Published in final edited form as:

Genet Epidemiol. 2010 December ; 34(8): 912–916. doi:10.1002/gepi.20535.

Distribution of Model-based Multipoint Heterogeneity Lod Scores

Chao Xing^{1,*}, Nathan Morris², and Guan Xing³

¹Department of Clinical Sciences, McDermott Center of Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas

²Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio

³Bristol-Myers Squibb Company, Pennington, New Jersey

Abstract

The distribution of two-point heterogeneity lod scores (HLOD) has been intensively investigated because the conventional χ^2 approximation to the likelihood ratio test is not directly applicable. However, there was no study investigating the distribution of the multipoint HLOD despite its wide application. Here we want to point out that, compared with the two-point HLOD, the multipoint HLOD essentially tests for homogeneity given linkage and follows a relatively simple

limiting distribution $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, which can be obtained by established statistical theory. We further examine the theoretical result by simulation studies.

Keywords

heterogeneity lod score; distribution; multipoint

Locus heterogeneity represents a form of genetic architecture of complex traits where alleles at more than one locus lead to the same phenotype. It adversely affects the power of linkage analysis if the heterogeneous disease genetic background of families is not taken into account. A natural way to model such heterogeneous data is by a mixture model, as first suggested by Smith [1963]. Under the mixture model framework one can either test for homogeneity given linkage [Ott, 1983] or test for linkage allowing for heterogeneity [Hodge et al., 1983] by a likelihood ratio test. The distribution of two-point heterogeneity lod scores (HLOD) has been intensively investigated [Abreu et al., 2002; Chernoff and Lander, 1995; Chiano and Yates, 1995; Faraway, 1993; Huang and Vieland, 2001; Lemdani and Pons, 1995; Liang and Rathouz, 1999] because the conventional χ^2 approximation to the likelihood ratio test is not directly applicable [Davies, 1977, 1987]. However, to our surprise, there was no study investigating the distribution of the multipoint HLOD despite its wide application. The multipoint HLOD is reported by popular software packages such as GENEHUNTER [Kruglyak et al., 1996] and MERLIN [Abecasis et al., 2002] without a *P*-value accompanying it. Here we want to point out that, compared with the two-point HLOD, the multipoint HLOD essentially tests for homogeneity given linkage and follows a

relatively simple limiting distribution, which can be obtained by established statistical theory. We further examine the theoretical result by simulation studies.

Denote by M the genotype data, by D the phenotype data, by α the admixture parameter, which indicates the proportion of families linked to the locus tested, and by x the map position of a putative disease locus. A general format of the likelihood for the i th family is defined as, $L_i(M, D; x, \alpha) = \alpha L_i(M, D; x) + (1 - \alpha)L_i(M, D; x = \infty)$, and the HLOD is defined as $HLOD = \log_{10}[\sup L(\alpha, x)/L_0(\alpha = 1; x = \infty)]$, where the subscript 0 denotes the null hypothesis of no linkage. In two-point analysis, x is parameterized as the recombination fraction, commonly denoted as $\theta \in [0, 0.5]$, between the marker and disease loci by a map function. The two-point analysis tests the hypotheses $H_0: \theta = 0.5$ versus $H_1: \theta < 0.5$. The likelihood ratio test statistic is $T_{\text{two-point}} = 2 \ln[L(\hat{\alpha}, \hat{\theta})/L(\theta = 0.5)]$. Note that when $\theta = 0.5$, α is unidentifiable, thus the conventional χ^2 approximation to the likelihood ratio test is not directly applicable [Davies, 1977, 1987], which causes difficulties in determining the distribution of $T_{\text{two-point}}$. In multipoint analysis, under the null hypothesis $x = \infty$, which is equivalent to $\theta = 0.5$ in the two-point analysis by a map function; however, the alternative hypothesis is that the putative disease locus is at a specific location c , i.e., x is not a free parameter. The likelihood ratio test statistic is expressed as $T_{\text{multipoint}} = 2 \ln[L(\hat{\alpha}, x = c)/L(\alpha = 1; x = \infty)]$, which indicates the multipoint analysis tests the hypotheses $H_0: \alpha = 0$ versus $H_1: \alpha > 0$. Thus, the multipoint HLOD essentially tests for homogeneity given linkage, in contrast to the two-point HLOD testing for linkage allowing for heterogeneity. The multipoint HLOD test statistic corresponds to *case 5* of Self and Liang [1987] with one parameter on the boundary of parameter space, and also corresponds to *example 11* of Lindsay [1995] that tests $H_0: \pi = 0$ versus $H_1: \pi > 0$ in the mixture model $(1 - \pi)f + \pi g$, where f and g are both known. Thus, following the theoretical arguments by both authors, the

