Am. J. Hum. Genet. 56:519–527, 1995

Rapid Multipoint Linkage Analysis of Recessive Traits in Nuclear Families, Including Homozygosity Mapping

Leonid Kruglyak,[1] Mark J. Daly,[1] and Eric S. Lander[1,2]

[1] Whitehead Institute for Biomedical Research and [2]Department of Biology, Massachusetts Institute of Technology, Cambridge

Summary

Homozygosity mapping is a powerful strategy for mapping rare recessive traits in children of consanguineous marriages. Practical applications of this strategy are currently limited by the inability of conventional linkage analysis software to compute, in reasonable time, multipoint LOD scores for pedigrees with inbreeding loops. We have developed a new algorithm for rapid multipoint likelihood calculations in small pedigrees, including those with inbreeding loops. The running time of the algorithm grows, at most, linearly with the number of loci considered simultaneously. The running time is not sensitive to the presence of inbreeding loops, missing genotype information, and highly polymorphic loci. We have incorporated this algorithm into a software package, MAPMAKER/HOMOZ, that allows very rapid multipoint mapping of disease genes in nuclear families, including homozygosity mapping. Multipoint analysis with dozens of markers can be carried out in minutes on a personal workstation.

Introduction

Since the early observations of Garrod (1902), human geneticists have recognized that rare recessive traits appear in children of consanguineous marriages more often than in the general population. Indeed, the appearance of such traits in inbred children usually is due to homozygosity by descent (HBD) for a single disease-causing allele inherited from a recent ancestor common to both the maternal and paternal lineage. This observation not only explains the incidence of such traits in consanguineous marriages but also provides a powerful tool for genetic mapping of the responsible genes.

Homozygosity mapping (Smith 1953; Lander and Botstein 1987) involves locating a gene causing a rare recessive trait by using multipoint linkage analysis to find regions of HBD shared among inbred affected children. The method is particularly powerful because it does not require the availability of families with multiple affected individuals but, rather, requires only unrelated affected singletons from consanguineous marriages. Linkage can be detected with a very small sample: in principle, three offspring from a first-cousin marriage suffice to obtain a LOD score of 3.0. Homozygosity mapping is thus well suited to a wide variety of recessive traits of medical or biological interest for which it is impractical or impossible to gather a large collection of multiplex families.

When homozygosity mapping was first proposed (Smith 1953; Lander and Botstein 1987), the biggest obstacle to its practical implementation was the lack of an adequate genetic map of the human genome. Without such a map, a scan for regions of homozygosity cannot be carried out. Recent efforts by Weissenbach and others have made dense genetic maps of polymorphic markers widely available (Weissenbach et al. 1992; Buetow et al. 1994; Gyapay et al. 1994). As a result, homozygosity mapping has recently been used to locate several recessive disease genes, and additional searches are under way (Goto et al. 1992; Ben Hamida et al. 1993; Farrall 1993; Pollak et al. 1993; German et al. 1994).

Despite these successes, broad application of homozygosity mapping has been hampered by a second serious problem: the inability of existing algorithms and software for genetic linkage analysis to perform multipoint calculations in reasonable time. Because currently used genetic markers have heterozygosity of ~70%, it is not enough to look for homozygosity at a single marker. Rather, one must perform multipoint analysis to detect HBD reliably. Unfortunately, even two-locus analysis with highly polymorphic markers is notoriously slow in the presence of inbreeding loops. Indeed, such loops make multipoint analysis with conventional linkage software so slow as to be infeasible. To quote from a recent editorial, "although most popular computer programs for linkage analysis can allow for such loops, calculations of an exact likelihood can be very complicated and may require large amounts of memory and CPU time. The problem is exacerbated if multiple markers are examined simultaneously, if marker genotypings in ancestors are not available or if the markers have many different alleles" (Farrall 1993, p. 108; see also Terwilliger and Ott 1994, concerning the difficulty of carrying

Received August 18, 1994; accepted for publication November 4, 1994.

Address for correspondence and reprints: Dr. Eric S. Lander, Whitehead Institute/MIT Center for Genome Research, 9 Cambridge Center, Cambridge, MA 02142.

0002-9297/95/5602-0021$02.00

out homozygosity mapping in practice). All of these conditions are likely to be true in any real search, and efficient analysis algorithms and software are essential if homozygosity mapping is to achieve its full potential.

Here, we describe a new algorithm for rapid multipoint likelihood calculations and its implementation in a new computer package, MAPMAKER/HOMOZ, designed especially for homozygosity mapping. The new algorithm allows very fast multipoint likelihood computations in small pedigrees with inbreeding loops, even in the presence of highly polymorphic loci and a great deal of missing genotype information. The algorithm can also be applied to nuclear families without inbreeding, with comparable speed. The MAPMAKER/HOMOZ program allows multipoint analysis with dozens of markers to be carried out in minutes on a personal workstation.

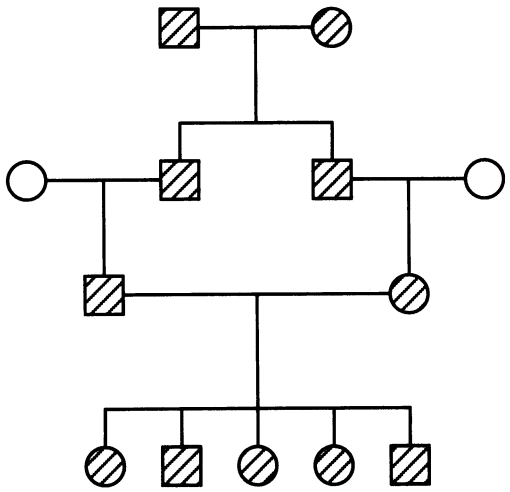

Figure 1 Typical consanguineous pedigree: a first-cousin marriage with five children.

