Am. J. Hum. Genet. 61:430–438, 1997

True and False Positive Peaks in Genomewide Scans: Applications of Length-Biased Sampling to Linkage Mapping

Joseph D. Terwilliger,[1,2] William D. Shannon,[3] G. Mark Lathrop,[1] John P. Nolan,[4] Lynn R. Goldin,[6] Gary A. Chase,[5] and Daniel E. Weeks[1,7]

[1]The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford; [2]Department of Psychiatry and Columbia Genome Center, Columbia University, New York; [3]Washington University School of Medicine, St. Louis; [4]Department of Mathematics and Statistics, American University, and National Center for Human Genome Research, National Institutes of Health, and [5]Georgetown University Medical Center, Washington, DC; [6]Clinical Neurogenetics Branch, National Institute of Mental Health, Bethesda; and [7]Department of Human Genetics, University of Pittsburgh, Pittsburgh

Summary

Disease-susceptibility loci are now being mapped via genomewide scans in which a linkage statistic is computed at each of a large number of markers. Such disease-susceptibility loci may be identified via a peak in the test statistic when the latter is plotted against the genetic map. In this paper we establish, by appealing to renewal theory, that true positive peaks are expected to be longer than false positive peaks. These results are verified by a realistic simulation of a genomewide linkage study based on the affected-sib-pair design. Since longer peaks are more likely to contain a gene of interest than are shorter peaks, these differences may aid in linkage mapping, justifying assignment of lower priority to shorter peaks. However, since these differences are generally small, statistics based on both peak length and height may not be much more powerful than those based on height alone. The results presented here also provide a theoretical framework for methods that use the length of shared haplotypes in populations to map disease genes.

Introduction

One method for genetically mapping disease-susceptibility loci involved in "complex" disease is to carry out a genomewide screen of a panel of affected sib pairs, testing for linkage with highly informative markers spaced evenly throughout the genome. The evidence for linkage may then be assayed by nonparametric tests of whether the siblings share marker alleles more often than expected. The results of such a genomewide scan are a collection of peaks for which the statistic exceeds a preset significance threshold. Some peaks, the true ones, are caused by the presence of a disease gene, whereas others (the false ones) are caused by random fluctuations alone. We have suggested that mean identical by descent (IBD) sharing is higher in a neighborhood around a true peak and that therefore peak length could be used to discriminate between true and false peaks of similar height (Shannon et al. 1995). In contrast, Lander and Kruglyak (1995) have claimed that there is no way to distinguish between true peaks and peaks of the same height that arise from random fluctuations. The motivation for this paper was to investigate the cause of the apparent discrepancy between these claims.

Our study was additionally motivated by the following observations: To evaluate different strategies for genomewide scans using the affected pedigree member (APM) method of linkage analysis, we simulated markers every 2.5 cM throughout the genome in families in which an autosomal dominant disease was segregating (Brown et al. 1994). When we examined the behavior of the APM statistic within each genomewide scan, often the true peak was longer than any false peaks of similar height. In addition, we (Goldin et al. 1995) noticed, when applying sib-pair methods to simulated data from the Genetic Analysis Workshop 9, that the region around a true susceptibility locus contained a long sequence of consecutive markers with P values $<.05$, although no single P value approached the standard threshold of .0001. We sought to determine whether these observations were based on a real difference between true and false peaks or were just due to random fluctuations. The theory of length-biased sampling provides the appropriate framework for exploring this area. Using this framework, we show here, by both analytical arguments and simulation experiments, that true peaks are, on average, longer than false peaks and that longer peaks are more likely to contain the gene of interest than are shorter peaks. Note that this paper provides the theoretical basis for these observations but leaves it to

Received February 2, 1996; accepted for publication May 14, 1997.

Address for correspondence and reprints: Dr. Daniel E. Weeks, The Wellcome Trust Centre, University of Oxford, Windmill Road, Oxford OX3 7BN, United Kingdom, or Department of Human Genetics, University of Pittsburgh, 130 DeSoto Street, Pittsburgh, PA 15261. E-mail: daniel.weeks@well.ox.ac.uk or dweeks@watson.hgen.pitt.edu

0002-9297/97/6102-0023$02.00

subsequent studies to explore how to take advantage of length-biased sampling in mapping. However, Goldin and Chase (in press) have recently developed some new length-biased sampling statistics that performed well in the context of a genomewide screen.

