

Genome Sharing in Large Pedigrees: Multiple Imputation of *ibd* for Linkage Detection

Elizabeth Thompson Saonli Basu

Department of Statistics, University of Washington, Seattle, Wash., USA

Key Words

Identity by descent · Inheritance vector · Multiple imputation · MCMC

Abstract

Our objective is the development of robust methods for assessment of evidence for linkage of loci affecting a complex trait to a marker linkage group, using data on extended pedigrees. Using Markov chain Monte Carlo (MCMC) methods, it is possible to sample realizations from the distribution of gene identity by descent (*ibd*) patterns on a pedigree, conditional on observed data \mathbf{Y}_M at multiple marker loci. Measures of gene *ibd* W which capture joint genome sharing in extended pedigrees often have unknown and highly skewed distributions, particularly when conditioned on marker data. MCMC provides a direct estimate of the distribution of such measures. Let W be the *ibd* measure from data \mathbf{Y}_M , and W^* the *ibd* measure from pseudo-data \mathbf{Y}_M^* simulated with the same data availability and genetic marker model as the true data \mathbf{Y}_M , but in the absence of linkage. Then measures of the difference in distributions of W and W^* provide evidence for linkage. This approach extracts more information from the data \mathbf{Y}_M than either comparison to the pedigree prior distribution of W or use of statistics that are expectations of W given the data \mathbf{Y}_M . A small example is presented.

Copyright © 2003 S. Karger AG, Basel

Introduction

Gene Descent in Pedigrees

Analyses of genetic marker data for purposes of trait linkage detection rely on imputed patterns of gene identity by descent (*ibd*) arising from the segregation of genes in the meioses of a pedigree. Meiosis is the biological process underlying the transmission of genetic information from a parent to an offspring. In diploid individuals, on a pedigree structure with m relevant meioses, meiosis outcomes are most conveniently summarized by binary indicators $\mathbf{S} = \{S_{i,j}; i = 1, \dots, m, j = 1, \dots, L\}$:

$$S_{i,j} = 1 \text{ or } 0 \text{ as the paternal or maternal parental gene is transmitted in meiosis } i \text{ at locus } j \quad (1)$$

Each non-founder individual has a maternal and a paternal meiosis, and the pedigree structure is determined by a specific labeling of these meioses. The indicators $S_{i,j}$ then specify the descent of all genes in a pedigree, and hence the founder origin of every allele at every locus j in each individual [1, 2]. For convenience, define $S_{.,j} = \{S_{i,j}; i = 1, \dots, m\}$ and $S_{i,.} = \{S_{i,j}; j = 1, \dots, L\}$. Gene identity by descent (*ibd*) at marker locus j is a function of the inheritance vector $S_{.,j}$ and more generally *ibd* at chromosomal location λ is a function of $S_{.,\lambda}$. The vectors $S_{i,.}$ are independent, while, in the absence of genetic interference, $S_{.,j}$ are first-order Markov in j [3].

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2003 S. Karger AG, Basel

Accessible online at:
www.karger.com/hhe

Dr. Elizabeth A. Thompson
Department of Statistics, University of Washington
Box 354322
Seattle, WA 98195-4322 (USA)
Tel. +1 206 543 7237, Fax +1 206 685 7419, E-Mail thompson@stat.washington.edu

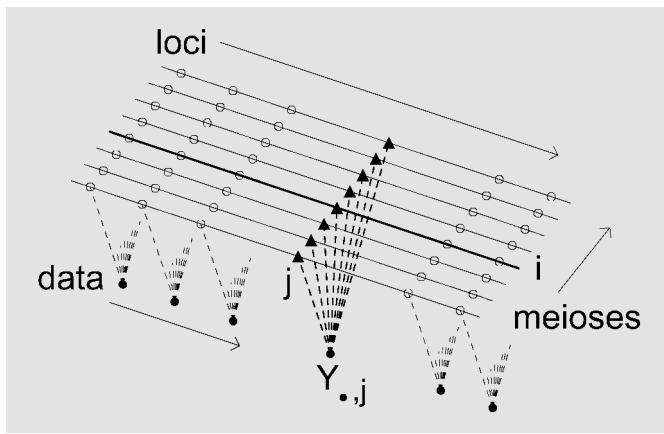


Fig. 1. The dependence structure of pedigree data. Any genetic marker data on the pedigree corresponding to locus j is denoted by $Y_{.,j}$. Given the full array \mathbf{S} , $Y_{.,j}$ depends only on the inheritance vector $S_{.,j}$ at locus j . In the absence of data, vectors $S_{i,.}$ are independent, and $S_{.,j}$ are Markov in j .

