A Bayesian LOD Score for Linkage Analysis of Complex Diseases

Ma'ayan Fishelson,* Dmitry Rusakov,* Dan Geiger
Computer Science Department, Technion
Haifa 32000, Israel

January 14, 2004

Keywords: Complex Diseases, MMLS, Linkage Analysis, Lod score, SUPERLINK.

Abstract

Motivation: The emphasis of genetic linkage analysis is rapidly shifting from the task of locating genes responsible for mendelian diseases, for which one major gene causes the disease, to identifying a collection of genes and their interactions which govern what is termed complex diseases.
Results: We propose a Bayesian averaging approach based on a known method called *MMLS-C*. The difference is in the way the likelihood of data, given a specific position of a disease gene, is computed. Given a certain inheritance model, either recessive or dominant, the likelihood of data is computed by averaging the likelihood of data given this model over all penetrance values, using a flat prior. We have implemented this approach, called Maximizing Bayesian *LOD* score (*MBLOD*), within SUPERLINK (Fishelson and Geiger, 2002), and demonstrated via simulations its advantages.
Availability: The program SUPERLINK is available at bioinfo.cs.technion.ac.il/superlink/temp/V1.4.html
Contact: {fmaayan,rusakov,dang}@cs.technion.ac.il

1 Introduction

The emphasis of genetic linkage analysis is rapidly shifting from the task of locating single genes causing mendelian diseases to locating a collection of disease genes and their interactions which govern what is termed complex diseases. One method that has been pursued in the genetics literature for analyzing complex diseases is to compute the likelihood of odds (*LOD* score) assuming two inheritance models of the disease: recessive or dominant, each with 50% penetrance, commit to the model that produces the higher maximum score, and correct for multiple tests by subtracting 0.3 from the maximum score (Hodge et al., 1997; Greenberg et al., 1998). This method, called *MMLS-C*, has been shown to work well in simulations (Greenberg et al., 1998; Abreu et al., 1999; Abreu et al., 1999; Greenberg and Abreu, 2001) and is used in real genetic linkage studies (Barrett et al., 2001; Silverman et al., 2002).

In this paper, we propose a Bayesian averaging approach based on *MMLS-C*. The difference is in the way the likelihood of data, given a specific position of a disease gene, is computed. According to the Bayesian averaging approach proposed herein, the likelihood of data given an inheritance model, either recessive or dominant, is computed by averaging the likelihood of data given this model over different penetrance values, using a flat prior. As in *MMLS-C*, we commit to the model that produces the higher maximum score, and correct for multiple tests by subtracting 0.3 from the maximum score. We have implemented this approach, which we call Maximizing Bayesian *LOD* score (*MBLOD*), within SUPERLINK (Fishelson and Geiger, 2002), a genetic linkage analysis program based on Bayesian networks, and demonstrated via simulations the advantages of Bayesian averaging in the context of genetic linkage analysis.

Our evaluation methodology follows that of (Greenberg et al., 1998; Durner et al., 1999). We gen-

*These authors contributed equally to this work.

erated pedigree data assuming various complex models of inheritance in which the disease is caused by two, three, or four genes. The location of one of the genes is recovered using our method and previous methods. We have shown that *MMLS-C* performs well on inheritance models that have not been tested before, and on larger families than used in previous simulation studies. For the complex inheritance models we tested, *MBLOD* has higher predictive power than *MMLS-C* and *NPL*, on the average.

The implementation of *MMLS-C* and *MBLOD* into SUPERLINK makes this program readily available to analyze a variety of complex genetic diseases. Its ability to perform computations on larger pedigrees than other programs (Fishelson and Geiger, 2002) raises the hope that the additional predictive power gained due to larger pedigrees and due to the *MBLOD* approach proposed herein will help decipher additional inheritance mechanisms of complex diseases.

The paper is organized as follows. Sections 2 and 3 introduce selected background on genetic linkage analysis. Section 4 describes the Bayesian averaging method we propose. In Section 5 we describe complex inheritance models used for simulation studies. Experimental results are reported in Section 6. Finally, Section 7 summarizes and outlines future work.

2 Genetic Linkage Analysis

A *locus* is the location on the chromosome of a *gene* or a *marker*. Common markers are short tandem repeats of two, three, or four nucleotides. Each gene, or marker, can appear in several different forms, or *alleles*. The different alleles of markers are the different number of repeats. In diploid organisms, such as humans, each individual carries two copies of each autosomal locus, one copy inherited from each parent. The sequence of alleles at different loci inherited by an individual from one parent is called a haplotype, and the two haplotypes of an individual constitute this individual's genotype. When genotypes are measured, the result is a list of unordered pairs of alleles, one pair for each locus. The standard measurement procedures do not distinguish between the paternal and maternal alleles at each locus.

A *recombination* is said to have occurred between two loci, if an haplotype of an individual contains alleles that resided in different haplotypes in the individual's parent. The probability that a recombination occurs between two loci is called the *recombination fraction* and is denoted by θ. The recombination fraction is a measure of the distance between two loci. For a comprehensive background on human genetic linkage analysis consult (Terwilliger and Ott, 1994; Lange, 1997; Ott, 1999).

The input for a genetic linkage problem consists of family trees, called pedigrees, phenotype description of people in the pedigree at the trait studied (affected, unaffected, or unknown), and partial unordered genotype information for people at marker loci. The output of genetic linkage analysis is the most probable location of a disease gene under study. This is traditionally done by placing a hypothetical disease locus somewhere within a map of markers and computing the probability of data given that specific location which is represented as the recombination fractions between the disease locus and its closest flanking markers. The position which produces the highest probability (and its vicinity) is the location declared to contain a disease gene, provided the result is statistically significant.

