Robustness and Power of the Maximum-Likelihood–Binomial and Maximum-Likelihood–Score Methods, in Multipoint Linkage Analysis of Affected-Sibship Data

Laurent Abel[1] and Bertram Müller-Myhsok[2]

[1]INSERM U.436, Mathematical and Statistical Modeling in Biology and Medicine, Hôpital Pitié-Salpêtrière, Paris; and [2]Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Hamburg

Summary

The maximum-likelihood–binomial (MLB) method, based on the binomial distribution of parental marker alleles among affected offspring, recently was shown to provide promising results by two-point linkage analysis of affected-sibship data. In this article, we extend the MLB method to multipoint linkage analysis, using the general framework of hidden Markov models. Furthermore, we perform a large simulation study to investigate the robustness and power of the MLB method, compared with those of the maximum-likelihood–score (MLS) method as implemented in MAPMAKER/SIBS, in the multipoint analysis of different affected-sibship samples. Analyses of multiple-affected sibships by means of the MLS were conducted by consideration of all possible sib pairs, with (weighted MLS [MLSw]) or without (unweighted MLS [MLSu]) application of a classic weighting procedure. In simulations under the null hypothesis, the MLB provided very consistent type I errors regardless of the type of family sample (sib pairs or multiple-affected sibships), as did the MLS for samples with sib pairs only. When samples included multiple-affected sibships, the MLSu led to inflation of low type I errors, whereas the MLSw yielded very conservative tests. Power comparisons showed that the MLB generally was more powerful than the MLS, except in recessive models with allele frequencies <.3. Missing parental marker data did not strongly influence type I error and power results in these multipoint analyses. The MLB approach, which in a natural way accounts for multiple-affected sibships and which provides a simple likelihood-ratio test for linkage, is an interesting alternative for multipoint analysis of sibships.

Received December 17, 1997; accepted for publication June 2, 1998; electronically published July 6, 1998.

Address for correspondence and reprints: Dr. Laurent Abel, INSERM U.436, Hôpital Pitié-Salpêtrière, 91 Bd de l'Hôpital, 75013 Paris, France. E-mail: abel@biomath.jussieu.fr

0002-9297/98/6203-0039$02.00

Introduction

Affected-sib-pair linkage studies are a very common design for the search for genetic components involved in complex traits. To analyze these data, a widely used approach is the maximum-likelihood–score (MLS) method proposed by Risch (1990*b*), which has been implemented in popular software packages such as MAPMAKER/SIBS (Kruglyak and Lander 1995), allowing performance of multipoint MLS analyses. Several approaches have been proposed for use of the MLS when affected-sib-pair samples include sibships with more than two affecteds (multiple-affected sibships). A simple strategy is to deconstruct the multiple-affected sibship into all possible constitutive sib pairs. However, this approach can lead to overestimation of significance levels, as was pointed out by Daly and Lander (1996) with regard to a linkage study on non–insulin-dependent (type 2) diabetes (Hanis et al. 1996). Another strategy for use of the MLS implemented in MAPMAKER/SIBS is to weight the LOD score of each sib pair by 2/*S*, where *S* is the total number of affected sibs from which the sib pair comes, but this approach was shown to provide very conservative tests (Meunier et al. 1997).

An alternative method for analysis of multiple-affected sibships, the "maximum-likelihood–binomial (MLB) method," recently has been studied in a two-point linkage analysis with complete parental marker data and has shown very consistent type I errors and good power performances when mixtures of sibships with different numbers of affected sibs were analyzed (Abel et al. 1998). The MLB method, which is based on the binomial distribution of parental alleles among affected offspring (Badner et al. 1984; Majumder and Pal 1987), takes into account, in a natural way, multiple-affected sibships and provides a simple likelihood-ratio test involving a single parameter. Furthermore, in an analysis of sibships with two affecteds, the MLB method was shown to be more powerful than the classic mean test, when a common asymptotic type I error was used (Abel et al. 1998).

The first goal of this article is to extend the MLB method to multipoint analysis, by means of hidden Markov chain models (Lander and Green 1987; Kruglyak and Lander 1995; Kruglyak et al. 1996). Then, a large simulation study is discussed, to study the type I errors and the power of both the MLS and MLB methods in the multipoint analysis of family data including (1) affected sib pairs only, (2) sibships with four affected sib pairs, and (3) a mixture of affected sib pairs and multiple-affected sibships. We also studied the influence of missing parental marker data on the analyses.

Methods

The MLS Method

The MLS method was devised originally by Risch (1990*b*) and is described briefly in the following discussion. For a given location along a chromosome, let z_0, z_1, and z_2 denote the probabilities that a sib pair shares 0, 1, or 2 alleles identical by descent, respectively; obviously, $z_0 + z_1 + z_2 = 1$. The test for linkage is expressed as a LOD score comparing the likelihood of the marker data maximized over (z_0, z_1, z_2) with the likelihood of the marker data under the null hypothesis (H_0); that is, $(z_0, z_1, z_2) = (.25, .5, .25)$. The properties of the MLS have been studied extensively by Holmans (1993), who showed that the power of the method could be improved by restricting maximization to genetically possible models only. The resulting asymptotic distribution is a mixture of χ^2 with 1 or 2 df, and this restricted test has a size of ~.001 for a $\log_{10}$ likelihood-ratio criterion of ~2.3 (Holmans 1993). The MLS test has been implemented in popular software packages such as MAPMAKER/SIBS (Kruglyak and Lander 1995), which allows performance of an efficient multipoint analysis and which was the software used in this study. Several options are available in MAPMAKER/SIBS, to take into account multiple-affected sibships: (1) consideration of all possible sib pairs and weighting of the LOD score of each sib pair by 2/*S,* referred to in this article as "weighted MLS" ("MLSw"); (2) consideration of all possible sib pairs, without any weighting, referred to as "unweighted MLS" ("MLSu"); (3) consideration of only independent sib pairs (the first sib pair or the sib pairs created by taking the first sib with other sibs), which was not studied extensively in the simulation because it led to serious decreases in power in our initial results (data not shown). For family samples including affected sib pairs only, all options are identical and will be referred to as "MLS," whereas samples containing multiple-affected sibships were analyzed by both the MLSu and the MLSw approaches.

