Effect of Genetic Heterogeneity and Assortative Mating on Linkage Analysis: A Simulation Study

Catherine T. Falk

The Lindsley F. Kimball Research Institute of The New York Blood Center, New York

Summary

Linkage studies of complex genetic traits raise questions about the effects of genetic heterogeneity and assortative mating on linkage analysis. To further understand these problems, I have simulated and analyzed family data for a complex genetic disease in which disease phenotype is determined by two unlinked disease loci. Two models were studied, a two-locus threshold model and a two-locus heterogeneity model. Information was generated for a marker locus linked to one of the disease-defining loci. Random-mating and assortative-mating samples were generated. Linkage analysis was then carried out by use of standard methods, under the assumptions of a single-locus disease trait and a random-mating population. Results were compared with those from analysis of a single-locus homogeneous trait in samples with the same levels of assortative mating as those considered for the two-locus traits. The results show that (1) introduction of assortative mating does not, in itself, markedly affect the estimate of the recombination fraction; (2) the power of the analysis, reflected in the LOD scores, is somewhat lower with assortative rather than random mating. Loss of power is greater with increasing levels of assortative mating; and (3) for a heterogeneous genetic disease, regardless of mating type, heterogeneity analysis permits more accurate estimate of the recombination fraction but may be of limited use in distinguishing which families belong to each homogeneous subset. These simulations also confirmed earlier observations that linkage to a disease "locus" can be detected even if the disease is incorrectly defined as a single-locus (homogeneous) trait, although the estimated recombination fraction will be significantly greater than the true recombination fraction between the linked disease-defining locus and the marker locus.

Received May 5, 1995; accepted for publication September 3, 1997; electronically published October 29, 1997.

Address for correspondence and reprints: Dr. Catherine T. Falk, The New York Blood Center, 310 East 67th Street, New York, NY 10021. E-mail: cathy@server.nybc.org

0002-9297/97/6105-0024$02.00

Introduction

Linkage studies of complex (i.e., non-Mendelian) genetic traits, such as psychiatric disorders, have raised questions about the effect of such problems as genetic heterogeneity and assortative mating on the outcome of linkage analysis. It is of great interest to try to identify loci that contribute to the genetic susceptibility of such traits as schizophrenia, affective disorder, Alzheimer disease, and countless others, but such efforts encounter difficulties that are not necessarily found in the analysis of Mendelian traits. For example, because of the known problems of genetic heterogeneity, it is often assumed that linkage studies of a single large pedigree will be most effective, since it is more likely that a single, homogeneous form of the disease will be segregating in the entire pedigree (e.g., see Egeland et al. 1987). Unfortunately, as that study and others have shown, identification of single large pedigrees does not necessarily make the task easier, particularly when one is dealing with relatively common traits. In fact, as Durner et al. (1992) have shown, sampling "high density" pedigrees may *increase* the probability of finding intrafamilial heterogeneity. In the Egeland et al. study, initial results identified a potential linkage between affective disorder and HRAS1 on chromosome 11. However, the initial LOD score (Z) of 4.32 did not hold up when new data were added to the analysis, and ultimately the evidence for linkage declined (Kelsoe et al. 1989). Various explanations for such fluctuations in results, even within a single large pedigree, have been proposed; these explanations include genetic heterogeneity, multilocus traits analyzed as single-locus traits, and nonrandom (i.e., assortative) mating for the trait.

We do not yet fully understand what the effect on linkage analysis will be if we assume a single-locus disease trait when disease state is actually determined by more than one locus and/or more than one genotype, although several simulation studies have been undertaken that explore the consequences of such assumptions and shed light on the problems (e.g., see Greenberg 1990; Durner et al. 1992; Goldin 1992; Vieland et al.

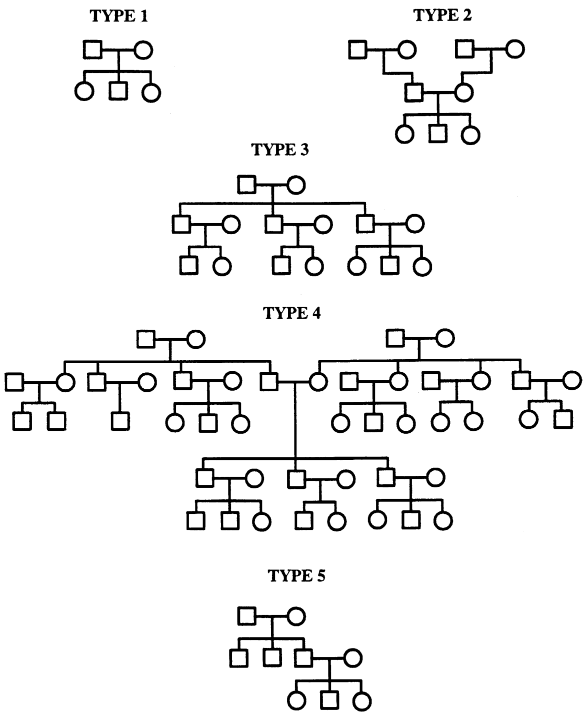

Figure 1 Representative examples of five family structures used in simulation studies.