Pedigrees are represented naturally using the following types of variables and probability tables:

- **Genetic Loci.** For each individual i and locus j, we define two random variables $G_{i,jp}$, $G_{i,jm}$ whose values are the specific alleles at locus j in individual i's paternal and maternal haplotypes, respectively.

- **Marker Phenotypes.** For each individual i and marker locus j, we define a random variable $P_{i,j}$ whose value is the specific unordered pair of alleles measured at locus j of individual i.

- **Disease Phenotypes.** For each individual i, we define a binary random variable P_i whose values are affected or unaffected.

- **Selector Variables.** For each individual i and marker locus j, we define two binary random variables $S_{i,jp}$ and $S_{i,jm}$, the values of which are determined as follows. If a denotes i's father and b denotes i's mother, then

$$S_{i,jp} = \begin{cases} 0 & \text{if } G_{i,jp} = G_{a,jp} \\ 1 & \text{if } G_{i,jp} = G_{a,jm} \end{cases}$$

$S_{i,jm}$ is defined in a similar way, with b replacing a. These variables are hidden, though their values can sometimes be inferred indirectly from pedigree data.

The probability tables which relate these variables are of the following forms:

- **Transmission Models:** $Pr(G_{i,jp} \mid G_{a,jp}, G_{a,jm}, S_{i,jp})$, $Pr(G_{i,jm} \mid G_{b,jp}, G_{b,jm}, S_{i,jm})$, where a and b are i's parents in the pedigree. These tables are deterministic, namely, consist solely of zeroes and ones. The first probability table equals 1, if $G_{i,jp} = G_{a,jp}$ and $S_{i,jp} = 0$ or if $G_{i,jp} = G_{a,jm}$ and $S_{i,jp} = 1$. In all other cases, this probability table equals 0. The second probability table is defined analogously. Note that if mutations need to be modelled, then these tables are no longer deterministic.

- **Marker Models:** $Pr(P_{i,j} \mid G_{i,jp}, G_{i,jm})$. These tables are also deterministic. The probability table equals 1 if $P_{i,j} = (G_{i,jp}, G_{i,jm})$ or if $P_{i,j} = (G_{i,jm}, G_{i,jp})$. Otherwise it equals 0. If measurement errors need to be introduced, then these tables are no longer deterministic.

- **Recombination Models:** $Pr(S_{i,1p}) = Pr(S_{i,1m}) = 0.5$, $Pr(S_{i,jp} \mid S_{i,j-1p}, \theta_{j-1})$ and $Pr(S_{i,jm} \mid S_{i,j-1m}, \theta_{j-1})$, where θ_{j-1} is the recombination fraction between locus $j-1$ and locus j. If either j or $j-1$ is a disease locus, then the parameter θ_{j-1} is unknown and is varied to maximize the probability of data.

- **Population Allele Frequencies:** $Pr(G_{i,jp})$, $Pr(G_{i,jm})$ are allele frequencies, where i is a *founder*, namely, an individual in the pedigree whose biological parents are not included in the pedigree.

- **Inheritance Models:** $Pr(P_i \mid G_{i,jp}, G_{i,jm})$. The inheritance models are the main focus of this paper. A *recessive inheritance model* M_r with penetrance probability α defines the probability table $Pr(P_i \mid G_{i,jp}, G_{i,jm})$ as follows:

$$\begin{cases} \alpha & \text{if } G_{i,jp} = d \textbf{ and } G_{i,jm} = d \\ 0 & \text{otherwise} \end{cases}$$

where h is the normal allele at disease locus j, and d is the mutant allele. A *dominant inheritance model* M_d with penetrance probability α defines $Pr(P_i \mid G_{i,jp}, G_{i,jm})$ as follows:

$$\begin{cases} \alpha & \text{if } G_{i,jp} = d \textbf{ or } G_{i,jm} = d \\ 0 & \text{otherwise} \end{cases}$$

The most general one-locus penetrance model defines $Pr(P_i \mid G_{i,jp}, G_{i,jm})$ as follows:

$$\begin{cases} f_1 & \text{if } G_{i,jp} = h \textbf{ and } G_{i,jm} = h \\ f_2 & \text{if } G_{i,jp} = h \textbf{ and } G_{i,jm} = d \\ f_2' & \text{if } G_{i,jp} = d \textbf{ and } G_{i,jm} = h \\ f_3 & \text{if } G_{i,jp} = d \textbf{ and } G_{i,jm} = d \end{cases}$$

Normally, f_2 is taken to be equal to f_2' except when the model takes into account whether the source of the mutant allele d is the father or mother, a phenomena that is termed *parental imprinting*. Furthermore, the disease status may depend on several loci which can interact in various ways to determine the penetrance of a disease, such as epistasis, suppression, complementary gene action, duplication of genes. These models, among others, are described in Section 5.

The above notation needs to be slightly modified in the case of markers and disease loci that reside on sex chromosomes.

Pedigree data is an assignment e of values to a subset E of marker and disease phenotype variables. The probability of data $Pr(E = e \mid \theta)$ is computed by assigning a specific value to each measured variable, and then summing the product of all the probability tables just described over all possible values for the rest of the variables. The likelihood function $Pr(E = e \mid \theta)$ is maximized w.r.t. the recombination fraction θ which determines the possible positions of a disease gene. The variables and probability tables described above define a Bayesian network (Friedman et al., 2000; Fishelson and Geiger, 2002).