Background

Length-biased sampling is based on the principle that, given a collection of random intervals of varying length covering a specific point, longer intervals are more likely to be sampled than shorter intervals. Although we here seek to take advantage of length-biased sampling to aid in distinguishing true from false peaks, it usually has negative ramifications in most studies—for example, in association studies (Simon 1980), segregation analyses (Ewens and Asaba 1984), population studies (Patil and Rao 1978), and cell genetics (Schotz and Zelen 1971). Length-biased sampling is the basis of the "waiting-time paradox" (Feller 1971) or "inspection paradox" (Ross 1983) (for an intuitive explanation, see Hemenway 1982). Feller (1971) described this paradox in terms of waiting times at a bus stop: Suppose buses arrive according to a Poisson process, with the interarrival times between buses distributed exponentially with mean $1/\lambda$. If a person arrives at time t, the expected time until the next bus arrives is $1/\lambda$, independent of when the previous bus had been there (the exponential distribution is "memoryless"). The expected time since the last bus arrived is also distributed exponentially with mean $1/\lambda$, yielding the surprising result that the waiting time from the previous bus to the next bus is the sum of two exponential random variables, or an Erlang$(2,\lambda)$ random variable with mean $2/\lambda$, twice as long as the standard mean interarrival time (note that an Erlang $[2,\lambda]$ is equivalent to a Gamma$[2,\lambda]$—the Erlang is used preferentially in the stochastic process literature).

This phenomenon of length-biased sampling can be rigorously explained by use of a well-developed mathematical framework known as "renewal theory" (see Smith [1958] also note that Owen [1948] and Bailey [1961] applied renewal theory to recombination processes). In a renewal process, events occur repeatedly, and the times between these events are independent and identically distributed (iid). The recombination process along a chromosome can be viewed as a renewal process in which "events" are recombination events and "time" is genetic distance along the chromosome (Owen 1948); "interarrival times" correspond to the distance between adjacent recombination events. If there is no interference, then the recombination process is a Poisson process; other processes can be used to model interference (Haldane 1938; Owen 1948; Bailey 1961; Feingold 1993; Feingold et al. 1993; Guo 1996; J. P. Nolan, unpublished data). Consistent with the bus example above, renewal theory indicates that the average length of an interval between recombination events covering a *specific* point t will be larger than the average interval length (see Corollary B below). This has been noted before in the context of human genetics (e.g., see Lange et al. 1985; Boehnke 1994). And, if there is no interference, then the mean of the interarrival times covering the *specific* time t is twice the mean of the arbitrary interarrival times (in large data sets, this result holds even if there is interference; see the Superposition section below). Thus, longer intervals are more likely to contain a true point (see Resnick 1994). In genetic-mapping studies, the definition of the disease prespecifies the point t, and ascertainment of disease families makes it more likely that this point t will actually be covered by a significant peak.

In order to verify that the theory of length-biased sampling does indeed apply to the length of chromosomal segments inherited IBD, we conducted a simple simulation study of segments, inherited IBD by a sib pair, around a gene with fixed map position; and the results (not shown) were as expected. Similar results obtain for more distantly related relatives as well. Note that single and multiple recombination frequencies do not change around the disease locus, since segregation and recombination are independent processes. Length-biased sampling has been implicitly invoked in linkage analyses that treat longer conserved haplotypes as evidence of linkage (Houwen et al. 1994). Our application of renewal-theory concepts to this area provides a more rigorous theoretical framework for the work of Houwen et al. (1994), supporting their ad hoc Monte Carlo approach.

Mathematical Theory

DEFINITION: Suppose that a stochastic process generates recurrent events with locations S_1, S_2, and so on, with $S_i \geq S_{i-1}$ for all i (fig. 1). This process is a *renewal process* if the interarrival times $Y_i = S_i - S_{i-1}$ are mutually independent and follow a common distribution F (Feller 1971).

McFadden (1962) points out that there is a distinction between the interarrival times Y_i indexed by labeling an arbitrary event with $i = 0$ and those interarrival times η indexed by starting with the first event before an arbi-

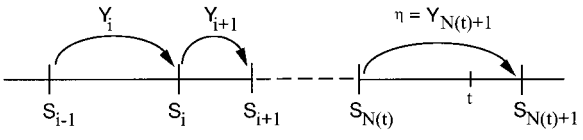

Figure 1 Pictorial representation of a renewal process, illustrating our notation.

trary time *t*. Indeed, "by starting with an arbitrary t we are more likely to choose a long interval than if we start with an arbitrary event" (McFadden 1962, p. 365). This is reflected in the following theorem.

THEOREM I: Suppose that we have a renewal process with interarrival times Y_i following the distribution *F*. Define $S_{N(t)}$ as the arrival time of the last event before some *fixed* time *t*, and let $\eta = Y_{N(t)+1} = S_{N(t)+1} - S_{N(t)}$ be the interarrival time covering *t* (fig. 1). Then η has the length-biased distribution function

$$G(y) = \frac{1}{E_F[Y]} \int_0^y x dF(x) ,$$

where $y \geq 0$ (see Sen 1987).

COROLLARY A: If the Y_i have the density function $dF(x)$ corresponding to *F*, then η has the density function (see Cox 1962; Sen 1987) $dG(x) = xdF(x)/E_F[Y]$, for $x \geq 0$ (McFadden 1962, eq. [2.1]).

