Chapter 19

SEMIPARAMETRIC AND NONPARAMETRIC GENE MAPPING

Fei Zou, Brian S. Yandell and Jason P. Fine

Department of Biostatistics
The University of North Carolina, Chapel Hill, NC, U.S.A

Departments of Statistics and Horticulture
University of Wisconsin, Madison, WI, U.S.A

Departments of Statistics and Biostatistics & Medical Informatics
University of Wisconsin, Madison, WI, U.S.A

E-mails: fzou@bios.unc.edu, byandell@wisc.edu, & fine@stat.wisc.edu

We review gene mapping, or inference for quantitative trait loci, in the context of recent research in semi-parametric and non-parametric inference for mixture models. Gene mapping studies the relationship between a phenotypic trait and inherited genotype. Semi-parametric gene mapping using the exponential tilt covers most standard exponential families and improves estimation of genetic effects. Non-parametric gene mapping, including a generalized Hodges-Lehmann shift estimator and Kaplan-Meier survival curve, provide a general framework for model selection for the influence of genotype on phenotype. Examples and summaries of reported simulations show the power of these methods when data are far from normal.

Keywords: Statistical genetics; Empirical process; Exponential tilt; Mixture model.

1. Introduction

Gene mapping concerns the statistical relationship between a phenotype, or measured response known as a trait, and the genotype, or heritable information measured at genetic markers scattered across the genome. Genetic information is incomplete, requiring consideration of mixture models across unknown genotypes. While gene mapping was initially developed

for normally distribution of traits, the framework extends readily to both semi-parametric and non-parametric models.

Commonly, individuals in a gene mapping study are sampled from an experimental cross such as a backcross or intercross. First, two inbred lines (A and B, say) are crossed to create the F1, which is heterogeneous everywhere. That is, at any selected genetic marker, the inbred parents are AA and BB, respectively, while the F1 is always AB. An F1 back-crossed to an inbred line, say A, produces backcross offspring that are either homozygous (AA) or heterozygous (AB) at every marker, with equal likelihood. The intercross, or F2, results from brother-sister mating of F1 children, yielding marker genotypes AA:AB:BB in an idealized 1:2:1 ratio. The backcross or intercross individuals are genetic mosaics of their inbred grandparents, due to meioses in the F1 parent(s). Other inbred experimental crosses are possible but are not considered further here (see Kao and Zeng 1997).

Each individual in a sample from an experimental cross is genetically unique. The different genetic patterns scored at markers spread across the genome allow us to associate the phenotype with genomic regions, or quantitative trait loci (QTL), where differences in genotype are inferred to affect the phenotype. QTL have great importance in revealing the genetic basis of phenotypic differences (Belknap *et al.*, 1997; Haston et al., 2002; Wang *et al.*, 2003). In plant and laboratory animals, backcross or F2 individuals are widely used for mapping quantitative traits (see Lynch and Walsh 1998).

The basic model selection problems for QTL mapping are: (i) detecting the presence of one or more QTLs, (ii) estimating QTL map position(s), and (iii) estimating the genetic effects of the QTLs. This model selection process is often referred to as inferring the genetic architecture (Mackay 2001). Complications arise due to lack of genotype data between genetic markers, leading to a likelihood based on a mixture of distributions across the possible QTL genotypes. Initially, Weller (1986), and later Lander and Botstein (1989), assumed the phenotype distribution given the genotype is normal. A general framework was sketched by Jansen (1992) and others.

The basic problem involves relating observed genetic marker information, m, to observed phenotypic trait measurements, y through two coupled models,

$$\mathrm{pr}(y|m,\lambda) = \sum_q \mathrm{pr}(y|q)\mathrm{pr}(q|m,\lambda),$$

with the sum over all possible genotypes, q, at the putative QTL(s), λ. In this paper, we allow the phenotype model, $\mathrm{pr}(y|q)$, to be semi-parametric (exponential tilt, including many generalized linear models) or fully non-parametric. The recombination model, $\mathrm{pr}(q|m,\lambda)$, can be directly calculated using the binomial based on markers, m, that flank the QTL, λ, and plays the role here of mixture weight (Kao and Zeng 1997).

The first gene mapping study involved single marker t-tests (Sax 1923), which was essentially the standard until the introduction of interval mapping (Lander and Botstein 1989; Haley and Knott 1992; Kruglyak and Lander 1995; see Doerge et al. 1997). The normal mixture, with a normal phenotype distribution, is the default in the widely used software Mapmaker/QTL (Lander et al. 1987), QTL/Cartographer (Basten et al. 1995) and R/qtl (Broman et al. 2003).

Nettleton (Nettleton and Praestgaard 1998; Nettleton 1999; Nettleton 2002) considered hypothesis testing for QTL against ordered alternatives, assuming an underlying normal model. Several investigators studied other, non-normal, parametric phenotype models, including binomial and threshold models (Visscher et al. 1996; Xu and Atchley 1996; McIntyre et al. 2000; Rao and Li 2000; Yi and Xu 2000; Broman 2003), Poisson (Mackay and Fry 1996; Shepel *et al.* 1998), negative binomial (Lan et al. 2001). Hackett and Weller (1995) considered ordinal threshold models. Broman (2003) proposed a two-part parametric model for phenotype with a spike at one value, including structural zeroes and type I censoring. Parametric Cox proportional hazard model with a specified baseline function was examined by Diao et al. (2004).

