Inter- and Intrafamilial Heterogeneity: Effective Sampling Strategies and Comparison of Analysis Methods

Martina Durner,* David A. Greenberg,*'† and Susan E. Hodge‡

Departments of *Psychiatry and †Biomathematics, Mount Sinai Medical Center; and †Departments of Psychiatry and Biostatistics, Columbia University and New York State Psychiatric Institute, New York

Summary

Heterogeneity, both inter- and intrafamilial, represents a serious problem in linkage studies of common complex diseases. In this study we simulated different scenarios with families who phenotypically have identical diseases but who genotypically have two different forms of the disease (both forms genetic). We examined the proportion of families displaying intrafamilial heterogeneity, as a function of mode of inheritance, gene frequency, penetrance, and sampling strategies. Furthermore, we compared two different ways of analyzing linkage in these data sets: a two-locus (2L) analysis versus a one-locus (SL) analysis combined with an admixture test. Data were simulated with tight linkage between *one* disease locus and a marker locus; the other disease locus was not linked to a marker. Our findings are as follows: (1) In contrast to what has been proposed elsewhere to minimize heterogeneity, sampling only "high-density" pedigrees will *increase* the proportion of families with intrafamilial heterogeneity, especially when the two forms are relatively close in frequency. (2) When one form is dominant and one is recessive, this sampling strategy will greatly decrease the proportions of families with a recessive form and may therefore make it more difficult to detect linkage to the recessive form. (3) An SL analysis combined with an admixture test achieves about the same lod scores and estimate of the recombination fraction as does a 2L analysis. Also, a 2L analysis of a sample of families with intrafamilial heterogeneity does not perform significantly better than an SL analysis. (4) Bilineal pedigrees have little effect on the mean maximum lod score and mean maximum recombination fraction, and therefore there is little danger that including these families will lead to a false exclusion of linkage.

Introduction

Genetic heterogeneity (the existence of two or more loci causing the same disease) has long been recognized as a factor that can affect a linkage analysis (Morton 1956). This factor particularly concerns workers in psychiatric genetics, where failure to replicate reported linkages (Egeland et al. 1987; Kennedy et al. 1988; Sherrington et al. 1988; Kelsoe et al. 1989) has been attributed by some (e.g., see Lander 1988) to the existence of genetic heterogeneity. Consequently, geneticists have been trying to determine ways to avoid genetic heterogeneity (Merikangas et al. 1989; Regier and Judd 1989; Goldin et al. 1991).

The difficulty that heterogeneity can cause is that disease form 1 may be linked to the marker being studied while disease form 2 is unlinked, yet neither we nor the linkage analysis can distinguish the two disease forms. Hence those families or individuals with disease form 2 provide evidence *against* linkage and can mislead investigators into rejecting linkage, although disease form 1 really is linked to the marker being studied. Methods have been developed for incorporating heterogeneity into linkage analysis (Smith 1963; Ott 1977; Risch and Baron 1982; Hodge et al. 1983), and the effects of heterogeneity on power and on the ability to detect linkage have been studied (Cavalli-Sforza and King 1986; Martinez and Goldin 1989; Durner and Greenberg 1992). These studies have focused on *interfamilial* heterogeneity—i.e., the

Received February 25, 1992; final revision received June 8, 1992.

Address for correspondence and reprints: Dr. David A. Greenberg, Department of Psychiatry, Box 1229, Mount Sinai Medical Center, One Gustave Levy Place, New York, NY 10029.

0002-9297/92/5104-0020$02.00

situation where some affected families have disease form 1 and others have disease form 2. These analysis methods incorporate the probability that any given family may have the unlinked form of the disease (disease form 2).

However, another, potentially more serious problem is the existence of *intrafamilial* heterogeneity—i.e., the situation where some members within an affected family have disease form 1 and other members of the *same family* have disease form 2 (anywhere in the pedigree). It is more difficult to incorporate intrafamilial heterogeneity into a linkage analysis, and, in fact, computer programs for analyzing such a situation have only recently become available (Lathrop and Ott 1990; N. Schork, M. Boehnke, J. Terwilliger, and J. Ott, personal communication). Concern about intrafamilial heterogeneity has been one of the factors behind the reluctance of many investigators to use bilineal pedigrees (i.e., pedigrees with disease exhibited on both sides of the family) in a linkage analysis (Hodge, in press).

However, little is known about (*a*) the actual magnitude of the intrafamilial heterogeneity problem or (*b*) the relative effectiveness of alternative methods of analyzing families with intrafamilial heterogeneity. Martinez and Goldin (1990) and Goldin (1992, and in press) have looked at both issues, assuming a fixed family size and fixed family structure and relatively restrictive sampling schemes. In the present study, we further address both of these issues, focusing specifically on two questions: (1) When disease heterogeneity exists, what proportion of families exhibit intrafamilial heterogeneity? This proportion is examined as a function of mode of inheritance, gene frequencies, penetrances, and sampling strategy. This is of the utmost importance because the sampling schemes that have been used to ascertain families with complex diseases have generally focused on those rare families with large numbers of affected people. If heterogeneity exists, does this ascertainment scheme increase the probability of finding heterogeneous families? (2) In those data sets exhibiting intra- and interfamilial heterogeneity, what is the best way to analyze the data for linkage? And what specific effect do families with intrafamilial heterogeneity have on the linkage analysis? We simulate families by assuming tight linkage between one of the two disease loci and a marker locus. We then compare a two-locus (2L) heterogeneity analysis, implemented in the program TMLINK (Lathrop and Ott 1990), with a single-locus (SL) approach combining LIPED (Ott 1974) with an admixture test (Hodge et al. 1983).