1992*a,* 1992*b,* 1993). Similarly, if there is nonrandom mating with respect to a disease trait (e.g., if individuals with the disease phenotype tend to mate assortatively, as is sometimes observed with respect to psychiatric traits [Merikangas 1982]), it is not known what effect, if any, this has on the outcome of a linkage analysis. Assortative mating has the effect of increasing the proportion of bilineal families in a sample—that is, families in which a trait is introduced into a pedigree on both the maternal and paternal sides. Many linkage analyses that have been performed on complex diseases have made use of one of the standard computer programs for linkage analysis. These programs were originally written under certain assumptions, some of which are violated when one is dealing with complex (non-Mendelian) traits. For example, the programs assume that the family sample is from a random-mating population and that the disease trait is determined by alleles at a single locus with a known mode of inheritance. More recently, programs have been developed that allow for a two-locus disease trait (Schork et al. 1993), and other methods of analysis, such as nonparametric sib-pair analysis, are not constrained by these assumptions. The problem of analyzing the data under the wrong assumption about mode of inheritance has been addressed by others (e.g., Greenberg and Hodge 1989), but questions remain about such characteristics of the data as assortative mating and genetic heterogeneity. It is possible that some of the inconsistencies and contradictions in results of linkage studies could be due to the failure to account for the proper underlying parameters.

To address these questions and gain further understanding of the problems, I have simulated family data for a complex genetic disease in which the disease phenotype is specified by a two-locus disease-determining model, the two disease-determining loci being unlinked to one another. Additional information on marker loci was generated to enable us to ask questions about the results of linkage analysis when only one of the two disease-determining loci is linked to the marker locus/loci. Both random-mating and assortative-mating samples were generated. Linkage analysis was then performed by use of standard methods that assume (*a*) a single-locus disease trait and (*b*) a random-mating population. Results were also compared with those from analysis of a single-locus homogeneous trait in samples with the same levels of assortative mating as those considered for the two-locus traits.

Methods

Computer simulation methods were developed that are capable of generating family data for disease models in which two or more unlinked loci contribute to disease status. In the present paper, two two-locus models are studied. Five family structures are used, ranging from simple two-generation families to relatively large four-generation pedigrees. One sample realization of each pedigree structure is shown in figure 1. The number of children in a sibship ranges from three to six. The size is chosen at random, with small sibships having higher probabilities. Thus the number of sibs may vary from realization to realization, but the overall structure of the pedigrees remains the same. For all models, genetic information is generated for the parents, on the basis of input allele frequencies at two disease-determining loci, B and C, as well as at one marker locus, A. Loci B and C have two possible alleles, giving rise to the two-locus genotypes and phenotypes shown in tables 1 and 2. In these examples, marker locus A also has two alleles. Other simulation runs, with more alleles at marker locus A, gave qualitatively similar results, although the number of informative families in a data set increased (data not shown). Marker A is linked, with a specified recombination fraction (θ), to disease locus B. Locus C is not linked to either A or B. Once parental genotypes are selected, the remaining members of the family are generated on the basis of family structure and size, the rules of Mendelian segregation, and the θ value between the marker locus and the linked disease locus. In the sets of families in which assortative mating occurs, it is assumed that there is a certain probability, β, that pairs with con-

Table 1

Input Parameters for Generation of Families for Disease Model 1

A. Disease Status for Two-Locus Genotypes[a]

	Locus B		
Locus C	11	12	22
11	0	0	0
12	0	0	1
22	0	1	1

B. Allele Frequencies for Loci A–C

	Locus		
Allele	A	B	C
1	.40	.70	.60
2	.60	.30	.40

C. Expected Frequencies for Two-Locus Disease Genotypes[b]

	Locus B		
Locus C	11	12	22
11	.18	.15	.03
12	.23	.20	.04 (D)
22	.08	.07 (D)	.01 (D)

[a] Data are probabilities of being affected.

[b] (D) = disease; all other genotypes are normal.

cordant disease phenotypes will mate and a probability, $1 - \beta$, that pairs will mate at random. This simulates data based on models presented by, for example, O'Donald (1960) and Falk (1971), reflecting a sample from a population in which a fraction, β, of the population consists of mating pairs exhibiting assortative mating and the remaining fraction, $1 - \beta$, mates at random. To implement this, a random number is chosen to determine whether a pair mates at random or assortatively. If the choice is random mating, generation of the nuclear family proceeds as described above; if the choice is assortative mating, then the disease phenotype of the second parent selected must match the first. Each disease genotype randomly generated is tested and is retained only if a phenotypic match is obtained. Thus the parents selected may have different disease *genotypes* but must be concordant for disease *phenotype*. This can be described as phenotypic assortative mating, in which affected individuals mate more often than would be expected in a random-mating population but in which the choice of mate is based on the phenotype and not on the underlying genetic basis of the disease in the two individuals (Spence et al. 1993). The simulation programs were written in a general way, so that more than two disease-determining loci can be specified, thereby producing any desired disease model, and so that additional marker genotypes, linked to one or more of the disease loci, can be generated.

Two two-locus disease models have been considered here. The first is a two-locus threshold model; the second is a heterogeneity model, in which a dominant phenotype at either of the two disease loci results in disease. In each case the θ value between the marker locus A and the linked disease locus B is .02. For purposes of comparison, a single-locus homogeneous model is included, in which disease locus B alone determines disease status and is again linked to locus A, with $\theta = .02$. The explicit details of each of the models are given below. For each model, the generation of family material to be analyzed was done in the same way.