asymptotic distribution of $T_{\text{multipoint}}$ is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, where χ_0^2 denotes a distribution degenerate at zero with probability one. It is of interest to note that, given linkage, the two-point homogeneity test statistic $2 \ln[L(\hat{\alpha}, \hat{\theta})/L(\alpha = 1, \hat{\theta})]$ follows the same asymptotic distribution as $T_{\text{multipoint}}$ [Ott, 1999].

Below we examine the theoretical result by simulation studies under varying simulation models, analysis models, and sample sizes. We simulated nuclear families consisting of two parents and four children, and two linked markers 10 cM apart with four alleles of equal frequency at each locus. Two randomly chosen children were set to be affected and the other two unaffected. The parental phenotypes were simulated under three models. In model I, one parent was set to be affected and the other unaffected, which mimicked a dominant trait; in model II, both parents were set to be unaffected, which mimicked a recessive trait; and in model III, both parents were set to be unknown, which mimicked a trait with mode of inheritance unclear. We generated 5,000 replicate samples under each of the four sample size scenarios—100, 500, 1,000 and 5,000 families in a sample dataset. Denote the penetrance by f_i , where $i \in \{0, 1, 2\}$ indicates the number of disease predisposing alleles. Always assuming a sporadic rate of 0.01 and a disease predisposing allele frequency of 0.01, we performed model-based multipoint linkage analysis on each dataset under 18 different genetic models: $(f_0, f_1, f_2) = (0.01, 0.1, 0.1), (0.01, 0.2, 0.2), \dots, (0.01, 0.9, 0.9), (0.01, 0.01, 0.1), (0.01, 0.01, 0.2), \dots, (0.01, 0.01, 0.9)$ denoted $D1, D2, \dots, D9, R1, R2, \dots, R9$, respectively, and recorded the multipoint HLOD at the middle of the two markers. The linkage analysis was performed using MERLIN [Abecasis et al., 2002], in which the mixture likelihood was maximized using the Brent's method [2002].

The simulation results confirmed the limiting distribution of $T_{\text{multipoint}} (= 2 \times \ln 10 \times \text{multipoint HLOD})$ to be $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. In Table I we estimated some parameters of the

empirical distribution under analysis models D2, D5, D8, R2, R5, and R8 with varying sample sizes in each simulation scenario. The proportion of $T_{\text{multipoint}}$ equal to zero under different models always approximated to a half. The mean and variance of a random variable of χ_1^2 are 1 and 2, respectively. We observed the mean and variance of non-zero $T_{\text{multipoint}}$ approximated to 1 and 2, respectively, under most dominant models and high penetrance recessive models when the sample size was large. Under each analysis model the mean and variance approached their expectations under the theoretical distribution as the sample size increased. However, both mean and variance were smaller than their expectations under low penetrance recessive models even if the sample size was 5,000. In contrast, the maximum likelihood estimate of mean approximated to 1 when sample size was 500 under any analysis model. Given the sample size and analysis model, the estimated mean of $T_{\text{multipoint}}$ for data generated under simulation model I approximated to 1 closer than that for data generated under models II and III, which was mostly clearly illustrated when analyzing the data by low penetrance recessive models. We calculated the empirical type I error rate of the multipoint HLOD under analysis models D2, D5, D8, R2, R5, and R8 with varying sample sizes in each simulation scenario assuming that $T_{\text{multipoint}}$ follows a

distribution of $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (Table II). The results were consistent with the empirical distribution as summarized in Table I. The dominant analysis models gave proper type I error rate. The recessive models were conservative except for high penetrance models with large sample size. Figure 1 illustrated the empirical distribution of the multipoint HLOD under analysis models D5 and R5 with the theoretical distribution.