Inference on the QTL map position(s) is fairly robust to normality. However, model misspecification may lead to reduced power to detect genes affecting a trait or to biased estimates of the genetic architecture (Hackett 1997; Wright and Kong 1997). Further, genetic differences may involve more than a mean shift, as modeled for normal data. Perhaps the phenotype has a different shaped distribution for individuals with different genotypes, as opposed to a difference in the means or center of location? While these issues have been widely studied with single QTL models, there has been little work on more complex multigene models. One might expect that naively using normal models with highly non-normal data might cause greater difficulty in this set-up, where inferences about subtle gene-gene interactions may be misleading. Therefore it is useful to consider semi-parametric and non-parametric generalizations for QTL, providing more robust inference about the genetic architecture, including insight about possible parametric models for the phenotype given the genotype.

Semi-parametric QTL were first considered by Zou and coauthors (Zou et al. 2000; Zou and Fine 2002; Jin et al. 2003) using the exponential tilt. Lange and Whittaker (2001) investigated QTL using generalized estimating equations; however, GEE may be biased for the mixture model necessary for QTL. Symons et al. (2002) and Epstein et al. (2003) considered a semi-parametric Cox proportional hazards model and a Tobit model, respectively, for gene mapping with censored survival data.

Kruglyak and Lander (1995) proposed model-free tests using Wilcoxon

rank statistics for a backcross, where there are two genotypes. Broman (2003) considered an omnibus generalization of the Wilcoxon test for the intercross. Poole and Drinkwater (1996) used the Jonckheere-Terpstra generalization of the Wilcoxon test to ordered alternatives for the intercross. Hoff et al. (2002) considered stochastic ordering with respect to genotype as an alternative to no QTL. Zou et al. (2003) and Fine et al. (2004) provided non-parametric estimators that generalize the Hodges-Lehman shift and the Kaplan-Meier survival curve to mixture models.

In this chapter, we present semi-parametric models for QTL in Section 2, and non-parametric inference applied to QTL model selection problems in Section 3. An example on tumour counts of rats is used to illustrate both semi-parametric and non-parametric inference for QTL.

2. Semi-parametric Models

It is well known that statistical methods work best when they use all available information, and in particular here, knowledge about the exact form of the phenotype model. In the best cases, this arises from extensive knowledge from previous studies and an understanding of the underlying mechanism. This ideally focuses attention on a few key parameters, such as the center (mean) and spread (variance) in a population of individuals with identical genotype. However, in many cases, a suitable parametric form is not known. We consider here semi-parametric models that encompass most common parametric models, allowing us to separate the question of model form from detection of QTL.

In the best situation, a researcher believes from previous research that a particular parametric model, such as binomial, is suitable. For instance, Poisson is often appropriate for counts of instances of some event, such as the number of offspring, while binomial is pertinent for proportions, such as germination success or disease resistance. Concentrations often follow a log-normal distribution. Generalizations that allow dispersion may be appropriate in other situations. Caution is in order if a model choice is made on the basis of raw phenotype data, as part of the histogram shape may be due to genetic variation in the sample.

When considering a model, there are three primary options: (1) just use the normal and hope it is satisfactory; (2) build a method streamlined to the 'correct' phenotype model; (3) find a transformation that makes the normal model more tenable. Instead, we propose using semi-parametric models, leaving validation of parametric form to a later investigation by the researcher once the genetics is better understood.

2.1. *Exponential tilt models*

A natural choice for the phenotype model is a common shape that is slightly modified by genotype through an 'exponential tilt':

$$\mathrm{pr}(Y = y|q, \theta) = f(y)\gamma(y|q, \beta)$$

with $\theta = (\beta, f)$, $\log(\gamma)$ a low-order polynomial tilt function that is usually linear or quadratic in y, β a vector with unknown polynomial coefficients and f an unknown density. Note that $\mathrm{pr}(y|q, \theta)$ must be a density for every genotype q, which places some technical constraints on β. If we estimate f with 'point mass' at the observed phenotypes for a sample of n individuals, these constraints become

$$\sum_{i=1}^{n} f(y_i)\gamma(y_i|q, \beta) = 1$$

regardless of the genotype q.

A test for QTL with this semi-parametric phenotype model is simply a test that $\beta = 0$ while leaving the shape of f unspecified. Many parametric models are special cases of this semi-parametric model, including normal, Poisson and binomial (Anderson 1979). Thus this approach can be used to aide in selection of a parametric model. Interestingly, we can even approximate parametric models that do not fit this form, such as negative binomial.

