European Journal of Human Genetics (2001) 9, 855–859

www.nature.com/ejhg

ARTICLE

Effect of misspecification of gene frequency on the two-point LOD score

Deb K Pal*[,1], Martina Durner[1] and David A Greenberg[1,2]

[1]*Department of Psychiatry, Mount Sinai Medical Center, New York, NY 10029, USA;* [2]*Department of Biomathematics, Mount Sinai Medical Center, New York, NY 10029, USA*

In this study, we used computer simulation of simple and complex models to ask: (1) What is the penalty in evidence for linkage when the assumed gene frequency is far from the true gene frequency? (2) If the assumed model for gene frequency and inheritance are misspecified in the analysis, can this lead to a higher maximum LOD score than that obtained under the true parameters? Linkage data simulated under simple dominant, recessive, dominant and recessive with reduced penetrance, and additive models, were analysed assuming a single locus with both the correct and incorrect dominance model and assuming a range of different gene frequencies. We found that misspecifying the analysis gene frequency led to little penalty in maximum LOD score in all models examined, especially if the assumed gene frequency was lower than the generating one. Analysing linkage data assuming a gene frequency of the order of 0.01 for a dominant gene, and 0.1 for a recessive gene, appears to be a reasonable tactic in the majority of realistic situations because underestimating the gene frequency, even when the true gene frequency is high, leads to little penalty in the LOD score. *European Journal of Human Genetics* (2001) **9**, 855–859.

Keywords: linkage analysis; parameter misspecification; gene frequency

Introduction

In LOD score analyses, one must assign certain parameters, eg the mode of inheritance, penetrance and gene frequency of the trait or disease under study. Assuming the correct mode of inheritance or penetrance values leads to a higher LOD score than if the incorrect parameters are specified.[1–3] We also know that misspecification of parameters can lead to an overestimation of the recombination fraction and a reduction in the power to detect linkage.[1] By simulating a simple Mendelian model, Greenberg *et al* showed that the mean maximum LOD score (ELOD) occurred at or near the analysis penetrance that matched the generating or 'true' penetrance.[4] As the true or generating penetrance dropped below approximately 0.6, the effect of varying the analysis penetrance on maximum LOD score became small. Only in the extreme case, when the true penetrance was high (0.9) and the analysis penetrance low, could the ELOD be halved. Furthermore, assuming a wrong penetrance would not lead to false negative evidence for linkage. Elston has shown analytically that maximizing the maximum LOD score with respect to genetic parameters, in order to infer the true mode of inheritance, is independent of family structure, marker frequencies and penetrance parameters.[3] Although ascertainment may play an as yet unquantified role, for purposes of detecting disease loci, the effect appears to be minimal.[5] Clerget-Darpoux has concluded that underestimating the gene frequency leads to a negligible change in maximum LOD score but a large overestimation of the recombination fraction.[1] However, the actual penalty in evidence for linkage remained to be determined.

Here we are not investigating the effect of misspecified gene frequency on the estimate of θ, only its effect on LOD score magnitude. Whilst there is strong evidence that the LOD score maximises at parameter values close to those of the generating model, the LOD score method has come under repeated criticism because of the perceived possibility

*Correspondence: Deb K Pal, Dept of Psychiatry, Mount Sinai Medical Center, Box 1229, Annenberg Bldg, One Gustave Levy Place, New York, NY 10029, USA. Tel: 212 241 6961; Fax: 212 831 1947;
E-mail: deb@shallot.salad.mssm.edu
Received 23 May 2001; revised 28 August 2001; accepted 28 August 2001

of producing false positive or false negative results when the model is misspecified. This in part, provided the motive for this study. In this study, we used computer simulation to ask two questions: (1) What is the penalty in evidence for linkage when the specified gene frequency is far from the true gene frequency?; (2) If the true model is misspecified in the analysis, are there any values of analysis gene frequency which would lead to a higher LOD score than under the 'true' generating parameters for inheritance model and gene frequency?

Methods

We simulated nuclear families according to a well-characterised family size distribution.[6] All matings were fully informative for the marker. Families were selected for linkage analysis if they had one or more affected members. Datasets were generated using our extensively-tested simulation program,[2] which uses a random process for each step in the simulation (eg selecting the mating type, family size, and segregation alleles from parents to offspring). For all analyses, ELODs were calculated by taking the mean maximum LOD score of all datasets.

