Am. J. Hum. Genet. 54:544–552, 1994

Efficient Strategies for Genomic Searching Using the Affected-Pedigree-Member Method of Linkage Analysis

Deborah L. Brown,* Michael B. Gorin,*,† and Daniel E. Weeks*

*Department of Human Genetics and †Department of Ophthalmology, University of Pittsburgh, Pittsburgh

Summary

The affected-pedigree-member (APM) method of linkage analysis is a nonparametric statistic that tests for nonrandom cosegregation of a disease and marker loci. The APM statistic is based on the observation that if a marker locus is near a disease-susceptibility locus, then affected individuals within a family should be more similar at the marker locus than is expected by chance. The APM statistic measures marker similarity in terms of identity by state (IBS) of marker alleles; that is, two alleles are IBS if they are the same, regardless of their ancestral origin. Since the APM statistic measures increased marker similarity, it makes no assumptions concerning how the disease is inherited; this can be an advantage when dealing with complex diseases for which the mode of inheritance is difficult to determine. We investigate here the power of the APM statistic to detect linkage in the context of a genomewide search. In such a search, the APM statistic is evaluated at a grid of markers. Then regions with high APM statistics are investigated more thoroughly by typing more markers in the region. Using simulated data, we investigate various search strategies and recommend an optimal search strategy that maximizes the power to detect linkage while minimizing the false-positive rate and number of markers. We determine an optimal series of three increasing cut-points and an independent criterion for significance.

Introduction

The development of technology to readily type individuals at molecularly based markers means that the genetic linkage maps of the human genome are rapidly becoming better in terms of the density of highly polymorphic markers (NIH/CEPH Collaborative Mapping Group 1992; Weissenbach et al. 1992). Thus, it is becoming easier and less costly to carry out a genomic search for a disease gene. In such a search, a large number of markers throughout the genome are tested for linkage to the disease. Then those regions with evidence for linkage are explored more thoroughly by typing more markers in the region (Elston 1992). When carrying out such a search, the investigator faces several choices involving the criteria for determining whether a region should be investigated further and how thoroughly to evaluate the evidence for linkage at each stage. For example, one could seek to reduce the costs and computational overhead of the initial stages by typing a reduced subset of people and by using a relatively rapid test for linkage, instead of typing everyone and calculating time-consuming multipoint lod scores. We investigate here the application of a rapid model-free test for linkage, the affected-pedigree-member (APM) method, which requires typing of only the affected individuals (Weeks and Lange 1988, 1991, 1992; Weeks et al. 1992). The APM method has the advantage of using marker information on all the affected relatives in each family, unlike the sib-pair methods which only use information on the siblings and their parents.

In addition to reducing the marker-typing requirements, the APM method has another advantage over the traditional lod score approach. The traditional

Received June 11, 1993; accepted for publication October 29, 1993.

Address for correspondence and reprints: Daniel E. Weeks, Ph.D., University of Pittsburgh, Department of Human Genetics, 130 DeSoto Street, A300 Crabtree Hall, Pittsburgh, PA 15261.

© 1994 by The American Society of Human Genetics. All rights reserved.

0002-9297/94/5403-0018$02.00

methods of linkage analysis, which test for cosegregation of disease and marker, were developed for Mendelian diseases (Morton 1955). These methods may be misleading or inconclusive when applied to complex non-Mendelian diseases, because they require simplifying assumptions about the inheritance of the disease in order to infer disease-gene location. In other words, an incorrect model of the disease may lead to incorrect conclusions. Model-free linkage-analysis methods, such as the APM method, avoid this problem by directly testing for nonrandom segregation of markers to affected individuals. The model-free methods have a crucial and important role in the search for disease-susceptibility loci involved in genetically complex diseases.

Ideally, an investigation of the statistical properties of a genomewide search should consider the correlations between closely linked markers. However, this correlational structure is difficult to define analytically. As a result, analytical analyses of the problem have taken two main approaches. In the first approach, the markers are assumed to be spaced far enough apart so that the correlations may be ignored (Risch 1991). Elston (1992) has developed an analytical approach for designing an optimal genome-searching strategy using data from a specific class (e.g., sibs, grandparent-grandchild, etc.) of pairs of affected relatives. He found that a two-stage strategy can save 25%–60% in the cost of a study, over a one-stage strategy (Elston 1992). In the second approach, there is no space between adjacent markers (Lander and Botstein 1989; Feingold et al. 1993), so that markers are almost perfectly correlated. While these approaches generate valuable insights into mapping with multiple markers, we chose here to use a simulation-based approach, which should accurately reflect the marker correlational relationships encountered in a real study.

