Uncertainty in inheritance: assessing evidence for linkage

E. A. Thompson

Department of Statistics, University of Washington
Box 354322, Seattle, WA 98195, USA

This paper is based on material presented at the Third University of Washington Biostatistics Symposium, November 2005, and submitted December 2005.

Technical Report no. 498
Department of Statistics
University of Washington

April 2006.

Uncertainty in inheritance: assessing evidence for linkage

E. A. Thompson

Department of Statistics, University of Washington
Box 354322, Seattle, WA 98195, USA
thompson@stat.washington.edu

1 Introduction

1.1 Linkage detection and estimation

Given trait and genetic marker data on sets of related individuals, the classical problem of genetic linkage analysis is to *map* the the location(s) in the genome of DNA affecting the trait relative to the known locations of the DNA markers. In this paper, we assume the relationships among the individuals are known. These relationships may be specified via the set of *pedigrees* of the individuals for whom data are observed. Not all members of the pedigrees are observed: some serve simply to specify the relationships among those who are observed (see Figure 1). We assume also that the DNA markers to be used are well characterized. That is, both their relative locations in the genome and the population frequencies of their alleles are assumed known. Finally, where necessary, we shall assume that the genetic basis of the trait of interest is also well characterized.

The classical problem of genetic linkage analysis is then two-fold. First is the issue of linkage detection: does any DNA on the chromosome(s) of the DNA markers affect the trait? This is a hypothesis testing question; the null hypothesis (H_0) is that it does not. Linkage detection does not require the specification of a trait model. However, if H_0 is rejected, then the relevant goal become estimation of the locations of DNA affecting the trait. Likelihood-based methods are typically used to address this question, and a probability model for the trait becomes necessary. Although the methods of genetic linkage analysis have been well established for over fifty years (Smith [1], Morton [2]), basic statistical issues remain, especially when the methods are applied to data on extended pedigrees where there are substantial missing data (unobserved individuals). Two primary issues are: What is the statistical significance of a given value of a test statistic used to test H_0? How should this be adjusted, to account for the fact that in considering multiple genetic markers we make multiple dependent tests?

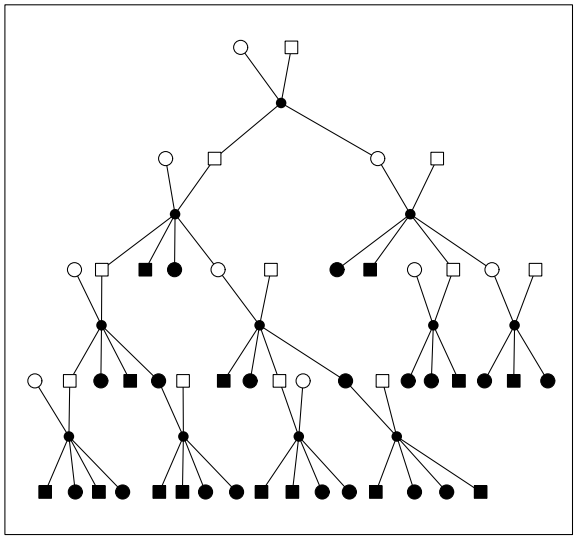

Fig. 1. The 52-member pedigree used in simulation. The 32 shaded individuals have marker and trait data available. Data were simulated on three copies of this pedigree

In this paper, we present a new approach to the assessment of statistical significance. This approach permits us to address the multiple testing issue, and also to express the level of uncertainty about this significance. This uncertainty derives from the fact that we can never actually observe the descent of DNA in a pedigree. This is the only source of uncertainty we consider here. Genetic marker and trait model misspecification are also important practical issues, but their inclusion in assessment of the strength of the evidence is beyond the scope of this paper.

Our approach will make use of the concept of a fuzzy p-values, as recently developed in a quite different context by Geyer & Meeden [3]. Thompson & Geyer [4] extended this idea to the latent variable context. In the detection and estimation of genetic linkage, the patterns of descent of DNA down a pedigree are latent variables that can never be precisely observed.

1.2 The example data set

We will discuss the approach in the context of data simulated on three copies of a 52-member, 5-generation pedigree (Figure 1). Data for several quantitative traits and a 12 genetic markers were simulated on these pedigrees. The markers were equally spaced at a genetic distance of 10 cM ($\sim 10^7$ bp), and each had 4 alleles, with frequencies 0.4, 0.3, 0.2, and 0.1. Trait and marker data are assumed available on 32 members of each pedigree.