A Novel Algorithm for Rapid Multipoint Likelihood Calculations

The Elston-Stewart algorithm (Elston and Stewart 1971), used to compute likelihoods by conventional linkage analysis software, has running times that blow up exponentially with the number of loci considered simultaneously. The explosion is worse in pedigrees with inbreeding loops, in the presence of missing genotype information, and when markers with many alleles are used. To avoid this combinatorial blowup, Lander and Green (1987) developed an alternative approach to multipoint likelihood calculations, using hidden Markov models (HMMs). With this approach, the computation time for a single likelihood scales linearly with the number of loci and is independent of the number of alleles—although it scales exponentially with the number of meioses in the pedigree. Accordingly, it is appropriate for pedigrees of modest size.

In principle, the Lander-Green algorithm could be used directly for homozygosity mapping, inasmuch as the pedigrees are typically not too large. However, it is possible to do even better. We have developed a new algorithm for speeding up the key step of the HMM calculation in the Lander-Green algorithm. As a result, it is possible to perform in minutes multipoint analysis with a large number of loci in pedigrees of the sort used for homozygosity mapping—despite the presence of inbreeding loops, missing information, and highly polymorphic markers.

In the Lander-Green algorithm (described in detail in the appendix), each locus i is assigned an inheritance vector P_i of length $N = 2^n$, where n is the number of meioses of interest. Each of the N components corresponds to one of the possible inheritance patterns in the pedigree at locus i. The key step in the HMM involves multiplying the inheritance vector by an $N \times N$ transition matrix $T_N(\theta)$, containing the transition probabilities between the inheritance patterns for loci at recombination fraction θ. In the general case, vector-by-matrix multiplication requires N^2 multiplications. However, the transition matrix $T_N(\theta)$ turns out to have a special structure that allows the calculation to be performed more rapidly. Specifically, we have developed an algorithm that requires only $N\log_2 N$ multiplications and $N(\log_2 N)^2$ additions. There is a dramatic difference in performance between algorithms that scale as N^2 and those that scale as $N\log_2 N$. To illustrate the improvement, consider a first-cousin marriage with three children. The number of meioses (n) under study is 12, so that $N^2 > 16{,}000{,}000$, while $N\log_2 N = 49{,}152$; the speedup amounts to a factor of ~ 340.

In a consanguineous marriage such as that shown in figure 1, the meioses that must be considered are those involving transmission from one shaded individual to another. For a kth-cousin marriage yielding l children, a total of $4 + 2k + 2l$ meioses must be considered. In fact, one can reduce this number by 2 (speeding up the overall calculation by slightly more than a factor of 4) by employing a minor approximation: arbitrarily select one founder as the source of the disease allele, and ignore the other founder. (The only consequence of ignoring the second founder is that one neglects the possibility of HBD occurring for the disease allele from one founder and for a marker allele from the other founder; this is, however, a rare event that involves two recombinations in an interval and makes only a small contribution to the likelihood.) In practice, this effect is so slight that it can be safely neglected, and we in general employ the approximation. To check the validity of this approximation, the program allows one to perform the calculation without this shortcut. Below, the number of meioses to be considered for a kth-cousin marriage with l children will be taken to be $2 + 2k + 2l$.

It is interesting to note that, while the time required to

Table 1

Running Time for 100 Calculations of Multipoint LOD Score in a First-Cousin Mating, as a Function of Number of Children and Number of Genetic Markers

No. of Children	No. of Meioses	Running Time (s)	
		5 Markers	20 Markers
1	6	<1	<1
2	8	<1	1.5
3	10	1.5	7
4	12	30	37
5	14	170	206
6	16	924	1,126

NOTE.—Calculations were performed on a DEC Alpha personal workstation (see text for specifications).

calculate the likelihood for a single putative location of the disease gene scales proportionally with the number of loci L, the time required to repeat this calculation for each of M possible locations along a chromosome scales only with $M + L$ (rather than ML, as one might expect). The additional efficiency is achieved because the calculation is performed in two steps. In an initialization step, two conditional inheritance vectors are computed and saved for each locus—one conditioned on all marker data at loci to the right and the other on all marker data at loci to the left (for details, see the appendix). This step takes time proportional to L but independent of M. After this initialization, the disease locus is placed at M positions on the map, and the corresponding likelihoods are computed. This calculation depends on only the conditional inheritance vectors at the closest informative flanking loci and thus takes time proportional to M but independent of L. As a result, calculating the LOD score at a fixed large number M of points across a chromosome is virtually independent of L. This is illustrated in table 1 for the case $M = 100$ and $L = 5$ or 20.

The details of the algorithm are described in the appendix. The key point is that the algorithm allows practical multipoint computations with an essentially unlimited number of loci considered simultaneously in the presence of missing information for early generations. It is limited to pedigrees with a reasonably small number of people: 20 meioses of interest is probably the practical limit. If more complete genotyping information is available, larger pedigrees may be handled. The algorithm is not restricted to pedigrees with loops: it allows multipoint likelihoods with many loci to be computed very rapidly in any small pedigree, and thus it is also suitable for mapping genes in nuclear families and other small pedigrees.