Here we explore the power of the APM-method to detect linkage in the context of a genomewide search (Weeks and Brown 1993). Using simulated data, we evaluated various search strategies and found an optimal strategy that maximizes power to detect linkage while minimizing the false-positive rate and number of markers. In the optimal strategy, if an APM statistic is above a certain cut-point, then that region is investigated with more markers. We have investigated different spacing of the starting grid of markers, different numbers of cut-points, and different criteria for when an area should be further investigated and for declaring a marker significant. We have found a search strategy that leads to a marked increase in power while keeping the overall false-positive rate below 5%.

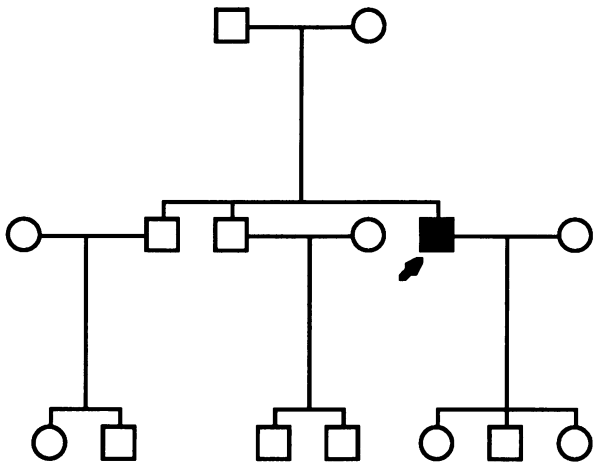

Figure 1 Pedigree used for simulation study (Boehnke et al. 1988). The proband, identified by the arrow, is always affected.

Methods

We simulated data on pedigrees that had 15 members in three generations (Boehnke et al. 1988) (fig. 1). We used a disease model with an autosomal dominant mode of inheritance, full penetrance, and no phenocopies. The disease allele had a frequency of .001, yielding an incidence of .001999. Markers with four equifrequent alleles were simulated every 2.5 cM on 22 autosomal chromosomes with realistic sex-averaged lengths, for a total of 1,891 markers. Lengths were obtained from the CEPH/NIH linkage map (NIH/CEPH Collaborative Mapping Group 1992). These lengths are believed to provide an upper limit for genetic distances, since undetected genotyping errors inflate map length. Note that overestimating the chromosome lengths would be expected to increase the false-positive rate in our simulation by providing more unlinked markers than would be encountered in reality. The disease locus was halfway between the 40th and 41st markers on chromosome 10 at 101.25 cM.

For each chromosome, data was simulated as follows: (1) Assign multilocus genotypes with phase to every founder in the pedigree. (A founder is a person whose parents are not in the pedigree.) These assignments are made assuming Hardy-Weinberg and linkage equilibrium. (2) Generate a gamete for those nonfounders whose parents have already been assigned ge-

notypes. The gamete is generated by moving locus by locus down the parental chromosomes, switching from one to the other when a recombination event occurs. In each interval, a recombination event is generated if a uniformly distributed random number is less than the recombination fraction. Interference is not taken into account. (3) Repeat step 2 until everyone has been assigned a genotype.

As a time-saving measure, the chromosome with the disease locus was simulated first, and the affected status of each individual was determined. Families were ascertained through an affected proband (fig. 1) and retained if there was at least one pair of affected individuals other than a parent-child pair. (Since a parent and child must share half their genes in common, they are uninformative for the APM method.) We ascertained families until at least 300 affected individuals had been collected; this comprised one replicate. Once a replicate was completed, we calculated and stored the APM statistic for each marker. We collected 1,000 sets of APM statistics for 1,891 markers.

Search strategies were then developed to be tested on the simulated data. Here we define the terms we use to describe a strategy. A stage is defined by how many times a region has been examined. The first stage always consists of a grid of equally spaced markers (fig. 2). The initial spacing of those first-stage markers must be specified. For each stage other than the last, a "cut-point" must be specified. The set of cut-points defines how the search proceeds: if a marker has an APM statistic above the cut-point, then its region will be investigated further at the next stage. After all the markers are typed, their APM statistics must be compared to the significance threshold; a typed marker is considered to be positive if its APM statistic is above the threshold. A positive result is called a "true positive" if it is within a certain distance from the disease locus and is a false positive otherwise. A strategy is defined by the spacing in the initial grid, the number of cut-points and their values, the significance threshold (which can be defined to be the last cut-point), and whether we carry out pairwise or nonpairwise comparisons (described below).