For each of the five family structures and for each of five mating schemes, including random mating (which is equivalent to $\beta = .0$) and four levels of assortative mating, 40 replicates of 50 families each were generated. Families (or pedigrees) were selected for inclusion only if they had at least two affected individuals in at least one sibship, thus assuring that the families would be somewhat informative for linkage analysis. The mating schemes represent five levels of assortative mating, where $\beta = .0$ (for random mating), .2, .4, .6, or .8. After it was determined that family structure did not appear to have an effect on the estimate of θ (see Results), families of all five family structures were pooled for random mating and for each level of assortative mating. The 10,000 resulting families for each mating scheme were then resampled in 100 replicates of 100 families each, for each of the genetic models. A second, independent run was performed for each of the two-locus disease models and for the one-locus model, in order to ascertain the stability of the results. Results from the two runs for each

Table 2

Input Parameters for Generation of Families for Disease Model 2

A. Disease Status for Two-Locus Genotypes[a]

	Locus B		
Locus C	11	12	22
11	0	1	1
12	1	1	1
22	1	1	1

B. Allele Frequencies for Loci A–C

	Locus		
Allele	A	B	C
1	.40	.98	.97
2	.60	.02	.03

C. Expected Frequencies for Two-Locus Disease Genotypes

	Locus B		
Locus C	11	12	22
11	.90 (n)[a]	.04 (d2)	<.0009
12	.06 (d1)	.002 (d3)	<.0009
22	<.0009	<.0009	<.0009

[a] n = normal.

model were comparable and were therefore combined, resulting in 200 replicates of 100 families each, for each model and for each mating scheme. The results presented are based on these final sets of 200 replicates.

Disease Model 1

Table 1a shows the disease status for each two-locus genotype at loci B and C, the two disease-determining loci. Individuals with at least three type-2 alleles at the two loci will have the disease; all other genotypes are normal. Allele frequencies at loci B and C were chosen so as to result in a disease prevalence of slightly >.10 in the population from which the sample was selected. Table 1b shows the input values for the allele frequencies at the three loci A–C, and table 1c gives the expected frequencies of the two-locus genotypes. This model results in the appearance of an intermediate mode of inheritance with respect to disease phenotype, with a fraction of disease "heterozygotes" being affected and the remainder being normal.

Disease Model 2

Table 2a shows the disease status for the two-locus genotypes for the heterogeneity model. When appropriate allele frequencies are chosen for loci B and C, the population prevalence is again set at ~.10. Table 2b shows the input values for the allele frequencies at the three loci A–C. The disease gives the appearance of a dominant disease, with an ~.05 disease-allele frequency in the population. By construction, only three of the disease genotypes have frequencies >.001. Table 2c represents these as d1 (B11/C12), d2 (B12/C11), and d3 (B12/C12). Of these, the two most frequent are types d1 and d2. Only the second of these, d2, with a heterozygote genotype at locus B,will be informative for linkage to the marker locus A, which is linked to B; the first, d1, is homozygous for B and thus is uninformative for linkage to A; and the third, d3, is a small, mixed class that may contribute a small amount of linkage information. This model is constructed in such a way that we can understand the consequences of performing linkage analysis by use of disease phenotypes, when we do not know the underlying genotype(s). Without prior knowledge of the nature of the trait, we can only see that the data suggest a dominant form of inheritance. We might, therefore, first analyze the data by assuming a single, dominant, disease locus. On the basis of the results, we might then look for genetic heterogeneity. By knowing the underlying disease *genotypes*—d1, d2, and d3—we can see how reliable our analysis will be.

Homogeneous Disease Model

In order to separate the effects of the two-locus disease structures from those of assortative mating, families were also generated under the assumption that only locus B, linked to marker locus A with $\theta = .02$, determines the disease state. Samples were generated as described above, for a combined total of 200 replicates of 100 families each, for each of the five values of β. It was assumed that locus B acts as a simple dominant in determining the disease state. Disease prevalence in the population was again set at ~.10.

Testing the Reliability of Generated Data

In order to be sure that the families generated reflected the input parameters reliably, I looked at population statistics, including parental-allele frequencies, disease-phenotype frequencies, and mating-class frequencies, in both the total (unselected) set of families generated and in the families selected for linkage analysis (i.e., families with at least two affected individuals in a sibship).

Methods of Analysis

Once the data were tested and considered to be reliable, linkage analysis was performed by use of LIPED (Ott 1974, 1976). The families were analyzed on the assumption that there was a single disease locus. A partially dominant mode of inheritance was assumed for model 1, and a dominant mode of inheritance with complete penetrance was assumed for both model 2 and for the single-disease-locus homogeneous model. Allele frequencies and other parameters of the disease model were estimated from the selected family material. Comparisons were then made between linkage results for the five levels of assortative mating, with $\beta = .0$, .2, .4, .6, or .8. For disease model 2, tests of heterogeneity were performed by use of the computer program HOMOG (Ott 1984), and estimates were made of both the level of heterogeneity and the θ value for the subset of families informative for linkage. Additionally, because the information on the true disease genotype was known, it was possible to estimate the reliability of predicting the disease-phenotype subclass to which each family belonged, on the basis of the posterior probability of linkage.