Overview of the MAPMAKER/HOMOZ Software Package

We implemented the new algorithm in a computer package called MAPMAKER/HOMOZ, designed for multipoint analysis of homozygosity mapping in inbred families. The program is not a general purpose linkage-analysis package designed for arbitrary pedigrees; rather, it is specifically designed for mapping of recessive traits in nuclear families. (The algorithm could, in principle, be used to map dominant traits as well, but small pedigrees are less useful in this case.) The only pedigree structure allowed is a nuclear family (with genotype information available for some or all of children and parents) in which the parents may be either unrelated or related by a specified degree of inbreeding (e.g., as uncle-niece, first cousins, etc.). The package computes and plots exact multipoint LOD scores. Incomplete penetrance can be included. In fact, the package allows liability classes to be specified, thereby allowing penetrance to be specified, for example, in an age-dependent manner. The allele frequencies for the disease locus and the marker loci can be specified. (In particular, the program does not assume that disease alleles are infinitely rare; the possibility of two disease alleles entering the pedigree is allowed.) Currently, the package does not allow sex-specific recombination fractions. Allowing sex-specific fractions would have only a minor effect on the power to detect a locus, but it turns out to significantly slow down the computation (see the appendix).

To compare the speed of this package with that of existing software, we used MAPMAKER/HOMOZ, LINKAGE (Lathrop et al. 1984), and FASTLINK (a faster version of LINKAGE; see Cottingham et al. 1993) to analyze the following relatively simple situation: given a fixed map with three genetic markers (two markers with three alleles each and one marker with four alleles) scored in a single first-cousin marriage with four children (with genotypes available for the parents and children but not earlier generations), calculate the LOD score for a disease locus tested in each of 20 positions along the map. The three packages were tested on the same machine, a SUN SPARC IPX personal computer (with the LINKMAP program used in both LINKAGE and FASTLINK). MAPMAKER/HOMOZ required 11 s to complete the analysis, FASTLINK took 2 h and 50 min, and LINKAGE took 36 h. All three programs obtained the same results. On this example, the new program was about 1,000-fold faster than FASTLINK and about 12,000-fold faster than LINKAGE.

It should be stressed that this example was particularly favorable for the traditional algorithms: adding additional markers, increasing the number of alleles per marker, or not specifying parental typings does not sig-

```
**********************************************************************
*                                                                    *
*                          MAPMAKER/HOMOZ                            *
*                           (version 0.9)                            *
*                                                                    *
**********************************************************************

homoz:1> load mydata.dat
Parsing Linkage marker data file...
1 affectation locus and 3 marker loci successfully read

homoz:2> prep mypedfile.dat
child 11 is untyped - dropped
Found 3 real kids - indivs: 10 12 13
Father=8, Mother=7
individual 1 is 2 generations above dad and 2 generations above mom
individual 2 is 2 generations above dad and 2 generations above mom

homoz:3> map haldane
The Haldane map function is now in use.

homoz:4> use 1 .1 2 .1 3
Current map (3 markers):
loc1 11.2 loc2 11.2 loc3

homoz:5> off end 0
Scanning will now be done 0.0 cM beyond the ends of the map

homoz:6> increment 1
scan increment is now set to 1.00

homoz:7> scan
1 pedigree input
PEDIGREE: 1
unlinked log_10 likelihood = -9.705433
position        log-like          LOD
   0.00         -8.9442          0.7612    ****
   1.00         -9.0176          0.6878    ****
   2.00         -9.0961          0.6093    ****
   3.00         -9.1810          0.5244    ***
   4.00         -9.2739          0.4315    ***
   5.00         -9.3772          0.3283    **
   6.00         -9.4943          0.2111    **
   7.00         -9.6309          0.0745    *
                   *
                   *
                   *
  19.00        -10.4616         -0.7562
  20.00        -10.5653         -0.8599
  21.00        -10.7638         -1.0583
  22.00        -11.3417         -1.6363

Maximum total LOD: 0.7612 (0.0 cM)
Total output data saved in scana.txt and scana.ps
```

Figure 2 Example of a MAPMAKER/HOMOZ session. Marker and pedigree data are loaded, the map function is chosen, the map of fixed markers is defined, the scan distance beyond the leftmost and rightmost markers is set to 0 cM, the scan increment is set to 1 cM, and multipoint LOD scores are computed.

nificantly alter the running time of MAPMAKER/HOMOZ, while the other programs become so slow that their performance can no longer be measured. Of course, the other packages are all-purpose linkage analysis packages and were not specifically designed to handle multipoint analysis in pedigrees with inbreeding loops. The new program has a narrower range of application, but it performs extremely well within this limited scope.

Interactive Shell

In order to allow easy exploration and analysis of experimental data, the package is interactive. The user interface is based on the genetic mapping package MAPMAKER (Lander et al. 1987). As in MAPMAKER, a simple command vocabulary is used to perform various types of analysis. On-line help is available at any time, and one can record a verbatim transcript of the session (see fig. 2).

Data

The data used by MAPMAKER/HOMOZ are contained in two types of files: pedigree files and locus-description files. Pedigree files contain genotype information on members of the pedigree, affectation status (and an optional liability class) for each, and a simple description of the degree of inbreeding in the pedigree. The package can read pedigree data in LINKAGE (Lathrop et al. 1984) format. The data may also be directly entered in an internal format, which is simpler, since it does not require information to be entered for any members of the pedigree other than the affected children and their siblings and parents. Pedigree data are automatically checked for inconsistencies in inheritance (non-Mendelian inheritance when parental genotypes are available and more than four distinct alleles among the children when they are not). Such checks are useful for detecting both laboratory and data-entry errors.

