Selecting Pedigrees for Linkage Analysis of a Quantitative Trait: The Expected Number of Informative Meioses

Michael Boehnke, Katherine H. Omoto, and Jean Marie Arduino

Department of Biostatistics, University of Michigan, Ann Arbor

Summary

With evidence of segregation at a major locus for a quantitative trait having been found, a logical next step is to select a subset of the pedigrees to include in a linkage study to map the major locus. Ideally this subset should include much of the linkage information in the sample but include only a fraction of the pedigrees. We previously described a strategy for selecting pedigrees for linkage analysis of a quantitative trait on the basis of a pedigree likelihood-ratio statistic. For quantitative traits controlled by a major locus with a rare dominant allele, the likelihood-ratio strategy extracted nearly all the information for linkage while typically requiring marker data on only about one-third of the pedigrees. Here, we describe a new strategy to select pedigrees for linkage analysis on the basis of the expected number of potentially informative meioses in each pedigree. We demonstrate that this informative-meiosis strategy provides an efficient and more general means to select pedigrees for a linkage study of a quantitative trait.

Introduction

Once we have found evidence of segregation at a major locus for a quantitative trait, we may then wish to attempt to map the locus. While the linkage study could be carried out on the entire sample of pedigrees, a useful intermediate step is to identify those pedigrees most likely to be most informative for linkage. These pedigrees can then be typed for a battery of genetic markers. However, for quantitative traits, an investigator cannot simply choose families in which the trait of interest is present, as can be done for simple Mendelian diseases; in general there is not any obvious pattern that identifies pedigrees that will be most informative.

In a previous paper (Boehnke and Moll 1989), we described a pedigree test statistic that compares (a) the likelihood of a pedigree under a model allowing for a major locus with (b) the likelihood of the pedigree under the model that results when the major locus is excluded. Such model pairs include the major-locus and random models and the mixed-genetic and polygenic models. To compare the likelihood of the pedigree under the major-locus and random models (see below), we defined the pedigree test statistic

\[ S = \log L_{ML}( \mu_{AA}, \mu_{AA}, \sigma; x) - \log L_{R}( \mu_{AA}, \sigma; x). \]  

(1)

Here, \( L_{ML} \) and \( L_{R} \) are the major-locus-model and random-model likelihoods for the pedigree; \( q \) is the frequency of the rarer allele A; \( \mu_{AA}, \mu_{AA}, \) and \( \mu_{AA} \) are the major-locus genotype means; \( \sigma \) is the within-distribution SD; \( x \) is the vector of quantitative trait values for the pedigree members; and \( A \) indicates the maximum-likelihood estimate for the major-locus model (Boehnke and Moll 1989). The larger the value of \( S \), the stronger the suggestion of the presence of a major locus with two alleles. A pedigree was selected for linkage analysis if \( S > 0 \), that is, if the pedigree was more likely under the model that included the major locus with two alleles than under the model that excluded major-locus segregation. Alternatively, those pedigrees with the largest test statistic values could be selected.

Given a major locus with one rare and one common allele, both the information for linkage and the pedigree likelihood-ratio statistic are strongly associated with the number of rare alleles present in the pedigree. For the rare-dominant major-locus and mixed models we
considered, the likelihood-ratio strategy made it possible to extract nearly all the linkage information in a set of pedigrees while typing genetic markers on only about one-third of the pedigrees.

However, for quantitative traits influenced by a major locus with two common alleles, both major-locus alleles are likely to be present in most pedigrees, making the grounds for the above likelihood-ratio strategy less secure. For such models, the linkage information is no longer strongly associated with the number of copies of a particular allele in the pedigree. Indeed, the presence of both major-locus alleles is needed, so that an alternative selection criterion is required.

In the present paper, we describe a second statistical strategy to select pedigrees for a linkage study of a quantitative trait. This new strategy estimates the expected number of potentially informative meioses in a pedigree by introduction of an extra constrained parameter in the pedigree likelihood. We find that the informative-meioses strategy provides an efficient means for selecting a subset of pedigrees for use in the linkage analysis of a quantitative trait both in the rare-dominant-allele case where the likelihood-ratio strategy works well and in the common-alleles case to which the likelihood-ratio strategy is not well suited.

A New Method to Select Pedigrees for Linkage Analysis

Since information for linkage is present only in families in which the parental origin of the offspring alleles can be inferred, we could define an informative pedigree as one in which the number of informative meioses is large. Unfortunately, for a quantitative trait generally we cannot even determine which parents are heterozygous at the trait locus, let alone which meioses are informative. However, as we presently show, it is possible to calculate the expected value of the number of potentially informative meioses conditional on the observed pedigree quantitative-trait values and the genetic model. We can then select those pedigrees with the largest expected number of potentially informative meioses for subsequent linkage analysis. In what follows, we drop “potentially” from “potentially informative meioses,” with the understanding that it is implied.