Results

Population Statistics for Genetic Model 1

Population statistics, generated to test the reliability of the simulations, showed that in all cases the observed frequencies in the unselected samples were in very good agreement with the expected values, with no significant χ^2's (data not shown). The selected samples show the expected increase in the frequency of disease-determining alleles, affected phenotypes, and matings involving one or two affected parents. Reassuringly, however, the allele frequencies of the uninvolved marker locus A remained close to the input values. No sex differences were

Table 3

Observed and Expected Frequencies of Founder Pairs Based on Disease Phenotypes: Disease Model 1 (200 Matings)

	Mating Class		
Type of Mating and Population	a × a	a × n[a]	n × n[a]
Random:			
Expected	.016	.219	.766
Unselected	.019	.208	.773
Selected	.035	.334	.631
20% Assortative:[b]			
Expected	.038	.175	.788
Unselected	.038	.177	.785
Selected	.074	.299	.627

[a] n = normal.

[b] $\beta = .2$.

seen. As an illustration, table 3 shows the observed and expected frequencies of founder mating pairs, on the basis of disease phenotypes.

The influence of assortative mating is also as expected in the unselected samples, with a higher proportion of affected × affected (a × a) matings than is seen in the random-mating sample (table 3). As before, the selected samples show a further increase in matings with affected parents. For example, with a disease prevalence of .125, the expected frequency of a × a matings in a random-mating sample would be $(.125)^2$, or ~.016. The observed, unselected frequency was .019, which is not significantly different from the expected value (table 3). In a population with assortative mating accounting for 20% of the matings, the expected frequency of a × a matings is $(.2)(.125) + (.8)(.125)^2$, or ~.038, more than double that of the random-mating sample; the observed frequency was .038. As expected, the selected samples with 20% assortative mating have even higher frequencies of a × a matings (table 3). The results of testing model 2 for the reliability of the simulation programs showed the same expected agreement as was shown by the results of testing model 1 (results not shown).

Single-Locus Homogeneous Model

Parameters for the linkage analysis were based on the selected sample of families. Disease-allele frequency was estimated to be .05, on the basis of both the estimated prevalence and the assumption of a dominant mode of inheritance with complete penetrance. Unrelated family members were used to estimate marker-allele frequencies. The average value of the expected θ (E[θ]), and of the expected Z (E[Z]), estimated at the position of the maximum Z (Z_{max}) for each replicate, were obtained for each level of assortative mating, on the basis of the 200 replicates of 100 families, as described above. Figure 2*a* shows the mean and standard error (SE) for E(θ) for the five levels of assortative mating, .0–.8. The values are not significantly different from one another ($P = .46$) and are quite close to the true θ value between loci A and B. Figure 2*b* shows the mean and SE of E(Z). With increasing assortative mating, there is a significant decrease ($P = .0001$) in average Z_{max}. The decrease is ~36% in the progression from random mating ($\beta = .0$) to $\beta = .8$. Since assortative mating increases the frequency of homozygous × homozygous (h × h) mating classes (table 3), increasing the level of assortative mating tends to increase the number of bilineal families (i.e., families in which disease is segregating on both sides of the family) in a sample. Our observation can therefore be compared with the conclusions drawn by Hodge (1992), who showed that, under random mating, there will be a loss of information when bilineal families are used, particularly when phase is unknown. In these simulations, we have a combination of phase-known and phase-unknown families. One would expect that we would therefore see an information-loss effect that is intermediate. Qualitatively, that is what we see, with a 36% decrease in average Z_{max}, somewhere between the

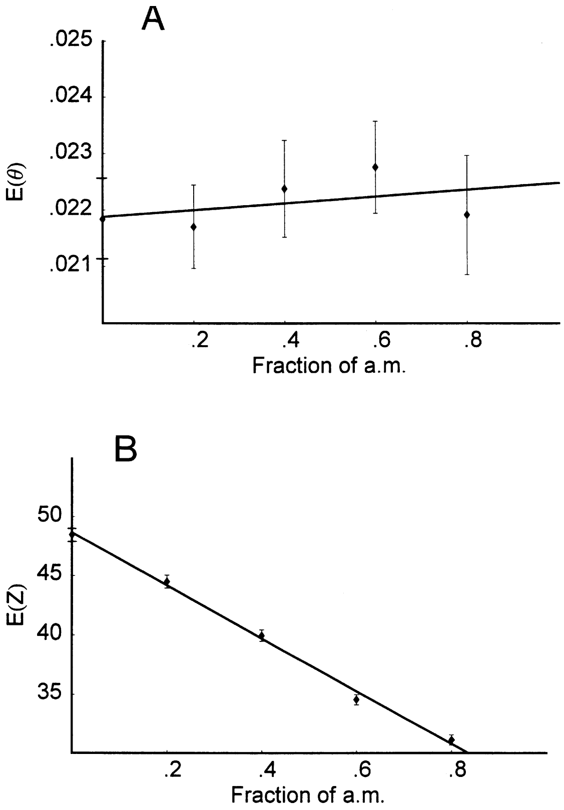

Figure 2 Results of linkage parameter estimates for single-locus homogeneous model, at five levels of assortative mating. Expected values and SEs are based on 200 replicates of 100 families each, at each level of assortative mating. Error bars represent ±1.96 SE. *a*, E(θ): .0218, .0217, .0224, .0228, and .0219. *b*, E(Z) (Z_{max}): 48.50, 44.49, 39.97, 34.55, and 31.12.

