_i(\gamma) \\ \log\left(\frac{f_0(y_i)}{f_1(y_i)}\right), & x_i(\gamma^*) = 0, x_i(\gamma) = 1 \\ \log\left(\frac{f_1(y_i)}{f_0(y_i)}\right), & x_i(\gamma^*) = 1, x_i(\gamma) = 0 \end{cases} \quad [3]$$

The last two cases correspond to recombinations between loci γ^* and γ , and each occurs with probability $\frac{1}{2}\theta_{\gamma^*, \gamma}$.

RESULT 1. The KL distance $K(p_{\gamma^*}, p_\gamma)$ is

$$\frac{n\theta_{\gamma^*, \gamma}}{2} \left\{ E_{f_0} \left[\log \left(\frac{f_0(y)}{f_1(y)} \right) \right] + E_{f_1} \left[\log \left(\frac{f_1(y)}{f_0(y)} \right) \right] \right\}$$

$$= \frac{n\theta_{\gamma^*, \gamma}}{2} \{K(f_0, f_1) + K(f_1, f_0)\}.$$

The key here is that the KL distance is proportional to $\theta_{\gamma^*, \gamma}$. Because $\theta_{\gamma^*, \gamma}$ and $|\gamma - \gamma^*|$ are approximately equal when small, if we let $n \rightarrow \infty$ and allow γ to depend on n , the KL distance diverges iff $n|\gamma - \gamma^*|$ diverges. We have

RESULT 2. Let γ_n be a sequence of locations which depend on the sample size n . If $n|\gamma_n - \gamma^*| \rightarrow \infty$ as $n \rightarrow \infty$, then $L(\gamma_n)/L(\gamma^*) \rightarrow_P 0$.

We now study the behavior of the log-likelihood function within a *shrinking window* of width $O(n^{-1})$. Consider a Poisson process on the interval $[0, T]$ with some constant intensity, and label the random points as $0 \leq T_1 \leq T_2 \leq \dots \leq T_L \leq T$, where L is the (random) number of events in the interval $[0, T]$. Let $J_1, J_2, \dots$ be iid random variables with common distribution F_J . Define $C(0) = 0$ and

$$C(t) = C(0) + \sum_{T_i \leq t} J_i$$

for $0 < t \leq T$. In other words, $C(t)$ is a random step function with random jump locations and random jump increments. It is called a *compound Poisson process* (2). Here we construct a specific F_J for the jump increments. Let $F_{(0,1)}$ be the distribution of $\log[f_1(y)/f_0(y)]$ with y having density $f_0(y)$, and let $F_{(1,0)}$ be the distribution of $\log[f_0(y)/f_1(y)]$ with y having density $f_1(y)$. Define $F_J = \frac{1}{2}(F_{(0,1)} + F_{(1,0)})$. The expectation of a jump increment is then $-[K(f_0, f_1) + K(f_1, f_0)]/2$ and hence negative.

Definition: For sample size n and some fixed $T > 0$, within the window $[\gamma^*, \gamma^* + (T/n)]$ to the right of γ^* , define for $0 \leq t \leq T$ the scaled log-likelihood

$$\xi_n(t) = \ell\left(\gamma^* + \frac{t}{n}\right) - \ell(\gamma^*).$$

RESULT 3. As $n \rightarrow \infty$, $\xi_n(t)$ converges weakly under the uniform metric to a compound Poisson process with jump point intensity 1 and increment distribution F_J .

Proof: For an individual i , let $\phi_i(\gamma) = \ell_i(\gamma) - \ell_i(\gamma^*) = \log[L_i(\gamma)/L_i(\gamma^*)]$. Hence

$$\xi_n(t) = \sum_{i=1}^n \left[\ell_i\left(\gamma^* + \frac{t}{n}\right) - \ell_i(\gamma^*) \right] = \sum_{i=1}^n \phi_i\left(\gamma^* + \frac{t}{n}\right).$$

Consider the behavior of $\phi_i(\gamma)$ as γ increases from $\gamma = \gamma^*$. Clearly $\phi_i(\gamma)$ is zero until it jumps at the first crossover point

to the right of γ^* . Because it has equal chance to be crossing over from $x_i = 0$ to $x_i = 1$ and the reverse, the jump increment has distribution F_J . After this first jump, $\phi_i(\gamma)$ again remains constant until the next crossover point for that individual. Consider the event that, for all individuals i , there is either 0 or 1 crossover between γ^* and $\gamma^* + (T/n)$. The probability of this event is

$$\left[\left(1 + \frac{T}{n}\right) e^{-T/n} \right]^n = \left(1 + \frac{T}{n}\right)^n e^{-T} \rightarrow 1 \text{ (as } n \rightarrow \infty).$$

