Published in final edited form as:
Ann Hum Genet. 2008 July ; 72(Pt 4): 566–574. doi:10.1111/j.1469-1809.2008.00442.x.

The Power and Robustness of Maximum LOD Score Statistics

Y. J. YOO and **N.R. MENDELL**

SUMMARY

The maximum LOD score statistic is extremely powerful for gene mapping when calculated using the correct genetic parameter value. When the mode of genetic transmission is unknown, the maximum of the LOD scores obtained using several genetic parameter values is reported. This latter statistic requires higher critical value than the maximum LOD score statistic calculated from a single genetic parameter value.

In this paper, we compare the power of maximum LOD scores based on three fixed sets of genetic parameter values with the power of the LOD score obtained after maximizing over the entire range of genetic parameter values. We simulate family data under nine generating models. For generating models with non-zero phenocopy rates, LOD scores maximized over the entire range of genetic parameters yielded greater power than maximum LOD scores for fixed sets of parameter values with zero phenocopy rates. No maximum LOD score was consistently more powerful than the others for generating models with a zero phenocopy rate. The power loss of the LOD score maximized over the entire range of genetic parameters relative to the maximum LOD score calculated using the correct genetic parameter value appeared to be robust to the generating models.

Keywords

linkage analysis; LOD; mod score; power; null distribution

INTRODUCTION

The LOD score statistic is a powerful statistic for human gene mapping that has been used for well over 50 years (Penrose, 1935; Morton, 1955). Since this statistic is a multiple of the likelihood ratio test statistic, its values are in part determined by the genetic model parameter values used in the analysis. The LOD score based on using the correct genetic model parameter value has been shown to be more powerful than affected sib pair statistics and other gene-model-free linkage statistics for detecting linkage (Greenberg et al., 1996; Greenberg et al., 1998; Abreu et al., 1999).

In spite of the popularity of case-control association analysis of population data in recent years, linkage studies using family data will continue to be very important. Clerget-Darpoux & Elston (2007) list several advantages of family based linkage studies over association studies and indicate that we will need both family based linkage studies and association studies to find disease susceptibility genes. Although Clerget-Darpoux & Elston's (2007) arguments are all in the context of comparing a family based linkage study using affected sib pairs to an association study, they also apply to family based linkage studies in general. Furthermore, the

Corresponding Author Yun Joo Yoo Samuel Lunenfeld Research Institute Mount Sinai Hospital Toronto, ON, M5T 3L9 Phone: 415-586-4800 (ext. 7358) Fax: 416-586-8404 Email: yoo@mshri.on.ca Nancy Role Mendell Department of Applied Mathematics and Statistics State University of New York at Stony Brook Stony Brook, New York 11794-3600 Phone: 631-632-8373 FAX: 631-632-8490 Email: Nancy.Mendell@Stonybrook.edu.

feasibility, efficiency, significance and the reliability of a family based linkage study may be even greater using the LOD score method in the analysis.

A large proportion of linkage analyses involve complex diseases, where the correct genetic model parameter values are unknown. The LOD scores based on incorrect genetic parameter values reduce power to detect linkage (Clerget —Darpoux et al., 1986). In order to help minimize this limitation, investigators have experimented with several genetic model parameter values and reported the result that generates the highest maximum LOD score (e.g., Sherrington, 1988), sometimes referred to as the 'mod score' (Clerget-Darpoux et al., 1986), the 'maximized maximum LOD score' (MMLS) (Greenberg, 1989) or LOD-M (Ulgen et al., 2004). These investigators have also noted that maximizing over *several* genetic parameter values can have a substantial effect on the null distribution. On the other hand the null distribution is unaffected by the parameter values used (correct or incorrect) when the maximized LOD score is based on a *single* parameter value (Williamson and Amos, 1990).

Several investigators conducted empirical studies on the null distribution and power of maximized LOD scores obtained through considering more than one genetic parameter value (Weeks et al., 1990; MacLean et al, 1993; Hodge et al., 1997; Greenberg at al., 1998; Ulgen et al., 2004). Hodge et al. (1997) investigated the null distributions of maximized LOD score statistics based on 2 ("dominance"), 10 ("penetrance") and 20 ("dominance and penetrance") parameter values. They concluded that after subtracting 0.3 (using either their "dominance" or "penetrance" maximized LOD score) or 0.6 (using their "dominance and penetrance" maximized LOD score) one could use the critical value for the LOD score based on one genetic parameter value.

Ulgen et al. (2004) investigated the LOD score obtained by maximizing over the entire range of genetic parameter values (with the constraint that the disease associated allele frequency is 0.5 or less). They concluded that the resulting null distribution of the cube root of this maximum LOD approximates a mixture in which 13% of the values equal 0 and the remaining non-zero values are normally distributed with mean equal to 0.8 and standard deviation equal to 0.2.