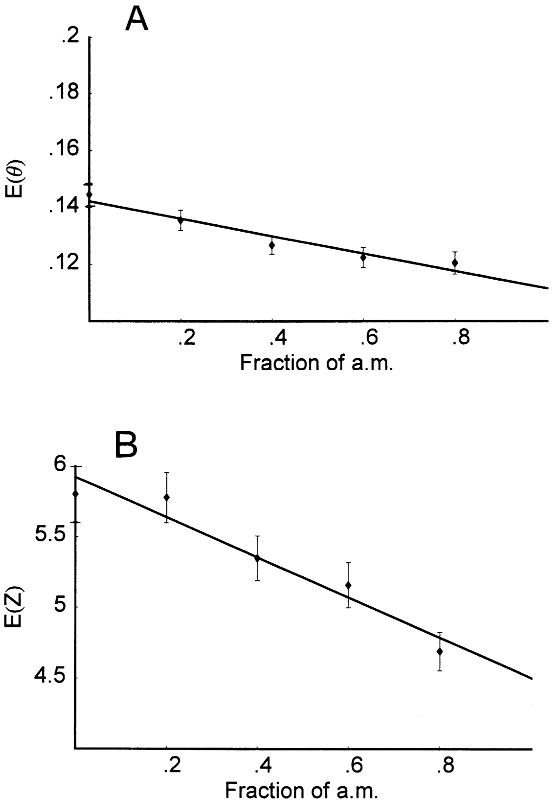

Figure 3 Results of linkage parameter estimates for model 1, at five levels of assortative mating. Expected values and SEs are based on 200 replicates of 100 families each, at each level of assortative mating. Error bars represent ±1.96 SE. *a*, E(θ): .144, .135, .127, .122, and .120. *b*, E(*Z*) ($Z_{\max}$): 5.80, 5.78, 5.35, 5.16, and 4.69.

10% loss in "less extreme" bilineal data sets and the 50% loss in "extremely bilineal" data sets, which were found by Hodge.

Genetic Model 1

Parameters for the linkage analysis were again based on the selected sample of families. The range of disease-allele frequency in the sample was .18–.26, increasing with the level of assortative mating in the sample. The probability of an affected individual, given a "disease" heterozygote, was assumed to be .65, which was consistent with the data. Unrelated family members were used to estimate marker-allele frequencies. Estimates of θ and SE for 40 replicates of 50 families per replicate were obtained for each family structure and for each level of assortative mating. In general there were no significant differences between family structures, although the large pedigrees (structure 4) tended to give higher estimates of E(θ) and smaller SEs, making the difference between E(θ) for family structure 4 of borderline significance when compared with those for the other family structures. The value of E(*Z*) varies, as might be expected for different family structures, with structure 1 being the least informative and structure 4 being the most informative. Because of the similarities in the estimates of E(θ), all family structures were pooled for each level of assortative mating and were resampled to give 100 replicates of 100 families each, for each of the five levels of assortative mating. As mentioned above, results from a second, independent run were comparable to those of the first run, so the two runs were combined, resulting in 200 replicates of 100 families each for each level of assortative mating.

Figure 3*a* shows the mean and SE for E(θ) for the five levels of assortative mating, .0–.8. With an increase in the level of assortative mating, there is a small but significant systematic decrease (.144–.120; $P = .007$) in the value of E(θ). This correlation cannot yet be explained, but, since it is not present in the homogeneous model, it may be an artifact of the method of analysis. For example, the correlation in this model is sensitive to the choice of disease-allele frequency. It may also be affected by the assumption of a single-locus disease model rather than the correct, two-locus model. This will be explored in future simulation studies, by comparing these results with those from analyses with different assumptions. The means and SEs for E(*Z*) are given in figure 3*b*. There is a significant ($P = .006$) decrease in average $Z_{\max}$, ~20% (5.8–4.7), in the progression from random mating ($\beta = .0$) to $\beta = .8$.

Genetic Model 2

Once again LIPED was used to analyze the data, and, on the basis of family-segregation information, a single-locus, dominant disease trait with complete penetrance was assumed. Despite the "complex" nature of this disease model, its construction is such that it gives the appearance of a single-locus, dominant trait. The disease-allele-frequency range was .12–.24, again increasing with the level of assortative mating in the sample and the other assumptions. Again, unrelated family members were used to estimate marker-allele frequencies. Evidence for linkage is present in all samples, and among family structures there are no significant differences in the estimates of E(θ). For this model the value of E(θ) is even greater than that in model 1. Once again, with an increase in the level of assortative mating, there is a slight, systematic decrease in E(θ) (.268–.251; $P = .02$) (fig. 4*a*). For this model there is a more pronounced decrease in power (fig. 4*b*), with average $Z_{\max}$ decreasing ~43% (7.2–4.2) in the progression from random mating ($\beta = .0$) to $\beta = .8$ ($P = .002$).

