# Distribution of the Admixture Test for the Detection of Linkage Under Heterogeneity

**Julian J. Faraway**

*Department of Statistics, University of Michigan, Ann Arbor, Michigan*

The admixture test for the detection of linkage under heterogeneity is considered. We show that the null distribution of this test statistic has half its weight concentrated on zero and the other half on a complicated distribution that can be approximated by max $(X_1,X_2)$ where $X_1$ and $X_2$ are independent $\chi_1^2$ variables. We also investigate the stability of the size of the test for small samples. The power of this test to detect linkage, when heterogeneity is present, can be substantially greater than the standard test that assumes homogeneity. Even when heterogeneity is not present, the test is only slightly less powerful than the homogeneous test. This would suggest the use of the admixture test in preference to the homogeneous test if the presence of heterogeneity is at all suspected. © 1993 Wiley-Liss, Inc.

**Key words: linkage heterogeneity, lod scores**

## INTRODUCTION

We consider the detection of linkage when linkage heterogeneity exists, that is when only a fraction of sibships may be linked to a given genetic marker. Smith [1963] introduced the admixture model based on the recombination fraction and the proportion of linked families. Ott [1983, 1985] and Risch [1988] consider tests for *heterogeneity* based on this model, whereas Hodge et al. [1983] and Risch [1989] consider tests for *linkage* based on this same model. The latter is discussed here. Martinez and Goldin [1989] discuss sample sizes needed for such tests.

The test for linkage is one-sided since recombination fractions greater than one half make no biological sense and should its estimated value be greater than one half, one would not take this as evidence of linkage. Hence, the true null distribution of the admixture statistic has half its weight concentrated at zero and the other half on some other distribution which is the subject of our interest here. Because of the symmetry of

Received for publication January 29, 1992; revision accepted October 22, 1992.

Address reprint requests to Julian J. Faraway, Department of Statistics, University of Michigan, Ann Arbor, MI 48109.

the problem, it's convenient and notationally simpler to just compute the null distribution for the two-sided test statistic to discover the aforementioned distribution. Bear in mind that although we shall be concerned with the two-sided test statistic in what follows, the true null distribution is as above.

Hodge claimed that the asymptotic (as the number of sibships becomes large) null distribution of the (two-sided) admixture test statistic was $\chi_1^2$ but Risch conjectured it was $\chi_2^2$. We claim here that neither is correct and that the true asymptotic distribution is quite complicated but can be adequately approximated by the max $(X_1,X_2)$ where $X_1$ and $X_2$ are independent $\chi_1^2$ variables. This distribution lies somewhere between the two previous claims and thus this result is of more than just technical interest given the popularity of the test. Ghosh and Sen [1986] study the asymptotic distribution of the likelihood ratio test statistic for a mixture model that is similar to the one here and obtained a result similar in form to ours.

## DISTRIBUTION OF THE TEST STATISTIC

Let the recombination fraction be $\theta$, the proportion of linked sibships be $\alpha$ and the sibship size be $s$. Let the number of sibships be $n$ and let $X_i$ be the number of recombinant gametes out of $s$ for sibship $i$. For now, let the phase be known.

Thus the likelihood for this set of sibships would be

$$L(\theta,\alpha) = \prod_{i=1}^{n} [\alpha\theta^{X_i}(1-\theta)^{s-X_i} + (1-\alpha)(1/2)^s].$$

Note that if we map $X_i \rightarrow s - X_i$ (producing an outcome that has equal probability under the hypothesis of no linkage) and $\theta \rightarrow 1 - \theta$ then the likelihood stays the same. This symmetry allows us to consider the two-sided test statistic in our computation of the actual one-sided admixture test. If we wish to test for linkage, the natural null and alternative hypotheses are

$$H_0 : \theta = 1/2 \quad H_A : \theta < 1/2$$

and the maximum likelihood-ratio test statistic is

$$T = 2\log [\mathrm{L}(\hat{\theta},\hat{\alpha})/L(1/2,\tilde{\alpha})]$$

where $\hat{\theta}$ and $\hat{\alpha}$ are the maximum likelihood estimates (mle) under the alternative hypothesis and $\tilde{\alpha}$ is the mle under the null. Note that when $\theta = 1/2$, $\alpha$ is unidentifiable, i.e., any value of $\alpha$ produces the same likelihood so the actual value of $\tilde{\alpha}$ is immaterial, although this unidentifiability is the source of the difficulty in determining the distribution of $T$. We use natural logs here for statistical convenience; lod scores will be discussed later. So

$$T = 2\max_{\alpha,\theta} T(\alpha,\theta) = 2\max_{\alpha,\theta} \sum_{i=1}^{n} \log\{\alpha[2^s\theta^{X_i}(1-\theta)^{s-X_i} - 1] + 1\}$$

where $0 \leqslant \alpha \leqslant 1$, $0 \leqslant \theta \leqslant 1$.

