Genome Scanning for Segments Shared Identical by Descent among Distant Relatives in Isolated Populations

L. Kathryn Durham* and Eleanor Feingold†

Department of Biostatistics, The Rollins School of Public Health, Emory University, Atlanta

Summary

In this paper, we address some of the statistical issues concerning false-positive rates that arise when the whole genome, or a portion thereof, is scanned in distantly related individuals, to search for a disease locus. We derive a method for correcting false-positive probabilities for the large number of comparisons that are performed when scanning a large portion of the genome. We consider both the idealized situation of a dense set of fully informative markers and the more realistic data-collection strategy of an initial scan at low resolution to identify promising areas, which then are typed with markers at high resolution. We also examine the accuracy of false-positive rates approximated using a conservative estimate of the separation distance between affected individuals in the current generation and the common ancestral couple. Calculation of false-positive rates when inbreeding is present in the pedigree also is considered.

Introduction

One standard strategy for detection of genetic linkage in humans involves the searching of the genomes of pairs of affected relatives for areas that they share identical by descent (IBD) or identical by state (IBS). The idea is that if affected pairs show greater-than-expected genetic agreement at a marker or over a continuous segment, then that region may have a high probability of containing the trait locus (e.g., see Haseman and Elston 1972; Weeks and Lange 1988; Feingold et al. 1993; Kruglyak et al. 1996). The fact that the relationship between the two individuals is known facilitates the determination both of expected genetic sharing and of how much excess sharing is considered to be significant evidence of linkage. Feingold (1993), Lander and Kruglyak (1995), and others have discussed a correction for linkage statistics that adjusts false-positive probabilities for the large number of comparisons that are made when a sizable portion of the genome is scanned. However, one possible drawback of these methods is that close relatives, on average, share relatively large areas of the genome. This means that many pairs may be required, to localize the gene to a reasonably small area.

Linkage-disequilibrium mapping is a technique that can result in a finer specification of the locus of a disease-causing allele. One approach is to apply this technique in a population that has grown in isolation since it was founded. It is assumed that most of the disease chromosomes in the current generation are descended from an ancestral chromosome in the founding generation, so that, in the immediate vicinity of a disease locus, a distinctive haplotype should be observed. Ideally, enough time, or generations, will have passed so that the shared region will be small enough to allow tight specification of the gene locus, but the mutation will be recent enough so that linkage equilibrium with the surrounding alleles has not yet been reached. This technique typically is used for fine mapping, once the gene has been localized to a region of a particular chromosome, and the association of a disease locus with several markers or haplotypes in the vicinity is considered (Jorde 1995). In some studies that were undertaken in Finland, for example, it was assumed that the population was founded 100 generations ago (Hästbacka et al. 1992; Lehesjoki et al. 1993). Thus, the extent of linkage disequilibrium (or the IBD region) between the disease locus and the surrounding markers is small, probably <1 cM. A full genome scan to detect a region of this size would require extremely dense genetic maps for each chromosome, making the number of markers required to localize the disease allele prohibitively large when the current genotyping technology is used.

In this article, we consider genome scanning for regions shared IBD by distantly related individuals, which is a strategy for gene localization that brings together some of the features of affected-pair analysis and of linkage-disequilibrium studies. It is applicable in the case of a young mutation or a population in which individu-

Received August 2, 1996; accepted for publication July 22, 1997.

Address for correspondence and reprints: Dr. Eleanor Feingold, Department of Human Genetics, University of Pittsburgh, 130 DeSoto Street, Pittsburgh, PA 15261.

* Present affiliation: Genetic Epidemiology Laboratory, Imperial Cancer Research Fund, St. James's University Hospital, Leeds.

† Present affiliation: Department of Human Genetics, University of Pittsburgh, Pittsburgh.

0002-9297/97/6104-0010$02.00

als share a relationship that is more distant than that usually considered for extended families (i.e., cousins, grandparents, etc.) but closer than that for individuals from an older population, such as in Finland. The assumption is that most of the disease chromosomes in the current population are IBD to a single founder mutation present in an earlier generation. Because the founding chromosome existed in the not-so-distant past, individuals in the current population still will share fairly large areas IBD. However, these areas will be smaller than those in individuals who usually are considered to be relatives, so that the sample size required to narrow the trait region, with reasonable probability, will be small. In general, the pedigree structure may not be known exactly, but, to calculate false-positive probabilities, it must be possible to at least approximate the number of generations between the probands and their common ancestor(s).

An example in which this mapping strategy was applied appears in the study by Houwen et al. (1994). The authors performed a genomewide scan for shared IBD segments, to search for a locus associated with benign recurrent intrahepatic cholestasis (BRIC), a recessive disorder. Using three affected individuals, they mapped the BRIC gene to a 20-cM region on chromosome 18. Since the complete pedigree was not available, they estimated that these individuals, all of whom were born from consanguineous relationships, were separated from a common ancestor by 6–10 generations. These patients were identified to be from a fishing community of several thousand people in the Netherlands, in which "most of its members descend from individuals who lived in the vicinity by the 17th century" (Houwen et al. 1994, p. 381). Houwen et al. (1994) used an initial scan of 256 microsatellite markers spaced at ~10–20-cM intervals, covering ~90% of the autosomal genome. They observed the number of individuals matching on all pairs of consecutive markers, and the segments exhibiting excess sharing then were typed with more markers, to distinguish those IBD from those IBS.

This article addresses some of the statistical issues that arise in studies for which the whole genome, or a portion thereof, is scanned in distantly related individuals, to search for a disease locus. We assume that the exact pedigree structure may be unknown but that some characteristic, such as the distance to common ancestors, can be estimated. In the first part of the article, we present methods for calculating false-positive probabilities. The second part of the article examines the possible error introduced by approximation of an unknown pedigree with a pedigree based on a conservative estimate of the separation distance between the affected individuals in the current generation and the common ancestral couple. We also consider a method for calculating false-positive rates when there is inbreeding in the pedigree (or in the population).

False-Positive Probability Calculations for an Approximate Pedigree Structure

The false-positive rate (or P value) that we wish to calculate is the probability that i or more of N chromosomes from affected individuals are IBD at some point on the genome, where i is the maximum number sharing that has been observed over the genome. The probability is calculated under the null hypothesis that there is no genetic locus for the trait. We initially consider the case in which virtually continuous IBD information is available, such as from a scan using densely spaced, highly polymorphic markers (Feingold et al. 1993; Guo 1995; Kruglyak and Lander 1995). We then consider a two-stage search strategy, for which an initial sparse scan is followed by a more exhaustive search of promising regions. We assume that the exact pedigree structure connecting the individuals under study is unknown and, therefore, is approximated by a simple pedigree, as was used in the study by Houwen et al. (1994).