The multipoint HLOD method is powerful to detect linkage even when the assumed heterogeneity model is incorrect [Greenberg and Abreu, 2001; Hodge et al., 2002]. However, the distribution of the multipoint HLOD has remained mysterious, possibly because it has been obscured by the complexity of the two-point HLOD. Similarly, the model-based multipoint lod score that is best suited in an evidential paradigm [Hodge et al., 2008] displays some asymptotic complexity and does not have a limiting distribution [Xing and Elston, 2006]. We have shown in this paper that, in contrast with the two-point HLOD and the multipoint lod score, the multipoint HLOD test statistics follows a relatively simple asymptotic distribution. That is, $2 \times \ln 10 \times \text{multipoint HLOD}$ follows an asymptotic

distribution of $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. This not only enables evaluating the significance level easily, but also facilitates further inferences such as multiple testing correction. The rate of convergence to asymptotic distribution depends on the informativeness of both markers and the trait. Given data, the pre-specified analysis model defines the informativeness of the trait. As the model-based linkage statistics generally do, the multipoint HLOD under low penetrance recessive models is conservative because the phenotype contributes little information under such trait models, which is reflected as a high proportion of zeros, low mean and small variance (Table I). Similarly, data simulated under model I contain more trait information than that simulated under models II and III; therefore, the multipoint HLOD under simulation models II and III is more conservative than that under simulation model I. In multipoint analysis, the marker information is relatively constant across a map; thus, the behavior of $T_{\text{multipoint}}$ should also be relatively stable across the map given an analysis model. We note that the proportion of HLODs equal to zero is always greater than, though close to, one half; thus, a nominal P -value is presumably conservative. In this study we did not investigate the performance of the multipoint HLOD test statistic in the extreme tails of its null distribution, which will be crucial in determining genome-wide significance. Considering the efficiency of the test depends on multiple factors such as the true, yet unknown, disease model, analysis model employed, and sample size, when a large

multipoint HLOD is observed in reality, it would be more appropriate to perform Monte Carlo simulations to evaluate the significance level [Lin and Zou, 2004].

Acknowledgments

We thank Dr. Robert Elston for critically reading the manuscript and helpful discussions, and thank Dr. Gonalo Abecasis for clarification and advice on the maximization procedure in MERLIN. C.X. was partially supported by a Pilot Award from UL1RR024982 from the National Center for Research Resources.