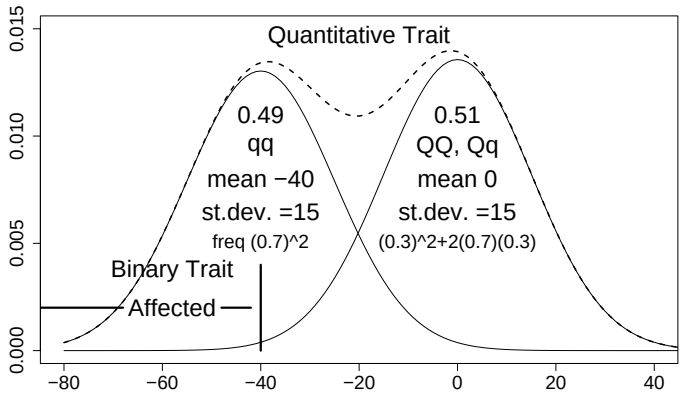

Fig. 2. The simulation model for trait-2, a quantitative trait simulated on the three 52-member pedigrees, each of the form shown in Figure 1

The particular quantitative trait we use in this paper is known as trait-2. It is controlled by a single genetic locus with two alleles, and this locus is mid-way between tenth marker (M10) and the eleventh (M11). The trait model is shown in Figure 2. The homozygous recessive genotype has mean -40, relative to the value 0 for the other two genotypes. The recessive allele has frequency 0.7, so the two phenotypic classes have almost equal frequency in the population. The individual variance about the genotypic mean is 225.0, so the two genotypic distributions have substantial overlap: samples of size 96 (3×32) from this mixture distribution rarely show significant bimodality.

Since the ideas of this paper are more easily introduced using binary trait, we define also such a trait by declaring all individuals with a quantitative trait-2 value less than -40 to be "affected". This gives 7, 9, and 6 affected individuals on the three pedigrees, including a sib pair and trio on the second copy and 3 affected sib pairs on the third.

1.3 Mendel's First Law and the inheritance of genome

Human individuals are diploid. Chromosomes in the cell nucleus are in homologous pairs, one deriving from the DNA of the individual's father (the paternal chromosome) and the other from the individual's mother (the maternal chromosome). At meiosis (the process of formation of gamete cells), the segregation of DNA from parent to offspring gamete any single genome location is specified by Mendel's First Law: each parent segregates a randomly chosen one of its two copies independently to each offspring.

In a pedigree, we will index the parent-offspring transmissions (meioses) by i, and the locations of interest by j ($i = 1, ..., m; \quad j = 1, ..., l$). Then the inheritance of DNA may be specified by the binary indicators $S_{i,j} = 0$ or 1 as in meiosis i at position j the maternal or paternal DNA (respectively) of the parent is transmitted to the offspring. Mendel's First Law may then be written as

$$\begin{aligned} \mathrm{P}(S_{i,j} = 0) = \mathrm{P}(S_{i,j} = 1) &= 1/2 \\ S_{i,\bullet} = \{S_{i,j}; j = 1, ..., l\} &\text{ are independent} \end{aligned}$$

At any location j, $j = 1, \ldots, l$, $S_{\bullet,j} = \{S_{i,j}; i = 1, \ldots, m\}$ is known as the *inheritance vector* at location j. This vector determines the founder origin of the DNA present in each individual at that location (Thompson [5]). Hence also determined is which individuals *share genome* at this location, in the sense of each receiving DNA copied from the same DNA in a founder of the pedigree.

Over locations j on the same chromosome, $S_{i,j}$ are dependent: this is the essence of genetic mapping. The gamete chromosomes formed ar meiosis consist of large ($\sim 10^8$ bp) alternating segments of the parent's paternal and maternal DNA. Hence for locations j and j' that are close together, there is probability close to 1 that $S_{i,j}$ and $S_{i,j'}$ are equal, and high probability that, if descendant individuals share genome at location j, they do also at location j'. The correlation in genome sharing decreases as the distance between the location j and j' increases. In essence, every test for genetic linkage for a trait is seeking evidence that genome sharing at a particular location in the genome is associated with sharing of trait characteristics.

2 Linkage detection for a binary trait

2.1 The case of observed inheritance patterns

Consider first the case of a binary trait and the approach to linkage detection that would be taken were the inheritance of DNA directly observable at the marker loci $j = 1, ..., l$. Let $\mathbf{S}$ denote the collection of $\{S_{\bullet,j}\}$, where $S_{\bullet,j}$ is the inheritance vector at marker locus j, or equivalently $\mathbf{S}$ is the matrix of $S_{i,j}$. Then, for marker j, we require a test statistic that reflects association in sharing of DNA at this location with sharing of trait phenotype. For example, we might consider all affected individuals and score pairwise their sharing at marker j, and sum this score over all pairs of related affected individuals. This would be the *Spairs* measure of genome sharing which is often used in the literature [6, 7]. In general, the test statistic would be some function of $S_{\bullet,j}$, say $t(S_{\bullet,j})$, and we would reject H_0 if the observed value t_{obs} of $t(S_{\bullet,j})$ is large.

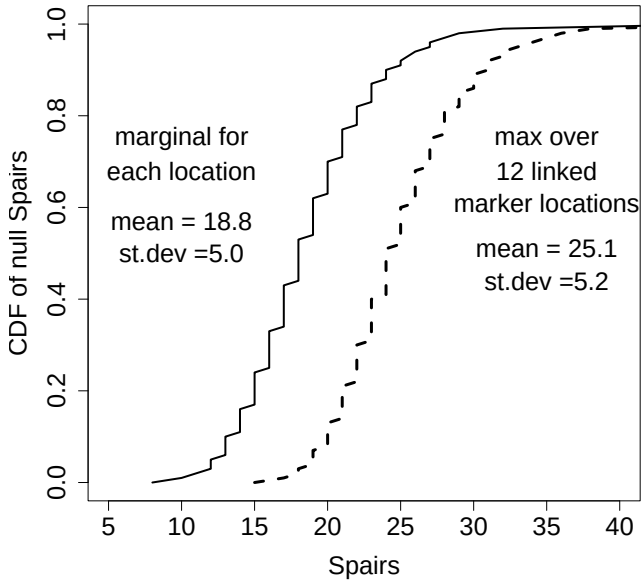

Fig. 3. CDFs of the empirical null distributions of the *Spairs* statistic for the example of this paper. The solid line gives the marginal distribution at any genome location, and the dashed line is for the maximum over the twelve marker locations