The significance of the best recombination fraction $\theta = \theta_1$ is normally determined using the *LOD* score defined for an inheritance model M (say recessive M_r or dominant M_d) along with a given penetrance probability α:

$$LOD(\theta_1) = \log_{10} \mathcal{L} \triangleq \log_{10} \frac{P(Data|\theta_1, \alpha, M)}{P(Data|\theta = 0.5, \alpha, M)} \quad (1)$$

where $\mathcal{L}$ is the likelihood ratio. The classical practice is to declare that linkage has been found between a disease gene and a location on a chromosome when the *LOD* score is above 3. The threshold of 3 approximately corresponds to a p-value of 0.05 for detecting true linkage (Ott, 1999). Genetics texts and papers warn from performing *LOD* score calculations with several different inheritance models and penetrances, and choosing the highest outcome, because in such practice a *LOD* score of 3 no longer corresponds to a p-value of 0.05, often yielding unacceptably many erroneous proposals for disease genes' locations. Such errors are said to be of type I.

Almost all analyses of linkage are done by comparing the probability of pedigree data assuming a specific disease gene location, $\theta = \theta_1$, versus the probability of the null hypothesis, $\theta = 0.5$, that the disease gene is on a different chromosome, or far on the same chromosome. There is some effort however to introduce a Bayesian approach which requires the specification of a prior on the true location of the disease genes (Vieland, 1998). The interpretation of the results of a linkage study in the Bayesian framework is simply the probability of linkage as a function of the location on the genetic map. Nevertheless, this method is not widely used in linkage studies.

3 Linkage Analysis for Complex Diseases

Complex diseases are diseases controlled by several genes, whose influence may or may not interact with external factors such as, say, obesity. For complex diseases the problem of type I errors in standard linkage analysis is magnified considerably. One normally does not know which inheritance model and penetrance values to use, and computing *LOD* scores for many different models will often yield eventually a high but meaningless *LOD* score.

There are several methods to detect linkage that try to avoid the model-based approach. Among these methods, which are called *non-parametric*, are Affected Sib-Pairs (ASP) (Suarez and Van Eerdewegh, 1984), Affected Sibships (Ethier and Hodge, 1985), Affected Pedigree Members (APM) (Weeks and Lange, 1988), and Non-Parametric Linkage (NPL) (Kruglyak et al., 1996). These methods are based on the fact that if a disease gene is located close on the chromosome to some marker, than, for all models of inheritance, it segregates with that marker to offsprings more often then random segregation. For example, two siblings share on the average one allele at each position on the chromosome. However, if one observes that *affected* siblings share on the average more than one allele at a specific location, one may conclude that a disease gene resides near that marker. Non-parametric methods are appealing because one does not have to specify a model of inheritance or penetrances, and yet, suspected areas for the location of disease genes are determined. Indeed these methods, and especially *NPL* using the GENEHUNTER software (Kruglyak et al., 1996), are currently the leading methods in the study of complex diseases.

The simplicity of these non-parametric methods comes at a statistical cost since they do not use the entire data. They only use affected individuals, neglecting information due to healthy members of the pedigrees.

Abreu et al. (1999) show that *NPL* scores follow a normal distribution, which allows one to transform *NPL* to *LOD* scores via $NPL^2/(2\ln 10) \approx LOD$ for sufficiently large samples. We refer to the transformed *NPL* score as *adjusted score* for *NPL*.

Greenberg et al. (1998) proposed the following method. Find the maximum *LOD* score for a given map of genetic markers assuming a dominant inheritance model with 50% penetrance. Repeat the same computation assuming a recessive inheritance model with 50% penetrance. Commit to the model that scores higher and use the *LOD* score of the selected model minus 0.3. This method is termed *MMLS-C* for Maximizing Maximum Lod Score using a Correction. The computed value is referred to as the *adjusted score* for *MMLS-C*, and is denoted by Z. The correction of 0.3, which equals $\log_{10} 2$, compensates for testing two models rather than one (Hodge et al., 1997). The correction of $\log_{10} g$ for g tests is a standard correction (Ott, 1999). In other words, a *LOD* score of 3.3 obtained using two tests gives the same or better level of type I errors than a *LOD* score of 3 obtained using one test.

Extensive simulations have been performed demonstrating convincingly that the simple model-based method *MMLS-C* has stronger power to detect disease genes even for complex diseases, than the described non-parametric methods (Greenberg et al., 1998; Abreu et al., 1999; Durner et al., 1999; Greenberg and Abreu, 2001). It has been shown for several complex inheritance models that when testing for linkage, the inheritance model at the linked locus being examined is important in the detection of linkage, and not the overall inheritance model of the disease. The effect of the other loci can be subsumed in the reduced penetrance with relatively little loss of power to detect linkage (Abreu et al., 1999). This is the apparent reason *MMLS-C* is so successful, albeit the 50% penetrance value selected by *MMLS-C* is a bit arbitrary.

4 Averaging of Penetrances

We propose to take uncertainty regarding both inheritance model and penetrance value into account in the prediction of disease loci position, as follows.