REFERENCES

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet.* 2002; 30:97–101. [PubMed: 11731797]
- Abreu PC, Hodge SE, Greenberg DA. Quantification of type I error probabilities for heterogeneity LOD scores. *Genet Epidemiol.* 2002; 22:156–169. [PubMed: 11788961]
- Brent, RP. Algorithms for Minimization Without Derivatives. New York: Dover Publications; 2002.
- Chernoff H, Lander E. Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *J Statist Plann Inference.* 1995; 43:19–40.
- Chiano MN, Yates JR. Linkage detection under heterogeneity and the mixture problem. *Ann Hum Genet.* 1995; 59:83–95. [PubMed: 7762986]
- Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika.* 1977; 64:247–254.
- Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika.* 1987; 74:33–43.
- Faraway JJ. Distribution of the admixture test for the detection of linkage under heterogeneity. *Genet Epidemiol.* 1993; 10:75–83. [PubMed: 8472936]
- Greenberg DA, Abreu PC. Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. *Genet Epidemiol.* 2001; 21:299–314. [PubMed: 11754466]
- Hodge SE, Anderson CE, Neiswanger K, Sparkes RS, Rimo DL. The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): linkage studies, two-locus models, and genetic heterogeneity. *Am J Hum Genet.* 1983; 35:1139–1155. [PubMed: 6580815]
- Hodge SE, Vieland VJ, Greenberg DA. HLODs remain powerful tools for detection of linkage in the presence of genetic heterogeneity. *Am J Hum Genet.* 2002; 70:556–559. [PubMed: 11791217]
- Hodge SE, Rodriguez-Murillo L, Strug LJ, Greenberg DA. Multipoint lods provide reliable linkage evidence despite unknown limiting distribution: type I error probabilities decrease with sample size for multipoint lods and mods. *Genet Epidemiol.* 2008; 32:800–815. [PubMed: 18613118]
- Huang J, Vieland VJ. The null distribution of the heterogeneity lod score does depend on the assumed genetic model for the trait. *Hum Hered.* 2001; 52:217–222. [PubMed: 11713418]
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet.* 1996; 58:1347–1363. [PubMed: 8651312]
- Lemdani M, Pons O. Tests for genetic linkage and homogeneity. *Biometrics.* 1995; 51:1033–1041. [PubMed: 7548688]
- Liang KY, Rathouz PJ. Hypothesis testing under mixture models: application to genetic linkage analysis. *Biometrics.* 1999; 55:65–74. [PubMed: 11318180]
- Lin DY, Zou F. Assessing genomewide statistical significance in linkage studies. *Genet Epidemiol.* 2004; 27:202–214. [PubMed: 15389929]
- Lindsay, BG. Testing for Latent Structure. Mixture Models: Theory, Geometry and Applications. Hayward: Institute of Mathematical Statistics; 1995.
- Ott J. Linkage analysis and family classification under heterogeneity. *Ann Hum Genet.* 1983; 47:311–320. [PubMed: 6651220]
- Ott, J. Analysis of Human Genetic Linkage, 3rd edition. Baltimore, MD: The Johns Hopkins University Press; 1999.
- Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Statist Assoc.* 1987; 82:605–610.

- Smith CAB. Testing for heterogeneity of recombination fraction values in human genetics. *Ann Hum Genet.* 1963; 27:175–182. [PubMed: 14081488]
- Xing C, Elston RC. Distribution and magnitude of type I error of model-based multipoint lod scores: implications for multipoint mod scores. *Genet Epidemiol.* 2006; 30:447–458. [PubMed: 16685712]

Fig. 1.

Empirical distribution of multipoint HLOD compared with the theoretical distribution under the null hypothesis of no linkage. (A) and (C) correspond to the probability density functions of HLOD analyzed under models D5 and R5, respectively; (B) and (D) correspond to the cumulative distribution functions of HLOD analyzed under models D5 and R5, respectively. Solid line represents the theoretical distribution of HLOD assuming $T_{\text{multipoint}}$ follows a

distribution of $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$; dashed line and histogram represent the empirical distribution of HLOD based on results from 5,000 replicates, each of which contained 5,000 families simulated under simulation model III—parental phenotype unknown.

TABLE I
Proportion of zeros among $T_{\text{multipoint}}$ and empirical distribution of non-zero $T_{\text{multipoint}}^a$

Sample size	Analysis model							
	D2	D5	D8	R2	R5	R8		
Simulation model I ^b								
100	Zero %	51.0	52.0	52.8	52.1	51.8	52.3	
	Mean ^c	1.08	1.08	1.06	0.20	0.80	1.01	
	Variance	2.28	2.23	2.15	0.03	0.74	1.77	
	df ^c	1.08	1.09	1.08	0.75	1.00	1.08	
500	Zero %	50.4	51.1	51.9	51.8	51.8	51.7	
	Mean	1.02	1.02	1.03	0.39	1.03	1.05	
	Variance	2.05	1.98	1.97	0.11	1.82	2.23	
	df	1.06	1.08	1.09	0.92	1.08	1.08	
1,000	Zero %	51.4	52.5	53.0	53.2	53.7	53.8	
	Mean	1.05	1.05	1.04	0.52	1.06	1.03	
	Variance	2.05	2.09	2.04	0.22	1.95	1.87	
	df	1.09	1.09	1.07	0.98	1.09	1.08	
5,000	Zero %	51.1	51.7	51.8	51.8	51.7	52.0	
	Mean	1.05	1.04	1.02	0.88	1.08	1.08	
	Variance	1.95	1.89	1.91	0.90	2.08	2.13	
	df	1.08	1.08	1.07	1.04	1.09	1.10	
Simulation model II ^b								
100	Zero %	52.1	52.8	53.5	51.6	52.8	53.7	
	Mean	1.04	1.04	1.03	0.17	0.71	1.00	
	Variance	1.90	1.93	1.90	0.02	0.50	1.52	
	df	1.08	1.08	1.08	0.72	0.96	1.09	
500	Zero %	51.4	51.7	51.8	52.1	52.0	52.5	
	Mean	1.03	1.05	1.06	0.35	0.98	1.07	
	Variance	1.94	2.10	2.24	0.09	1.50	2.26	
	df	1.08	1.10	1.08	0.89	1.06	1.09	