The p-value associated with the observation t_{obs} is straightforward

$$\text{p} = \text{P}(t(\mathbf{S}_0) \ \geq \ t_{\text{obs}}) \tag{1}$$

where $\mathbf{S}_0$ is generated under H_0. Although the probability distribution of $t(\mathbf{S}_0)$ is not known explicitly, Monte Carlo estimation of this distribution is trivial. The descent of genome in a pedigree may be simulated, either marginally at any genome location or jointly over dependent loci. For the example of this paper, and using the *Spairs* measure of genome sharing scored pairwise among affected individuals, the marginal cumulative distribution function is shown in Figure 3 (solid line). The p-value may be read from the curve. For example, if $t_{\text{obs}} = 27$ the p-value is about 0.05.

More importantly, the issue of correction for multiple dependent tests is also easily addressed. If the $S_{\bullet,j}$ are observed, then a single omnibus statistic may be formed. In the current context a natural choice is

$$t^*(\mathbf{S}) = \max_{j=1,\ldots,l} t(S_{\bullet,j}). \tag{2}$$

This choice will be powerful against alternatives in which the DNA affecting the trait is close to one j. Another alternative, with similar justification, might

be to sum adjacent pairs $t(S_{\bullet,j-1}) + t(S_{\bullet,j})$ and to take the maximum of these pairwise sums, but here we have used the simpler choice of equation (2). In any event, the distribution of the omnibus statistic under H_0 may be found by simulating the descent of genome in the pedigrees. For the example of this paper, and using the *Spairs* measure of genome sharing scored pairwise among affected individuals, the cumulative distribution function for the maximum value over the 12 linked marker locations is shown in Figure 3 (dashed line). The p-value may be read from the curve. For example, if the observed value of t^* is 35 the p-value for the omnibus test is about 0.05. Of course, the distribution of $t^*(\mathbf{S})$ is stochastically larger than of each marginal $t(S_{\bullet,j})$. The extent of the difference depends both on the number of loci and their relative locations.

2.2 Sampling unobserved inheritance patterns

In reality, the latent variables $\mathbf{S}$ are not observed, and $t(S_{\bullet,j}))$ is a *latent test statistic*. We observe only marker genotypes $\mathbf{Y}$ of some individuals, and our objective is thus to impute $\mathbf{S}$ from marker data $\mathbf{Y} = \{Y_{\bullet,j}; j = 1, ..., l\}$.

On small pedigrees there are standard methods, based on the Baum algorithm [8] for computation of $\mathrm{P}(S_{\bullet,j} \mid \mathbf{Y})$ [7, 9]. However, since this is only a marginal distribution at location j, we do not then have $\max_j t(S_{\bullet,j})$ or other statistic based jointly on the $S_{\bullet,j}$. Better is to use analogous methods to obtain i.i.d. realizations of the complete matrix $\mathbf{S} = \{S_{i,j}\}$ from $\mathrm{P}(\mathbf{S} \mid \mathbf{Y})$. On large pedigrees, exact computations are infeasible, but several Markov chain Monte Carlo (MCMC) methods [10, 5] have been developed to successively sample dependent realizations of $\mathbf{S}$ from $\mathrm{P}(\mathbf{S} \mid \mathbf{Y})$. Hence, in either case, we have Monte Carlo estimate of the distribution of $t(\mathbf{S})$ given marker data $\mathbf{Y}$.

2.3 The standard approach to linkage detection

In the standard approach to this problem, a test statistic is based on the marker data $\mathbf{Y}$, and it has been proposed [6, 7] that a sensible form of this test statistic at each marker location j is the expectation of the genome sharing measure (or latent test statistic) given the observed marker data:

$$W_j(\mathbf{Y}) = \mathrm{E}(t(S_{\bullet,j}) \mid \mathbf{Y}). \tag{3}$$

Given realizations of $\mathbf{S}$ given $\mathbf{Y}$, Monte Carlo estimates of statistics such as those of equation (3) may be obtained by simply averaging the realized values of $t(S_{\bullet,j})$. If this approach is applied to the example of the binary trait in this paper, the resulting values of $W_j,\ j = 1, ..., 12$ are as shown in Figure 4.

Unfortunately, on extended pedigrees with missing data, almost nothing is known about the distribution of $W_j(\mathbf{Y})$ under H_0. Only its expectation is known since

$$\mathrm{E}(W_j) = \mathrm{E}(\mathrm{E}(t(S_{\bullet,j}) \mid \mathbf{Y})) \quad = \quad \mathrm{E}(t(S_{\bullet,j})).$$

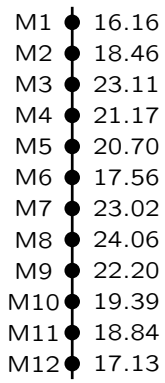

Fig. 4. Values of the expected Spairs measure at each marker location given the joint data $\mathbf{Y}$ at all markers

With regard to $W^* \equiv \max_j W_j$ not even the expectation under H_0 is known. Thus we have no idea whether the values observed in Figure 4 are significantly higher than the null mean 18.8, nor how the maximum value 24.06 at marker M8 should be interpreted.

