Genetic Epidemiology 22:156–169 (2002)

Quantification of Type I Error Probabilities for Heterogeneity LOD Scores

Paula C. Abreu,[1,2] Susan E. Hodge,[1,3,4]* and David A. Greenberg[3–5]

[1] Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York
[2] Schering-Plough Research Institute, Kenilworth, New Jersey
[3] Department of Psychiatry, Columbia College of Physicians and Surgeons, New York, New York
[4] New York State Psychiatric Institute, New York, New York
[5] Departments of Psychiatry and Biomathematics, Mount Sinai School of Medicine, New York, New York

Locus heterogeneity is a major confounding factor in linkage analysis. When no prior knowledge of linkage exists, and one aims to detect linkage and heterogeneity simultaneously, classical distribution theory of log-likelihood ratios does not hold. Despite some theoretical work on this problem, no generally accepted practical guidelines exist. Nor has anyone rigorously examined the combined effect of testing for linkage and heterogeneity *and* simultaneously maximizing over two genetic models (dominant, recessive). The effect of linkage phase represents another uninvestigated issue. Using computer simulation, we investigated type I error (*P* value) of the "admixture" heterogeneity LOD (HLOD) score, i.e., the LOD score maximized over both recombination fraction θ and admixture parameter α and we compared this with the *P* values when one maximizes only with respect to θ (i.e., the standard LOD score). We generated datasets of phase-known and -unknown nuclear families, sizes $k = 2$, 4, and 6 children, under fully penetrant autosomal dominant inheritance. We analyzed these datasets (1) assuming a single genetic model, and maximizing the HLOD over θ and α; and (2) maximizing the HLOD additionally over two dominance models (dominant vs. recessive), then subtracting a 0.3 correction. For both (1) and (2), *P* values increased with family size *k*; rose less for phase-unknown families than for phase-known ones, with the former approaching the latter as *k* increased; and did not exceed the one-sided mixture distribution $\xi = (½)\ \chi^2_1 + (½)\ \chi^2_2$. Thus, maximizing the HLOD over θ and α appears to add considerably *less* than an additional degree of freedom to the associated χ^2_1 distribution. We conclude with practical guidelines for linkage investigators. Genet. Epidemiol. 22:156–169, 2002. © 2002 Wiley-Liss, Inc.

Key words: heterogeneity; LOD scores; hetlods; linkage analysis; complex disease

Contract grant sponsor: NIH; Contract grant numbers: MH-48858, DK-31813, MH-28274, DK-31775, NS-27941, and DK-52464.

*Correspondence to: Susan E. Hodge, D.Sc., NYSPI, Unit 24, 1051 Riverside Drive, New York, NY 10032.
E-mail: seh2@columbia.edu

Received 6 June 2001; Accepted 24 August 2001

DOI 10.1002/gepi.0155

INTRODUCTION

Locus heterogeneity represents a major confounding factor in linkage analysis. Locus heterogeneity exists when disease alleles at two or more independently-acting loci each cause the same disease phenotype. Thus, if we perform linkage analysis with a marker located near one of these disease loci, some families will yield a small recombination fraction estimate ($\hat{\theta}$) with that marker, whereas other families will show independent segregation of disease and marker ($\theta = 0.5$). The combination of families will result in a reduction of the linkage signal, and may lead to failure to detect a true linkage altogether.

One way to detect locus heterogeneity is to use a form of linkage analysis that models the heterogeneity by using an additional "admixture" parameter, α, in the analysis. Along these lines, Smith [1963] proposed a model in which a proportion (α) of families have a genetic form linked to the marker of interest ($\theta < 0.5$) whereas the remaining ($1\text{-}\alpha$) families are unlinked ($\theta = 0.5$). Hodge et al. [1983] and Ott [1983] subsequently put this admixture approach into a likelihood-ratio (LR) test framework.

The admixture LR can be used to detect the existence of heterogeneity, once linkage has already been established, and this application has been examined [for example, see Ott, 1991]. Alternatively, one may use this approach to attempt to detect linkage *and* heterogeneity, when no prior knowledge of linkage exists. Testing for genetic linkage under heterogeneity is closely related to the problems of testing for mixture models [e.g., Faraway, 1993; Chernoff and Lander, 1995; Lindsay, 1995].

Consider the usual linkage situation without heterogeneity. For testing the hypothesis of linkage, the LR is $L(\theta)/L(.5)$, where θ denotes the recombination fraction. We define the max LOD score as $\log_{10}(\mathrm{LR})$, maximized with respect to θ. The conventional criterion for statistical significance is that the max LOD score should be ≥ 3.0. Asymptotically, the max LOD score multiplied by $2(\mathit{ln}\ 10)$, i.e., $2 \times \max_\theta (\mathit{ln}\ \mathrm{LR})$, follows a one-sided chi-square distribution with one degree of freedom (χ^2_1), and thus the 3.0 LOD score criterion corresponds asymptotically to a significance level of ≈ 0.0001.

However, these asymptotic χ^2 relationships no longer hold under heterogeneity. It has long been recognized that standard asymptotic results do not hold for a heterogeneity or mixture model [e.g., Davies, 1977, 1987]. The null hypothesis H_0, of no linkage (i.e., $\theta = 0.5$ or, equivalently, $\alpha = 0$) corresponds to an infinite set of parameter points. When one tests simultaneously for linkage and heterogeneity, one aims to declare a linkage significant while simultaneously allowing for α as a nuisance parameter. (Or in some situations, one may be interested in estimating α as well, although see Whittemore and Halpern [2001] and Vieland and Logue [2001] for problems inherent in that goal.) Although there are two parameters (θ and α) under the alternative hypothesis H_A, one parameter (α) disappears under H_0. This leads to a degenerate situation and a test statistic that is no longer asymptotically distributed as χ^2 [Wald, 1949].