COROLLARY B: $E_G[\eta] = E_F[Y] + \mathrm{Var}(Y)/E_F[Y] = E_F[Y](1 + I_Y) = E_F[Y^2]/E_F[Y]$, where the dispersion index I_Y is $\mathrm{Var}(Y)/\{E_F[Y]\}^2$. For higher moments, $E_G[\eta^k] = E_F[Y^{k+1}]/E_F[Y]$ (Cox 1962; Cox and Lewis 1966; Patil and Rao 1978; Nelson 1995).

Since $I_Y \geq 0$, we have $E_G[\eta] \geq E_F[Y]$. Note that, if the dispersion index $I_Y < 1$, then Corollary B implies that (see Cox and Isham 1980) $E_G[\eta] < 2E_F[Y]$, and so $E_G[\eta]$ is between $E_F[Y]$ and $2E_F[Y]$. Also, for a Poisson process, the dispersion index $I_Y = 1$, and so we have $E_G[\eta] = 2E_F[Y]$; that is, the mean of the interarrival times covering the *specific* time *t* is twice the mean of the arbitrary interarrival times.

Superposition

The superposition of a large number of independent renewal processes is approximately a Poisson process, according to the Palm-Khintchine theorem (Palm 1943; Khinchin 1960; Nelson 1995). The conditions required for the Palm-Khintchine theorem to hold have been outlined by Grigelionis (1963). However, Samuels (1974) showed that the superposition of a small number of independent renewal processes is itself a renewal process if and only if the component processes are all Poisson processes themselves.

Applicability to Linkage Statistics

Although we have established that the distance between two crossovers flanking a specific disease gene is expected to be larger than the distance between any two adjacent crossovers, this is not equivalent to showing that true peaks are longer than false peaks. Peaks are defined in terms of statistics that are themselves based on the combined effects of multiple independent recombination renewal processes, so it is not immediately obvious that peak lengths are themselves governed by a renewal process. To discuss this further, we narrow our discussion to the mean sib-pair-sharing statistic

$$W_k = \frac{\left[2\left(\sum_{n=1}^{N} A_{k,n}\right) - M\right]}{\sqrt{M}} ,$$

where $A_{k,n}$ is the number of alleles shared IBD at locus *k* by the *n*th sib pair, *M* is the number of informative meioses, and *N* is the number of sib pairs.

Recall that our main query is: Can the length-biased sampling effect aid in distinguishing true peaks from false peaks *of the same height,* in contrast to the claim of Lander and Kruglyak (1995) to the contrary? To answer this by using the theory of length-biased sampling, we need to define the peak lengths, *L,* as interarrival times from some renewal process, which means that the *L*'s have to be independent and identically distributed (peak length is defined as the genetic distance for which the statistics are continuously significant—i.e., length of the excursion above the significance threshold). However, the peak lengths are not identically distributed throughout the genome if there is a disease gene, since peaks are generally higher (and, therefore, longer) in the region of a disease-susceptibility locus. The way out of this difficulty is to condition on height, where two peaks are "of the same height" if they each have the same maximal values of W_k. Then, if we define our renewal-process event as "crossing the significance threshold *c*," we have a process alternating between upward and downward excursions (fig. 2). If L_j is the

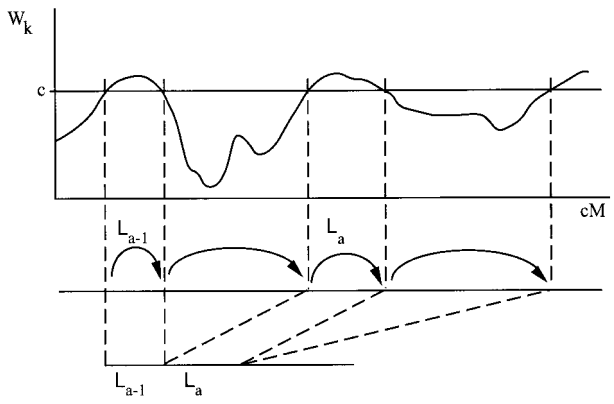

Figure 2 Pictorial representation of an alternating renewal process, illustrating our notation. The upper portion graphs the sib-pair statistic W_k against the genetic map, where *c* is the significance threshold. The lower portion shows how the renewal process is defined as a function of the sib-pair process.

length of the jth peak of a given height, then we can construct a new stochastic process in which the ith event's location is given by $S_i = L_1 + L_2 + \ldots + L_i$. All the L_i's are independent and follow the same distribution, since each one is a function of exactly the same underlying recombination processes. Thus, renewal theory does apply to lengths of peaks, and so, conditional on height, the mean length of a true peak should be larger than the corresponding mean length of a false peak (as discussed in the Mathematical Theory section above) (Scheaffer 1972).