Only some form of simulation will provide the p-value for a test based on the values of W_j or W^*. The simplest is a parametric bootstrap. That is, new marker datasets $\mathbf{Y}^{(k)},\ k = 1, ..., N$ are resimulated under the null hypothesis of no trait linkage, and each is analyzed to determine $W(\mathbf{Y}^{(k)})$. An empirical p-value is then given by

$$\mathrm{p} = (N+1)^{-1}(1 + \sum_{k=1}^{N} I(W(\mathbf{Y}^{(k)}) \ \geq \ W(\mathbf{Y}))). \tag{4}$$

This procedure is very computationally intensive. Simulation of new marker datasets under a specified marker model is easy, but MCMC analysis is required for each of the N resimulated datasets.

The resimulation approach also lacks robustness to marker-model misspecification. When marker data on the early generations of a pedigree are not available, imputed genome sharing can be quite sensitive to misspecified allele frequencies, and this is typically one of the greatest uncertainties in the marker model. If the resimulated datasets use incorrect alleles frequencies, either type-1 or type-2 error may be impacted. Since the linkage hypothesis is of an association between the inheritance of trait and marker phenotypes, an alternative simulation approach is to condition on the marker data, and resimulate trait data on the pedigrees. However, this raises even greater model misspecification issues. The testing approach developed to this point requires no specification of a marker model. Even where a researcher is able to specify a trait model, pedigrees are normally ascertained through trait phenotypes,

so in resimulation this would have to be included in the simulation process. In reality, precise ascertainment distributions may be even harder to specify than the trait model itself.

2.4 Introducing the fuzzy p-value

The procedures of the previous section are valid, but appear quite wasteful of the available information. MCMC provides an estimate of the full distribution of $t(S_{\bullet,j})$ or of any omnibus statistic $t^*(\mathbf{S})$ given $\mathbf{Y}$. To simply average over the chain, obtaining a single set of numbers $W_j(\mathbf{Y})$, $j = 1, ..., l$ and $W^*(\mathbf{Y}) = \mathrm{E}(t^*(\mathbf{S}) \mid \mathbf{Y})$ seems suboptimal. This is especially so, since we know (almost) nothing about the distributions of W_j and W^*, but (up to Monte Carlo error) everything about the distributions of $t(S_{\bullet,j})$ and $t^*(\mathbf{S})$ given $\mathbf{Y}$. By averaging over realizations of $\mathbf{S}$, information that $\mathbf{Y}$ provides about $t^*(\mathbf{S})$ is confounded with the evidence $t^*(\mathbf{S})$ provides about H_0.

Geyer & Meeden [3] have introduced, in a quite different context, the idea of a *fuzzy p-value*. This is a random variable with the distribution of $(Q|\mathbf{Y})$, where Q is U(0,1) (unconditionally) under the null hypothesis H_0. Then, over data sets $\mathbf{Y}$ under H_0, $\mathrm{E}(\mathrm{P}(Q \leq \alpha|\mathbf{Y})) = \alpha$ where E(). That is, under H_0, the fuzzy p-value has a U(0,1) distribution, which is the defining canonical characteristic of any p-value.

Now in the current context consider the complete-data p-value of equation (1). This may be equivalently written as

$$\pi(\mathbf{S}) = \mathrm{P}(t(\mathbf{S}_0) > t(\mathbf{S})|\mathbf{S}),$$

where $\mathbf{S}_0$ is generated under the null hypothesis H_0. By definition, this has a distribution uniform on (0,1) under H_0, and hence $\mathrm{E}(\mathrm{P}(\pi(\mathbf{S}) \leq \alpha)|\ \mathbf{Y}) = \alpha$ under H_0. That is, a random variable with the distribution of $\pi(\mathbf{S})$ given $\mathbf{Y}$ is a fuzzy p-value.

Further, a Monte Carlo estimate of the fuzzy p-value distribution is immediately available to us. Note that

$$\pi(\mathbf{S}) = \mathrm{P}(t(\mathbf{S}_0) \geq t(\mathbf{S})|\mathbf{S}) \quad = \quad \mathrm{P}(t(\mathbf{S}_0) \geq t(\mathbf{S})|\mathbf{S}, \mathbf{Y}).$$

Consider a set of realizations under H_0 (unconditional on $\mathbf{Y}$), $\mathbf{S}_0^{(h)}$, $h = 1, ..., H$, and a single set of MCMC realizations $\mathbf{S}^{(k)}$, $k = 1, ..., K$, from the conditional distribution given $\mathbf{Y}$. Then, for each k,

$$\eta(\mathbf{S}^{(k)},\ \mathbf{Y}) = \mathrm{P}(t(\mathbf{S}_0) \geq t(\mathbf{S}^{(k)})|\mathbf{S}^{(k)},\ \mathbf{Y}), \qquad k = 1, ..., K$$

is estimated by

$$H^{-1} \sum_{h=1}^{H} I(t(\mathbf{S}_0^{(h)}) \geq t(\mathbf{S}^{(k)}))$$

providing K realizations from the fuzzy p-value distribution.