Methods

Generating Models

Families were generated who had a disease which could be caused by either one of two different loci. The first locus (disease form 1) was tightly linked to the marker locus, with a recombination fraction (θ) of 0. The second locus (disease form 2) was unlinked to both the marker and the first locus. The linked form of the disease (disease form 1) was always dominant, whereas the unlinked form (disease form 2) could be dominant (D + D model) or recessive (D + R model). The gene frequency and penetrance were varied in each model. We use the notation D + D, D + R, etc., to denote 2L *heterogeneity* models, as opposed to DD, DR, etc., for 2L *epistatic* models (see Greenberg 1984).

Generation of Family Data

We expanded the previously reported simulation program (Greenberg 1984, 1989) to generate three-generation families. Three-generation pedigrees were generated according to the following scheme: (1) An at-risk mating type is chosen at random according to the weighted frequency of that mating type in the population. (2) The number of offspring is then chosen for that family. Family sizes were determined according to a negative binomial family size distribution with mean 2.8 and SD 2.3 (Cavalli-Sforza and Bodmer 1971, pp. 310–313). Families with zero offspring or with only one offspring were not included. (3) Recombination status was then determined. For the linked locus, θ = 0, so a recombination event could not occur, and the marker and disease alleles were always inherited together. (4) The chromosomes were then allowed to segregate randomly to the offspring. (5) Once the genotypes of the second generation were determined, the third generation was generated. Unlike the first generation of simulated offspring, zero and one were allowed numbers of children. If a second-generation child was to have children, the genotype of the mate was chosen according to the population allele frequencies. Then steps 2–4 were repeated, with numbers of offspring for each second-generation family member varying from 1 to 10.

The pedigrees that result from the simulation are varied and quite "real" looking (for examples of pedigrees that the simulation produces, see fig. 1). For families selected for the presence of ≥1 affected member in generations 2 and 3, with full penetrance and an ascertainment probability of 1, the average family size was 11, with a minimum of 4 and a maximum of 34.

Selection Procedures

A family was ascertained through generations 2 and 3, on the basis of having a *minimum number* of affected members. Under the "most relaxed" selection scheme, all families in the population with ≥1 affected member were included in the sample. Under the "more stringent" schemes, a family had to have ≥3, ≥5, or ≥9 affected members in order to be included. The different ascertainment schemes follow scenarios that have been proposed to maximize the yield of linkage information (see below).

Proportion of Families with Both Disease Forms

To determine the proportion of families with disease form 1, disease form 2, or both disease forms, we generated data sets under both D + D and D + R models. Different gene frequencies (.01, .05, .10, and .15) and penetrance values (30%, 50%, 70%, and 90%) were used to generate the families. Some combinations of these gene frequencies and prevalences lead to what may be unrealistic population disease prevalences. However, investigators are now looking not

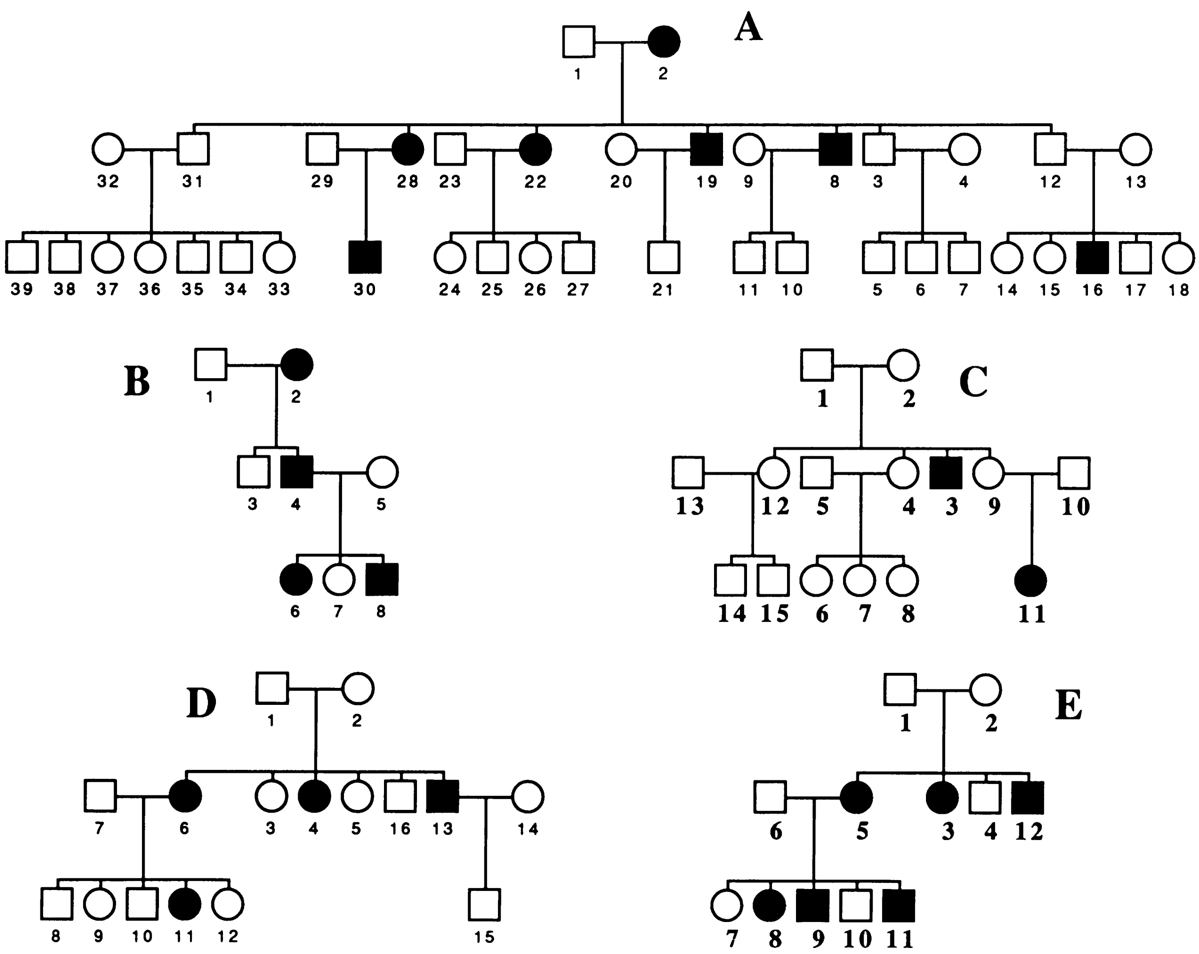

Figure 1 Example of pedigrees produced by the simulation program. These sample families were generated under a dominant mode of inheritance with 50% penetrance and a disease allele frequency of .01. Pedigrees A and E were ascertained under the condition that ≥5 members in generations two and three be affected; pedigrees B and D required that ≥3 be affected; and pedigree C required that ≥1 be affected.