Assume now that genetic marker genotypes are available on some individuals at some of the L loci. We assume a known genetic model for the markers, including the genetic map and marker allele frequencies, and assume absence of linkage disequilibrium in the founders of the pedigree. Then, any observed genetic data at marker locus j depends only on the inheritance vector $S_{.,j}$ at that locus. Thus, in the absence of genetic interference, we have the hidden Markov probability structure shown in figure 1 in which the pedigree structure is implicit in the labeling of the meioses [4]. Given genetic marker data $Y_{.,j}$ at locus j on some pedigree members, $\Pr(Y_{.,j}|S_{.,j})$ can be easily computed [1, 5]. In principle, the Baum algorithm [6] permits computation of probabilities $\Pr(S_{.,j}|\mathbf{Y}_M)$ of inheritance patterns conditional on all the marker data \mathbf{Y}_M observed on a pedigree. In practice, however, even with modern programs [7] pedigree sizes are limited and computation is slow. Using MCMC methods, patterns of *ibd* can be realized jointly over individuals and jointly over loci, conditional on all marker data \mathbf{Y}_M observed on the pedigree [1, 2]. Such MCMC methods have the advantage that *ibd* can be scored jointly over loci.

In the current work we use a two block-Gibbs samplers to perform the MCMC, combined to form the ‘lm-sampler’ [1]. This sampler allows for gender-specific maps. Further, using importance sampling or incorporating Metropolis-Hastings acceptance steps, the samplers can allow for genetic interference and linkage disequilibrium among marker loci in pedigree founders [1]. However, in

the example of this paper we assume a sex-averaged map, no genetic interference, and no population-level allelic associations among loci.

Methods

Linkage Detection Using Gene *ibd*

Linkage detection tests based on the correlation in concordance and discordance of trait and marker phenotypes in pairs of sibs date back to Penrose [8], but it was Suarez et al. [9] who placed the problem in the context of probabilities of gene *ibd* at trait and marker loci. Gene *ibd* underlies patterns of phenotypic similarity among related individuals. Individuals concordant for a trait phenotype have increased probability of *ibd* at causal loci, and hence also at linked markers.

Whittemore and Halpern [10] developed an approach based on a scalar measure W of the patterns of *ibd* among affected individuals, at a particular chromosomal location λ . For example, W might be the sum over all pairs of affected individuals of the number of genes shared *ibd* between them. This approach is applicable to data on any pedigree structure provided the relevant probabilities and expectations can be computed. Whittemore and Halpern [10] proposed test statistics T of the form

$$T(\mathbf{Y}_M, \lambda) = E(W_\lambda | \mathbf{Y}_M) \text{ where } W_\lambda = W(S_{.,\lambda}). \quad (2)$$

There are two null distributions of interest. We use the subscript 0 to denote pedigree-based expectations in the absence of both linkage and data, and the superscript * to denote expectations over datasets \mathbf{Y}_M^* having the same genetic map, marker loci characteristics, and data availability as the actual data \mathbf{Y}_M , but absence of linkage. For any random variables, the well known properties of conditional expectation provide that

$$E^*(T) = E^*(E(W | \mathbf{Y}_M)) = E_0(W), \quad (3)$$

and

$$\begin{aligned} \text{var}_0(W) &= \text{var}^*(E(W | \mathbf{Y}_M^*)) + E^*(\text{var}(W | \mathbf{Y}_M^*)) \\ &= \text{var}^*(T) + E^*(\text{var}(W | \mathbf{Y}_M^*)) \geq \text{var}^*(T). \end{aligned} \quad (4)$$

Since $\text{var}^*(T) \leq \text{var}_0(W)$, using the pedigree-based moments of the latent W results in a conservative test for linkage based on the test statistic T .