Since the main motivation for using maximum LOD scores is to minimize the loss of power caused by using incorrect genetic parameter values, the challenge is to identify the genetic model parameter values or parameter space that will result in the greatest power. Clearly, a LOD score based on fully maximizing over the genetic parameter space will be greater than a maximum LOD score based on maximizing over a subset of 2 or 20 genetic parameter values in that space. However we also know that the increase in maximum LOD associated with the increase in the number of parameter values is fairly substantial in the *null case* (Hodge et al., 1997; Ulgen et al., 2004). Thus maximizing over additional genetic parameter values may not result in an increase in power after adjusting for inflated type I error.

Only a few studies have evaluated the power of maximum LOD scores based on more than one genetic parameter value. Greenberg et al. (1998) investigated the power of the "corrected dominance maximized LOD score", obtained by subtracting 0.3 from the maximum LOD based on the two genetic parameter values given by Hodge et al. (1997). They noted that (1) the corrected dominance maximized LOD score had slightly lower power than the maximized LOD calculated using one (correct) genetic parameter value and (2) the greatest loss in power occurred when the actual generating model parameter values were quite different from the two analysis values.

No study has compared the power of the "fully maximized LOD score", the LOD score obtained through maximization over the entire range of genetic parameter values, with the power of "partially maximized LOD scores", the LOD scores obtained by maximizing the LOD score over a fixed set of genetic parameter values (for example, 2 or 20). One issue of particular

interest is whether numerical maximization of the LOD score over the entire range of the genetic parameters results in a large enough increase in the LOD to make up for the required increase in the critical value associated with full maximization. The second issue is the choice of the set of genetic parameter values to use in the calculation of a partially maximized LOD score. This issue is only important if the partially maximized LOD scores obtained using 20 (or 2) parameter values result in the same or greater power than the fully maximized LOD. Of particular interest is whether the partially maximized LOD score based on only two parameter values suggested by Hodge et al. (1997) is as powerful as the partially maximized LOD score based on 20 genetic parameter values for the low penetrance, high phenocopy, incomplete dominance situations that are characteristic of complex diseases.

The focus of this paper is to compare the power of the fully maximized LOD score to the power of partially maximized partially maximized LOD score statistics. We want to determine whether it is worth it, given the required increase in the critical value, to fully maximize the LOD score over the range of genetic parameter values. We consider LOD scores partially maximized over 3 fixed sets of genetic parameter values and a fully maximized LOD score obtained through maximization over the range of genetic parameter values. Our primary purpose is to compare the power of two of the adjusted maximum LOD scores proposed by Hodge et al. (1997) with the adjusted maximum LOD score proposed by Ulgen et al. (2004) which requires a maximization algorithm and a much higher critical value for a significant finding. The third partially maximized LOD score we will consider includes penetrance values that allow for additivity and phenocopies. The power of this maximum LOD score will be also compared with other partially and fully maximized LOD scores.

METHODS

Maximum LOD scores

The LOD score likelihood calculations are based on the assumption of a biallelic disease locus with alleles *D* and *d* that are in Hardy Weinberg equilibrium. The genetic parameter value is a vector, $\varphi = \{q, f_0, f_1, f_2\}$, where q denotes the frequency of the disease allele, *D*, and f_i denotes the penetrance (the probability of being affected) of the disease for the genotype with $i = 0, 1, 2$ copies of *D*.

We consider 3 partially maximized LOD score statistics, which we denote as LOD-F2, LOD-F20 and LOD-F20$_{AP}$. The designation ‘F’ is used because these maximized LOD scores are based on a *fixed* set of values for φ. The numbers 2 and 20 refer to the number of genetic parameter values used in the partial maximization. The subscript ‘AP’ is used to differentiate LOD-F20$_{AP}$ from LOD-F20. The LOD-F20$_{AP}$ statistic maximizes over fewer values of f_2 than LOD-F20 (4 vs. 10) but it additionally computes the LOD for settings of f_1 and f_0 that correspond to ***a***dditive effects of the *D* allele $f_1 = (f_2 + f_0)/2$, and ***p***henocopies ($f_0 = 0.1$).

The genetic parameter values used for each of these 3 partially maximized LOD score statistics are given in Table 1. The values of the genetic parameters for LOD-F2 and LOD-F20 are the same as those considered by Hodge et al. (1997). We propose LOD-F20$_{AP}$ with the objective of obtaining greater power in those situations where we have additivity or phenocopies.

The power values of these partially maximized LOD score statistics are compared to the power of LOD-M, the fully maximized LOD score statistic proposed by Ulgen et al. (2004). The genetic parameter space for LOD-M is defined as $0 < q \leq 0.5$, and $0 \leq f_0 \leq f_1 \leq f_2 \leq 1$. Of course, all of these LOD scores are maximized over values of the recombination fraction θ, for $0 \leq \theta \leq 0.5$. The power values of all of these maximized LOD score statistics are compared to the “***c***lassical” maximum LOD score LOD-C, the LOD score obtained by setting φ equal to the correct genetic parameter value and maximizing over $0 \leq \theta \leq 0.5$.

Data Simulation

Two sets of nuclear family data were generated and investigated. The first was generated through single ascertainment (one affected proband) and the second through double ascertainment (two affected probands who are siblings). The SLINK (Ott, 1989; Weeks et al., 1990) software simulated the phenotype data for the proband's (or probands') remaining family members and the marker genotype data for the proband(s) and their siblings. The parents were set to be heterozygous for 4 different alleles at the marker locus.

