Am. J. Hum. Genet. 56:1212–1223, 1995

High-Resolution Genetic Mapping of Complex Traits

Leonid Kruglyak[1] and Eric S. Lander[1,2]

[1]Whitehead Institute for Biomedical Research, and [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA

Summary

Positional cloning requires high-resolution genetic mapping. To plan a positional cloning project, one needs to know how many informative meioses will be required to narrow the search for a disease gene to an acceptably small region. For a simple Mendelian trait studied with linkage analysis, the answer is straightforward. In this paper, we address the situation of a complex trait studied with affected-relative-pair methods. We derive mathematical formulas for the size of an appropriate confidence region, as a function of the relative risk attributable to the gene. Using these results, we provide graphs showing the number of relative pairs required to narrow the gene hunt to an interval of a given size. For example, we show that localizing a gene to 1 cM requires a median of 200 sib pairs for a locus causing a fivefold increased risk to an offspring and 700 sib pairs for a locus causing a twofold increased risk. We discuss the implications of these results for the positional cloning of genes underlying complex traits.

Introduction

Positional cloning of disease genes depends on high-resolution mapping. With current technology, a gene must be localized to <1 cM—corresponding to ~1 million bp of DNA—before it becomes practical to systematically scour the region to identify it. Occasionally, the fortuitous presence of a smoking gun—such as a chromosomal deletion or a trinucleotide repeat expansion—may allow one to pinpoint the position of the gene. In general, however, localization depends on genetic mapping. Before undertaking a positional cloning project, it is thus essential to ask: How many meioses will be required to achieve a required degree of genetic resolution?

For a simple Mendelian trait, the problem is straightforward. Since the trait must show perfect cosegregation with the gene, even a single crossover is enough to exclude a region from consideration. The critical interval containing the gene is thus defined by the closest crossover on either side. The size of this critical interval is easily calculated in terms of the crossover rate and the number of informative meioses studied.

In this paper, we explore the analogous problem for complex traits. Complex traits are those that do not show perfect cosegregation with any single locus—owing to such problems as incomplete penetrance, phenocopy, genetic heterogeneity, and polygenic inheritance (Lander and Schork 1994). Individuals carrying a susceptibility allele may have a higher relative risk of disease, but some carriers may be unaffected and some noncarriers may be affected. The lack of a perfect correspondence between genotype and phenotype complicates the task of genetic mapping. Although there are many techniques for the genetic dissection of complex traits (Lander and Schork 1994), allele-sharing methods offer a particularly robust approach.

Allele-sharing methods are based on the notion that a predisposing locus for a complex trait can be recognized by virtue of the fact that a pair of affected relatives will tend to have inherited the same allele more often than expected under random Mendelian segregation. Below, we will consider a collection of affected relative pairs of a fixed type R—e.g., grandparent-grandchild, half sibs, sibs, first cousins, etc. Each pair shares either 0 or 1 allele identical by descent (IBD) at any locus L, and the allele-sharing proportion z_L is defined to be the proportion of affected relative pairs that share an allele IBD at L. (The one exception is sib pairs, which can share 0, 1, or 2 alleles IBD. However, sib pairs may be regarded as two half sib pairs—corresponding to the paternal and maternal chromosome pair, respectively. The allele-sharing proportion z_L for sibs is defined as $z_L = (z_1 + 2z_2)/2$, where z_1 and z_2 are the proportions of affected sib pairs sharing 1 and 2 alleles IBD, respectively. See the appendix, part A for details.) Letting α $(=\alpha_R)$ denote the expected value of the allele-sharing statistic under random Mendelian segregation, the condition $z_L > \alpha$ suggests that there is a predisposing locus at or near L. As we discuss below, the allele-sharing proportion z_L can be related to the relative risk λ for the trait, under certain hypotheses.

In this paper, we consider the following problem. Suppose that affected-relative-pair studies show convincing evidence of a predisposing locus L in a given region. How precisely can one localize L? Whereas a simple

Received November 17, 1994; accepted for publication February 7, 1994.

Address for correspondence and reprints: Dr. Eric S. Lander, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142-1479. E-mail: lander@genome.wi.mit.edu

0002-9297/95/5605-0026$02.00

Mendelian trait can be localized by a single crossover, the allele-sharing proportion can fluctuate considerably and, in any given sample, may not attain its maximum exactly at L. One cannot define a critical region that is *certain* to contain L, but only a confidence region having a given probability (e.g., 95% or 99%) of including L. Our problem thus becomes: How large is a confidence region for L?

Because our focus is on high-resolution mapping rather than initial detection of linkage, we will make two assumptions: (1) The number N of affected relative pairs is large. The assumption of large sample size permits asymptotic approximations that greatly simplify the formal analysis while still providing a good description of the cases of interest. (2) A perfect genetic map of the region is available. Mathematically, this means an infinitely dense map consisting of infinitely polymorphic markers. In practice, the genetic map should be dense enough that IBD can be recognized unambiguously and that crossovers can be localized with arbitrary precision. These two assumptions accurately model the terminal phases of a positional cloning effort: One would not undertake positional cloning without a relatively large number of meioses and a relatively dense genetic map. Our results thus reflect the maximum attainable resolution. We also discuss below the consequence of relaxing these assumptions.

Given these assumptions, we determine the exact distribution of the size of a confidence region for L. The size turns out to follow the Gamma distribution $\Gamma(S,\nu)$, where S is a random variable (depending on z_L) that describes the number of transitions between allele sharing and nonsharing needed to exit the confidence region and ν is the rate of these transitions. The analysis is based on the theory of random walks. Using these results, we present graphs showing the number of affected relative pairs needed to narrow the confidence region to a given size. We discuss the implications of these results for the positional cloning of genes underlying complex traits.

The paper is organized as follows. We begin with the mathematical results on genetic resolution—first for a simple trait and then for a complex trait. Proofs of the results, along with some caveats, are deferred to the appendix. We then present the results in graphical form and discuss their implications for positional cloning. The readers interested only in the graphs and practical implications are invited to skip directly to those sections.

Simple Traits

For a simple Mendelian trait, the gene must lie within the critical interval defined by the closest flanking crossovers on either side. With a perfect genetic map, the LOD score is constant and positive in the critical interval and drops to $-\infty$ beyond the closest flanking crossovers

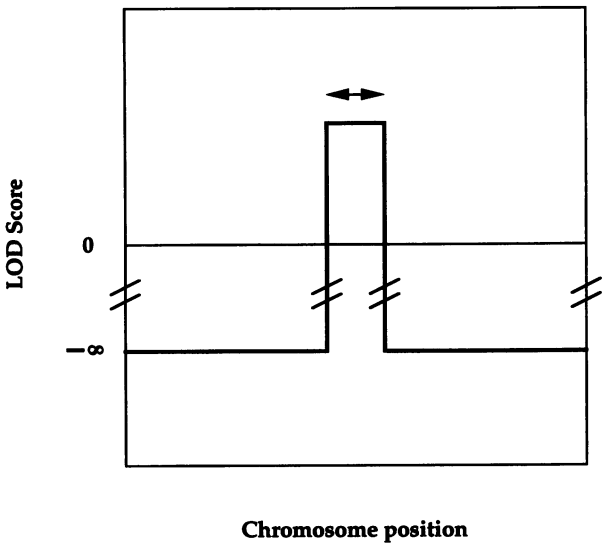

Figure 1 LOD score behavior for a simple Mendelian trait. The LOD score peaks in the region containing the gene (denoted by the arrow) and then drops to $-\infty$ at the first crossover on each side.