only at disease but also at subclinical phenotypes that may be related to the disease genotype. These subclinical phenotypes would generally be much more prevalent than the disease itself. It is important to consider what data would look like under these circumstances.

Several different selection schemes were considered, i.e., requiring ⩾1, ⩾3, ⩾5, or ⩾9 affected family members per family. For each mode of inheritance, gene frequency, penetrance value, and selection scheme, 2,000 families were generated. We tabulated the proportion of families with intrafamilial heterogeneity. We also calculated the theoretically expected proportions of families, under the D + D model with full penetrance, using a selection scheme requiring ⩾1 affected member, and we used these proportions to check our simulations (see Appendix).

Analyzing Data Sets with Intra- and Interfamilial Heterogeneity

To evaluate the better way to analyze the data, we generated two different kinds of samples and then analyzed each sample type in two different ways. Sample type A was a population mixture of families with disease form 1, disease form 2, or both disease forms, and sample type B consisted solely of families with *both* disease forms segregating within *each* family (anywhere within the pedigree). Obviously, this type of sample is completely artificial and would not be expected to arise in actual data collection. It is designed to test how well the 2L analysis deals with intrafamilial heterogeneity.

The generating penetrance values were 50%, 70%, and 90%. The gene frequencies for both dominant disease forms were $q_1 = q_2 = .01$, and that for the recessive form was $q_2 = .15$, resulting in approximately equal population prevalences for the two different disease forms. We anticipated that this would represent the "worst" case, in the sense of leading to the highest proportion of families with both disease forms segregating. As before, the linked disease form always segregated dominantly, whereas the unlinked disease form was either dominant or recessive. The true θ between the linked disease and the marker was always 0. Only fully informative matings were generated at the marker locus.

Lod scores for the two sample types were calculated under an SL analysis model and a 2L analysis model. For the SL analysis we used LIPED (Ott 1974). The analyzing gene frequency was set to $q = q_1 + q_2$, where q_1 was .01 for the dominant linked form of the disease and q_2 was either .01 for the dominant unlinked or .15 for the recessive unlinked disease form. The 2L analysis used the 2L version of LINKAGE, TMLINK (Lathrop and Ott 1990), using the same—i.e., correct—gene frequencies, q_1 and q_2, which were used to generate the data. The other analysis parameters (penetrance and mode of inheritance) were identical to the generating parameters in both analyses. We report mean maximum lod scores, maximized over $\theta = .01$, .05, .1, .2, .3, .4, and .5. In sample type A the output of the SL model analysis was used to perform a heterogeneity test (Hodge et al. 1983) and to calculate a lod2 score, which is a lod score maximized with respect to θ and α (percentage of linked families).

Note that when we analyzed the data under the 2L models, we used the correct parameters for the analysis, i.e., the correct gene frequencies, modes of inheritance, and penetrance. Since the SL analysis is, by definition, not the model under which the data were generated, the choices of gene frequency, etc. are not correct but represent the best guess that a researcher might make given that the analysis is SL. Thus, when the results are examined, it must be remembered that every advantage is being given to the 2L analysis.

Three thousand families were generated for each set of parameter values. They were then analyzed as 300 data sets consisting of 10 families each, and means and SDs were computed across the 300 data sets.

Results

1. *Proportions of Families with Both Disease Forms*

Table 1 shows the proportions of families segregating both disease forms, as a function of selection scheme, mode of inheritance, penetrance, and gene frequency.

Selection scheme.—The more stringent the selection scheme, the higher the proportion of families with both disease forms segregating within the family. The change in this proportion can be quite dramatic. In extreme cases there was a 6.5-fold increase when ⩾9 affected family members were required, compared with the most relaxed selection scheme (⩾1 affected). These patterns were observed in both the D + D and D + R cases.

In the D + R case, selecting for many affected family members not only increased the proportion of families with both disease forms but also *decreased* the number of families with the pure recessive disease form (not shown in table 1). For example, at a gene frequency of $q_1 = .01$ for the dominant disease gene and $q_2 =$

Table 1

Proportions (in %) of Families with Both Disease Forms Segregating within Each Family

Model, Gene Frequencies, and Penetrance	Selection Scheme[a]			
	≥1 Affected	≥3 Affected	≥5 Affected	≥9 Affected
D + D				
.01 and .01:				
90%	6.6	10.5	13.6	23.5
70%	6.5	12.4	16.8	27.7
50%	7.5	11.6	20.4	35.4
30%	5.5	12.5	24.2	37.1
.05 and .01:				
90%	10.8	14.6	16.8	24.8
70%	10.6	15.3	18.4	29.1
50%	8.6	15.9	23.2	35.4
30%	7.8	16.4	25.7	43.1
.10 and .01:				
90%	10.8	13.1	15.6	20.1
70%	10.4	13.6	16.7	23.8
50%	9.3	14.3	19.0	28.2
30%	6.4	13.9	22.2	30.1
.15 and .01:				
90%	8.1	12.6	11.4	15.4
70%	7.4	11.5	14.6	18.3
50%	9.4	13.6	16.7	22.5
30%	7.0	12.9	19.6	31.5
D + R:				
.01 and .15:				
90%	10.2	17.7	23.3	33.8
70%	9.2	17.6	25.9	36.6
50%	7.7	18.5	26.7	42.4
30%	6.3	19.7	29.3	53.1
.01 and .10:				
90%	6.8	9.8	12.8	17.4
70%	5.9	10.1	13.4	18.2
50%	5.4	9.6	12.6	18.9
30%	4.4	10.6	15.9	24.2
.01 and .05:				
90%	2.5	4.5	3.8	5.3
70%	2.3	3.3	3.5	5.1
50%	2.4	3.2	4.2	5.0
30%	1.3	3.7	4.3	3.8

[a] Number of affected family members required for ascertainment.