As an example, consider a nonpairwise two-cut-point strategy with an initial grid of 20 cM (fig. 2*a*). In the first stage, we start with a grid of markers every 20 cM and type markers A, B, and C. If B is above the first cut-point, then we type the two markers 10 cM to either side—i.e., markers D and E. In the second stage, the APM statistic of the originally examined marker, B,

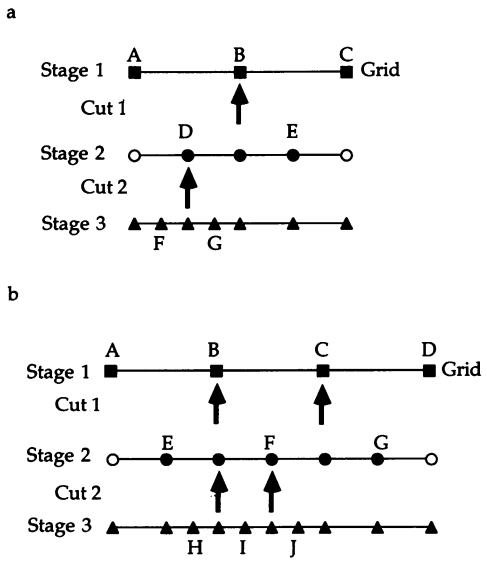

Figure 2 Nonpairwise (*a*) and pairwise (*b*) (two-cut-point) strategies. In each stage, lettered markers are the ones newly typed, shaded markers are evaluated with respect to the cut-point (stage 3 markers are compared with the threshold), arrows indicate those markers with APM statistics greater than the current cut-point, and unshaded markers are carried over to the final stage but are not reevaluated at intervening stages.

and those of the two on either side will then be considered relative to the second cut-point. If the APM statistic for D is higher than the second cut-point, we then examine the markers 5 cM to either side—i.e., markers F and G. Finally, the APM statistic of each marker we have typed is compared with the threshold, and the numbers of true positives and false positives are counted.

In a nonpairwise strategy, a region is explored more if only a single marker is above the current cut-point. In contrast, a pairwise strategy requires that two adjacent markers (in the current stage) both be above the cut-point. A pairwise strategy may more specifically detect true-positive regions, since many of the markers in the region of the disease-susceptibility locus should have high APM statistics. In addition, on an unlinked chromosome, the odds of more than one high value at nearby markers should be small. Figure 2*b* displays an example of a two-cut-point, pairwise strategy. Markers A, B, C, and D are examined in the initial grid. If B and C are both above the first cut-point, then E, F, and G are typed in the second stage. We then make four pairwise considerations: E and B, B and F, F and C, and C and G. If, for example, only B and F are above the

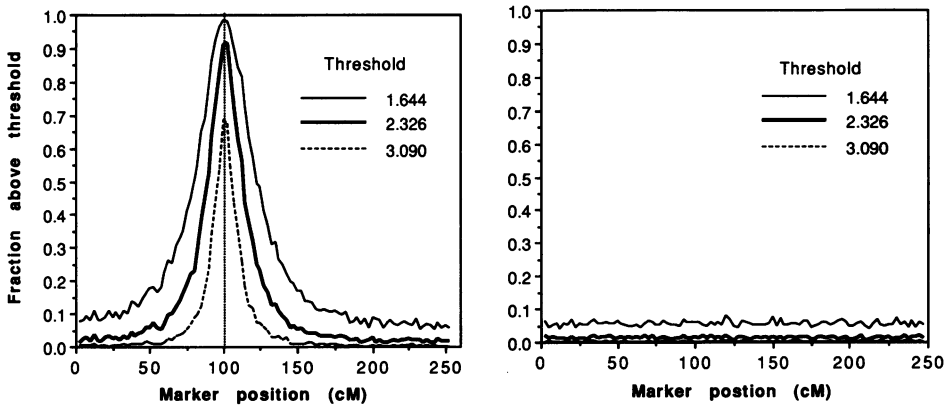

Figure 3 Distribution of the APM statistics along two chromosomes. The linked chromosome 10 is on the left, with the disease locus at 101.25 cM (vertical line). On the right is chromosome 11, which is unlinked to the disease.

second cut-point, we then look at H, I, and J in the third stage. We compare the APM statistics of A–J to the threshold to determine which, if any, are true positives and which are false positives.