Conditioned on this event, the jump locations of $\xi_n(t)$ within the shrinking window are all *first* crossovers of different individuals. As a consequence, the jump increments of $\xi_n(t)$ have common distribution F_J and are independent. The probability of the event approaches 1, and $\xi_n(t)$ converges to a compound Poisson process. Because there are n individuals, the jump locations of the overall log-likelihood function follow a Poisson process with intensity n . Because $\xi_n(t)$ is rescaled by the factor n , its jump locations have intensity 1. ■

Definition: To the left of γ^* , define, for $0 \leq t \leq T$,

$$\xi_n^\circ(t) = \ell\left(\gamma^* - \frac{t}{n}\right) - \ell(\gamma^*).$$

RESULT 4. As $n \rightarrow \infty$, $\xi_n^\circ(t)$ converges weakly to a compound Poisson process with jump point intensity 1 and increment distribution F_J . Also, $\xi_n(t)$ and $\xi_n^\circ(t)$ are asymptotically independent.

Proof: From symmetry, $\xi_n(t)$ and $\xi_n^\circ(t)$ have the same distribution. Because the crossover process is Poisson, the jump locations of $\xi_n(t)$ and $\xi_n^\circ(t)$ are independent. Furthermore, as $n \rightarrow \infty$, the probability that for all individuals i there is 0 or 1 crossover in the shrinking window $[\gamma^* - (T/n), \gamma^* + (T/n)]$ approaches 1. Hence, within the window, crossovers on the left and right of γ^* come from different individuals and the jump increments are independent. ■

We refer to the limit behavior of the scaled log-likelihood function around γ^* as a *mirror-image* compound Poisson process. The jump increments have negative expectations and so the log-likelihood drifts downward away from γ^* .

3.2. The Case with $\delta = o(n^{-1})$. Suppose the markers are not dense, but the number of markers m increases at a faster rate than n . The distance between two adjacent markers is $\delta = O(m^{-1}) = o(n^{-1})$. Denote the m -marker log-likelihood by $\ell^\delta(\gamma)$, while $\ell(\gamma)$ continues to be the log-likelihood with dense markers. The key message here is that asymptotically we have essentially the same information as if we had dense markers. One way to establish this is to show that $\ell^\delta(\gamma)$ is close to $\ell(\gamma)$ asymptotically (3). An alternative treatment which avoids technical details is taken here. In place of $\ell^\delta(\gamma)$, we construct a pseudo-log-likelihood $\ell^{\delta*}(\gamma)$ which can also be computed using only the marker data. The latter can be interpreted as an estimate of the unobserved $\ell(\gamma)$.

Let γ_1 and $\gamma_1 + \delta$ be any two adjacent marker locations. For individual i , $\ell_i(\gamma_1) = \ell_i^\delta(\gamma_1)$ and $\ell_i(\gamma_1 + \delta) = \ell_i^\delta(\gamma_1 + \delta)$.

Definition: For γ between two marker locations γ_1 and $\gamma_1 + \delta$, let

$$\ell_i^{\delta*}(\gamma) = \begin{cases} \ell_i^\delta(\gamma_1), & \gamma_1 \leq \gamma < \gamma_1 + \delta/2 \\ \ell_i^\delta(\gamma_1 + \delta), & \gamma_1 + \delta/2 \leq \gamma \leq \gamma_1 + \delta. \end{cases}$$

Definition: For sample size n and some fixed $T > 0$, define, for $-T \leq t \leq T$,

$$\zeta_n(t) = \ell\left(\gamma^* + \frac{t}{n}\right), \quad \zeta_n^{\delta*}(t) = \ell^{\delta*}\left(\gamma^* + \frac{t}{n}\right).$$

RESULT 5. Let ρ be the Skorohod (J_1) metric. With $\delta = o(n^{-1})$, as $n \rightarrow \infty$,

$$P\left(\rho(\xi_n(\gamma), \xi_n^{\delta*}(\gamma)) \leq \frac{n\delta}{2}\right) \rightarrow 1.$$