.15 for the recessive disease gene (equal population prevalences), the proportion of families with the recessive disease was ~50% under the most relaxed selection scheme (requiring ≥1 affected member). This proportion dropped to ~25%–30% in data sets with ≥3 affected per family and to 10%–20% when more stringent selection schemes (requiring ≥5 affected) were applied. This would presumably make it more difficult to detect linkage in families with the recessive form.

Penetrance.—Reduced penetrance tended to accentuate the increase in intrafamilial heterogeneity when only heavily loaded pedigrees were selected. Under the more relaxed selection schemes (≥1 or ≥3 affected), the proportion of families with both forms of the disease did not change significantly as the penetrance went down. However, when ≥5 affected members were required, the proportion of families with both forms of the disease increased as the penetrance fell, particularly under the D + D models. For example, when both diseases were dominant and had equal gene frequency, this proportion increased from 13.6% at 90% penetrance to 24.2% at 30% penetrance. The one exception occurred for the D + R model with frequencies .01 and .05, i.e., where the D form of the disease is eight times more prevalent than the R form. Therefore, in most cases the combination of low penetrance and more stringent selection schemes will increase intrafamilial heterogeneity.

Gene frequency. — In the D + D model, an increase in the gene frequency of disease form 1 (and therefore the ratio of disease form 1 vs. disease form 2) increased the proportion of families with disease form 1 and lowered the proportion of families with disease form 2. But the proportion of families with both disease forms was relatively constant. Only when the gene frequency of disease form 1 was much higher than the gene frequency of disease form 2 did the proportion of families with intrafamilial heterogeneity become smaller. Decreasing the gene frequency of the second recessive disease in the D + R model, and thereby also increasing the ratio of disease form 1 versus disease form 2, led to a smaller proportion of families with both disease forms within each family. These patterns also agree with our theoretical calculations (Appendix).

We had anticipated that, under a D + D model, the highest proportion of families showing both forms of the disease would be seen when the gene frequencies — and thus the prevalences — were equal. To our surprise, the worst overall case occurred when the gene frequencies were .05 and .01. For the D + R model, the worst case, i.e., the highest proportion of families with both disease forms, was seen when the population prevalences were equal.

In summary, among the models considered, the proportion of families with intrafamilial heterogeneity increases when only heavily loaded families are selected and when the disease penetrance is low. When the most relaxed selection scheme is used (requiring ⩾1 affected), then the proportion of families with both disease forms never exceeds 10%–11%, no matter what the mode of inheritance, gene frequency, or penetrance.

2. *Sample Type A: Lod Scores of a Population Mixture of Families with Disease Form 1, Disease Form 2, or Both Disease Forms*

In this section we compare, in heterogeneous data sets, the results from a 2L analysis with those from an SL analysis. Table 2 shows mean maximum lod scores and mean θ values from the SL and 2L analyses, for the D + D model and for the D + R model.

The mean maximum lod scores derived from the SL versus the 2L analyses are surprisingly close. The SL analysis before a heterogeneity test performs quite well, even though it is not the correct model for analyzing the data. The mean maximum lod score obtained from an SL analysis was closest to the mean maximum lod score of a 2L analysis when the disease under consideration had a low penetrance and when the most relaxed selection scheme(s) was (were) used. In general, the maximum lod score is higher under the D + R models than under the D + D models, an observation we also made in data sets exhibiting *only* interfamilial heterogeneity (Durner and Greenberg 1992). Estimation of θ is more influenced by the kind of analysis than is the magnitude of the lod score. The 2L analysis of the data yields a θ estimate closer to the generating value than is the θ estimate yielded by the SL analysis, as expected from theory. The estimate of θ under an SL analysis was better when the unlinked disease form was recessive ($\hat{\theta}$ = .07–.1) than when the unlinked form of the disease in the mixture was also dominant ($\hat{\theta}$ = .2). This is not surprising, since under D + R models the dominant form of the disease tends to be preferentially selected.

When a heterogeneity test is performed on the data analyzed under an SL model, the maximized lod2 score (lod score maximized with respect to θ and α) was almost identical to the lod score found under a 2L analysis (table 2). Also, the corresponding estimate of θ was as good as the estimate of θ obtained from a 2L analysis. The estimates of α and 1 – α (i.e., percentage of unlinked families) were quite accurate (for the type A sample). The mean maximum lod2 score and also the number of data sets with a significant lod2 score (>3) were higher when the linked disease showed high penetrance than when the penetrance was low. When more affected family members per family were required for the study, the mean maximum lod2 score increased and significant evidence for heterogeneity could be demonstrated in nearly every data set.