We draw on empirical likelihoods, which use distributions that have point mass at the observed phenotypes. Recent work (see Owen 2001) shows how we can use much of the standard likelihood machinery for point mass empirical distributions with only slight modification. Thus we can use already developed QTL interval mapping for normal data once we can evaluate the likelihood, which is

$$\begin{aligned}\mathrm{pr}(y|m, \theta, \lambda) &= \textstyle\prod_{i=1}^{n} \sum_q \mathrm{pr}(q|m_i, \lambda) f(y_i)\gamma(y_i|q, \beta) \\ &= \textstyle\prod_{i=1}^{n} w(y_i|m_i, \beta, \lambda) f(y_i)\end{aligned}$$

with weights $w(y_i|m_i, \beta, \lambda) = \sum_q \mathrm{pr}(q|m_i, \lambda)\gamma(y_i|q, \beta)$ that rely only on the phenotype and on flanking markers around the QTL. Ideally, we profile the likelihood across loci λ in the genome. Unfortunately, the profile empirical likelihood may not exist for all β in a small compact neighborhood of the null value. That is, there may be no β that make $f(y)\gamma(y|q, \beta)$ a density for all possible q.

Zou et al. (2002) proposed a partial empirical likelihood, treating markers m as fixed, by noting that the profile empirical log-likelihood can be factored as

$$\log(\mathrm{pr}(y|m, \theta, \lambda)) = \ell_1(\beta, \alpha(\beta)) + \ell_2(\beta) - n \log n.$$

The first term involves a nuisance parameter to enforce the density constraints. It uses a clever trick concerning the Lagrange multiplier α for the constraints on β, leading to point mass estimates

$$\hat{f}(y_i|m,\beta,\lambda) = \left[\sum_q \gamma(y_i|q,\beta)\sum_{i=1}^n \mathrm{pr}(q|m_i,\lambda)\right]^{-1}.$$

The second term is the partial empirical likelihood,

$$\ell_2 = \sum_{i=1}^n \log(w(y_i|m_i,\beta,\lambda)) - \sum_{i=1}^n \log\left(\sum_q w(y_i|m_i,\beta,\lambda)\rho(m_i)\right),$$

with $\rho(m_i)$ estimated as the empirical proportion of flanking markers with the genotype agreeing with m_i (for a backcross, there are four possible flanking marker genotype combinations). Notice that the partial empirical likelihood ℓ_2 does not depend on the shape of the density f.

Zou and Fine (2002) justified this partial empirical likelihood using a conditioning argument. They assumed that the marker genotypes are random, as in breeding experiments, and that the flanking marker probabilities $\rho(m_i)$ may be determined directly by the breeding design, the map function and the marker map, which are typically known. They then demonstrated that one may construct a conditional likelihood based on distribution of flanking marker genotype given phenotype not involving the baseline density f. The partial empirical likelihood is this conditional likelihood with $\rho(m_i)$ replaced by estimates. Zou and Fine (2002) and Jin et al. (2003) showed that ℓ_2 gives valid inferences regardless whether or not m_i are treated as fixed or random.

Thus we profile ℓ_2 with respect to λ, maximizing β for each possible locus. This semi-parametric profiling yields the same formal behavior as the normal-based profile likelilhood the maximum profile likelihood (see Discussion). This semi-parametric approach can be used to examine the robustness of normal or other parametric phenotype models. First, does the estimated QTL, at the maximum LOD, agree between normal and semi-parametric approaches? Second, are the data consistent with a particular parametric model, using the cumulative distributions conditional on QTL genotype in a graphical goodness-of-fit test?

Mammary Tumors in Rats

Study has shown that female rats from the Wistar-Kyoto (WKy) strain resistant to carcinogenesis were crossed with male rats from the Wistar-Furth (WF) strain (Lan et al. 2000). To identify carcinogenesis resistant genes, 383 female BC rats were generated by mating F1 progeny to WF animals. These backcross rats were scored for number of mammary carcinomas and were genotyped at 58 markers on chromosome 5. Using Mapmaker/QTL,

Lan et al. (2000) found that marker D5Rat22 was strongly associated with lower tumor counts. The mean numbers of counts estimated from the normal mixture are 2.68 and 5.43 for the WKy/WF and WF/WF genotypes, respectively at the putative QTL identified.

Zou et al. (2002) applied the semiparametric method to this rat data and the results are summarized in Figures 1 and 2. In Figure 1, the partial likelihood ratio statistic is shown as a function of location on chromosome 5. The LOD score calculated from the partial likelihood ratio statistic is also given. For comparison, the profile from a normal mixture using MapMaker/QTL is displayed. Both curves are very similar with peaks near D5Rat22. The estimated distribution functions for Wky/WF and WF/WF genotypes were computed at the locus giving the maximum LOD score under the semiparametric and normal mixtures. These are displayed in Figure 2 along with 0.95 pointwise confidence intervals. The plots exhibit that WF/WF rats have higher tumor counts. The estimated means for carcinomas in WKy/WF and WF/WF rats are 2.69 and 5.45, respectively. The estimated distributions from the normal mixture are rather different from the semiparametric estimates and may lie outside the confidence intervals. Other estimates (not shown) from a negative binomial model (Drinkwater and Klotz 1981) fall entirely within the 0.95 limits.