Of course, all our statistics are discrete, as indeed is inheritance itself. In [4], we have shown how this discreteness may be dealt with exactly in the fuzzy p-value computation, providing a distribution over data sets $\mathbf{Y}$ which is exactly U(0,1) up to Monte Carlo error.

2.5 Application to the binary trait data

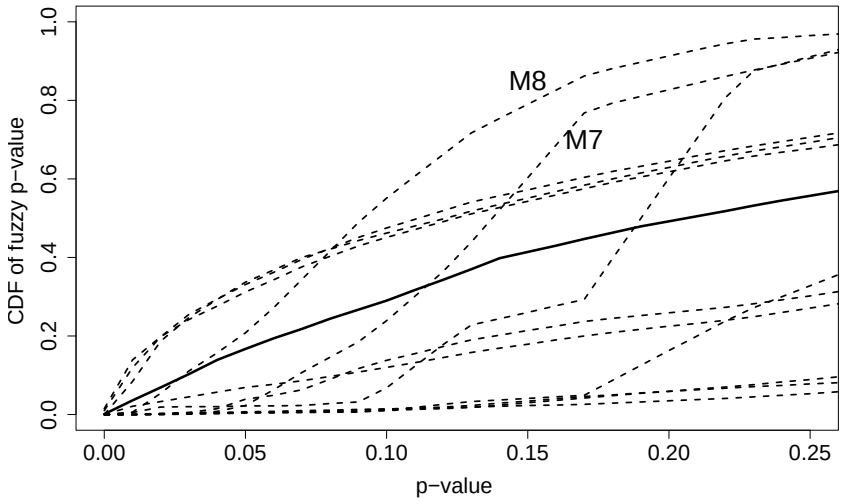

Fig. 5. CDFs of fuzzy p-values for the binary trait example of this paper. One CDF is estimated for each marker location (dashed lines) and for the multiple-testing corrected version (solid line)

The advantages of the fuzzy p-value are apparent. First, it can be easily estimated from two Monte Carlo samples (one unconditional, and one conditional on $\mathbf{Y}$). This does not require resimulation of data, which is both a computational and a statistical (robustness) advantage. Second, it provides a valid p-value, including any correction desired for testing at multiple linked markers. Third, and most importantly, it separates the uncertainty about $t(\mathbf{S})$ from the evidence in $t(\mathbf{S})$.

In Figure 5 the approach of section 2.4 is applied to the binary trait data described at the end of section 1.2. One fuzzy p-value distribution is estimated for the *Spairs* latent statistic at each marker location, and one for the omnibus test using the latent test statistic that is the maximum over the marker locations (equation (2)). In this example, the evidence for linkage is very weak.

Only marker M8 gives a probability greater than 0.5 that the fuzzy p-value is less than 0.1, while the omnibus test has probabiity about 0.25 that the fuzzy p-value is less than 0.1. However, the evidence is not only weak but also uncertain. Consider again the test based on the inheritance at marker M8. The mass of the fuzzy p-value distribution is spread fairly uniformly over the range 0.05 to 0.2.

3 Fuzzy p-values for linkage lod scores

3.1 MCMC estimation of the linkage lod score

It is not surprising that our dichotomized quantitative trait provides little evidence for linkage. To use quantitative trait data, or to estimate trait locus positions, a model for the trait data Z is required. In this case, a linkage *lod score* is often used. The lod score at a position γ on the marker map is simply a (base-10) log-likelihood-ratio of the hypothesis that the trait locus is at position γ relative to the alternative that the trait locus is unlinked to the markers (H_0). The probabilities of marker data $\mathbf{Y}$ do not depend on γ, and under H_0 the trait data Z and marker data $\mathbf{Y}$ are independent. Thus the lod score may be written

$$\mathrm{lod}(\gamma) = \log_{10}(\mathrm{P}_\gamma(Z \mid \mathbf{Y}) \;/\; \mathrm{P}(Z))$$

In this form, it may appear that the lod score may be affected by the ascertainment of pedigrees through trait phenotypes. However, this is not the case, provided ascertainment is only through trait (or only marker) phenotypes. Any correction for ascertainment appears in both numerator and denominator: the association between the inheritance of trait and marker phenotypes is not affected by ascertainment.