(see fig. 1). (With less dense maps, LOD score plots show numerous peaks because double crossovers cannot be excluded with certainty. However, as noted above, the case of arbitrarily dense maps is most relevant to the end stages of positional cloning.) The resolution of genetic mapping depends only on the distance from the gene to these flanking crossovers, which is straightforward to analyze.

In any given meiosis, nearby crossovers are distributed randomly with respect to genetic distance (although not necessarily with respect to physical distance) with a rate of ρ = one crossover per Morgan. In a collection of N meioses, crossovers are randomly distributed with a rate of $N\rho$ per Morgan. It follows immediately that the distance to the closest flanking crossover on one side is exponentially distributed with a mean of $1/N\rho$ Morgans. The size of the critical interval is the sum of two such exponentially distributed variables, corresponding to the distance to the nearest crossover on each side. The mean size of critical interval is thus $2/N\rho$. To narrow the critical interval to an expected size of 1 cM, some 200 informative meioses are required.

Although the results are completely straightforward, we record them as proposition 1 to facilitate comparison with the situation of complex traits developed in proposition 3 below.

PROPOSITION 1. For a simple Mendelian trait studied with linkage analysis, suppose that the maximum LOD score occurs at point x. Let ρ be defined as above and let $\nu = N\rho$. Then:

(*a*) The critical region C is defined as $[x_1, x_2]$, where x_1 and x_2 are the closest points to the left and right of x at which the LOD score drops to $-\infty$.

(*b*) The points x_1 and x_2 are the locations of the closest crossover on the left and on the right, respectively.
(*c*) The size of the critical region is distributed as the sum of two independent exponentially distributed variables, corresponding to the distances to the left and right. Thus, the total size is distributed as $\Gamma(2,\nu)$.

In the proposition, $\Gamma(2,\nu)$ is the well-known Gamma distribution (see the appendix, part B). Essentially the same result was stated by Lange et al (1985). The situation of simple Mendelian traits has been explored in further detail by Boehnke (1994).

Complex Traits

We now turn to the situation of a complex trait. Suppose that one studies a collection of N affected relative pairs of a particular type R using allele-sharing methods and finds that the number of pairs sharing an allele IBD at a locus x is $N_s(x)$. The allele-sharing proportion at position x thus has an observed value of $z \equiv z(x) = N_s(x)/N$. As above, α will denote the proportion of allele sharing expected under Mendelian segregation. The evidence for linkage at locus x can then be summarized in terms of a LOD score (Risch 1990*b*):

$$\mathrm{LOD}(x) = N_s(x)\log\frac{z(x)}{\alpha} + [N - N_s(x)]\log\frac{1 - z(x)}{1 - \alpha}.$$

The allele-sharing statistic—and consequently the LOD score—fluctuates along the length of a chromosome, as individual relative pairs undergo "transitions" from sharing to nonsharing and vice versa (fig. 2). We refer to "transitions" rather than "crossovers" because not all crossovers change the sharing status. This point is illustrated in figure 3. Crossovers always produce transitions in the case of grandparent-grandchild or half sib pairs, but not for uncle-nephew pairs or cousin pairs. The density of ρ_+ of "up-transitions" from nonsharing to sharing and the density ρ_- of "down-transitions" from sharing to nonsharing depend on the type of relative pair, as shown in table 1.

Suppose that a region shows a high degree of allele sharing. Let x^* denote the location at which the allele-sharing statistic attains its maximal value z^*, yielding the maximal LOD score Z^*. Moreover, suppose that the LOD score is far above the minimum threshold for statistical significance (see, e.g., Feingold et al. 1993; Lander and Schork 1994), so that the evidence is unambiguous that the region contains a predisposing locus L.

Because the LOD score never drops to $-\infty$, a critical region cannot be defined as in the case of a simple Mendelian trait. Instead, resolution is a statistical matter. A positional cloner can at most hope to know a confidence region C_γ having probability γ of containing the true locus L. Two questions arise: What is an appropriate definition for C_γ? What is the likely size of C_γ?

Definition of Confidence Region, C_γ

The first question has been extensively studied in the statistical literature (Siegmund 1988). One approach is to define a confidence region C_γ that contains all points at which the LOD is within an appropriately chosen threshold of the maximum (Feingold et al. 1993). The problem is that such a confidence region may not be connected. The LOD score can drop below threshold and then rise above threshold, producing a "hole" in the confidence region (see fig. 2, *right panel*). In positional cloning, we would consider it prudent not to exclude such a "hole" from consideration as a possible location for the disease locus.

A more natural approach for positional cloning is to define the confidence region C_γ to be the *smallest interval* containing all points at which the LOD score is within threshold of the maximum. In the appendix, part E, we show that an appropriate threshold T_γ is defined by the equation $T_\gamma = -\log_{10}[(1 - \gamma)/2] + \log_{10}(1 - \pi)$, where

$$\pi = \frac{(1 - z_L)\alpha}{(1 - z_L)\alpha + z_L(1 - \alpha)}.$$

(As we note below, π is the probability of an upward step in a random walk.) For a 95% confidence region, the threshold is $T_{.95} = 1.6$ when $\pi = 1$ (high excess sharing) and $T_{.95} = 1.3$ when $\pi = \frac{1}{2}$ (low excess sharing). The corresponding values for a 99% confidence region are $T_{.99} = 2.3$ and $T_{.99} = 2.0$.

Size of Confidence Region C_γ

The second question requires determining the distribution function of the size of the confidence region C_γ. To do this, we exploit the fact that the behavior of the LOD score can be described in terms of random walks. Specifically, one has the following proposition:

PROPOSITION 2. Consider a predisposing locus L with true allele-sharing proportion $z_L > \alpha$, studied in a large number N of affected relative pairs. In the neighborhood of L, the LOD score follows a random walk that starts at L and proceeds outward in both directions. The walk has constant step size δ and downward drift away from L, with probability of upward and downward steps being π and $1 - \pi$, respectively. Here, $\delta = \log_{10}(z_L/\alpha) - \log_{10}[(1 - z_L)/(1 - \alpha)]$, and π is defined above. (Note that $\pi/(1 - \pi) = (1 - z_L)\alpha/z_L(1 - \alpha) = 10^{-\delta}$.) See the appendix, part C for a justification; the assumption of large N guarantees that the observed allele-sharing proportion does not deviate significantly from the true value z_L in the neighborhood of L.

How many total transitions S will be needed for the LOD score to drop permanently below $Z^* - T_\gamma$? Since each step changes the LOD score by $\pm\delta$, a drop of T_γ