Locus-description files contain the following information on mapped marker loci: locus name, possible alleles at the locus (either as ordinal numbers or as actual allele sizes), population frequency of each allele, and map position of the marker. This information may be entered directly in the specified format. Alternatively, it may be read in from a LINKAGE parameter file.

Computing LOD Scores

Once the data are loaded, analysis can begin. The fundamental analysis is to calculate the multipoint LOD score (for each individual pedigree and for all pedigrees together) at specified intervals across a chromosome. This is accomplished by typing "scan." The LOD scores for each pedigree at each location are displayed as they are computed. The combined scores are then displayed in numerical format, along with a terminal graphics display that plots any scores >0. The results for individual pedigrees and the totals are saved to ASCII text files for future review as well as to postscript files for easy display. An excerpt from a session is shown in figure 2.

A number of options are available for customizing the LOD-score computation. The chromosomal region to be studied, as well as the map positions of markers, can be specified by the "use" command (e.g., "use loc1 0.05 loc2 0.15 loc7 0.09 loc5" specifies the use of markers loc1, loc2, loc7, and loc5, in that order, with the corresponding recombination fractions between them). In addition, the user can specify the increment between consecutive points in a scan (or the number of computations between consecutive markers), the distance of the scan beyond the leftmost and rightmost markers, a subset of pedigrees to include, and the map function to employ.

Running Time

As discussed in the section describing the algorithm, the running time scales linearly with $M + L$, where L is the

number of genetic markers used and M is the number of points in a scan and is slightly faster than 2^n, where n is the number of meioses studied. To evaluate the running time of MAPMAKER/HOMOZ, the computer package was applied to data from a single first-cousin mating with $k = 1, 2, \ldots, 6$ affected children (corresponding to $n = 6$, 8, 10, 12, 14, and 16 meioses) with $L = 5$ or 20 genetic markers. In each case, the multipoint LOD score was calculated at 100 points along a chromosome. The running time for each case was measured on a DEC 3000 Alpha workstation with a 64-bit RISC processor (clock speed 190 MHz) The results are shown in table 1.

The running time changes only slightly when the number of markers is increased from 5 to 20, reflecting the fact that the computation time is dominated by calculating LOD scores at 100 points. The running time also shows the expected behavior of scaling slightly faster than 2^n, increasing by a factor of ~ 5 when n increases by 2. In absolute terms, the running times are very fast for small families. Calculating 100 six-locus multipoint LOD scores in a first-cousin mating required < 1 s with one affected child and 1.5 s with three affected children but ~ 15 min with six affected children. This is still extremely fast compared with the speed of conventional linkage software in inbred families, even with only two markers.

Analyzing >16 meioses would strain available memory of most workstations and could result in considerably slower computation times. Accordingly, MAPMAKER/HOMOZ normally limits the allowed number of meioses to 16 (although this number may be changed by the user). As discussed earlier, this corresponds to eight children for a noninbred mating, six for a first-cousin mating, five for a second-cousin mating, and four for a third-cousin mating. If greater speed is desired, some children can be dropped from a large sibship—starting with unaffecteds, who contribute little linkage information. Of course, individuals to be dropped must be specified prior to the analysis, either by phenotype, by the amount of genotype information available, or at random, but never on the basis of the results of the analysis.

Availability

The MAPMAKER/HOMOZ software is written in standard ANSI C. It is freely available from the authors.

Population-Genetic Issues in Applying Homozygosity Mapping

In applying homozygosity mapping, it is important to recognize that—as with all linkage analysis—the results may be sensitive to incorrect assumptions about key parameters. For homozygosity mapping, the most important parameters are the allele frequencies at the marker loci. Since greater evidence of HBD is contributed by finding homozygosity for a rare allele than for a common one, *under*estimating the frequency of an allele will lead one to *over*estimate the LOD score. (Note that this is a general feature of linkage analysis in any pedigree with missing founders and not of any specific method used to compute LOD scores.)

Ideally, allele frequencies should be estimated in the study population. For example, one could examine their occurrence on normal chromosomes from unaffected relatives in the pedigree or on unselected chromosomes in the same ethnic group or regional population. Alternatively, allele frequencies may be estimated from complete pedigrees by maximum likelihood (Boehnke 1991). Frequently, this may be impractical, and published allele frequencies are used. These allele frequencies are determined in a reference population and may differ considerably from the allele frequencies in the population(s) from which the study pedigrees are drawn.

One approach that is often used when allele frequencies at a locus are not known is to assign all alleles equal frequencies. Unfortunately, this assumption is biased: it can inflate positive LOD scores or even produce positive expected LOD scores in the absence of linkage (Ott 1992; Terwilliger and Ott 1994). The reason is not difficult to see: the LOD-score increase caused by erroneously assuming that common alleles are rare more than offsets the LOD-score decrease caused by assuming that rare alleles are common.