Unfortunately, the asymptotic distribution under the null is not simply $\chi_1^2$ as it would be if the usual theory were applicable. This is because a regularity condition regarding the identifiability of the parameters is not satisfied; see Wald [1949]. This means the asymptotic distribution of $T$ must be derived by other means. We give a heuristic justification of our result and verify it by simulation.

Since $\alpha$ is unidentifiable at the null, the likelihood will be rather flat in the $\alpha$ direction and since the range of $\alpha$ is restricted, the value of $\alpha$ maximizing $T(\alpha,\theta)$ will tend to occur at the boundary of the range of $\alpha$ for large $n$. To see this, expand $T(\alpha,\theta)$ in $\theta$ about 1/2, with $\alpha$ bounded away from 0,

$$T(\alpha,\theta) \approx -8\alpha \sum_{i=1}^{n} (X_i - s/2)(\theta - \tfrac{1}{2}) + 4\alpha[(1-\alpha) \sum_{i=1}^{n} (2X_i - s)^2 - ns](\theta - \tfrac{1}{2})^2$$

(where $\approx$ means approximately). Maximizing over $\theta$ gives

$$\max_{\theta} T(\alpha,\theta) \approx \frac{4\alpha[\Sigma_{i=1}^{n}(X_i - s/2)]^2}{ns - 4(1-\alpha)\Sigma_{i=1}^{n}(X_i - s/2)^2}. \tag{1}$$

Let $Z = \Sigma_{i=1}^{n} (X_i - s/2)$ and $S^2 = \Sigma_{i=1}^{n}(X_i - s/2)^2$ and now differentiating with respect to $\alpha$

$$\frac{d}{d\alpha} \max_{\theta} T(\alpha,\theta) \approx \frac{-4Z^2(4S^2 - ns)}{[ns - 4S^2(1-a)]^2}$$

which will be positive or negative depending on whether $S^2$ is less or more than $ns/4$, independent of the value of $\alpha$ so for $n$ sufficiently large $\max_{\theta} T(\alpha,\theta)$ will be maximized at $\alpha = 1$ or for $\alpha$ small ($T = 0$ when $\alpha = 0$). Since $S^2 \rightarrow ns/4$ as $n \rightarrow \infty$, both cases will be roughly equally likely. So we consider the distribution of $\max_{\theta} T(\alpha,\theta)$ for $\alpha = 1$ and for $\alpha$ small.

When $\alpha = 1$, $\max_{\theta} T(1,\theta) \approx (4/ns)[\Sigma_{i=1}^{n}(X_i - s/2)]^2$ using Eq. (1). Since $EX_i = s/2$ and $\mathrm{Var}\, X_i = s/4$, $\max_{\theta} T(1,\theta)$ is asymptotically $\chi_1^2$, just applying the central limit theorem.

However, when $\alpha$ is small the distribution of $T$ is not so clear.

Write $k_j = \{\text{number of } X_i = j\}$ for $j = 0,1,\ldots s$ then

$$T = 2\max_{\alpha,\theta} \sum_{i=0}^{s} k_i \log\{\alpha[2^s\theta^i(1-\theta)^{s-i} - 1] + 1\}.$$

Now since $\alpha$ is small, we can expand log in terms of $\alpha$ [$\log(1+x) \approx x - x^2/2$]:

$$T \approx 2\max_{\alpha,\theta} \sum_{i=0}^{s} k_i\{\alpha[2^s\theta^i(1-\theta)^{s-i} - 1] - \tfrac{1}{2}\alpha^2[2^s\theta^i(1-\theta)^{s-i} - 1]^2\}$$

and maximizing this over $\alpha$ gives

$$T \approx \max \frac{\{\Sigma_{i=0}^{s} k_i[2^s\theta^i(1-\theta)^{s-i}-1]\}^2}{\Sigma_{i=0}^{s} k_i[2^s\theta^i(1-\theta)^{s-i}-1]^2}.$$