Proof: Consider the event that every marker interval within the window $[\gamma^* - (T/n), \gamma^* + (T/n)]$ has either 0 or 1 crossover aggregating over all individuals. Noting that the aggregated crossover points follow a Poisson process with intensity n , that the marker intervals have width $\delta = o(n^{-1})$, and that there are $2T/n\delta$ marker intervals within the window, the probability of this event is

$$[e^{-n\delta}(1+n\delta)]^{2T/n\delta} = e^{-2T(1+n\delta)} \rightarrow 1 \quad (\text{as } n \rightarrow \infty).$$

We proceed by conditioning on this event. For an individual i and a marker interval $[\gamma_1, \gamma_1 + \delta]$, first consider the case where $x_i(\gamma_1) = x_i(\gamma_1 + \delta)$. Conditioned on no multiple crossovers within marker intervals, this implies that $\ell_i(\gamma) = \ell_i(\gamma_1) = \ell_i^{\delta}(\gamma_1) = \ell_i(\gamma_1 + \delta) = \ell_i^{\delta}(\gamma_1 + \delta)$ for all γ in the interval $[\gamma_1, \gamma_1 + \delta]$, and so $\ell_i^{\delta*}(\gamma)$ equals all of these in the interval also. Hence, $\ell(\gamma) = \ell^{\delta*}(\gamma)$ for all marker intervals where there are no crossovers for all individuals. Now, turn to the case where $x_i(\gamma_1) \neq x_i(\gamma_1 + \delta)$. Conditioned on no multiple crossovers, there is exactly one crossover. Hence $\ell_i(\gamma)$ jumps from $\ell_i(\gamma_1)$ to $\ell_i(\gamma_1 + \delta)$ at the crossover point. $\ell_i^{\delta*}(\gamma)$ will also jump from $\ell_i(\gamma_1)$ to $\ell_i(\gamma_1 + \delta)$, but the jump location is the midpoint of the interval. The absolute difference $|\ell_i(\gamma) - \ell_i^{\delta*}(\gamma)|$ is the magnitude of the jump for γ in the interval between the real crossover point and the artificially constructed jump point. Hence the distance between $\ell_i(\gamma)$ and $\ell_i^{\delta*}(\gamma)$ will not be small under the uniform metric. This is why we need the Skorohod J_1 metric, which allows small perturbations of the time scale to match up two functions (4). In the entire window $[\gamma^* - (T/n), \gamma^* + (T/n)]$, $\ell_i(\gamma)$ and $\ell_i^{\delta*}(\gamma)$ are identical except at the jump locations. The difference of any two corresponding jump locations is smaller than $\delta/2$, and so the Skorohod distance between $\ell_i(\gamma)$ and $\ell_i^{\delta*}(\gamma)$ within the window is smaller than $\delta/2$. On the scale of $\xi_n(t)$ and $\xi_n^{\delta*}(t)$ the Skorohod distance is bounded by $n\delta/2$. This still goes to zero, as $\delta = o(n^{-1})$. Because of the assumption that at each marker interval at most one person has a jump, the result extends to the overall log-likelihood $\ell(\gamma)$ and $\ell^{\delta*}(\gamma)$. ■

3.3. The Case with $\delta = cn^{-s}$, $0 \leq s < 1$. In this subsection, the number of markers is assumed to be fixed or increasing at a rate slower than n . For simplicity, we assume $m = an^s$, where a is some constant and $0 \leq s < 1$. This implies that $\delta = cn^{-s}$ for some constant c . For all n , let γ_1 and $\gamma_1 + \delta$ denote the two adjacent markers flanking γ^* . It is assumed that $\gamma^* = \gamma_1 + \alpha\delta$ for some α which is strictly between 0 and 1 and does not depend on n .

When m is increasing at a slower rate than n , then for large n the log-likelihood of any location outside of the marker interval $[\gamma_1, \gamma_1 + \delta]$ will become negligible relative to locations within the interval. (This can be shown by using KL distances.) The expected number of individuals who have recombinations between γ_1 and $\gamma_1 + \delta$ is $n\delta = cn^{1-s}$, which goes to infinity as n goes to infinity. It will be demonstrated that, for $s > 0$, it is these recombination cases which provide extra resolution within the marker interval. Indeed, if n_r is the number of recombinations between γ_1 and $\gamma_1 + \delta$ and the interval $[\gamma_1, \gamma_1 + \delta]$ is always rescaled to have width 1, then the likelihood function is concentrated in a neighborhood around γ^* with width of order $n_r^{-1/2}$. Essentially, with appropriate scaling and taking n_r as the effective sample size, standard asymptotics apply. Since $n_r = O_p(n^{1-s})$ and the width of the marker interval is actually $\delta = O(n^{-s})$, the convergence is of order $n^{-(1-s)/2}n^{-s} = n^{-(1+s)/2}$.