We consider a class of inheritance models $\mathcal{M}$, currently consisting of recessive M_R and dominant M_D models. For each inheritance model $M \in \mathcal{M}$, we specify a prior probability $\mu(\alpha|M)$ of penetrance values given the model M. In this paper we select a flat prior for the penetrances given an inheritance model, namely, $\mu(\alpha|M_d) = \mu(\alpha|M_r) = 1$ for $0 \leq \alpha \leq 1$. The likelihood of the data given an inheritance model and a specific location of a disease gene is computed as follows:

$$P(Data|\theta, M) = \int P(Data|\theta, \alpha, M)\mu(\alpha|M)d\alpha \quad (2)$$

Eq (2) is based on the assumption that the prior $\mu(\alpha|\theta, M) = \mu(\alpha|M)$ does not depend on the true location of a disease gene θ. We now define the following *BLOD* score for evaluating linkage for any particular disease locus for a specific model of inheritance:

$$BLOD(\theta, M) = \log_{10} \frac{P(Data|\theta, M)}{P(Data|\theta = 0.5, M)}, \quad (3)$$

where $P(Data|\theta, M)$ is computed via Eq. (1). Note that we fall short of specifying a prior for M and for θ, hence this is not a fully Bayesian approach.

Our approach, called *MBLOD*, computes the Maximum *BLOD* score assuming two inheritance models, recessive or dominant, commits to the inheritance model that produced the higher maximum *BLOD*, and corrects for multiple testing by subtracting 0.3 from the maximum *BLOD*. The value computed is referred to as the *adjusted score* for *MBLOD*.

Note that *BLOD* score generalizes the standard *LOD* score computations. Had we placed all the mass on a specific penetrance value α, then the *BLOD* score would have reduced to the LOD score for the specified inheritance model with the specified penetrance.

From a user's perspective our method is non-parametric in the sense that a user need not specify a model of inheritance or penetrances, however, the computations done are clearly model-based, taking inheritance models and penetrances into account.

A *LOD* score of 3 is the standard threshold for declaring linkage. It is based on the well known fact that $P(\mathcal{L} > 10^3) < 10^{-3}$ where $\mathcal{L}$ is the likelihood ratio. The same reasoning yields a threshold of 3 for $BLOD$, provided that the probability of the null hypothesis can be written as

$$P(Data|\theta = 0.5) = \int P(Data|\theta = 0.5, \alpha)\mu(\alpha)d\alpha.$$

5 Inheritance Models

We generated data under different complex disease inheritance models. The first nine are two-locus models (Fig 1). The next four are three-locus and four-locus disease models, which generalize the respective two-locus models. For all models presented below we describe the condition necessary for an individual to be affected with probability α, which is the penetrance value.

Additive2 Model (Add2). A total of at least two mutant alleles at the two disease loci are required for a person to be affected.

Additive3 Model (Add3). A total of at least three mutant alleles at the two disease loci are required for a person to be affected.

Additive4 Model (Add4, also termed Epistatic Model RR). A total of four mutant alleles at the two disease loci are required for a person to be affected. This model corresponds to the known phenomenon of *gene duplication*.

Epistatic Model DD. At least one mutant allele at each of the two disease loci is required for a person to be affected.

Epistatic Model DR. At least one mutant allele at the linked disease locus and two mutant alleles at the unlinked disease locus are required for a person to be affected.

Add2

0	0	α
0	α	α
α	α	α

Add3

0	0	0
0	0	α
0	α	α

Add4 (RR)

0	0	0
0	0	0
0	0	α

DD

0	0	0
0	α	α
0	α	α

DR

0	0	0
0	0	α
0	0	α

RD

0	0	0
0	0	0
0	α	α

R+R

0	0	α
0	0	α
α	α	α

D+D

0	α	α
α	α	α
α	α	α

R⊕R

0	0	α
0	0	α
α	α	0

Figure 1: **Penetrance Tables for the Two-locus Inheritance Models.** The rows correspond to the linked disease locus, and the columns correspond to the unlinked disease locus. Both relate to the following phenotypes in the respective order: *hh, hd, dd* (*h* represents the normal allele and *d* the mutant allele).

Epistatic Model RD. Two mutant alleles at the linked disease locus and at least one mutant allele at the unlinked disease locus are required for a person to be affected.

Complementary Action Model R+R. At least two mutant alleles at one of the two disease loci are required for a person to be affected. This model corresponds to the known phenomenon of *complementary gene action*.

Model D+D. At least one mutant allele at one of the two disease loci is required for a person to be affected. This model also corresponds to the phenomenon of *complementary gene action*.

Suppression Model R⊕R. Two mutant alleles at one of the two disease loci and at most one mutant allele at the other disease locus are required for a person to be affected. This model corresponds to the phenomenon of *suppression*.

The following four models are three-locus and four-locus disease models.

Gene Duplication Model ($\mathbf{R^3}$, $\mathbf{R^4}$). Two mutant alleles at each of the three, or four, disease loci are required for a person to be affected.

Complementary Gene Action Model ($\sum_3 \mathbf{R}$, $\sum_4 \mathbf{R}$). At least two mutant alleles at one of the three, or four, disease loci are required for a person to be affected.

6 Experimental Results

The experiments compare the predictive power of *MBLOD* versus *MMLS-C* and *NPL* on a variety of realistic simulated data sets, and show its superiority for most generating models examined. Run times of *MBLOD* are 2-5 fold slower compared to *MMLS-C*, which may limit its applicability for large data sets. All experiments use SUPERLINK (Fishelson and Geiger, 2002) for *MBLOD* and *MMLS-C* scores and GENEHUNTER (Kruglyak et al., 1996) for *NPL* scores via the option NPL_{all}.

Data was generated using a simulation program, retaining only families which included at least two affected siblings. In all generating models, one of the disease loci is linked to the marker map at $\theta = 0.01$, and the other disease loci are unlinked. The frequencies of the of mutant alleles at the disease loci were set so that the population prevalence is approximately 1%. For all models, the penetrance $\alpha = 1.0$ was used. For each inheritance model in each experiment, 100 sets of 20 random families were simulated. The inheritance models along with the parameters used for generating the data are termed generating models (GMs).