To guard against inflated LOD scores, we would recommend performing "sensitivity analysis," to see how the LOD score responds to changes in the allele frequencies. One possible approach is to perform a "bounded influence" calculation by imposing a lower bound on all allele frequencies. For example, any allele frequency $<10\%$ could be set to 10%. (The fact that the sum of adjusted allele frequencies is >1 does not interfere with the analysis; it has the effect of increasing the likelihood of seeing homozygosity for an allele, without decreasing other probabilities. In the case of singleton affecteds, one can show that the calculation is always conservative.)

MAPMAKER/HOMOZ incorporates a built-in feature for performing such sensitivity analysis. The user can override the allele frequencies in the locus-description file, with the command "allele thresholded x," which sets all allele frequencies less than x to be equal to x. (Originally specified allele frequencies may be restored with the command "allele given.") The LOD scores are then recomputed with the new frequencies.

If the sensitivity analysis yields results that differ substantially from those obtained with the originally specified allele frequencies, the results should be treated with caution. This is most likely to occur when a high LOD score is produced by rare alleles at one or two markers; one should be certain that the allele is indeed rare in the relevant population,

before accepting the LOD score as evidence of linkage. The best solution is to genotype additional markers in the region, since the evidence for HBD with multiple markers is much less sensitive to variation in any given allele frequency.

Application of MAPMAKER/HOMOZ to Familial Mediterranean Fever (FMF)

We tested the package by analyzing data from nine consanguineous pedigrees used in earlier mapping studies of FMF. FMF is an autosomal recessive disorder characterized by intermittent attacks of fever with abdominal pain, pleurisy, and/or arthritis; it affects primarily members of non-Ashkenazi Jewish, Armenian, Turkish, and Middle Eastern Arab populations (Pras et al. 1992). The gene causing FMF, designated "MEF," has been mapped to the short arm of chromosome 16 (Pras et al. 1992; Aksentijevich et al. 1993*b*). These studies used homozygosity mapping, as well as traditional linkage analysis, to establish linkage between MEF and chromosome 16 markers in consanguineous and nonconsanguineous non-Ashkenazi Jewish families.

We analyzed the nine consanguineous pedigrees by computing nine-point LOD scores for MEF against a fixed genetic map with eight markers in a 25-cM region surrounding MEF. The pedigrees include seven first-cousin marriages and two uncle-niece marriages. The results of the analysis, which required several minutes on a DEC Alpha workstation, are shown in figure 3. The LOD score peaks at 15.5, between D16S246 and D16S523, where MEF is known to map (D. Kastner, personal communication). The LOD score drops off sharply at D16S246 and D16S80. This analysis is presented only as a test of the package; full details of multipoint mapping of MEF will be presented elsewhere (D. Kastner, personal communication).

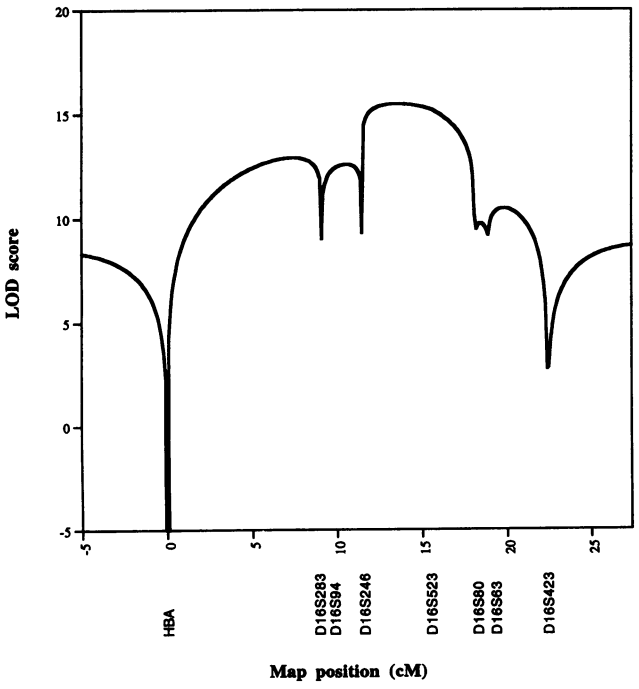

Figure 3 Nine-point LOD scores for MEF, relative to a map of eight fixed markers on the short arm of chromosome 16. Marker names and map distances are shown. LOD scores were computed at 325 points separated by 0.1 cM.

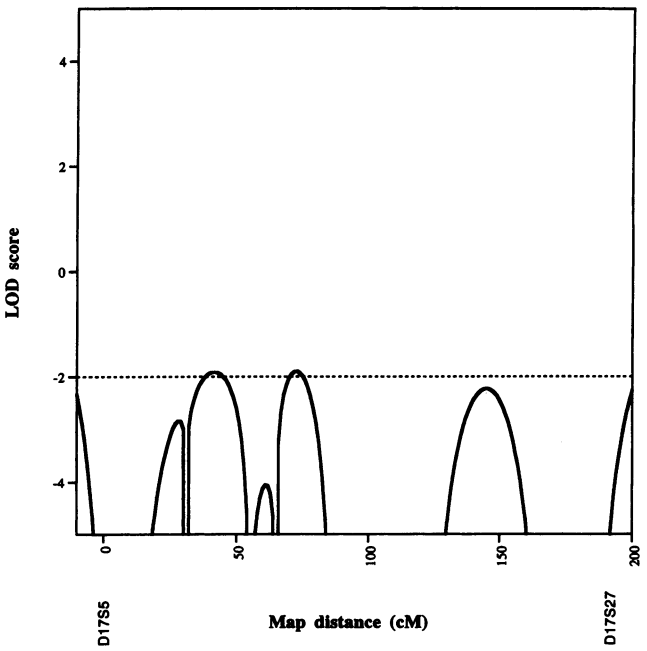

Figure 4 Seventeen-point LOD scores for MEF, relative to a map of 16 fixed markers covering most of chromosome 17. The leftmost and rightmost markers are shown. LOD scores were computed at 210 points separated by 1 cM.