The MLB Method

The likelihood of the marker data of *S* affected children can be expressed under H_0 as a product, over *j* (the sibships) and *k* (the parents), of binomial distribution (S_j, .5), where S_j is the number of affected offspring in sibship *j* (Majumder and Pal 1987; Abel et al. 1998). A simple linkage test then can be constructed by assessing the departure from .5 of the probability parameter of these binomial distributions among affected sibs who have received the same marker allele (Abel et al. 1998). When this parameter is denoted as α, the likelihood of the sample, $L(\alpha)$, can be written as $L(\alpha) = \prod_{jk} L_{jk}(\alpha)$, where $L_{jk}(\alpha)$ denotes the likelihood of a sibship in which the distribution of the number of affected sibs who have received allele A from a heterozygous AB parent, denoted as n_A, is binomial (S_j, α). Taking into account unknown phase matings, the relevant part of the likelihood depending on α, denoted as $f_{jk}(\alpha)$, can be expressed as follows (Abel et al. 1998):

$$f_{jk}(\alpha) = [\alpha^{n_A}(1-\alpha)^{S_j - n_A} + (1-\alpha)^{n_A}\alpha^{S_j - n_A}] \ .$$

Note that this expression is totally symmetric for α, and, before proceeding to the test itself, it is necessary to decide the meaning of α. The parameter α can be defined as either the probability that an affected sib has received the marker allele not transmitted with the disease allele or the probability that an affected sib has received the marker allele transmitted with the disease allele. In the first case, the appropriate alternative hypothesis is $\alpha < .5$, whereas in the second case it is defined as $\alpha > .5$. In either case, the test is one-sided, once the interpretation of α has been determined. In the discussion that follows, we arbitrarily have used the second, intuitively probably more appealing meaning of α—namely, the alternative hypothesis $\alpha > .5$. The corresponding test described below is one-sided, as already stated above. Let *a* be the maximum-likelihood estimator of α; therefore, the test for linkage is a standard likelihood-ratio statistic Λ:

$$\Lambda = 2\ln\left[\frac{L(\alpha = a)}{L(\alpha = .5)}\right] = 2\sum_{jk} \ln\left[\frac{f_{jk}(\alpha = a)}{f_{jk}(\alpha = .5)}\right] \ .$$

Λ asymptotically has a mixture distribution of .5 χ^2 (0 df) and .5 χ^2 (1 df); that is, $\Lambda^{1/2}$, denoted as Z_{MLB}, is a one-sided standard normal deviate. The reasoning for the 50% point mass at 0 is that the test is one-sided; that is, *a* is bounded at .5 when the unrestricted maximum is <.5, and the probability that this situation occurs is .5, under H_0. The test also can be expressed as a LOD-score criterion [equal to $\Lambda/2\ln(10)$], which, in this case, has the same distribution as a classic LOD score based on the estimation of the recombination fraction (θ).

There is a direct relationship between α and π, the proportion of alleles shared by the sib pairs. For sibships with $S = 2$, $\pi = 1 - 2\alpha(1 - \alpha)$, and a stronger result is shown in the study by Abel et al. (1998)—namely, that this equality holds regardless of the size of the sibship. By use of this relationship, for a sample of sibships with $S = 2$ (sib pairs only), the likelihood-ratio test is equivalent to the classic mean test, in the sense developed by Knapp et al. (1994); in particular, Λ and the mean test statistic are monotonally increasing with an increasing proportion of shared alleles. However, this equivalence does not necessarily imply the equality of the power of the tests derived from these two statistics, when a common asymptotic type I error is used. In appendix B of their study, Abel et al. (1998) demonstrate that, for a given value of π, Λ is always greater than the corresponding χ^2 of the mean test, indicating that the MLB test is expected to be more powerful than the mean test when a common asymptotic type I error is used.

The multipoint extension of the MLB method uses the general framework of hidden Markov models developed previously in this context (Lander and Green 1987; Kruglyak and Lander 1995; Kruglyak et al. 1996), and we will use notations from the article describing the GENEHUNTER program (Kruglyak et al. 1996). For the case of nuclear families, there are two founders (the parents) and n nonfounders, including S affected and $n - S$ unaffected children. Let $\boldsymbol{v}(x)$ be the inheritance vector at point x of the genome $\boldsymbol{v}(x) = (p_1, m_1,..., p_n, m_n)$, where p_i and m_i are binary variables (0; 1) indicating the outcomes of paternal and maternal meioses. The set of the 2^{2n} possible inheritance vectors is denoted V, and the probability distribution of all inheritance vectors given the marker data at any point of the genome, which corresponds to $P[\boldsymbol{v}(x) = \boldsymbol{w}]\forall \boldsymbol{w} \in V$, computed by use of hidden Markov models (Lander and Green 1987; Kruglyak and Lander 1995; Kruglyak et al. 1996). For a given inheritance vector $\boldsymbol{w}$, the relevant part of the likelihood in the MLB approach, for sibship j,—denoted as $f_{jk}(\alpha; \boldsymbol{w})$—becomes