We generated 100 datasets with 20 nuclear families each under simple dominant and recessive models, with and without reduced penetrance, and under one complex model (2 locus additive). For the simple models, we generated family data with traits showing single-locus dominant or recessive inheritance with 90% and 50% penetrance. The recombination fraction was fixed at 0.01. The families were generated with true gene frequencies of 0.001, 0.01, 0.1, 0.3, and 0.5 for dominant models and 0.001, 0.01, 0.1, 0.5 for recessive models. A high generating gene frequency increases the chance that the disease is inherited from both sides of the family, as well as increasing the chance of homozygosity in the parents, thereby reducing the informativeness of the marker. We then analysed the data using LIPED[7] under both the correct and incorrect dominance models (eg dominant generated data analysed recessive), under a range of different assumed gene frequencies. The resultant ELODs for each generating model are shown in Figures 1 and 2.

We also generated data under a two-locus additive trait model ('additive2')[8] in which a total of at least two disease-related alleles in any combination at the two loci are necessary for disease expression. The penetrance for the disease genotype was set at 90%. We chose a combination of generating gene frequencies at loci 1 and 2 each of 0.001, 0.006, 0.043, 0.100, 0.300 and 0.500. Greenberg and colleagues have shown that in LOD score analysis of complex disorders where marker loci are analysed for linkage to a trait one at a time, it is the inheritance assumptions at the *linked locus*, not the disease inheritance per se, that are critical.[8] Furthermore, complex traits can be satisfactorily approximated in single-locus analysis by subsuming the effect of other loci under reduced penetrance at the locus being analysed. We therefore analysed the families generated using the additive model assuming simple dominant and recessive models with 50% penetrance, under a range of assumed gene frequencies.

Lastly, we generated three-generation pedigrees of random size with at least four affected members under dominant and recessive, both with full and reduced penetrance, using the same gene frequencies and recombination fraction as for the nuclear families. We also generated pedigrees under the additive2 model with 90% penetrance using gene frequencies of 0.1/0.05 and 0.2/0.1 (locus 1/locus 2).

Results

Figure 1 shows results of analyses for data generated under a dominant model and a variety of generating gene frequencies. We see little dependence of ELODs on the assumed analysis gene frequency. When the generating gene frequencies are low (between 0.001 and 0.1), assuming a gene frequency above 0.1 leads to a slight fall in ELOD compared to analysing at gene frequencies between 0.001 and 0.1. For example, when the generating frequency is 0.01, the ELOD is 9.0 when analysed at the correct gene frequency. When analysed at the extreme gene frequency of 0.5, ie a 1.5 order of magnitude higher gene frequency than the true, the ELOD=7.69. In comparison, when the generating gene frequency is 0.1, the ELOD varies from 7.2 when analysis gene frequency is the same as the generating gene frequency, to 7.0 at an assumed gene frequency of 0.001, and 6.5 at an assumed gene frequency of 0.5. The ELOD also varies little when the generating gene frequency is much higher (0.5), eg from 2.2 when analysed under the correct parameters, to 1.75 when analysed assuming a gene frequency of 0.01. We also observe that, as expected, the ELODs peak when the analysis gene frequency approaches the true gene frequency. Examining the broad range of generating gene frequencies of 0.001–0.5, we find that the maximum penalty in ELOD analysed at any analysis gene frequency $\leqslant 0.5$ is 10–22% for a dominant trait.

Figure 2 shows analysis of data generated under the recessive model. There is even less fluctuation in ELOD caused by misspecification of gene frequency than in the dominant model. For a generating gene frequency of 0.01, the LOD score varies only from 7.7 to 7.4 as the analysing gene frequency is changed from 0.001 to 0.5. At a generating gene frequency of 0.5, the ELOD ranges from 6.3 to 6.6 assuming gene frequencies from 0.001 to 0.5. By comparison with the dominant model, the penalty in ELOD for the same range of generating and analysis gene frequencies for a highly penetrant, single-locus, recessive trait is only 3–4%.