Several researchers have considered the asymptotic distribution of the max HLOD statistic [Hodge et al., 1983; Ott, 1983; Risch, 1989; Faraway, 1993; Chernoff and Lander, 1995; Chiano and Yates, 1995; Liu and Shao, 1999]. Hodge et al. [1983] claimed that the asymptotic null distribution of the (two-sided) admixture test statistic was χ^2_1, whereas Risch [1989] and Ott [1991] have stated that it is χ^2_2. However, neither is correct, and it is now generally understood that the appropriate distribution is a mixture of two χ^2's [see, for example, Nyholt, 2000]. For k = 2 (i.e., 2 children per family), Chernoff and Lander [1995] introduced an alternative parametrization for the phase-known situation and thereby identified the asymptotic distribution of the test statistic as a mixture of two χ^2 distributions, χ^2_1 and χ^2_2. Unfortunately, for k > 2, their regularity conditions no longer apply even under this new parameterization.

Recently, Liu and Shao [1999] have developed a unified approach to characterize the asymptotic distribution for different values of k and for both phase-known and phase-unknown families. They showed that the distribution of 2 *ln* LR always lies between a χ^2 distribution with one degree of freedom and one with two degrees of freedom, i.e., a mixture distribution denoted as $\lambda\chi^2_1 + (1 - \lambda)\chi^2_2$, where $\lambda \in [0, 1]$. However, they did not specify λ, since its value may differ for each case; thus, the true asymptotic distribution of the test statistic is quite complicated. This controversy is of practical interest, given that heterogeneity is one of the most common confounding problems in human genetic linkage studies. In fact, unlike most "nonparametric" methods, LOD-score-based methods allow for testing and detecting heterogeneity.

In this work we used computer simulations to investigate the behavior of the max HLOD statistic, combining the statistical theory of the mixture of distributions with theory of human genetic linkage analysis. We first (1) determined the type I error of the max HLOD statistic, assuming a single model, i.e., maximizing the HLOD over both recombination fraction θ and admixture parameter α. We then (2) performed the same analyses for max HLODs that had also been maximized over two dominance models, dominant and recessive (i.e., the "MMLS" approach to linkage analysis of complex diseases [Greenberg et al., 1998]). For these "MMLS HLODs," we *first* subtracted 0.3 from the scores, before recording them [Hodge et al., 1997].

In both analyses, we examined both phase-unknown and phase-known families. The work has two aims: to determine the actual magnitude of the increase in significance levels for the mixture test in these models; and to give guidelines of how much an investigator should add to the cutoff value used as a test criterion in order to achieve a given probability of type I error.

METHODS

We used computer simulation to generate data sets under the autosomal dominant model with full penetrance. There was no linkage between the disease and the marker, i.e., $\theta = 0.5$ (and thus no heterogeneity), since we were interested only in evaluating significance levels in this study. The disease allele frequency used in all simulations was 0.01. All matings were fully informative for the marker.

We evaluated the max HLOD statistic, based on the admixture likelihood: Consider a population of *n* independent nuclear families, of which α are linked and 1-α are unlinked. Then the likelihood for the ith family, $L_i\ (\alpha,\theta)$, is defined as

$$L_i(\alpha,\theta) = \alpha L_i(\theta) + (1-\alpha)L_i(\theta = .5). \qquad (1)$$

where $L_i\ (\theta)$ represents the likelihood of θ for a linked family. The likelihood of the whole sample is

$$L(\alpha,\theta) = \prod_i^n L_i(\alpha,\theta),\ i = 1,\ldots,n$$

where $L(\alpha,\theta)$ can be maximized with respect to θ and α simultaneously, yielding MLEs $\hat{\alpha}$ and $\hat{\theta}$.

Consider an H_0 of no linkage against an H_A of linkage and heterogeneity. The max HLOD score is defined as:

$$\text{Log}_{10}\ \text{max}_{\alpha,\theta}\ \text{LR} = \log_{10}\left[\frac{L(\hat{\alpha},\hat{\theta})}{L(\alpha=0,\theta=\frac{1}{2}}\right] \tag{2}$$

Multiplying this score by 2(*ln* 10) yields the max HLOD statistic.

In what follows, to avoid confusion, we refer to LODs based on $\log_{10}$ as LOD "scores," whereas when they are multiplied by 2(*ln* 10), they are LOD "statistics." Thus, the expression in equation (2) is the "max HLOD score," whereas "max HLOD statistic" denotes 2(*ln* 10) times that quantity.

Note that when $\theta = 0.5$, α is unidentifiable; i.e., any value of α produces the same likelihood in equation (1). Thus, although in the denominator of equation (2) we assumed α to be equal to zero (0% linked families), the actual value of α is immaterial under H_0.