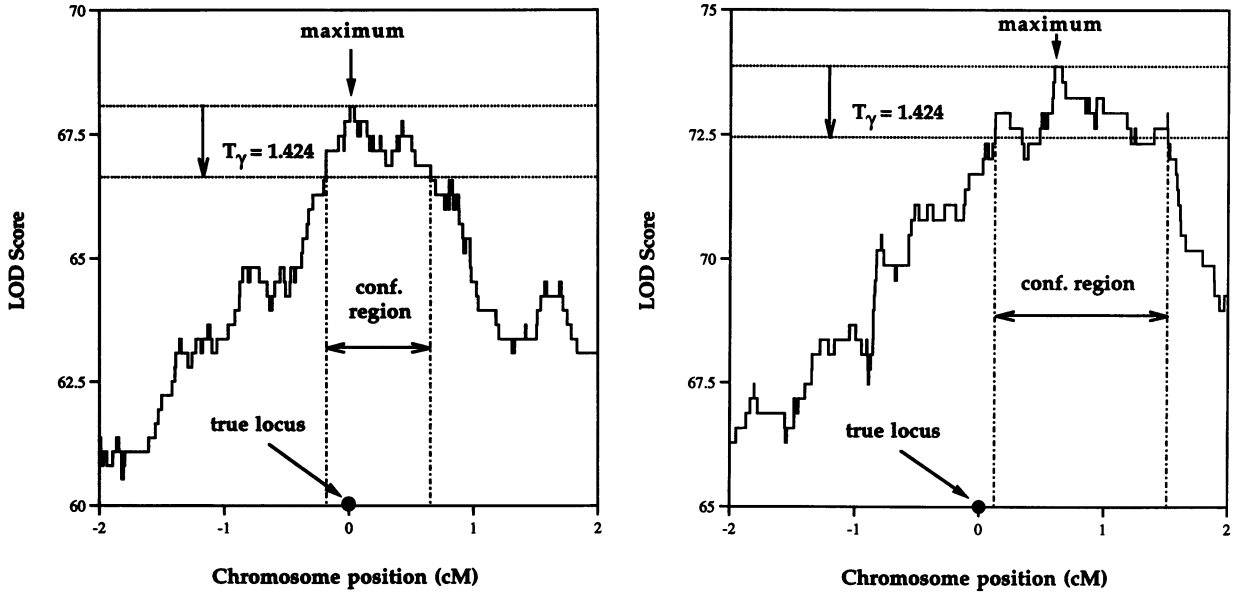

Figure 2 LOD score behavior for a complex trait. The LOD score follows a random walk in the neighborhood of its peak, with steps occurring at transitions between sharing and nonsharing. Distances between transitions are exponentially distributed. Unlike the simple trait case, the LOD score never drops off to $-\infty$. The trait locus is shown at the origin. The maximum LOD score, the drop in LOD score for the 95% confidence region, and the extent of the confidence region are indicated. Two cases are shown: *Left,* The true locus lies within the confidence region. *Right,* The true locus lies outside the confidence region. Note the "holes" in the confidence region where the LOD score drops below the threshold T_γ.

corresponds to a *net* of $\Delta = \lceil T_\alpha/\delta \rceil$ downward transitions, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. It is not hard to see that our problem is equivalent to the question of when the doubly infinite random walk described in proposition 2 makes its last crossing to below the level $Z^* - T_\gamma$ on either side of its maximum. Let the random variable $N_{\pi,\Delta}$ denote the number of steps between such last crossings to the left and to the right of the maximum. The number S of transitions occurring between the ends of the confidence region C_γ is then distributed as $N_{\pi,\Delta}$. The exact distribution of $N_{\pi,\Delta}$ can be computed by using results from the theory of random walks (see the appendix, part D).

It remains only to describe the corresponding distance, given the number S of transitions. In the neighborhood of L with allele-sharing proportion z_L, a collection of N affected relative pairs will have new upward transitions arriving randomly with density $\rho_+(1 - z_L)N$ per Morgan and new downward transitions arriving randomly with density $\rho_- z_L N$ per Morgan (see table 1; see also Feingold 1993). It is not difficult to show that total transitions (up and down) arrive randomly with a rate of $\nu = (\rho_+(1 - z_L) + \rho_- z_L)N$ per Morgan, and that the distance for the arrival of S transitions is distributed as the sum of S exponential random variables with mean $1/\nu$—that is, as a random variable with distribution $\Gamma(S,\nu)$ (see the appendix, part F). An explicit formula for $\Gamma(S,\nu)$ is given in the appendix, part B.

We summarize the results in the following proposition, which is phrased to facilitate comparison with proposition 1 for simple Mendelian traits above:

PROPOSITION 3. For a complex trait studied with allele-sharing methods using a large number N of affected relative pairs of type R, suppose that the maximum LOD score occurs at point x. Let z, α, Δ, and ν be defined as above. Then:

(*a*) A confidence region C_γ is defined by $[x_1, x_2]$, where x_1 and x_2 are the leftmost and rightmost points, respectively, at which the LOD score is within T_γ of its maximum. The threshold T_γ is given by $T_\gamma = -\log_{10}[(1 - \gamma)/2] + \log_{10}(1 - \pi)$.

(*b*) The number S of transitions occurring between x_1 and x_2 is distributed as the random variable $N_{\pi,\Delta}$.

(*c*) The size of the confidence region (the distance between x_1 and x_2) is distributed as $\Gamma(S,\nu)$, where S is distributed as in (*b*).

This result provides an exact mathematical description of the confidence region for a complex trait locus, for the case of large N. The distribution of the random variables in (*b*) is given in the appendix, part D. The distance distribution in (*c*) is the Gamma distribution with a random parameter; its cumulative distribution function is straightforward to compute from this description (see the appendix, part B). Finally, we note that relaxing the assumption of large N has only a minor effect on these results (see the appendix, part G) and that relaxing the assumption of a dense genetic map

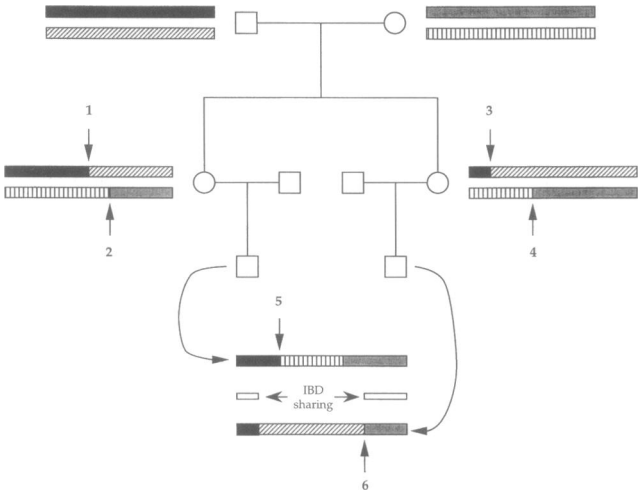

Figure 3 An illustration of the relationship between transitions and crossovers for the case of first cousins. Transmission of the four grandparental chromosomes (distinguished by shading) is shown; the chromosomes of the two unrelated fathers are irrelevant and are omitted. The outcomes of six meioses are relevant to the sharing status of the cousins. In the example shown, a crossover has occurred in each of these meioses; these crossovers are numbered 1–6. The maternally derived chromosomes of the two cousins are compared at the bottom; the regions of IBD sharing are indicated by white rectangles. The state of the cousin pair at every point in the genome can be described by the outcomes of the six meioses, i.e., by whether the maternally derived or the paternally derived chromosome was transmitted in each meiosis. Every crossover changes the outcome of a meiosis and hence results in a change of state. However, in this example only crossovers 3 and 6 lead to transitions between sharing and nonsharing. Crossovers 1 and 4 have no effect on sharing because they occur in chromosomal regions that are not passed on. Crossovers 2 and 5 also do not change the sharing status, because they merely lead to switches between different modes of nonsharing. (Figure adapted from Feingold 1993; additional details may be found in this reference.)

does not fundamentally change the analysis (see the appendix, part H).

Numerical Results: Allele-Sharing Proportion

One can directly calculate the properties of a 95% confidence region $C_{.95}$, based on proposition 3. Specifically, one can derive the distribution of: (i) the number of transitions between the endpoints of the region and (ii) the number of relative pairs needed to narrow the region to 1 cM. We will focus first on the situation in which initial linkage has already been detected, by virtue of the presence of excess allele sharing. In this case, the allele-sharing proportion, z_L, can be assumed to be known (or, at least, well estimated) based on the preliminary linkage data. For example, Berrettini et al. (1994) report that sib pairs affected with bipolar affective disorder show 58% IBD sharing for marker D18S21.