Genetic Models

We consider the case of a quantitative trait determined by the summed effects of a major locus with alleles A and a and a normally distributed individual-specific environmental term. Hence, the population trait distribution is a mixture of normals. Parameters of the major-locus model include the frequency $q$ of the allele A, means $\mu_{AA}$, $\mu_{Aa}$, and $\mu_{aa}$ for the major-locus genotype distributions, and within-distribution SD $\sigma$. If $\mu_{AA} = \mu_{AA}$, the model is dominant at the major locus; if $\mu_{AA} - \mu_{aa} = \mu_{AA} - \mu_{AA}$, the model is additive at the major locus. The model that results when the gene frequency $q = 0$ is the random model. Under the random model, the trait is normally distributed, and each individual can be thought of as having the same major-locus genotype $aa$.

Calculating the Expected Number of Informative Meioses

For parent-offspring trios in which all three individuals have been measured for the quantitative trait, we count as an informative meiosis each transmission of an allele from parent to offspring for which we can infer with certainty which parental allele was passed. Thus, offspring of homozygote × homozygote matings provide no informative meioses, offspring of heterozygote × homozygote matings provide one informative meiosis, and offspring of heterozygote × heterozygote matings provide zero informative meioses if heterozygous or two informative meioses if homozygous. There is a direct correspondence between informative meiosis as defined here and the PIC (Botstein et al. 1980).

To calculate the expected number of informative meioses, we require the following notation: Let $x = (x_1, \ldots, x_n)$ and $g = (g_1, \ldots, g_n)$ be the observed quantitative-trait vector and a possible major-locus genotype vector, respectively, for a pedigree of size $n$. Let the random variable $M$ be the number of informative meioses in a pedigree, and let $m(g)$ be the fixed number of informative meioses corresponding to the specific major-locus genotype vector $g$. Then the conditional expectation of $M$ given the observed quantitative-trait vector $x$ can be calculated as

$$E[M|x] = \sum_g m(g) P(g|x)$$

$$= \frac{\sum_g m(g) P(x|g) P(g)}{P(x)} , \quad (2)$$

where $P(\cdot)$ is probability (density), $P(\cdot|\cdot)$ is conditional probability (density), and the summations range over all possible major-locus genotype vectors $g$. The denominator in equation (2) is the likelihood of the observed quantitative-trait vector $x$ for the pedigree and can be evaluated using any of the standard pedigree-analysis programs.

To calculate the numerator of equation (2), we in-
introduce a dummy parameter $\lambda$ into the pedigree likelihood. Define

$$H(\lambda) = \sum_g \left[ P(x|g) \prod_i P(g_i) \prod_j P(g_j|g_k,g_l) \right] \lambda^{m(g)} ,$$

where the sum again ranges over all possible major-locus genotype vectors $g$, $i$ ranges over all pedigree members without parents in the pedigree, $j$ ranges over all pedigree members with parents in the pedigree, and $k$ and $l$ represent the parents of $j$. Since whether the meioses that produced a particular offspring are informative can be determined solely on the basis of parent and offspring genotypes, $H(\lambda)$ can be calculated by multiplying each transmission probability $P(g_i|g_k,g_l)$ by $\lambda$ raised to the power 0, 1, or 2, depending on whether offspring $j$ provides 0, 1, or 2 informative meioses. Computation of $H(\lambda)$ can then be carried out using the Elston and Stewart (1971) algorithm. Setting $\lambda = 1$, $H(1)$ is equal to the likelihood of the pedigree data $P(x)$. Differentiating $H(\lambda)$ with respect to $\lambda$ at $\lambda = 1$ gives the numerator of equation (2).

In fact, most pedigree-analysis programs calculate the logarithm of the pedigree likelihood. Since the derivative $d[\log H(\lambda)]/d\lambda = H'(\lambda)/H(\lambda)$ is equal to $E[M|x]$ when $\lambda = 1$, the expected number of informative meioses can be calculated by introducing the parameter $\lambda$ into the transmission probabilities as described above, constraining $\lambda$ to equal one, and numerically evaluating the score with respect to $\lambda$. If model parameters are known, calculation of $E[M|x]$ can be carried out by numerically evaluating the score with respect to $\lambda$ with model parameters set to their known values. Since model parameters generally are unknown, constraining $\lambda = 1$ and estimating model parameters by iterative maximum-likelihood methods yields $E[M|x]$ as the score with respect to $\lambda$ at the final iteration. This approach has been used by Lange and Matthysse (1989) to calculate the expected number of offspring of heterozygote $\times$ heterozygote matings for a recessive trait and by Weeks and Lange (submitted) to calculate the expected number of offspring of major-locus heterozygotes.