Tests based on statistics T of the above form have been extended to pedigrees [5], and McPeck [11] has considered a variety of different measures W appropriate for small pedigrees. Where the marker data are highly informative as to the latent *ibd*, $E^*(\text{var}(W | \mathbf{Y}_M^*))$ is small and a useful test results. However, on extended pedigrees with substantial missing data, test statistics standardized by the upper bound $\text{var}_0(W)$ on the variance of T lead to unduly conservative tests [12, 13]. Moreover, the measures W and statistics T that take into account joint sharing among multiple affected individuals often have highly skewed distributions, making normality assumptions untenable [14]. Proposals to overcome these drawbacks have been made. For example, Churchill and Doerge [12] have proposed permutation tests, which are well suited to the standardized multi-offspring pedigrees of experimental organisms, but less so to extended pedigrees sampled from natural populations. The approach most closely analogous to the one adopted here is that of Davis et al. [13], where marker

data are resimulated for affected individuals in the assumed absence of trait loci in order to obtain an empirical distribution for T in the presence of data but absence of linkage. The estimated value of T given the observed marker data can be assessed against this empirical distribution.

Instead of using statistics of the form $T \equiv E(W|Y_M)$ (equation (2)), we propose to use directly the imputed distribution of W conditional on Y_M . Given actual data Y_M , we simulate additional N pseudo-datasets $Y_M^{*(i)}$, $i = 1, \dots, N$. As in equation (3), each pseudo-dataset $Y_M^{*(i)}$ has the same genetic marker model and data availability pattern as Y_M , but assumes absence of trait linkage. For any given chromosomal location λ , let W_λ denote the *ibd* measure given data Y_M , and $W_\lambda^{*(i)}$ the same measure conditioned on pseudo-dataset $Y_M^{*(i)}$. Of interest is then the difference between the distribution of W_λ and each $W_\lambda^{*(i)}$, $i = 1, \dots, N$. Two measures of such difference in distribution are considered below. Additionally, for comparison, we compute $T(Y_M, \lambda) = E(W_\lambda|Y_M)$ and each $T^*(Y_M^{*(i)}, \lambda) = E(W_\lambda^{*(i)}|Y_M^{*(i)})$ from the estimated distributions of W_λ and each $W_\lambda^{*(i)}$, $i = 1, \dots, N$, respectively.

The first measure of the imputed distribution of *ibd* measure W is based on the probabilities that a realization from a distribution conditioned on data arising in the absence of linkage would exceed a realization given the observed data Y_M . Specifically, we propose to measure evidence for linkage at location λ using probabilities of the form

$$p_0 = \Pr(W_0 \geq W_\lambda | Y_M)$$

and

$$p^{*(i)} = p^*(Y_M^{*(i)}) = \Pr(W_\lambda^{*(i)} \geq W_\lambda | Y_M^{*(i)}, Y_M) \quad i = 1, \dots, N, \quad (5)$$

where the subscript 0 and superscript * have the same meanings as above, and thus W_0 is a realization from the pedigree-based null distribution of W . Note that

$$E^*(p^*(Y_M)) = \Pr(W_0 \geq W_\lambda | Y_M) = p_0$$

where the expectation E^* is taken over dataset realizations Y_M^* generated in the absence of linkage.

Let F_0 denote the cumulative distribution function (cdf) of W_0 , F_λ denote the cdf of W_λ given the data Y_M , and $F_\lambda^{*(i)}$ denote the cdf of $W_\lambda^{*(i)}$ given $Y_M^{*(i)}$, $i = 1, \dots, N$. The cdf's F_0 , F_λ and $F_\lambda^{*(i)}$ are estimated for each of a set of chromosomal locations $\lambda \in \Lambda$ by MCMC given the pedigree structure only, the observed marker data Y_M , and each pseudo-dataset $Y_M^{*(i)}$, respectively. Then p_0 and each $p^*(Y_M^{*(i)})$ is computed from equation (5). The distribution of p^* values may be estimated by realizing multiple pseudo-datasets Y_M^* .

The probabilities p_0 and p^* values are analogous to p-values, in the sense of being a probability that a value under the null exceeds a value under the actual data. However, neither is a true p-value in the sense of having a $U(0, 1)$ distribution under the null hypothesis of absence of linkage but presence of marker data. The value p_0 is based on the (often used but incorrect) no-data no-linkage null. The distribution of p^* values in the absence of linkage in the actual data Y_M will be symmetric about 1/2, taking more extreme values at locations at which the data are more informative about *ibd*. For a location at which there is no *ibd* information $p^* = 1/2$, while for a location at which *ibd* is completely determined p^* is either 0 or 1.