Methods

To test whether our theoretical model applies to reality, we simulated a genomewide linkage study of a genetically complex trait, using an affected-sib-pair design. We present here a specific example of many different simulations that we have undertaken. This simulation was intended to show that the length of a peak that covers one of the five true loci tends to be larger than the lengths of the false peaks. Simulation details are as follows.

a. Disease Model

We simulated a trait under the control of five loci such that each one contributes additively 6% of the trait variance and independent environmental factors control the remaining 70% of the variance. The trait was assumed to have a prevalence of 5%, and the disease-predisposing allele had a frequency of .1 at each locus. The five disease loci were located arbitrarily, at map position 50 cM on the five longest chromosomes, and were thus segregating independently. For each disease locus, we simulated parents' genotypes according to population allele frequencies, and then we simulated the segregation of alleles to the two children. On the basis of these disease genotypes, quantitative-trait phenotypes were simulated for the sib pair. A child was "affected" if his or her phenotype was in the upper 5th percentile. Five hundred affected sib pairs were ascertained.

b. Marker Simulation

Markers were spaced every 1 cM throughout the human genome, where each chromosome had a realistic length as given by Morton (1991). If a family was ascertained, then recombination events on each chromosome were simulated, from parents to children, according to the Sturt (1976) mapping function. Once the crossover positions had been simulated, the segregation to the offspring was simulated randomly for nondisease chromosomes, whereas, for disease chromosomes, segregation was determined by the previously simulated disease genotypes, according to the chromosome-based simulation method of Terwilliger et al. (1993). For each marker k and each sib pair, we simulated whether the father was informative, where the probability of being informative was ψ. (Note that a marker is "informative" if the parent's marker genotype is heterozygous and different from the other parent's genotype. Our "informativity" is approximately the same as the PIC, and this approximation improves as the number of alleles increases.) If the father was informative, we incremented the number of informative parents, M, by 1, and, if the sibs inherited identical chromosomes from the father at the marker, we incremented the IBD count, $A_{k,n}$, by 1. We then repeated the same process for the mother. Then the affected-sib-pair mean test W_k was computed for each marker k, and the P value was computed according to the standard normal distribution.

c. Peak Definition

A peak was defined as an excursion of the W_k statistics above the significance threshold c. A peak was "true" if it contained at least one point within ξ cM of a disease locus. To count multiple excursions very near one another as one peak, W_k was permitted to fall below c for $<2\xi$ cM (e.g., a peak may have a brief gap). Peaks were grouped into height classes based on rounding of $-\ln(P \text{ value})$ to the nearest integer, where the P value is based on the maximum height of the peak. This classification scheme was fine enough that, within any class, the distribution of the peak heights did not vary significantly between the true and false groups.

Note that none of our simulation assumptions should cause true peaks to be spuriously longer than false peaks. The assumptions were that (1) the mode of inheritance of the trait was fixed as described above; (2) segregation was independently simulated for each chromosome; (3) recombination events were simulated according to the Sturt (1976) model; (4) if a peak was truncated at a telomere, simulation of markers was continued beyond the telomere until the peak decayed below the threshold c, which may slightly bias toward longer false peaks; and (5) the disease loci were on the five longest chromosomes, which may bias slightly toward shorter true peaks because, under the assumptions of the Sturt mapping function, interference is stronger on shorter chromosomes. Thus, we are confident that our simulation results are, if anything, conservative.

Results

Peak Lengths

The simulation results are consistent with the theoretical expectation that true peaks should be longer than false peaks, as shown, in table 1, for two levels of marker informativity and several definitions of "peaks." Note also that peak-length differences are greater for partially informative markers than for fully informative markers,

Table 1

Mean *L*'s from a Simulation of 1,000 Genomes, with Five Disease Genes per Genome, by Height Class

	$\psi = 1$						$\psi = .70$			
	$\xi = 10$		$\xi = 5$		$\xi = 1$		$\xi = 10$		$\xi = 5$	
P	False	True	False	True	False	True	False	True	False	True
.01	3.8	6.4	3.1	5.0	1.9	2.5	3.1	7.3	2.2	4.3
.0025	10.2	15.2	8.6	12.2	3.1	6.9	8.2	17.8	5.5	11.9
.0009	17.3	23.5	14.6	19.5	7.8	12.1	14.1	26.0	9.6	18.9
.0003	20.8	31.3	18.5	26.1	9.8	17.4	18.9	32.2	14.3	25.4
.0001	26.7	34.1	23.8	30.2	10.6	21.8	23.6	38.8	17.6	30.8
.00005	31.8	37.4	27.7	34.2	10.4	25.7	26.6	43.1	20.7	35.3
.00001	31.3	41.3	29.0	37.4	9.0	29.5	35.8	45.2	25.0	38.1
.000006	40.4	44.9	37.6	42.0	11.8	34.2	24.7	49.6	22.3	41.2

NOTE.—*L* is defined as the length of time that the statistic stays significant at the .01 level.

because the variance of the false peak lengths is larger when markers are less informative. Also, when $\xi > 1$ cM, the theory for the waiting-time paradox does not directly apply, since there is not a specific point *t* which all "true" peaks must cover—they must cover some point in a region near the point *t*—hence the difference between true and false peak lengths is smaller than when $\xi = 1$ cM.