To show the effect of "bilineal" pedigrees (i.e., pedigrees with affected members on both sides of the family) on the analysis, we repeated our simulation and deliberately did not ascertain either families with affected married-ins or families where both founders were affected (i.e., each data set still contained 10 families, but none of these 10 were bilineal). It is surprising that this did not change the lod scores very

Table 2

Mean Maximum Lod Score ± SD and Mean $\hat{\theta}$ ± SD from SL and 2L Analysis—Population Mixture of Disease Form 1, Disease Form 2, and Both Disease Forms

Model, Penetrance, Selection Scheme[a]	Bilineal Pedigrees Included						Bilineal Pedigrees Excluded			
	2L Analysis		SL Analysis		Heterogeneity Test		2L Analysis		SL Analysis	
	Lod Score	$\hat{\theta}$	Lod Score	$\hat{\theta}$	Lod2 Score	$\hat{\theta}$	Lod Score	$\hat{\theta}$	Lod Score	$\hat{\theta}$
D + D:										
50%:										
⩾1 Affected ...	1.6 ± 1.3	.07 ± .1	1.4 ± 1.4	.2 ± .1	1.8 ± 1.5	.07 ± .1	1.6 ± 1.3	.08 ± .1	1.5 ± 1.4	.2 ± .1
⩾3 Affected ...	2.6 ± 1.6	.05 ± .1	2.0 ± 1.7	.2 ± .1	2.8 ± 1.7	.05 ± .07	2.9 ± 1.7	.05 ± .1	2.3 ± 2.7	.2 ± .1
⩾5 Affected ...	4.5 ± 2.6	.03 ± .07	3.4 ± 2.6	.2 ± .1	4.7 ± 2.7	.04 ± .05	4.9 ± 2.6	.02 ± .06	3.6 ± 2.7	.2 ± .1
70%:										
⩾1 Affected ...	2.7 ± 1.9	.06 ± .1	2.3 ± 1.9	.2 ± .1	2.8 ± 2.0	.05 ± .08	2.6 ± 2.0	.05 ± .1	2.1 ± 2.1	.2 ± .1
⩾3 Affected ...	4.3 ± 2.5	.03 ± .07	3.4 ± 2.7	.2 ± .1	4.5 ± 2.6	.04 ± .06	4.2 ± 2.5	.03 ± .07	3.3 ± 2.6	.2 ± .1
⩾5 Affected ...	6.2 ± 3.4	.02 ± .07	4.7 ± 3.5	.2 ± .1	6.3 ± 3.5	.03 ± .05	6.2 ± 3.3	.02 ± .06	4.4 ± 3.2	.2 ± .1
90%:										
⩾1 Affected ...	4.5 ± 2.9	.03 ± .08	3.4 ± 2.9	.2 ± .1	4.7 ± 3.0	.04 ± .06	4.2 ± 2.7	.04 ± .09	3.1 ± 2.6	.2 ± .1
⩾3 Affected ...	6.9 ± 3.4	.03 ± .08	4.2 ± 3.3	.2 ± .1	6.1 ± 3.5	.03 ± .05	6.6 ± 3.5	.02 ± .05	4.7 ± 3.4	.2 ± .1
⩾5 Affected ...	8.0 ± 4.0	.02 ± .05	5.0 ± 3.8	.2 ± .09	8.0 ± 4.0	.02 ± .03	9.8 ± 4.1	.01 ± .03	6.4 ± 4.1	.2 ± .09
D + R:										
50%:										
⩾1 Affected ...	2.1 ± 1.5	.05 ± .1	2.0 ± 1.6	.1 ± .1	2.2 ± 1.6	.06 ± .09	2.1 ± 1.4	.05 ± .1	2.0 ± 1.5	.1 ± .1
⩾3 Affected ...	4.8 ± 2.3	.02 ± .05	4.3 ± 2.4	.1 ± .08	4.7 ± 2.4	.04 ± .05	5.3 ± 2.9	.02 ± .05	5.0 ± 2.6	.1 ± .07
⩾5 Affected ...	8.3 ± 2.9	.01 ± .02	7.3 ± 3.3	.09 ± .07	8.1 ± 3.1	.03 ± .03	9.8 ± 3.0	.01 ± .02	9.1 ± 3.4	.07 ± .05
70%:										
⩾1 Affected ...	3.4 ± 2.1	.03 ± .08	2.9 ± 2.1	.1 ± .1	3.2 ± 2.1	.06 ± .08	3.5 ± 2.0	.04 ± .09	3.1 ± 2.1	.1 ± .1
⩾3 Affected ...	6.5 ± 2.8	.02 ± .05	5.5 ± 3.0	.1 ± .08	6.1 ± 2.9	.03 ± .05	6.7 ± 2.8	.02 ± .03	5.8 ± 2.9	.1 ± .07
⩾5 Affected ...	10.8 ± 3.4	.01 ± .02	9.6 ± 3.8	.07 ± .06	10.4 ± 3.5	.02 ± .03	12.3 ± 3.7	.01 ± .02	11.2 ± 4.3	.06 ± .05
90%:										
⩾1 Affected ...	5.6 ± 3.1	.02 ± .06	4.1 ± 3.0	.1 ± .09	4.9 ± 3.1	.04 ± .06	5.6 ± 2.8	.03 ± .04	4.1 ± 2.8	.2 ± .1
⩾3 Affected ...	9.0 ± 3.7	.01 ± .03	6.8 ± 3.8	.1 ± .08	8.0 ± 3.7	.03 ± .04	10.0 ± 3.6	.01 ± .3	7.8 ± 3.7	.1 ± .07
⩾5 Affected ...	14.2 ± 4.5	.01 ± .02	11.2 ± 5.0	.1 ± .07	13.0 ± 4.7	.03 ± .03	14.9 ± 4.3	.01 ± .01	13.0 ± 5.3	.1 ± .06

[a] Number of affected family members required for ascertainment.

much. The mean maximum lod score was, in some cases, even slightly lower than or equal to the mean maximum lod score in the mixture where bilineal pedigrees were allowed. Only under a more stringent selection scheme, where heterogeneity is more pronounced to begin with, does the exclusion of bilineal pedigrees increase the lod score (~0.7–1.8 lod-score units in a sample size of 10 families). This effect was more pronounced in the D + R model than in the D + D model. Excluding families with affected members on both sides had no effect on the estimate of θ.