Several sets of experiments were performed. In the first experiment (Tables 1, 2), we used all 13 models listed in Section 5 to analyze pedigree data consisting of nuclear families with 2 to 5 children. The analysis used one marker with 2 alleles of equal frequencies. In the next two experiments, 6 inheritance models, $\mathbf{RR}, \mathbf{RD}, \mathbf{DR}, \mathbf{DD}, \mathbf{R+R}, \mathbf{R \oplus R}$, were used. In the second experiment (Table 3, Figure 3), pedigree data consisted of three-generation families of 12 individuals. The third experiment (Tables 3, 4) was based on a multipoint analysis with 5 markers, each with 2 alleles of equal frequencies, with recombination of 0.1 between the markers. Pedigree data for this experiment consisted of nuclear families. For each experiment we present the average adjusted LOD-score computed by each analysis method, and the power to detect linkage achieved by each one. Power is defined as the percentage of data sets in which the given statistic reached a given significance level. For each analysis, the average adjusted score, denoted by Z, is computed as an average over 100 data sets.

Tables 1-4 and Figure 3 show that the power and average adjusted score (Z) are usually highest using

MBLOD and lowest using *NPL*.

In the fourth experiment we used the **R+R** Model to test the power increase of *MBLOD*, *MMLS-C*, and *NPL*, as the number of nuclear families segregating the disease increased. Nuclear families with progeny generated according to the negative-binomial distribution with mean 2.8 and variance 5.3 (Cavalli-Sforza and Bodmer, 1971) were analyzed using one marker with 5 alleles of equal frequencies. Figure 2 summarizes the results showing that starting from 3 families, *MBLOD* retained the highest power for all levels of evidential support.

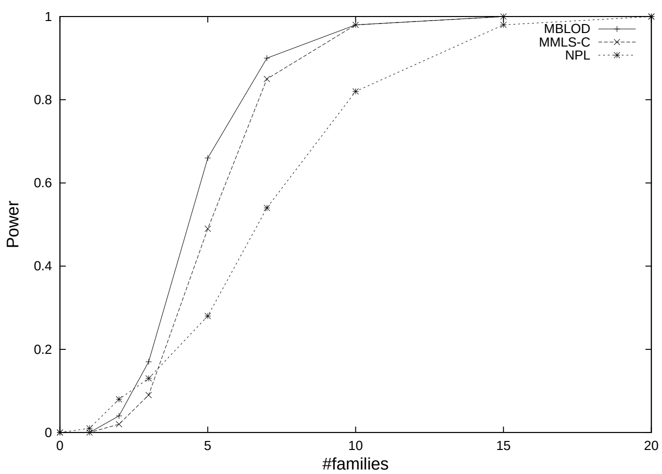


Figure 2: Power of each method for the **R+R** model as a function of the number of nuclear families segregating the disease. Each point is based on 100 data sets of x nuclear families.

We repeated the fourth experiment for a semi-multiplicative model with three disease loci. If an individual is homozygous for the mutant allele in all three disease loci, then the penetrance is α^3. If an individual is homozygous for the mutant allele in only two of the disease loci, then the penetrance is α^2. In all other cases the penetrance is zero. A model in this spirit may be used for explaining the inheritance of the short-segment Hirschsprung disease (Gabriel et al., 2002). The power results for *NPL*, *MMLS-C*, and *MBLOD*, assuming $\alpha = 0.8$, were 0.32, 0.60, and 0.58, respectively, for 20 families, and for 40 families, all methods reached practically the same power of 0.94, 0.96, and 0.96.

A final experiment tested whether the rate of type I errors introduced by the *MMLS-C* and *MBLOD* methods is indeed under the required level of 0.001, as guaranteed by theory. We generated 10,000 data sets each consisting of 20 nuclear families with 2-11 children, assuming dominant inheritance model with 50% penetrance and assuming the disease locus is *unlinked*. There were only two data sets for which the adjusted score for *MBLOD* exceeded the value 3, leading to a measured type I error rate of 0.0002 in this sample, far below the requirement. *MMLS-C* and *NPL* erred only once.

7 Discussion

In this paper we defined the *MBLOD* method for genetic linkage analysis and tested it on simulated data sets. The results confirm that *MMLS-C* should be preferred over *NPL* in analyzing complex diseases. Furthermore, *MBLOD* provided higher power than *MMLS-C* for two-locus disease models (Tables 2 and 4, Figure 3), at the cost of larger run time. For three-locus and four-locus models tested, *MMLS-C* and *MBLOD* had similar (high) power.

The flat prior chosen for the penetrance α in the *MBLOD* method can be replaced by other prior distributions. A particularly appealing choice is a Beta prior $\mu(\alpha) \propto \alpha^{a-1}(1-\alpha)^{b-1}$, where $n_0 = a + b$ is the imaginary sample size and $a/(a+b)$ is the maximum of the Beta distribution. The Beta distribution is a special case of the Dirichlet distribution which serves as a conjugate prior for multinomial sampling. Choosing large values for n_0 sharpens the peak around the maximum and decreases the variance, reflecting higher confidence of the user regarding the prior penetrance α. Choosing large values for the hyper-parameter a and making b equal to a is asymptotically equivalent to asserting that $\alpha = 0.5$ with certainty. If the number of loci known to cause a complex disease is above a certain number c, one may want to place the maximum of $\mu(\alpha)$, say at c^{-1}, or another function of c, with a value for n_0 that depends on the evidence supporting the constant c. Such priors have not been incorporated nor tested so far. Furthermore, one may use more than one penetrance value for each inheritance model, in which case a more sophisticated choice of a Dirichlet prior can be introduced.

