European Journal of Human Genetics (1999) 7, 103–109
© 1999 Stockton Press All rights reserved 1018–4813/99 $12.00
http://www.stockton-press.co.uk/ejhg

ARTICLE

Multipoint genomic scanning for quantitative loci: effects of map density, sibship size and computational approach

Yin Y Shugart[1] and David E Goldgar[2]

[1]*Department of Family Medicine and Clinical Epidemiology, University of Pittsburgh, Pittsburgh, PA, USA*
[2]*Unit of Genetic Epidemiology, International Agency for Research on Cancer, Lyon, France*

Multipoint interval mapping (MIM) and the MAPMAKER/SIBS program (M/S) are two methods of mapping quantitative loci by examining identity by descent (IBD) sharing in a region spanned by multiple microsatellite DNA markers. For the purpose of comparison, we simulated a quantitative trait controlled by a two-locus model, and evaluated the power and genome-wide false positive rate of both approaches. Based on our simulation, we examined the effects of marker density (5 cM, 10 cM and 20 cM) and sibship size (2, 3, 4 and 5) on the power to detect linkage. Our results indicate that a 10 cM map provides the optimal trade-off between power and type I error, and that the power of MIM increases with sibship size and, in general, performs better than MAPMAKER/SIBS. Furthermore, we conclude that using a reasonable sample of randomly ascertained sibships, it is possible to map a quantitative trait locus (QTL) which accounts for 25% of the phenotypic variance.

Keywords: multipoint genome scanning; quantitative trait; sibship; sib pair

Introduction

After a decade of successes in mapping simple Mendelian diseases, traits with complex etiology are becoming the center of attention; many of these are quantitative phenotypes. Unlike qualitative traits where one can classify a specific individual as 'affected' or 'normal', many quantitative phenotypes may not exhibit multimodality in populations. With a trait such as obesity, for example, it is philosophically ambiguous to define an 'all or none' character, dividing the subjects into 'obese' and 'lean' groups. Rather, the useful information one obtains is the variation of the weights between different individuals. And the fundamental question of concern to human geneticists is what proportion of the individual difference in weight is due to genetic factors and how we can best identify specific genes which contribute to this phenotypic variability.

The basic idea of detecting linkage with quantitative traits was first described by Penrose[1] who developed a method of detecting linkage between a gene responsible for a significant amount of variation in a graded (ordered categorical) trait and another gene which was responsible for variation in another graded trait. The logic behind his method is straightforward: if genes A and B are linked, then the range of the covariance of their effects within the sibship should be increased. Therefore, the degree of linkage can be measured by the observed correlation between two graded traits. More recently, several other methods have been developed to study linkage between quantitative traits and

Correspondence: Yin Yao (Shugart), Johns Hopkins University, Center for Inherited Disease Research, 333 Cassell Drive, Suite 2000, USA. Fax: 410 550 3559; E-mail: shugart + @pitt.edu or shugart@iarc.fr

Received 28 May 1998; revised 28 September 1998; accepted 14 October 1998

genetic marker loci. Haseman and Elston[2] proposed a test (H-E) based on the regression of the squared sib-pair difference for the trait on their estimated genetic correlation at the marker locus, conditional on the joint distribution of the number of genes identified by descent (IBD) at a marker locus and the number of genes IBD at a hypothesized locus. Such a method is rooted in the assumption that if two sibs share the same allele (IBD), then it is more likely they will have similar phenotypic values compared with the ones who do not. Several variance components methods have been studied by various authors.[8] Goldgar[4] developed the first multipoint approach of partitioning genetic variance of a quantitative trait to loci in specific chromosomal regions. More precisely, the method estimates the proportion of genetic material shared identical by descent between two siblings in a given region conditional on their IBD and recombination pattern at a number of marker loci in the region. Those estimates are used to form the predicted covariance matrix of the sibship trait values as a function of the proportion of trait variance due to loci in the region of interest. In contrast to the H-E approach, this method used sibship information instead of sib-pair information. A similar maximum variance components approach proposed by Fulker *et al*[8] suggested calculating IBD sharing at each point in an interval using the expectation based on the genotypes at the two closest flanking markers. This approach was applied to data on reading disabilities in a set of families ascertained through probands who have dyslexia; a significant signal for linkage in a region on chromosome 6p was detected.[9] Most recently, Kruglyak and Lander[10] provided a computer program MAPMAKER/SIBS (M/S) in which they provide three tests: maximum likelihood quantitative trait locus (QTL) estimation, a multipoint H-E method and a non-parametric approach. The maximum likelihood estimation (MLE) variance option of the program computes the variance for a given QTL trait for sibs sharing 0, 1 and 2 (σ_0^2, σ_1^2, σ_2^2) alleles IBD, testing whether $\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2$, based upon the assumption that conditional on the IBD sharing for a pair, the distribution of squared pair difference is normal.