Note that p^* can be written in terms of the imputed cdf's:

$$p^{*(i)} = \int_w F_\lambda(w) dF_\lambda^{*(i)}(w)$$

It is therefore of interest to consider other measures of the difference between F_λ and the set of $F_\lambda^{*(i)}$ which may serve to extract a linkage signal. In particular we consider the Anderson-Darling measure [15]. Define the average of the $N + 1$ cdf's as

$$G_\lambda(w) = (N + 1)^{-1} (F_\lambda(w) + \sum_{i=1}^N F_\lambda^{*(i)}(w)).$$

Then the distance measure

$$d(F, G) = \int_w \frac{(F(w) - G(w))^2}{G(w)(1 - G(w))} dG(w) \quad (6)$$

may be used to assess the deviation of F_λ and each $F_\lambda^{*(i)}$ from the average G_λ . These distances are well defined since the support of G_λ includes that of F_λ and each $F_\lambda^{*(i)}$. Due to the factor $G(w)(1 - G(w))$ in the denominator, differences in the tails of the distributions are emphasized.

Results

A Small Example

As an example consider the pedigree of figure 2. The pedigree derives from a real study [16], but for the purposes of this example all marker data are simulated. In accordance with the marker loci characteristics, genetic map, and true patterns of marker availability, marker data for 10 Chromosome 1 markers were simulated conditional on a partial specification of inheritance at a trait locus located close to the marker D1S479 (table 1). These data constitute Y_M , the actual marker data under study, and the goal is to assess the evidence for linkage provided by Y_M . As described above we resimulate datasets $Y_M^{*(i)}$, $i = 1, \dots, N$, analogous to Y_M in the marker model and data availability, but in the absence of linkage.

For the purposes of illustration we chose the following simple measure W_λ at each chromosomal location λ :

$$W_\lambda = \max_{x \in H} (\# \text{ affected individuals carrying genetic material at } \lambda \text{ from founder haplotype } x)$$

where H is the set of all founder haplotypes. Thus W_λ measures joint sharing of genetic material at chromosomal position λ among the affected individuals. The set of locations Λ consists of the 10 marker locations themselves plus an unlinked location which is routinely included as a check. Using MCMC to obtain multiple imputations of *ibd*, we estimate the cdf's F_0 of W_0 , F_λ of W_λ given Y_M , and $F_\lambda^{*(i)}$ of $W_\lambda^{*(i)}$ given $Y_M^{*(i)}$, $i = 1, \dots, N$. For this small example, each set of eleven cdf's over Λ either given Y_M or given a $Y_M^{*(i)}$ can be reliably estimated with 30,000 scans of the lm-sampler. This takes less than 5 min on a Compaq Alphaserver DS10 with a 466-MHz processor. Nonetheless we restrict our analysis to $N = 8$ pseudo-datasets $Y_M^{*(i)}$.

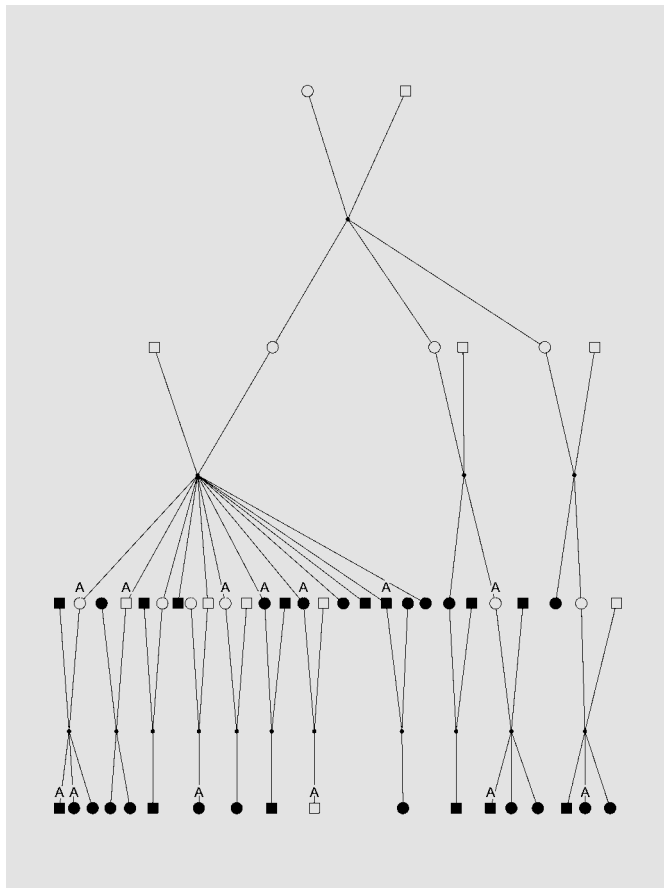


Fig. 2. Pedigree used for example. Individuals marked 'A' are affected, while those who are shaded have substantial, although not necessarily complete, marker data.