$$f_{jk}(\alpha;\boldsymbol{w}) = \alpha^{n_{1k}(\boldsymbol{w})}(1 - \alpha)^{S_j - n_{1k}(\boldsymbol{w})} + (1 - \alpha)^{n_{1k}(\boldsymbol{w})}\alpha^{S_j - n_{1k}(\boldsymbol{w})} ,$$

where $n_{1k}(\boldsymbol{w})$ is the number of affected sibs for whom $p_i = 1$ ($k = 1$, father) or $m_i = 1$ ($k = 2$, mother), and the likelihood of the whole sample will be denoted as $L(\alpha; \boldsymbol{w})$. Note that when $\alpha = .5$, $f_{jk}(\alpha; \boldsymbol{w})$ is independent of $\boldsymbol{w}$ and is equal to $.5^{S_j} + .5^{S_j}$.

The likelihood at location x is obtained by summing over all possible inheritance vectors, and, for the whole sample, the multipoint likelihood-ratio test at position x, $\Lambda(x)$, is computed as

$$\Lambda(x) = 2\ln\left\{\frac{\sum_{w\in V} L(\alpha = a;\ \boldsymbol{w})P[\boldsymbol{v}(x) = \boldsymbol{w}]}{\sum_{w\in V} L(\alpha = .5;\ \boldsymbol{w})P[\boldsymbol{v}(x) = \boldsymbol{w}]}\right\}$$

$$= 2\sum_{jk}\ln\left\{\frac{\sum_{w\in V} f_{jk}(\alpha = a;\ \boldsymbol{w})P[\boldsymbol{v}(x) = \boldsymbol{w}]}{\sum_{w\in V} f_{jk}(\alpha = .5;\ \boldsymbol{w})P[\boldsymbol{v}(x) = \boldsymbol{w}]}\right\} ,$$

which has the same distribution as described above. Note that the denominator of $\Lambda(x)$ is independent of the probability distribution of the inheritance vectors, since $\alpha = .5$. This expression has the same form as the likelihood-ratio statistic $\overline{LR}(x)$ described by Kruglyak et al. (1996) for parametric linkage analysis. What is different in this article is the use of a nonparametric binomial approach to express the likelihood of a sibship. A procedure for MLB calculations, in programming language C, was developed and linked to the GENEHUNTER program (Kruglyak et al. 1996), which provides the distribution of inheritance vectors, given marker data for any location in the genome map considered for the analysis.

Simulation Study

Simulation Settings

Genetic model.—Simulations were conducted to investigate the type I error and the power of the MLS and MLB methods in the analysis of multipoint data. To generate the family data, we considered a disease locus G with alleles D and d with population frequencies q and $1 - q$, respectively, and three penetrances f_{DD}, f_{Dd}, and f_{dd}, corresponding to the three possible genotypes. Three dominance effects were considered for allele D: recessive ($f_{Dd} = f_{dd}$), additive [$f_{Dd} = (f_{dd} + f_{DD})/2$], and dominant ($f_{DD} = f_{Dd}$). The overall prevalence of the disease, K, generally was fixed at .05, but, to assess the influence of prevalence on the power of the methods, we also considered K values of .02 and .10. Given K and the dominance effect of D, penetrances were computed for different values of q and λ_S, the sibling recurrence-risk ratio (Risch 1990*a*). Relationships between penetrances, genotypic relative risks, and the probability that an affected child has received a D allele from a Dd parent have been described elsewhere (Abel et al. 1998). The penetrances corresponding to all genetic models that were considered in the simulation study, according to prevalence, λ_S value, and dominance effects, are shown in table 1. We also introduced genetic heterogeneity in the simulation model, so that the disease was controlled by locus G in a proportion β of families, whereas it was due to another locus (not linked to G) with the same characteristics as G in a proportion $1 - \beta$ of families. Note that, in this case, the sibling relative recurrence risk

Table 1

Genetic Models Considered in the Simulation Study, According to *K*, λ_S, and the Type of Dominance Model

			Penetrance		
Model and *K*	λ_S	Frequency of D Allele	f_{DD}	f_{Dd}	f_{dd}
Dominant:					
.05	2	.005	.755	.755	.043
.05	2	.01	.547	.547	.040
.05	2	.05	.266	.266	.027
.05	2	.10	.198	.198	.015
.05	1.5	.01	.401	.401	.043
.05	2.5	.01	.659	.659	.038
.05	3	.01	.753	.753	.036
.10	2	.05	.533	.533	.053
.02	2	.05	.107	.107	.011
Additive:					
.05	2	.02	.750	.393	.036
.05	2	.05	.486	.256	.027
.05	2	.10	.350	.183	.017
.05	2	.20	.250	.125	.000
.05	1.5	.05	.358	.196	.034
.05	2.5	.05	.584	.303	.022
.05	3	.05	.666	.342	.018
.10	2	.10	.700	.367	.033
.02	2	.10	.140	.073	.007
Recessive:					
.05	2	.10	.965	.041	.041
.05	2	.20	.474	.032	.032
.05	2	.30	.313	.024	.024
.05	2	.40	.233	.015	.015
.05	1.5	.20	.350	.037	.037
.05	2.5	.20	.570	.028	.028
.05	3	.20	.650	.025	.025
.10	2	.20	.948	.065	.065
.02	2	.20	.190	.013	.013

due specifically to locus G, denoted as $\lambda_S(G)$, is equal to $1 + \beta(\lambda_S - 1)$.