In summary, among the models considered, in heterogeneous data sets with two genetically different disease forms, an SL analysis combined with a heterogeneity test performs as well as a 2L analysis. There seems to be no need to exclude bilineal families from the analysis.

3. *Sample Type B: Lod Scores of Families with Both Disease Forms Segregating within Each Family*

We tested the 2L analysis in an extreme situation in order to determine more precisely how the 2L analysis compares with the SL analysis, which is the "wrong" model for this situation.

The difference in lod score achieved under a 2L analysis versus an SL analysis was small. On average, the lod score was 0.2–2.7 lod-score units higher under a 2L analysis than under an SL (table 3). The difference between the mean maximum lod scores of the two analyses was lower when both disease forms were dominant (0.2–1.6 lod-score units) than when one disease form was dominant and the other recessive (0.2–2.6 lod-score units). It was smaller for disease forms with low penetrance (0.2–0.8 lod-score units at 50% penetrance) than for those with high penetrance (1.1–2.7 lod-score units at 90% penetrance). For both the SL and 2L analyses, the lod score was higher when the unlinked form of the disease was recessive than when the unlinked disease form was dominant.

However, the estimate of θ was much closer to the correct, i.e., generating, value under the 2L analysis than under the SL analysis. As expected from theoretical considerations, θ was estimated more accurately when the diseases had a high penetrance than when they had a low penetrance, especially in the analysis done under a 2L model.

In summary, in the extreme case of intrafamilial heterogeneity *only,* a 2L analysis achieves slightly higher lod scores and somewhat better estimates of θ than does the SL analysis. The 2L analysis was run with the correct parameters. In a real-life situation, knowledge of the correct parameter values is unlikely. We did not test how sensitive the 2L analysis is to parameter misspecification.

Discussion

1. *Proportions of Families with Both Disease Forms*

While extended pedigrees with many affected people definitely provide a great deal of information for linkage, selecting specifically for such high-density families has certain pitfalls when heterogeneity is present. When the number of affected members required for selection of the family is increased (i.e., when selection schemes are more stringent), the number of families with intrafamilial heterogeneity increases as well. Especially in diseases with low penetrance, as may be the case in psychiatric familial diseases, collecting only "high-density" pedigrees can bias the sample toward those families with more than one disease gene segregating. Families with disease genes coming into them from more than one side are likely to have more affected members than are families where only one founder has the disease gene. This is especially true for diseases with low penetrance. Therefore, when only families with many affected family members are selected, those families with more than one form of the disease may be overrepresented.

Because such high-density pedigrees are rare, a data set may consist of just a few very dense pedigrees. The data in table 1 indicate that, for the models we have looked at, up to one-fourth to one-half of the families may have intrafamilial heterogeneity. Whether the disadvantages of having this much intra-familial heterogeneity in the data set outweigh the advantages of high-density pedigrees is a complex question beyond the scope of this study. However, in any case, the investigator should be aware of these issues when devising a data collection strategy (also see Greenberg, in press).

In the case of the D + R model, selecting families with many affected members also should presumably make it more difficult to detect linkage to the recessive disease form. This is because the families with the recessive form of a disease have fewer affected children, on average, and hence will be underrepresented in a sample of heavily loaded families. Similarly, the proportion of families with both disease forms increases as more affected family members are required. A low disease penetrance has a most unfortunate effect and compounds both of these trends. These same gen-

Table 3

Mean Maximum Lod Score ± SD and Mean $\hat{\theta}$ ± SD from SL and 2L Analysis—Only Families with Both Disease Forms within Each Family

Model, Penetrance, and Selection Scheme[a]	2L Analysis		SL Analysis	
	Lod Score	$\hat{\theta}$	Lod Score	$\hat{\theta}$
D + D:				
50%:				
≥1 Affected	1.1 ± 1.0	.1 ± .2	0.9 ± 1.1	.3 ± .1
≥3 Affected	1.3 ± 1.2	.1 ± .2	1.1 ± 1.2	.3 ± .1
≥5 Affected	1.8 ± 1.2	.1 ± .1	1.6 ± 1.5	.3 ± .1
70%:				
≥1 Affected	1.9 ± 1.7	.1 ± .1	1.7 ± 1.6	.2 ± .1
≥3 Affected	2.1 ± 1.7	.1 ± .1	1.9 ± 1.7	.2 ± .1
≥5 Affected	3.3 ± 1.8	.06 ± .09	2.6 ± 2.0	.2 ± .1
90%:				
≥1 Affected	4.0 ± 2.8	.07 ± .1	2.8 ± 2.4	.2 ± .1
≥3 Affected	4.1 ± 2.7	.05 ± .08	3.0 ± 2.4	.2 ± .1
≥5 Affected	5.3 ± 3.1	.05 ± .08	3.7 ± 2.7	.2 ± .09
D + R:				
50%:				
≥1 Affected	1.5 ± 1.4	.1 ± .1	1.3 ± 1.1	.2 ± .1
≥3 Affected	2.3 ± 1.5	.1 ± .1	1.8 ± 1.4	.2 ± .1
≥5 Affected	4.4 ± 2.2	.08 ± .06	3.6 ± 2.3	.2 ± .09
70%:				
≥1 Affected	2.9 ± 2.1	.1 ± .1	2.4 ± 1.9	.2 ± .1
≥3 Affected	3.7 ± 2.3	.07 ± .07	3.2 ± 2.2	.2 ± .1
≥5 Affected	6.2 ± 3.0	.06 ± .06	5.2 ± 2.7	.1 ± .07
90%:				
≥1 Affected	5.8 ± 3.0	.06 ± .07	4.0 ± 2.5	.2 ± .08
≥3 Affected	7.3 ± 3.3	.04 ± .04	4.6 ± 2.6	.2 ± .08
≥5 Affected	9.6 ± 3.6	.04 ± .04	7.0 ± 3.4	.2 ± .06

[a] Number of affected family members required for ascertainment.

eralizations probably also apply even when the unlinked form of the disease is nongenetic.