We attempted to determine the optimal strategy in terms of the number of stages, the spacing of typed markers, the cut-points, and the threshold. In an ideal strategy, we would like the false-positive rate alpha (α) low, the power ($1-\beta$) high, and the number of markers typed per replicate low. More precisely, α is the fraction of replicates having a false-positive result, and $1-\beta$ is the fraction of replicates having a true-positive result.

Therefore, we constructed a cost function designed to minimize α and the number of markers while maximizing power. A strategy is optimal if it has the minimum cost. Note that the results we obtain are optimal under this particular cost function. If a different cost function were used, with, for example, more emphasis placed on power, we would expect different optimal results. The power, α, and the number of markers typed per replicate must be weighted appropriately in the cost function so that the function value can be used to assess the strength of a search strategy. We designed our cost function so that the false-positive rate (α) would be kept under 5% by imposing a substantial penalty if α is above 5%. We also defined our cost function so that a power below 75% was penalized. To make the power term comparable to the alpha term, the power term was divided by four. In our cost function, we weight the marker term μ by a constant k; the magnitude of k reflects how critical it is to keep the number of markers low. We report strategies where we weight the number of markers by differing amounts.

We used the function $F(C, T) = \gamma(\alpha) - [\varepsilon(\beta)/4] + \mu/k$, where C = the set of cut-points, T = the significance threshold,

$$\gamma(\alpha) = \begin{cases} \alpha & \text{if } \alpha < .05 \\ 10\alpha & \text{if } \alpha \geq .05 \end{cases},$$

$$\varepsilon(\beta) = \begin{cases} 2(1-\beta) & \text{if } (1-\beta) \geq .75 \\ (1-\beta) & \text{if } (1-\beta) < .75 \end{cases},$$

μ = average number of markers over all replicates, and k = weighting term for the number of markers (40,000).

Optimization Design

For each strategy, a number of selected sets of cut-points were input to start a downhill simplex optimization routine (AMOEBA; Press et al. 1986). The program evaluates the function value for each set of cut-points and iteratively changes each set to minimize the function value. A direction set method (POWELL; Press et al. 1986) was also used for confirmation and to come closer to the global minimum. While the strategy reported may not yield the global minimum, the differences in function values among the local minima are very small. It appears that there are multiple ways to achieve similar power ($1-\beta$) and α and that the truly optimal cut-points would probably not yield function values very different from the values obtained.

In the first stage, we type evenly spaced markers starting at a given marker on each chromosome. In any real situation, the actual position of the disease locus is

Table 1

The Best Alpha, Power, Average Number of Markers, and Cost-Function Value for Each Search Strategy, with Corresponding Cut-Points and Significance Threshold

Strategy	No. of Cut-Points	Grid Spacing (cM)	Pairwise	Independent Threshold	Cut 1	Cut 2	Cut 3	Threshold[a]	α	Power	Markers	Function Value
A	2	10	No	No	2.08	4.22	...	LCP	.031	.521	502.0	−.0867
B	2	20	No	No	1.96	3.99	...	LCP	.035	.387	260.3	−.0552
C	3	20	No	No	1.67	2.08	3.99	LCP	.039	.651	287.8	−.1166
D	3	20	No	Yes	1.67	2.00	2.26	3.98	.044	.838	305.4	−.3674
E	3	20	Yes, first	Yes	.72	.85	2.34	4.00	.046	.814	305.8	−.3534
F	3	20	Yes, all	Yes	.72	1.35	1.56	4.00	.046	.839	299.9	−.3660
G	3	40	No	Yes	.73	1.57	2.01	3.99	.031	.643	212.3	−.1244
H	3	40	Yes, first	Yes	−.24	1.28	2.08	3.99	.031	.642	216.0	−.1241
I	3	40	Yes, all	Yes	.36	1.51	1.98	3.92	.030	.578	166.2	−.1103

[a] LCP = last cut-point.

unknown with respect to these first-stage markers. However, in our simulation study, the disease location is known and fixed. Therefore, the choice of the starting point for our grid may influence the power estimates. For example, if we start a 10-cM grid at marker 1, then the marker nearest the disease locus is 1.25 cM away, while if we start the grid at marker 3, then the nearest marker is 3.75 cM away. In the first case, we have a marker in the initial grid that is very close to the disease locus and therefore would expect to have more power than in the second case, where the closest marker in the initial grid is 3.75 cM away.