The disease phenotype and marker genotype data were generated under several genetic parameter value settings. To make the power comparison of the maximum LOD score statistics meaningful, we chose recombination parameter (θ) values and heterogeneity parameter (a) values so that the power of the LOD-C was between 85% and 95% for most of the genetic parameter values considered. In every case the family structure consisted of sets of nuclear families with 4 offspring per family and every data set consisted of 240 individuals (or 40 nuclear families). In most cases, θ was set equal to 0.1. Whenever the power of LOD-C was too low, θ was set at 0.01. For those situations that resulted in very high power with $\theta = 0.1$, the heterogeneity parameter, a, was set equal to 0.5. The SLINK software simulates heterogeneity by setting the genetic parameter value for the unlinked disease locus equal to that of the linked disease locus.

We generated data sets for each of 9 genetic parameter values as follows:

1. 'D100', a completely dominant disease allele with full penetrance: (q=0.01, f_0=0.0, f_1=f_2=1.0),
2. 'R100', a completely recessive disease allele with full penetrance: (q=0.01, f_0= f_1=0.0, f_2=1.0),
3. 'D80', a completely dominant disease allele with 80% penetrance: (q=0.01, f_0=0.0, f_1=f_2=0.8),
4. 'R80', a completely recessive disease allele with 80% penetrance; (q=0.01, f_0= f_1=0.0, f_2=0.8),
5. 'D100-Q10', a completely dominant disease allele with full penetrance and disease allele frequency equal to 0.1 instead of 0.01 (as in D100 of (1)) (q=0.1, f_0=0.0, f_1= f_2=1.0),
6. 'A50', an additive model with full penetrance in the disease allele homozygote: (q=0.01, f_0=0.0, f_1=0.5, f_2=1.0),
7. 'D100-P10', a completely dominant disease allele with full penetrance and a nonzero phenocopy rate: (q=0.01, f_0=0.1, f_1= f_2=1.0),
8. 'D90-P05', a completely dominant disease allele with 90% penetrance, and a nonzero phenocopy rate: (q=0.01, f_0=0.05, f_1= f_2=0.9), and
9. 'I80-P05', a dominant disease allele, full penetrance in the disease homozygote and a nonzero phenocopy rate: (q=0.01, f_0=0.05, f_1=0.8, f_2=1.0).

We set q=0.01 in every case but (5) (D100-*Q10*). Parameter values (7)-(9) set $f_0 > 0.0$ and hence are referred to as "phenocopy values" whereas parameter values (1)-(6) are referred to as "no phenocopy values".

We also generated the data sets under the null hypothesis, *i.e.* θ=0.5, in order to obtain the null distribution of the three partially maximized LOD scores. The simulated samples consisted of n=200 families with 4 offspring per family. This was done for 10 generating genetic parameter values (chosen from the analysis parameter settings of LOD-$F20_{AP}$). The simulated samples

were constructed using double and single ascertainment methods resulting in 20 simulations. For each of these simulations we obtained *N*=400 samples. The result was 8,000 simulated samples for which we computed each of the partial maximum LOD scores under the null hypothesis.

Computational Methods

For each data set, we computed LOD-F2, LOD-F20, LOD-F20$_{AP}$ and LOD-M scores as well as the LOD score for the correct genetic parameter value (LOD-C). In order to calculate LOD scores, we developed a program to compute the likelihood of the nuclear family data using Elston-Stewart algorithm (1971), which was embedded in the program to calculate LOD-M scores. The LOD-F2, LOD-F20 and LOD-F20$_{AP}$ scores were obtained as the maximum of the 2 or 20 LOD scores with different genetic analysis parameter values as defined above. The validity of these computations was confirmed by comparing the results with the output from the MLINK program (Lathrop & Lalouel, 1984).

We developed a separate program, based on Powell's direction set method (Press et al., 1993), to compute the fully maximized LOD, LOD-M. This program finds a minimum or maximum of a function on *n*-dimensional space by moving along the lines with changing directions. Powell's algorithm may converge to a local maximum. We used several strategies to obtain the global maximum. These include repeating the search using randomly selected starting points, searches through the direction of unit vectors after computing a local maximum, and searches that used the twenty parameter values of LODF-20 as starting points. Application of the algorithm using 100 different random starting values 10 times per sample, resulted in the program achieving results which varied less than 0.05 lod units in 99 out of 100 samples. After confirming that the empirical distributions were essentially the same for 20 random starting values and for *1000* random starting values (Kolmogorov-Smirov test), we used 20 random starting points for the simulation so that the required computation time was feasible.