To determine whether peak length can aid in the categorization of peaks as true or false (within a height class), we computed the posterior probability of a peak being true, conditional on both height and length. Note that we define this posterior probability as the proportion, T/(T + F), of all simulated peaks that are true. The results (fig. 3) indicate that the most efficient use of length information is to exclude very short peaks from further consideration. In addition, these findings support the common strategy of preferentially exploring the longest peaks first, since they are more likely to be true than the shorter peaks.

Marker Informativeness

The observed numbers of true and false peaks under different assumptions about the probability ψ of a parent being informative and about ξ are shown in table 2. Note that the number of true peaks increased as the informativeness decreased but that the number of false peaks increased even more rapidly, so that the posterior probability of a true peak dropped when ψ decreased (e.g., from .32 at $\psi = 1$ to .19 at $\psi = .70$, when $\xi = 5$ cM). When $\psi = 1$, the W_k are strongly correlated along the chromosome, because of linkage. However, as ψ decreases, the W_k become less correlated, since different subsets of the simulated meioses are informative at tightly linked loci. As they become less correlated, the observed statistic will vary more from one marker to the next, whereas in the fully informative case peaks are much smoother (because of the correlation). Thus, a positive test result with a low-heterozygosity marker has a greater chance of being a false positive than has one with a fully polymorphic marker. (In other words, in a genomewide scan, a LOD score of 3 with a marker with 50% heterozygosity is much less impressive than a LOD score of 3 with a fully informative marker.)

Table 2 indicates, as expected, that the posterior probability of a peak being true is strongly influenced by its maximum height, H_k. For example, when $\psi = 1$ and $\xi = 10$ cM, the posterior probability of a true peak was .37, based on *all* peaks regardless of height. However, the posterior probability of a true peak conditional on H_k is only 484/3,996, or .12, when H_k barely exceeds *c* and is as high as 1 for large values of H_k. In other words, if the *P* value is highly significant, then the peak is almost certainly true.

Lengths of General Shared IBD Segments

Length-biased sampling is not only applicable to IBD-based linkage analysis using large numbers of sib pairs but can also be applied to small samples of distantly related relatives. Recently Houwen et al. (1994) presented an empirical argument for use of the lengths of regions shared IBD between relatives to isolate true genes from a background of segments shared IBD by chance alone. If the relatives are sufficiently distantly related, then the distribution of the lengths of segments shared IBD by *all* these individuals is independent of the specific relationships between them and is simply a function of the sum of the number of meioses connecting them; for example, two second-cousins should have the same shared segment distribution as do two sets of siblings who are first-cousins—these four individuals rep-

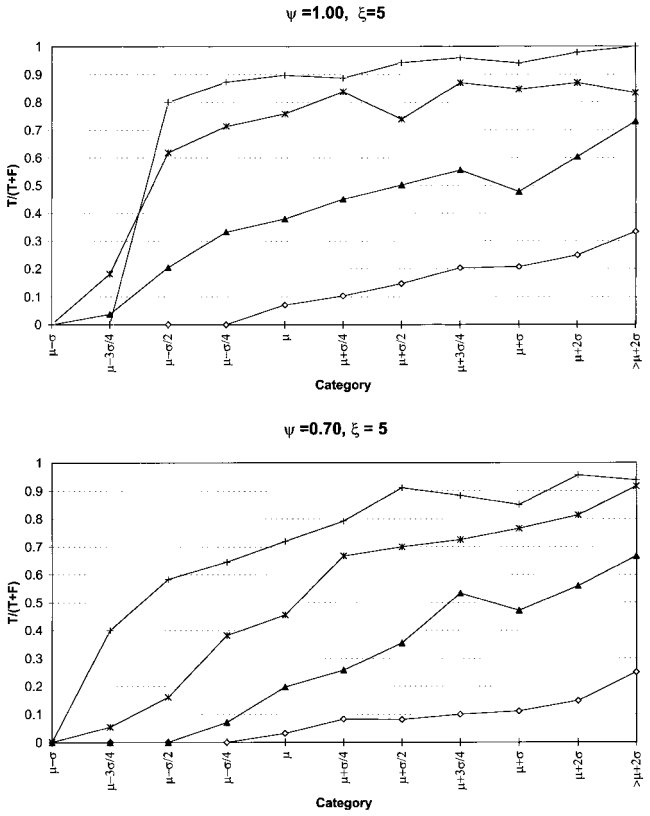

Figure 3 Probability of a peak being true, conditional on length and height, for different heights (◇ = *P*-value class centered on .01; ▲ = *P*-value class centered on .001; | = *P*-value class centered on .0001; and + = *P*-value class centered on .00001). The peak classes are defined in terms of the mean μ and variance s^2 of the lengths of the false peaks.