Genotype data were generated for a genetic map of 20 cM, with five markers spaced every 5 cM. Each of the marker loci had five equally frequent alleles. The disease locus G was located in the middle of the map, at $\theta = 0$, with marker 3. To compare the observed statistics to a known asymptotic distribution, we considered the values obtained for only one position, which was in the middle of the map and corresponded to the actual location of G. With respect to parental marker data, two situations were considered: either entirely known (complete data) or fully missing (incomplete data) parental genotypes. For the situation of incomplete data, analyses were performed with the correct marker allele frequency, .20.

Generation of families.—Monte Carlo methods were used to simulate data from nuclear families with a number of children per sibship, following the distribution of sibship sizes provided by Speer et al. (1995). Genotypes and affected status were assigned randomly under the different genetic models defined above. For these families, three types of family samples were considered. The first type (the 2AS sample) consisted of affected sib pairs only and included 100 or 200 families. The second type (the 4AS sample) included 30 families with four affected children, corresponding to 180 possible affected sib pairs. The last type (the MAS sample) was a mixture of sibships, with the number of affecteds varying from two to five, in a proportion close to that observed by Hanis et al. (1996): for 76% of the sibships, $S = 2$; for 16%, $S = 3$; for 6%, $S = 4$; and for 2%, $S = 5$. MAS samples of 50 and 100 families were simulated, corresponding to 90 and 180 possible affected sib pairs, respectively, and also including available unaffected children, up to a maximum of five sibs (affected and unaffected) per family.

Simulations under H_0

Simulations under H_0 were performed to explore the type I error of the tests. For these simulations, only one genetic model was considered (dominant with $q = .01$ and $\lambda_S = 3$), and 50,000 replicates of family samples were generated for each situation, according to the sampling scheme (2AS, 4AS, or MAS), the sample size, and the presence or absence of parental genotypic data. For the MLB, the observed .05, .001, and .0001 type I errors were the proportion of replicates that provided a value for Z_{MLB}, the one-sided standard normal deviate of the MLB statistic, above 1.645, 3.090, and 3.717, respectively. For the LOD-score statistic of the MLS, we used the asymptotic thresholds provided by Holmans (1993) for the case of a fully informative marker—namely, 0.742, 2.324, and 3.289 for .05, .001, and .0001 type I errors, respectively. The 95% confidence intervals of

Table 2

Results of the Simulation Study under H_0, for the 2AS Samples, by Use of 50,000 Replicates of the MLS and MLB Tests

Expected Type I Error and Test	Observed Type I Error, When Parental Data Complete		Observed Type I Error, When Parental Data Missing[a]	
	100 Sib Pairs	200 Sib Pairs	100 Sib Pairs	200 Sib Pairs
.05:				
MLB	.05084	.05068	.05092	.04940
MLS	.05156	.04966	.04552	.04726
.001:				
MLB	.00100	.00092	.00080	.00092
MLS	.00098	.00094	.00106	.00092
.0001:				
MLB	.00006	.00014	.00010	.00008
MLS	.00004	.00012	.00004	.00008

[a] Results obtained by use of correct allele frequencies.

Table 3

Results of the Simulation Study under H_0, for the MAS and 4AS Samples, by Use of 50,000 Replicates of the MLB, MLSw, and MLSu Tests

	Observed Type I Error, When Parental Data Complete			Observed Type I Error, When Parental Data Missing[a]		
Expected Type I Error and Test	30 4AS Families	50 MAS Families	100 MAS Families	30 4AS Families	50 MAS Families	100 MAS Families
.05:						
MLB	.04880	.04972	.05024	.04866	.04818	.04958
MLSw	.01144	.03538	.03384	.00858	.03276	.03290
MLSu	.05802	.05622	.05520	.04844	.05084	.04970
.001:						
MLB	.00090	.00086	.00110	.00080	.00110	.00104
MLSw	.00002	.00040	.00034	.00002	.00040	.00046
MLSu	.00234	.00194	.00168	.00130	.00156	.00146
.0001:						
MLB	.00010	.00004	.00010	.00004	.00016	.00008
MLSw	.00000	.00004	.00002	.00000	.00004	.00004
MLSu	.00034	.00034	.00018	.00008	.00034	.00026

[a] Results obtained by use of correct allele frequencies.

the .05, .001, and .0001 type I errors observed for 50,000 replicates are .0481–.0519, .00072–.00128, and .00001–.00019, respectively.

Results for the 2AS samples are shown in table 2. Both the MLB and MLS statistics provide very consistent results, regardless of whether the parents are genotyped. In particular, observed .001 and .0001 type I errors are always within their 95% confidence intervals, for both 100- and 200-family samples. The mean information content, as defined by Kruglyak et al. (1996), was .938 when parents were genotyped and .516 when parental data were missing. For the situation of 200 2AS families with missing parental data, we also performed the analysis using, for markers 1, 3, and 5, a misspecified vector of allele frequencies, (.35, .30, .15, .10, .10), whereas the correct vector, (.20, .20, .20, .20, .20), was used for markers 2 and 4. This misspecification led to an inflation of the type I error, for both methods, with an observed .001 type I error equal to .0021 and .0020 for the MLB and the MLS, respectively (data not shown). Table 3 presents the results for the MAS and 4AS samples. The mean information content was .952 and .956 when parents were genotyped and .710 and .755 when parental data were missing, for the MAS and 4AS samples, respectively. The MLB always yields observed type I errors within their 95% confidence intervals, regardless of family-sample size or of the presence or absence of marker parental data (when correct allele frequencies are used). The MLSu provides inflation of type I errors, especially for low type I errors and for the 50 MAS family samples. The MLSw leads to conservative type I errors, especially for the 4AS samples. This is an unexpected result, since, for a sample including sibships with the same number of affecteds, the weighted and the unweighted statistic are identical when the classic mean test is used (Suarez and Van Eerdewegh 1984; Abel et al. 1998). For this reason, only the MLSu was used for power comparisons with the MLB, for the MAS samples.