For nuclear families with a fully penetrant autosomal dominant condition, matings are either (1) phase-unknown or (2) phase-known; or they fall "between," if there is partial phase information. In our simulations we examined both phase-unknown and phase-known families:

1. In a *k*-child phase-unknown family, say we observe *r* children in one "concordance class" and *k-r* in the other. That is, depending on the phase of the informative parent, either *r* children would be recombinant, and *k-r*, nonrecombinant; or vice versa. Then the likelihood for that family is the sum of two binomial terms, one for each possible phase, weighted relatively by the probability of the corresponding phase (assumed to be ½ in the absence of linkage disequilibrium), plus a third binomial term representing the probability of no linkage. Thus, the heterogeneity likelihood for the ith family is a mixture of three weighted binomial distributions with parameters B(θ, k), B(1-θ, k), and B(0.5, k), respectively:

$$L_i(r;\theta,\alpha) = \frac{\alpha}{2}\left[\theta^{r_i}(1-\theta)^{k_i-r_i} + \theta^{k_i-r_i}(1-\theta)^{r_i}\right] + (1-\alpha)(.5)^{k_i} \tag{3}$$

2. In contrast, a mating is phase-known if the distribution of alleles on chromosome pairs in the parents can be determined without ambiguity. To generate phase-known families we used three-generation families, i.e., grandparents, parents, and offspring. One parent was informative (i.e., was affected and was heterozygous at the marker locus), and this informative parent had grandparental information. The other parent was uninformative at the disease locus. Now the recombinants (*r*) can be counted directly, from *k* offspring in a family, and the heterogeneity likelihood is a mixture of only two binomial distributions: B(θ, k) and B(0.5, k):

$$L_i(r;\theta,\alpha) = \alpha\theta^{r_i}(1-\theta)^{k_i-r_i} + (1-\alpha)(.5)^{k_i},\ i = 1,\ldots,n. \tag{4}$$

We generated data sets of either 40 phase-unknown or 40 phase-known families. The phase-unknown families consisted of two parents plus a fixed number of children (k = 2, 3, 4, and 6), as described above, whereas the phase-known families consisted of three-generation pedigrees with two grandparents and two parents plus a fixed number of children (k = 2, 3, 4, and 6), as described above.

For k = 2 and k = 3, we simulated N = 3,000 datasets, whereas for k = 4 and k = 6 there were N = 5,000 data sets. For k > 2, we required at least two affected children in order for a family to be ascertained, whereas for k = 2 only one affected child was required. Because families were generated under dominant inheritance with full penetrance, affected and unaffected offspring provided equal information for linkage.

In Maximizing HLOD Under a Single Model, all datasets were analyzed by maximizing the HLOD over θ and α. We used GENEHUNTER [Kruglyak et al, 1996], which implements the admixture test described above, the HLOD score.

In Maximizing the Maximum HLOD Over Dominance Model, we maximized the HLOD as above, then maximized this max HLOD over two dominance models. Following the guidelines of Hodge et al. [1997], we chose an arbitrary penetrance of 50% (even though the data were generated under full penetrance), then analyzed the data once assuming dominant and once assuming recessive inheritance (MMLS approach, [Greenberg et al, 1998]). We determined the larger of the two maximum HLOD scores from these two dominance models, then subtracted 0.3 for multiple testing, so that the resultant score would be comparable to the "MMLS-C" (MMLS-Corrected) score [Hodge et al., 1997]. We will refer to this maximized HLOD, with 0.3 already subtracted, as our "baseline MMLS HLOD score." We calculated these baseline MMLS HLODs, then determined the corresponding distribution of *P* values.

For each simulation, the observed significance levels, P(Z), were determined as a function of the cutoff value *Z*, as follows:

$$P(Z) \equiv (\text{number of data sets yielding max HLOD} \geq Z)/N,$$

where *N* represents the number of data sets generated for that simulation (i.e., 3,000 or 5,000).

These significance levels were plotted vs. *Z*, and were compared with the corresponding one-sided critical values from χ^2 curves with 1 and 2 degrees of freedom. For the figure, rather than transform max HLOD "scores" to "statistics," we plotted them as they were, and instead transformed the χ^2 curves, by dividing by 2(*ln* 10) = 4.605. Throughout, all χ^2 distributions are one-sided, since the test for linkage is one-sided ($\theta < 0.5$). The tables also give 95% confidence intervals around those *P* values, calculated using the normal approximation to the binomial when $N \times P \geq 5$, and exact binomial calculations for $N \times P < 5$ (where *N* indicates the number of datasets per simulation). All confidence intervals are two-sided, except when the observed number was zero, in which case an exact one-sided confidence interval was calculated.

RESULTS

Maximizing HLOD Under a Single Model

Phase-unknown families

For k = 2 and k = 3 families, it can be shown analytically that significance levels for the HLOD are *identical* to those for the LOD. Details are in the Appendix. Thus, no correction is needed for including heterogeneity in the likelihood. However, also note that there is no additional information for heterogeneity in those small phase-unknown families (see Discussion). For k = 4 and k = 6, Figure 1 shows observed significance levels as curves plotted against Z. For k = 4, these significance levels are raised only slightly above the one-sided χ^2_1; they are raised slightly more for k = 6. Table I gives the results for specified values of Z, along with 95% confidence intervals.

As predicted by Liu and Shao [1999] (see Introduction above), these observed significance levels are bounded below by a one-sided χ^2_1 and above by a one-sided χ^2_2. Moreover,