2.2. *Measuring the shift of center*

Another way to generalize the normal model is to suppose that QT genotypes can shift the center but not otherwise change the shape of the model. That is,

$$\mathrm{pr}(y|q,\theta) = F(y + q\beta)$$

with $\theta = (\beta, F)$, β consisting of a few parameters and F a completely unspecified distribution. This semi-parametric shift model has a natural estimator of shift suggested by Hodges and Lehmann. All one has to do is divide the phenotypes into groups based on QT genotype q and find β that shifts the medians of all groups to coincide.

Suppose we knew the shift, say β, and we knew the genotypes q. Then the shifted values $y_i(\beta) = y_i + (q_i - \bar{q}_\cdot)\beta$ would all have the same distribution F. Consider the linear rank statistic

$$T(b|y,q) = \sum_{i=1}^{n} (q_i - \bar{q}_\cdot) \frac{\mathrm{rank}(y_i(\beta))}{n+1} \ ,$$

which depends on the phenotypes only through the ranks of their shifted values. In the next section, we develop this into a formal test for $\beta = 0$, but here we are interested in estimating the shift. If we knew q, then we

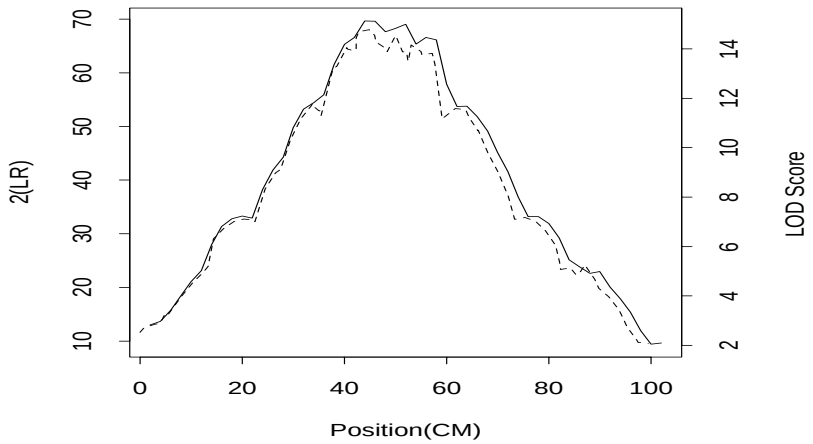

Figure 1 Likelihood ratio statistics and LOD score on chromosome 5. Solid line is the semiparametric mixture and the dashed is the normal mixture. (From Zou et al. 2002.)

could use the Hodges-Lehmann estimator $\hat{\beta} = \text{median}\{b|T(b) \approx 0\}$. Note that the linear rank statistic may not reach zero, so in practice we take the closest values on either side and average them.

This seems rather difficult to do in practice since q are unknown. However, Haley-Knott regression provides a decent approximation. In other words, we can substitute unknown q with its expectation when estimating β:

$$\text{pr}(y|q, \theta) = F(y + E(q)\beta),$$

with $E(q)$ the expectation of q given flanking markers to the loci λ (Haley and Knott 1992). Haley-Knott least squares estimators are consisten, but may be inefficient, while modified Hodges-Lehmann (HL) estimators may have bias, since they are nonlinear in q, depending on the median. Nevertheless, our HL estimators perform well in simulations. Our investigation for a single QTL shows that (Zou et al. 2003) the approximation works well for linkage maps that are relatively dense (when the average marker distance is no larger than 20 cM) which is true for most of the modern QTL mapping studies. The proposed estimator of β is more efficient than its traditional estimator based on the normality assumption when the data

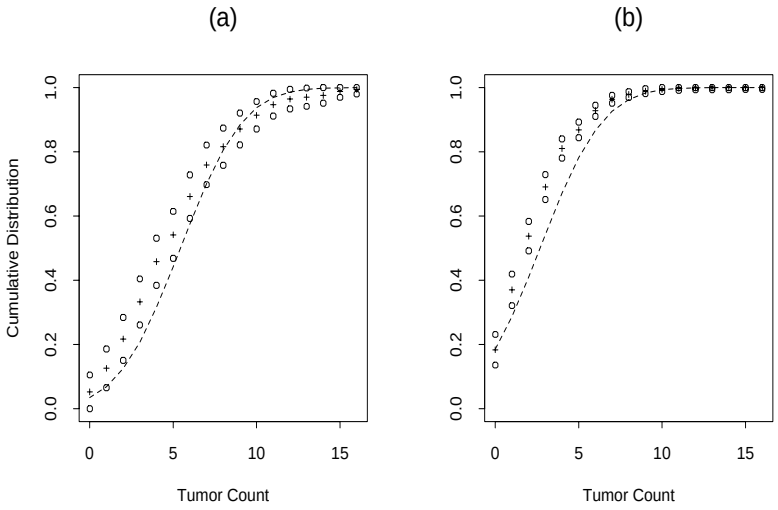

Figure 2 Point estimates (+) and 0.95 pointwise confidence limits (0) for cumulative distributions at location of maximum partial likelihood ratio statis tic. Dashed lines are point estimates from the normal mixture model. (a) WF/WF; (b) WKy/WF. (From Zou et al. 2002.)

is not normally distributed. Further, Haley-Knott (1992) regression gives valid estimates and testing when data are not normal.