Having calculated the expected number of informative meioses in each of a set of pedigrees, we can then select for linkage analysis those pedigrees with the largest expected number of informative meioses. The number of pedigrees to select can be determined on the basis of the available resources or on the basis of a subsequent power calculation (Ploughman and Boehnke 1989; also see Discussion).

**Simulation**

**Methods**

To compare the performance of the likelihood-ratio and informative-meioses pedigree-selection strategies, we turned to computer simulation. We generated pedigree data with a quantitative trait under a major-locus model and also with a fully informative genetic marker totally linked ($\theta = 0$) to the major locus. Two major-locus models were considered for the quantitative trait: (1) a dominant model with genotype-specific means 100 and 120, within-distribution $SD = 10$, and allele frequency $q = .012579$ (2.5% of the population in the upper distribution) (Boehnke and Moll 1989) and (2) an additive major-locus model with genotype specific means 100, 115, and 130, within-distribution $SD = 10$, and allele frequency $q = .50$. Four pedigree structures were considered (fig. 1). They included a nuclear family of size 5, 3-generation pedigrees of sizes 9 and 15, and a large, multigeneration pedigree of size 45 (Burns et al. 1984; Boehnke and Moll 1989).

For each combination of the two genetic models and the four pedigree structures, 100 samples of 450 individuals each were generated. For the rare dominant major-locus model, pedigrees were singly ascertained through the upper 5% of the quantitative-trait distribution, independent of marker phenotype. Probands are noted by arrows (fig. 1). For the common additive major-locus model, pedigrees were randomly sampled. Generation of the quantitative-trait values and marker phenotypes and sampling of the pedigrees were as described by Boehnke et al. (1988) and Boehnke and Moll (1989).

![Pedigree Structures](image.png)

**Figure 1** Pedigree structures of size 5, 9, 15, and 45. For the dominant model, pedigrees were singly ascertained through the probands indicated by the arrows.
For each replicate data set, we calculated the major-locus and random-model log likelihoods and the pedigree likelihood-ratio statistics by using PAP (Hasstedt and Cartwright 1981). Log likelihoods were maximized using GEMINI (Lalouel 1979). For the rare dominant model, we corrected for ascertainment by calculating the likelihoods conditional on the trait value of the probands (Boehnke and Lange 1984). We used MENDEL (Lange et al. 1988) to calculate the expected number of informative meioses for each pedigree as described above. Lod scores (Morton 1955) were calculated using LI PED (Ott 1974, 1976).

**Results**

Results of the computer simulation are presented in figures 2 and 3. These figures display the proportion of the maximum summed lod score for the combined data sets, obtained as a function of the proportion of pedigrees selected for our two pedigree-selection strategies. For the rare dominant model (fig. 2), both the likelihood-ratio and the informative-meioses strategies very efficiently selected a subset of the pedigrees for linkage analysis. For all four pedigree structures, both strategies obtained at least 92% of the linkage information for the entire data set, as measured by maximum summed lod score, when selecting only one-third of the pedigrees. While the efficiencies of the two strategies were very similar, pedigree subsets selected using the informative-meioses strategy were generally slightly more informative for linkage than were the pedigree subsets of the same size selected by the likelihood-ratio strategy. This slight increase in efficiency was more apparent for the smaller pedigrees.

For the common additive model, the informative-meioses strategy was clearly superior to the likelihood-ratio strategy for all pedigree structures considered, particularly for the smaller pedigree structures (fig. 3). Indeed, the likelihood-ratio strategy resulted in subsets of pedigrees no more informative for linkage than were a random sample of the available pedigrees. While the informative-meioses strategy provided less improvement in the common additive case than in the rare dominant case, subsets could be selected that gave substantially more linkage information per pedigree than would have resulted from a random selection of pedigrees.

**Discussion**

Our results suggest that the expected number of informative meioses provides a useful criterion for selecting pedigrees for a linkage study of a quantitative trait. For the rare-dominant model we considered, both
pedigree-selection strategies provided a very efficient means of selecting a subset of the pedigree data for linkage analysis, with the informative-meioses strategy resulting in slightly more efficient subsets of the data. Results for the additive model strongly suggest that the informative-meioses strategy is useful in the case of the common additive major locus, while the likelihood-ratio strategy does no better than random selection of pedigrees.