Sample size	Analysis model						
	D2	D5	D8	R2	R5	R8	
1,000	Zero %	51.8	52.7	53.3	52.9	53.0	53.4
	Mean	1.04	1.05	1.04	0.47	1.04	1.05
	Variance	2.06	2.08	2.08	0.17	1.85	1.95
	df	1.07	1.07	1.06	0.96	1.09	1.10
5,000	Zero %	50.7	51.1	52.3	51.5	51.3	51.8
	Mean	1.04	1.03	1.05	0.82	1.08	1.08
	Variance	1.97	2.03	2.09	0.74	2.04	2.06
	df	1.07	1.07	1.09	1.01	1.07	1.08
Simulation model III ^b							
100	Zero %	51.8	52.7	53.7	52.4	53.3	54.0
	Mean	1.05	1.05	1.04	0.17	0.71	0.99
	Variance	2.26	2.27	2.24	0.02	0.49	1.46
	df	1.06	1.05	1.04	0.72	0.97	1.08
500	Zero %	51.0	51.1	52.2	51.6	51.9	52.1
	Mean	1.04	1.03	1.03	0.36	0.99	1.04
	Variance	2.15	2.13	2.10	0.09	1.54	2.07
	df	1.07	1.06	1.06	0.88	1.06	1.07
1,000	Zero %	52.0	52.6	52.9	52.6	53.0	53.3
	Mean	1.03	1.02	1.00	0.47	1.04	1.05
	Variance	2.11	1.94	1.82	0.17	1.84	1.90
	df	1.07	1.06	1.05	0.95	1.09	1.10
5,000	Zero %	50.7	51.1	52.1	51.4	51.1	51.5
	Mean	1.03	1.03	1.05	0.81	1.07	1.06
	Variance	1.96	2.03	2.09	0.74	2.09	2.10
	df	1.06	1.06	1.08	1.00	1.07	1.07

^aResults based on 5,000 replicates.

^bSimulation models I, II, and III correspond to parental phenotypes of one affected the other unaffected—a dominant model, both unaffected—a recessive model, and both unknown—an ambiguous model, respectively.

^c Assuming the non-zero $T_{\text{multipoint}}$ follow a χ^2 distribution, “mean” and “df” (degree of freedom) correspond to the mean estimates by the method-of-moments and the maximum likelihood estimation, respectively.