Simulations under the Alternative Hypothesis

These simulations were conducted to investigate the respective power of the different statistics. Family data were generated under the different genetic models indicated in table 1. Two heterogeneity levels were considered, with $\beta = .75$ or $\beta = .5$. Samples consisted of 200 2AS families or 100 MAS families. For each genetic model, 1,000 replicates of family samples were generated, and results are presented in terms of power for a .001 type I error—that is, the proportion of replicates that yielded a test value above a given .001 threshold. For the 2AS families, the asymptotic .001 thresholds defined above were used. For the MAS families, we used empirical .001 thresholds, with respect to the overestimated type I errors observed for the MLSu. From the 50,000 replicates of 100 MAS families simulated previously under H_0, the empirical .001 threshold was considered to be the value above which 50 values of test statistics were observed. For Z_{MLB}, the empirical .001 threshold was equal to 3.1210 and 3.1212 for samples with and those without parental genotypic data, respectively, and, for the LOD-score statistic of the MLSu, these values were 2.632 and 2.5104, respectively.

Results are shown, in figure 1, for 200 2AS families (75% of linked families), simulated under the 12 genetic models of table 1 with $\lambda_S = 2$ and a prevalence of .05. Power levels depend on q and the dominance model, with a range of 15%–80% for dominant models,

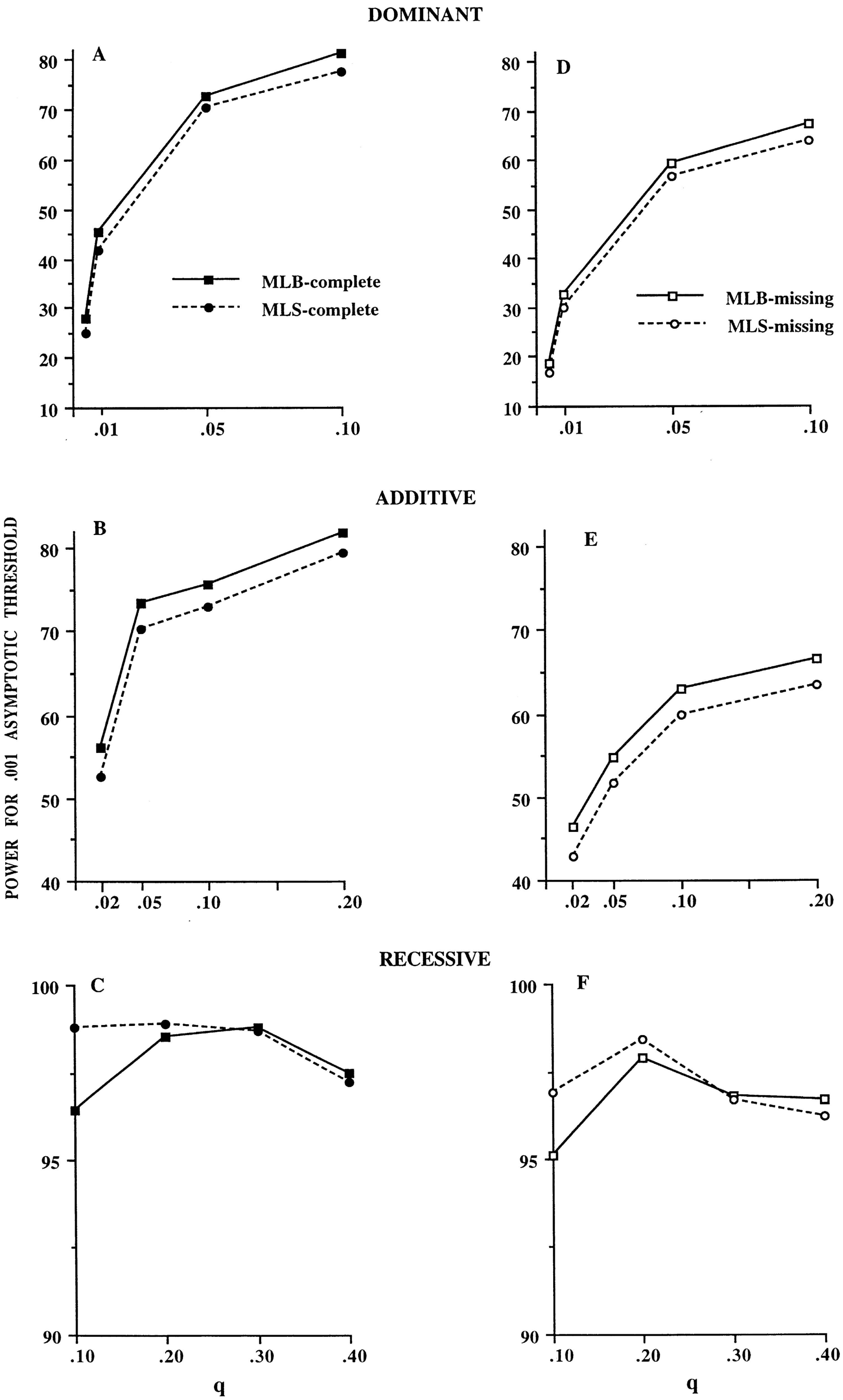

Figure 1 Power, in percentages, for an asymptotic .001 threshold observed with the MLB and the MLS, for 200 2AS families with parental marker data (*panels A–C*) and without parental marker data, when correct allele frequencies were used (*panels D–F*). The proportion of linked families was 75%, and data were generated under the genetic models in table 1 with $\lambda_S = 2$ and $K = .05$. *A and D,* Dominant model. *B and E,* Additive model. *C and F,* Recessive model.