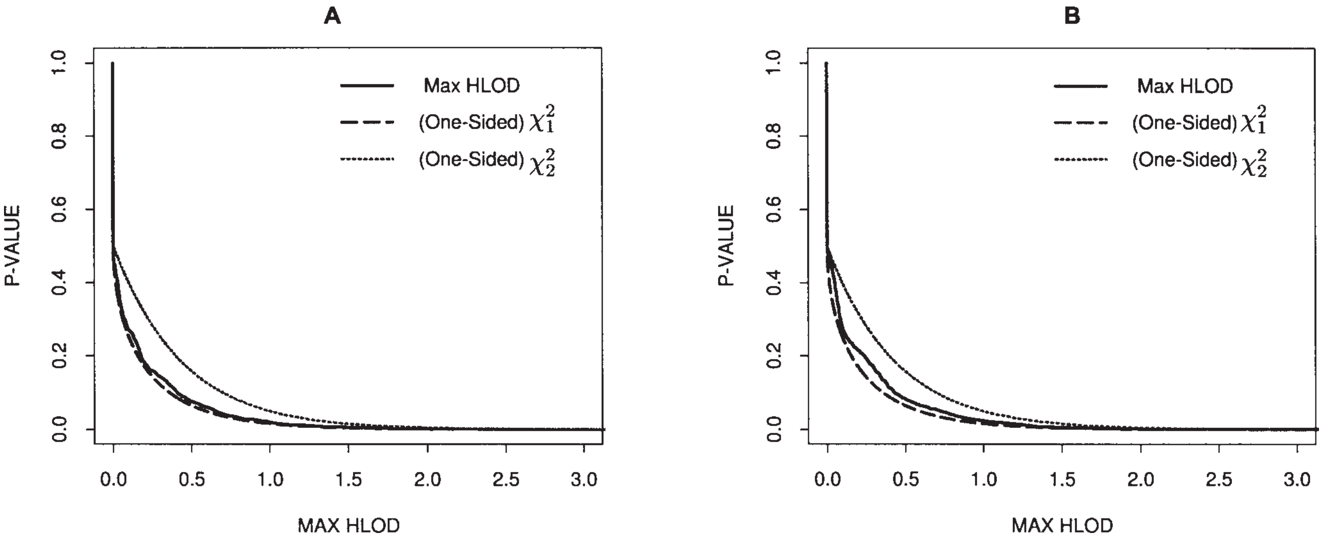

Fig. 1. Significance levels for HLOD scores for phase-unknown families (**A**) of four children, (**B**) of six children. Also see Table I. The χ^2 curves are graphed on a "lod" scale; see text for details.

TABLE I. Maximizing HLODs Under a Single Genetic Model, in Phase-Unknown Families: Observed Significance Levels (P) and Associated 95% Confidence Intervals (See Fig. 1)*

Test size		k = 2		k = 4		k = 6		*P* value
(one-sided)	Z	*P*	95% CI	*P*	95% CI	*P*	95% CI	for ξ
0.05	0.59	0.039	[0.032, 0.046]	0.061	[0.054, 0.068]	0.066	[0.059, 0.073]	0.089
0.025	0.83	0.019	[0.014, 0.024]	0.029	[0.024, 0.034]	0.037	[0.032, 0.042]	0.050
0.01	1.17	0.009	[0.006, 0.012]	0.012	[0.009, 0.015]	0.017	[0.013, 0.021]	0.022
0.005	1.44	0.004	[0.002, 0.006]	0.007	[0.005, 0.009]	0.007	[0.005, 0.009]	0.012
0.001	2.0	0.001	[0.000, 0.003][a]	0.001	[0.000, 0.003][a]	0.001	[0.000, 0.003][a]	0.003
0.0001	3.0	0	[0, 0.0010][a]	0	[0, .0007][a]	0	[0, .0007][a]	0.0003

*k = number of offspring per family. Confidence intervals are two-sided, except when observed number equaled zero; see text.

[a] Exact binomial confidence interval.

not only do these significance levels *not* reach χ^2_2; they do not even reach (½)χ^2_1 + (½)χ^2_2 (i.e., an equally weighted mixture of the two χ^2 curves). In what follows, we denote this mixture distribution by ξ. The *P* values for this mixture distribution are shown in the rightmost column in Tables I–IV. In most cases, not even the upper limit of the 95% confidence interval reaches the *P* value for the mixture distribution. Rather, our observed significance levels for the phase-unknown cases appear closer to the χ^2_1 distribution, at least up to size k = 6. Thus, "adding" a whole degree of freedom would be extremely overconservative; even adding "half" a degree of freedom appears conservative.

Phase-known families

Phase-known families consistently display slightly higher significance levels than phase-unknown families, but even these levels do not exceed those for the ξ mixture distribution, within sampling variation. Table II gives the results and 95% confidence intervals for specified Z values. For phase-known cases, we do not see an increase in significance levels as family size increases. Again, for Z values of interest (up to ~ Z = 3.0), the mixture curve ξ provides a reasonable and somewhat conservative upper bound for the type I error probabilities.

Comparing Tables I and II also reveals that the phase-unknown significance levels approach the phase-known ones as *k* increases, as one would expect.

TABLE II. Maximizing HLODs Under a Single Genetic Model, in Phase-Known Families: Observed Significance Levels (P) and Associated 95% Confidence Intervals*

Test size		k = 2		k = 4		k = 6		*P* value
(one-sided)	Z	*P*	95% CI	*P*	95% CI	*P*	95% CI	for ξ
0.05	0.59	0.067	[0.058, 0.076]	0.085	[0.077, 0.093]	0.073	[0.077, 0.093]	0.089
0.025	0.83	0.036	[0.029, 0.043]	0.044	[0.038, 0.050]	0.042	[0.038, 0.050]	0.050
0.01	1.17	0.013	[0.008, 0.017]	0.017	[0.013, 0.021]	0.018	[0.013, 0.021]	0.022
0.005	1.44	0.008	[0.005, 0.011]	0.009	[0.006, 0.012]	0.010	[0.006, 0.012]	0.012
0.001	2.0	0.002	[0.001, 0.003]	0.003	[0.002, 0.005]	0.003	[0.002, 0.005]	0.003
0.0001	3.0	0	[0, 0.0010][a]	0	[0, 0.0010][a]	0.001	[0.000, 0.003][a]	0.0003

*k = number of offspring per family. Confidence intervals are two-sided, except when observed number equaled zero; see text.

[a] Exact binomial confidence interval.