The program "MLOD" that we have developed to calculate LOD-M for a given data set of *n* nuclear families allows for an arbitrary number of random starting points and is available at no cost at www.ams.sunysb.edu/~nmendell/mlod

Null Distributions

We obtained the empirical null distributions for LOD-F2, LOD-F20 and LOD-F20$_{AP}$ through simulation of 8,000 data sets of *n=200 families with k=4 offspring per family* (4,000 with single ascertainment and 4,000 with double ascertainment). We fit the cumulative distribution $F\left(\sqrt[3]{x}\right)=\pi+(1-\pi)\,\Phi\left(\frac{\sqrt[3]{x}-\mu}{\sigma}\right)$ proposed by Ulgen (2004) to the results obtained for each of the partially maximized LODS. We estimated π, μ, and σ using proportions of LOD values equal to 0.0 *(p)*, the mean (*m)* and the standard deviation (*s)* of the non-zero values of the cube root of the maximum LODS, respectively.

The null distribution for LOD-M proposed in Ulgen et al. (2004), LOD-M $\left(F\left(\sqrt[3]{x}\right)=0.13+0.87\quad\Phi\left(\frac{\sqrt[3]{x}-0.78}{0.24}\right)\right.$, was used to obtain critical values for LOD-M. A 50:50 mixture distribution of χ^2_0 and χ^2_1 was used to obtain the critical values for LOD-C (Ott, 1974).

Power comparisons

The power of an α level test based on each of the maximum LOD statistics (given in Table 1) was estimated for each of the generating genetic parameter values (in Table 2) as the relative

frequency in 1000 simulated data sets in which the maximum LOD score exceeded its α level critical value (for α=0.05, 0.01, 0.0001).

We compared the power of the maximum LOD score statistics pairwise (for each genetic parameter value and ascertainment method) using McNemar tests at the 0.05 level. We also obtained ELODs of adjusted maximum LOD score statistics. We estimated the ratio of the power of each of the fully and partially maximized LODs relative to the power of LOD-C for each of the genetic parameter values considered.

Results

Null distributions obtained through simulation

The null distributions observed for the data generated using single and double ascertainment were not significantly different based on a Kolmogorov-Smirov test. Therefore, we combined the data sets and report parameters for a single null distribution for each of the LOD statistics. The resulting null distribution parameter values and critical values are summarized in Table 3. The 0.0001 level critical values for LOD-F2 and LOD-F20 are estimated to equal 3.1 and 3.7, respectively. The corresponding critical value for LOD-F20$_{AP}$ is 3.9. The critical value of LOD-M based on the null distribution proposed by Ulgen et al. (2004) equals 4.6 for α=0.0001.

Power comparison using the critical values from approximated null distributions

We next used the critical values in Table 3 to estimate the power of the maximized LOD scores for each of the 9 generating genetic parameter values in Table 2 (*N*=1000 samples for each parameter value). Table 4 reports the observed power values corresponding to three type I error values (0.05, 0.01, 0.0001) for the maximized LOD scores and the LOD-C score for the data generated using single ascertainment. Table 4 also presents the "adjusted" ELOD values of these maximum LOD scores for each generating genetic parameter value. The adjustment for each ELOD (subtracted from each ELOD) was calculated as the 0.0001 level critical value minus 3.0 (the 0.0001 critical value for LOD-C). Thus the adjusted ELODs were obtained by subtracting 0.11, 0.68, 0.91 and 1.61 from the ELODS for LOD-F2, LOD-F20, LOD-F20$_{AP}$ and LOD-M, respectively.

Pairwise power comparisons of the maximum LOD score statistics using the McNemar test at the 0.05 level provided rankings of the four maximum LOD statistics ('1' denoting the most powerful). These ranks are shown in Table 4 (superscripts) for the 0.0001 level tests. Examination of the ranks shows that, with the exception of D100 and D100-Q10, the power values of LOD-F2, LOD-F20 and LOD-F20$_{AP}$ were significantly greater than LOD-M for all genetic parameter values with a zero phenocopy rate. For D100 and D100-Q10, the power values of LOD-M were the same as or greater than LOD-F2. However, LOD-F20$_{AP}$ was essentially as powerful as either LOD-F2 or LOD-F20 for all of these zero phenocopy rate situations. Further, power values of LOD-M and LOD-F20$_{AP}$ were significantly greater than those obtained for LOD-F20 for the data generated using non-zero phenocopy rates (D100-P10, D90-P05, I80-P05). The power of LOD-F20$_{AP}$ was slightly higher or the same as LOD-M for all three of these models. Results based on adjusted ELODs for each of the maximum LOD scores were consistent with the findings on the power values.

The ranking of the maximum LODs obtained for the simulations under the double ascertainment method were essentially the same as those described above for the single ascertainment method.

The relative power of LOD-M and each of the partially maximized LOD scores was then calculated as the ratio of the power of the maximized LOD divided by the power of LOD-C. These relative power values are plotted against the power of LOD-C in Figures 1a, 1b, 1c, and

1d. (Figure 1). The varying power values are obtained by varying the Type I error. From the plotted graphs, we observed that the ratio is not constant across the power values of LOD-C and is lower for the lower power values of LOD-C.