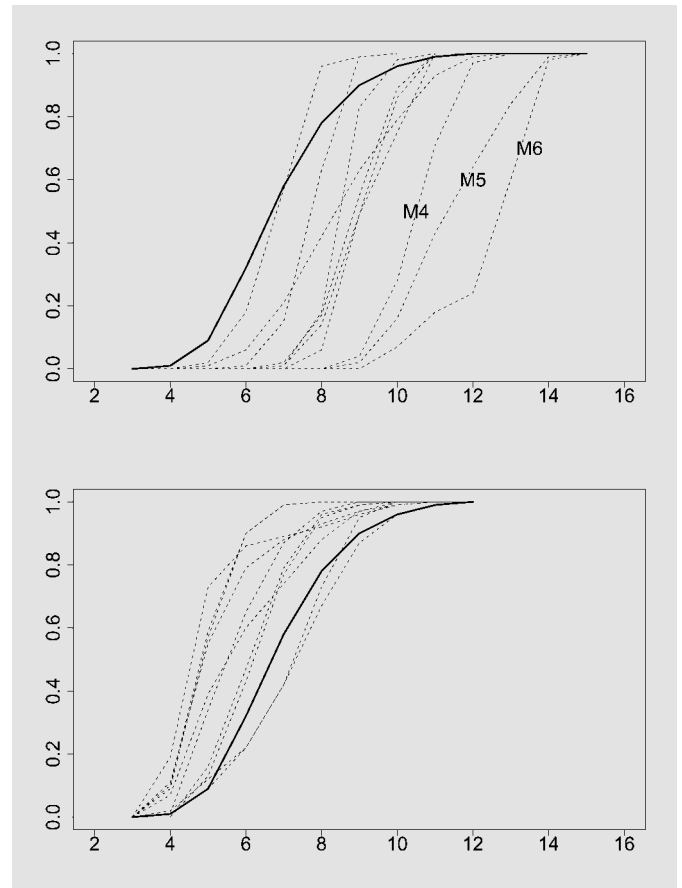


Fig. 3. Cumulative conditional distribution functions of W at each of 10 marker locations. Each distribution is conditional on the 10-locus marker data, which are simulated either in the presence (L: above) or absence (U: below) of linkage. The solid line is the pedigree-based null distribution. For convenience the integer-valued random variable W is convolved with a random variable uniform on (0,1). In the case of linkage (L) the three marker locations showing the highest values of the *ibd* measure are indicated.

Table 1. Markers and trait location used in the analysis

Label	Locus	Position (sex avg. cM)	Number of alleles
M1	D1S306	215	12
M2	D1S249	221	15
M3	D1S245	228	10
M4	D1S237	233	13
M5	D1S229	238	8
M6	D1S479	242	11
T	Trait	244	
M7	D1S446	252	13
M8	D1S235	255	9
M9	D1S180	267	11
M10	D1S102	276	6

For a larger dataset consisting of several extended pedigrees estimation of the cdf's will become computationally intensive, and only a small number of pseudo-datasets will be practical.

The top panel of figure 3 shows the cdf, F_0 , of W_0 under the no-data null and F_λ at each marker locus conditional on the linked (L) marker data \mathbf{Y}_M . Since W is integer-valued, for convenience its distribution is convolved with a random variable uniform on (0,1) to make the distributions absolutely continuous. Below the same is shown for a single set of unlinked (U) pseudo-data $\mathbf{Y}_M^{*(i)}$. In the case of linkage, several loci show a shift in the distribution of W , the most extreme being markers M6, M5 and M4. The particular unlinked dataset shown gives a slight shift to lower values of