Approximate Pedigree Structures

It is useful to divert our thinking somewhat from the human pedigree structure and to focus on the pedigree of relevant chromosomes (Donnelly 1983). In this framework, individuals in a pedigree are represented by two chromosomes if they are bilineally related to the founding couple and by one chromosome if they are unilineally related. The chromosome pedigrees that we consider here appear to be most similar to that of Donnelly's (1983) cousin-type relationship, but we are interested in the relationship among several descendants.

In the article by Houwen et al. (1994), the exact relationships among the individuals selected from the present population, on the basis of their disease status, were unknown. However, to approximate probabilities, a conservative pedigree structure was assessed on the basis of the likely number of meioses separating the affected individuals from a single ancestral couple. Also, Houwen et al. (1994) calculated false-positive probabilities by treating each line of descent as independent from the others, implying that the individuals share no intermediate ancestors.

In our calculations, we use the approximate pedigree structure used by Houwen et al. (1994). Figure 1*A* displays the pedigree for a dominant disease, for which four chromosomes are separated from the common ancestral couple, by four meioses. Figure 1*B* shows the pedigree for a recessive disease, for which four chromosomes are separated from the common ancestral couple, by four meioses. This is similar to the pedigree approximation used by Houwen et al. (1994), except that, in their study, six chromosomes in three affected individuals were considered. Note that, without loss of generality, we can represent all members of the pedigree as the same sex, except when matings are implied.

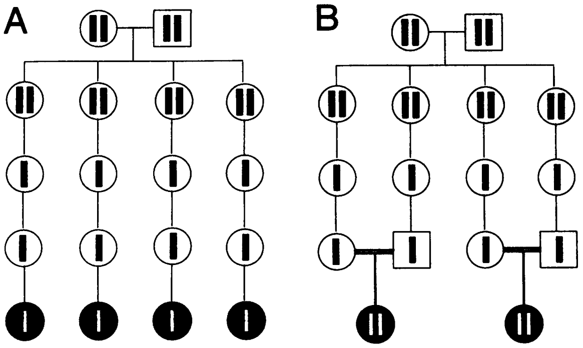

Figure 1 *A*, Chromosome pedigree for four chromosomes for a dominant disease, with a separation distance of four meioses. Blackened circles represent affected individuals. *B*, Chromosome pedigree for four chromosomes for a recessive disease, with a separation distance of four meioses. Blackened circles represent affected individuals.

For a dominant disease, if the separation distance from the ancestral couple is approximated as m meioses, the actual pedigree structure is approximated by the representation of $(m - 1)$th cousins. For a recessive disease, however, as was the case in the study by Houwen et al. (1994), IBD matching on chromosomes within the same individual necessitates that the affected individuals be bilineally related to the ancestral couple. If the first inbred individuals occur in this last generation, this should not affect the probability calculations, since the comparison of all chromosomes for individuals with a recessive trait is equivalent to the comparison of single chromosomes in individuals affected by a dominant trait (see Discussion).

False-Positive Probabilities for a Dense Scan

To approximate false-positive probabilities for a chromosome pedigree like those described above, we use ideas from Donnelly's (1983) model and results from the Poisson clumping heuristic (Aldous 1989). Donnelly (1983) used a random walk on a hypercube model to represent all meioses connecting two relatives. He used this model to compute the probability that two individuals with a given relationship share any genetic material IBD, when continuous information is available. (See Appendix A for more background details. For other applications of Donnelly's hypercube model, see the studies by Guo 1995 and Bickeböller and Thompson 1996.) Related work also was done by Thomas et al. (1994), who used a slightly different approach to calculate the probability that relatives in a pedigree of arbitrary size all share some genetic material IBD along a continuous genome scan. They define a random variable S, which represents the number of segments where all individuals are IBD, and a false-positive probability is given by $P(S > 0)$. A useful feature of their results is that they also did not depend on the exact pedigree structure but only on the total number of meioses separating the individuals.

Of course, for an area of the genome to be considered as a candidate region for a particular gene locus, especially for a complex trait, it is not necessary that all individuals be IBD at that region. The strategy adopted by Thomas et al. (1994) was to classify as a sporadic case an individual who did not match the others, to remove this person from the pedigree under consideration, and to recalculate the probabilities. Houwen et al. (1994) approximated false-positive probabilities for subsets of individuals matching, using the initial spacing between markers. To approximately correct for multiple testing, the probabilities of false positives were added for all segments typed throughout the genome.

Our method, based on Donnelly's (1983) ideas, allows us to approximate the probability that i or more of N chromosomes from affected individuals are IBD somewhere on the genome. This is done by application of the Poisson clumping heuristic (Aldous 1989). The idea is that some events, such as the maxima or minima of certain processes, can be modeled as sparse clumps randomly scattered over an area, with their positions determined by a Poisson process. In this case, we are interested in the probabilities of rare excursions of the process that models the number of chromosomes matching as a function of the crossovers along the genome. Feingold (1993) used these ideas to approximate false-positive probabilities for linkage analysis, using pairs of affected relatives.

We define X_t to be the number of N chromosomes matching an ancestral chromosome IBD at a point t along the genome. A false-positive rate based on a threshold i can be approximated as

$$\alpha \approx 4 \times P[\max_{0 \leqslant t \leqslant L}(X_t) \leqslant i] \, ,$$

where L is the length of the genome and where the probability is calculated with the assumption that there is no trait locus. We multiply the probability of matching for a single ancestral chromosome by 4 because there generally is a common ancestral couple, resulting in four opportunities for sharing by chance. This factor of 4 is approximate but is reasonable for a large i, since the probability of exceeding i somewhere in the scan for two or more ancestral chromosomes is small.

We approximate the probability by assuming that the time T_i it takes until the process first reaches the level i is distributed exponentially (this exponential approximation has a long history in the probability and the engineering literature; see Aldous 1989). Then, the false-positive rate is

$$\alpha \approx 4 \times P[\max_{0 \leqslant t \leqslant L}(X_t) \geqslant i] = 4 \times P(T_i \leqslant L) \, .$$

The idea behind these methods is that if excursions of the process X_t above i are approximately a Poisson process, then the mean time between excursions should be $[\pi(i)]^{-1}$, where π is the stationary distribution of X_t. The approximation also involves the expected clump size, EC_i, which is the expected number of visits to i within an excursion to that level.