After performing linkage analysis, I also analyzed the results by using HOMOG (Ott 1984), to see whether I could detect significant genetic heterogeneity. In all cases

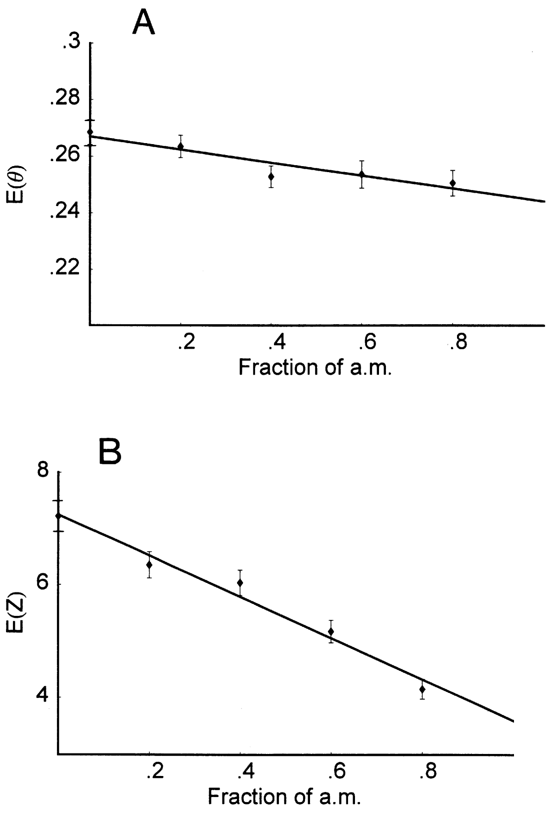

Figure 4 Results of linkage parameter estimates for model 2, at five levels of assortative mating. Expected values and SEs are based on 200 replicates of 100 families each, at each level of assortative mating. Error bars represent ± 1.96 SE. *a,* E(θ): .268, .263, .253, .254, and .251. *b,* E(Z) ($Z_{\max}$): 7.22, 6.35, 6.03, 5.17, and 4.15.

the family data gave strong evidence of genetic heterogeneity. For example, in the random-mating sample of 200 families for family structure 2, the family structure used by CEPH, χ^2 for heterogeneity was 9.61 (P = .002). An estimated 45% of the families were informative for linkage, with θ = .05. These estimates are in reasonable agreement both with the generated values of ~40% of families being informative for linkage and with θ = .02.

As part of the output from HOMOG, one has a posterior probability for each family, indicating the probability that it is part of the subset of families informative for linkage to marker A. This probability is based on the Z values for the family and on its ranking among families. Since we know the disease *genotype* for each individual—that is, whether the disease is of type d1 (i.e., heterozygous at locus C), type d2 (i.e., heterozygous at locus B), or type d3 (i.e., heterozygous at both loci)—we can see what fraction of families can, on the basis of posterior probabilities, be correctly assigned to the informative subset (d2) or the uninformative subset (d1). Figure 5 shows the results for a sample with family structure 2. We would expect that most of the families with disease genotype d1 would have low posterior probabilities of linkage and that those with disease genotype d2 would have high probabilities. In general, this is seen to be the case, although, by chance, some families will be incorrectly assigned by this designation. For this model, ~80% of the sibships, on average, were correctly classified as part of the subset of families "informative" or "uninformative" for linkage to marker A, when they were separated by the value of α, the estimated fraction of linked families, for that sample. Based on results of linkage to a single marker locus with only a moderate level of heterozygosity, the ability to correctly identify the linked subset of families is therefore somewhat limited.

Discussion

Linkage analysis has moved from the realm of linking sets of genetic loci, known to segregate as single-locus Mendelian traits, to the challenges of mapping a locus or loci known to have an effect on the determination of disease state for a complex genetic trait. There is no single clear-cut strategy to follow in such studies, and many promising approaches are being explored. Since complex traits by their very nature may have very different genetic mechanisms, it is unlikely that the same approach will prove to be best for all such traits. When family data are available, it is possible to attempt a classical linkage analysis, in which all members of the family are included and adjustments are made, if possible, for the assumptions generally required when linkage analysis is performed. Approaches that have been proposed

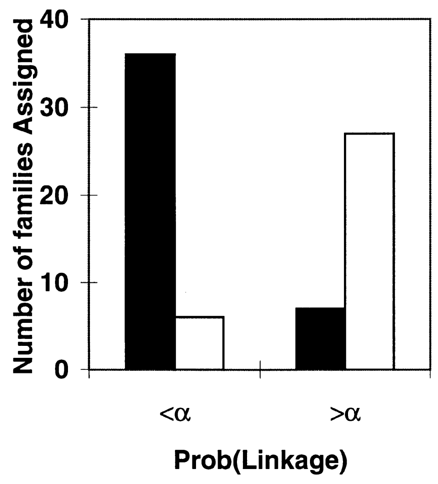

Figure 5 Separation of families, on the basis of posterior probabilities of linkage and known disease genotypes d1 (*blackened bars*) and d2 (*unblackened bars*). Number of families of each disease type with posterior probability <α or >α, estimated fraction of linked families, is displayed. Families uninformative for linkage are not shown.