Now under the null $\theta = 1/2$, and so $Ek_i = n\binom{s}{i}2^{-s}$ so replacing $k_i$ by its expectation and then by applying the binomial theorem, we see that the denominator is approximately

$$\sum_{i=0}^{n} n\binom{s}{i}2^{-s}[2^s\theta^i(1-\theta)^{s-i}-1]^2 = n\{[\theta^2+(1-\theta)^2]^s2^s-1\}.$$

Hence

$$T \approx \max_{\theta}\left[\sum_{i=0}^{s} k_ic_i(\theta)\right]^2$$

where

$$c_i(\theta) = \frac{[2^s\theta^i(1-\theta)^{s-i}-1]}{\sqrt{n\{[\theta^2+(1-\theta)^2]^s2^s-1\}}}.$$

The distribution of this cannot be explicitly stated for general $s$, but given that $k_i$ is asymptotically normal as $n \to \infty$, $\Sigma_{i=0}^{s} k_ic_i(\theta)$ is asymptotically a weighted sum of normals and is hence normal for given $\theta$. This might suggest a $\chi_1^2$ as a possible approximation and simulation shows that this is indeed a good fit but it should be emphasized that this is not the exact distribution.

Now when $T$ is maximized for $\alpha$ small, $T$ is a weighted sum of the $k_i$ and when maximized for $\alpha = 1$, $T$ is a function of the sample mean, so the maximizing values at these two points will tend to be independent especially for large $s$. This suggests a distribution for $T$ as the maximum of two independently distributed $\chi_1^2$ variables. Again, this is not an exact result but simulation indicates that it is a good approximation. The true asymptotic distribution a function of the maximum of a particular Gaussian process but since this cannot be explicitly calculated, the suggested approximation will be of more practical utility.

To check the validity of this suggested approximating distribution, the exact distribution has been calculated, by computing $T$ for all possible data, for several values of $s$ and $n$. The likelihood was maximized by first transforming $\alpha$ and $\theta$ to a logit scale $\{x \to \log[x/(1-x)]\}$ so that the constraints on $\alpha$ and $\theta$ can be removed and then using the Nelder–Mead simplex method described in Press et al. [1988] to find the maximum. The maximum at $\alpha = 1$ and $\alpha$ small as indicated in the discussion above was also calculated.