By use of Aldous' (1989) Poisson clumping heuristic, the false-positive rate is approximated by

$$\alpha \approx 4 \times \mathrm{P}(T_i \leqslant L) \approx 4\left\{1 - \exp\left[-L\frac{\pi(i)}{EC_i}\right]\right\} .$$

We use ideas from Donnelly's (1983) model for chromosome pedigrees to derive the expression for EC_i. These results and the derivation of $\pi(i)$ appear in Appendix A. By use of equations (A1) and (A2), in Appendix A, for EC_i and $\pi(i)$, respectively, the false-positive rate α is approximated by

$$\mathrm{P}(T_i \leqslant L) \approx 4\left\{1 - \exp\left[\frac{-L\binom{N}{i}\left(\frac{1}{2^m}\right)^i\left(1-\frac{1}{2^m}\right)^{N-i}\lambda m(i2^m - N)}{2^m - 1}\right]\right\} , \tag{1}$$

where λ is the average rate of crossovers per genetic length, for a single meiosis ($\lambda = 1$ for L measured in morgans, and $\lambda = .01$ for L measured in centimorgans). The approximation should hold for any N and for values of i large enough so that $\mathrm{P}(T_i \leqslant L)$ is fairly small.

Note that an alternative form of this approximation has the factor of 4 in the exponent instead of outside the rest of the expression. That form should be more accurate for small i (larger $\mathrm{P}[T_i \leqslant L]$). For large i, the two forms give virtually the same probabilities.

Accuracy of the Approximation

As an example, we considered the case with a separation distance of six meioses, as in the Houwen et al. (1994) approximate pedigree. Table 1 compares our approximation from equation (1) with simulated probabilities using a genome length of $L = 33$ M. The simulation methods are described in Appendix B. The simulation results indicate that use of the approximation is reasonable, to estimate the probabilities associated with larger numbers of chromosomes matching IBD in approximate pedigree structures.

Two-Stage Scanning

An efficient strategy for scanning the chromosomes of affected individuals involves typing markers along a sparse map, throughout the genome, and then following up on promising areas from the initial scan with a dense array of markers, to extract the full inheritance information (Lander and Kruglyak 1995). Whereas the initial scan indicates areas where the individuals share common alleles, the second stage of typing should reveal whether the identity is by state or by descent, as was done in the study by Houwen et al. (1994). Thus, alleles that are IBS in the initial scan will increase the number of areas indicated for follow-up but otherwise should not interfere with the gene localization.

False-positive rates calculated on the basis of sparse scanning inherently are functions of the marker spacing used: for a given observed maximum number of chromosomes IBD, the farther apart the markers, the smaller the P value. Figure 2 shows an example of the possible relationship between continuous information and that from an initial sparse scan, when markers are spaced every 10 cM. The height of the graph represents the number of chromosomes matching IBD along a continuous scan, and the arrowheads indicate locations where marker data are observed. If a maximum number IBD of i is found in a dense scan, it is known that this is the true maximum, and the P value is calculated accordingly. However, seeing i matching is a more rare event in a sparse scan, and it is likely that the maximum of the continuous process was missed. The only argument that could be made for reporting of the sparse-scan P values would be in the case for which the investigator did no further typing after the initial scan, but, in practice, this is rarely the case. Interesting areas almost always are examined in a subsequent, more focused search.

Once areas in the second stage of the analysis have been covered by a dense marker map, the appropriate probabilities of false positives should be calculated as if a dense genome scan had been conducted. We advocate

Table 1

Comparison of False-Positive Probabilities Approximated by Use of Equation (1) with Empirical False-Positive Probabilities from Simulations

		α		
N	*i*	From Equation (1)	From Simulations	No. of Simulations
3	2	.9866	.8758	1×10^6
	3	.0091	.0077	
4	3	.0353	.0331	1×10^6
	4	.0002	.0002	
5	3	.0860	.0792	1×10^7
	4	.0009	.0009	
	5	3.7×10^{-6}	4.8×10^{-6}	
6	3	.1666	.1518	1×10^8
	4	.0027	.0026	
	5	2.2×10^{-5}	2.1×10^{-5}	
	6	6.9×10^{-8}	8×10^{-8}	

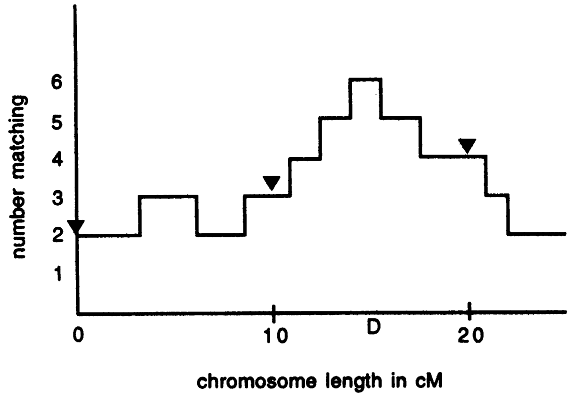

Figure 2 Example of a continuous-matching process. The horizontal axis indicates the length along a segment of the genome, and the vertical axis indicates how many chromosomes are matching the ancestral chromosome. The arrowheads point to the numbers observed to be matching along a 10-cM marker map. "D" indicates the disease locus.

the reporting of false-positive rates only in terms of dense scanning, for the following reasons. When a hierarchical scan is performed, the areas that would indicate false-positive results from the continuous chromosome almost invariably are included as areas for follow-up in the secondary dense search (Lander and Kruglyak 1995). This means that the probabilities of observation of large numbers matching are essentially the same whether a hierarchical scan or a complete dense scan is used. The two-stage strategy can be thought of as an efficient way to approximately scan the entire genome, since the interesting areas can be focused on without having to look at every marker. Another appealing feature of *P* values calculated on the basis of dense scanning is that they are the most conservative estimates. Thus, there is no correction, in terms of multiple-testing considerations, for the inclusion of as many regions as possible from the initial scan in the second stage of dense marker typing, if the conservative corrections for dense data are used at the second stage.

The Effect of Pedigree Approximation on False-Positive Rates

In this section, we consider implications for the resulting false-positive probabilities of estimating an actual pedigree with an approximate structure and what effect inbreeding within the pedigree has on these estimates.

Comparison of Probabilities from Approximate and Exact Pedigrees

The probabilities in table 1 are based on the approximate pedigree structure used by Houwen et al. (1994), described earlier, for which not all relationships among individuals are known and for which a conservative estimate of the separation distance from the common ancestral couple is used. Also in this approximation, the lines of descent connecting the affected individuals in the current generation to the ancestral couple are independent.