and/or tested include (1) limiting the analysis to a single large pedigree, to maximize the possibility of a homogeneous genetic basis for the trait (Egeland et al. 1987); (2) allowing for reduced penetrance with respect to the trait, so that heterogeneity, if present, or misclassification can be absorbed as noise by the model (Vieland et al. 1992*b,* 1993); (3) using high disease-allele frequencies in multipoint analyses, as a means of increasing the robustness of results when other genetic parameters may be misspecified (Risch and Giuffra 1992); and (4) analyzing the data under a spectrum of models, to see how the results vary. Other assumptions, such as the mating structure of the underlying population, are often ignored, despite the fact that, for some complex traits, particularly behavioral traits, a certain level of assortative mating may be observed. In this study I have attempted, by using a simulation model, to assess the consequences that some of these strategies have on the outcome of linkage analyses. I have chosen to look at two models in which disease is defined by two, unlinked loci. One model resembles a threshold model, in which a certain number of "disease" alleles at either or both of the loci are required for expression of the disease phenotype. The second is a model with genetic heterogeneity, constructed so that we are able to analyze the consequences of making certain assumptions at the analytic stage. Since the goal of many studies of complex traits is to try to determine the location of a locus or loci contributing to disease state—and to do so with enough precision so that other methods can be used to identify and ultimately clone the gene(s)—it is important to know whether standard techniques of linkage analysis enable us to reach such goals.

On the basis of the results of my simulation studies, there is little evidence to suggest that assortative mating has a major effect on the θ estimate obtained in linkage analysis. In the homogeneous model there was no significant difference between the estimated values of $E(\theta)$ for the five levels of assortative mating. Although we saw some correlation between the level of assortative mating and the value of $E(\theta)$ in the two-locus models, the variation in $E(\theta)$ was minor when compared with the $E(\theta)$ inflation attributable to the two-locus disease models. The correlation could be an artifact due to the analysis of a two-locus disease under the assumption of a single-locus disease, or it could be an incorrect specification of disease-allele frequency. The latter possibility is suggested by the observation that the estimated disease-allele frequency varies with the level of assortative mating and that the value used in the analysis had some influence on the correlation between the level of assortative mating and $E(\theta)$ (data not shown). The reason for the correlation is not yet clear and will be explored in ongoing studies. Although, as with all simulation studies, such observations are limited to the models studied, the characteristics of our two models are somewhat different and resemble situations that have been observed. In both models and for all family structures considered, estimates of θ are close, whether the selected families are from a random-mating population or from a population in which there is a degree of assortative mating for the trait being studied. In the homogeneous model, the estimates of $E(\theta)$ are very close to those expected, and there is absolutely no difference based on the level of assortative mating. On the other hand, all three models studied indicate that assortative mating has an effect on the power of the analysis, with high levels of assortative mating resulting in lower Z values. The magnitude of this difference varies with the model, and even the simple, single-locus model shows a significant decrease in power. Thus, when assortative mating is suspected, larger sample sizes may be necessary for detection of linkage. This conclusion agrees with that of Sribney and Swift (1992) in a study showing that assortative mating together with genetic heterogeneity can greatly increase the sample size required for detection of linkage when sib-pair analysis is used. Although the disease models and methods of analysis differ from those presented here, the basic conclusion is in agreement with that presented here.

Since assortative mating increases the frequency of h × h mating classes (table 3), the effect is to increase the number of bilineal families (i.e., families in which disease is segregating on both sides of the family) in a sample. Hodge (1992) studied the effect of bilineality on the results of linkage analysis for a single-gene disease with no heterogeneity. She concluded that, although there may be some loss of power, particularly in phase-unknown families, it was not so great that bilineal families should be excluded from analyses. The results of the present study, both those for a single-locus homogeneous trait and those for the two two-locus models, are consistent with her results. The level of assortative mating (which reflects the proportion of bilineal families) has an effect on power. The magnitude of that effect will vary with the underlying genetic model. Other studies, discussed in Spence et al. (1993), using other methods (e.g., affected-relative pairs; D. T. Bishop and R. C. Elston, personal communication), have reached similar conclusions. Positive assortative mating does not compromise the validity of linkage analysis, even if its presence is not explicitly assumed in the analysis. Estimates of θ are not drastically altered, although some loss of power is likely (Spence et al. 1993; members of MacArthur Research Network I, personal communication).

The results of the present study also confirm observations of earlier studies: that (1) linkage to a disease trait can be detected even if the disease is incorrectly defined as a single-locus (homogeneous) trait rather than as the correct, two-locus trait (e.g., see Greenberg 1990;