Analysing a dominant gene assuming recessive inheritance causes a dramatic fall in ELOD by a factor ranging from 10–70 (Figure 1, bottom). Varying the analysis gene frequency makes little difference in this situation. Conversely, analysing a recessive gene assuming dominant inheritance causes a

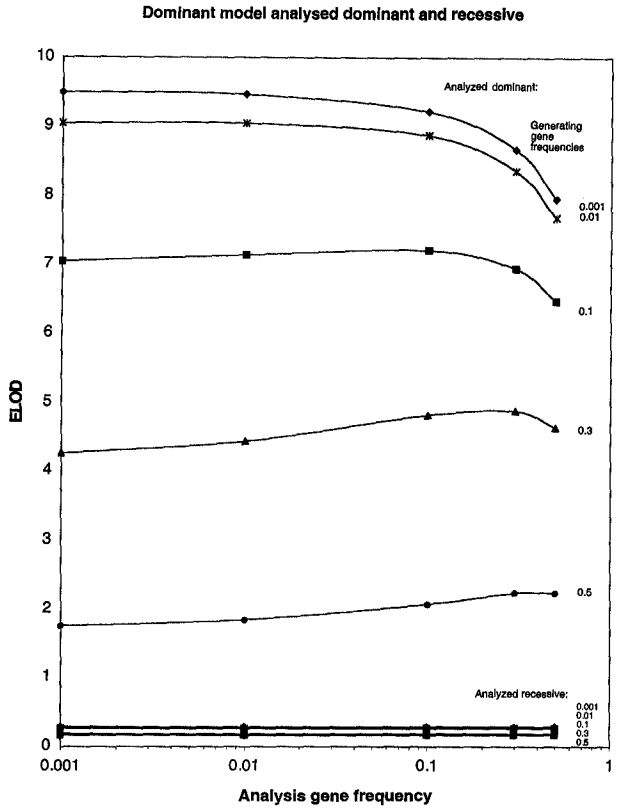

Figure 1 ELOD curves for 100 datasets of 20 nuclear families each generated under single locus 90% dominant model and variety of gene frequencies, analysed at variety of gene frequencies under dominant and recessive assumptions. The top five curves show results from datasets analysed as dominant; the bottom five (which appear to be only two curves) show results when analysed as recessive.

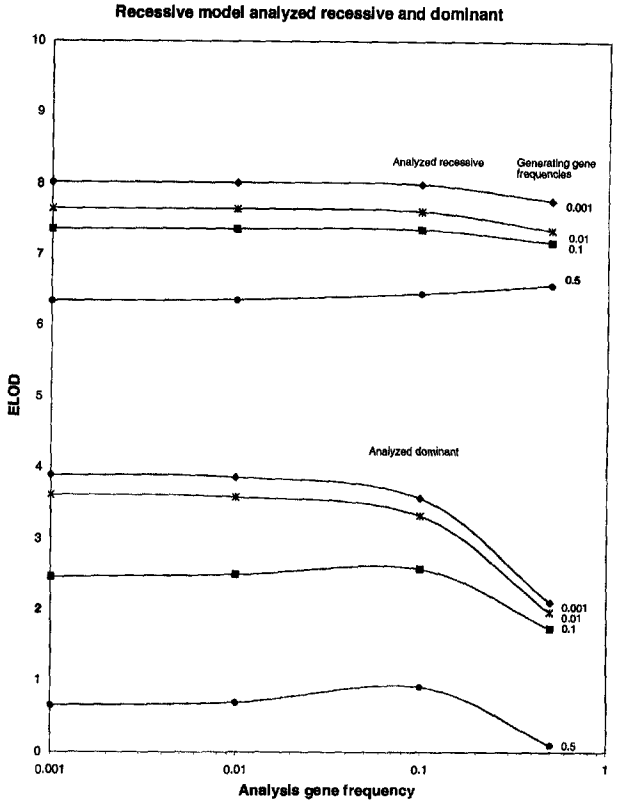

Figure 2 ELOD curves for 100 datasets of 20 nuclear families each generated under single locus recessive 90% penetrant model and variety of gene frequencies. analysed at variety of gene frequencies under dominant and recessive assumptions. The top four curves show results from datasets analysed as recessive; the bottom four show results when analysed as dominant.

lesser reduction in LOD score, by a factor of 3–12 (Figure 2, bottom). When a dominant trait is analysed recessive, varying the analysis gene frequency has negligible effect on the ELOD. For a recessive trait analysed as dominant, ELODs do not change when analysed between assumed gene frequencies 0.001 to 0.1, but fall when analysed at 0.5. When analysed under an incorrect inheritance model, varying the gene frequency never resulted in a higher LOD score than that obtained under the 'true' inheritance model. It is misspecifying the dominance model that leads to a serious underestimate of the LOD score, not misspecifying the gene frequency.