An important question is how accurately such approximations might reflect false-positive probabilities for the true pedigree structure. One obvious difference between the approximation and a realistic pedigree is that if the minimum separation distance is used in the approximation, then there will be some chromosomes that are separated from the ancestral couple by a greater number of meioses. This indicates that more meioses are involved in the actual structure than in the approximation. On the other hand, it is likely that there are subsets of affected individuals who share other ancestors more closely than they share the ancestral couple, which indicates that fewer meioses are involved in the actual structure than in the approximation. Our goal is to determine what effect this overestimation and underestimation, with respect to different parts of the actual pedigree, has on false-positive probabilities.

This question is difficult to answer in full generality, but we can compare the false-positive rates for a large extended pedigree with known relationships to those from an approximation with a conservative estimate of the number of meioses separating affecteds from their common ancestor. We considered the extended pedigree from the study by Thomas et al. (1994) (hereafter called "the Thomas pedigree"), which connected six cases of colorectal cancer to an ancestral couple (note that we considered only one individual from the youngest sibship in lineage 1) (fig. 3*A*). We approximated this pedigree with one in which six chromosomes are separated from the ancestral couple by 3 meioses, since this is the minimum separation represented in the actual pedigree (fig. 3*B*). Note that this adds extra meioses to the pedigree, in terms of the first-generation sibship, but underestimates meioses with respect to separation from the founders, since the greatest separation is 5 meioses. In this case, the total number of meioses in the exact and the approximate pedigree structures are the same (18 meioses).

Table 2 compares results from our approximation for continuous IBD information with those from simulations based on variations of the Thomas-pedigree structure. The simulation methods are described in Appendix B. In addition to the approximation to the actual structure described above, we also considered the Thomas pedigree with one extra meiosis added to each lineage, so that the minimum separation distance was 4 meioses. We approximated this pedigree with one in which six chromosomes are separated from the ancestral couple by 4 meioses, so that, again, the total number of meioses in the exact and the approximate pedigree structures

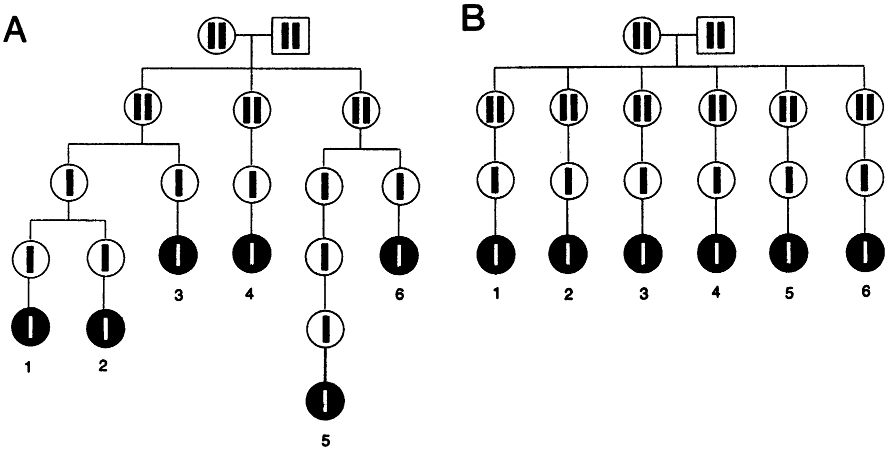

Figure 3 *A*, Thomas pedigree, including only one individual from the sibship at lineage 1. Blackened circles represent affected individuals. *B*, Approximate pedigree based on the minimum separation from the ancestral couple. Blackened circles represent affected individuals.

was the same (24 meioses). For both of these cases, the approximation results were fairly close to the simulated results based on the actual pedigree.

We then considered the case in which the approximation and the actual structure have a different total number of meioses. To do this, we performed simulations based on pedigrees almost identical to the Thomas pedigree and to the variation described above but with one more or one less meiosis. For example, in the actual Thomas pedigree, addition of an extra meiosis to lineage 1 increases the total number of meioses in the pedigree to 19, but the approximation structure based on the minimum separation in the pedigree does not change. The results in table 2 include those for the addition to and the subtraction from two different lines, to demonstrate that the probabilities do not depend significantly on which lineage was changed. The results indicate that false-positive probabilities from the approximation will be too large if the approximation underestimates the total number of meioses in the actual pedigree and too small if the approximation overestimates the total meioses.

We also considered whether the best approximation always would be achieved by using the minimum separation represented in the actual pedigree. Consider the Thomas pedigree with an extra meiosis added to every line, except to that of lineage 4. Thus, the minimum separation is still 3 meioses, but the total number of actual meioses (i.e., 23) is closer to that for the approximation in which six

Table 2

Comparison of False-Positive Rates for Variations on the Thomas Pedigree

	α, Based on Thomas Pedigree					
			One Meiosis Added to		One Meiosis Deleted from	
i	Approximation from Equation (1) (*m* = 3)	Actual Structure	Lineage 1	Lineage 5	Lineage 1	Lineage 5
5	.2966	.2136	.1302	.1580	.3595	.3030
6	.0091	.0092	.0050	.0046	.0154	.0160
	α, Based on Thomas Pedigree with an Extra Meiosis Added to Each Lineage					
	Approximation from Equation (1) (*m* = 4)					
4	.3903	.2456	.1655	.2099	.3881	.3110
5	.0139	.0106	.0062	.0079	.0192	.0164
6	.0002	.0002	.0001	.0001	.0004	.0004

Note.—See figure 3*A* for an illustration of the Thomas pedigree.

individuals are separated from the ancestral couple by four generations, giving 24 total meioses. Table 3 indicates that, whereas use of the minimum separation as an approximation yields a conservative result, the approximation in which the total number of meioses is one more than the actual total results in a much closer estimate of the probabilities for the real pedigree.

It appears that the best strategy for estimation of false-positive rates that are as accurate as possible, while remaining conservative, will depend on how much information about the actual pedigree is available. If the exact relationships are unknown, it may be best to use a conservative estimate of m in equation (1), especially if it is not possible to estimate the total number of meioses represented in the pedigree.

If the exact pedigree relating the individuals is available, there are several options for obtaining estimates of false-positive rates. If a conservative value for m can be chosen, yielding an approximate pedigree with a close match to the total number of meioses in the actual pedigree, then this value for m can be used in equation (1) to yield a reasonably close approximation to the P value. Another option for obtaining estimates of false-positive probabilities is to perform simulations for the exact pedigree structure, as were generated here for the Thomas pedigree, as outlined in Appendix B. However, it is important to note that these results represent a kind of weighted average of the probabilities that different actual subsets of i chromosomes are matching IBD. For example, the probability of a false positive for a subset of i chromosomes for individuals who are more closely related than most will be somewhat higher than that of the average simulation result.