Goldin 1992; Vieland et al. 1992*a*) and (2) the estimated θ will be much larger than the true θ between the linked disease-defining locus and the marker locus (e.g., see Clerget-Darpoux et al. 1986; Risch and Giuffra 1992; Vieland et al. 1992*b,* 1993). Such confirmation in new models is worthwhile, since results from simulation studies are, by their very nature, restricted to the models used in the simulations. Agreement of general observations derived from use of different models and different strategies provides evidence for somewhat robust conclusions, making them more relevant to the analysis of real data. The presence of the second disease locus in the current models makes it extremely difficult to correctly identify the distance between the marker and the linked disease locus, because of the noise caused by the second disease locus. With linkage analysis, therefore, it may not be possible to achieve the goal of narrowing the location of the disease locus sufficiently so that other techniques can be effectively used to identify the gene. If the disease is found to be genetically heterogeneous, as in the second model discussed herein, linkage analysis combined with heterogeneity analysis (using, e.g., HOMOG) will provide both an estimate of the level of heterogeneity and a more accurate estimate of the θ value between the marker and the disease locus. It does not, however, identify with certainty which of the families exhibit the disease form that is informative for linkage to the marker. In the absence of this information, it is not possible to know which families to study further to narrow the range of chromosomal material containing the disease locus. As we have seen, in the model studied, only ~80% of the families, on average, are correctly assigned to the informative (for linkage to marker A) subset of families when posterior probabilities of linkage are used, and ~20% of the uninformative set are, by chance, assigned to the "linked" subset. In this study, we have limited the analysis to a single marker locus, and this limits the amount of information that we can obtain. By increasing the number of marker loci included in the analysis and by choosing highly informative markers, we can devise strategies that will increase the chances of identifying a more homogeneous sample of families with the linked form of the disease and, hence, of identifying more precisely the chromosomal region of interest (Falk 1993, and unpublished data). Ott (1983) has examined the effectiveness of different strategies of family classification, given a heterogeneous trait, and has concluded that the admixture test, used in the program HOMOG, is a reasonably successful strategy.

Other strategies for identifying loci contributing to disease may be more successful when one is dealing with certain complex disease models. For example, methods of affected-sib-pair analysis or affected-pedigree-member analysis (e.g., see Weeks and Lange 1988) are not affected by incomplete penetrance, since such analyses are based on individuals known to be affected (the assumption being that the diagnosis can be reliably made). Such an approach was recently used to successfully identify genes contributing to type 1 diabetes that were not in the HLA region of chromosome 6 (Davies et al. 1994). As a trade-off, however, the power of such nonparametric methods is generally lower than that of classical likelihood methods of linkage analysis (Goldin and Weeks 1993). It is likely that no single approach will prove to be the best approach for the analysis of all complex traits, and an understanding of the problems and limitations of each will be important in deciding how to analyze data for traits of interest.

Acknowledgments

This work was supported by National Institutes of Health grant GM29177 and by a pilot research grant from the John D. and Catherine T. MacArthur Foundation Mental Health Research Network I. I would like to thank Neil Risch, Sue Hodge, and other members of the MacArthur Foundation Mental Health Research Network I for helpful and productive discussions. Reviewers of an earlier version of this paper made many constructive suggestions; their comments are gratefully acknowledged. Most of the programming for the simulation studies was done by Lynh Wu; her help is gratefully acknowledged.

References

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–399

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, et al (1994) A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371: 130–136

Durner M, Greenberg DA, Hodge SE (1992) Inter- and intrafamilial heterogeneity: effective sampling strategies and comparison of analysis methods. Am J Hum Genet 51:859–870

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

Falk CT (1971) The combined effects of positive assortative mating and selection. Heredity 27:125-136

——— (1993) Effective strategies for linkage analysis of genetically heterogeneous traits. Am J Hum Genet Suppl 53: 65

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

Goldin LR, Weeks DE (1993) Two-locus models of disease: comparison of likelihood and nonparametric linkage methods. Am J Hum Genet 53:908–915

Greenberg DA (1990) Linkage analysis assuming a single-locus mode of inheritance for traits determined by two loci: inferring mode of inheritance and estimating penetrance. Genet Epidemiol 7:467–479

Greenberg DA, Hodge SE (1989) Linkage analysis under "random" and "genetic" reduced penetrance. Genet Epidemiol 6: 259–264

Hodge SE (1992) Do bilineal pedigrees represent a problem for linkage analysis? Basic principles and simulation results for single-gene diseases with no heterogeneity. Genet Epidemiol 9:191–206

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

Merikangas KR (1982) Assortative mating for psychiatric disorders and psychological traits. Arch Gen Psychiatry 39: 1173–1180

O'Donald P (1960) Assortive mating in a population in which two alleles are segregating. Heredity 15:389–396

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

——— (1976) A computer program for linkage analysis of general human pedigrees. Am J Hum Genet 28:528–529

——— (1983) Linkage analysis and family classification under heterogeneity. Ann Hum Genet 47:311–320

——— (1984) Analysis of human genetic linkage, 1st ed. Johns Hopkins University Press, Baltimore

Risch N, Giuffra L (1992) Model misspecification and multipoint linkage analysis. Hum Hered 42:77–92

Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53:1127–1136

Spence MA, Bishop DT, Boehnke M, Elston RC, Falk C, Hodge SE, Ott J, et al (1993) Methodological issues in linkage analyses for psychiatric disorders: secular trends, assortative mating, bilineal pedigrees: report of the MacArthur Foundation Network I Task Force on Methodological Issues. Hum Hered 43:166–172

Sribney WM, Swift M (1992) Power of sib-pair and sib-trio linkage analysis with assortative mating and multiple disease loci. Am J Hum Genet 51:773–784

Vieland VJ, Greenberg DA, Hodge SE (1993) Adequacy of single-locus approximations for linkage analyses of oligogenic traits: extension to multigenerational pedigree structures. Hum Hered 43:329–336

Vieland V, Greenberg DA, Hodge SE, Ott J (1992*a*) Linkage analysis of two-locus diseases under single-locus and two-locus analysis models. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) Genetic Analysis Workshop 7: issues in gene mapping and detection of major genes. Cytogenet Cell Genet 59:145–146

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

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