With more and more linkage studies of complex genetic traits being done, it is important to clarify the strength and weaknesses of the different computational approaches. Some power comparisons were reported in the past; for example, it was demonstrated that the multipoint interval mapping (MIM) method was considerably more powerful than the H-E method.[4] Simulations performed by Goldgar and Oniki[11] compared MIM with the traditional lod score method which also can be used for analyzing quantitative traits, but assumes that the users have some knowledge of trait mean and variance for each genotype, which is a situation unlikely to occur in practice. Recently, Kruglyak and Lander[10] claimed the superiority of their approach to methods such as MIM which does not carry out the full IBD computation. Therefore, it is of interest to examine the operating characteristics of the two methods and to provide guidelines for the users who are interested in the utilizing different approaches in quantitative mapping.

In the present paper, we will first examine the issue of optimal marker density, power and type I error corresponding to various sibship sizes. Subsequently, we will report the results of power and false positive rate associated with MIM in comparison with M/S. The power of using independent or all possible pairs in M/S with pedigrees containing more than two sibs will be evaluated and the effect of having sibships with or without parental marker genotype is to be investigated.

Methods

The Underlying Genetic and Chromosomal Model

In all cases two additive trait loci were simulated independently. Each trait locus was assumed to have two alleles, *A* and *a*, with frequencies 0.1 and 0.9 and genotypic effect chosen so that each locus accounted for 25% of the phenotypic variance. Thus, those two major trait loci accounted for 50% of the total phenotypic variation, whilst the remaining 50% was due to individual specific random, normally distributed, environmental factors. We assumed that the first major locus was located at position 70 cM on chromosome 1 and the second major locus at position 20 cM on chromosome 2. For the sake of simplicity, we fixed all 22 chromosome lengths at 150 cM and assumed all markers had four equally frequent alleles giving a heterozygosity of 75%.

Simulation Methods

The detailed algorithm of simulating disease genotypes, phenotypes and marker phenotypes were described elsewhere.[12] The simulation procedure can be summarized as follows:

1) simulation of all crossover points on the maternally and paternally derived chromosomes for each offspring according to Sturt mapping function;
2) simulating of founder genotypes according to the Hardy-Weinberg law, for any given marker at a given map position. Transmission of the marker and trait loci through the pedigrees was uniquely determined by the previously simulated crossover point in each meiosis;
3) a normal random deviate with mean 0 and variance 1 is obtained by adding the genotypic effect at each trait locus and the individual-specific normal environmental component.

To investigate the optimal marker density, 100 pedigrees with two parents and four offspring were generated 500 times with marker spacing of 5 cM, 10 cM and 20 cM. And the same samples were used with different marker spacing. After the optimal marker spacing was picked, the same two-locus genetic model was examined for three additional sibship sizes. The sibship sizes used in the simulation study are two, three, four and five. In all cases the total number of independent pairs generated was held constant at 300, resulting in:

a) 300 families, each consisting of two parents and two offspring;
b) 150 families, each consisting of two parents and three offspring;
c) 100 families, each consisting of two parents and four offspring;
d) 75 families, each consisting of two parents and five offspring;

One thousand replicates were generated for each sibship size.

Implementation in MIM and in M/S

MIM (version 1.2) was downloaded from morgan.med.utah.edu/pub/Mim, and M/S (version 2.0) was obtained from the Whitehead Institute's network site genome.wi.mit.edu.