Listeria Monocytogene Time-to-Death in Mice

Our second example relates to the date on the time-to-death following infection with Listeria monocytogenes of 116 F2 mice from an intercross between the BALB/cByJ and C57BL/6ByJ strains (Boyartchuk et al 2001). The histograms of the log time-to-death of the non-survivors are given in Figure 3. 31 mice which is roughly 30% of mice survive beyond 264 days. From the histogram it is hard to justify that the log time-to-death of the non-survivors is normally distributed. Broman (2003) applied four different methods, including both the standard interval mapping and non-parametric interval mapping, to this data set and showed that the locus on chromosome 1 appears to have effect only on the average time-to-death among the non-survivors. For this reason, our analysis will be restricted on chromosome 1 for those non-survivors.

The additive and dominance estimators from standard interval mapping are 0.262, 0.059, respectively while they are 0.257, 0.038, respectively

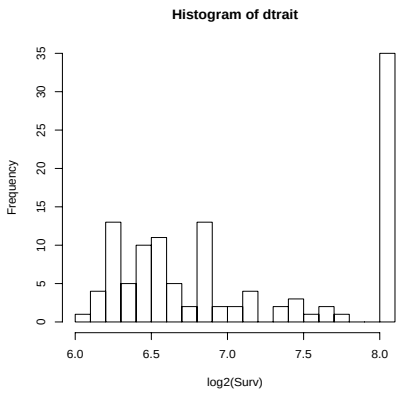

Figure 3 Histogram of $log2$(survival time), following infection with *Lis teria Monocytogenes*. 31 mice recovered from the infection and survived to the end of experiment $264hr$ ($log2(264) = 8$).

based on the rank based method. Therefore, the non-parametric rank based analysis confirms the results by Broman (2003).

3. Non-parametric Models

The semi-parametric models are quite useful, but they still rely on some common shape in some sense. What if we want to allow completely arbitrary shaped distributions with different QTL genotypes?

Here we examine non-parametric methods that make no assumptions about the shape of the distribution, that is we focus on cumulative distributions conditional only on the QT genotype

$$\mathrm{pr}(Y \leq y|q) = F_q(y) \ .$$

This approach is more robust to heavy-tailed phenotype distributions and to occasional outliers.

Estimates of shift discussed in the previous section could be useful here, but they are actually semi-parametric. We wish to estimate the conditional

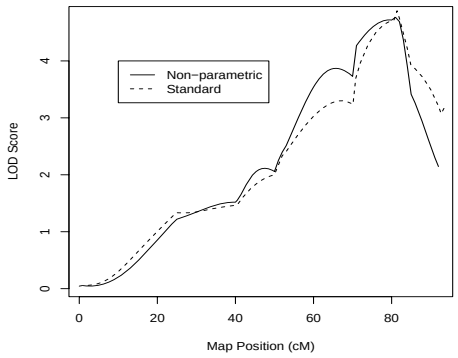

Figure 4 LOD score curves from standard interval mapping (dashed line) and nonparametric interval mapping (solid line).

distributions F_q without any assumptions of shape. Here is the basic idea. We estimate the cumulative distributions given flanking markers, $\text{pr}(Y \leq y|m, \lambda)$, by dividing phenotypes into groups based on flanking markers and summing up the corresponding histograms (details below). Now notice that the phenotype distributions conditional on QTL are mixtures of these flanking-marker distributions:

$$\text{pr}(Y \leq y|m, \lambda) = \sum_q \text{pr}(q|m, \lambda) F_q(y) \ .$$

Given QTL λ, we can calculate $\text{pr}(q|m, \lambda)$. If there are m QTL, then in a backcross there are 2^{2m} possible flanking marker values but only $L = 2^m$ possible QT genotypes. Thus we have fewer unknowns (F_q) than knowns in a set of linear equations, and we can estimate. This argument can be extended to handle missing marker genotypes and other types of experimental crosses.

To be specific, consider the cumulative distributions

$$H_i(y) = \text{pr}(y_i \leq y|m_i, \lambda) \ .$$

Here is a way to get the estimator of H_i. Let $N_i(y) = I(y_i \leq y)$, being 1 if $y_i \leq y$ or 0 if $y_i > y$. Divide experimental units up into sets based on the

value of their flanking markers around the loci λ. Let s be one such set. For each unit i in this set s, average the indicators across this set:

$$\hat{H}_i(y) = \sum_{k \in s} N_k(y)/n_s$$

with n_s being the size of the set s. This gives an empirical estimator of $H_i(y)$ which increases from 0 to 1 as y increases, taking steps of size $1/n_s$. All individuals in set s have this same estimator. Thus,

$$\sum_{i \in s} \hat{H}_i(y) = \sum_{i \in s} N_i(y) \ .$$

Let $H = (H_1, \cdots, H_n)^{\mathrm{T}}$ be the cumulative phenotype distributions conditioned on flanking markers, and F be a column vector across the QT genotypes of F_q. Combine the segregation model into an $n \times 2^m$ matrix R with $R_{iq} = \mathrm{pr}(q|m_i, \lambda)$. Thus

$$H(y) = RF(y) \ .$$

In the case of fully informative flanking markers, the 'best' (least squares) estimator of $F_q(y)$ given QTL λ is

$$\hat{F}(y|\lambda) = (R^{\mathrm{T}}R)^{-1}R^{\mathrm{T}}\hat{H}(y) = W\hat{H}(y) = WN(y)$$

with $N = (N_1, \cdots, N_n)^{\mathrm{T}}$. The last equality holds since we are effectively summing first over individuals with the same flanking markers. This makes sense, since we can think of the problem as having the cumulative distribution as the phenotype of interest, with data being $N_i(y) = I(y_i \leq y)$. The least squares estimator of $F_q(y)$ minimizes the following sum of squares:

$$\sum_{i=1}^{n} \left[I(y_i \leq y) - \sum_q \mathrm{pr}(q|m_i, \lambda) F_q(y) \right]^2 \ .$$