Within the marker interval $[\gamma_1, \gamma_1 + \delta]$, the log-likelihood $\ell^{\delta}(\gamma)$ is infinitely differentiable and hence its asymptotic behavior can be understood through its derivatives at γ^* .

RESULT 6. If $\delta = cn^{-s}$, $0 \leq s < 1$, and the distance between γ^* and the left flanking marker γ_1 is $\alpha\delta$, $0 < \alpha < 1$, then

$$E(\ell^{\delta'}(\gamma^*)) = E\left(\frac{\partial \ell^{\delta}(\gamma)}{\partial \gamma}\bigg|_{\gamma=\gamma^*}\right) = 0.$$

The following scaled Fisher information exists and is greater than zero:

$$I^{\diamond}(\gamma^*) \equiv \lim_{n \rightarrow \infty} \frac{1}{n^{1+s}} E(-\ell^{\delta n}(\gamma^*)) = \lim_{n \rightarrow \infty} \frac{1}{n^s} E(-\ell_1^{\delta n}(\gamma^*)) > 0.$$

When $s > 0$, it reduces to

$$I^{\diamond}(\gamma^*) = \frac{1}{2c} \left[\int \frac{(f_1(y) - f_0(y))^2}{\alpha f_0(y) + (1 - \alpha)f_1(y)} dy + \int \frac{(f_1(y) - f_0(y))^2}{(1 - \alpha)f_0(y) + \alpha f_1(y)} dy \right]. \quad [4]$$

Proof: For each individual i , define an indexing random variable K_i as follows:

$$K_i = \begin{cases} 1, x_i(\gamma_1) = 0, x_i(\gamma_1 + \delta) = 0 & (1 - \theta_{\gamma_1, \gamma_1 + \delta})/2 = \frac{1 - \delta}{2} + o(\delta) \\ 2, x_i(\gamma_1) = 0, x_i(\gamma_1 + \delta) = 1 & (\theta_{\gamma_1, \gamma_1 + \delta})/2 = \frac{\delta}{2} + o(\delta) \\ 3, x_i(\gamma_1) = 1, x_i(\gamma_1 + \delta) = 0 & (\theta_{\gamma_1, \gamma_1 + \delta})/2 = \frac{\delta}{2} + o(\delta) \\ 4, x_i(\gamma_1) = 1, x_i(\gamma_1 + \delta) = 1 & (1 - \theta_{\gamma_1, \gamma_1 + \delta})/2 = \frac{1 - \delta}{2} + o(\delta) \end{cases}$$

Define $p_k^{\delta}(\gamma) = P(x_i(\gamma) = 1 | K_i = k; \delta)$. By symmetry, $p_3^{\delta}(\gamma) = 1 - p_2^{\delta}(\gamma)$ and $p_4^{\delta}(\gamma) = 1 - p_1^{\delta}(\gamma)$. It follows from Eqs. 1 and 2 in Section 2, with $\alpha\delta$ substituted for u , that

$$\begin{aligned} p_1^{\delta}(\gamma^*) &= \alpha(1 - \alpha)\delta^2 + o(\delta^2) & p_2^{\delta}(\gamma^*) &= \alpha + o(\delta) \\ p_1^{\delta'}(\gamma^*) &= (1 - 2\alpha)\delta + o(\delta) & p_2^{\delta'}(\gamma^*) &= \frac{1}{\delta} + O(\delta), \end{aligned} \quad [5]$$

where $p_k^{\delta'}$ is the derivative of p_k^{δ} with respect to γ . After some algebraic manipulations, the Fisher information for an individual (individual 1, say) can be shown to be