For each implementation of MIM, sliding three-point analyses of overlapping sets of adjacent markers were used. For a 10 cM genome scanning, a genetic region stretching 5 cM distal and proximal to the three marker region was used, thus the total coverage of each run was 30 cM and analysis of each chromosome resulted in 8 MIM analyses. MIM requires a fixed h^2 for the trait under analysis, which was set to be true value (50%) used in simulation for all analyses. MIM also estimates a parameter P, which is the proportion of genetic variance of the trait due to loci in a chromosomal region determined by a set marker loci. In the output, a χ^2 statistic is reported for a maximized P and saved for each interval. The sliding technique with a 10 cM map is illustrated as follows:

M1___M2___M3

M3___M4___M5

M5___M6___M7

The MLE variance option of the M/S program was used in our analysis to compare with MIM because it was claimed to be the most powerful test among the three provided in the M/S package. The simulated data sets containing families with different sibship sizes were analyzed by both methods simultaneously. As noted earlier, M/S uses genotypic information along the whole chromosome, and thus causes a computationally intensive scanning process. For computational efficiency, we set the step size for scanning in M/S to be 10 cM so that it performs fewer tests and runs faster. Thus, for M/S, lod scores were calculated at 8 points along each chromosome.

To demonstrate the effect of using independent pairs only versus all possible pairs, for all sibship sizes, we examined both 'independent pairs' and 'all possible pairs' options when the analyses were performed with M/S.

The investigation regarding marker spacing and sibship size associated with MIM and M/S was performed on the same data set. Trait values on parents were not used in the analysis but parental genotypes were used when they were available. The critical value for MIM was chosen to be χ^2 of 9.2 (asymptotically equivalent to a lod score of 2), and for M/S a lod score of 2. Both programs were run on a DEC-Alpha Unix workstation.

Defining True and False Positives

With genome scanning data, it is often not an easy task to define false or true positives. For instance, when two peaks appear to be close, whether one should call it one or two positives can be ambiguous. For the sake of simplicity, we used the following definitions: on each chromosome, we only allow for a maximum of one 'hit' (lod score or χ^2 exceeding the threshold) per chromosome. If the hit is located on one of the two true trait locus, it is defined as true positive, otherwise it is counted as a false positive. Given that the primary purpose of the study was a comparison of methods and study design, this somewhat simplistic definition should be adequate.

Effect of Parental Genotyping

As it was also our goal to compare the difference in power when the pedigrees were analyzed with or without parental genotypes, we performed the following simulation. We generated 100 pedigrees with four offspring for each family, and wrote out two sets of files. In the first set, we provided both parental and offspring genotypes, but in the second we only provided genotypes for the offspring. Both data sets were analyzed simultaneously.

Results

Table 1 shows that reasonable power was achieved by MIM at a variety of marker densities. With MIM, the true and false positive counts were generated using a threshold of χ^2 of 9.2 (asymptotically equivalent to a lod score of 2). With M/S, only the independent pairs were used in the analysis and the threshold was also picked to be a lod score of 2. With either program, 10 cM provided a good trade off between power and type I error under the assumed model. It can be directly observed that MIM had higher power and generated fewer false positives. However, one could argue that the relatively poor performance of M/S is due to the use of only independent pairs and is particular to the sibship size used. More simulations were carried out for data sets with a variety of sibship sizes which were analyzed by MIM and by M/S with different ‘pair used’ options.

Table 1 Comparison of power and false positive rate

	MIM		*MAPMAKER/SIBS*	
Map density	*Empirical Power*[a]	*FP*[b]	*Empirical Power*[a]	*FP*[b]
5cM	87%	1.01	60%	2.92
10cM	86%	0.99	56%	2.81
20cM	78%	0.96	46%	2.01

[a]Percentage of 1000 replicates in which at least one QTL was detected; [b]Average number of false positives per genome.