On an extended pedigree, with multiple genetic markers and many missing observations, exact computation of $\mathrm{P}_\gamma(Z \mid \mathbf{Y})$ is infeasible. However, following [11] we may write

$$\begin{aligned}\mathrm{P}_\gamma(Z \mid \mathbf{Y}) &= \sum_{\mathbf{S}} \mathrm{P}_\gamma(Z \mid \mathbf{S})\mathrm{P}(\mathbf{S} \mid \mathbf{Y}) \\ &= \mathrm{E}(\mathrm{P}_\gamma(Z \mid \mathbf{S}) \mid \mathbf{Y})\end{aligned}$$

where as before $\mathbf{S}$ is the complete set of binary inheritance vectors at the marker locations, and the expectation is over the distribution of $\mathbf{S}$ given $\mathbf{Y}$, which does not involve γ. Equation (5) provides a route to a Monte Carlo estimate of the complete lod score curve as a function of γ given a single Monte Carlo sample from $\mathrm{P}(\mathbf{S} \mid \mathbf{Y})$. We have already seen that such a sample may be obtained. For each realized $\mathbf{S}$, $\mathrm{P}_\gamma(Z \mid \mathbf{S})$ must be computed for each γ of interest. This is feasible using a modification of standard pedigree peeling algorithms [12]. Under the assumption of no genetic interference, vectors $S_{\bullet,j}$

are Markov over j. Then, under the model indexed by γ, $(Z \mid \mathbf{S})$ depends only on the (at most 2) inheritance vectors at the marker positions flanking the position γ.

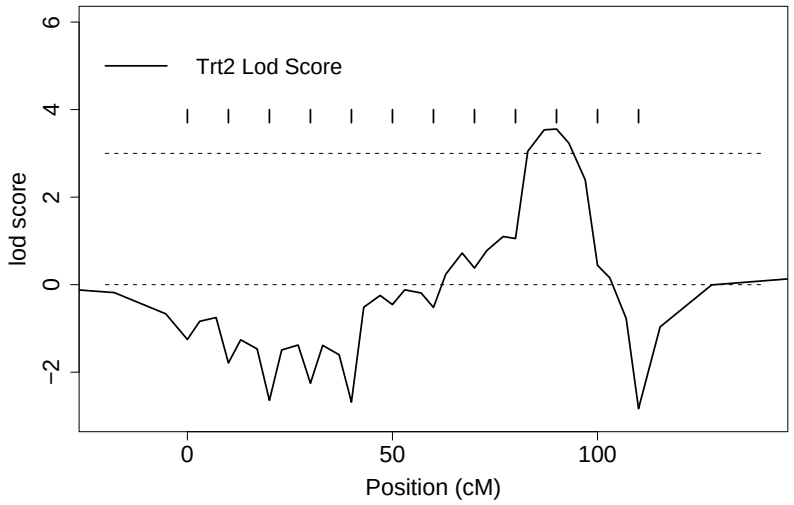

Fig. 6. The MCMC-estimated lod-score curve for quantitative trait-2. The twelve tick-marks show the positions of the markers

An MCMC-based estimate of the lod score curve is shown in Figure 6. The estimate is based on 30,000 MCMC realizations, and Monte-Carlo error is minimal. Lod scores are estimated at the marker positions, at 2 points in each marker interval, and at 5 points at each end of the marker map. The entire run of 30,000 MCMC scans and related computations took about 17 minutes on a 2.0 GHz linux laptop. We see that in the neighborhood of the marker M10 there is some evidence for linkage: the lod score exceeds the traditional target of 3, required to declare that a gene has been localized!

3.2 Fuzzy p-values for the linkage lod score

When originally introduced [1, 2], the lod score target of 3 had better justification than it does now. Trait models were straightforward, there were no genetic maps, and testing for linkage was done among a set of mostly unlinked marker and trait loci. Now, with more complex trait models and the multiple dependent tests of a genome scan many more questions arise.

The approach of section 2.4 provides an immediate solution. We take exacly the same approach as before, now using as our latent test statistic the lod

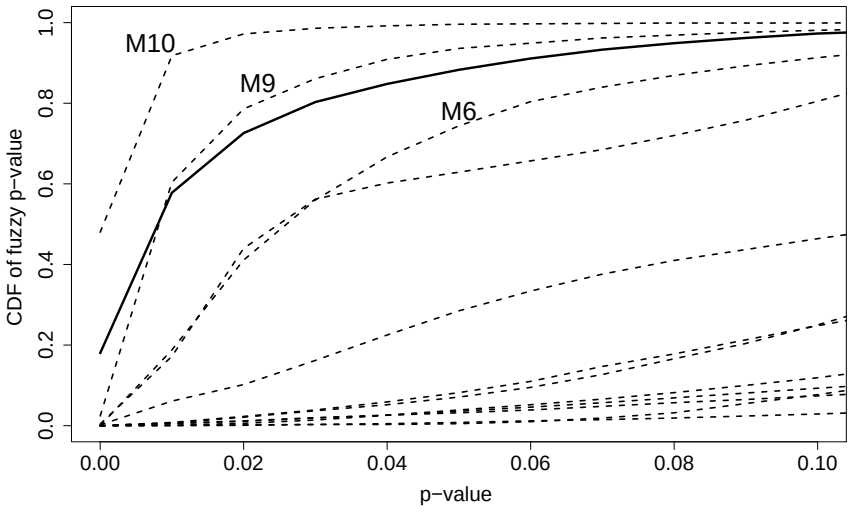

Fig. 7. The CDFs of fuzzy p-values estimated for each marker location (dotted lines), and for the maximum over marker locations (solid line)

score we would have were **S** observed:

$$t_\gamma(\mathbf{S}) = \log_{10}\left(\mathrm{P}_\gamma(Z \mid \mathbf{S})/\mathrm{P}(Z)\right) \tag{5}$$

for each location γ. Hence we can compute the fuzzy p-value both pointwise and adjusted for multiple testing, exactly as before. Indeed, no additional computation is required, since we already have the (MCMC) realizations from $\mathrm{P}(\mathbf{S} \mid \mathbf{Y})$ and we already compute $\mathrm{P}_\gamma(Z \mid \mathbf{S})$ in computing the MCMC estimate of the lod score.