$$\begin{aligned} i^{\delta}(\gamma^*) &\equiv E(-\ell_1^{\delta n}(\gamma^*)) = E_K[E(-\ell_1^{\delta n}(\gamma^*) | K_1)] \\ &= \sum_{k=1}^4 P(K_1 = k; \delta) i_k^{\delta}(\gamma^*) = \sum_{k=1}^4 P(K_1 = k; \delta) \\ &\times \left\{ [p_k^{\delta'}(\gamma^*)]^2 \int \frac{[f_1(y) - f_0(y)]^2}{f_0(y)(1 - p_k^{\delta}(\gamma^*)) + f_1(y)p_k^{\delta}(\gamma^*)} dy \right\} \\ &= \sum_{k=1}^4 P(K_1 = k; \delta) \{ [p_k^{\delta'}(\gamma^*)]^2 \mathfrak{Z}_k^{\delta} \}. \end{aligned}$$

where $i_k^{\delta}(\gamma^*) = E(-\ell_1^{\delta n}(\gamma^*) | K_1 = k)$ is the conditional Fisher information given $K_1 = k$. The integrals $\mathfrak{Z}_k^{\delta}$ in the sum all exist for any fixed $\delta > 0$, although the integrals for $k = 1$ and $k = 4$ can be $O(\delta^{-2})$ for some choices of f_0 and f_1 (when their

supports are different). If $s > 0$, δ goes to zero as n goes to ∞ , and by the values of $P(K_1 = k; \delta)$ and the expressions in Eq. 5 it follows that the product $P(K_1 = k; \delta) i_k^{\delta}(\gamma^*)$ is $O(\delta\delta^{-2}) = O(\delta^{-1})$ for $k = 2$ or 3 and is at most $O(\delta^2\delta^{-2}) = O(1)$ for $k = 1$ or 4 . So $i^{\delta}(\gamma^*)$ is $O(\delta^{-1})$ and is dominated by the contributions from $K_1 = 2$ or 3 . Applying this result to all n individuals implies that although only a small fraction ($\approx \delta$) of the individuals recombine between γ_1 and $\gamma_1 + \delta$, they provide most of the information. In particular,

$$I^{\circ}(\gamma^*) = \lim_{n \rightarrow \infty} \frac{i^{\delta}(\gamma^*)}{n^s} = \lim_{n \rightarrow \infty} \frac{\delta}{c} \left[\left(\frac{\delta}{2} + o(\delta) \right) \left(\frac{1}{\delta} + O(\delta) \right)^2 \mathfrak{Z}_2^{\delta} + \left(\frac{\delta}{2} + o(\delta) \right) \left(\frac{1}{\delta} + O(\delta) \right)^2 \mathfrak{Z}_3^{\delta} \right] = \frac{1}{2c} \lim_{n \rightarrow \infty} [\mathfrak{Z}_2^{\delta} + \mathfrak{Z}_3^{\delta}],$$

which is formula 4. ■

By considering a neighborhood around γ^* of the appropriate scale, in this case with width of order $n^{-(1+s)/2}$, all the standard asymptotic results can be obtained. For example, with Result 6, standard techniques can be used to show.

RESULT 7. *The limiting distribution of $\sqrt{n^{1+s}}(\hat{\gamma} - \gamma^*)$, where $\hat{\gamma}$ is the maximum likelihood estimate, is $N(0, 1/I^{\circ}(\gamma^*))$. Moreover, the likelihood is asymptotically normal around $\hat{\gamma}$.*

The scaled Fisher information $I^{\circ}(\gamma^*)$ depends on α . $I^{\circ}(\gamma^*)$ is convex with respect to α and achieves its minimum value at $\alpha = 1/2$. For some choices of f_0 and f_1 , the value of $I^{\circ}(\gamma^*)$ can go to infinity as α approaches 0 or 1. Irregular results can occur when the gene is exactly at a marker location. The rate of convergence can be of order n^{-1} even with a fixed number of markers (3).

3.4. The Case with $\delta = cn^{-1}$. Suppose here that the marker density increases at the same rate as n . For simplicity we assume that γ^* lies exactly at a marker location. It is clear from earlier results that the convergence rate must be of order n^{-1} . We first consider the limit law for the right shrinking window $[\gamma^*, \gamma^* + (T/n)]$ for some fixed $T > 0$.