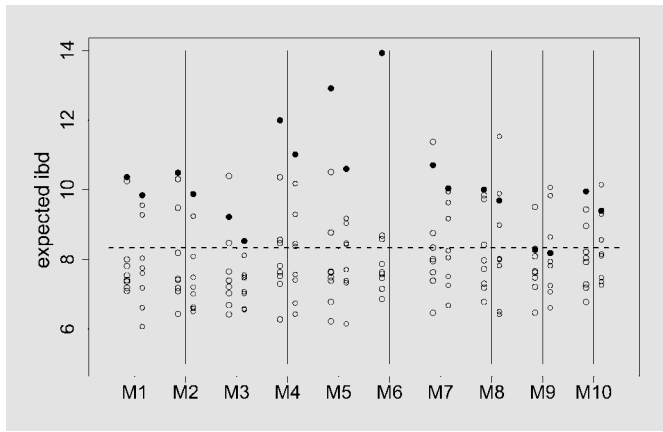


Fig. 4. Values of statistics $T = E(W|Y_M)$ (solid circle) and $T^* = E(W^*|Y_M^*)$ for each pseudo-dataset Y_M^* (open circles), plotted for each of 10 marker locations. The values to the left of each marker include data from all loci. Those to the right exclude the M6 (D1S479) data. The broken horizontal line shows the pedigree-based null expectation $E_0(W)$.

W at most marker locations, as compared to the pedigree null cdf F_0 . In expectation for an unlinked dataset there is no shift (equation (3)), but the reductions in variance seen both for linked (L) and unlinked (U) datasets are always present (equation (4)). This reduction in variance of W in the presence of marker data leads to increased power to detect linkage [13, 17], as compared to a test against the no-data null F_0 . It is also of interest to consider the extent to which the marker data at marker M6 (D1S479) influence the cdf's at other marker locations. This marker is closest to the trait (table 1) and produces the most extreme cdf in figure 3. The apparent signals at neighboring markers M4 and M5 may be due only to the strong signal provided by M6. All following analyses below were therefore duplicated, both including and excluding the data on M6.

First, we perform the 'standard' analysis (equation (2)). Figure 4 shows the estimates of $T(Y_M, \lambda)$ and $T^*(Y_M^*, \lambda)$ for each marker location λ and each dataset. Values are shown for analyses that include (to the left) and exclude (to the right) the data at marker D1S479 (M6). The values corresponding to the actual linked dataset are often larger than for any of the unlinked pseudo-datasets, particularly when the data at marker D1S479 (M6) are included. However, large numbers of unlinked pseudo-datasets U (i.e. simulations of Y_M^*) would be needed to obtain a reliable p value estimate for the observed values of T in the linked dataset L .

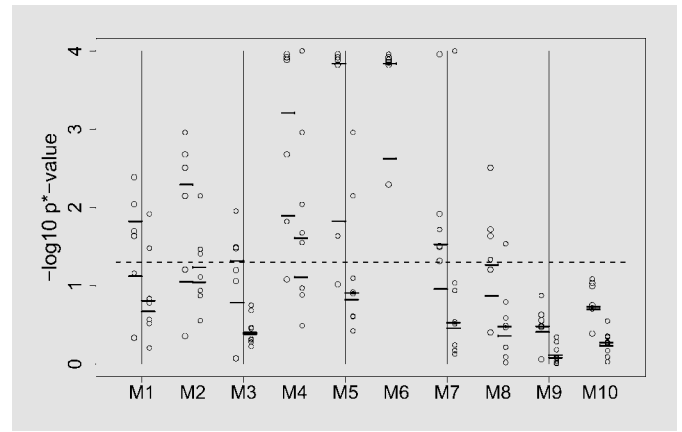


Fig. 5. Negative log (base 10) p^* values for linkage from eight unlinked pseudo-datasets, plotted for each of 10 marker locations. The values to the left of each marker include data from all loci. Those to the right exclude the M6 (D1S479) data. The datasets are the same as those used in figure 4. The broken horizontal line is at height 1.301 corresponding to $p^* = 0.05$. The bars indicate log10 of the mean and median p^* values at each marker over the pseudo-datasets. In all cases the median is larger than the mean.

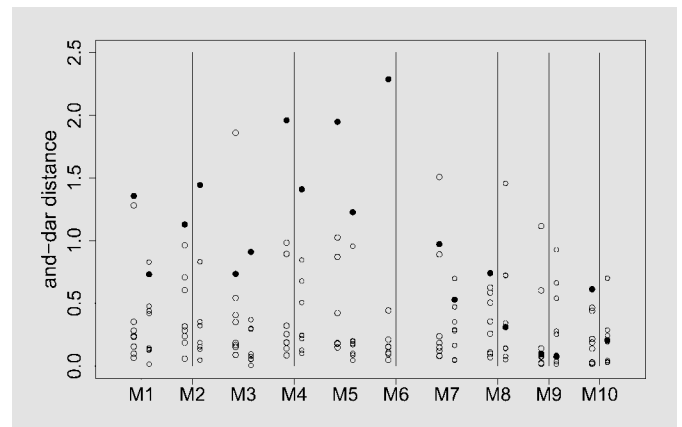


Fig. 6. Values of the Anderson-Darling distance measure between the cumulative distribution functions of W given data Y_M (solid circle) and given each pseudo-dataset Y_M^* (open circles), plotted for each of 10 marker locations. The values to the left of each marker include data from all loci. Those to the right exclude the M6 (D1S479) data. The datasets are the same as those used in figures 4 and 5.