Finally, the performance of *MBLOD* should be further evaluated as a function of pedigree size and structure, number of markers and their degree of polymorphism, true location of the linked disease gene, and additional inheritance models.

Table 1: Average Z scores for different Generating Models (GMs), for one-marker analysis on nuclear family data, using three types of analysis methods. Standard deviations are given in parenthesis.

GM	*MBLOD*	*MMLS-C*	*NPL*
Additive2:	**4.01**(1.86)	3.65(1.64)	2.35(1.13)
Additive3	**1.74**(1.27)	1.25(1.06)	1.54(0.91)
Epistatic (RR)	**4.35**(1.69)	3.97(1.54)	2.46(1.18)
Epistatic (DD)	**2.47**(1.27)	2.09(1.10)	2.13(1.00)
Epistatic (DR)	**3.80**(1.75)	3.51(1.50)	2.21(0.93)
Epistatic (RD)	**3.16**(1.51)	2.97(1.36)	2.34(0.98)
Complementation (R+R)	**3.91**(1.81)	3.66(1.64)	2.31(1.10)
Complementation (D+D)	**2.21**(1.52)	1.50(1.15)	1.90(1.11)
Suppression (R $\oplus$ R)	**4.00**(1.77)	3.80(1.62)	2.39(1.08)
$\mathbf{R^3}$	**5.28**(1.63)	4.81(1.41)	2.35(0.96)
$\mathbf{R^4}$	**5.40**(1.71)	4.92(1.49)	2.47(1.10)
$\sum_3 \mathbf{R}$	**5.21**(1.82)	4.77(1.54)	2.50(0.97)
$\sum_4 \mathbf{R}$	**5.38**(1.59)	5.03(1.39)	2.53(1.02)

Table 2: Power to achieve a given Z value, for one-marker analysis on nuclear family data.

	Z = 2.0			Z = 3.0			Z = 4.0		
GM	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*
Add2	**0.87**	0.83	0.56	**0.65**	0.60	0.24	**0.47**	0.42	0.06
Add3	**0.37**	0.24	0.31	**0.20**	0.07	0.06	**0.04**	0.02	0
RR	**0.93**	0.9	0.62	**0.77**	0.70	0.27	**0.54**	0.48	0.08
DD	**0.6**	0.51	0.48	**0.35**	0.19	0.20	**0.14**	0.05	0.07
DR	**0.84**	0.82	0.56	**0.66**	0.63	0.18	**0.47**	0.38	0.05
RD	0.74	**0.75**	0.63	**0.54**	0.47	0.24	**0.32**	0.26	0.08
R+R	**0.87**	0.85	0.53	**0.66**	0.62	0.18	**0.43**	0.35	0.07
D+D	**0.49**	0.29	0.41	**0.30**	0.10	0.17	**0.12**	0.06	0.04
R$\oplus$R	**0.86**	0.85	0.62	**0.70**	0.68	0.22	**0.45**	0.44	0.08
$\mathbf{R^3}$	**0.99**	0.98	0.56	**0.91**	0.90	0.24	**0.77**	0.69	0.06
$\mathbf{R^4}$	**0.97**	**0.97**	0.61	**0.92**	0.91	0.26	**0.8**	0.71	0.07
$\sum_3 \mathbf{R}$	0.95	**0.96**	0.67	**0.88**	0.86	0.25	**0.76**	0.7	0.06
$\sum_4 \mathbf{R}$	**0.99**	**0.99**	0.64	0.92	**0.93**	0.29	**0.79**	0.78	0.11

Table 3: Average Z scores for different Generating Models, using three types of analysis methods. Standard deviations are given in parenthesis. The first 3 columns refer to one-marker analysis on 3-generation family data. The next 3 columns refer to five-marker analysis on nuclear family data.

	Experiment 2			Experiment 3		
GM	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*
Epistatic (RR)	**3.53**(1.72)	3.08(1.50)	1.90(0.96)	**8.72**(2.27)	7.80(1.83)	4.66(1.22)
Epistatic (DD)	1.45(1.17)	1.32(1.04)	**1.57**(0.84)	**6.21**(1.74)	4.30(1.65)	4.53(1.23)
Epistatic (DR)	**3.59**(1.65)	3.19(1.42)	1.84(0.78)	**5.52**(2.15)	5.21(1.94)	4.62(1.35)
Epistatic (RD)	**2.46**(1.41)	2.19(1.24)	1.75(0.95)	**4.18**(1.82)	3.81(1.70)	3.20(0.88)
Complementation (R+R)	**1.99**(1.32)	1.69(1.30)	1.36(1.07)	**8.83**(2.11)	7.92(1.82)	4.79(1.23)
Suppression (R$\oplus$R)	**2.26**(1.41)	1.93(1.29)	1.35(0.88)	**8.43**(1.90)	7.57(1.74)	**8.43**(1.90)

Table 4: Power to achieve a given Z value, for five-marker analysis on nuclear family data.

	Z = 2.0			Z = 3.0			Z = 4.0		
GM	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*
RR	0.99	**1.0**	**1.0**	0.98	**0.99**	0.93	**0.97**	**0.97**	0.67
DD	**0.99**	0.9	**0.99**	**0.97**	0.81	0.91	**0.9**	0.59	0.62
DR	0.94	0.94	**0.99**	0.9	0.88	**0.91**	**0.85**	0.78	0.69
RD	0.87	0.86	**0.91**	**0.7**	0.68	0.61	**0.56**	0.41	0.2
R+R	0.99	**1.0**	**1.0**	**0.99**	**0.99**	0.96	**0.99**	0.98	0.69
R⊕R	**1.0**	**1.0**	0.98	**1.0**	**1.0**	0.91	**0.98**	0.97	0.71

Acknowledgements

This research was supported by the Israel Science Foundation. We thank Anna Tzemach for computer assistance.