Figure 4, *top left* describes the confidence region in terms of the number of transitions between allele sharing and nonsharing, for the case of relatives with $\alpha_R = 1/2$ (e.g., grandparent-grandchild, half sib, sib, and uncle-nephew pairs). The graph shows the minimum, the median, and the 95% upper confidence limit on the number of transitions between the ends of $C_{.95}$. Let us focus on the median numbers. For $z > .975$, the confidence region is two transitions wide; that is, a single transition on each side is sufficient to eliminate a region from consideration—just as in the case of a simple trait. For $z > .855$, the median size of the region increases to four transitions. The number of transitions then grows rapidly: 14 for $z = .75$; 28 for $z = .67$; 76 for $z = .60$; and 368 for $z = .55$. In short, high-resolution mapping becomes more difficult as z approaches $\alpha_R = .50$. Since these are medians, there is a 50% chance that the required number of transitions will be higher. The upper 95% confidence limit is roughly twofold higher.

By calculating the size of the confidence region as a function of the number of affected relative pairs, one can determined the number of pairs needed to narrow $C_{.95}$ to a given size. Figure 4, *top right* shows the number of pairs required so that Prob[size($C_{.95}$) $\leq$ 1 cM] is 50% or 95%. For $z > .975$, a total of 170 meioses is required to achieve a median size of 1 cM, just as in the case of a simple trait. The 170 meioses could be 170 grandparent-grandchild pairs, 85 half sib pairs, 43 sib pairs or 68 uncle-nephew pairs. The sample size roughly doubles to ~400 meioses in the range $z = .855–.975$. For smaller z's the sample size grows dramatically: 1,500 meioses for $z = .75$; 2,800 meioses for $z = .67$; 7,600 meioses for $z = .60$; and 37,000 meioses for $z = .55$. Although these numbers pertain to narrowing the confidence region to 1 cM, the corresponding numbers for decreasing the region to ($1/k$) cM are simply k-fold larger. Figure 4, *bottom left* and *bottom right* shows the corresponding graphs for the case of relative pairs with $\alpha_R = 1/4$ (e.g., first cousins).

Table 1

Parameters for Various Relative Pairs

Relative Type, R	α_R	ρ_+	ρ_-
Grandparent-grandchild	1/2	1	1
Half sibs[a]	1/2	2	2
Uncle-nephew	1/2	5/2	5/2
First cousin	1/4	4/3	4
Second cousin	1/16	2/5	6

NOTE.—The parameter α_R denotes the expected allele-sharing statistic under random Mendelian segregation, and ρ_+ and ρ_- denote the density per Morgan of upward and downward transitions, respectively, across the genome under random Mendelian segregation. (Note well that the rates ρ_+ and ρ_- satisfy the relation $\rho_+(1 - \alpha) = \rho_-\alpha$, reflecting the fact that there is, on average, no net average change in the degree of allele sharing under random Mendelian segregation.)

[a] Sib pairs are treated as equivalent to two half sib pairs, corresponding to sharing on the maternally and paternally derived chromosomes, respectively.

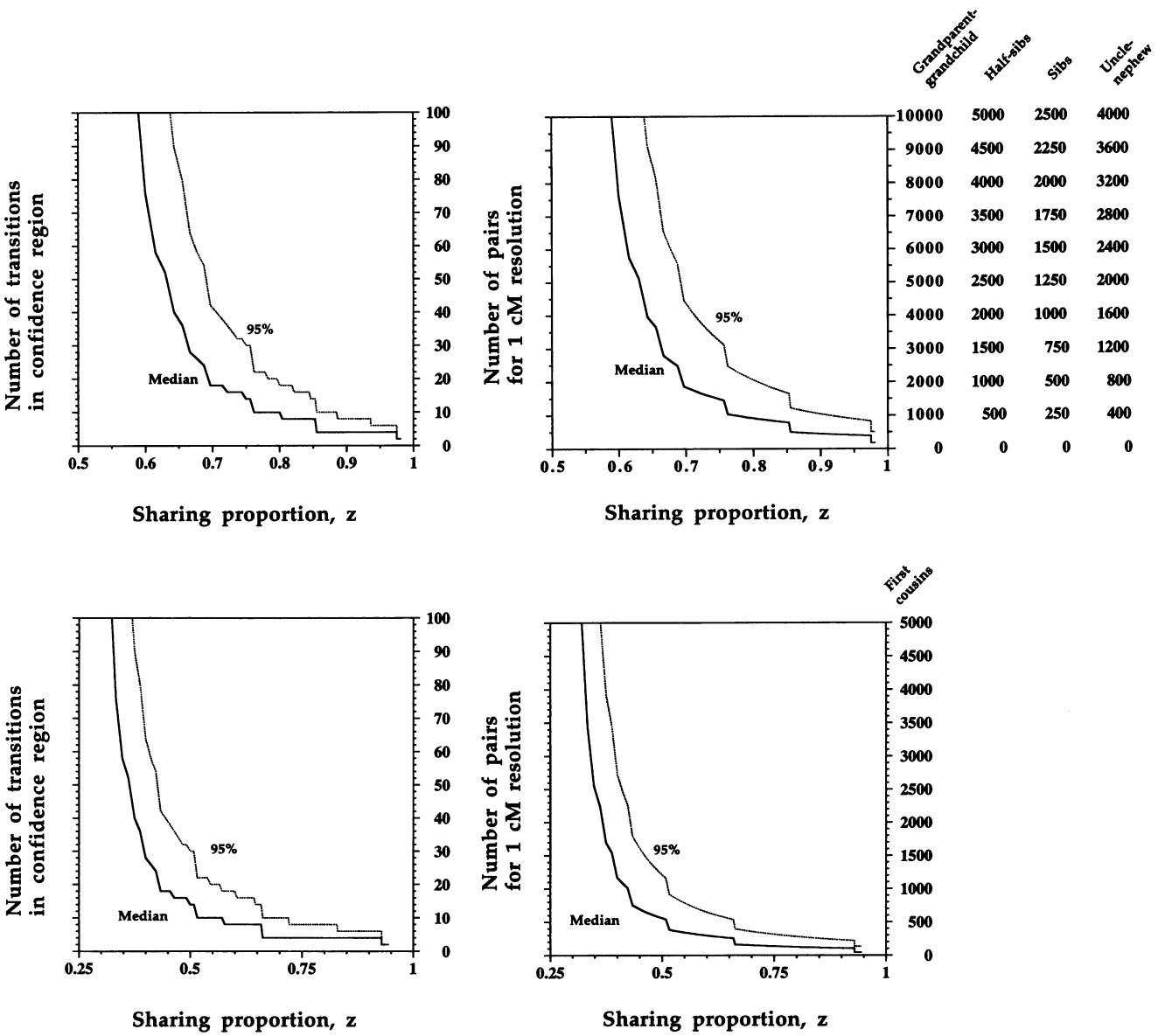

Figure 4 Size of the 95% confidence region, plotted against allele sharing proportion. *Top left,* The number of transitions needed for the LOD score to drop by the threshold T_Y for first-degree relatives (grandparent-grandchild, half sib, sib, or uncle-nephew pairs). Graph shows the median number of transitions and the 95% upper confidence limit on the number of transitions. *Top right,* The number of relative pairs required to narrow the 95% confidence region to 1 cM, for first-degree relatives. Graph shows the median number of relative pairs and the upper 95% confidence limit on the number of relative pairs. *Bottom left,* The number of transitions for second-degree relatives (e.g., first cousins). *Bottom right,* The number of second-degree relative pairs.

Numerical Results: Relative Risk

The results above are expressed in terms of the allele-sharing proportion, z_L, for the susceptibility locus. It would also be useful to express the results in terms of the relative risk, λ_R, of the trait for a relative of type R, defined by:

$$\lambda_R = \frac{\text{Prob}(X \text{ is affected} \mid \text{relative of type } R \text{ is affected})}{\text{Prob}(X \text{ is affected})}.$$

Because λ can be directly estimated by epidemiology, it would be particularly useful to express results in terms of λ for situations in which loci have not yet been mapped.