Figure 5 shows the p^* values for the single linked dataset L against eight unlinked pseudo-datasets. The p^* values are plotted on the negative log base-10 scale. As for figure 4, analyses are done with and without the data on marker D1S479 (M6). The lower bar shows the average

p^* value, which is close to the estimated p_0 value for each marker (not shown). The upper bar denotes the median p^* value for each marker. Including the data on M6, markers M4, M5 and M6 show several p^* values as small as 10^{-4} . Without the data on marker M6, there is little evidence for linkage. Only the closest highly informative marker M4 achieves a p^* values less than 0.05 at least 50% of the time when the M6 data are excluded. Marker M2, which is more distant from the true trait location but also highly informative has a median p^* value of 0.06.

Figure 6 shows the Anderson-Darling [15] distances of each imputed cdf from the corresponding average G_λ (equation 6). Including the data on marker M6, there is a strong signal, with several markers picking up the difference between the cdf F_λ and the no-linkage $F_\lambda^{*(i)}$, $i = 1, \dots, 8$. Without the M6 data, the two most informative markers M2 and M4 still show a large deviation for F_λ , while the less informative but closest remaining marker M5 also gives F_λ as having the largest deviation from the average cdf G_λ .

Discussion

We have developed a testing procedure for robust linkage detection on extended pedigrees, using directly the imputed distribution of an *ibd* measure W given marker data \mathbf{Y}_M , rather than first computing statistics T which are expectations of W given \mathbf{Y}_M (equation (2)). Using $T_\lambda^* = E(W_\lambda^{*(i)} | \mathbf{Y}_M^*)$, each \mathbf{Y}_M^* gives rise to only an expectation T_λ^* at each location λ . An advantage of the approach using the p^* values based on the imputed cdf of W is that it provides, for each pseudo-dataset \mathbf{Y}_M^* and at each location λ , a direct estimate of the probability of a realized null (no-linkage) *ibd* value exceeding a value realized given the observed-data. However, although p^* values are thus analogous to p-values, they do not have a $U(0, 1)$ distribution under the null hypothesis, and should not be over-interpreted. When multiple locations λ are considered, issues of multiple testing in any case remain.

Our method of simulation of marker pseudo-data \mathbf{Y}_M^* has some similarities to that used in SimIBD [13]. For a single marker, these authors showed the increased power of using the true marker-based null distribution rather than the pedigree-based null distribution. However, the SimIBD statistic is (like those of other authors) an expected *ibd* measure based on realized or computed probabilities of *ibd* conditional on marker data. In contrast, the current approach uses the full cumulative distribution function of an *ibd* measure estimated conditional on

marker data. One way to view the difference between a method such as that of SimIBD and the current proposal is that other methods consider

$$T_\lambda - T_\lambda^{*(i)} = E(W_\lambda - W_\lambda^{*(i)} | \mathbf{Y}_M, \mathbf{Y}_M^{*(i)})$$

while the current approach is based on

$$p^{*(i)} = E(I(W_\lambda \leq W_\lambda^{*(i)} | \mathbf{Y}_M, \mathbf{Y}_M^{*(i)}))$$

where $I(\cdot)$ is the indicator function. The two statistics will emphasize different aspects of the difference in the distributions F_λ of W_λ and the $F_\lambda^{*(i)}$ of $W_\lambda^{*(i)}$. This leads to a consideration of additional alternate measures of the deviation of the cdf conditioned on actual data from those based on analogous unlinked pseudo-datasets \mathbf{Y}_M^* , such and the Anderson-Darling measure illustrated here. The use of the expectation may be insensitive to distributional differences in the tails: this is seen at some markers in figure 4 particularly in the absence of the M6 data. Other measures may be more effective at detecting such differences as seen in figures 5 and 6.