References

Abreu, P., Greenberg, D. A. and Hodge, S. E. (1999). Direct power comparisons between simple lod scores and npl scores for linkage analysis in complex diseases., *Am J. Hum Genet* **65**(3): 847–857.

Barrett, S., Beck, J., Bernier, R., Bisson, E., Braun, T., Casavant, T., Childress, D., Folstein, S., Garcia, M., Gardiner, M., Gilman, S., Haines, J., Hopkins, K., Landa, R., Meyer, N., Mullane, J., Nishimura, D., Palmer, P., Piven, J., Purdy, J., Santangelo, S., Searby, C., Sheffield, V., Singleton, J., Slager, S., Struchen, T., Svenson, S., Vieland, V., Wang, K. and Winklosky, B. (2001). An autosomal genomic screen for autism., *Am J. Hum Genet* **105**(8): 609–615.

Cavalli-Sforza, L. L. and Bodmer, W. F. (1971). *The genetics of human populations.*, Freeman., San Francisco.

Durner, M., Vieland, V. and Greenberg, D. A. (1999). Further evidence for the increased power of lod scores compared with nonparametric methods., *Am J. Hum Genet* **64**(1): 281–289.

Ethier, S. N. and Hodge, S. E. (1985). Identity-by-descent analysis of sibship configurations., *Am J. Hum Genet* **22**: 263–272.

Fishelson, M. and Geiger, D. (2002). Exact genetic linkage computations for general pedigrees., *Bioinformatics* **18**(Suppl. 1): S189–S198.

Friedman, N., Geiger, D. and Lotner, N. (2000). Likelihood computation with value abstraction., *Proc. 16th Conf. on Uncertainty in Artificial Intelligence (UAI)*.

Gabriel, B. S., Salomon, R., Pelet, A., Angrist, M., Amiel, J., Fornage, M., Atti-Bitach, T., Olson, J. M., Hofstra, R., Buys, C., Steffann, J., Munnich, A., Lyonnet, S. and Chakravarti, A. (2002). Segregation at three loci explains familial and population risk in hirschsprung disease., *Nat. Genet* **31**(1): 89–93.

Greenberg, D. A. and Abreu, P. (2001). Determining trait locus position from multipoint analysis: accuracy and power of three different statistics., *Genet Epidemiol* **21**(4): 299–314.

Greenberg, D. A., Abreu, P. and Hodge, S. E. (1998). The power to detect linkage in complex disease by means of simple lod-score analyses., *Am J. Hum Genet* **63**(3): 870–879.

Hodge, S. E., Abreu, P. and Greenberg, D. A. (1997). Magnitude of type i error when single-locus linkage analysis is maximized over models: a simulation study., *Am J. Hum Genet* **60**(4): 217–227.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach., *Am. J. Hum. Genet.* **58**(6): 1347–1363.

Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis.*, Springer-Verlag, New York.

Ott, J. (1999). *Analysis of Human Genetic Linkage.*, Johns Hopkins University Press, Baltimore, Maryland.

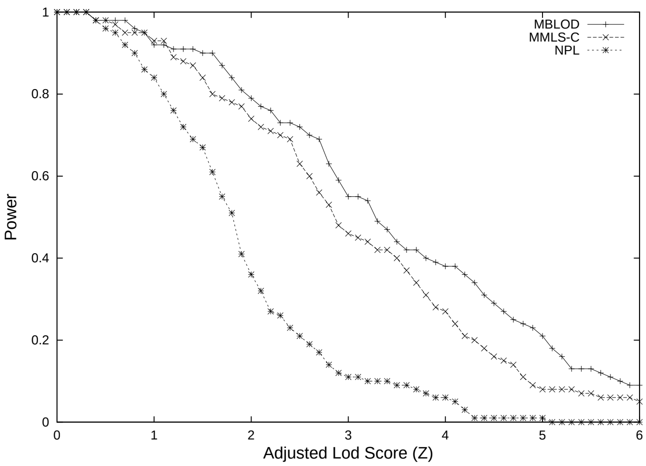
Silverman, E., Mosley, J., Palmer, L., Barth, M., Senter, J., Brown, A., Drazen, J., Kwiatkowski, D., Chapman, H., Campbell, E., Province, M., Rao, D., Reilly, J., Ginns, L., Speizer, F. and Weiss, S. (2002). Genome-wide linkage analysis of severe, early-onset chronic obstructive pulmonary disease: airflow obstruction and chronic bronchitis phenotypes., *Hum Mol Genet* **11**(6): 623–632.

Suarez, B. K. and Van Eerdewegh, P. (1984). A comparison of three affected-sib-pair scoring methods to detect hla-linked disease susceptibility genes., *Am J. Hum Genet* **18**: 135–146.