Table 2 summarizes the results of power comparison with the false positive rate fixed at one per genome. As expected, the relative advantage of MIM over M/S appeared to increase with sibship size. For example, the power of detecting at least one locus is 59% with 300 sib pairs, and jumps to 84% when 100 pedigrees of sibship size of four were genotyped. Clearly, when the false positive rate is held as a constant, the order of power performance is as the follows: MIM > all possible pairs with correction > independent pairs. Even with a sibship size of two, to our surprise, given its feature of using all markers for computing accurate IBD sharing probability, the power of M/S is lower than MIM.

Table 3 is generated using the same data set as Table 2, excluding all the parental genotypes. Interestingly, the change in power or false positive rate is minimal in all cases.

Discussion

Evaluation of MIM

The methodology for partitioning genetic variance of a quantitative trait to specific chromosomal regions was first proposed by Goldgar in 1990.[4] Although preliminary power studies of this method were subsequently carried out,[11,13] this paper provides a more complete evaluation of its power and significance level for genome search and its performance relative to another

Table 2 Sib pairs *vs* sibships with parental genotypes

	MIM		*MAPMAKER/SIBS*			
Sibship			*IND*[c]		*APWC*[d]	
size	*Threshold*[a]	*Power*	*Lod*[b]	*Power*	*Lod*[b]	*Power*
2	1.73	59%	2.61	31%	N/A	
3	1.79	71%	2.62	30%	1.77	42%
4	2.00	83%	2.66	30%	1.81	59%
5	2.02	89%	2.73	33%	1.86	69%

[a]lod threshold (converted from □^{2}) to satisfy average number of false positives of 1 per genome; [b]lod threshold to satisfy average number of false positives of 1 per genome; [c]independent pairs; [d]all pairs with correction.

Table 3 Sib pairs *vs* sibships without parental genotypes

	MIM		*MAPMAKER/SIBS*			
Sibship			*IND*		*APWC*	
size	*Threshold*[a]	*Power*	*Lod*[b]	*Power*	*Lod*[b]	*Power*
2	1.84	48%	2.71	28%	N/A	
3	1.93	61%	2.74	28%	1.78	36%
4	2.07	79%	2.90	29%	1.90	53%
5	2.08	83%	2.97	28%	1.91	62%

[a]lod threshold (converted from □^{2}) to satisfy average number of false positives of 1 per genome; [b]lod threshold to satisfy average number of false positives of 1 per genome.

multipoint quantitative mapping approach based on sib-pair IBD sharing.

First of all, our simulation shows that the power gained by a 5 cM map is not significantly greater than what is gained by a 10 cM map, but resulted in a slightly higher false positive rate. Therefore, a 10 cM map seems to be optimal, at least for the model investigated. The same conclusion is also reached when M/S is used in the analysis. Second, the simulation also shows the power of detecting linkage with MIM increases with sibship size; the power difference between two sibs and three sibs, three sibs and four sibs appears to be similar, but the difference between four sibs and five sibs is less dramatic. With 100 pedigrees of sibship size of four, the empirical power of detecting linkage under the assumed model is over 80% with a false hit of approximately one per genome. It should be noted that all the MIM analyses were performed under true parameters h^2 = 0.50); however, it has been our experience that an overestimation of heritability may result in a biased estimate of P, the proportion of genetic variance due to loci of the tested region, but does not cause loss of power or inflation of type I error.[11]

In summary, based on our simulation results, we believe typing a reasonable number of nuclear families with a map of genetic markers evenly spaced at 10 cM intervals is a good strategy for mapping QTLs using MIM in a genomic scan. If a χ^2 of 9.2 or higher is obtained in a data set at one or a few regions, one should consider the linkage results worth reporting and, more importantly, such regions should be revisited with denser markers. We recommend that the investigators include large size of sibships whenever possible to examine the complex traits under scrutiny.

MIM versus M/S

Our results indicated that when the threshold is chosen to give an average number of false positives of one per genome, with various sibship size, MIM consistently has higher power, as compared with M/S. When only independent pairs were used, the power loss is clearly due to the fact that the sibship information is not fully used. Even though power increases as all possible pairs are included in the analyses but corrected for statistically, it still remains significantly lower than those with MIM. The observation may be explained as follows:

1) MIM uses sibship as a whole instead of breaking it down into sibships, thus avoiding the correction problem existing in sib-pair oriented approach.
2) M/S calculates IBD sharing at each single point whilst MIM provides interval estimates; consequently M/S performs twice as many tests as does MIM (for the 10 cM map).