We observe from Figure 1a that the ratio of the power of LOD-M to the power of LOD-C is essentially the same for all generating parameter values. For example, at about 60% power for LOD-C, the ratio of power of LOD-M to the power of LOD-C equals the 0.7 regardless of genetic parameter value. However, in Figure 1b we see that the corresponding ratios for LODF2 when LOD-C has 60% power equals 0.9 for generating parameter value 'A50', 0.6 for generating parameter value D100 and about 0.3 for generating parameter value 'D100-P10'. LOD-F20 showed the similar large differences in relative power associated with the generating parameter value as LOD-F2 (Figure 1c). LOD-F20$_{AP}$ (Figure 1d) shows a lot less dependence on the generating parameter value than LOD-F2 and LOD-F20. This indicates that the *power loss of LOD-F2 and LOD-F20 are highly dependent on the actual parameter values*, LODF-20$_{AP}$ is slightly less dependent whereas the power loss of LOD-M appears to be independent of the true values of the genetic parameters.

DISCUSSION

Our results show that two factors affect the power of partially maximized LOD score statistics. The first factor is the extent to which the candidate parameter values depart from those of the generating genetic model. The second factor is the magnitude of the increase in the critical value of the test statistic as the number of genetic parameter values used in the maximizing process increases. This second factor affects the power of the fully maximized LOD score, LOD-M, as well as the partially maximized LOD score statistics, but the magnitude of the effect is much greater for LOD-M than for the other scores. LOD-M requires an increase of 1.6 lod units. In contrast, the increase in the critical value compared with LOD-C at α=0.0001 is quite small, 0.1 lod units, for LOD-F2, and gets larger for LOD-F20 and LOD-F20$_{AP}$, 0.7 lod units and 0.9 lod units, respectively.

The genetic generating parameter values corresponding to phenocopy models (D100-P10, D90-P05 and I80-P05) and incomplete dominance (A50 and I80-P05) are most relevant for evaluating the effect of large departures of parameter values from analysis values. These results are also particularly relevant to complex diseases involving several genes and non-genetic factors in which a sizable proportion of affected individuals will not have the disease predisposing allele.

We considered the D100, D80, R100 and R80 parameter values in this study because the same generating parameter values were used by other investigators to incorporate the effects of heterogeneity (Greenberg et al., 1998, Abreu et al., 1999). The "D100-Q10" parameter value was considered in order to estimate the effect of misspecification of gene frequency on the power of partially maximized LOD scores.

The results suggest that partially maximised LOD scores based on 20 values (LOD-F20 and LOD-F20$_{AP}$) have greater power than LOD-F2 whenever the true genetic parameter values depart substantially from the two LOD-F2 analysis values. For example, LOD-F20 is more powerful than LOD-F2 for the D100 generating parameter value, because the D100 generating parameter value is not included in the LOD-F2 calculation. More interestingly, LOD-M is more powerful (at the α=0.0001 level) than the maximum LOD scores based only on zero phenocopy values (LOD-F2 and LOD-F20) *whenever there are phenocopies*, even when f_0 equals 0.1. This finding suggests that the maximum LOD is more sensitive to differences between the analysis values of f_0 and the true value of f_0 than it is to the corresponding difference for f_2. Since the parameter f_0 represents one dimension of the four-dimensional parameter space, the

partially maximized LOD-F2 and LOD-F20 may not pick up the varying signal generated by the f_0 dimension.

The LOD-F20$_{AP}$ statistic was the most powerful for one of genetic parameter values that resulted in non-zero phenocopy rates and close in power to LOD-M for the other two non-zero phenocopy rate parameter values. The LOD-F20$_{AP}$ statistic was also more powerful than LODM for genetic parameter values with zero phenocopy rates and genetic heterogeneity. Thus, for the genetic parameter values considered in this study, the performance of LOD-F20$_{AP}$ yielded the best or close to the best power overall.

Finally, we observed that the power of LOD-M relative to LOD-C is largely unaffected by the genetic generating parameter values whereas the relative power of the partially maximized LODs is substantially more sensitive to the genetic generating parameter values. Thus, a nonsignificant finding based on LOD-M is not going to be due to using the wrong set of parameter values in the analysis. On the other hand, a nonsignificant finding based on LOD-M can be a result of the required inflation of the critical value.

This study has several limitations. First, we did not simulate under a realistic heterogeneity model or an epistatic model. Second, we limited our pedigree structure to nuclear families of uniform size and simulated samples in which every family had completely informative marker information (both parents typed and heterozygous for different marker alleles). Third, we did not consider the effects of genotyping or diagnostic errors.

We anticipate that larger pedigrees would result in greater power for all of the maximum LOD scores whereas allowing for missing marker data and homozygous parents would lower power of these maximum LOD scores. We conjecture that the power ranking of these LOD score statistics is unlikely to be substantially affected by missing marker data and homozygous parents.

A major contribution of this study is that our simulations have identified situations in which LOD-M performs the same or better than some partially maximized LOD statistics. This encouraging result was unexpected, making the finding all the more meaningful (Elston, 1989).