Figure 7 shows the result of this computation and may be compared with Figure 5 for the dichotomized binary trait. Again, each dotted line corresponds to taking γ to be a marker location, and the solid line is the CDF of the fuzzy p-value for the maximum of the values of equation (5) over marker locations. We see that the linkage signal is much stronger than before. At marker M10 the probability that the fuzzy p-value is less than 0.05 is close to 1, while for the multiple-testing corrected version this probability is about 0.8. Uncertainty due to uncertainty in **S** is much reduced, but is still substantial for the overall test.

3.3 Fuzzy confidence intervals for a trait locus location

Statistical geneticists would not normally use the lod score (only) at the marker locations. Indeed, the decreases in lod score typically seen at markers

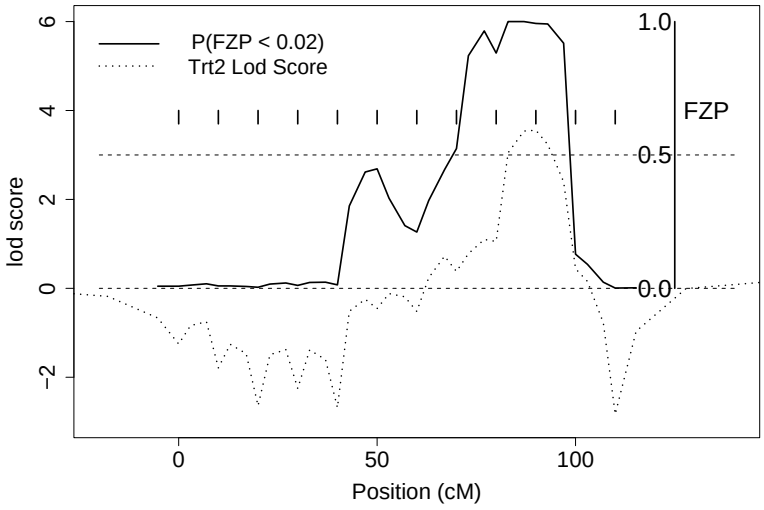

Fig. 8. Lod score curve (dotted curve) with added 2% quantile of the pointwise fuzzy p-value distributions (solid curve). The tick marks show the marker locations.

(Figure 6) arise because it is unlikely that the DNA affecting the trait segregates exactly with the marker DNA. An alternative representation of the fuzzy p-value based on equation (5) considered as a function of γ is given in Figure 8. The pointwise probability that the fuzzy p-value of the lod-score statistic for that location γ is less that 0.02 is plotted above (solid line), with the lod score curve of Figure 6 repeated (as a dotted line) for comparison. Again, we see strong evidence of a linkage signal in the neighborhood of marker M10, and again we see the apparent weaker signal at marker M6 (see Figure 7).

It is tempting to interpret the 2% quantile of the fuzzy p-value distributions shown in Figure 8 as the (fuzzy) level of a 98% confidence interval for the location γ of the trait. However, it is not. A non-fuzzy level-$(1-\alpha)$ confidence interval consists of those parameter values γ_0 that fail to be rejected in a test of size α of $H_0 :\ \gamma = \gamma_0$ against alternatives $\gamma \neq \gamma_0$. The fuzzy version assigns a probability of inclusion to each such γ_0. What we have constructed in Figure 8 is the probability of inclusion of each γ in a set formed by values γ_0 for which the null hypothesis of no linkage is rejected in a size $\alpha = 0.02$ test again the alternative $\gamma = \gamma_0$.

How then should such a confidence interval be constructed? First, for each γ we need a test of H_γ that the trait locus is at location γ against the alternative of all other locations. A reasonable such test, given $\mathbf{S}$ the set of inheritance vectors at the marker loci, is to reject H_γ if

$$t_\gamma(\mathbf{S}) = -\log(\mathrm{P}_\gamma(Z|\mathbf{S})/\sup_{\gamma^*}\mathrm{P}_{\gamma^*}(Z|\mathbf{S}))$$

is too large. That is, the likelihood of the simple hypothesis H_γ is too small.

Note, under H_γ, Z and $\mathbf{S}$ are not independent. Note also that linkage involves an association in the inheritance patterns of Z and of $\mathbf{Y}$ so that any test for linkage may be conditioned on either. Thus an analogue of our previous test of H_0 is to first realize $\mathbf{S}$ conditional only on Z, and then also conditional on both the marker data $\mathbf{Y}$ and Z, under H_γ. The first is straightforward: we first realize the latent inheritance vector at the trait locus given Z, and then inheritance at other genome locations conditional on this latent trait inheritance. The sampling conditional on Z and $\mathbf{Y}$ can be done as follows: first use MCMC as before to sample marker locus inheritance patterns $\mathbf{S}$ conditional on $\mathbf{Y}$, and then, for each γ, use importance sampling reweighting to incorporate the conditioning also on Z.