resent six meioses from the founder mating, and the two second-cousins also represent six meioses. If there is no interference, the collective outcome from all the meioses follows a Poisson process (since the superposition of Poisson processes is also a Poisson process). Thus, this permits us to apply our mathematically based framework to this area and to conclude that a true shared segment should be twice as long as the average false positive shared segment. However, there is not very much power to distinguish a true peak from a large set of false peaks. To examine this, we simulated a 4,000-cM genome in many different types of relative pairs separated by a fixed number of generations, conditional on sharing a disease gene IBD from one founder. The simulation results, based on 5,000 replicates, show that the observed distributions conform to the predicted ones (fig. 4). In length-biased sampling, the harmonic mean of the true positive length distribution is equal to the arithmetic mean of the false positive distribution, and the ratio of first and second moments of the false positive distribution accurately predicts the arithmetic mean of the true positive distribution (see Sen 1987), and, since this is a Poisson process, the mean length of the true positives should be twice that of the false positives. However, note that, in any given genome, ~25% of the false segments are longer than the single true segment.

Now consider the effects of interference on these shared chromosomal regions. If there is interference, then the recombination process is no longer Poisson, but, according to the Superposition section above, as the number of connecting meioses increases, the effective number of superimposed renewal processes increases, and the limiting distribution should approach a Poisson process. However, for relatives separated by only a small number of generations, the behavior may not be consistent with that expected for a renewal process, since the interarrival times are no longer identically distributed (Mecke 1969). To explore this, we repeated the shared-segment simulation mentioned above under a number of different Erlang renewal-process models of interference; as expected, there was less difference between true and false peak lengths than when there was no interference. However, as the number of meiotic steps increased, the length ratio between true and false shared segments gradually approached the ratio of 2, expected under a Poisson process.

Discussion

Length-biased sampling occurs whenever one chooses to observe a renewal process at a specific point *t*. So the peaks (if there are any) covering any arbitrarily prespecified point *t* will be longer than average. However, in our case, the definition of the disease prespecifies the point *t,* and ascertainment of pedigrees segregating for the disease increases the chance that this point *t* will actually be covered by a peak above the significance threshold. Nature determines which peaks are true and which are false, and so true peaks are longer (on average) than false peaks of the same height.

If true peaks are, on average, longer than false peaks, then a test based on both length and height might perhaps be more powerful than a test based on height alone, since it is using more information. However, such a test would have an additional df, as compared with a test based on height alone. This would have to be compensated for—typically each additional df increases the required likelihood ratio by a factor of 2 (see Terwilliger and Ott 1994). The ratio of the density functions, $dG(x)/dF(x)$, for a given length x, can be thought of as the ratio of the likelihood of a given length coming from the true distribution G versus the likelihood of it coming from the false distribution F. Note that this likelihood ratio equals $x/E_F[X]$ (Corollary A). More than half the time this likelihood ratio will be <2 (since $E_G[X] \leqslant 2E_F[X]$ if the index of dispersion is <1; see Corollary B), and so using length and height jointly may often

Table 2

Number of False and True Peaks Observed in 1,000 Simulated Genomes, with Five Disease Genes per Genome, by Height Class

	$\psi = 1$						$\psi = .70$			
	$\xi = 10$		$\xi = 5$		$\xi = 1$		$\xi = 10$		$\xi = 5$	
P	False	True	False	True	False	True	False	True	False	True
.01	3,512	484	4,355	450	7,518	393	7,572	471	10,156	463
.003	1,791	633	2,106	616	3,634	546	4,207	732	5,273	691
.0009	860	723	980	702	1,806	633	1,884	853	2,323	786
.0003	315	570	354	547	771	496	820	784	968	752
.0001	120	453	141	454	377	435	348	651	423	622
.00004	46	353	55	358	225	350	140	431	159	429
.00001	19	266	21	273	131	273	55	311	73	296
.000006	7	126	8	130	54	129	25	195	27	190
.000002	3	133	5	136	47	135	6	129	8	125
.0000008	1	54	1	59	28	60	1	63	2	64
.0000003	0	42	3	41	16	40	2	41	2	41
.0000001	0	22	0	21	11	21	1	28	1	29
.00000004	0	23	1	23	10	23	2	14	1	15
.00000001	0	7	0	7	2	7	0	10	0	10
.000000005	0	6	0	6	1	6	0	5	0	5
.000000002	0	7	0	7	1	7	0	3	0	3
.0000000008	0	3	0	4	0	4	0	2	0	2
.0000000003	0	5	0	5	0	5	0	1	0	1
Total	6,674	3,910	8,030	3,839	14,632	3,563	15,063	4,724	19,416	4,524
$P < .01$:										
E[genes detected][a]	3.910		3.839		3.563		4.724		4.524	
P(true|detected)[b]	.37		.32		.20		.24		.19	

[a] Average number of peaks per genome scan.