That is, we find the best fit to the cumulative distribution of phenotypes y based on the segretation model and on phenotype model given QTL at λ.

The covariances of the phenotype cumulative distribution arise directly from the binomial model, since we are estimating a probability. For $y \leq y\prime$,

$$\mathrm{cov}\left(\hat{F}(y|\lambda), \hat{F}(y\prime|\lambda)\right) = WH(y)(I - H(y\prime))W^{\mathrm{T}} \ ,$$

which can be estimated by $WN(y)(I - N(y\prime))W^{\mathrm{T}}$.

The linear rank test provides a formal non-parametric testing framework to infer QTL, assuming common shape. Following this localization, the above estimators can provide graphical assessment of the shape of the distribution for each genotype.

The rank tests of Kruglyak and Lander (1995) may have low power to detect differences between the phenotypic distributions. A test for homogeneity of the components may also be conducted using the proposed nonparametric estimators. For given y, the null hypothesis is $H_0 : A\hat{F}(t) = 0$, where A is an $(L-1) \times L$ matrix containing $(L-1)$ linearly independent contrasts of $F_Q(y)$s corresponding all possible QTL genotypes q. Under H_0, the statistic

$$\mathcal{L}(y) = \{A\hat{F}(y)\}\{A\hat{\Sigma}(y,y)A^T\}^{-1}\{A\hat{F}(y)\}^T$$

has a chi-squared distribution with $L-1$ degrees of freedom. Evaluating the distribution of $\mathcal{L}$ as a process in $y \in [0, \tau]$ (τ is the maximum y value observed would) enable omnibus testing procedures which are sensitive to differences amongst the component distributions at all time points. For example, using $\sup_y \mathcal{L}(y)$ would provide a statistic which is sensitive to all alternatives, unlike the test of Kruglyak and Lander (1995). The theoretical developments of $\sup_y \mathcal{L}(y)$ appear to be rather challenging and deserves further investigation. In practice, one might consider using the bootstrap to approximate the distribution of the sup test under H_0 across the genome.

Again, the proposed method has been applied to the mammary tumor rat data (Fine et al. 2004). We compute nonparametric estimates of the carcinoma distributions for the WKy/WF and WF/WF genotypes at the estimated QTL and the estimated distributions are displayed in Figure 5 along with 0.95 pointwise confidence intervals. The plots exhibit that WF/WF rats have higher tumor counts. Further, the estimated distribution $\hat{F}(y)$ provides another goodness of fit method of the traditional parametric QTL mapping. The estimated means in the WKy/WF and WF/WF groups are 2.64 and 5.46, respectively, which agrees with Mapmaker/QTL. However, the estimated distributions from the normal mixture are rather different from the nonparametric estimates; these are not shown. Instead, the estimated components from a model with $F_{WKy/WF}$ and $F_{WF/WF}$ assumed to be negative binomial, which was fitted by Lan *et al.* (2001), are displayed in Figure 5. These fall entirely within the 0.95 limits, indicating that this model matches the data well.

4. Discussion

The Wilcoxon rank-sum test was extended to interval mapping by Kruglyak and Lander (1995). For related sum of scores tests that might be used as alternatives, see Puri and Sen (1985) or other texts on non-parametric statistics.

Technical details for the QTL exponential tilt can be found in Zou,

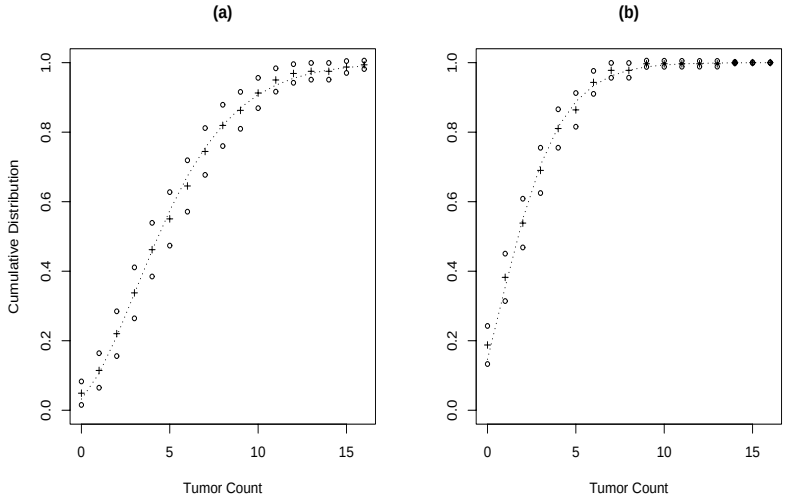

Figure 5 Point estimates (+) and 0.95 pointwise confidence limits (0) for cumulative distributions at location of maximum partial likelihood ratio statis tic. Dashed lines are point estimates from the negative binomial mixture model. (a) WF/WF; (b) WKy/WF. (From Fine et al. 2004.)