TABLE III. Maximizing HLODs Over Two Dominance Models (Baseline MMLS HLODs), in Phase-Unknown Families: Observed Significance Levels (P) and Associated 95% Confidence Intervals*

Test size		k = 2		k = 4		k = 6		*P* value
(one-sided)	Z	*P*	95% CI	*P*	95% CI	*P*	95% CI	for ξ
0.05	0.59	0.039	[0.032, 0.046]	0.054	[0.048, 0.060]	0.068	[0.061, 0.075]	0.089
0.025	0.83	0.021	[0.016, 0.026]	0.029	[0.024, 0.034]	0.037	[0.032, 0.042]	0.050
0.01	1.17	0.009	[0.006, 0.012]	0.013	[0.010, 0.016]	0.012	[0.009, 0.015]	0.022
0.005	1.44	0.007	[0.004, 0.010]	0.006	[0.004, 0.008]	0.006	[0.004, 0.008]	0.012
0.001	2.0	0.002	[0.000, 0.004][a]	0.001	[0.000, 0.002]	0.002	[0.001, 0.003]	0.003
0.0001	3.0	0	[0, 0.0019][a]	0	[0, 0.0014][a]	0	[0.000, 0.002][a]	0.0003

*k = number of offspring per family. Confidence intervals are two-sided, except when observed number equaled zero; see text.

[a] Exact binomial confidence interval.

Maximizing the Maximum HLOD Over Dominance Model

Phase-unknown families

Table III gives observed significance levels and 95% confidence intervals for all phase-unknown simulations of the "baseline MMLS HLOD" score. For k = 2, the resulting significance levels also closely match those from a one-sided χ_1^2, just as MMLS-C curves without heterogeneity had done [Hodge et al. 1997]. Thus, no additional correction for type I error is needed for maximizing HLOD for k = 2 phase-unknown families, although again there is very little information for heterogeneity in these small phase-unknown families. (In fact, the "baseline" MMLS HLOD scores are slightly conservative here, as were MMLS-C scores without heterogeneity in our earlier work [Hodge et al., 1997].) We observe higher significance levels as family size increases (Table III), though still well below the equally weighted mixture distribution ξ.

Phase-known families

Finally, we considered baseline MMLS HLODs for phase-known families. As in Maximizing HLOD Under a Single Model, phase-known families consistently display higher significance levels than phase-unknown families, but again these levels do not exceed those for the ξ mixture distribution. Table IV gives the results and 95% confidence intervals for specified Z values, for *k* = 2, 4, and 6. Thus, for Z values of interest (at least up to ~ Z = 3.0, and

TABLE IV. Maximizing HLODs Over Two Dominance Models (Baseline MMLS HLODs), in Phase-Known Families: Observed Significance Levels (P) and Associated 95% Confidence Intervals*

Test size		k = 2		k = 4		k = 6		*P* value
(one-sided)	Z	*P*	95% CI	*P*	95% CI	*P*	95% CI	for ξ
0.05	0.59	0.053	[0.045, 0.061]	0.068	[0.061, 0.075]	0.063	[0.056, 0.070]	0.089
0.025	0.83	0.026	[0.020, 0.032]	0.034	[0.029, 0.039]	0.033	[0.028, 0.038]	0.050
0.01	1.17	0.011	[0.007, 0.015]	0.011	[0.008, 0.014]	0.015	[0.012, 0.018]	0.022
0.005	1.44	0.005	[0.002, 0.008]	0.006	[0.004, 0.008]	0.007	[0.005, 0.009]	0.012
0.001	2.0	0.001	[0.000, 0.002]	0.002	[0.001, 0.003]	0.003	[0.001, 0.005]	0.003
0.0001	3.0	0	[0, 0.0010][a]	0.001	[0, 0.0018][a]	0	[0, 0.002][a]	0.0003

*k = number of offspring per family. Confidence intervals are two-sided, except when observed number equaled zero; see text.

[a] Exact binomial confidence interval.

presumably beyond), the ξ mixture curve continues to provide a reasonable upper bound for the type I error probabilities. Again, too, the phase-unknown significance levels approach the phase-known ones as *k* increases, as can be seen in Tables III and IV.

DISCUSSION

Summary of Findings

As outlined in the Introduction, several researchers have addressed the testing and detection of genetic heterogeneity, not always correctly. Other researchers have considered the theoretical distribution of a mixture of χ^2 distributions [e.g., Faraway, 1993; Chernoff and Lander, 1995; Chiano and Yates, 1995; Liu and Shao, 1999]. However, to our knowledge no one has determined the actual magnitude of type I error of the HLOD in realistic genetic situations, or examined the distinction between phase-unknown and -known families. Nor has anyone considered the effect of testing for heterogeneity when simultaneously maximizing the LOD scores over two models. In this study, we explored the distribution of the maximum HLOD under the H_0 of no linkage, for all these circumstances.

In Maximizing HLOD Under a Single Model, we assumed a single genetic model (the correct one) and maximized the HLOD over θ and α. For phase-unknown families of sibship size k = 2 or 3, *P* values are not increased at all for the HLOD compare to the LOD (as proven in the Appendix). For larger *k*, the simulations reveal that *P* values for the max HLOD score are slightly higher than for the max LOD score, increasing as *k* increases. For the same-size phase-known families, the increases in the max HLOD are somewhat higher than for phase-unknown families. Moreover, as *k* increases, phase-unknown families approach phase-known ones in informativeness, so it is not surprising that the significance levels for the former approach those for the latter. In addition, the differences between the significance levels for k = 4 and k = 6 were almost negligible. We would expect the difference between significance levels to converge to zero as $k \to \infty$. Therefore, even for larger families, with k > 6, the significance levels for families with k = 6 probably represent close to the upper limit for both phase-unknown and phase-known families. In any case, since even the k = 6 phase-known families do not actually reach the equally weighted mixture distribution $\xi = (½)\chi_1^2 + (½)\chi_2^2$, it seems reasonable to take that mixture distribution as a practical upper limit.