The null distribution values may vary slightly from the values we used, depending on sample sizes, family structures and possibly, the genetic generating parameter values. Nevertheless, the values given here seem to be accurate estimates of the relative ranking of the critical values of these maximum LOD scores. We were surprised, however, to observe in our simulation, that two partially maximized LOD scores based on the same number of analysis values, LOD-F20 and LOD-F20P, differ in null distribution. The implication is that the null distribution needs to be determined through simulation when using a partially maximized LOD score if the genetic parameter set differs from those of LOD-F2, LOD-F20 and LOD-F20$_{AP}$.

To conclude, our guidelines to investigators based on these simulations are as follows. (1) Power is optimized by using LOD-F20$_{AP}$ for linkage analyses of complex diseases and/or traits with completely unknown genetic parameter values. LOD-M can also be used and will result in the same or little change in power. (2) For simple monogenic traits with no phenocopies and reduced penetrance, power is optimized by LOD-F2. (3) For monogenic traits that may have full penetrance, LOD-F20$_{AP}$ and LOD-F20 have a slight advantage over LOD-F2. These guidelines apply only to the specific situations simulated by this study -- a sample of nuclear families that are all approximately the equal in size, and obtained through either single or double ascertainment.

ACKNOWLEDGEMENTS

This work is supported in part by grants (R01-MH071523) from the National Institute of Mental Health, the National Alliance for Research on Schizophrenia and Depression, the Essel Foundation, and the Sidney R. Baer, Jr. Foundation. The authors would like to acknowledge Drs. Derek Gordon, Deborah Levy and. Stephen Finch for their helpful comments and suggestions.

REFERENCES

Abreu PC, Greenberg DA, Hodge SE. Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex disease. Am J Hum Genet 1999;65:847–857. [PubMed: 10441591]

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J. Effects of misspecifying genetic parameters in lod score analysis. Biometrics 1986;42:393–399. [PubMed: 3741977]

Clerget-Darpoux F, Elston RC. Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. Hum Herd 2007;64:91–96.

Elston RC, Stewart J. General model for the genetic analysis of pedigree data. Hum Hered 1971;21:523–542. [PubMed: 5149961]

Elston RC. Man Bites Dog? The validity of maximizing lod scores to determine mode of inheritance. Am J Hum Genet 1989;34:487–488.

Greenberg DA. Inferring mode of inheritance by comparison of lod scores. Am J Med Genet 1989;34:480–486. [PubMed: 2624256]

Greenberg DA, Hodge SE, Vieland VJ, Spence MA. Affected-only linkage methods are not a panacea. Am J Hum Genet 1996;58:892–895. [PubMed: 8644756]

Greenberg DA, Abreu PC, Hodge SE. The power to detect linkage in complex disease by means of simple lod-score analysis. Am J Hum Genet 1998;63:870–879. [PubMed: 9718328]

Hodge SE, Abreu PC, Greenberg DA. Magnitude of Type I error when single-locus linkage analysis is maximized over models: a simulation study. Am J Hum Genet 1997;60:217–227. [PubMed: 8981965]

Lathrop GM, Lalouel JM. Easy calculation of lod scores and genetic risks on small computers. Am J Hum Genet 1984;36:460–465. [PubMed: 6585139]

Morton NE. Sequential test for the detection of linkage. Am J Hum Genet 1955;7:277–318. [PubMed: 13258560]

McLean CJ, Bishop DT, Sherman SL, Diehl SR. Distribution of lod scores under uncertain mode of inheritance. Am J Hum Genet 1993;52:354–361. [PubMed: 8430696]

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. [PubMed: 4422075]

Ott J. Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 1989;86:4175–4178. [PubMed: 2726769]

Penrose LS. The detection of autosomal linkage in a data which consists of pairs of brothers and sisters of unspecified parentage. Ann Eugen 1935;6:133–138.

Press, WH.; Flannery, BP.; Teukolsky, SA.; Vetterling, WT. Numerical Recipes in C: The art of scientific computing, second Edition. Cambridge University Press; 1993.

Sherrington R, Brynjolfsson J, Peterson H, Potter M, Dudleston K, Barraclough B, Wasmuth J, Dobbs M, Gurling H. Localization of a susceptibility locus for schizophrenia on chromosome 5. Nature 1988;336:164–167. [PubMed: 2903449]

Ulgen A, Yoo YJ, Gordon D, Finch S, Mendell NR. Percentiles of the null distribution of 2 maximum lod scores tests. Hum Hered 2004;57:39–48. [PubMed: 15133311]

Weeks DE, Lehner T, Squires-Wheeler E, Kaufmann C, Ott J. Measuring the inflation of the lod score due to its maximization over model parameter values in human linkage analysis. Genet Epidemiol 1990;7:237–243. [PubMed: 2227370]

Weeks DE, Ott J, Lathrop GM. SLINK: a general simulation program for linkage analysis. Am J Hum Genet 1990;47(suppl):A204.