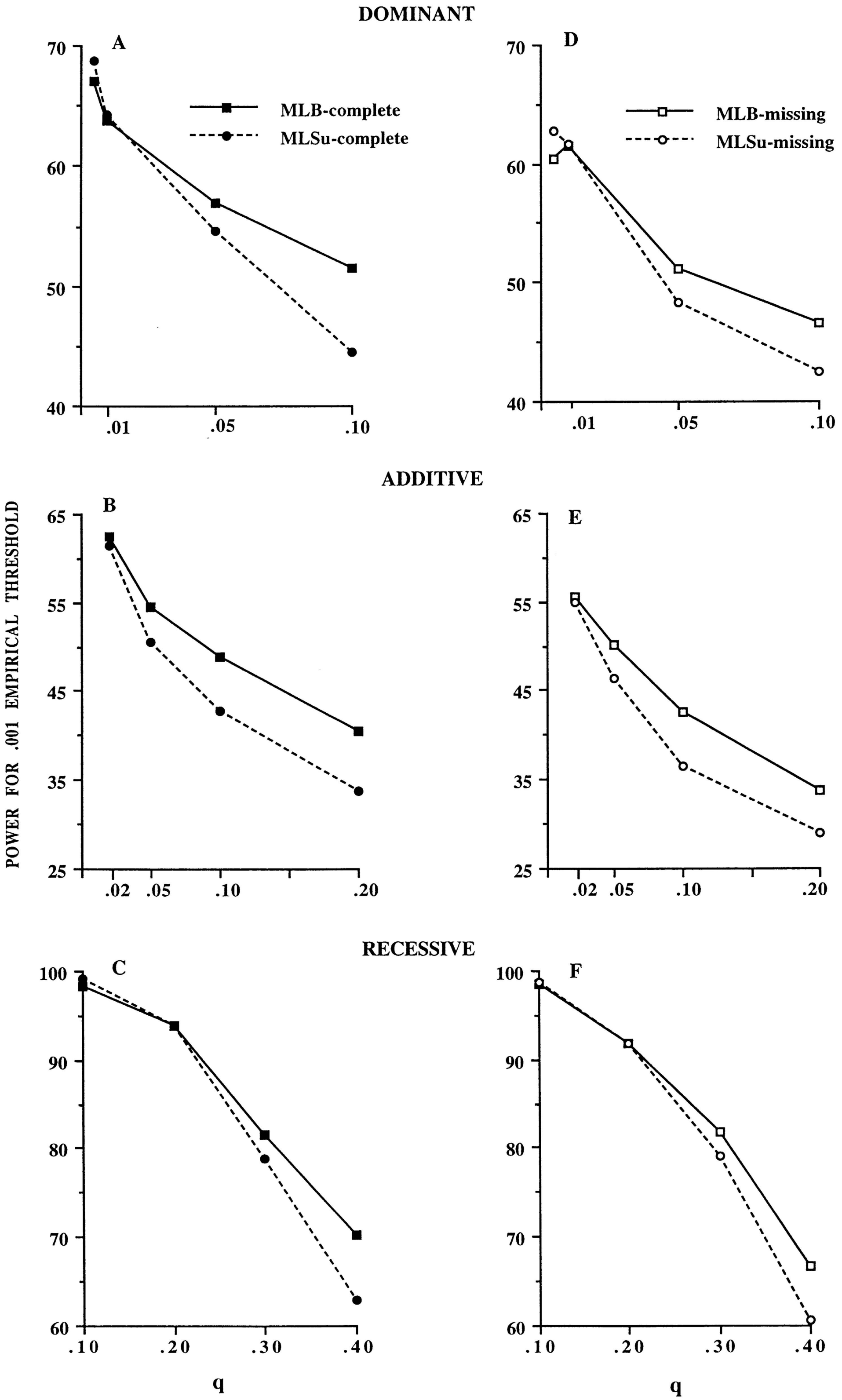

Figure 2 Power, in percentages, for an empirical .001 threshold observed with the MLB and the MLSu, for 100 MAS families with parental marker data (*panels A–C*) and without parental marker data, when correct allele frequencies were used (*panels D–F*). The proportion of linked families was 75%, and data were generated under the genetic models in table 1 with $\lambda_S = 2$ and $K = .05$. *A and D,* Dominant model. *B and E,* Additive model. *C and F,* Recessive model.