To alleviate this problem, we attempted to pick the most conservative starting point. However, we found that the "worst" starting point is actually variable and is dependent on which cut-points are chosen. Since it is one of the worst starting points, we start the initial grid at the second marker on each chromosome. Also, in some real cases, the disease-susceptibility locus may be very near the end of a chromosome. In such a case, there would be fewer flanking markers, and we would expect the power to be reduced.

A typed marker is considered a true positive if its APM statistic is above the final threshold and if it is within a certain distance from the disease locus. This distance is a critical criterion. One could imagine a situation where every significant marker syntenic to the disease locus was scored as a true positive. Allowing this would obviously inflate the true-positive rate. Alternatively, we could define true positives as only those significant results within a very small distance from the disease locus. However, the high APM statistics do spread rather far from the disease locus (fig. 3), and thus it would not be necessary to penalize the positive results resulting from that spread. We evaluated the effect of changing the size of the interval in which a significant result is considered a true positive. For 20-, 30-, and 40-cM intervals, the power was essentially the same, while the 20-cM interval gave a false-positive rate more than double the rates generated by the 30- and 40-cM intervals. On the basis of these results, we chose to call a significant result a true positive if it was within 30 cM of the disease locus. While 30 cM is a relatively large distance in terms of genetic mapping, we felt that it did not make sense to classify a significant result as a "false positive" if it was within 30 cM of the true location of the disease locus.

Experimental Issues

Two-cut-point strategy versus three-cut-point strategy.—Because we observed that the first two cut-points in a three-cut-point strategy tend to be similar, we investigated two-cut-point strategies. A two-cut-point strategy could be beneficial because it would be simpler to implement and would require fewer markers than a three-cut-point strategy would require.

Initial grid spacing.—The separation of markers in the original grid has the greatest effect on how many markers must be typed for the whole search. The finer grid would be expected to pick up more true positives but could also increase the false-positive rate. Optimal cut-point sets were determined for 10-cM versus 20-cM initial spacings for a two-cut-point strategy and for 20 cM versus 40 cM for a three-cut-point strategy.

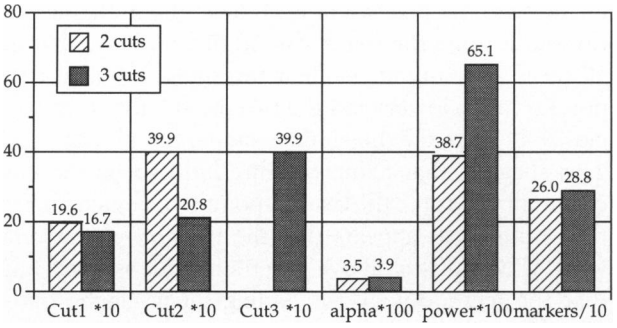

Figure 4 Comparison of two cut-points vs. three cut-points in a 20-cM grid, with the threshold-equals-last-cut-point, nonpairwise strategy.

Independent threshold versus last cut-point threshold.—Once we type a number of markers, we need to decide which of them are significant. A typed marker is considered to be positive if its APM statistic is above the final threshold. It is customary to use the last cut-point as that threshold. However, it may be beneficial to optimize a separate independent threshold in addition to the cut-points. We examined whether it is better to use an independent threshold or to use the last cut-point as the criterion for significance.

Pairwise versus nonpairwise.—We tested a pairwise strategy in which, at each stage, we simultaneously looked at two markers. If they were both above a certain cut-point, then the region was typed with more markers in the next stage (see example above, fig. 2*b*). Another strategy, the first-pairwise strategy, involves doing the first stage in a pairwise manner but then doing the subsequent stages using the original (single-comparison) method.

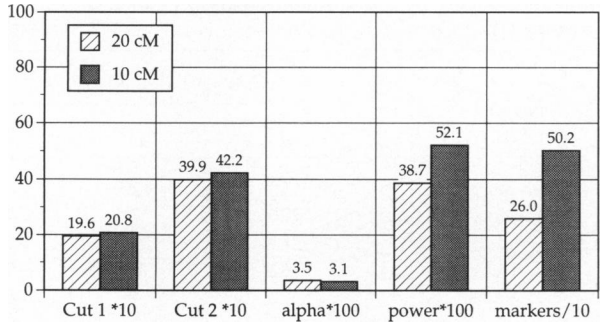

Figure 5 Comparison of initial grid spacing of 20 cM vs. 10 cM in a two-cut-point, threshold-equals-last-cut-point, nonpairwise strategy.