To avoid genetic heterogeneity in the common diseases, especially psychiatric diseases, many have proposed selecting only families with many affected members. Our simulations show that, on the contrary, these sampling strategies can *increase* the proportion of families with intrafamilial heterogeneity.

2. *Lod Scores of a Population Mixture of Families with Disease Form 1, Disease Form 2, or Both Disease Forms*

An SL analysis with a subsequent heterogeneity test performs almost as well as a 2L analysis. The lod score and the estimate of θ are almost identical. On purely theoretical grounds, one might favor the 2L analysis. However, an additional point to consider is that we have little experience with how robust a 2L analysis is when the input parameters are misspecified, whereas we know that SL linkage-analysis models are fairly robust. In this study we gave the 2L analysis method the "best case," in that we analyzed the data with all the correct input parameters (correct mode of inheritance, gene frequencies, and penetrances). However, for most common diseases, these parameters are unknown.

In comparing the 2L and SL methods, we gave the 2L an advantage over the SL analysis by using the correct parameter values. Despite that, the 2L analysis did little better than the SL method. Thus, the SL analysis has potential advantages of known robustness and simplicity. This is also true in the case of 2L epistatic models (Vieland et al. 1992). Therefore, given the choice between two methods that perform equally well, we would opt for the SL analysis plus heterogeneity test, which has fewer parameters. This may change when more is known about the robustness and behavior of a 2L analysis. These recommendations are also in line with those of Goldin (1992, and in press).

(Note, however, that we considered only the case where *one* disease locus is linked to a marker being studied. The conclusions concerning the relative merits of the SL and 2L analyses may differ if one has a linked marker for *each* disease locus and if one analyzes both linkages simultaneously [N. Schork, M. Boehnke, J. Terwilliger, and J. Ott, personal communication].)

It has been policy for some time to exclude bilineal pedigrees from linkage studies. Our study shows that this policy has (*a*) little effect on the mean maximum lod score and (*b*) essentially no effect on the θ estimate. Therefore, we recommend against it. Our observation was that the transmission of the other gene, which was brought into the family mostly through the affected married-in, was limited to a small part of the pedigree and often was restricted to the affected married-in only. We conclude, therefore, that these families still provide positive information for linkage analysis, as Hodge (in press) has also shown for homogeneous diseases. There is no evidence that inclusion of these bilineal families will lead to a false exclusion of linkage. However, we did not examine the case where there is assortative mating, i.e., families in which affected people preferentially marry other affected people.

3. *Lod Scores of Families with Both Disease Forms Segregating within Each Family*

On average, families with intrafamilial heterogeneity provide positive information for linkage. The amount of information is surprisingly high. When families are not selected for multiple affected members, the difference in the lod score compared with the population mixture is only ~ ½ lod-score unit for a given sample size of 10 families. This difference becomes greater as more affected members per family are required, i.e., the more stringent the selection schemes. We also noticed that the estimate of θ is higher in those data sets consisting only of families with both disease forms than it is in data sets of a population mixture of families each of whom has only one of the two disease forms.

The main purpose of this artificial sample of only families with intrafamilial heterogeneity was to test whether a 2L analysis would be able to detect linkage in spite of the heterogeneity. The fact that, for this sample, the 2L analysis did not do significantly better than the SL, despite the use of the correct parameter values, indicates that the 2L analysis will not completely solve the problem of identifying linkage in the presence of intrafamilial heterogeneity.

Conclusions

1. The problem of intrafamilial heterogeneity in heterogeneous diseases will worsen under more stringent selection schemes (i.e., those requiring many affected members). When a disease can be caused by either a dominant or a recessive gene (i.e., in the D + R model), most selection schemes will preferentially select the dominant form. Moreover, we believe that the more stringent sampling strategies will further complicate the detection of linkage to the recessive form by reducing even further the proportion of families with the recessive form and increasing the proportion of families with intrafamilial heterogeneity.

2. Linkage in diseases caused by two different genes can be studied by either a 2L analysis or an SL analysis combined with an admixture test. The results in terms of lod score and estimate of θ are about the same. An SL analysis might be preferable because of its known robustness and simplicity. There is little harm in including bilineal pedigrees in the analysis.

3. To the extent that intrafamilial heterogeneity represents a problem for linkage studies, the 2L analysis does not provide a panacea. This is not because of a failure in the analysis method but because of the inherent situation of genetic heterogeneity.

Acknowledgments

This work supported in part by grants DK31775, DK31813, NS27941, NS21908, MH48858, MH30906, MH43878, MH45212, MH28274, and MH46981, and by a Fellowship in the Neurosciences from the Klingenstein Foundation (to D.A.G.).

Appendix

Simple Formula for Proportions of Family Types

For some special cases, a simple and elegant formula can be derived for proportions of different types of families (as in table 1). This formula then provides a useful way to help check the correctness of the computer simulations. We define a type 1 (or type 2) family as one in which ≥1 disease form 1 (or disease form 2) allele is present but in which no allele of the other disease form are present, and we define a type 3 family as one with ⩾1 allele of each disease form. If both disease forms are autosomal dominant (i.e., a D + D model) and fully penetrant, then these three family types correspond to families with ⩾1 affected member and with either disease form 1 only (family type 1), disease form 2 only (family type 2), or both forms of the disease (family type 3).

For any family, the only individuals relevant for calculating these probabilities are the married-ins, because the disease gene(s) can enter the family only via these individuals. Hence, we define the effective family size, *n*, as the number of married-in individuals. (For example, for a nuclear family consisting of two parents and any number of children, $n = 2$.)