In Maximizing the Maximum HLOD Over Dominance Model—maximizing HLOD over dominance models (dominant vs. recessive), as well as over θ and α, and subtracting 0.3 from that maximum—similar patterns emerge. Significance levels for the phase-unknown case are increased very little in the smaller families, increasing as family size increases, and approaching significance levels for phase-known families, but not exceeding ξ. For phase-known families, significance levels increase as *k* increases, but again not exceeding the ξ mixture distribution.

The significance levels from our simulations for phase-unknown and phase-known matings compare quite well with the theoretical results of Liu and Shao [1999], who examined k = 2 and k = 3. Our results are also consistent with those of Chernoff and Lander [1995] for phase-known families with k = 2. Our demonstration that *P* values are higher in phase-known families than in phase-unknown ones is also in agreement with Liu and Shao's [1999] theoretical findings. Furthermore, we have shown that as *k* increases, the discrepancy between the significance levels for phase-unknown and phase-known families narrows. This suggests that corrections for large phase-unknown nuclear families should be similar to those for phase-known ones.

Recently, Whittemore and Halpern [2001] suggested that investigators should not use HLODs to detect linkage in the presence of heterogeneity, when certain assumptions of the HLODs are violated. However, they do not cite any evidence to support this recommendation, and in fact there is a fair body of published work supporting the use of HLODs to detect linkage, even when the assumed heterogeneity model is incorrect [Goldin, 1992; Durner et al., 1992; Goldin and Weeks, 1993; Abreu et al., 1999]. Recently, Huang and Vieland [2001] and Vieland and Logue [in press] have provided further evidence; also see Huang and Vieland [in press]. Certainly we agree with Whittemore and Halpern that *estimates* of the admixture parameter α (their p) are unreliable when the genetic models are unknown, but evidence indicates that the HLOD can be very useful for *detecting* linkage in the presence of heterogeneity. Now, we have also demonstrated that one does not pay much of a price in type I error by using HLODs to detect linkage.

Simulations

We make three comments about the simulations. (1) Number of replicates: We arrived at what we believed was an optimum number of replicates by plotting the distribution for 1,000–5,000 replicates. We observed that the results after about 3,000 replicates changed very little. This led us to conclude that 3,000 replicates would be sufficient to make inferences about the asymptotic distribution of HLODs. We did perform more replicates (5,000) for the k = 4 and k = 6 case, because we were interested in comparing our results to the study by Chiano and Yates [1995], which was based on 5,000 replicates. (2) Choice of genetic model: We considered only dominant with full penetrance because we were interested in simulating phase-known families as well as phase-unknown families. Since the data were simulated under no linkage, there is no mode-of-inheritance information in the linkage data [Williamson and Amos, 1990], and therefore, *significance levels* (as opposed to power) should not be affected by the choice of the generating model (also see Hodge et al., [1997]). (3) Mixture of family sizes and markers: We did not consider mixtures of different family sizes or of markers with different levels of informativity because we are setting baselines in this study.

Guidelines and Recommendations

So far, we have presented our results in terms of χ^2 distributions. However, one may also ask the practical question: How much would an investigator need to add to the Z value (in LOD score units) used as a test criterion, in order to achieve a desired test size? To answer this, we took the mixture distribution $\xi = (½)\chi^2_1 + (½)\chi^2_2$ as a *conservative* upper bound for all the situations considered here, and used this mixture distribution to derive guidelines. Since the mixture distribution does not appear in standard statistical tables, we give the correct Z cutoff values for several standard test sizes in Table V, and the corresponding values of Δ, the increase. Thus, for example, for a type I error of 0.05, an investigator testing simultaneously for linkage and heterogeneity should increase the LOD score test criterion by $\Delta = 0.24$; for a test size of 0.001, one should increase it by $\Delta = 0.41$; and so on. Recall that for the MMLS HLOD situation, where one is maximizing over two dominance models, one would already have increased the test criterion by 0.3, as discussed above; these values of Δ are *in addition to* that. For example, for a test size of 0.001, one would increase the simple LOD score test criterion by 0.30 + 0.41 = 0.71 when maximizing over θ, α, *and* dominance model.

Finally, what should an investigator do with a real dataset? Actual common-disease data sets often contain a mixture of different-size families and different amounts of phase infor-

TABLE V. New Values of Z to Be Used as Test Criterion, and Increase (Δ) in Z, for Given Test Sizes

	Corresponding Z (lod)		
Test size (one-sided)	χ_1^2	$\xi = (½)\ \chi_1^2 + (½)\ \chi_2^2$	Δ
0.05	0.59	0.83	0.24
0.025	0.83	1.12	0.29
0.01	1.17	1.50	0.33
0.005	1.44	1.80	0.36
0.001	2.07[a]	2.48	0.41
0.0005	2.35	2.78	0.43
0.0001	3.00	3.47	0.47

[a] For most purposes, investigators use Z ≈ 2.00 for the 0.001 test size, but for the purposes of this table we use the more precise value Z = 2.07.

mation in the families. At one extreme, if the data consist exclusively or mainly of small, phase-unknown nuclear families, then it is hardly necessary to increase the LOD test criterion at all (Tables I and III). However, one should also realize that very small, phase-unknown families do not contain much additional information for heterogeneity over linkage, in the first place. At the other extreme, if most of the families are large and phase-known, one should use the correction based on ξ, remembering that even in this case, that correction is conservative. In general, we recommend that the investigator examine his/her data set and apply a correction that is appropriate to the proportion of sibship sizes, informativeness, and phase information in the data. In studies of common disease, there is often limited phase information in families, particularly if the data consist of nuclear families or affected sib pairs. In these cases, one could afford to be even less conservative. In short, we cannot give hard-and-fast rules for real datasets of mixed family types but suggest that investigators look at their data and use common sense.