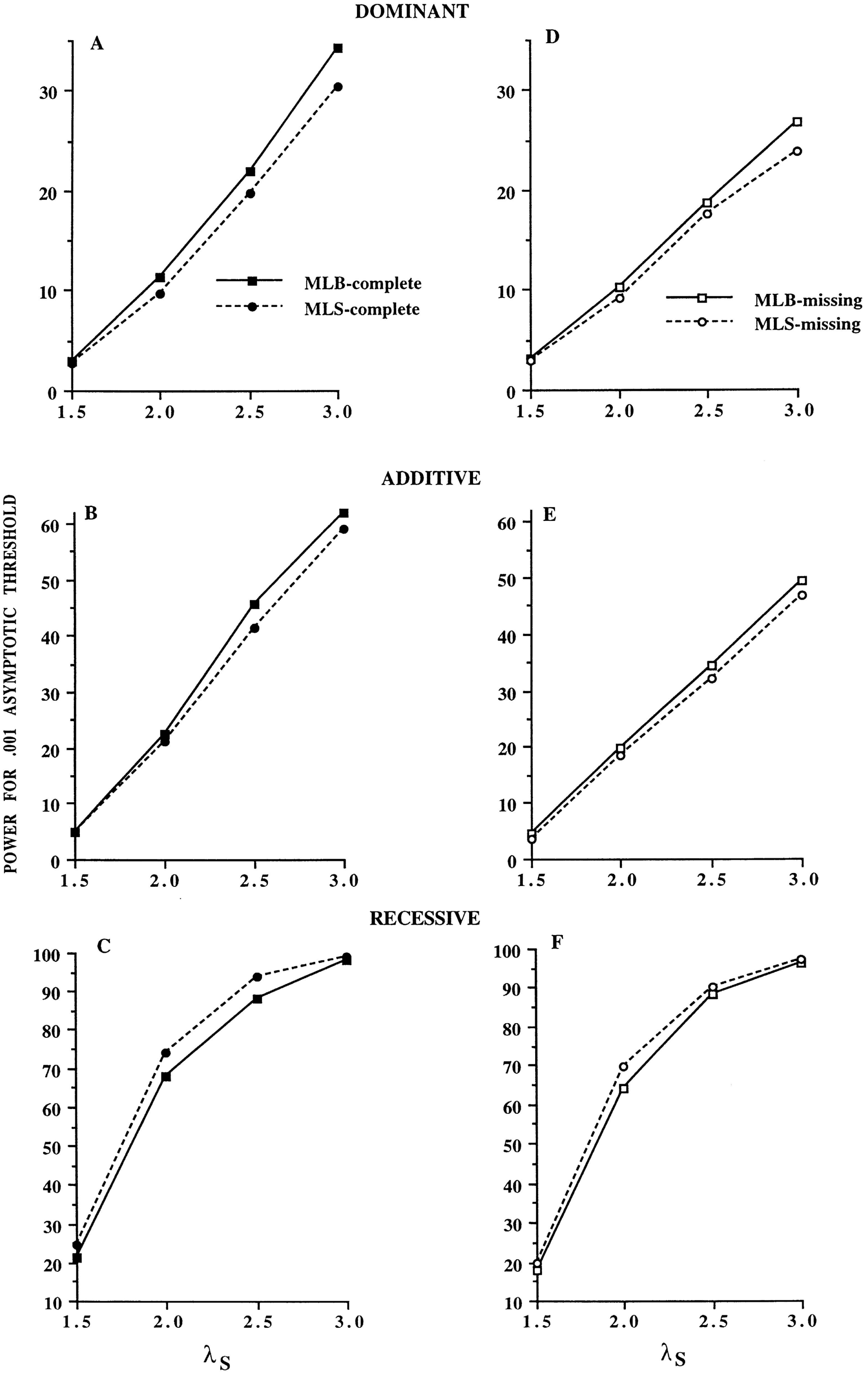

Figure 3 Power, in percentages, for an asymptotic .001 threshold observed with the MLB and the MLS, for 200 2AS families with parental marker data (*panels A–C*) and without parental marker data, when correct allele frequencies were used (*panels D–F*). The proportion of linked families was 50%, and data were generated for different λ_S values ($K = .05$). *A and D,* Dominant model with $q = .01$. *B and E,* Additive model with $q = .05$. *C and F,* Recessive model with $q = .2$.

Table 4

Power (for a .001 Type I Error) of the MLB and MLS Methods, for 200 2AS Families (with Genotyped Parents), under Three Genetic Models with λ_S = 2, According to *K*

	Power, for *K* = (%)		
Model (*q*) and Test	.02	.05	.10
Dominant (.05):			
MLB	73.8	72.5	67.2
MLS	69.5	70.4	63.3
Additive (.10):			
MLB	73.1	75.6	80.8
MLS	70.5	73.0	78.0
Dominant (.05):			
MLB	98.4	98.5	97.6
MLS	98.9	98.9	98.6

40%–80% for additive models, and >95% for recessive models. Except for recessive models with $q < .3$, the MLB statistic is slightly more powerful than the MLS statistic, regardless of whether the parents have been genotyped. Power studies also were performed for the 2AS samples with missing parental data, by use of the previously defined misspecified allele frequencies. In this case, empirical thresholds were used, since large inflation of .001 type I errors was observed under H_0, and power levels were very close to those obtained for samples analyzed with correct allele frequencies and asymptotic thresholds (data not shown). Under these 12 genetic models, the same pattern of results between the MLB and the MLSu was observed for 100 MAS families (fig. 2). The MLB outperforms the MLS statistic, except for a dominant model with $q < .01$ and a recessive model with $q < .2$. Similar results were obtained with 50% of linked families (data not shown); for this 50% heterogeneity level, figure 3 displays the results for different λ_S values for the 2AS samples. The MLB provides greater power than the MLS, for dominant (q = .01) and additive (q = .05) models, whereas the reverse was observed for a recessive (q = .2) model. Table 4 shows the power results obtained for 200 2AS families, when the overall prevalence was varied. Whereas the power decreases as the prevalence increases for a dominant model (and to a lower extent for a recessive model), the reverse is observed for an additive model. However, the differences observed between the MLB and MLS methods remain quite similar regardless of the prevalence values.