We assume two independent autosomal diseases, with gene frequencies q_1 and q_2 for disease alleles 1 and 2, respectively. Also define $r_1 = 1 - q_1$ and $r_2 = 1 - q_2$. The *n* married-ins have $2n$ loci at which a disease form 1 allele could appear. Hence, the probability that *no one* of them has a disease allele at locus 1 is r_1^{2n}. (For form 2, this quantity is r_2^{2n}.)

Define Q_0 as the probability that a family of effective size *n* has no disease alleles of either disease form, and let Q_i, for $i = 1$, 2, and 3, represent the population probabilities of the three family types defined above. Then, since loci 1 and 2 are independent, $Q_0 = r_1^{2n} r_2^{2n}$, $q_1 = (1 - r_1^{2n}) r_2^{2n}$, $Q_2 = (1 - r_2^{2n}) r_1^{2n}$, and $Q_3 = (1 - r_1^{2n})(1 - r_2^{2n})$. The relative probabilities, R_i, that a family is of type *i*, given that it has at least one disease allele at some locus, are $R_i = Q_i/(1 - Q_0)$ for $i = 1$, 2, and 3. These R_i's give the proportions of families of each type, in a sample obtained by truncate sampling, i.e., with ascertainment probability $\pi = 1$ (Morton 1959). This corresponds to our "most relaxed" sampling scheme, the one requiring ⩾1 affected. As mentioned above, the model is D + D, with 100% penetrance. These formulas check with the corresponding numbers in table 1.

References

Cavalli-Sforza LL, Bodmer WF (1971) The genetics of human populations. WH Freeman, San Francisco

Cavalli-Sforza LL, King M-C (1986) Detecting linkage for genetically heterogeneous diseases and detecting heterogeneity with linkage data. Am J Hum Genet 38:599–616

Durner M, Greenberg DA (1992) Effect of heterogeneity and assumed mode of inheritance on lod scores. Am J Med Genet 42:271–275

Egeland JA, Gerhard DS, Pauls DL, Sussex JN, Kidd KK, Allen CR, Hostetter AM, et al (1987) Bipolar affective disorders linked to DNA markers on chromosome 11. Nature 325:783–787

Goldin LR (1992) Detection of linkage under heterogeneity: comparison of the two-locus vs. admixture models. Genet Epidemiol 9:61–66

——— Genetic heterogeneity and other complex models: a problem for linkage determination. In: Gershon ES, Cloninger CR, Barrett JE (eds) New genetic approaches to mental disorders. American Psychiatric (in press)

Goldin LR, Martinez MM, Gershon ES (1991) Sampling strategies for linkage studies. Eur Arch Psychiatry Clin Neurosci 240:182–187

Greenberg DA (1984) Simulation studies of segregation analysis: application to two-locus models. Am J Hum Genet 36:167–176

——— (1989) Inferring mode of inheritance by comparison of lod scores. Am J Med Genet 34:480–486

——— There is more than one way to collect data for linkage analysis: what a study of epilepsy can tell us about linkage strategy for psychiatric disease. Arch Gen Psychiatry (in press)

Hodge SE. Do bilineal pedigrees represent a problem for linkage analysis? basic principles and simulation results for single-gene diseases with no heterogeneity. Genet Epidemiol (in press)

Hodge SE, Anderson CE, Neiswanger K, Sparkes RS, Rimoin DL (1983) The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): linkage studies, two-locus models, and genetic heterogeneity. Am J Hum Genet 35:1139–1155

Kelsoe JR, Ginns EI, Egeland JA, Gerhard DS, Goldstein AM, Bale SJ, Pauls DL, et al (1989) Re-evaluation of the linkage relationship between chromosome 11p loci and the gene for bipolar affective disorder in the Old Order Amish. Nature 342:238–243

Kennedy JL, Giuffra LA, Moises HW, Cavalli-Sforza LL, Pakstis AJ, Kidd JR, Castiglione CM, et al (1988) Evidence against linkage of schizophrenia to markers on chromosome 5 in a northern Swedish pedigree. Nature 336:167–170

Lander E (1988) Splitting schizophrenia. Nature 336:105–106

Lathrop GM, Ott J (1990) Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. Am J Hum Genet 47 [Suppl]: A188

Martinez MM, Goldin LR (1989) The detection of linkage and heterogeneity in nuclear families for complex disorders: one versus two marker loci. Am J Hum Genet 44: 552–559

——— (1990) Power of the linkage test for a heterogeneous disorder due to two independent inherited causes: a simulation study. Genet Epidemiol 7:219–230

Merikangas KR, Spence MA, Kupfer DJ (1989) Linkage studies of bipolar disorder: methodologic and analytic issues. Arch Gen Psychiatry 46:1137–1141

Morton NE (1956) The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. Am J Hum Genet 8:80–96

——— (1959) Genetic tests under incomplete ascertainment. Am J Hum Genet 11:1–16

Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26:588–597

——— (1977) Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. Ann Hum Genet 40:443–454

Regier DA, Judd LL (1989) Request for applications: Diagnostic Centers for Psychiatric Linkage Studies, National Institute of Mental Health. National Institute of Mental Health, Rockville, MD

Risch N, Baron M (1982) X-linkage and genetic heterogeneity in bipolar-related major affective illness: reanalysis of linkage data. Ann Hum Genet 46:153–166

Sherrington R, Brynjolfsson J, Petursson H, Potter M, Dudleston K, Barraclough B, Wasmuth J, et al (1988) Localization of a susceptibility locus for schizophrenia on chromosome 5. Nature 336:164–167

Smith CAB (1963) Testing for heterogeneity of recombination fraction values in human genetics. Ann Hum Genet 27:175–182

Vieland VJ, Hodge SE, Greenberg DA (1992) Adequacy of single-locus approximations for linkage analyses of oligogenic traits. Genet Epidemiol 9:45–59