This can be done exactly in the case of a trait caused by a single susceptibility locus. By a simple application of Bayes's Theorem, one has the equations:

$$\frac{z_L}{\alpha_R} = \frac{\lambda_O}{\lambda_R}$$

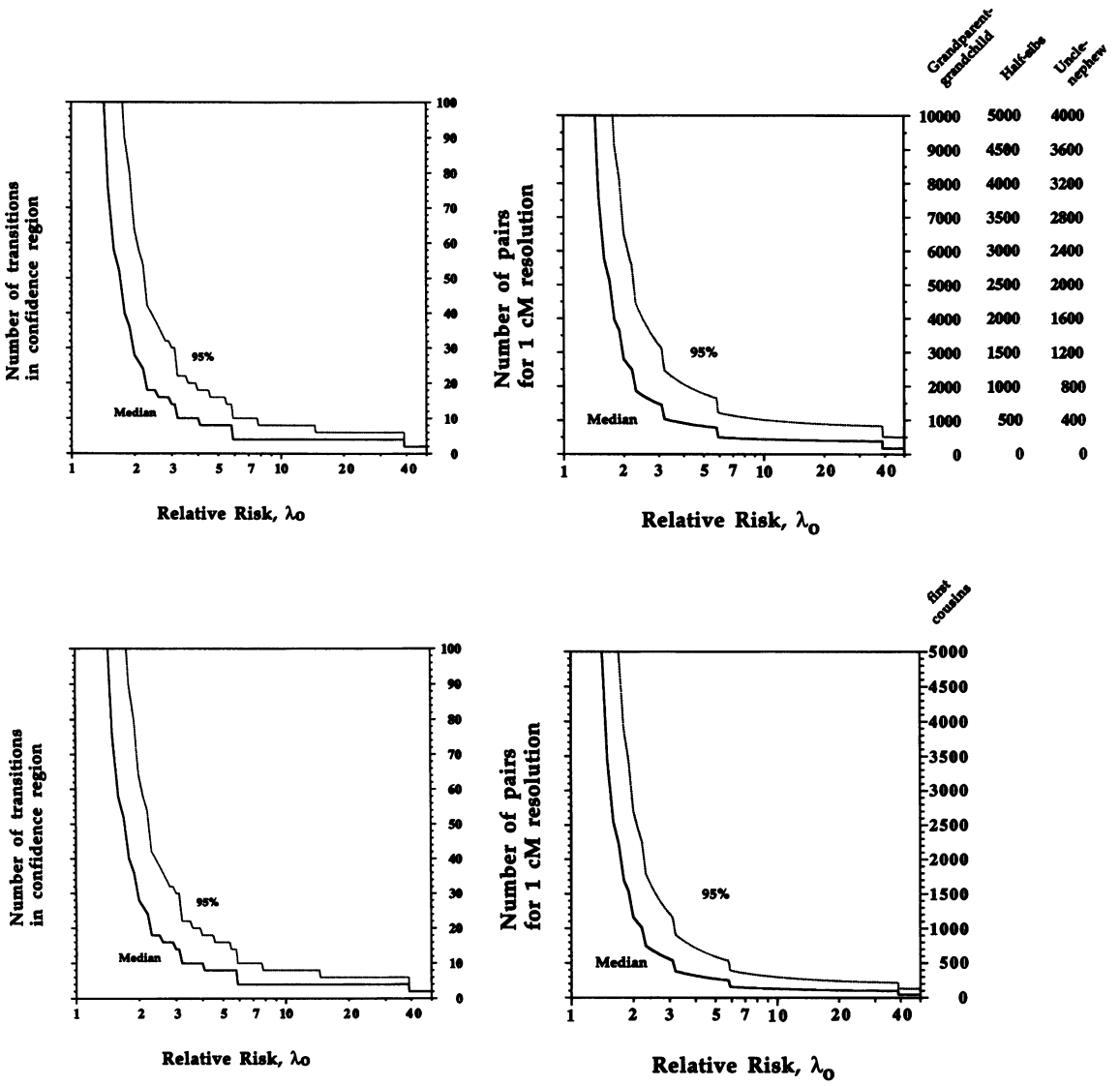

Figure 5 Size of the 95% confidence region, shown as a function of the risk ratio λ_O to an offspring (for the case of a single susceptibility locus). *Top left,* The number of transitions for first-degree relatives (grandparent-grandchild, half-sib, or uncle-nephew pairs). *Top right,* The number of relative pairs for first-degree relatives. *Bottom left,* The number of transitions for second-degree relatives (e.g., first cousins). *Bottom right,* The number of second-degree relative pairs.

and

$$\lambda_R - 1 = \alpha_R(\lambda_O - 1)\,,$$

where λ_O represents the relative risk to offspring (see Risch 1990*b*). It is thus possible to express z_L directly in terms of λ_O, by the equation

$$z_L = \frac{\alpha_R \lambda_O}{1 + \alpha_R(\lambda_O - 1)}\,.$$

The allele-sharing proportions discussed above, z_L = .975, .855, .75, .67, .60 and .55, translate to relative risks of λ_O = 40, 6, 3, 2, 1.5, and 1.2, respectively. Figure 5 shows the corresponding graphs redrawn in terms of λ_O. (In the case of sib pairs, the equation is slightly different: $z_L = (\lambda_M + \lambda_O)/(\lambda_M + 2\lambda_O + 1)$, where λ_M denotes the relative risk to an MZ twin.)

In the case of a trait involving multiple susceptibility loci, there is no absolute dependence between z_L and λ_O or between λ_O and λ_R. The relationship depends on the precise nature of the complex trait—including the number of loci and the epistatic relationship among them. Nonetheless, some useful observations can be made:

(i) Regardless of the genetic details, it is always true that

$$\frac{z_L}{\alpha_R} \leq \frac{\lambda_O}{\lambda_R}\,.$$

(ii) For a trait following a multiplicative model (Risch 1990*a*) involving loci $L_1, L_2, \ldots, L_k$, the sharing proportion z_{iL} at the ith locus is given by

$$z_{iL} = \frac{\alpha_R \lambda_{iO}}{1 + \alpha_R(\lambda_{iO} - 1)},$$

where λ_{iO} is the "risk ratio factor" for the ith locus, with $\lambda_O = \lambda_{1O}\lambda_{2O} \ldots \lambda_{kO}$ (see Risch 1990*a* for a precise definition of λ_{iO}). Note that this is the same relation as for a single locus, but with λ_O replaced by λ_{iO}. The results in figure 5 corresponding to each λ_O thus provide a lower bound on the required number of relative pairs, in the event that multiple loci are involved.