Conclusions

In conclusion, we have quantified the magnitude of the increase in the significance levels when using the HLOD statistic, and we took a practical approach, similar to the approach we used in Hodge et al. [1997]. Taking the equally weighted mixture of the two chi-square distributions, ξ, as the working approximation, the increase needed for some values of the Z cutoff ranges from 0.24 for a test size of 0.05, to 0.47 for a test size of 0.0001 (see Table V).

Since one is maximizing the log-likelihood with respect to two parameters, α and θ, it might appear that one is introducing another "whole" degree of freedom. This is in fact what several authors have assumed. However, in a two-point linkage analysis, the estimates of θ and α are highly correlated, and H_0 is "degenerate," as outlined in the Introduction. Our simulations confirm that the additional parameter a does not represent a whole degree of freedom. Rather, it appears to correspond more closely to at most "half" a degree of freedom, i.e., our ξ.

We believe that the corrections suggested here (Table V) are reasonable; and we advocate, as a general scientific philosophy, that one should not be *too* statistically "conservative." As discussed above, these corrections are already on the conservative side, since in most cases the simulated *P* values did not in fact reach the ξ distribution; and they are more conservative than those suggested by Nyholt [2000].

The critical values that we suggest are considerably lower than the critical value of 3.70 for maximizing the LOD over θ and α for a single genetic model as suggested by Risch [1989]. By our guidelines, the usual critical value of 3.00 should be increased, at a *maximum*, only to 3.47, when a single genetic model is used. Only when one simultaneously maximizes over θ and α *and* over two genetic models should one then increase the critical value of 3.00 to 3.77 (results from adding first 0.30, then 0.47).

We have noted here that no additional correction for HLOD over the simple LOD is needed for small phase-unknown families. That is, there is no additional information reflecting heterogeneity in those families beyond the information for recombination fraction. However, that is not necessarily the case for multipoint analysis, in which information is available from adjacent markers. In such cases, small phase-known families probably do provide information for heterogeneity, although we did not examine the multipoint case in this study. Recent work, however, does provide some information with regard to the multipoint case. Greenberg and Abreu [2001] showed that, in a system consisting of 10 markers with a trait locus not linked to any of them, the multipoint HLOD critical value should be increased by one full LOD unit to be comparable to a LOD for the specific 10-marker system that these authors investigated. As for the broader question of appropriate type I error levels for genome screens, we believe the same kinds of considerations would apply here as for any other types of genome screens.

Our work included Z values up to around 3.0. However, if Z is much higher than 3.0, indicating strong evidence of linkage, then the linkage results are of great interest in any case, and the question of correction becomes moot. We do not advocate using *P* values as measures of “strength of evidence” [Vieland and Hodge, 1998], but only as indicators of the probability of type I error. Therefore, we are not particularly concerned with distinctions between *P* values of, say, 0.00001 and 0.000001.

Finally, given that family data are so difficult, time-consuming, and expensive to collect, it behooves us to extract the maximum amount of information from them. Using overly conservative guidelines may guard one from accepting the alternative hypothesis too readily but may also make it more difficult to learn anything from the data. The danger of false negatives (failing to find a true linkage that is there) must be considered, as well as the danger of false positives. We have shown here that testing for admixture heterogeneity with HLOD scores inflates type I error less than other authors have claimed, and we propose that investigators using HLODs be cognizant of the less conservative guidelines presented here.

ACKNOWLEDGMENTS

We thank Dr. Shao for helpful discussions.

REFERENCES

Abreu PC, Greenberg DA, Hodge SE. 1999. Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases. Am J Hum Genet 65:847–57.

Chernoff H, Lander E. 1995. Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. J Stat Plan Infer 43:19–40.

Chiano MN, Yates JRW. 1995. Linkage detection under heterogeneity and the mixture problem. Ann Hum Genet 59:83–95.

Davies RB. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64:247–54.

Davies RB. 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 74:33–43.

Durner M, Greenberg DA, Hodge SE. 1992. Inter- and intrafamilial heterogeneity: effective strategies and comparison of analysis methods. Am J Hum Genet 64:859–70.

Faraway JJ. 1993. Distribution of the admixture test for the detection of linkage under heterogeneity. Genet Epidemiol 10:75–83.

Goldin LR. 1992. Detection of linkage under heterogeneity: comparison of the two-locus vs. admixture models. Genet Epidemiol 9:61–6.

Goldin LR, Weeks DE. 1993. Two-locus models of disease: comparison of likelihood and nonparametric linkage methods. Am J Hum Genet 53:908–15.

Greenberg DA, Abreu PC. 2001. Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. Genet Epidemiol 21:299–314.

Greenberg DA, Abreu PC, Hodge SE. 1998. The power to detect linkage in complex disease using simple genetic models. Am J Hum Genet 63:870–9.

Hodge SE, Abreu PC, Greenberg DA. 1997. Magnitude of Type I error when single-locus linkage analysis is maximized over models: a simulation study. Am J Hum Genet 60:217–27.

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 35:1139–55.

Huang J, Vieland VJ. 2001. Comparison of “model-free” and “model-based” linkage statistics in the presence of locus heterogeneity: single data set and multiple data set applications. Hum Hered 51:217–25.