Power to detect linkage depends on (i) the true distortion of *ibd* among affected individuals relative to that expected in the absence of trait linkage, (ii) the ability of the scalar measure W to reflect this distortion, and (iii) the information in marker data \mathbf{Y}_M regarding the latent W at any given chromosomal location. The first depends on the trait model, while an optimal measure W will depend also on the form of distortion to be detected. Generally, easily computed measures that summarize joint sharing among pedigree members will be the most effective. The measure of this paper was chosen primarily for simplicity; other measures are considered by [14] and are the subject of current work. Finally, with regard to (iii), note that equation (4) may be rewritten as

$$E^*(\text{var}(W | \mathbf{Y}_M^*)) = \text{var}_0(W) - \text{var}^*(E(W | \mathbf{Y}_M^*)) = \text{var}_0(W) - \text{var}^*(T)$$

That is, the extent to which the pedigree-based $\text{var}_0(W)$ overestimates the data-based null variance $\text{var}^*(T)$ is the expected residual variance of W given marker data. This residual variance $\text{var}(W | \mathbf{Y}_M^*)$ can also be estimated from the MCMC run for given data \mathbf{Y}_M^* .

Although a location at which the data are highly informative as to *ibd* leads to small $E^*(\text{var}(W | \mathbf{Y}_M^*))$ and hence high power to detect linkage, the most extreme p^* values should not be taken as indicating the most likely trait locus location. Whereas the informativeness of marker data leads to reduction in the expected conditional variance of W , linkage will in general lead to a shift in the distribution of W . Both contribute to the detection of linkage, but only the latter is relevant to estimation of location. To estimate the location of the trait gene, an esti-

mation procedure for the degree of shift is required [17]. In essence, the location estimate should be the chromosomal position at which the latent *ibd* distribution is maximally distorted: this also is the subject of ongoing work.

Acknowledgement

Research supported in part by NIH grant GM-46255, and by a Guggenheim Foundation fellowship to E.A.T. while visiting North Carolina State University. E.A.T. is grateful for the hospitality of the Department of Statistics and the Bioinformatics Research Center, NCSU. We are grateful to two referees for helpful comments.

References

- 1 Thompson EA: Statistical Inferences from Genetic Data on Pedigrees, vol. 6 of NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics, Beachwood, OH, 2000.
- 2 Sobel E, Lange K: Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 1996;58:1323–1337.
- 3 Donnelly KP: The probability that related individuals share some section of genome identical by descent. *Theor Popul Biol* 1983;23:34–63.
- 4 Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 1987;84:2363–2367.
- 5 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996;58:1347–1363.
- 6 Baum LE, Petrie T, Soules G, Weiss N: A maximization technique occurring in the statistical analysis of probabilistic functions on Markov chains. *Ann Math Stat* 1970;41:164–171.
- 7 Gudbjartsson D, Jonasson K, Frigge M, Kong A: Allegro, a new computer program for multipoint linkage analysis. *Nat Genet* 2000;25:12–13.
- 8 Penrose LS: The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann Eugenics* 1935;6:133–138.
- 9 Suarez BK, Rice J, Reich T: The generalized sib pair *IBD* distribution: Its use in the detection of linkage. *Ann Hum Genet* 1978;42:87–94.
- 10 Whittemore A, Halpern J: A class of tests for linkage using affected pedigree members. *Biometrics* 1994;50:118–127.
- 11 McPeck MS: Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet Epidemiol* 1999;16:225–249.
- 12 Churchill GA, Doerge RW: Empirical threshold values for quantitative trait mapping. *Genetics* 1994;138:963–971.
- 13 Davis S, Schroeder M, Goldin LR, Weeks DE: Nonparametric simulation based statistics for detecting linkage in general pedigrees. *Am J Hum Genet* 1996;53:867–880.
- 14 Basu S, Wijsman EM, Thompson EA: Allele-sharing methods on large pedigrees. *Genet Epidemiol* 2002;23:267.
- 15 Anderson TW, Darling DA: A test of goodness of fit. *J Am Stat Ass* 1954;49:765–769.
- 16 Levy-Lahad E, Wijsman EM, Nemens E, Anderson L, Goddard KA, Weber JL, Bird TD, Schellenberg GD: Familial Alzheimer's disease locus on chromosome 1. *Science* 1995;269:970–973.
- 17 Thompson EA: Linkage detection for complex traits. In *Invited Proceedings of the 54th Session of the International Statistical Institute*. 2003, in press.