Implications for Positional Cloning

These results have important implications for the design of positional cloning experiments. For loci with high λ_O, the situation closely resembles the case of a simple Mendelian trait. For loci with intermediate values of λ_O in the range of 6–40, the resolution is twofold lower for the same sample size; alternatively, twice the sample is necessary to obtain the same resolution. For loci with low λ_O, <3, the sample size needed for high resolution explodes. For example, a locus with IBD sharing proportion of .58, such as the one reported recently for bipolar affective disorder (Berrettini et al. 1994), would require a median of ~3,000 sib pairs for 1 cM resolution. In such cases, it may be difficult to collect the required number of relative pairs. Alternative strategies may be desirable or necessary to achieve the required resolution. We mention two possibilities:

i) Redefining disease. After detecting initial linkage by allele-sharing methods, it may be possible to identify clinical features (such as early onset of disease) or epidemiological features (such as presence of significant family history beyond the initial affected relative pair) that increase the degree of genetic homogeneity and thereby yield families showing a higher degree of allele sharing in the candidate region (Lander and Schork 1994). This effectively increases λ_O and decreases the number of affected pairs (having the more restricted phenotype) that must be considered.

ii) Fine-structure linkage disequilibrium mapping in isolated populations. If one can identify an isolated human population in which a significant fraction of affected individuals have inherited the predisposing allele from a common ancestor, one can exploit the tremendous power of linkage disequilibrium mapping. In brief, this involves looking for allele sharing not just within each affected relative pair, but across the entire population of affected individuals. By treating all affected individuals as distant relatives, one can exploit not just the meioses in the current generation but all meioses in the history of the population. This strategy has recently been applied to a simple Mendelian disease, diastrophic dysplasia. Although the study of all multiplex affected families only narrowed the gene to ~2 Mb, the study of linkage disequilibrium in Finnish patients pinpointed the gene to ~40 kb and made possible its positional cloning (Hastbacka et al. 1994). Similar studies should be possible for complex traits, although the fraction of affected individuals descending from a single common ancestor would not be expected to be as high.

Conclusion

Genetic mapping of simple Mendelian traits in humans is now straightforward, having been successfully accomplished in >400 cases. Positional cloning of such traits remains a major undertaking but is clearly within reach of today's technology and has been accomplished in ~40 cases. The situation is very different for complex traits, where genetic mapping is quite difficult and positional cloning is uncharted territory. A key issue in the genetic dissection of complex traits is the degree of resolution that can be achieved through genetic mapping. Whereas a single crossover on each side suffices to delimit the region containing simple trait gene, localization of a gene involved in a complex trait may require many crossovers—yielding a much larger region that must be examined.

With the Human Genome Project producing detailed maps showing the location and sequence of all or most genes, it may become practical to systematically scan somewhat larger regions to find genes. Nonetheless, it is essential for the planning of a positional cloning project to know the likely size of the region that will need to be examined. We have addressed this question by computing the size of the region likely to contain a complex-trait gene, as a function of relative risk attributable to the locus. For relative risks >40, a single crossover suffices: such traits are simple when it comes to resolution. For traits with risk ratios in the range of 6–40, the sample size required to achieve the same resolution as in the simple case roughly doubles. For traits with risk ratios <6, the sample size requirements grow rapidly—making positional cloning of such traits increasingly difficult. For traits with very low risk ratios, other strategies will probably be needed to achieve the required resolution. These calculations provide explicit guidelines for the design and evaluation of studies to clone genes underlying complex traits.

Acknowledgments

We thank Carl Rosenberg for useful discussions and two anonymous referees for helpful comments on the manuscript.

This work was supported in part by National Institutes of Health grants HG00098 to E.S.L. and HG00017 to L.K.

Appendix

A. The Case of Sib Pairs

Since sibs may share 0, 1, or 2 alleles IBD, a suitable and commonly used statistic is $z = \frac{1}{2}(z_1 + 2z_2)$, where z_1 is the number of pairs sharing one allele and z_2 is the number sharing two alleles. This statistic is simply the overall proportion of alleles shared, which is equal to the average of the sharing statistics for the two half sib pairs defined by the maternal and paternal chromosomes. It is easy to show that z makes transitions up and down at the rate of $\nu = 4N$ per Morgan.

B. Gamma Distribution

The Gamma distribution $\Gamma(k,\nu)$ is defined to be the distribution of a sum of k independent exponentially distributed variables with mean $1/\nu$ and has the probability density function

$$g_{k,\nu}(t) = \frac{\nu^k t^{k-1}}{(k-1)!} e^{-\nu t} .$$

$\Gamma(1,\nu)$ is the exponential distribution and has density $\nu e^{-\nu t}$. Exponential and Gamma distributions are useful for describing the distance between crossovers (assuming no interference). Suppose that crossovers are distributed with a density of ρ per Morgan. The distance to the first crossover lying to the right of a given point is distributed as $\Gamma(1,\rho)$; the distance to the kth crossover is distributed as $\Gamma(k,\rho)$. We will find it useful below to denote by $\Gamma(S,\nu)$ the distribution of the sum of S exponentially distributed variables, where S is itself a random variable with probability $\text{Prob}(S = k) = p(k)$. By this we mean the distribution with probability density function

$$d(t) = \sum_{k=0}^{\infty} p(k) g_{k,\nu}(t) .$$

The cumulative distribution function for this distribution is easily obtained by integrating over t from 0 to x or by noting that the cumulative distribution function of $\Gamma(k,\nu)$ is the incomplete Gamma function $P(k,\nu x)$.

C. Local Random-Walk Description of LOD Score

The local random-walk description of the LOD score is essentially contained in Feingold et al. (1993) and Feingold (1993). In brief, the LOD score across the genome follows a Markov process in the case of grandparent-grandchild and half sib pairs and is locally well approximated (over distances such that the chance of two transitions occurring in the same pair is small) by a Markov process in the case of sib, uncle-nephew and cousin pairs; see Feingold (1993) for a complete discussion of these processes. Consider a predisposing locus L with expected allele-sharing statistic z_L. Treating the observed allele-sharing statistic z_{obs} at L as a nuisance parameter, the LOD score near L behaves as the Markov process conditioned to pass through z_{obs} at L. Because z_{obs} tends to z_L for large n, we can ignore this nuisance parameter. Focusing on a region sufficiently small that the allele-sharing statistic does not change significantly, the Markov process reduces to a simple random walk as described in proposition 2.

D. Formulas for Random Walks

(This section draws on results about random walks that are described in Feller 1970, chapters III and XIV.) Consider a doubly infinite random walk $(\ldots, S_{-2}, S_{-1}, S_0, S_1, S_2, \ldots)$ passing through the origin (i.e., $S_0 = 0$) and having downward drift in both directions from the origin. That is, starting at the origin, the height of the walk changes at each step away from the origin in either direction by $+1$ with probability π and by -1 with probability $1 - \pi$, with $\pi < \frac{1}{2}$. Let Y denote the maximum height reached by the walk (this height may be reached more than once). Let i_1 denote the left step and i_2 denote the right step at which the walk makes its last crossing from the level $Y - \Delta + 1$ to the level $Y - \Delta$ (precisely, $S_i < Y - \Delta + 1$ for $i < i_1$, $S_{i_1} = Y - \Delta$, $S_{i_1+1} = Y - \Delta + 1$, and $S_{i_2-1} = Y - \Delta + 1$, $S_{i_2} = Y - \Delta$, $S_i < Y - \Delta + 1$ for $i > i_2$; note that level Y is attained between i_1 and i_2. For the appropriate choice of Δ, see below). Let the random variable $N_{\pi,\Delta} = i_2 - i_1$ denote the number of steps between i_1 and i_2; we seek to compute the distribution of $N_{\pi,\Delta}$.

Let the random variables y_1, y_2 denote the maximum heights reached in the left and right half-walks, respectively: that is, $y_1 = \max_{(i \leq 0)} S_i$ and $y_2 = \max_{(i \geq 0)} S_i$. (Note that either or both of y_1, y_2 may occur at the origin and that $Y = \max(y_1,y_2)$.) The random variables y_1, y_2 are independent and identically distributed with probability

$$\text{Prob}(y_i = A) = \frac{1 - 2\pi}{1 - \pi} \left(\frac{\pi}{1 - \pi}\right)^A ,$$

where the second term is the probability that a half-walk reaches the value A, and the first term is the probability that once it is reached, it is never exceeded (both terms follow from Feller 1970, chapter XIV, eq. [2.8]).