Huang J, Vieland VJ. The null distribution of the heterogeneity lod score does depend on the assumed genetic model for the trait. Hum Hered (in press).

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

Lindsay BG. 1995. Mixture Models: theory, geometry and applications. NSF-CBMS Regional Conference Series in Probability and Statistics. Vol. 5. Hayward, CA: Institute for Mathematical Statistics.

Liu X, Shao Y. 1999. Asymptotic distribution of the likelihood ratio test that a mixture of binomials is a single binomial. Technical Report, Department of Statistics, Columbia University, New York.

Nyholt DR. 2000. All LODs are not created equal. Am J Hum Genet 67:282–8.

Ott J. 1991. Analysis of human genetic linkage, 2nd ed. Baltimore: Johns Hopkins University Press.

Ott J. 1983. Linkage analysis and family classification under heterogeneity. Ann Hum Genet 47:311–20.

Risch N. 1989. Linkage detection tests under heterogeneity. Genet Epidemiol 6:473–80.

Smith CAB. 1963. Testing for heterogeneity of recombination values in human genetics. Biometrics 49:151–61.

Vieland VJ, Hodge SE. 1998. Review of statistical evidence: a likelihood paradigm, by R. Royall. Am J Hum Genet 63:283–9.

Vieland VJ, Logue M. 2001. HLODs, trait models, and ascertainment: Implications of admixture for parameter estimation and linkage detection. Hum Hered (in press).

Wald A. 1949. Note on the consistency of the maximum likelihood estimate. Ann Math Stat 20:595–601.

Whittemore AS, Halpern J. 2001. Problems in the definition, interpretation, and evaluation of genetic heterogeneity. Am J Hum Genet 68:457–65.

Williamson JA, Amos CI. 1990. On the asymptotic behavior of the estimate of the recombination reaction under the null hypothesis of no linkage when the model is misspecified. Genet Epidemiol 7:309–18.

APPENDIX

Analytical Results for Small Nuclear Phase-Unknown Families

In this Appendix, we demonstrate that for two- and three-child phase-unknown families, the significance levels of HLODs are identical to those for “regular” LOD scores (i.e., LOD scores that do not allow for heterogeneity). These demonstrations complement our simulations.

Recall that in this study, the disease is assumed to be fully penetrant autosomal dominant). Let the affected (informative) parent have marker alleles 1 and 2. We arbitrarily define a child as being in “concordance class 1” if s/he is either affected and received the 1 allele from the affected parent *or* unaffected and received the 2 allele, whereas a child is in “concordance class 2” if s/he is affected and received the 2, or is unaffected and received the 1 allele.

Two-child families

There are only two possible outcomes for the pair of children: Either both children are in the same concordance class (outcome 1), or one child is in one concordance class and the other child is in the other (outcome 2). The probabilities of these two outcomes for regular LOD scores (without heterogeneity) and for HLODs (admixture model) are shown in Table VIa.

In a sample of n phase-unknown two-child families, let n_1 represent the number of families displaying outcome 1, and n_2, the number displaying outcome 2. For the regular LOD score, there is one unknown parameter (θ) and one degree of freedom, so that the maximized log likelihood ratio is simply

$$\max \log LR = n_1 \log (n_1/n) + n_2 \log (n_2/n) - n \log (½) \qquad \text{(A.1)}$$

For the heterogeneity LOD score (HLOD), there is still one d.f., but two unknown parameters (θ and α). Therefore, it is true that for *estimation*, those two parameters would be confounded. However, the *likelihood* for the HLOD is still given by equation (A.1). Thus, for these families, the maximized LRs are identical, and therefore, so are the LOD and HLOD scores. Therefore, their significance levels are also identical. (That is, they are not merely asymptotically equivalent but are actually identical in any sample, large or small.)

Three-child families

The situation is similar for three-child phase-unknown families, although the formulas differ from those for the two-child families. Define the same two concordance classes as before. Now the two possible outcomes for a family are that either all three children are in the same concordance class, or two children are in one concordance class and the third child is in the other class. Table VIb shows the probabilities for the two outcomes.

Again let n_1 represent the number of families displaying outcome 1, and n_2, the number displaying outcome 2, in a sample of n families. As with the two-child families, for the regular LOD score, there is one unknown parameter (θ) and one degree of freedom. The maximized likelihood ratio is

$$\max \log LR = n_1 \log (n_1/n) + n_2 \log (n_2/n) - [n_1 \log(¼\ ) + n_2 \log(¾)] \qquad \text{(A.2)}$$

For the HLOD there is also one d.f., but two unknown parameters (θ and α). The maximized LR is given by (A.2), and, therefore, the LOD and HLOD scores are identical in any sample, and so are the exact significance levels.

TABLE VI. Possible Outcomes and Their Probabilities, for Two- (k = 2) and Three-Child (k = 3) Phase-Unknown Families (in Appendix)

		Probability[a]	
Outcome	Description	Lods	Hlods
(a) k = 2:			
1	Both children in same concordance class	ψ	$\alpha\psi + (1-\alpha)\cdot½$
2	Children in two different concordance classes	$1-\psi$	$1-[\alpha\psi + (1-\alpha)\cdot½]$
(b) k = 3:			
1	All three children in same concordance class	ϕ	$\alpha\phi + (1-\alpha)\cdot¼$
2	Two children in one concordance class, third child in other concordance class	$1-\phi$	$1-[\alpha\phi + (1-\alpha)\cdot¼]$

[a] Define $\psi \equiv \theta^2 + (1-\theta)^2$. Define $\phi \equiv \theta^3 + (1-\theta)^3$.