We also looked at 16 markers on chromosome 17, covering almost the entire length of the chromosome: 190 cM between D17S5 and D17S27 (Nakamura et al. 1988). Before MEF was definitively mapped to chromosome 16, there was tentative evidence of linkage to chromosome 17. This appears to have been a false positive, on the basis of three-point analysis that included both consanguineous and nonconsanguineous families (Aksentijevich et al. 1993*a*). When we computed 17-point LOD scores for the nine consanguineous families, we found no evidence of linkage of MEF to chromosome 17 (fig. 4). In fact, almost the entire chromosome could be excluded at the LOD threshold of −2. This result is consistent with the earlier observation of a lack of excess homozygosity at chromosome 17 markers (Aksentijevich et al. 1993*a*), and the LOD scores from the present analysis would have substantially strengthened the argument against linkage on chromosome 17. Once again, the analysis was performed in several minutes on a DEC Alpha workstation.

Conclusion

Human recessive traits are an important subject of study, and mapping the genes underlying these traits is of great interest. Homozygosity mapping was proposed as an efficient strategy for mapping such traits (Smith 1953; Lander and Botstein 1987). Originally, the most severe obstacle to practical applications of the strategy was the lack of a good genetic linkage map of the human genome. Recent availability of such maps has placed homozygosity mapping within reach of many researchers (Farrall 1993). Until now, the remaining bottleneck has been computational: available software tools cannot provide adequate multipoint analysis in the presence of inbreeding loops. MAPMAKER/HOMOZ should eliminate this bottleneck by allowing multipoint homozygosity mapping to be carried out rapidly in a user-friendly environment.

More generally, multipoint analysis is increasingly important for every kind of linkage analysis, with the recent availability of high-quality genetic maps for the entire human genome. Conventional linkage analysis programs are not necessarily optimally designed for this task and, in some cases, may be too slow to be practical. The algorithm described in this paper allows very rapid multipoint likelihood calculation in nuclear families (with or without parental consanguinity), and the accompanying software package makes multipoint mapping feasible in many experimental contexts.

Acknowledgments

We thank David Botstein and Michele Gschwend for many discussions concerning homozygosity mapping and for sharing unpublished data. We thank Daniel Kastner and his colleagues for sharing the pedigree and genotype data from their FMF studies. We thank Robert Elston, Michael Boehnke, Augustine Kong, and an anonymous referee for comments on the manuscript. This work was supported in part by National Institutes of Health grant HG00098 to E.S.L.

Appendix

Description of Algorithm

Consider a fixed map of M ordered marker loci with known recombination fractions θ_i between loci i and $i + 1$. We wish to compute the likelihood for a given pedigree. According to Lander and Green (1987), the inheritance pattern at each locus i ($i = 1, 2, \ldots, M$) can be described by an n-bit vector v_i. Each bit describes the outcome of one of the n meioses in the pedigree: the bit is 0 if the paternally derived allele is transmitted and 1 if the maternally derived allele is transmitted. The set of all possible n-bit vectors will be identified with $(Z_2)^n$, the additive vector space over the field with two elements (i.e., with component-wise addition modulo 2).

Because bits reflect the inheritance pattern at each locus, a given bit in v_{i+1} differs a priori from the corresponding bit in v_i, with probability equal to the recombination fraction θ_i. Let $H(\alpha,\beta)$ denote the Hamming distance (the number of bits that differ) between two vectors $\alpha,\beta \in (Z_2)^n$. The Hamming distance is a metric on $(Z_2)^n$. We will define a $2^n \times 2^n$ transition matrix $T(\theta)$ having rows and columns indexed by elements $\alpha \in (Z_2)^n$, with elements $T_{\alpha,\beta}(\theta) = \theta^j(1-\theta)^{n-j}$, where $j = H(\alpha,\beta)$. The inheritance vectors $v_1, v_2, \ldots, v_n$, then follow an (inhomogeneous) Markov chain, with the transition matrix between loci i and $i + 1$ being

$$T_{v_i,v_{i+1}}(\theta_i).$$

Typically, the true inheritance vector v_i at locus i cannot be uniquely determined from the data; a number of vectors may be consistent. One can, however, easily compute the probability $P_i(\alpha)$ of the observed data for locus i, given that $v_i = \alpha$. (In brief, this probability is 0 for vectors inconsistent with the phenotypic data and involves the appropriate combination of allele frequencies and penetrances for vectors that are consistent.) We define a $2^n \times 2^n$ diagonal matrix Q_i having rows and columns indexed by elements $\alpha \in (Z_2)^n$, with elements $Q_{\alpha,\alpha} = P_i(\alpha)$ and $Q_{\alpha,\beta} = 0$ for $\alpha \neq \beta$.

The likelihood can then be computed from an HMM in which the hidden states are the inheritance vectors, the observed states are the data, and the transition matrix (for the transition between loci i and $i + 1$) is $\mathbf{T}(\theta_i)$ given above. The likelihood L is then given by

$$L = \mathbf{1}^R Q_1 T(\theta_1) Q_2 T(\theta_2) \ldots T(\theta_{M-1}) Q_M \mathbf{1}^C,$$

where $\mathbf{1}^R$ and $\mathbf{1}^C$ are, respectively, the 1×2^n row vector and $2^n \times 1$ column vector with all coordinates equal to 1.