TABLE II

Empirical type I error rates^a for multipoint HLOD

Sample size	Nominal level ^b				Analysis model			
	D2	D5	D8	R8	R2	R5	R8	
Simulation model I ^c								
100	0.05	0.0530	0.0530	0.0502	–	0.0004	0.0472	
	0.01	0.0124	0.0118	0.0100	–	–	0.0064	
	0.001	0.0012	0.0010	0.0010	–	–	0.0004	
	0.0001	0.0004	0.0004	0.0002	–	–	–	
500	0.05	0.0502	0.0460	0.0470	–	0.0466	0.0498	
	0.01	0.0110	0.0094	0.0100	–	0.0088	0.0096	
	0.001	0.0012	0.0012	0.0006	–	0.0004	0.0018	
	0.0001	–	0.0004	0.0002	–	–	0.0002	
1,000	0.05	0.0516	0.0496	0.0500	0.0002	0.0490	0.0498	
	0.01	0.0096	0.0104	0.0094	–	0.0100	0.0088	
	0.001	0.0010	0.0008	0.0010	–	0.0004	0.0006	
	0.0001	0.0002	0.0002	0.0002	–	–	–	
5,000	0.05	0.0552	0.0512	0.0506	0.0272	0.0518	0.0502	
	0.01	0.0080	0.0082	0.0094	0.0002	0.0102	0.0108	
	0.001	0.0008	0.0006	0.0006	–	0.0012	0.0012	
	0.0001	0.0004	–	–	–	0.0004	0.0002	
Simulation model II ^c								
100	0.05	0.0470	0.0534	0.0484	–	0.0086	0.0464	
	0.01	0.0088	0.0092	0.0078	–	–	0.0058	
	0.001	0.0006	0.0008	0.0004	–	–	–	
	0.0001	–	–	–	–	–	–	
500	0.05	0.0506	0.0506	0.0504	–	0.0500	0.0508	
	0.01	0.0094	0.0112	0.0112	–	0.0056	0.0118	
	0.001	0.0010	0.0012	0.0016	–	–	0.0016	
	0.0001	–	0.0004	0.0006	–	–	–	

Sample size	Nominal level ^b				Analysis model			
	D2	D5	D8	R8	R5	R2	R5	R8
1,000	0.05	0.0508	0.0468	0.0476	–	–	0.0532	0.0502
	0.01	0.0100	0.0114	0.0118	–	–	0.0084	0.0094
	0.001	0.0008	0.0010	0.0014	–	–	0.0002	0.0008
	0.0001	–	–	–	–	–	–	–
5,000	0.05	0.0500	0.0488	0.0498	0.0172	0.0544	0.0538	
	0.01	0.0094	0.0102	0.0110	0.0004	0.0128	0.0120	
	0.001	0.0006	0.0010	0.0012	–	0.0008	0.0008	
	0.0001	–	0.0002	–	–	0.0002	0.0002	
Simulation model III ^c								
100	0.05	0.0506	0.0498	0.0468	–	0.0072	0.0452	
	0.01	0.0132	0.0128	0.0108	–	–	0.0046	
	0.001	0.0012	0.0012	0.0016	–	–	–	
	0.0001	0.0002	0.0002	–	–	–	–	
500	0.05	0.0518	0.0480	0.0450	–	0.0468	0.0476	
	0.01	0.0112	0.0114	0.0102	–	0.0066	0.0104	
	0.001	0.0022	0.0016	0.0018	–	–	0.0010	
	0.0001	–	–	–	–	–	–	
1,000	0.05	0.0500	0.0442	0.0450	–	0.0518	0.0492	
	0.01	0.0102	0.0112	0.0096	–	0.0086	0.0092	
	0.001	0.0012	0.0006	0.0006	–	0.0002	0.0008	
	0.0001	0.0002	–	–	–	–	–	
5,000	0.05	0.0492	0.0488	0.0494	0.0170	0.0538	0.0530	
	0.01	0.0094	0.0102	0.0110	0.0002	0.0120	0.0112	
	0.001	0.0006	0.0010	0.0012	–	0.0010	0.0008	
	0.0001	–	0.0002	–	–	0.0002	0.0002	

^aResults based on 5,000 replicates. “–” means no HLOD greater than the corresponding critical value.

^bThe nominal levels of 0.05, 0.01, 0.001 and 0.0001 approximately correspond to the multipoint HLOD of 0.59, 1.17, 2.07, and 3.00, respectively under the assumption that $T_{\text{multipoint}}$ follows a distribution of $\frac{1}{2}X_0^2 + \frac{1}{2}X_1^2$.

^cSimulation models I, II, and III correspond to parental phenotypes of one affected, the other unaffected—a dominant model, both unaffected—a recessive model, and both unknown—an ambiguous model, respectively.