While the informative-meioses strategy improved the efficiency of a linkage study for the common additive model, it was not as helpful as in the case of the rare-dominant model. This is not surprising. For the dominant model, the rare allele A was present in only a small subset of the pedigrees (about one-third of the 5-, 9-, and 15-person pedigrees and about one-half of the 45-person pedigrees). Only these pedigrees could provide linkage information. Thus, a subset of the pedigrees could potentially provide essentially all the linkage information. In contrast, for the common additive model most of the pedigrees provided some information for linkage, and this proportion increased with pedigree size. Hence, even if the most informative pedigrees were selected, some information was bound to be lost, particularly as pedigree size increased. This is consistent with the observation that in the case of the additive model, the informative-meioses strategy was most useful for the smaller pedigrees.

Explaining the observation that the informative-meioses strategy resulted in more informative subsets of the data, particularly for the common additive case, is not difficult. The pedigree likelihood-ratio statistic specifically assays the presence of one of the alleles at the major locus. If that allele is rare, its presence suggests an informative pedigree. In contrast, when both alleles are common, the presence of a particular allele is in no way predictive of an informative pedigree. The presence of both alleles is required if the pedigree is to be informative for linkage. Indeed, it is not simply the presence of both alleles but also the genotypic combinations in which they appear that is critical. For example, for the 5-person nuclear family, an Aa × Aa mating with three aa offspring is maximally informative for the additive model, despite including only two A alleles and thus resulting in a modest pedigree likelihood-ratio statistic. In contrast, an AA × AA mating with all AA offspring provides no information for linkage but results in a very large pedigree likelihood-ratio statistic.

While we considered only the cases of dominant and additive major-locus models, an analogous selection strategy could be used for a recessive major-locus model. In that case, one might consider the expected number of offspring of major-locus heterozygote × heterozygote matings (Lange and Matthysse 1989; Weeks and Lange, submitted). Sample sizes required to detect linkage would likely be much greater.

Variants of the informative-meioses strategy might be considered. In particular, when defining informative, it might be preferable to differentiate between (a) phase-known meioses that might result given 3 generations of quantitative trait information and (b) phase-unknown meioses that might result given only 2 generations. For a fully informative marker, it might be possible to write down formulas similar to equation (2) to calculate the expected number of such phase-known and phase-unknown informative meioses, given quantitative-trait information for the pedigree. However, major-locus genotypes on 3 generations would be required. The resulting formulas would not factor nicely in the framework of the Elston and Stewart (1971) algorithm, making computation much more difficult.

If pedigrees are of variable size and structure, simply comparing expected number of informative meioses may not be an appropriate strategy. Pedigree size should probably also be taken into account. One approach would be to select pedigrees or pedigree branches on the basis of the expected number of informative meioses divided by a measure of the cost of sampling,—e.g., the number of members of the pedigree or pedigree branch.

In principle, the best approach would be to directly estimate the expected lod score for each available pedigree by using the method of Ploughman and Boehnke (1989). Those pedigrees with the largest expected lod scores—or largest expected per-person lod scores, if the pedigrees were of variable size—could then be selected for linkage analysis. This strategy would guarantee that the most informative set of pedigrees is selected. However, such an approach would be much more time-consuming than the informative-meioses strategy described here. A logical intermediate might be to (1) order pedigrees according to the informative-meioses strategy, (2) calculate the expected lod scores for the pedigrees that appear better on the basis of the expected number of informative meioses, and (3) use expected lod score as the final selection criterion.

With any likelihood-based strategy to select pedigrees for a linkage study, success obviously depends on both the correctness of the model and the accuracy of the parameter estimates. Thus, a smaller sample may result in less informative selection, since parameters will
generally be less accurately estimated. More important, if the genetic model itself is incorrect, we cannot expect these pedigree-selection strategies to work well at all. However, given good evidence for a genetic model and even the relatively modest sample size of 450 individuals used in the computer simulation, our results suggest that the informative-meoise strategy offers an attractive means of selecting pedigrees for the linkage analysis of a quantitative trait under the influence of a major locus.

Acknowledgments

We thank Kenneth Lange for suggesting a similar strategy for selecting pedigrees for linkage analysis of a quantitative trait and for reading and commenting on a previous version of this paper. We thank Sandra Hasstedt, Jean-Marc Lalouel, Kenneth Lange, and Jurg Ott for developing and supporting PAP, GEMINI, MENDEL, and LIPEP, respectively. Support for this work was provided by NIH grant R29-GM41440 to M.B. Computing support was provided by the University of Michigan.

References


Hasstedt SJ, Cartwright PE (1981) PAP: pedigree analysis package. Tech rep no 13, Department of Medical Biophysics and Computing, University of Utah, Salt Lake City

Lalouel JM (1979) GEMINI: a computer program for optimization of general nonlinear functions. Tech rep no 14, Department of Medical Biophysics and Computing, University of Utah, Salt Lake City


Weeks DE, Lange K (1989) Trials, tribulations, and triumphs of the EM algorithm in pedigree analysis (submitted)