Because matrix multiplication is associative, the product can be computed from either direction. To map a disease locus relative to a map consisting of marker loci, one can precompute all partial products from the left and right sides. It is then trivial to compute the likelihood when the disease locus is placed at any position in the map.

The computationally intensive step is computing vector-matrix products of the form $\mathbf{P}\,T(\theta)$, where the coordinates are indexed by 2^n elements $\alpha \in (Z_2)^n$. Naively, this would appear to require 2^{2n} multiplications. Lander and Green (1987) conjectured that it should be possible to perform the computation in roughly $n\,2^n$ multiplications, but they did not provide an algorithm to do so. Our new algorithm for computing the product achieves comparable performance (see below) by taking advantage of the fact that the

transition matrix T has only $n + 1$ independent elements, corresponding to the possible number of recombinations.

Given a vector $\mathbf{P} = (P_\alpha)$, we seek to compute a vector S with elements

$$S_\beta = \sum_\alpha P_\alpha T_{\alpha,\beta}(\theta) ,$$

with $\alpha,\beta \in (Z_2)^n$. We compute it recursively as follows. We first form the $2^n \times (n + 1)$ matrix of products $\mathbf{W}^0$, which contains all possible products of the 2^n elements of P and the $n + 1$ independent elements of T. We then perform a recursive operation to compute $\mathbf{W}^1, \ldots, \mathbf{W}^n$. The first column of $\mathbf{W}^n$ is the desired vector S. In order to describe the recursive operation and to show that it gives the desired result, we introduce the following notation.

Let c_k denote the n-bit vector whose first k bits are 0 and whose remaining bits are 1. The set

$$\{c_k\}_{k=0,\ldots,n-1}$$

is a basis for the vector space $(Z_2)^n$. Let $\langle c_i, \ldots, c_j\rangle$ denote the subspace spanned by $c_i, \ldots, c_j$. Then the following lemma holds:

LEMMA. Let $\sigma \in \langle c_1, \ldots, c_{k-1}\rangle$ and let $\gamma, \delta \in \langle c_k, \ldots, c_{n-1}\rangle$ such that $H(0,\gamma) = H(0,\delta)$. Then $H(\sigma,\gamma) = H(\sigma,\delta)$.

PROOF OF LEMMA. By the assumptions, γ and δ must have 0's in the first k coordinates and must contain an equal number of 1's in the last $n - k$ coordinates. Since $\sigma \in \langle c_1, \ldots, c_{k-1}\rangle$, the last $n - k$ coordinates of σ are equal and the result follows immediately.

We now come to the main result. Let $\mathbf{W}^0$ have elements $W^0_{\beta,j} = P_\beta t_j$, where $\beta \in (Z_2)^n$ and $j = 0, 1, \ldots, n$, counting the number of recombinations. Here, t_j is defined as the transition probability between any pair of vectors α,β with $H(\alpha,\beta) = j$—that is, $t_j(\theta) = \theta^j(1 - \theta)^{n-j}$ (as before, t_j depends on θ, but, for convenience, we suppress this dependence here and in what follows). Recursively define

$$W^{k+1}_{\beta,j} = W^k_{\beta,j} + W^k_{(\beta+c_k),(n-k-j)} , \tag{1}$$

where $j = 0, 1, \ldots, n - k - 1$. Also define a subset of vectors $\boldsymbol{\Omega}_k(\boldsymbol{\beta})$ by

$$\Omega_k(\beta) = \{\beta + \varepsilon_0 c_0 + \ldots + \varepsilon_{k-1}c_{k-1} \mid \varepsilon_0, \ldots, \varepsilon_{k-1} \in \{0,1\}\} . \tag{2}$$

Note that $\boldsymbol{\Omega}_k(\boldsymbol{\beta})$ is a coset of $\langle c_1, \ldots, c_{k-1}\rangle$. In particular, $\Omega_n(\beta) = (Z_2)^n$. Then the following result holds:

PROPOSITION 1.

$$W^k_{\beta,j} = \sum_{\alpha\in\Omega_k(\beta)} P_\alpha T_{\alpha,(\beta+c_{n-j})} .$$

In particular,

$$W^n_{\beta,0} = \sum_{\alpha\in(Z_2)^n} P_\alpha T_{\alpha,\beta} = S_\beta .$$

PROOF. The proof is by induction. To prove the base case ($k = 0$), note that, from the definition of $W^0_{\beta,j}$, we have

$$W^0_{\beta,j} = \sum_{\alpha\in\Omega_0(\beta)} P_\alpha T_{\alpha,(\beta+c_{n-j})} = P_\beta T_{\beta,(\beta+c_{n-j})} = P_\beta T_j .$$

The last equality follows because $H(\beta,\beta + c_{n-j}) = H(0,c_{n-j}) = j$. To prove the general case, assume that the proposition holds for $W^k_{\beta,j}$. From the recursion relation (1), we have

$$W^{k+1}_{\beta,j} = \sum_{\alpha\in\Omega_k(\beta)} P_\alpha T_{\alpha,(\beta+c_{n-j})} + \sum_{\alpha\in\Omega_k(\beta+c_k)} P_\alpha T_{\alpha,(\beta+c_k+c_{j+k})} .$$