Fine and Yandell (2002), based on empirical likelihood work of Qin (Qin & Lawless 1994; Qin 1999). See Owen (2001) for a comprehensive treatment of empirical likelihoods. Zou and Fine (2002) showed how the partial empirical likelihood is closely related to the conditional likelihood. This connection raises interesting robustness issues with respect to selective genotyping and selective phenotyping that are discussed in Jin et al. (2003).

Fine, Zou and Yandell (2001) developed non-parametric cumulative distributions for QTL phenotypes for uncensored and censored data. Speed (pers. comm.) developed a QTL version of the Cox proportional hazards. Recent research has touched on time series and repeated measures analysis in the QTL context.

Calculating thresholds and power are important practical issues in the design and analysis of any QTL study. However, the usual point-wise significance level based on chi-square approximation is inadequate because the entire genome is tested for the presence of a QTL. Theoretical approximations based on the Ornstein-Uhlenbeck diffusion process have been developed to determine threshold and power (Lander and Botstein 1989; Dupuis and Siegmund 1999; Rebai et al. 1994, 1995; Zou et al. 2001, 2002)

in some simple experimental crosses. However, permutation procedure is time consuming and may not be applicable under some conditions. The theoretical approximation is not readily available for any study designs and hard to obtain for complicated models. Empirical permutation procedures to estimate genome-wide threshold values for traditional interval mapping proposed by Churchill and Doerge (1994) and widely used for normal data can be readily applied to the semiparametric and nonparametric methods reviewed here. Recently, Zou et al. (2004) proposed a new resampling procedure to assess the significance of genome-wide QTL mapping that is computationally much less intensive than Churchill and Doerge (1994). Further, it is applicable to complicated QTL mapping models that the permutation and theoretical methods cannot handle.

Acknowledgements

This work was supported in part by United States Department of Agriculture Cooperative State Research, Education, and Extension Service and by National Institutes of Health/NIDDK, 5803701 and 66369-01.

References

1. Anderson J. A(1979). Multivariate logistic compounds. *Biometrika* **66**, 17-26.
2. Basten CJ, Weir BS, Zeng ZB (1995) QTL Cartographer: A Reference Manual and Tutorial for QTL Mapping. Center for Quantitative Genetics, NC State Univeristy.
3. Belknap JK, Richards SP, O'Toole LA, Helms ML, and Phillips TJ (1997). Short-term selective breeding as a tool for QTL mapping: ethanol preference drinking in mice. *Behavior Genetics* **27**, 55-66.
4. Broman KW (2003). Quantitative trait locus mapping in the case of a spike in the phenotype distribution. *Genetics* **163**, 1169-1175.
5. Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-890.
6. Churchill GA and Doerge RW(1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-971.
7. Diao G, Lin DY, and Zou F (2004). Mapping quantitative trait loci with censored observations. *Genetics* **168**, 1689-1698.
8. Doerge RW, Zeng ZB, Weir BS (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statist Sci* **12**, 195-219.

9. Dupuis J, Siegmund D. (1999) Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**: 373-386.

10. Epstein MP, Lin X, Boehnke M (2003) A Tobit variance-component method for linkage analysis of censored trait data. *Amer J Hum Genet* **72**, 611-620.

11. Fine JP, Zou F, and Yandell BS (2004). Nonparametric estimation of mixture models, with application to quantitative trait loci. *Biostatistics* **5**, 501-513.

12. Hackett CA(1997). Model diagnostics for fitting QTL models to trait and marker data by interval mapping. *Heredity* **79**, 319-328.

13. Hackett CA and Weller JI (1995) Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**, 1254-1263.

14. Haley C. and Knott S. (1992) A simple regression method for mapping quantitative trait loci of linked factors. *J Genetics* **8**, 299-309.

15. Haston CK, Zhou X, Gumbiner-Russo L, Irani R, Dejournett R, Gu X, Weil M, Amos CI, and Travis EL(2002). Universal and radiation-specific loci influence murine susceptibility to radiation-induced pulmonary fibrosis. *Cancer Research* **62**, 3782-3788.

16. Hoff PD, Halberg RB, Shedlovsky A, Dove WF, Newton MA (2002) Identifying carriers of a genetic modifier using nonparametric Bayes methods. in *Case Studies in Bayesian Statistics* **5**, Springer-Verlag, 327-342.

17. Jansen RC (1992) A general mixture model for mapping quantitative trait loci by using molecular markers. *Theor Appl Genet* **85**, 252-260.

18. Jin C, Fine J, Yandell B (2003) A unified semiparametric framework for QTL analyses, with application to spike phenotypes. *J Amer Statist Assoc* (in review).