The Effect of Inbreeding

To this point, we have ignored the effect that inbreeding may have on calculations of this type, except to note that, in this pedigree structure, the individuals under study for a recessive disease are assumed to be bilineally related to the ancestral couple. In fact, young isolated populations may be strongly inbred, and it is clear that if intermediate ancestors also are bilineally related to the ancestral couple, then this will tend to increase the amount of genetic material that those in the current generation share with the founding couple.

Table 3

Comparison of False-Positive Probabilities for Two Different Approximations for the Thomas Pedigree with an Extra Meiosis for Each Lineage, Except for Lineage 4

	α		
	Approximation from Equation (1)		From Simulations Based on Actual Structure
i	m = 4	m = 3	
5	.0139	.2966	.0181
6	.0002	.0091	.0004

Consider an individual who is part of a pedigree in which there are m meioses between him/her and the common ancestors. If none of the intervening relatives (besides the first-generation siblings) are bilineally related to the ancestral couple, then the probability that a particular allele is inherited from one of the four chromosomes is $(1/2)^m$.

Now suppose that the affected individuals are each related to the ancestral couple by d lines of descent. Then, the probability that a particular allele is inherited is

$$\begin{aligned} &\mathrm{P(allele\ inherited)} \\ &\quad = \sum_{j=1}^{d} \mathrm{P(passed\ through\ line\ }j) \\ &\quad - \sum_{j<k} \mathrm{P(passed\ through\ lines\ }j\mathrm{\ and\ }k) \\ &\quad + \ldots + (-1)^{d+1}\mathrm{P(passed\ through\ all\ }d\mathrm{\ lines)} \quad (2) \\ &\quad \leqslant \sum_{j=1}^{d} \mathrm{P(passed\ through\ line\ }j) \\ &\quad \leqslant \frac{d}{2^m} . \quad (3) \end{aligned}$$

The direction of the inequality assures that the use of equation (3) to estimate the expected proportion of genetic material shared with the ancestor's chromosome will be conservative. Also, equation (3) will be a good approximation to equation (2), because the probabilities that the allele is passed through two or more lines become small, particularly as m becomes moderately large.

To adjust the false-positive rates for inbreeding, we can incorporate equation (3) into $\pi(i)$, given in equation (A2) of Appendix A, so that

$$\mathrm{P}(T_i \leqslant L) \approx 4 \left\{ 1 - \exp\left[\frac{-L\binom{N}{i}\left(\frac{d}{2^m}\right)^i \left(1 - \frac{d}{2^m}\right)^{N-i} \lambda m(i2^m - N)}{2^m - 1} \right] \right\} , \quad (4)$$

when the current-generation individuals are related to the ancestors by d lines of descent. Because of inbreeding, EC_i, as approximated in Appendix A, should be somewhat smaller than the actual EC_i. This also will result in a slight overestimation of $\mathrm{P}(T_i \leqslant L)$, that is, in a conservative P value.

To assess the accuracy of this inbreeding correction,

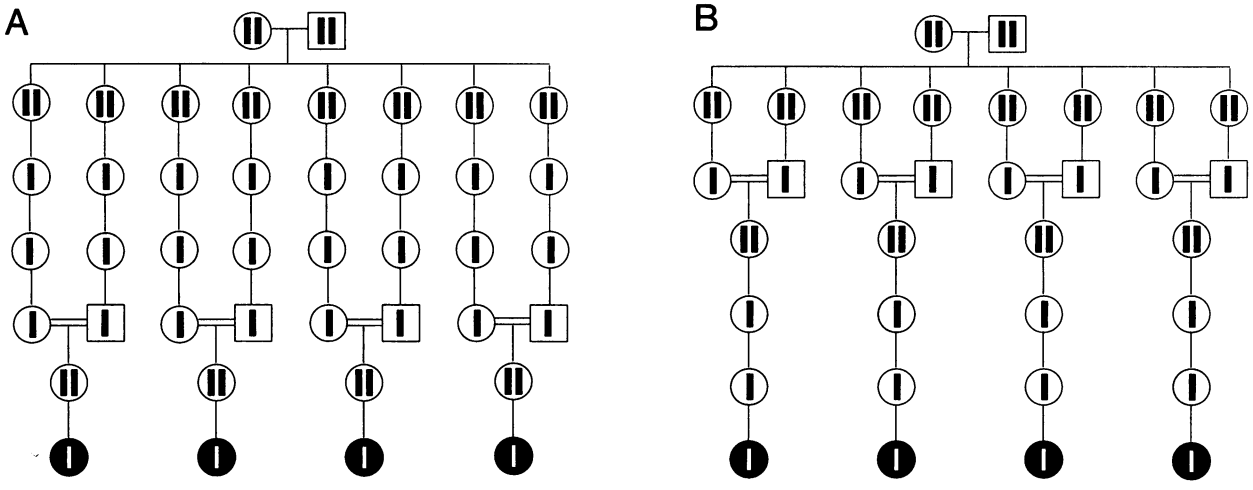

Figure 4 *A,* Chromosome pedigree for four chromosomes, in which the parents of the affecteds are bilineally related to the ancestral couple. Blackened circles represent affected individuals. *B,* Chromosome pedigree for four chromosomes, in which the great-grandparents of the affecteds are bilineally related to the ancestral couple. Blackened circles represent affected individuals.

we first compared the approximation in equation (4) with simulations based on the pedigree structures shown in figure 4, in which four chromosomes are separated from common ancestors by six meioses. In the pedigree shown in figure 4*A*, the parents of the affected individuals are bilineally related to the common ancestors, and, in the pedigree shown in figure 4*B*, their great-grandparents are inbred. Notice that we initially are considering independent lines of descent, similar to those in the pedigrees shown in figure 1. For a recessive disease, the approximate structures would be similar, except that the affected individuals would be bilineally related to the ancestors, as shown in figure 1*B*. Table 4 compares the simulated and the approximated probabilities, calculated with $N = 4$, $m = 6$, and $d = 2$. Equation (4) appears to offer a good approximation to the simulated results, which do not differ greatly according to where the inbreeding actually occurred in the pedigree structure. Thus, matching of the total number of meioses does not appear to be important when this inbreeding correction strategy is used, since the total number of meioses represented in figure 4*A* and *B* are quite different (44 and 36 meioses, respectively). For comparison, the false-positive approximation using equation (1) without inbreeding also is presented. Notice that there is a significant difference between the results calculated with inbreeding and those calculated without inbreeding.