In the reduced penetrance dominant and recessive models (not shown), we found the results to be very similar to the 90% penetrance examples in Figures 1 and 2. Assuming the recessive model, there was little effect on the LOD score of varying the analysis gene frequency; for the dominant model, there was a slight drop in ELOD (about 30%) when the analysis gene frequency was set above 0.1. When the dominant trait was analysed assuming recessive inheritance, LOD scores fell dramatically, for example from ELOD=2.5 at a generating frequency of 0.001 when analysed assuming a dominant model, to ELOD=0.5 when a recessive model was assumed. Varying the analysis gene frequency had negligible effect on the ELOD in this situation. When the recessive trait was analysed under a dominant assumption, there was a small drop in LOD score (about 30%) and no marked variation in LOD score resulted from changing the analysis gene frequency.

The additive2 model is neither dominant nor recessive. Its properties are dependent on the frequencies of the two component loci; therefore it may behave, for analysis purposes, as either a dominant or a recessive. For this reason, we analysed it under both dominant and recessive assumptions and present the maximum LOD scores from either analysis, a method known as the maximized maximum LOD score (MMLS).[8] We found very little dependence of the MMLS on gene frequency when the analysis gene frequency was below 0.1 (see Figure 3). For clarity of graphing, we have

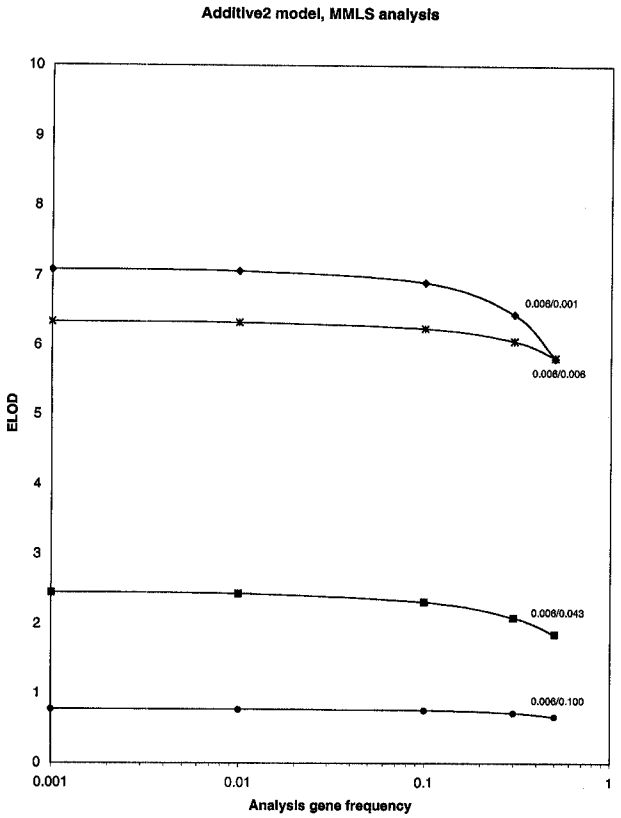

Figure 3 ELOD curves for 100 datasets of 20 nuclear families each generated under additive2 90% penetrant model and variety of gene frequencies (labelled at right of curves). The four curves show results from datasets analysed using maximised maximum LOD score (MMLS) method, taking the maximum LOD score from either dominant or recessive analysis.

shown only the example in which gene 1 has frequency 0.006 and gene 2 takes the values 0.001, 0.006, 0.043 and 0.1. As expected, when both genes were of low frequency (eg 0.001/0.001) higher LOD scores were obtained by recessive analysis; at higher gene frequencies (0.006/0.006 or more), higher LOD scores were obtained by dominant analysis. As in the case of the simple and reduced penetrance examples, the LOD score in the additive2 model dropped by up to 50% when analysis gene frequencies above 0.1 were assumed.

We found that the results for three-generation pedigrees were almost identical to those for nuclear families in the simple dominant, recessive, full and reduced penetrance, and additive models generated (not shown).