[b] Posterior probability of a true peak, calculated as the total number of true peaks divided by the total number of peaks.

not even compensate for the extra df. Even so, length information can be helpful: our results indicate that the most efficient use of length information is to exclude very short peaks from immediate consideration. In fact, if the false interval lengths are exponentially distributed, then a length threshold that excludes 25% of the shortest peaks will exclude only 3% of the true peaks; excluding 50% of the shortest peaks will exclude only 15% of the true peaks. Note that the common practice of screening a genome with a sparse map effectively excludes the shortest peaks—and so biases toward finding true positives.

It is important to note that false peak rate and behavior stay the same as sample size increases (provided that the sample size is "big enough" to begin with); increasing the sample size only influences the true peak behavior. To show that false peak rates remain stable, let us consider the behavior of the sequence W_k as a function of genetic distance. Let us make the simplifying assumption that the recombination events occur according to a Poisson process. A change in W_k occurs when there is a change in $A_{k,n}$ in some sib pair n. If we consider the transmission from each parent to sib pair independently, the probability of a recombination changing a given sib pair from IBD to not IBD, or vice versa, is $1 - R = 2\theta(1 - \theta)$, which is $\sim$.02 when the intermarker distance is 1 cM (as in our simulation). For fully informative loci typed on N sib pairs, if, at marker k, there are α_k alleles IBD and $(2N - \alpha_k)$ alleles not IBD, then the expected number of alleles IBD at the next marker is just $E(\alpha_{k+1}|\alpha_k) = R\alpha_k + (1 - R)(2N - \alpha_k)$, and $\mathrm{Var}(\alpha_{k+1}|\alpha_k) = 2NR(1 - R)$. Therefore, $E(W_{k+1}|\alpha_k) = (2R - 1)W_k = .96W_k$, and $\mathrm{Var}(W_{k+1}|\alpha_k) = 4R(1 - R) = .0784$, when $\theta = .01$. Note that $\mathrm{Var}(W_{k+1}|\alpha_k)$ is independent of N. In other words, as we move along the chromosome, the value of the current statistic is correlated with the previous value, and the level of correlation is independent of sample size. We verified this via simulation and found that the expected number of false peaks and their lengths were not changed as a function of the sample size—an observation that holds empirically for partially informative markers as well. The limiting behavior of false positive statistics has been studied by Lander and Schork (1994) and Lander and Kruglyak (1995).

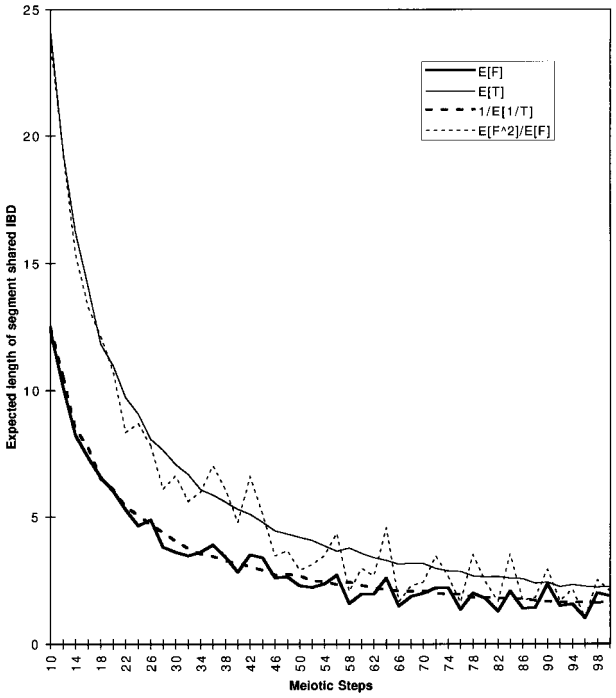

Figure 4 Expected lengths of IBD sharing for true (T) and false (F) peaks, as a function of degree of relationship (i.e., meiotic steps). These results are from a simulation of a 4,000-cM genome (5,000 replicates). For the true peaks, the members of the current generation had to share a disease gene IBD from one founder. The harmonic mean $1/E[1/T]$ of the true-peak lengths should equal the mean of the false peaks $E[F]$. $E[F^2]/E[F]$ should equal the mean of the true peaks $E[T]$.

Our main simulation study here was done by single-marker analyses. It is important to consider how our results might have been altered had we defined peaks in terms of multipoint statistics, rather than in terms of single-point statistics. For the fully informative case, the peak shape would stay essentially the same, because we used such a dense map of markers, and so our results would remain the same. For the partially informative case, multilocus analysis should increase the informativity of the analysis, and so the results would be more similar to the fully informative results than to the partially informative single-point results. In either case, length-biased sampling holds.

Conclusion

We have established, by analytical arguments and by simulation experiments, that true peaks are in fact longer than false peaks of similar height and that longer peaks are more likely to contain the gene of interest than are shorter peaks. We have shown that these differences have the potential to aid in linkage mapping, mainly by permitting us to exclude from immediate consideration the shortest peaks; however, we do not know how much these differences will aid in distinguishing true peaks from false peaks of the same height; this merits further investigation, since preliminary results by Goldin and Chase (in press) indicate that statistics that use both height and length may have more power than do those based on height alone.