Another advantage of the sibship-oriented approach is its capability of reducing the amount of genotyping. For instance, when one uses 100 pedigrees with sibship size of four, the individuals to be typed amount to 600 instead of 1200 when 300 sibpairs are selected. If parents are also collected, naturally the genotyping saving advantage with larger sibship size becomes even more evident. Another interesting issue raised by this simulation is whether it is necessary to extract genotypic information from the whole chromosome to compute IBD sharing since reasonable amount of power was achieved with a 30 cM interval by MIM, although this may somewhat depend on the variation of map density and marker polymorphism.

The good news conveyed by this exercise is that parental genotypic information does not seem to be critical to achieving high power. From Table 3, we see that, when the parents are not genotyped in a sibship, the power loss is not as substantial using either MIM or M/S, even when sib pairs are analyzed. A similar conclusion was reached by Holmans and Craddock[14] based on their simulation results, which were analyzed with a single point approach. This fact can not only cut down a substantial amount of genotyping, but also be of benefit to investigations on any late onset traits such as rheumatoid arthritis where the parents are often unavailable. But it should be noted that when parents' genotypes are available, they are helpful in terms of error checking for marker genotypes, and that power may drop substantially if allele frequencies are misspecified.

General Problems of Sib-pair Oriented Approaches

Another important question to be addressed in the sib-pair based method is how to extract information on each sib pair from the sibship. Since the sib pairs are not independent of each other, the genome wide type I error rate based upon such dataset will be biased upwards, if all pairs are treated as independent pairs. One solution furnished by Suarez and Hodge[15] is to scale down the contribution of each pair by a factor of 2/*a*, where *a* is the number of sibs. In the most recent version of M/S, it allows three options in terms of the use of sib pairs breaking for sibship data:

1) use the first pair;

2) use independent pairs;

3) use all possible pairs.

The third option applied the Suarez-Hodge correction. We observed that when we fixed the false positive rate at 1 per genome, we found that the power obtained using all possible pairs is higher than that for independent pairs, which implies the correction scheme is conservative. However, to obtain a false positive of 1, the thresholds are much different depending upon which 'pairs used' option one chooses. This discrepancy was not discussed in M/S documentation and therefore may confuse some inexperienced users. One possible solution is to consider using a lower threshold jointly with a second variable such as the length of the region containing suggestive signals.[16] Of course, one can routinely carry out a simulation based on a specific data set to determine an appropriate threshold for a desired false positive rate.

Recently, a study by Sham *et al*[17] concluded that the optimal weighting scheme may depend on the frequency of the susceptibility gene and mode of transmission, based on the comparisons of different correction schemes. They also challenged the statistical validity of the Suarez-Hodge correction scheme and proposed a weighting scheme based on the means and variances of the contribution from a number of independent observations. Therefore, as a general recommendation, for the sib-pair oriented softwares, the choice for different correction schemes is better left to the user to define although in the new version of M/S, if the user chose the 'all pairs' option, then the likelihood was automatically corrected with the Suarez-Hodge weighting scheme.

Directions for MIM Extensions

M/S was developed as a novel multipoint method which accommodates substantial numbers of markers and sib pairs. Even though it is computationally more intensive than Goldgar's multipoint approach which is based on computing the average sharing over a region, our simulations indicate that MIM has higher power at all sibship sizes. This result is hardly surprising since MIM is a sibship-oriented approach and it uses covariance structure of the observations. It is somewhat striking to observe that the gain in power based on presumably more complete IBD information is not substantial. This was also observed by Fulker and Cherny[2] in their simulations. We will further explore the issue of whether or not an approach which combines the strength of the sibship-oriented method with accurate IBD sharing calculation would optimize the efficiency of a genome search and develop a strategy of using the multipoint IBD information computed by GENEHUNTER (GH) in MIM for multipoint interval mapping.

We are also making an effort to extend the MIM method to handle extended pedigrees which may involve using IBD sharing probability calculated by GH or GH$^+$ since Williams *et al*[19] demonstrated that use of complete pedigree information greatly increases power and efficiency.