Let $I_1(\cdot), I_2(\cdot), \dots$ be iid random functions on the interval $[0, c]$, with $I_j(0) = 0$ for all j . Define a random process $M(t)$ on $[0, T]$ as follows. For $0 \leq t \leq c$, $M(t) = I_1(t)$. For $kc < t \leq (k+1)c$, $k = 1, 2, \dots$,

$$M(t) = I_{k+1}(t - kc) + \sum_{j=1}^k I_j(c).$$

We call $M(t)$ an I^4 process (for *independent and identically distributed interval increments*), and we refer to the distribution of $I_j(\cdot)$ as the *interval increment distribution*. We now construct a specific interval increment distribution. Let $G_{(0,1)}$ be the distribution of $[f_1(y)/f_0(y)]$ with y having the density $f_0(y)$, and $G_{(1,0)}$ be the distribution of $[f_0(y)/f_1(y)]$ with y having density $f_1(y)$. Define $G_R = 1/2(G_{(0,1)} + G_{(1,0)})$. Let W be a Poisson random variable with mean c and let $R_1, R_2, \dots$ be iid random variables with distribution G_R . For $0 \leq t \leq c$, let

$$I(t) = \begin{cases} 0, & W = 0 \\ \sum_{w=1}^W \log \left[1 + \left(\frac{t}{c} \right) (R_w - 1) \right], & W > 0 \end{cases} \quad [6]$$

with distribution H . The following result is again stated using the scaled log-likelihood $\xi_n(t) = \ell^{\delta}(\gamma^* + t/n) - \ell(\gamma^*)$.

RESULT 8. *Under the conditions that $f_0(y)$ and $f_1(y)$ are absolutely continuous with respect to each other and $\delta =$*

cn^{-1} , as $n \rightarrow \infty$, $\xi_n(t)$ converges weakly under the uniform metric to an I^4 process with interval increment distribution H .

The limit process in the left window $[\gamma^* - (T/n), \gamma^*]$ is a mirror image of that in the right, and the two are independent. Collectively, we call the two mirror-image I^4 processes the MI^5 process. The proof, which is in ref. 3, is omitted.

Section 4. Discussion

We feel that much has been gained from this mathematical treatment of the problem. A key message which follows from the mathematical results is that the efficient way of gathering information is to have n and m be of comparable sizes, or $n = b/\delta$ for some constant b . The practical problem of choosing b deserves further exploration.

The actual data collection process is often sequential. After the gene is localized to within one or two (neighboring) marker intervals, data on a new marker in the region are gathered to further narrow down the location. With a sequential design, for δ to decrease at a rate of $O(n^{-1})$ about the true location, the number of markers must increase at a rate of only $\log(n)$.

The phenotype distributions are often not known exactly. Suppose f_0 and f_1 are used to compute the likelihood function, while the true phenotype distributions are g_0 and g_1 , respectively. All of our results still hold if the pseudo-KL distances $E_{g_0}[\log(f_0(y) - \log(f_1(y)))]$ and $E_{g_1}[\log(f_1(y) - \log(f_0(y)))]$ are positive. The scaled log-likelihood is still a mirror-image compound Poisson process around γ^* , and the maximum likelihood estimate of the gene location is still consistent and converges at a rate of $O_p(n^{-1})$. At worst, the constant of the convergence rate is affected. The problem, however, is that posterior probability intervals can be too wide or too narrow. The situation is different for $m = o(n)$ [or $\delta > O(n^{-1})$]. Here misspecification of the phenotype distributions can lead to bias within the interval containing the gene. There will be convergence to the correct interval, but the extra resolution within the interval is not dependable.

In some cases, f_0 and f_1 are only specified up to a parametric form with (nuisance) parameter β . The *profile likelihood* $L_p(\gamma) = \max_{\beta} L(\gamma, \beta)$ is often used in place of the likelihood. Because the entire sample is informative for estimating β , we believe that if the parametric form is specified correctly, one will often get the same convergence as if β were known.

For a polygenic trait (1), one strategy is to search for one gene at a time. If the multiple genes are on different chromosome pairs, all the results still apply with the understanding that the distributions f_0 and f_1 incorporate the extra variation caused by genes which lie on chromosome pairs other than the one pair under study. Further work is necessary to understand the situation when two or more genes are on the same chromosome pair.

We thank the referees for valuable suggestions. This research was supported in part by National Institutes of Health Grant R01-GM46800.

1. Lander, E. S. & Botstein, D. (1989) *Genetics* **121**, 185–199.
2. Feller, W. (1971) *An Introduction to Probability Theory and Its Applications* (Wiley, New York), 2nd Ed., Vol. 2.
3. Kong, A. & Wright, F. (1993) *Department of Statistics Technical Report 369* (Univ. of Chicago, Chicago).
4. Billingsley, P. (1968) *Convergence of Probability Measures* (Wiley, New York).