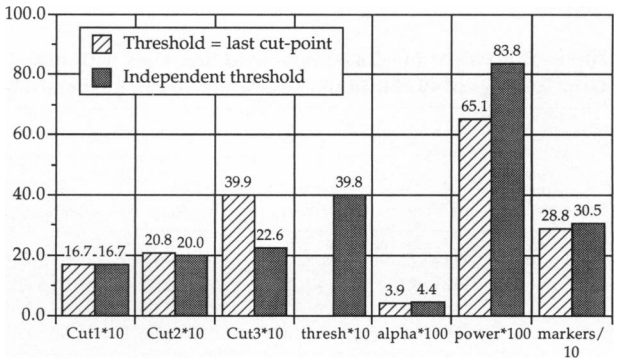

Figure 6 Comparison of a nonindependent threshold vs. independent threshold in a 20-cM grid, three-cut-point, nonpairwise strategy. Note the substantial increase in power with slight effect on alpha and the number of markers.

Table 1 displays the strategies we evaluated. We compare the results of these various strategies to those obtained by performing a complete search in which every marker is typed and also to those results obtained by typing only the initial grid of markers.

Results and Discussion

In order to collect at least 300 affected individuals in each replicate, we ascertained an average of 110 ± 18 (mean ± SD) probands. Of the families of those probands, an average of 54 (± 2) had the required affected-affected pair other than a parent-child pair. These families yielded an average of 302.6 (± 2.1) affected individuals, with an average number of affected individuals per family of 5.64.

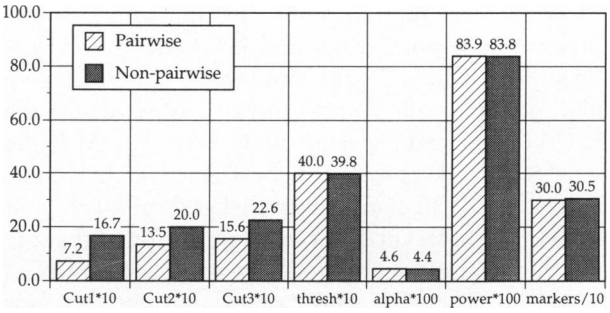

Figure 7 Comparison of pairwise vs. nonpairwise in a 20-cM grid, three-cut-point, independent threshold strategy. Note the similarity in alpha, power, and the number of markers between the two strategies. However, there is a substantial difference in the cut-points used to obtain the same results.

Table 2

Alpha and Power for One-Stage Grid Searches with Initial Grid Spacings of 40 cM, 30 cM, 20 cM, and 10 cM, Respectively

	ALPHA (power) FOR			
THRESHOLD	127 Markers	166 Markers	243 Markers	474 Markers
1.644	.999 (.827)	1.000 (.956)	1.000 (.989)	1.000 (1.000)
2.326	.839 (.527)	.914 (.789)	.973 (.928)	.999 (.992)
3.090	.256 (.184)	.351 (.482)	.431 (.626)	.669 (.845)
3.981	.017 (.034)	.025 (.163)	.035 (.256)	.076 (.397)

The APM statistics at unlinked loci should be asymptotically normally distributed with a mean of 0.0 and a variance of 1.0. Results in figure 3 show the distribution of the statistics on the linked chromosome and a representative unlinked chromosome. The probabilities of obtaining statistics greater than certain values are illustrated. We would expect 0.1% of the random APM statistics to be greater than 3.090 in a normal distribution, 1% to be above 2.326, and 5% to be above 1.644; this is approximately what we observed. Note that the peak of elevated APM statistics has a 25–50-cM spread around the disease locus.

Two-Cut-Point Strategy versus Three-Cut-Point Strategy

The three-cut-point strategy is much more powerful than the two-cut-point strategy (table 1, strategies B and C; and fig. 4). The three-cut-point method has more power than the two-cut-point method because interesting regions are explored in more detail.

Initial Grid Spacing

In the strategy with a 10-cM grid, the average number of markers approximately doubles, with a similar but lower false-positive rate and 35% better power than the strategy with a 20-cM grid (fig. 5). If we include a mild penalty for the number of markers, we determine that the initial grid spacing could be at 10 cM if the number of markers were not constrained (table 1, strategies A and B). In comparing initial grids of 20 cM and 40 cM in a three-cut-point strategy, we found that the 40-cM grid gives much worse power (table 1, strategies D and G). The high APM values apparently do not spread far enough to make 40 cM an effective grid.