We then considered whether this strategy could accurately approximate false-positive rates for the more realistic pedigree shown in figure 5, which we developed on the basis of the Thomas pedigree but with added inbreeding. Table 5 shows empirical false-positive rates for simulations based on this pedigree, compared with approximated rates calculated by use of equation (4), with $N = 6$, $m = 5$, and $d = 2$. In table 5, we also compare approximated rates calculated by use of equation (1), with $m = 6$ and $N = 6$. This approximation

Table 4

Comparison of False-Positive Probabilities Approximated by Use of Equation (4) ($d = 2$) and Equation (1) (no inbreeding), Both with $N = 4$ and $m = 6$, with Empirical False-Positive Probabilities from Simulations Based on the Inbred Pedigrees Shown in Figure 4

	α			
		From Simulations		
i	Approximation from Equation (4)	Figure 4*A* Pedigree	Figure 4*B* Pedigree	Approximation from Equation (1)
3	.2700	.2428	.2371	.0353
4	.0030	.0027	.0029	.0002

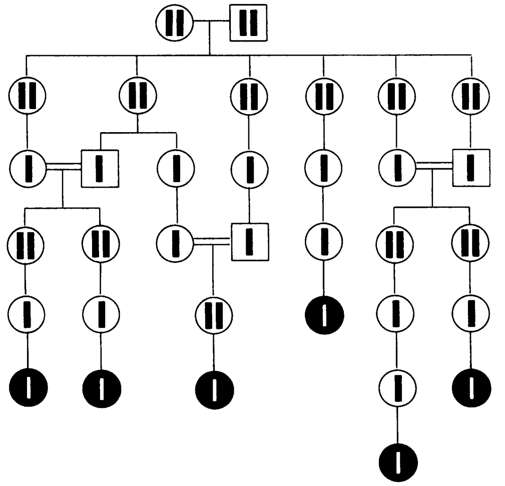

Figure 5 Pedigree based loosely on the Thomas pedigree but with inbreeding added. Blackened circles represent affected individuals.

ignores inbreeding but gives a close conservative match to the total meioses in the actual pedigree (37 meioses, compared with 36 meioses in the approximation). Although the empirical rates are close to those calculated under inbreeding, by use of equation (4), these approximate results appear to underestimate somewhat the *P* values. This likely is due to the nonindependence of the lines of descent in the realistic pedigree. However, these approximate rates are much closer to the empirical rates than those calculated by use of equation (1), which match the total number of meioses but which ignore inbreeding.

In calculating false-positive rates under inbreeding, using equation (4), we used the *average* or best estimate of *m*, instead of the most conservative value. We applied this same strategy in estimating the average *d*, without the intermediate step of developing an approximate pedigree and attempting to match the total number of meioses. In the situation in which the exact pedigree is unknown, it only may be possible to use a reasonable approximation for *m* and for *d*, on the basis of the parts of the pedigree that can be reconstructed. If this is not possible, false-positive rates for the observed number of chromosomes matching can be calculated with and without reasonable values for *d*, to see if the statistical significance is robust, that is, if it still holds under inbreeding.

Discussion

In this paper, we derived a method for correcting false-positive probabilities for the large number of comparisons that are performed in a dense scan of a large portion of the genome in affected distant relatives, to search for a disease locus. We also discussed guidelines for a realistic data-collection strategy in which areas are typed with markers at high resolution, only after an initial scan at low resolution is performed, and we advocated the reporting of dense-scan *P* values after this final stage. We investigated the accuracy of the approximation of an unknown pedigree with a pedigree based on a conservative estimate of the separation distance between affected individuals in the current generation and the common ancestral couple. We also proposed a simple method for calculation of false-positive rates under inbreeding that accounts for the extra genetic material introduced by chance in such a case.

We would like to be able to recommend a standard procedure for the application of this approximation for estimation of false-positive rates for IBD sharing among distant relatives. This is virtually impossible, because the amount of information and, in fact, the pedigree structure itself are different in every situation. However, we believe that the important pieces of information for estimation of false-positive probabilities include the following:

1. Total meioses in the pedigree;
2. Number of lines of descent (*d*) per affected (or per chromosome); and
3. Number of meioses (*m*) separating affecteds (or chromosomes) from common ancestors.

Where there is little or no inbreeding, the best strategy for approximation appears to be use of a value for *m* in equation (1) that results in an approximate pedigree that most closely matches the estimate of the total meioses in the real pedigree. In some cases, this will correspond to use of the estimate of minimum separation from the ancestral couple, but this strategy alone, without consideration of the total meioses, may give misleading results. The same qualitative results should hold for pedigrees larger than the ones considered here, although a difference of a few meioses will not have as large an effect in such a case. When inbreeding is present, the intermediate step of the development of an approxi-

Table 5

Comparison of False-Positive Probabilities Approximated by Use of Equation (4) (N = 6, d = 2, and m = 5) and Equation (1) (N = 6 and m = 6) with Empirical False-Positive Probabilities from Simulations Based on the Pedigree Shown in Figure 5

	α		
i	From Simulations	Approximation from Equation (4)	Approximation from Equation (1)
4	.5626	.4898	.0027
5	.0346	.0175	2.2×10^{-5}
6	.0008	.0002	6.9×10^{-8}

mate pedigree and the matching of meioses did not appear to be as useful. Instead, we used estimates of the average *d* and *m* from the real pedigree, in equation (4), to match most closely the empirical false-positive rates.

Approximated false-positive probabilities such as those described here can be important tools for the localization of genes, by use of young isolated populations, but it is important to use caution and common sense when the results are interpreted. For an unknown pedigree, the approximation will be only as good as the best guess about the pedigree structure. If subgroups of affected individuals are much more closely related than the estimate of *m* implies, then their probabilities for IBD sharing could be severely underestimated. For the case in which there is more information about the pedigree, an area for future work involves how to combine information and to design mapping strategies for pedigrees with subsets of close relatives who all share a common ancestor.

An example of such a pedigree can be seen in the study by Puffenberger et al. (1994). They performed a genome scan for a recessive form of Hirschsprung disease (HSCR) in a large, inbred, Mennonite kindred and reported the mapping to chromosome 13q22 of a new locus for HSCR. Information on the complete pedigree structure was available, and all HSCR cases were 8–12 generations removed from a single ancestral couple. Also in this pedigree, the parents of diseased individuals in the last generation are related to the ancestral couple by about seven lines of descent, on average. Although the affected individuals were descended from a single ancestral couple, a scanning strategy different from that employed by Houwen et al. (1994) was used. A low-resolution genomewide screen was performed for sib pairs from three nuclear families, and segments identified as candidates for further dense searching were those for which the sib pairs had high IBD scores. Thus, alleles were allowed to vary between different nuclear families, as in linkage analysis, and the fact that all individuals were descended from common ancestors was not used in the initial scan. Targeted regions then were saturated with more markers, at high resolution, for one sibling from each family in the original scan as well as for other cases included from the pedigree, and frequency differences between transmitted and untransmitted parental alleles were used to test for linkage disequilibrium.