Willamson JA, Amos CI. On the asymptotic behavior of the estimate of the recombination fraction under the null hypothesis of no linkage when the model is misspecified. Genet Epidemiol 1990;7:309–318. [PubMed: 2253866]

Wilson EB, Hilferty MM. The distribution of chi-square. Proc Natl Acad Sci 1931;17:684–688. [PubMed: 16577411]

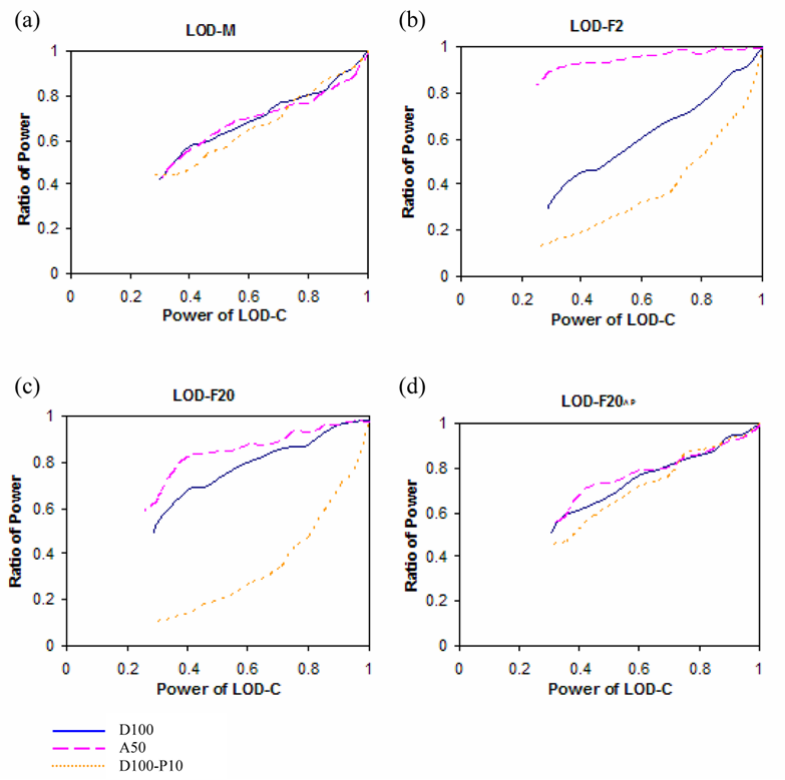

Figure 1. Ratio of the power of LOD-M, LOD-F2, LOD-F20 and LOD-F20$_{AP}$ to the power of LOD-C, for three generating models

The ratio of power is obtained using the critical values for LOD-C and for maximum LOD scores corresponding to the same Type I error varying from 0.5 to 10^{-7}.

Table 1

Genetic parameter values used for the five maximum LOD score statistics

LOD scores	Genetic parameter values in maximization (in the order of $\{q,f_0,f_1,f_2\}$)*
LOD-C	$\varphi_c = \{q_c, f_{c0}, f_{c1}, f_{c2}\}$; true genetic parameter vector for φ
LOD-F2	$\varphi_{1,F2} = \{0.01, 0.0, 0.0, 0.5\}$, $\varphi_{2,F\text{-}2} = \{0.01, 0.0, 0.5, 0.5\}$
LOD-F20	$\varphi_{i,F20} = \{0.01, 0.0, 0.0, 0.1\times i\}$ for i=1, …, 10 ; $\{0.01, 0.0, 0.1\ (i\text{—}10)\ , 0.1\times(i\text{—}10)\}$ for i=11, …, 20
LOD-F20$_{AP}$	$\varphi_{i,F20AP} = \{0.01, 0.0, 0.0, 0.25\times i\}$ for i=1, …, 4; $\{0.01, 0.0, 0.25\times(i\text{—}4), 0.25\times(i\text{—}4)\}$ for i=5, …,8; $\{0.01, 0.1, 0.1, 0.25\times(i\text{—}8)\}$ for i=9, …, 12; $\{0.01, 0.1, 0.25\times(i\text{—}12)\ , 0.25\times(i\text{—}12)\}$ for i=13, …,16; $\{0.01, 0.0, 0.5, 1.0\}$, $\{0.01, 0.1, 0.55, 1.0\}$, $\{0.01, 0.0, 0.3, 0.6\}$, $\{0.01, 0.1, 0.35, 0.6\}$ for i=17, …, 20
LOD-M	$0 < q \leq 0.5$, $0 \leq f_0 \leq f_1 \leq f_2 \leq 1$

* All LOD scores were also maximized over θ on [0, 0.5].

Table 2

Genetic model parameter values used to generate the simulated data

Group	Models	Genetic parameter value					
		q	f_2	f_1	f_0	θ^{**}	a^{*}
No Phenocopy Models	**D100**	0.01	1	1	0	0.1	0.5
	R100	0.01	1	0	0	0.1	0.5
	D80	0.01	0.8	0.8	0	0.1	0.5
	R80	0.01	0.8	0	0	0.1	0.5
	D100-Q10	0.1	1	1	0	0.1	0.5
	A50	0.01	1	0.5	0	0.1	NA
Phenocopy Models	**D100-P10**	0.01	1	1	0.1	0.01	NA
	D90-P05	0.01	0.9	0.9	0.05	0.01	NA
	I80-P05	0.01	1	0.8	0.05	0.01	NA

* a denotes the proportion of unlinked families to the marker assuming heterogeneity.