19. Kao CH and Zeng ZB (1997) General formulae for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 653-665.

20. Kruglyak L and Lander ES (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421-1428.

21. Lan H, Kendziorski CM, Haag JD, Shepel LA, Newton MA, and Gould MN (2001). Genetic loci controlling breast cancer susceptibility in the Wistar-Kyoto rat. *Genetics* **157**, 331-339.

22. Lander E, Green P, Abrahamson J, Barlow A, Daley M, Lincoln S, and Newburg L (1987). MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174-181.

23. Lander ES, Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-199.

24. Lange C, Whittaker JC (2001) Mapping quantitative trait loci using generalized estimating equations. *Genetics* **159**, 1325-1337.

25. Lynch M and Walsh B (1998). *Genetics and analysis of quantitative traits*. Sunderland, Mass., Sinauer.

26. Mackay TFC (2001) The genetic architecture of quantitative traits. *Ann Rev Genet* **35**, 303-339.

27. Mackay TF and Fry JD (1996) Polygenic mutation in Drosophila melanogaster: Genetic interactions between selection lines and candidate quantitative trait loci. *Genetics* **144**, 671-688.

28. McIntyre LM and Coffman C Doerge RW (2000) Detection and location of a single binary trait locus in experimental populations. *Genet Res* **78**, 79-92.

29. Nettleton D (1999) Order restricted hypothesis testing in a variation of the normal mixture model. *Can J Statist* **27**, 383-394.

30. Nettleton D (2002) Testing for ordered means in a variation of the normal mixture model. *J Statist Plan Infer* **107**, 143-153.

31. Nettleton D and Doerge RW (2000) Accounting for variability in the use of permutation testing to detect quantitative trait loci. *Biometrics* **56**, 52-58.

32. Nettleton D and Praestgaard J (1998) Interval mapping of quantitative trait loci through order-restricted inference. *Biometrics* **54**, 74-87.

33. Owen AB (2001) *Empirical Likelihood*. Monographs on Statistics and Applied Probability, v. 92.

34. Poole TM and Drinkwater NR (1996) Two genes abrogate the inhibition of murine hepatocarcinogenesis by ovarian hormones. *Proc Nat Acad Sci USA* **93**: 5848-5853.

35. Puri ML and Sen PK (1985). *Nonparametric methods in general linear models*. John Wiley & Sons.

36. Rao SQ and Li X (2000) Strategies for genetic mapping of categorical traits. *Genetica* **109**, 183-197.

37. Rebai A, Goffinet B, and Mangin B (1994) Approximate thresholds of interval mapping tests for QTL detection. *Genetics* **138**: 235-240.

38. Rebai A, Goffinet B, and Mangin B (1995) Comparing power of different methods of QTL detection. *Biometrics* **51**: 87-99.

39. Shepel LA, Lan H, Haag JD, Brasic GM, Gheen ME, Simon JS, Hoff P, MA Newton, and Gould MN (1998) Genetic identification of multiple loci that control breast cancer susceptibility in the rat. *Genetics* **149**, 289-299.

40. Qin, J (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *The Annals of Statistics* **27**, 1368-84.

41. Qin J and Lawless JF (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22**, 300-25.

42. Sax K (1923). The association of size differences with seed-coat pattern and pigmentation in Phaseoulus vulgaris. *Genetics* **8**, 552-560.

43. Symons RCA, Daly MJ, Fridlyand J, Speed TP, Cook WD, Gerondakis S, Harris AW, and Foote SJ (2002). Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in E5-v-abl transgenic mice. *Proc Natl Acad Sci USA* **99**, 11299-11304.

44. Visscher PM, Haley CS, and Knott SA (1996) Mapping QTLs for binary traits in backcross and F-2 populations. *Genet Research* **68**, 55-63.

45. Wang X, Le Roy I, Nicodeme E, Li R, Wagner R, Petros C, Churchill GA, Harris S, Darvasi A, Kirilovsky J, Roubertoux PL, and Paige B (2003). Using Advanced Intercross Lines for High-Resolution Mapping of HDL Cholesterol Quantitative Trait Loci. *Genome Research* **13**, 1654-1664.

46. Weller JI (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**, 627-640.

47. Wright FA and Kong A (1997). Linkage mapping in experimental crosses: the robustness of single-gene models. *Genetics* **146**, 417-425.

48. Xu S and Atchley WR (1995). A random model approach to interval mapping of quantitative genes. *Genetics* **141**, 1189-1197.

49. Yi N and Xu S (2000) Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics* **155**, 1391-1403.

50. Zou F and Fine JP (2002) Note on a partial empirical likelihood. *Biometrika* **89**, 958-961.

51. Zou F, Fine JP, Hu J, and Lin DY (2004). An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics* **168**, 2307-2316.

52. Zou F, Fine JP, and Yandell BS (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika* **89**, 61-75.

53. Zou F, Yandell BS, and Fine JP (2003). Rank-based statistical methodologies for quantitative trait locus mapping. *Genetics* **165**, 1599-1605.

54. Zou F, Yandell BS, and Fine JP (2001). Statistical issues in the analysis of quantitative traits in combined crosses. *Genetics* **158**, 1339-1346.