Since we are advocating calculating probabilities only on the basis of the last stage of dense scanning, the methods presented in this article could be applied to a study similar to that conducted by Puffenberger et al. (1994), in which a search for IBD segments due to the common ancestor was conducted in the final dense scan. Note that Puffenberger et al. (1994) were able to achieve in one study what often takes several studies to accomplish—that is, initial scanning using linkage techniques and fine localization using linkage-disequilibrium mapping. If probability calculations of this type are appropriate for such an all-in-one study, it could be argued that they also are appropriate for many linkage-disequilibrium studies that take advantage of earlier linkage results. The extent to which these results can be applied to different types of multiple-stage studies and studies conducted in older populations is an area of further research.

We stated that, in the pedigree, if the first individuals who are bilineally related to the ancestral couple (after the first-generation siblings) occur in this last generation, this should not affect the probability calculations, since the comparison of all chromosomes of individuals with a recessive trait is equivalent to the comparison of single chromosomes of individuals affected by a dominant trait. One possible exception to this generalization could occur if there are person-level factors, associated with the trait under study, that affect the classification of chromosomes. One example of this could be an excessive presence of phenocopies in the population. For a recessive disease, this may exclude chromosomes in pairs, whereas, for a dominant disease, it will exclude only one chromosome per phenocopy.

There are notable differences between our probability approximation, which controls for multiple testing by application of the ideas of the Poisson clumping heuristic under a continuous scan, and that of Houwen et al. (1994). They calculated probabilities for the information only in the initial sparse scan, using 256 microsatellite markers spaced at 10–20-cM intervals, spanning ~3,260 cM. For example, for six chromosomes at a separation distance of 6 meioses, we approximated the probability that three or more chromosomes match somewhere in the continuous genome to be .167 (table 1). Houwen et al. (1994) approximated the probability that their technique locates a segment for which three or more chromosomes match to be .011. Their probability is much smaller because it indicates how likely they were to find the shared region for which they were searching on the basis of its size and the marker spacing that was used.

By use of the Poisson model for crossovers, the size of the segment shared by a particular set of *i* chromosomes separated from the ancestral couple by *m* meioses has a γ distribution with parameters of 2 and $im\lambda$. This is because, from the trait gene, the distance to the nearest crossover on each side is distributed exponentially, and the sum of two independent exponentials is distributed as a γ random variable. Since the median size of a region where three chromosomes, at a separation distance of six meioses, match is only ~9 cM, it is not surprising that the calculations from the study by Houwen et al. (1994) indicate that observation of this amount of sharing is a rare event when a 10–20-cM map is used. Also, the median size of a segment shared IBD by five chromosomes, with a separation distance of six meioses, is only ~5.5 cM. By use of the conservative dense-scan approximation, the P value is 2×10^{-5} (table 1), as compared with 5×10^{-7}, given by Houwen et al. (1994) for their technique. It is important to consider whether another study conducted with a separation distance of six meioses also could be expected to detect significant 20-cM shared segments. It appears that the segment size that Houwen et al. (1994) were able to detect may be quite unusual, since the probability of a 20-cM or longer segment shared by five chromosomes, with a separation of six meioses, is <.02.

Acknowledgments

We are grateful to Stephanie Sherman and David Siegmund for their helpful comments on an earlier version of the manuscript. This work was supported, in part, by NIH grants RO1-HG00848-01A1 (to E.F.) and T32-AI07442 (to L.K.D.).

Appendix A

Calculations for $\pi(i)$ and for EC_i

Donnelly (1983) computed the probability that two individuals with a given relationship will share any genetic material IBD when continuous information is available. His model involves the representation of each meiosis by the assigning of either a 0 or a 1 along the continuous length of a chromosome, according to which of the parent's parents' (i.e., the child's grandparents) genetic material is present. The process takes the value 0 if the grandmother's genetic material is present and 1 if the grandfather's is present. Donnelly (1983) used the idea that the genetic sharing states resulting from the meioses that connect two individuals can be thought of as the vertices of a hypercube, which is a cube-like structure with more vertices than a three-dimensional cube. The arrangement of vertices on the hypercube structure indicates how many crossovers are needed to move between different genetic sharing states, so that only adjacent vertices can be reached in a single crossover. For example, in a three-dimensional cube (fig. A1), the vertices (001), (010), and (100) could be reached by a single crossover from the vertex (000), but the vertex (111) could be reached only after a minimum of three crossovers. (Note that the vertices do *not* represent markers along a chromosome but, rather, the outcome of all relevant meioses in the pedigree, for a fixed point along the genome.) In general, if the pedigree representing the relationship between two individuals involves m meioses, the process of crossovers can be represented as a random walk on an m-dimensional cube, and certain vertices of the hypercube can represent IBD sharing between the two individuals. These vertices are called the "hitting set." This walk on the hypercube is a continuous-time Markov process for which the time element is actually the length of the genome.

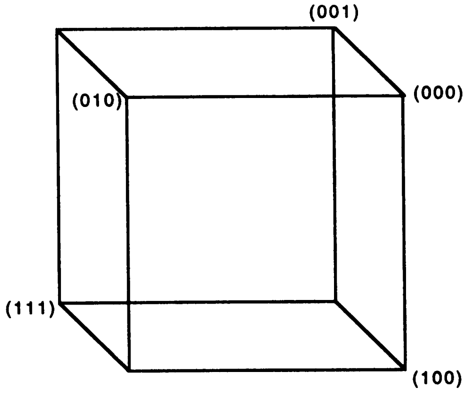

Figure A1 Hypercube representing a three-meiosis pedigree (m = 3).

To approximate EC_i, we follow Aldous' (1989) example B4, in which the Markov chain is approximated near i by a chain with an up transition rate $\mu(N - i)$ and a down transition rate ϕi, where μ is the rate at which chromosomes that do not match move to matching and where ϕ is the rate at which matching chromosomes move to not matching. Then, the approximation for the clump size is $EC_i \approx [\phi i - \mu(N - i)]^{-1}$. We assume that crossovers occur along a chromosome as a Poisson process.