Another task demanded by the nature of the complex traits is to extend MIM for mapping two or more QTLs simultaneously. The computational algorithm based on this idea has initially been tested by Lewis and Kort.[20] Compared with most parametric methods (TLINK, for example), MIM has the advantage of being developed into a method to handle multiple loci because it has a likelihood-computing engine and contains fewer genetic parameters. As shown by Lewis and Kort, one can propose a simplistic but practical extension of MIM to test for linkage to a quantitative trait in multiple (more than two) nonlinked genetic regions, using a genome scan approach.

Acknowledgements

This work was supported in part by a National Institutes of Health grant HG00571 (to DEG), by a special training grant from the International Agency for Research on Cancer (to YYS), and by an ACS grant IRG-60-002-39-IRG (to YYS).

References

1 Penrose: 1938.
2 Haseman JK, Elston RC: The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 1972; **2**: 3–19.
3 Amos CI: Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* 1994; **54**: 535–543.
4 Goldgar DE: Multipoint analysis for human quantitative genetic variation. *Am J Hum Genet* 1990; **47**: 957–967.
5 Schork NJ: Extended multipoint identity-by-descent analysis of human quantitative and qualitative traits. *Am J Hum Genet* 1993; **57**: 439–454.
6 Fulker DW, Cardon LR: A sib-pair approach to interval mapping of quantitative trait loci. *Am J Hum Genet* 1994; **54**: 1092–1103.
7 Blangero J: Multivariate oligogenic linkage analysis of quantitative traits in general pedigrees. *Am J Hum Genet* 1995; **57**: A11.
8 Fulker EW, Cherny SS, Cardon LR: Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am J Hum Genet* 1995; **56**: 1224–1233.

9 Cardon LR, Smith SD, Fulker DW, Kimberling WJ, Pennington BR, DeFries JC: Quantitative trait locus for reading disability on chromosome 6. *Science* 1994; **266**: 276–279.

10 Kruglyak L, Lander ES: Complete multipoint sib-pairs analysis of qualitative and quantitative traits. *Am J Hum Genet* 1995; **57**: 439–454.

11 Goldgar DE, Oniki R: Comparison of a multipoint IBD method with parametric multipoint linkage analysis for mapping quantitative traits. *Am J Hum Genet* 1992; **50**: 598–606.

12 Terwilliger JD, Speer M, Ott J: Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genet Epidemiol* 1993; **10**: 217–224.

13 Goldgar, Lewis: 1994.

14 Holmans P, Craddock N: Efficient strategies for genome scanning using maximum-likelihood affected-sib-pair analysis. *Am J Hum Genet* 1997; **60**: 657–666.

15 Suarez BD, Hodge SE: A simple method to detect linkage for rare recessive diseases: an application to juvenile diabetes. *Clin Genet* 1997; **15**: 126–136.

16 Shugart YY, Ott J: Using lod-score peak length to distinguish true and false positives. In: Clementi M, Forabosco P (eds). *Mathematical Genetics*. CLEUP: Padua, 1996, pp 127–133.

17 Sham PC, Zhao JH, Curtis D: Optimal weighting scheme for affected sib-pair analysis of sibship data. *Ann J Hum Genet* 1997; **61**: 61–69.

18 Fulker EW, Cherny SS: Improved multipoint sib-pair analysis of quantitative traits. *Behav Genet* 1996; **26**: 527–532.

19 Williams JT, Duggirala R, Blangero J: Statistical properties of a variance-components method for quantitative trait linkage analysis in nuclear families and extended pedigrees. In: Goldin LR, Beiley-Wilson JE, Borecki IB *et al* (eds). Genetic analysis workshop 10: detection of genes for complex traits. *Genet Epidemiol* 1997; **14**: 987–992.

20 Lewis CM, Kort EN: Multilocus quantitative trait analysis using the multipoint identity-by-descent method. In: Goldin LR, Beiley-Wilson JE, Borecki IB *et al* (eds). Genetic analysis workshop 10: detection of genes for complex traits. *Genet Epidemiol* 1997; **14**: 839–844.