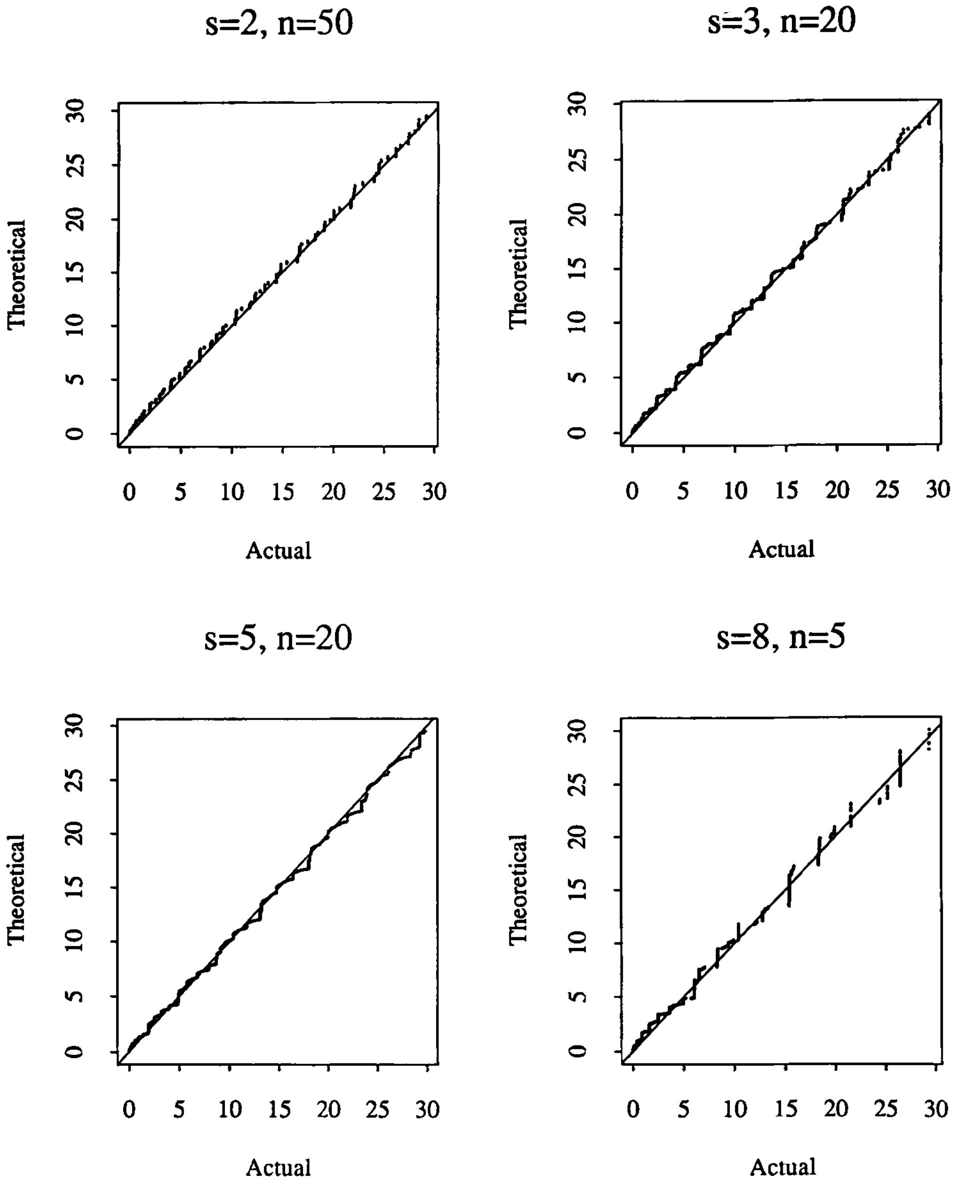

Fig. 1. Quantile–quantile plots of the actual distribution of $T$ against its supposed asymptotic distribution for family sizes $s$ and number of families $n$.

The quantiles of the suggested distribution of $T$ may be simply calculated by noting that

$$P(T < q) = P(\chi_1^2 < q)^2.$$

In Figure 1, we show quantile–quantile plots of the actual test-statistic against its claimed distribution.

The agreement between the simulated and theoretical distributions is good. Since for finite $n$, $T$ is discrete, we cannot expect an exact match. As $n$ increases, the fit will improve.

## CRITICAL VALUES AND SIZE

Recall that if $\hat{\theta} \geqslant 0.5$ we have no evidence for linkage, otherwise we can determine the significance of the observed $T$ by referring to the approximate null distribution that we have calculated. If lod scores are preferred, one would use

$$T' = 2\log(10)T.$$

If the same level of test is desired as for a $\chi_1^2$ distributed statistic, lod scores of 2, 3, and 5 correspond to scores for $T'$ of 2.28, 3.28, and 5.27, respectively. (Compare the values given by Risch [1989] of 2.62, 3.70, and 5.80, respectively.)

This result is asymptotic in nature and may not necessarily be good for the small samples used in practice. It is computationally feasible to calculate the exact significance level for small samples and this is to be preferred. However, if the asymptotic critical values are to be used then the stability of the size of the test is important. In Figure 2, we show the size of the heterogeneous test and, by way of comparison, the size of the standard test where homogeneity is assumed. Critical values of 3.28 and 3 were used, respectively, which both correspond to a nominal significance level of 0.020166%. As can be seen, the true size of the test can vary somewhat although the variation is about the same for the heterogeneous test as the homogeneous test.

## POWER

Risch [1989] compared the power of the heterogeneous test against the standard test where homogeneity is assumed ($\alpha = 1$), and concluded that the homogeneous test was more powerful in most circumstances. Contrary to this, we demonstrate here, by using the correct critical value and computing the power exactly, that the heterogeneous test is generally preferable.