To determine the rates at which chromosomes that do not match move to matching and those that match move to not matching, we use an extension of Donnelly's (1983) grandparent-type relationship. Since the lines of descent are independent under the approximate pedigree structure, each affected individual has his/her own grandparent-type process determining the amount of IBD sharing with the ancestral chromosome. Without loss of generality, we assume that all the individuals in the pedigree structure are female, so that the affected individual in the present generation is matching the ancestor IBD only where the random walk hits the vertex (00 . . . 0), that is, an m-dimensional vector of 0's. Note that we add one meiosis to Donnelly's (1983) version of this chain, since we are interested in matching a particular chromosome of the ancestor.

A usual square transition-rate matrix, or Q matrix, has dimensions equal to the number of possible states and would contain the rates of switching from a vertex to any other vertex, that is, from any vector of 0's and 1's, representing the outcomes of each meiosis, to any other such vector. Thus, the Q matrix would be the square of dimension 2^m, which can be quite large. Donnelly (1983) circumvented this problem by considering instead a Q matrix of orbits, for which the orbits can be thought of as mutually exclusive sets of states that are similar. We also use this approach, since the orbits for the grandparent-type relationship are straightforward; a state is simply classified into an orbit according to the number of 1's that it contains. When m is the number of meioses separating an individual from the common ancestor, the Q matrix for the process of orbits is

$$\begin{bmatrix} -m & 1 & 0 & \cdots & & 0 \\ m & -m & 2 & & & \vdots \\ 0 & m-1 & -m & \ddots & & \\ \vdots & & m-2 & \ddots & m-1 & 0 \\ & & & \ddots & -m & m \\ 0 & \cdots & & 0 & 1 & -m \end{bmatrix},$$

so that the matrix of transition probabilities for the embedded process is

$$\begin{bmatrix} 0 & \frac{1}{m} & 0 & \cdots & & 0 \\ 1 & 0 & \frac{2}{m} & & & \vdots \\ 0 & \frac{m-1}{m} & 0 & \ddots & & \\ \vdots & & \frac{m-2}{m} & \ddots & \frac{m-1}{m} & 0 \\ & & & \ddots & 0 & 1 \\ 0 & \cdots & & 0 & \frac{1}{m} & 0 \end{bmatrix}.$$

Note that we have preserved Donnelly's (1983) notation, in which the columns sum to 1. As an example, the probability that a jump from a state with two 1's results in a state with one 1 is $2/m$. There is only one orbit from which a chromosome can move to match the genetic material of the ancestor's chromosome, that is, to hit the vertex (00 . . . 0). This orbit consists of states with only one 1, that is, states in which there is only one nonmatching meiosis. Since all states are equally probable, the probability of being in one of these states and moving to match is $[m/(2^m - 1)](1/m)$. Since the rate at which crossovers are occurring is $m\lambda$, $\mu = (m\lambda)/(2^m - 1)$. If a chromosome is matching, then, after a crossover, a move to nonmatching is certain. Thus, $\phi = m\lambda$, and EC_i is approximated by

$$[\phi i - \mu(N - i)]^{-1} = \frac{2^m - 1}{\lambda m(i2^m - N)} . \qquad \text{(A1)}$$

Since which chromosomes are matching an ancestral chromosome is independent under the pedigree approximation described in the section entitled "Approximate Pedigree Structures," we also will use the binomial distribution used by Houwen et al. (1994), for the calculation of $\pi(i)$. Note that $\pi(i)$ can be interpreted as the probability that the number of chromosomes at a separation distance m that match an ancestral chromosome is i at a particular location t along the genome. Then,

$$\pi(i) = \binom{N}{i}\left(\frac{1}{2^m}\right)^i\left(1 - \frac{1}{2^m}\right)^{N-i} . \qquad \text{(A2)}$$

Appendix B

Simulation Methods

The simulations discussed in this article were performed by use of a Poisson-process model for crossovers. Donnelly's (1983) chromosome pedigree, as described in Appendix A, was constructed to represent the relationships among relevant chromosomes. Each simulation consisted of the following:

1. A random number of crossovers occurring throughout the genome, for the pedigree, was generated based on a Poisson distribution, with the mean determined by N, m, an assumed genome length (L = 33 morgans), and the average rate of crossovers for a single meiosis (one per morgan).
2. An initial state was chosen for each chromosome in the pedigree by the random assigning of either a 0 or a 1 to represent which parent's genetic material was present.
3. At each crossover, it was determined randomly which chromosome switched either from a 0 to a 1 or from a 1 to a 0. Then, the number of chromosomes matching the ancestor was determined by consideration of the values of the 0–1 processes. Without loss of generality, we assumed all matching to be through maternal lines. Only if all the chromosomes connecting the current-generation individual to the ancestor took the value 0 was a match indicated. This corresponds to the random walk hitting the vertex (0, . . . , 0), as described in Appendix A. For the simulations in table 1, a different sequence of chromosomes connected each sampled individual to the ancestor. For simulations reflecting an actual pedigree, as in table 2, the outcomes of the meioses were connected as in the actual structure.
4. The maximum number observed matching i was recorded from each simulation and was used to calculate empirical probabilities. These were multiplied by 4, in order to represent the fact that there are actually four chromosomes, associated with the ancestral couple, that could result in matching.

References

Aldous D (1989) Probability approximations via the Poisson clumping heuristic. Springer-Verlag, New York

Bickeböller H, Thompson EA (1996) Distribution of genome shared IBD by half-sibs: approximation by the Poisson clumping heuristic. Theor Pop Biol 50:66–90

Donnelly KP (1983) The probability that related individuals share some section of genome identical by descent. Theor Pop Biol 23:34–63

Feingold E (1993) Markov processes for modeling and analyzing a new genetic mapping method. J Appl Prob 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

Guo S-W (1995) Proportion of genome shared identical by descent by relatives: concept, computation, and applications. Am J Hum Genet 56:1468–1476

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and marker locus. Behav Genet 2:3–19

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, Freimer NB (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat Genet 8:380–386

Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet 56:11–14

Kruglyak L, Lander ES (1995) High-resolution genetic mapping of complex traits. Am J Hum Genet 56:1212–1223

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Lander ES, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Lehesjoki AE, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, de la Chapelle A (1993) Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. Hum Mol Genet 2:1229–1234

Puffenberger EG, Kauffman ER, Bolk S, Matise TC, Washington SS, Angrist M, Weissenbach J, et al (1994) Identity by descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. Hum Mol Genet 3:1217–1225

Thomas A, Skolnick MH, Lewis CM (1994) Genomic mismatch scanning in pedigrees. IMA J Math Appl Med Biol 11:1–16

Weeks DE, Lange K (1988) The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315–326