Let the random variable $C(a_1,a_2)$ denote the distance between x_1 and x_2, conditional on the walk having $y_1 = a_1$ and $y_2 = a_2$. Then,

$$\text{Prob}(N_{\pi,\Delta} = x) = \sum_{a_1=0}^{\infty} \sum_{a_2=0}^{\infty} \text{Prob}(y_1 = a_1)\, \text{Prob}(y_2 = a_2)$$

$$\times \text{Prob}[C(a_1, a_2) = x] .$$

Hence, once we know the distribution of $C(a_1,a_2)$, we can obtain the distribution of $N_{\pi,\Delta}$ by performing the sum.

We start with a lemma. We define a nonnegative integer random variable $T_R(\pi)$ by the distribution

$$\text{Prob}[T_R(\pi) = x] = \frac{R}{x}\binom{x}{\frac{1}{2}(x+R)}\pi^{(x-R)/2}(1-\pi)^{(x+R)/2} .$$

In this formula, the binomial coefficient is understood to be zero if $\frac{1}{2}(x + R)$ is not an integer between 0 and x.

LEMMA. Consider a one-sided random walk with downward drift starting at the origin. Let the random variable V_R denote the time of the first visit to level R (that is, $V_R = k$ if $S_i < R$ for $i < k$ and $S_k = R$), conditional on the walk reaching this level. Let random variable U_R denote the time of the last crossing from the level $-(R - 1)$ to the level $-R$, conditional on the walk being completely nonpositive (that is, $U_R = k$ if $S_k - 1 = -(R - 1)$, $S_k = -R$, $S_i < -(R - 1)$ for $i > k$, and $S_i \leq 0$ for all i). Below, we will refer to U_R as the last exit to $-R$. Then, both V_R and U_R are distributed as $T_R(\pi)$.

PROOF OF LEMMA. The term $(R/x)\binom{x}{\frac{1}{2}(x+R)}$ is the number of paths that first reach level R at step x (Feller 1970, chapter III, eq. [7.5]). To obtain the result for V_R, we first multiply by the probability of such a path, $\pi^{(x+R)/2}(1-\pi)^{(x-R)/2}$, which is the probability of taking a net of R upward steps in x total steps. In order to condition on the walk actually reaching R, we divide by the probability of this occurring, which, as we saw above, is $[\pi/(1-\pi)]^R$.

To obtain the result for U_R, we note that the number of paths that start at the origin, remain negative, and reach level $-R$ at step x is exactly the same as the number of paths that first reach level R at x (this is most easily seen by switching the roles of the origin and the point (x,R)). This number is also the same as the number of paths that start at the origin, remain ≤ 0, and reach level $-(R - 1)$ at step $x - 1$. The probability of such a path is equal to the number of paths multiplied by $\pi^{(x-R)/2}(1-\pi)^{(x+R)/2}$, the probability of taking a net of R downward steps in x total steps. Conditioning on the walk never exceeding 0 involves dividing by $(1 - 2\pi)/(1 - \pi)$ and requiring that the walk never returns to level $-(R - 1)$ after step x involves multiplying by $(1 - 2\pi)/(1 - \pi)$; these operations cancel out and yield the desired result.□

With the lemma in hand, we can return to computing the distribution of $C(a_1,a_2)$. Two cases are of interest (see fig. A1):

Case 1: $|y_1 - y_2| \geq \Delta$.—In this case, the origin is *not* included in the confidence region. $C(a_1,a_2)$ is the sum of two random variables: V, the time of the first visit from level $Y - \Delta + 1$ to level Y (plus one step for the crossing) and U, the time of the last exit to level $Y - \Delta$, starting at level Y. V is distributed as $T_{\Delta-1}(\pi)$, and U is distributed as $T_\Delta(\pi)$. Note that in this case the distribution of $C(a_1,a_2)$ is independent of a_1 and a_2; we can compute it once and weigh by the probability that $|y_1 - y_2| \geq \Delta$, that is, by the probability that the origin is not included in the confidence region (see below).

Case 2: $|y_1 - y_2| < \Delta$.—In this case, the origin *is* included in the confidence region. $C(a_1,a_2)$ is the sum of four random variables: V_1, the first visit to a_1; V_2, the first visit to a_2; U_1, the last exit to $Y - \Delta$, starting at a_1; and U_2, the last exit to $Y - \Delta$, starting at a_2. V_1 is distributed as $T_{a_1}(\pi)$; V_2 is distributed as $T_{a_2}(\pi)$; U_1 is distributed as $T_{\Delta+a_1-Y}(\pi)$; and U_2 is distributed as $T_{\Delta+a_2-Y}(\pi)$ (note that one of the last two terms is just $T_\Delta(\pi)$, since $Y = \max(a_1,a_2)$).

E. Drop in LOD Score Needed for a γ Confidence Region

We saw above that the probability of excluding the origin (the true position of the gene) from the confidence region is just $\text{Prob}(|y_1 - y_2| \geq \Delta)$, where y_1,y_2 are the maxima reached on the left and right sides of the origin, respectively. Using the expression for $\text{Prob}(y_i = A)$ given above, straightforward algebra yields

$$\begin{aligned}\text{Prob}(|y_1 - y_2| \geq \Delta) &= 2\left(\frac{1-2\pi}{1-\pi}\right)^2 \sum_{y_1=0}^{\infty}\sum_{y_2=y_1+\Delta}^{\infty}\left(\frac{\pi}{1-\pi}\right)^{y_1+y_2} \\ &= 2\left(\frac{\pi}{1-\pi}\right)^{\Delta}(1-\pi) .\end{aligned}$$

Setting $\text{Prob}(|y_1 - y_2| \geq \Delta) = 1 - \gamma$ and using the fact that $[\pi/(1 - \pi)] = 10^{-\delta}$ (from proposition 2), we have $1 - \gamma = 2\,(10^{-\delta\Delta})\,(1 - \pi)$. We thus obtain the critical LOD threshold $T_\gamma = \delta\Delta = -\log_{10}(1 - \gamma)/2 + \log_{10}(1 - \pi)$. When T_γ/δ is not an integer, the random walk must drop by $\Delta = T_\alpha/\delta$ steps in order to leave the confidence region. The resulting confidence region then corresponds to a slightly larger γ because of the discreetness of the walk.

F. Transition Arrival Distribution

For a relative pair, sharing versus nonsharing is determined by the outcomes of k meioses in the pedigree: an outcome is 0 if the paternally derived allele is transmitted and 1 if the maternally derived allele is transmitted. Hence, there are 2^k possible states of a pair, described