Exact critical values for a significance level of 0.020166% (which corresponds to the asymptotic level of lod 3 for the homogeneous test) for both tests were computed and the exact power to detect linkage was calculated for a range of values of $\alpha$ from 0 to 1 and of $\theta$ from 0 to 1/2. Because $T$ (and the homogeneous test statistic) are discrete for finite $n$, some adjustment is necessary to ensure that the level of significance is exactly obtained. It is possible to choose a $T_c$ and $p_a$ and define the test to reject the null (no linkage) when $T > T_c$ or reject with probability $p_a$ when $T = T_c$ and otherwise accept, so that the significance level is exactly 0.020166%. A similar adjustment is done for the homogeneous test. Of course, one would not do this in practice, but here it is desirable for the purpose of a fair comparison of the two tests.

Figure 3 shows the power of the heterogeneous test minus that of the homogeneous test. The lines show contours of equal difference in power (probability expressed as a percentage) and "=" denotes the region where there is a less than a 0.01 difference in the power. In the case of sibship size 2 and 50 sibships, the homogeneous test exceeds the power of heterogeneous test by no more than 0.01 and can be 0.1 less

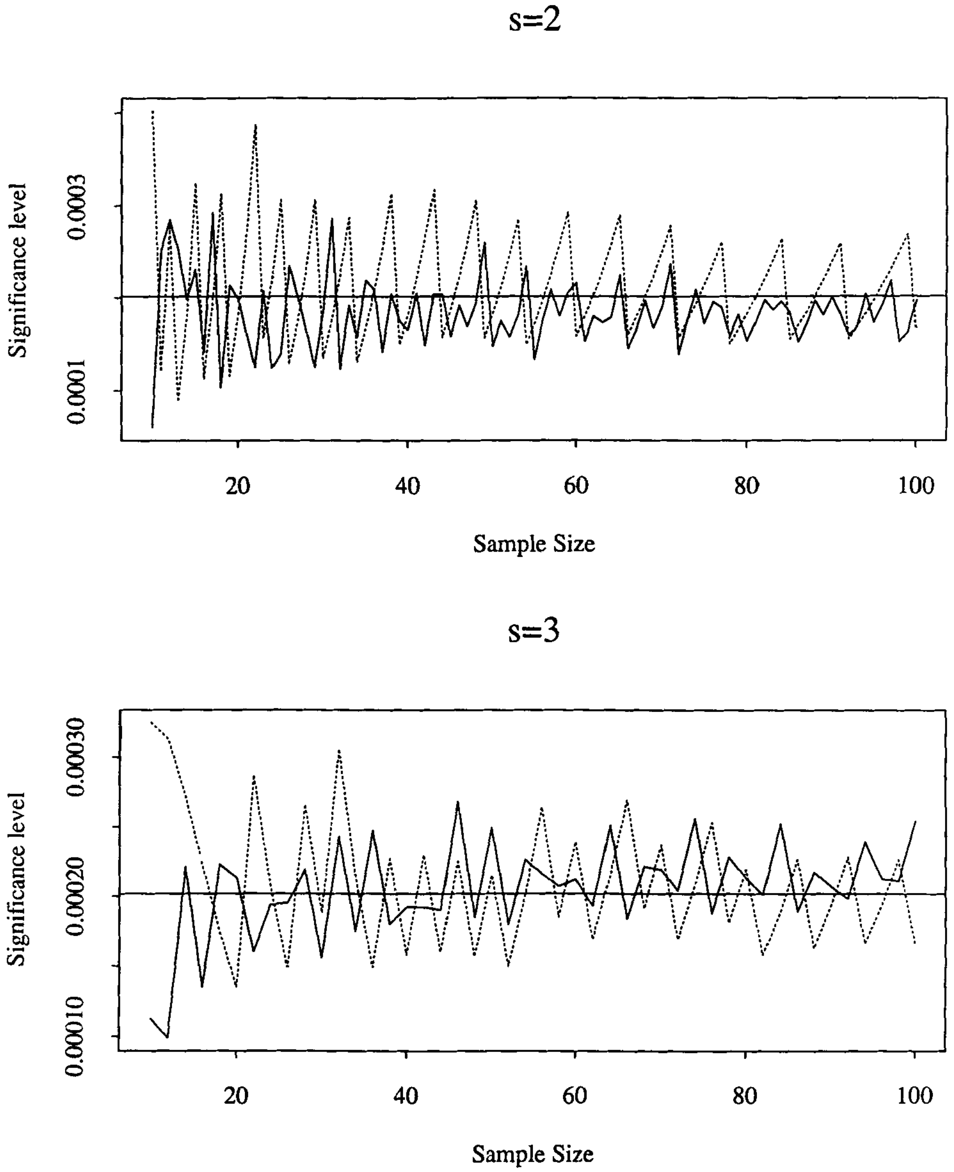

Fig. 2. The true significance level of the heterogeneous test (solid) and the homogeneous test (dashed) corresponding to critical values of 3.28 and 3, respectively.