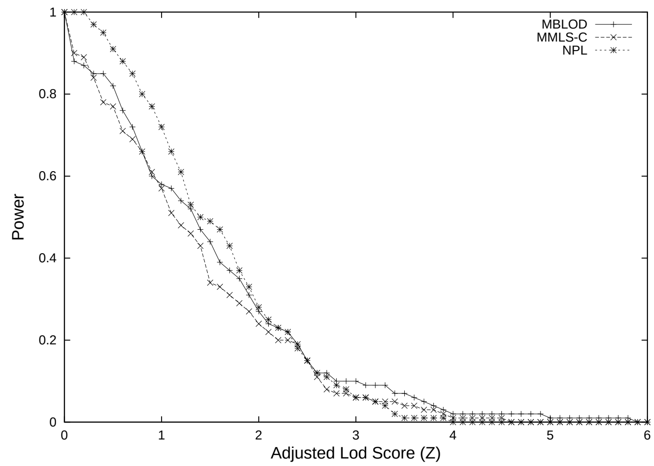
Terwilliger, J. D. and Ott, J. (1994). *Handbook of Human Genetic Linkage.*, Johns Hopkins University Press, Baltimore, Maryland.

Vieland, J. V. (1998). Bayesian linkage analysis, or: How i learned to stop worrying and love the posterior probability of linkage., *Am J. Hum Genet* **63**(4): 947–954.

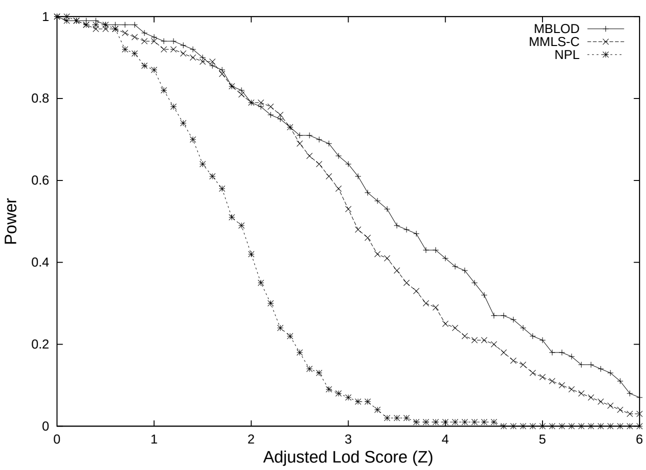
Weeks, D. E. and Lange, K. (1988). The affected-pedigree member method of linkage analysis., *Am J. Hum Genet* **42**(2): 315–326.



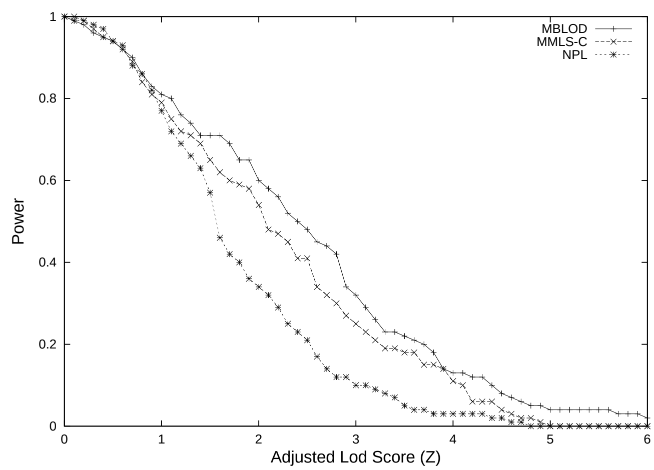
(a) RR Model



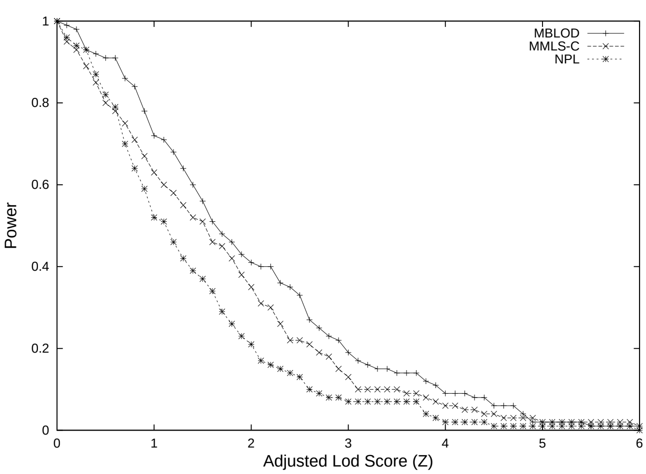
(b) DD Model



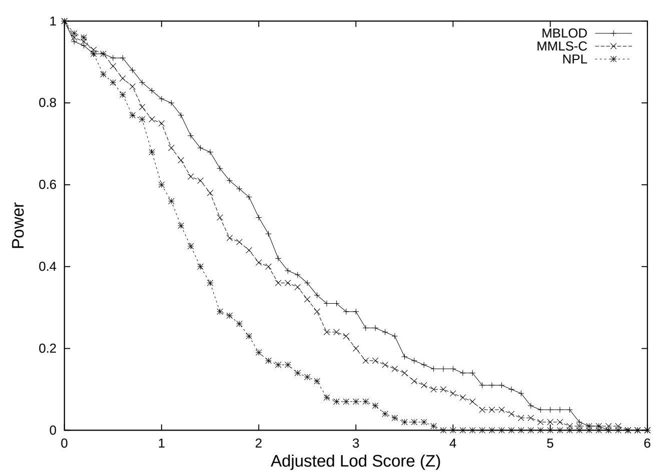
(c) DR Model



(d) RD Model



(e) R+R Model



(f) R$\oplus$R Model

Figure 3: Power curves to achieve a given Z value for the Epistatic RR, Epistatic DD, Epistatic DR, Epistatic RD, Complementation R+R, and Suppression R$\oplus$R Models, for one-marker analysis on 3-generation family data. In all models, *MBLOD* shows the highest power for $Z \geq 3$.