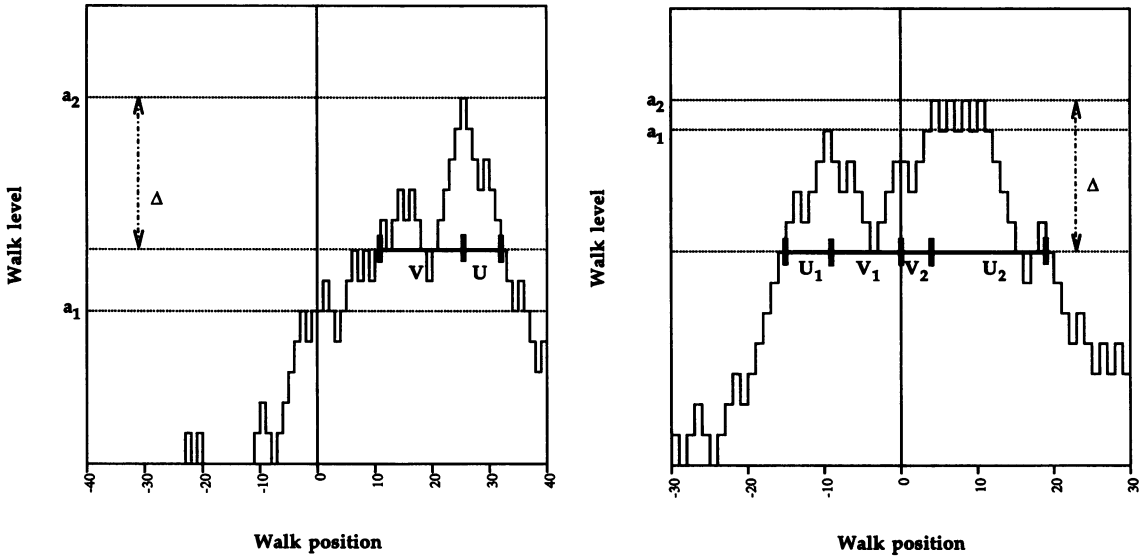

Figure A1 Two cases of the appendix, part D, illustrated. *Left*, $|y_1 - y_2| \geq \Delta$; the origin is not included in the confidence region. *Right*, $|y_1 - y_2| < \Delta$; the origin *is* included in the confidence region. The variables referred to in the appendix are indicated.

by the outcome of each meiosis. Some of the states are sharing; others are nonsharing. If every crossover (change of state) leads to a change in sharing status, the Poisson distribution of transitions follows trivially from the assumption of no crossover interference.

When not every crossover changes the sharing status, the following argument applies: In a collection of N pairs, where N is large, a fraction z_L of pairs will be in a sharing state and a fraction $1 - z_L$ will be in a nonsharing state. Within the sharing and nonsharing categories, different states are equally probable. Hence the overall up and down transition rates may be obtained by averaging the rates for each state. The rates for different states, as well as the averages, are given by Feingold (1993). In the limit of large N, the number S of transitions between the ends of the confidence region is small (that is, $o(N)$) and thus (with probability 1) at most one crossover occurs on each chromosome. Accordingly, transition arrivals are locally well described by a Poisson process with rate $\nu = [\rho_+(1 - z_L) + \rho_- z_L]N$ per Morgan.

G. Relaxing the Assumption of Large Sample Size

Our formulas are based on the assumption that the sample size N is large, in which case the LOD score Markov process reduces to a simple random walk in the region of interest. To explore the effect of smaller N, one can explicitly simulate the actual Markov process itself. We briefly discuss the results of such simulations.

Given a sample size N and a value of z_L (α was fixed at $\frac{1}{2}$), we directly simulated the process as follows. We chose the initial number N_s of pairs that share alleles at the locus from a binomial distribution binomial (N,z_L). We then took steps in both directions from the locus. At each step, we changed the number of pairs that share by $+1$ with probability $\pi = (N - N_s)/N$ and by -1 with probability $1 - \pi$ (the updated value of N_s was used at each step). We carried out the walk for a sufficient number of steps for N_s to drop far enough below threshold that the probability of returning was negligible. We then looked for the leftmost and rightmost exits from the confidence interval and accumulated a histogram of the number of steps between exits. A total of 100,000 walks were generated in each simulation.

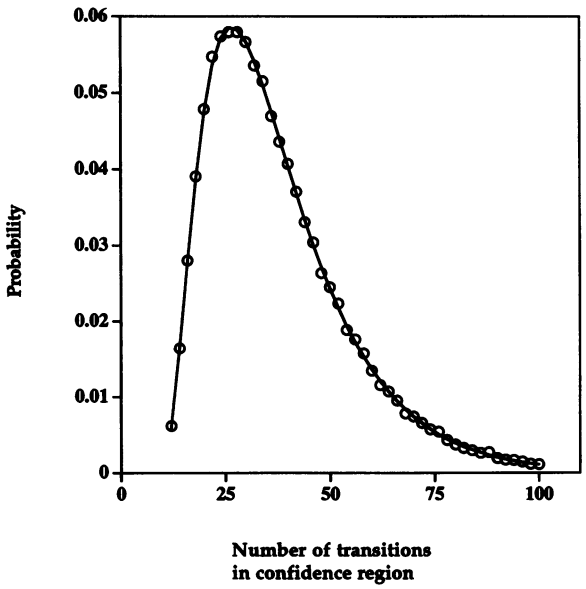

Figure A2 Agreement between calculated distribution and simulations for large N. Graph shows the distribution of the number of steps within the 95% confidence region for $z_L = \frac{2}{3}$. Solid line is computed as in proposition 3 and the appendix, part D. Unblackened circles are simulated as in the appendix, part G.

Table A1

Comparison of the Median Number of Steps within the 95% Confidence Region for Finite $N = N^*$ and the Large N Limit (See the Appendix, Part G, for Details)

Sharing Fraction z_L	N^* for 1 cM	Median for $N = N^*$	Median for Limit of Large N
.600	7642	80	76
.667	2742	28	28
.750	1480	14	14
.800	910	10	10
.900	447	4	4
.950	403	4	4
.975	175	4	2
.980	173	2	2

To verify proposition 3 by simulation, we first ran the simulations for infinite N, i.e., with π fixed at $(1 - z_L)$, producing a random walk with constant downward drift. In figure A2, we show the computed and simulated probability distributions for the number of steps in $C_{.95}$ for $z_L = 2/3$. The agreement is excellent. As a quantitative test, we computed the χ^2 statistic for the 45 bins shown in the figure. The value of χ^2 was 35.15 with 45 df, indicating no significant difference between the two distributions ($P > .85$).

We then studied the effect of finite N, as opposed to large (infinite) N. Using the large N assumption in proposition 3, we determined the number N^* of meioses needed to reduce $C_{.95}$ to 1 cM. We then carried out simulations with $N = N^*$ to see if the result was different. In fact, we expected no significant effect since in every case N^* is large compared to the net change Δ in N_s needed to exit the confidence region. This expectation was confirmed. The median number of steps within $C_{.95}$ is shown in table A1 for both the finite case ($N = N^*$) and in the limit of large N. The numbers are very close, and identical in most cases. Note that the differences are always in the direction of more steps in the finite N case: since the sharing fraction decreases away from the locus, the drift decreases, and it takes somewhat longer for the LOD score to drop. Hence, the large N limit provides a lower bound on the number of steps within the confidence region (and on the number of pairs to narrow the region to 1 cM).

H. Incomplete Genetic Map

Confidence regions are defined above in terms of the actual location of transitions. If the map is not infinitely dense, one will only be able to localize each transition to the interval between two available markers. This slightly enlarges the region that must be searched. If markers are randomly distributed, the size of the region will be increased by the addition of exponentially distributed variables corresponding to the distance to the first marker beyond the confidence region on either end.

References

Berrettini WH, Ferraro TN, Goldin LR, Weeks DE, Dtera-Wadleigh S, Nurnberger JI, Gershon ES (1994) Chromosome 18 DNA markers and manic-depressive illness: evidence for a susceptibility gene. Proc Natl Acad Sci USA 91:5918–5921

Boehnke M (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Hum Genet 55:379–390

Feingold E (1993) Markov processes for modeling and analyzing a new genetic mapping method. J Appl Probability 30:766–779

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234–251

Feller W (1970) An introduction to probability theory and its applications. Vol 1, 3d ed. John Wiley & Sons, New York

Hastbacka J, de la Chapelle A, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, et al (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1073–1087

Lander ES, Schork N (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lange K, Kunkel L, Aldridge J, Latt SA (1985) Accurate and superaccurate gene mapping. Am J Hum Genet 37:853–867

Risch N (1990*a*) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228

Risch N (1990*b*) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241

Siegmund D (1988) Confidence sets in change-point problems. Int Stat Rev 56:31–48