powerful when $\alpha = 0.4$ and $\theta = 0$. When the sibship size is 5 and with 20 sibships, the heterogeneous test exceeds the power of the homogeneous test by 0.37 when $\alpha = 0.3$ and $\theta = 0$. The region where the homogeneous test is mildly preferable is confined to an area of low mixing and moderate linkage. Other comparisons show that the region where the heterogeneous test is clearly preferable expands with sibship size and number of sibships. Even when there is no mixing the homogeneous test is only mildly more (0.05–0.1 at best) powerful than the heterogeneous test, but if there is some mix-

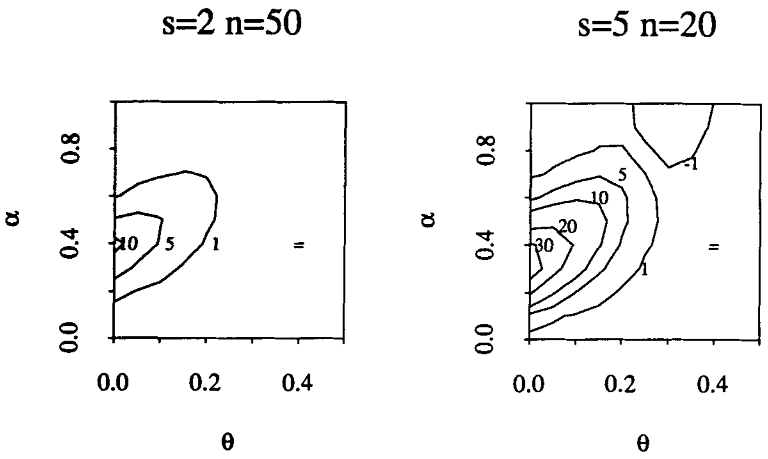

Fig. 3. Contour plots showing the difference in power between the heterogeneous test and the homogeneous test.

ing the heterogeneous test can be substantially more powerful. This leads one to recommend the heterogeneous test in preference to the homogeneous test when heterogeneity is at all suspected.

## DISCUSSION

We have approximated the null distribution of the admixture test for the detection of linkage and demonstrated that if the possibility of heterogeneity exists, this admixture test is generally more powerful than the usual test which takes no account of heterogeneity.

We have considered constant sibship size here for simplicity of the exposition but this is not crucial and the same asymptotic result would follow even if the sibship size were allowed to vary. Furthermore, the same result holds even when the meioses are not completely informative. For the least informative, phase unknown, case, the test statistic is

$$T = \max_{\alpha,\theta} 2 \sum_{i=1}^{n} \log\{\alpha[2^{s-1}[\theta^{X_i}(1-\theta)^{s-X_i} + \theta^{s-X_i}(1-\theta)^{X_i}] - 1] + 1\}$$

and a similar reasoning to the one above may be used to get the same result.

## ACKNOWLEDGMENTS

Thanks to Michael Boehnke of the Department of Biostatistics, University of Michigan for bringing my attention to this problem and offering helpful comments and thanks also to two referees for improving the initial draft.

## REFERENCES

Ghosh JK, Sen PK (1986): On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. "Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer," Volume II. Wadsworth.

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.

Martinez M, Goldin L (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.

Ott J (1983): Linkage analysis and family classification under heterogeneity. Ann Hum Genet 47:80–96.

Ott J (1985): "Analysis of Human Genetic Linkage." Baltimore: The Johns Hopkins University Press.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1988): "Numerical Recipes." Cambridge: Cambridge University Press.

Risch N (1988): A new statistical test for linkage heterogeneity. Am J Hum Genet 42:353–364.

Risch N (1989): Linkage detection tests under heterogeneity. Genet Epidemiol 6:473–480.

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

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