Acknowledgments

This work was supported by the Wellcome Trust Center for Human Genetics, National Institutes of Health (NIH) grant HG00719 (to D.E.W.), the Association Française Contre Les Myopathies, the University of Pittsburgh, NIH grant HG00008, a Hitchings-Elion Fellowship from the Burroughs-Wellcome Foundation (to J.D.T.), and the W. M. Keck Center for Advanced Training in Computational Biology at the University of Pittsburgh, Carnegie Mellon University, and the Pittsburgh Supercomputing Center. Fruitful discussions and input from Cyrus Derman, Fan-Hui Kong, Janet Sinsheimer, and Martin Farrall are gratefully acknowledged. We would also like to thank the reviewers for their help in improving this paper.

References

Bailey NTJ (1961) Introduction to the mathematical theory of genetic linkage. Clarendon Press, Oxford

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

Brown DL, Gorin MB, Weeks DE (1994) Efficient strategies for genomic searching using the affected-pedigree-member method of linkage analysis. Am J Hum Genet 54:544–552

Cox DR (1962) Renewal theory. Methuen, New York

Cox DR, Isham V (1980) Point processes. Methuen, New York

Cox DR, Lewis PAW (1966) The statistical analysis of series of events. Barnes & Noble, New York

Ewens WJ, Asaba B (1984) Estimating parameters of the family-size distribution in ascertainment sampling schemes: numerical results. Biometrics 40:367–374

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

Feller W (1971) Introduction to probability theory and its applications, 2d ed. Vol 2. John Wiley & Sons, New York

Goldin LR, Chase GA. Improvement of the power to detect complex disease genes by regional inference procedures. Genet Epidemiol (in press)

Goldin LR, Chase GA, King TM, Badner JA, Gershon ES (1995) Use of exact and adjusted liability scores to detect genes affecting common traits. Genet Epidemiol 12:765–769

Grigelionis B (1963) On the convergence of sums of random step processes to a Poisson process. Theory Prob Appl 8: 177–182

Guo S-W (1996) Gametogenesis processes and multilocus gene identity by descent. Am J Hum Genet 58:408–419

Haldane JBS (1938) The estimation of the frequency of recessive conditions in man. Ann Eugenics 7:255–262

Hemenway D (1982) Why your classes are larger than 'average.' Math Magazine 55:162–164

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

Khinchin AI (1960) Mathematical methods in the theory of queueing (in Russian). Translated by Andrews DM, Quenouille MH. Vol 7 in: Griffin's statistical monographs and courses. Hafner, New York

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

Lander ES, Schork NJ (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

McFadden JA (1962) On the lengths of intervals in a stationary point process. J R Stat Soc [B] 24:364–382

Mecke J (1969) Verscharfung eines Satzes von McFadden. Wiss Z Friedrich-Schiller-Universitat Jena 18:387–392

Morton NE (1991) Parameters of the human genome. Proc Natl Acad Sci USA 88:7474–7476

Nelson R (1995) Probability, stochastic processes, and queueing theory. Springer-Verlag, New York

Owen ARG (1948) The theory of genetical recombination. I. Long chromosome arms. Proc R Soc Lond B Biol Sci 136: 67–94

Palm C (1943) Intensitatsschwankungen im Fernsprechverkehr. Ericsson Tech 44:1–189

Patil GP, Rao CR (1978) Weighted distributions and size-biased sampling with applications to wildlife populations and human families. Biometrics 34:179–189

Resnick SI (1994) Adventures in stochastic processes: the random world of Happy Harry. Birkhauser, Boston, Basel

Ross SM (1983) Stochastic processes. John Wiley & Sons, New York

Samuels SM (1974) A characterization of the Poisson process. J. Appl Prob 11:72–85

Scheaffer RL (1972) Size-biased sampling. Technometrics 14: 635–644

Schotz WE, Zelen M (1971) Effect of length sampling bias on labeled mitotic index waves. J Theor Biol 32:383–404

Sen PK (1987) What do the arithmetic, geometric, and harmonic means tell us in length biased sampling? Stat Prob Lett 5:95–98

Shannon WD, Goldin LR, Chase GA, Weeks DE (1995) Distinguishing true and false peaks in allele-sharing statistics. Am J Hum Genet Suppl 57:A35

Simon R (1980) Length-biased sampling in etiologic studies. Am J Epidemiol 111:444–451

Smith WL (1958) Renewal theory and its ramifications. J R Stat Soc B 20:243–302

Sturt E (1976) A mapping function for human chromosomes. Ann Hum Genet 40:147–163

Terwilliger JD, Ott J (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore

Terwilliger JD, Speer MC, Ott J (1993) Chromosome based method for rapid computer simulation in human genetic linkage analysis. Genet Epidemiol 10:217–224