** For double ascertainment models, $\theta = 0.1$ for every model.

Table 3

Null distributions of maximum LOD score statistics obtained using simulated data

	Estimated parameters*			Null distribution approximated**	Critical Values at Type I error†		
	p	*m*	*s*		0.05	0.01	0.001
LOD-F2	37.6%	0.52	0.26	$0.376 + 0.624\Phi((\sqrt[3]{x}-0.52)/0.26)$	0.70	1.26	3.11
LOD-F20	25.3%	0.57	0.27	$0.253 + 0.747\Phi((\sqrt[3]{x}-0.57)/0.27)$	0.91	1.57	3.68
LOD-F20$_{AP}$	14.0%	0.59	0.27	$0.140 + 0.860\Phi((\sqrt[3]{x}-0.59)/0.27)$	1.04	1.72	3.91
LOD-M	13.3%	0.78	0.24	$0.133 + 0.867\Phi((\sqrt[3]{x} - 0.78) / 0.24)$	1.55	2.33	4.61

* *p* (percentage of zero values), *m* (mean of cube root of non-zero values), *s* (standard deviation of cube root of non-zero values) obtained from 8000 samples.

** Approximated null distribution using method of Ulgen et al. (2004).

† Critical values computed based on approximated null distributions.

Table 4

Power values and adjusted ELODs for the five maximum LOD scores

Type I error	Maximum Lod scores	Critical Value	No Phenocopy Models						Phenocopy Models		
			D100	R100	D80	R80	D100-Q	A50	D100-P10	D90-P05	I80-P05
0.05	**LOD-C**	0.59	98.9%	91.9%	89.1%	83.1%	96.4%	98.0%	92.4%	98.7%	94.1%
	LOD-F2	0.70	94.8%	88.2%	85.2%	80.4%	86.4%	97.6%	67.2%	93.5%	83.6%
	LOD-F20	0.91	97.4%	88.1%	85.9%	79.3%	93.5%	97.6%	68.6%	94.1%	83.6%
	LOD-F20$_{AP}$	1.04	96.7%	85.5%	84.0%	75.1%	91.8%	96.9%	87.2%	97.6%	90.0%
	LOD-M	1.55	95.3%	84.3%	80.9%	74.2%	91.2%	95.1%	84.7%	95.1%	88.5%
0.01	**LOD-C**	1.18	94.4%	78.5%	72.9%	64.4%	87.5%	92.9%	81.7%	95.2%	84.1%
	LOD-F2	1.26	86.5%	73.5%	68.2%	63.2%	71.4%	91.9%	46.7%	83.8%	65.8%
	LOD-F20	1.57	91.2%	74.2%	68.1%	60.9%	80.6%	89.2%	44.4%	81.3%	63.5%
	LOD-F20$_{AP}$	1.72	90.6%	70.5%	66.7%	56.5%	78.6%	86.5%	73.6%	91.1%	73.4%
	LOD-M	2.33	87.1%	67.7%	59.5%	52.5%	75.5%	81.5%	69.0%	88.7%	72.8%
0.0001	**LOD-C**	3.00	65.8%	36.8%	26.1%	19.4%	48.6%	49.7%	42.0%	70.3%	38.7%
	LOD-F2	3.11	[3]43.0%	[1]28.4%	[1]16.9%	[1]17.7%	[3]23.4%	[1]46.7%	[3]9.3%	[3]35.7%	[2]18.6%
	LOD-F20	3.68	[1]54.4%	[2]26.3%	[1]18.3%	[2]14.8%	[1]33.2%	[2]40.3%	[4]6.5%	[4]29.4%	[4]14.1%
	LOD-F20$_{AP}$	3.91	[2]51.4%	[3]21.7%	[1]16.6%	[3]12.1%	[2]30.3%	[3]36.3%	[1]23.5%	[2]45.1%	[3]17.6%
	LOD-M	4.61	[3]46.9%	[3]21.8%	[2]13.4%	[4]10.2%	[2]29.4%	[4]32.2%	[2]20.4%	[1]50.2%	[1]23.2%
ELOD*	**LOD-C**		4.02	2.62	2.19	1.93	3.15	3.27	2.77	4.09	4.09
	LOD-F2		2.90	2.28	1.89	1.82	2.05	3.16	1.30	2.60	2.60
	LOD-F20		3.44	2.09	1.73	1.49	2.47	2.81	0.97	2.30	2.30
	LOD-F20$_{AP}$		3.30	1.85	1.57	1.24	2.31	2.60	1.93	2.95	2.95
	LOD-M		3.04	1.69	1.27	1.03	2.15	2.30	1.71	3.22	3.22

1, 2, 3 and4 are the power rankings for maximum LOD scores based on the McNemar test at the 0.05 level. Highest observed power is ranked as 1. Results showing non-significant differences are ranked as tied.

*Adjusted ELODs are calculated as Mean — δ(LODj), where δ(LOD-j) =0, 0.11, 0.68, 0.91 and 1.61 for j=C, F2, F20, F20$_{AP}$ and M, respectively.