Discussion

The simulation studies under H_0 for the 2AS samples indicate that asymptotic thresholds can be used in this context, for both the MLB and MLS methods. For samples including multiple-affected sibships (4AS and MAS), the MLB yielded very consistent type I errors, confirming the results observed in two-point analysis with complete parental data (Abel et al. 1998). In MAS samples, the MLSu led to overestimated low type I errors, which requires the use of Monte Carlo methods to obtain reliable significance levels, as discussed elsewhere (Daly and Lander 1996; Kong et al. 1997). In contrast, the MLSw, as implemented in MAPMAKER/SIBS, provides very conservative tests, as has been noted elsewhere for two-point analyses (Meunier et al. 1997). In particular, results observed for the 4AS samples were unexpected, since, for a sample including sibships with the same number of affecteds, the weighted and the unweighted statistic are identical when the classic mean test is used (Suarez and Van Eerdewegh 1984; Abel et al. 1998). We also noted that, in our multipoint analyses performed with correct allele frequencies, missing parental data did not strongly influence the results obtained under H_0, whereas they have been shown to lead to a decrease of MLS type I errors in two-point analysis (Meunier et al. 1997). However, we noted a large inflation of type I errors for the MLB and MLS methods, when incorrect allele frequencies were considered in this study, and further studies will investigate more extensively the influence of misspecifying these frequencies on the robustness of the tests.

Power comparisons between the MLB and the MLS led to similar results regardless of the type of family sample (2AS or MAS), the heterogeneity level, or whether parental marker data were present or absent; that is, the MLB statistic was slightly more powerful than the MLS statistic, except for recessive models with $q < .3$ and dominant models with $q < .01$. We also noted that missing parental data led to only a small power decrease, especially for recessive models, in the multipoint analyses performed with correct allele frequencies, an observation consistent with the results obtained by Holmans (1993). The main difference between the 2AS and MAS samples is in the influence of q, the frequency of the deleterious allele, on the power of the statistics. For the 2AS samples, the power increases with an increase of q, whereas the reverse is observed for the MAS samples. These results are explained by the difference in the proportion of parents heterozygous for the deleterious allele, according to the number of affected sibs and q (Abel et al. 1998), and emphasize the importance of analyzing samples including a mixture of sibships of different sizes, in order to cover a large range of disease allele frequencies. Finally, it is interesting to note that, for a same λ_S value (e.g., λ_S = 2), there are large differences in power results, according to q and the dominance effect of the deleterious allele, with the higher power being observed in recessive models.

The multipoint extension of the method was performed easily by use of the general framework of hidden Markov models developed previously in this context

(Lander and Green 1987; Kruglyak and Lander 1995; Kruglyak et al. 1996). This development also could be applied to the MLB method recently proposed for linkage analysis of quantitative traits (Alcais and Abel 1997). Studies of extensions that take into account genetic heterogeneity are ongoing. It is, for example, straightforward to include and to consider additional parameters according to some measured factor (e.g., sex of parent). The MLB approach, which in a natural way accounts for multiple-affected sibships and leads to a simple likelihood-ratio test for linkage, which provides very consistent type I errors and good power performances, appears to be a quite interesting alternative method for multipoint linkage analysis of sib-pair data.

Acknowledgments

We thank Namik Taright for helpful assistance in modifying the GENEHUNTER program. The modified GENEHUNTER program that computes the MLB test is available, by request, from the authors (e-mail: abel@biomath.jussieu.fr or bmm@bni.uni-hamburg.de).

References

Abel L, Alcais A, Mallet A (1998) Comparison of four sib-pair linkage methods for analyzing sibships with more than two affecteds: interest of the binomial maximum likelihood approach. Genet Epidemiol 15:371–390

Alcais A, Abel L (1997) A maximum likelihood binomial method for nonparametric linkage analysis of quantitative traits in sibships. Paper presented at the 47th annual meeting of the American Society of Human Genetics, Baltimore, October 28–November 1

Badner JA, Chakravarti A, Wagener DK (1984) A test of nonrandom segregation. Genet Epidemiol 1:329–340

Daly MJ, Lander ES (1996) The importance of being independent: sib pair analysis in diabetes. Nat Genet 14:131–132

Hanis CL, Boerwinkle E, Chakraborty R, Ellsworth DL, Concannon P, Stirling B, Morrisson VA, et al (1996) A genome-wide search for human non–insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nat Genet 13:161–166

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374

Knapp M, Seuchter SA, Baur MP (1994) Linkage analysis in nuclear families. 2. Relationship between affected sib-pair tests and lod-score analysis. Hum Hered 44:44–51

Kong A, Frigge M, Bell Gi, Lander ES, Daly M, Cox NJ (1997) Diabetes, dependence, asymptotics, selection and significance. Nat Genet 17:148

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Lander ES, Green P (1987) Construction of multilocus genetic maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Majumder PP, Pal N (1987) Nonrandom segregation: uniformly most powerful test and related considerations. Genet Epidemiol 4:277–287

Meunier F, Philippi A, Martinez M, Demenais F (1997) Affected sib-pair tests for linkage: type I errors with dependent sib-pairs. Genet Epidemiol 14:1107–1111

Risch N (1990*a*) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228

——— (1990*b*) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253

Speer MC, Terwilliger JD, Ott J (1995) Data simulation for GAW9 problems 1 and 2. Genet Epidemiol 12:561–564

Suarez BK, Van Eerdewegh P (1984) A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. Am J Med Genet 18:135–146