Independent Threshold versus Last Cut-Point Threshold

Allowing the significance threshold to vary independently greatly increased the power (to 83.8%) and reduced the false-positive rate, compared with fixing the threshold to be the last cut-point (fig. 6). Note that in all cases with an independent threshold, the final cut-point is much lower than the threshold (table 1, strategies D–I). Also, the threshold is similar for all strategies. This suggests that fixing the threshold to be the last cut-point will force the last cut-point to be much higher than optimal. It appears that the fraction of markers with APM statistic above a certain value is fixed but that the methodology for finding them can be optimized.

Pairwise versus Nonpairwise

We found it surprising that a pairwise strategy is not markedly better than a nonpairwise strategy (table 1, strategies D–F and G–I; fig. 7); a pairwise strategy should be more powerful, since many of the markers in the region of the disease-susceptibility locus should have high APM statistics. In addition, on an unlinked chromosome, the odds of obtaining more than one high value at nearby markers should be small. In other words, we would expect that on the disease chromosome there would be correlations between the APM statistics for adjacent markers and that the pairwise strategy would take advantage of these. To determine why the pairwise strategy was not better than the nonpairwise strategy, we computed the correlations of the APM statistics across replicates and found that the correlations were not markedly elevated on the disease chromosome, as compared with those on the nondisease chromosomes. This is probably due to the fact

Table 3

The Distribution of the 1,000 Replicates in Terms of How Many True Positives and False Positives Were Obtained in Each Replicate by the Optimal Strategy (Table 1, Strategy D)

	NO. OF FALSE POSITIVES		
NO. OF TRUE POSITIVES	0	1	2
0	150	12	0
1	270	7	1
2	285	9	3
3	152	6	0
4	72	4	0
5	23	1	0
6	3	1	0
7	0	0	0
8	1	0	0

Table 4

Cut-Point Series, Thresholds, and Results for Various Values of *k*

k	Cut-Point 1	Cut-Point 2	Cut-Point 3	Threshold	Alpha	Power	No. of Markers
100	7.6790	.4006	.9849	4.0037	.032	.245	243.000
400	2.4646	.7964	2.4008	3.9983	.038	.757	271.644
1,000	2.2298	1.2907	2.4412	3.9914	.039	.796	279.506
40,000	1.6727	1.9993	2.2578	3.9814	.044	.838	305.387
100,000	1.6593	1.9992	2.2178	3.9912	.042	.834	306.949
Infinity	1.8119	1.7484	2.3777	4.0118	.044	.822	303.782

NOTE.—A low value of *k* emphasizes the number of markers in the cost function, while a value of infinity ignores the number of markers. These are evaluated in the context of a nonpairwise strategy with a 20-cM grid, three cut-points, and an independent threshold.

that, for a rare dominant disease, affected individuals tend to share one marker allele, while the nonshared marker allele tends to enter into the pedigree at random. Thus, the value of the APM statistic can change quite a bit from one marker to the next, even in the absence of recombination. For a rare recessive disease, affected individuals should tend to share both marker alleles, and the sharing at the next adjacent marker should be similar except for recombination. Therefore, for a recessive disease, the correlations between adjacent statistics are much (4–20 times) higher in the disease region (as verified by a small simulation study). Thus we would expect that a pairwise strategy would be better than a nonpairwise strategy for diseases that are recessive in nature. Since a pairwise strategy is not markedly different from a nonpairwise strategy (table 1, strategies D–F and G–I; fig. 7), we chose the nonpairwise strategy for simplicity.

The Optimal Strategy

Among the strategies considered here, the optimal strategy is a nonpairwise strategy with 20-cM initial grid, three cut-points at 1.67, 2.00, 2.26, and an independent threshold of 3.98 (table 1, strategy D). On 1,000 replicates of simulated data, the optimal strategy yields a false-positive rate of .044 and a true-positive rate of .838, and it requires the typing of an average of 305.4 markers. Note that this is a genomic search strategy and assumes that all areas of interest are explored further.

It is useful to compare the performance of the optimal strategy to a naive strategy of simply evaluating only the markers in the initial grid. If only the 20-cM grid of 243 markers is examined, power drops to .256 with an average reduction of only 62.4 markers and .009 in alpha (table 2). If we evaluate all 1,891 markers and use the same threshold, 3.981, then alpha rises to .248, while power reaches only .866.