The performance of this procedure, and properties of the resulting confidence sets remain to be investigated. Although the above seems to be the most direct analogue of our previous testing procedure, in the sense that it reduces to that procedure when the null hypothesis is absence of linkage ("$\gamma = \infty$"), it is not the only possible way to construct a test of H_γ and hence a (fuzzy) confidence set.

4 Discussion

Data at additional markers loci, or on additional members of the pedigree structure, can reduce uncertainty in $\mathbf{S}$. The fuzzy-p-value distribution can guide this collection of additional data on the pedigrees. Figure 9 shows four hypothetical omnibus fuzzy p-value distributions from a study. In the case of A, the entire probability mass of the distribution is below 0.05: no additional data are needed. In case B, there is considerable uncertainty, but the probability that the fuzzy p-value is less than 0.05 is small. It is improbable that additional data will provide the desired evidence of linkage. In case C, the probability mass of the fuzzy p-value distribution is around 0.07, and there is little uncertainty. An optimistic researcher might wish to follow up this signal, but additional pedigrees would be required: the available evidence has been extracted from the current pedigrees. In case D, there is both sufficient uncertainty and a sufficient signal to make collection of additional data on these same pedigrees quite desirable. One clear advantage of the fuzzy p-value is that it puts uncertainty directly onto the p-value (evidence) scale.

Other advantages of the fuzzy p-value include robustness to the distribution of marker data $\mathbf{Y}$ and ease of computation. Both these advantages arise from the fact that there is no resimulation of marker data. In this, the approach shares similar characteristics to the permutation test of Churchill & Doerge [13]. However, this permutation test is readily applied only in experimental populations, since it requires large numbers of individuals exchangeable with

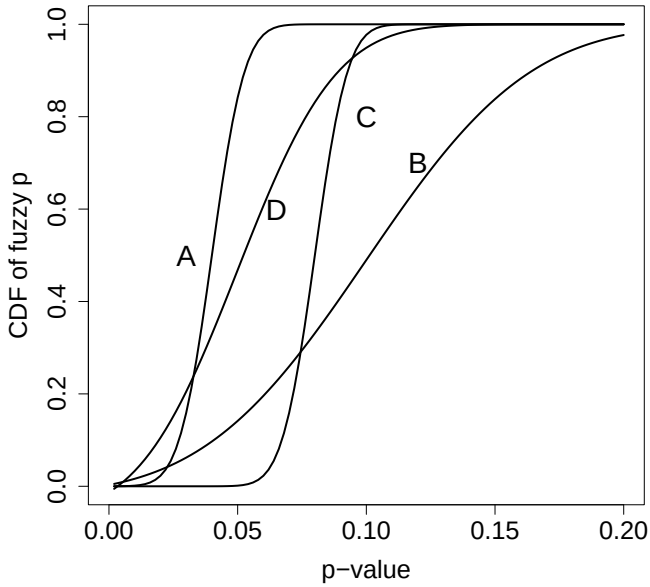

Fig. 9. Example of potential fuzzy p-value summaries of the evidence

respect to their inheritance of trait characteristics. It is often not possible to construct a valid permutation test for data on an extended pedigree.

However, the key advantage of a fuzzy p-value is that it separates the evidence for linkage from the uncertainty about **S**. It is latent inheritance patterns **S** that provides evidence for genetic hypotheses such as linkage, but marker data **Y** are a very imperfect reflection of **S**. Basing p-values on statistics constructed from data **Y** is very computationally intensive, requires detailed marker model, and raises unsolved multiple testing issues. Fuzzy p-values address these issues, putting uncertainty in **S** directly on evidence scale.

Moreover, the developments in this paper show how fuzzy p-values can be applied to lod scores in any trait-model based analysis of a quantitative or qualitative trait, as well as to linkage detection tests. In principle, at least, estimation can also be addressed through the construction of fuzzy confidence sets.

Acknowledgement

This research was supported in part by NIH grant GM-46255. I am grateful to Dr. C. J. Geyer for many helpful discussions.

References

1. Smith CAB. Detection of linkage in human genetics. Journal of the Royal Statistical Society (Series B) 1953;15:153–192.
2. Morton NE. Sequential tests for the detection of linkage. Am. J. Hum. Gen. 1955;7:277–318.
3. Geyer CJ, Meeden GD. Fuzzy and randomized confidence intervals and p-values. Statistical Science 2005;??:in press.
4. Thompson EA, Geyer CJ. Fuzzy p-values in latent variable problems. Biometrika 2006;p. submitted.
5. 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.
6. Whittemore A, Halpern J. A class of tests for linkage using affected pedigree members. Biometrics 1994;50:118–127.
7. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: A unified multipoint approach. Am. J. Hum. Gen. 1996; 58:1347–1363.
8. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions on Markov chains. Annals of Mathematical Statistics 1970;41:164–171.
9. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics 2002;30:97–101.
10. Sobel E, Lange K. Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Gen. 1996; 58:1323–1337.
11. Lange K, Sobel E. A random walk method for computing genetic location scores. Am. J. Hum. Gen. 1991;49:1320–1334.
12. Elston RC, Stewart J. A general model for the analysis of pedigree data. Human Heredity 1971;21:523–542.
13. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. Genetics 1994;138:963–971.