It is clear from the definition that $\Omega_k(\beta)\cup\Omega_k(\beta + c_k) = \Omega_{k+1}(\beta)$. We need to show that

$$T_{\alpha,(\beta+c_k+c_{j+k})} = T_{\alpha,(\beta+c_{n-j})} ,$$

or $H(\alpha,\beta + c_k + c_{j+k}) = H(\alpha,\beta + c_{n-j})$, for $\alpha \in \Omega_k(\beta + c_k)$. Note that α is of the form $\sigma + \beta + c_k$, where $\sigma \in \langle c_1, \ldots, c_{k-1}\rangle$. Therefore, $H(\alpha,\beta + c_k + c_{j+k}) = H(\sigma,c_{j+k})$, and $H(\alpha,\beta + c_{n-j}) = H(\sigma,c_k + c_{n-j})$. Note that $c_k, c_{j+k}, c_{n-j} \in \langle c_k, \ldots, c_{n-1}\rangle$, and that when the lemma is used, $H(\sigma,c_{j+k}) = H(\sigma,c_k + c_{n-j})$ if $H(0,c_{j+k}) = H(0, c_k + c_{n-j})$. The last equality holds since both c_{j+k} and $c_k + c_{n-j}$ differ from 0 by $n - j - k$ bits. Therefore we have,

$$W^{k+1}_{\beta,j} = \sum_{\alpha\in\Omega_{k+1}(\beta)} P_\alpha T_{\alpha,(\beta+c_{n-j})} .$$

This completes the proof.

Using the proposition, we iteratively compute the matrix $W^k_{\beta,j}$ for $j = n, n - 1, \ldots, 0$, until we obtain $W^n_{\beta,0}$, which is the desired vector-matrix product P T. The entire calculation can be performed with $(n + 1)2^n$ multiplications and $(n + 1)(n + 2)2^{n-2}$ additions, instead of 2^{2n} multiplications for simple matrix multiplication.

To include sex-specific recombination fractions in this algorithm, we would need to keep track of not only how many crossovers occur but whether they occur in male or female meioses. While a similar recursive algorithm can be constructed for this case, it would be slower by roughly a factor of n. In view of the minimal LOD-score differences between analyses with and without sex difference (see, e.g., Terwilliger and Ott 1994), we de-

cided to neglect sex-specific recombination fractions in favor of greater computational efficiency.

References

Aksentijevich I, Gruberg L, Pras E, Barlow JE, Kovo M, Gazit E, Dean M, et al (1993*a*) Evidence for linkage of the gene causing familial Mediterranean fever to chromosome 17q in non-Ashkenazi Jewish families: second locus or type I error? Hum Genet 91:527–534

Aksentijevich I, Pras E, Gruberg L, Shen Y, Holman K, Helling S, Prosen L, et al (1993*b*) Refined mapping of the gene causing familial Mediterranean fever, by linkage and homozygosity studies. Am J Hum Genet 53:451–461

Ben Hamida C, Doerflinger N, Belal S, Linder C, Reutenauer L, Dib C, Gyapay G, et al (1993) Localization of Friedreich ataxia phenotype with selective vitamin E deficiency to chromosome 8q by homozygosity mapping. Nat Genet 5:195–200

Boehnke M (1991) Allele frequency estimation from data on relatives. Am J Hum Genet 48:22–25

Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Duyk GM, Sheffield VC, Wang Z, et al (1994) Integrated human genome-wide maps constructed using the CEPH reference panel. Nat Genet 6:391–393

Cottingham RW Jr, Idury RM, Schäffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53:252–263

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Farrall M (1993) Homozygosity mapping: familiarity breeds debility. Nat Genet 5:107–108

Garrod AE (1902) A study in chemical individuality. Lancet 1902:1616ff

German J, Roe AM, Leppert M, Ellis NA (1994) Bloom syndrome: an analysis of consanguineous families assigns the locus to chromosome band 15q26.1. Proc Natl Acad Sci USA 91:6669–6673

Goto M, Rubenstein M, Weber J, Woods K, Drayna D (1992) Genetic linkage of Werner's syndrome to five markers on chromosome 8. Nature 355:735–738

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, et al (1994) The 1993–94 Généthon human genetic linkage map. Nat Genet 7:246–339

Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570

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

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181

Lathrop GM, Lalouel JM, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443–3446

Nakamura Y, Lathrop M, O'Connel P, Leppert M, Barker D, Wright E, Skolnick M, et al (1988) A mapped set of markers for human chromosome 17. Genomics 2:302–309

Ott J (1992) Strategies for characterizing highly polymorphic markers in human gene mapping. Am J Hum Genet 51:283–290

Pollak MR, Chou YH, Cerda JJ, Steinmann B, La Du BN, Seidman JG, Seidman CE (1993) Homozygosity mapping of the gene for alkaptonuria to chromosome 3q2. Nat Genet 5:201–204

Pras E, Aksentijevich I, Gruberg L, Balow JE, Prosen L, Dean M, Steinberg AD, et al (1992) Mapping of a gene causing familial Mediterranean fever to the short arm of chromosome 16. N Engl J Med 326:1509–1513

Smith C (1953) The detection of linkage in human genetics. J R Stat Soc B 15:153–184

Terwilliger JD, Ott J (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, et al (1992) A second-generation linkage map of the human genome. Nature 259:794–801