Discussion

We conclude that there is little penalty to pay, in terms of LOD score, for misspecifying the gene frequency in the dominant single-locus model, and even less so in the recessive model, for the range of parameters studied in this simulation. Furthermore, it would seem that specifying an arbitrary gene frequency is reasonable in LOD score analysis. A high analysis gene frequency, eg 0.5 may lead to some penalty (10–22%) in ELOD if the true model is dominant, but not if the true model is recessive. Conversely, there seems to be an even smaller penalty if an analysis gene frequency lower than the true (0.001–0.1) is chosen. Specifying a lower or higher gene frequency will not lead to an incorrect conclusion about the true mode of inheritance at that locus when the underlying gene frequency is unknown. Moreover, the incorrectly assumed gene frequency combined with the incorrect dominance model will appear unlikely to lead to false evidence for linkage. These conclusions hold true even when considering high-frequency genes that are assumed to underlie complex traits.

We have also shown that assuming 'incorrect' analysis gene frequencies has little effect on the LOD score in the reduced penetrance examples we have considered, when the analysis gene frequency is of the order of 0.1 or less. The same holds true for the additive2 model under MMLS analysis. The findings for all of these models apply equally to nuclear families and pedigrees.

Several authors have now shown, for a wide range of complex inheritance models, that single-locus approximations yield LOD scores very close to the values when data are analysed under the 'true' complex model.[4,8–13] When the 'true' mode of inheritance is unknown, Hodge and colleagues advocated analysing linkage data under both dominant and recessive assumptions, and adjusting for multiple testing, while keeping penetrance fixed at 50% to minimise type I error.[14] Since we have shown that the effect of varying analysis gene frequency is minimal, we advocate fixing the gene frequency at, say, 0.01 for a dominant gene and 0.1 for a recessive gene, when analysing either a simple or a complex trait by simple LOD score methods. We have used two-point analyses for these calculations, but in work in press, we show that these results apply equally to multipoint analysis.[15]

Acknowledgements

This work was supported in part by a Royal Society-Fulbright Distinguished Postdoctoral Scholarship (DKP); the Dunhill Medical Trust (DKP); and NIH grants NS37466 (MD) and DK31775, NS27941, MH48858 (DAG).

References

1 Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J: Effects of misspecifying genetic parameters in LOD score analysis. *Biometrics* 1986; **42**: 393–399.
2 Greenberg DA: Inferring mode of inheritance by comparison of LOD scores. *Am J Hum Genet* 1989; **34**: 480–486.
3 Elston RC: Man bites dog. The validity of maximizing lod scores to determine mode of inheritance. *Am J Hum Genet* 1989; **34**(4): 487–488.
4 Greenberg DA, Hodge SE: Linkage analysis under 'random' and 'genetic' reduced penetrance. *Genet Epidem* 1989; **6**: 259–264.
5 Vieland VJ, Hodge SE: The problem of ascertainment for linkage analysis. *Am J Hum Genet* 1996; **58**(5): 1072–1084.

6 Cavalli-Sforza LL, Bodmer WF: *The genetics of human populations*. San Francisco: Freeman, 1971; p310–313.
7 Ott J: Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet* 1974; **26**: 588–597.
8 Greenberg DA, Abreu P, Hodge SE: The power to detect linkage in complex disease by means of simple LOD-score analyses. *Am J Hum Genet* 1998; **63**(3): 870–879.
9 Greenberg DA: Linkage analysis assuming a single-locus mode of inheritance for traits determined by two loci: inferring mode of inheritance and estimating penetrance. *Genet Epidem* 1990; **7**(6): 467–479.
10 Vieland VJ, Greenberg DA, Hodge SE: Adequacy of single-locus approximations for linkage analyses of oligogenic traits. *Genet Epidemiol* 1992; **9**: 45–49.
11 Durner M, Greenberg DA: Effect of heterogeneity and assumed mode of inheritance on lod scores. *Am J Hum Genet* 1992; **42**: 271–275.
12 Vieland VJ, Greenberg DA, Hodge SE: Adequacy of single-locus approximations for linkage analyses of oligogenic traits: extension to multigenerational pedigree structures. *Hum Heredity* 1993; **43**(6): 329–336.
13 Goldin LR: Genetic heterogeneity and other complex models: a problem for linkage detection. In: Gershon ES, Cloninger CR, eds. Genetic approaches to mental disorders. Washington DC: American Psychiatric Press, 1994; 77–87.
14 Hodge SE, Abreu P, Greenberg DA: Magnitude of type I error when single-locus linkage analysis is maximized over models: a simulation study. *Am J Hum Genet* 1997; **60**(1): 217–227.
15 Greenberg DA, Abreu P: Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. *Genet Epidemiol* 2001; in press.