The probabilities of finding true and false positives were determined experimentally for several sets of equally spaced markers by using standard significance levels and the optimal strategy threshold (table 2). The optimal strategy is clearly both more specific and more sensitive than any one-stage method. Also, we determined the threshold required to maintain an overall false-positive rate below 5% for one-stage grid strategies of various size. For a 40-cM grid of 127 markers, a threshold of 3.65 is required, while for a 10-cM grid of 474 markers, the threshold must be 4.10.

In an ideal strategy, we would hope to have no false positives. However, we may not be bothered as much by false positives as long as we find a true positive (if it is assumed that there is, in fact, a true disease-susceptibility locus to be found). Table 3 displays the relationship between the number of true positives detected and the number of false positives obtained in the optimal strategy D (table 1). As might be desired, it is rare to have a false positive without at least one true positive: of the replicates with one or more false positives, 73% have one or more true positives. There are only 12 replicates with just one false positive and no true positives, and only 4 with more than one false positive. Note that of 1,000 replicates, the vast majority of replicates (838) have at least one true positive. An average of 1.5 true positives were detected per replicate.

Within the context of the optimal strategy, we investigated the effect of changing the weighting term, k, for the average number of markers (table 4). As expected, with decreasing penalty on the average number of

markers, the number of markers typed in the optimal strategy increases and then plateaus.

Using a large-scale simulation study, we have investigated several strategies for mapping a disease gene by a genomic search. We found that the optimal strategy is a nonpairwise, 20-cM initial grid, three-cut-point strategy with an independent threshold. While the disease model we investigated here is definitely simple, the best strategy may also be optimal for mapping more complex diseases. However, we would not expect that the exact numerical values of the cut-points and the significance threshold would apply to different modes of inheritance. For example, we carried out exactly the same simulation study on a dominant disorder with markedly reduced (50%) penetrance (results not shown) and found that a nonpairwise, 20-cM initial grid, three-cut-point strategy with an independent threshold is again the optimal strategy. In the future, we plan to investigate optimal strategies for mapping more complex diseases, such as diseases due to the interaction of two or more susceptibility loci. Also, this simulation study used only the single-locus APM statistic. A more powerful strategy, which we plan to investigate, may be to first use the single-locus statistic for the initial grid of markers and then to use multilocus APM statistics in subsequent stages.

Acknowledgments

We would like to thank Dr. Kenneth Lange and two anonymous reviewers for their helpful comments and Mark Schroeder for his excellent programming assistance. This work was supported by the University of Pittsburgh, Research to Prevent Blindness, Inc., National Center for Human Genome Research grant HG00719 (to D.E.W.), and National Eye Institute grant EY09859 (to M.B.G.).

References

Boehnke M, Young MR, Moll PP (1988) Comparison of sequential and fixed-structure sampling of pedigrees in complex segregation analysis of a quantitative trait. Am J Hum Genet 43:336–343

Elston RC (1992) Designs for the global search of the human genome by linkage analysis. In: Proceedings of the 16th International Biometric conference: Hamilton, New Zealand, December 7–11. Ruakura Agricultural Center, Hamilton, New Zealand pp 39–51

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

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318

NIH/CEPH Collaborative Mapping Group (1992) A comprehensive genetic linkage map of the human genome. Science 258:67–86

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1986) Numerical recipes: the art of scientific computing. Cambridge University Press, New York

Risch N (1991) A note on multiple testing procedures in linkage analysis. Am J Hum Genet 48:1058–1064

Weeks DE, Brown DL (1993) Genomic searching using the affected pedigree member method of linkage analysis. Paper presented at The Biometric Society spring meetings, Philadelphia, March 21–24

Weeks DE, Harby LD, Sarneso CA, Gorin MB (1992) Using the affected pedigree member method of linkage analysis. INSERM atelier 44: Linkage analysis of single gene and polygenic traits. INSERM, Le Vesinet, France

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

——— (1991) An overview of the affected-pedigree-member method of linkage analysis. In: Keramidas EM, Kaufman SM (eds) Proceedings of the 23rd Symposium on the Interface. Interface Foundation of North America, Seattle, pp 386–391

——— (1992) A multilocus extension of the affected-pedigree-member method of linkage analysis. Am J Hum Genet 50:859–868

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, et al (1992) A second-generation linkage map of the human genome. Nature 359:794–801