Copyright © 1997 by the Genetics Society of America

Linkage Mapping in Experimental Crosses: The Robustness of Single-Gene Models

Fred A. Wright* and Augustine Kong†

**Family and Preventive Medicine and the Cancer Center University of California, San Diego, California 92093-0622 and †Department of Statistics, The University of Chicago, Chicago, Illinois 60637*

Manuscript received May 15, 1996
Accepted for publication January 21, 1997

ABSTRACT

The robustness of parametric linkage mapping against model misspecification is considered in experimental breeding designs, with a focus on *localization* of the gene. By examining the expected LOD across the genome, it is shown that single-gene models are quite robust, even for polygenic traits. However, when the marker map is of low resolution, linked polygenes can give rise to an apparent "ghost" gene, mapped to an incorrect interval. The results apply equally well to quantitative traits or qualitative (categorical) traits. The results are derived for backcross populations, with a discussion of extensions to intercross populations and relative-pair mapping in humans.

IN linkage mapping, researchers often begin by adopting a parametric model for the dependence of the phenotype on the unknown (single) gene of interest. The LOD score is then used to estimate the gene location. A natural question arises: how robust is the LOD score? In other words, if one or more trait genes exist but the assumed phenotype model is incorrect, will the LOD score still tend to be maximized near the true gene or genes? This issue has not been extensively investigated in the experimental breeding setting, although some robustness is implicitly assumed in the common practice of estimating multiple gene locations using a single-gene model (LANDER and BOTSTEIN 1989, *e.g.*, JACOB *et al.* 1991). Related robustness investigations in human mapping include CLERGET-DARPOUX *et al.* (1986), VIELAND *et al.* (1992), RISCH and GIUFFRA (1992), and HODGE and ELSTON (1994). This paper is strictly concerned with the localization of genes, assuming that linkage has already been established. Other investigators (WILLIAMSON and AMOS 1990; FREIMER *et al.* 1993) have examined another form of robustness involving the adequacy of linkage testing where some parameters (*e.g.*, marker genotype frequencies) are misspecified.

The LOD score is based on a likelihood model for the phenotype, and if the model is specified correctly it follows from the consistency property of likelihood (COX and HINKLEY 1974) that as the number of individuals increases, the estimate of the gene location will converge to the true location. However, under model misspecification a likelihood-based estimate may converge to an *incorrect* location on the genome. Furthermore, some degree of misspecification may be unavoidable, and the trait may in fact be polygenic. A main point of this paper is to show that with large sample sizes the LOD score tends to reach its maximum within the marker interval containing the gene, as long as no more than one trait gene resides on the chromosome under study. This result may appear to contrast with the finding of CLERGET-DARPOUX *et al.* (1986) that misspecification may result in highly biased location estimates. However these authors assumed that only a *single* marker is present, while the present paper demonstrates that additional markers will greatly reduce the bias.

Even stronger results hold for the *dense marker* case in which a marker resides at every locus. Although not feasible with current technology, this hypothetical scenario provides insight into the effect of marker maps of very high resolution. The results are established here for backcross populations, (see LANDER and BOTSTEIN 1989 for an introduction to mapping in experimental crosses) and apply to doubled haploids with trivial modification. Extensions to intercross populations and human relative pair designs are considered in the DISCUSSION. Some finer statistical details are contained in a technical report (WRIGHT and KONG 1995) and are omitted here.

Throughout this paper, the term *assumed model* refers to the model under which the likelihood is constructed, while *true model* refers to the true state of nature. To avoid confusion, the term *gene* will be used to refer only to loci that influence the phenotype of interest.

Likelihood estimation: LANDER and BOTSTEIN (1989) provided an important development of the likelihood method for quantitative trait locus (QTL) estimation in experimental populations. The extension of this likelihood approach to more general settings (*e.g.*, categorical traits) is relatively straightforward (JANSEN 1993,

Corresponding author: Fred A. Wright, Department of Family & Preventive Medicine, The UCSD Cancer Center, 9500 Gilman Dr., 0622, La Jolla, CA 92093-0622. E-mail: fwright@ucsd.edu

CHURCHILL and DOERGE 1994). The examples in this paper involve QTL mapping, but the theoretical development is equally applicable to quantitative and categorical traits.

Let P0 and P1 denote two parental inbred lines, from which a population of backcross individuals is created, with (say) P0 as the recurrent parent. Let y_i represent the phenotype for the ith individual and g_i the genotype at a particular locus. Following LANDER and BOTSTEIN (1989), g_i may be coded as a (0, 1) indicator variable for the number of P1 alleles. The assumed model supposes that the gene lies at the locus, and that the phenotype distribution is either f_0 (if $g_i = 0$), or f_1 (if $g_i = 1$). These two phenotype distributions are themselves unknown, but are specified by a parameter vector $\boldsymbol{\lambda}$ that can be estimated from the data. We will use $\hat{\boldsymbol{\lambda}}$ to denote this maximum likelihood estimate (MLE), *i.e.*, the parameter value that maximizes the likelihood or probability $L(\boldsymbol{\lambda})$ of the observed data. Finally, we use $\mathbf{m}_i$ to represent the marker genotype information for the ith individual. It is straightforward to show (*e.g.*, equation 7 of LANDER and BOTSTEIN 1989) that the likelihood is

$$L(\boldsymbol{\lambda}) = \prod_i [f_0(y_i; \boldsymbol{\lambda}) P(g_i = 0 \mid \mathbf{m}_i) + f_1(y_i; \boldsymbol{\lambda}) P(g_i = 1 \mid \mathbf{m}_i)].$$

The quantities $P(g_i \mid \mathbf{m}_i)$ appear because the genotype is not observed directly, and the probabilities are computed using the genotypes and positions of the flanking markers. We will assume throughout that Haldane's map function applies, *i.e.*, there is no interference. (Note that in the special instance that the locus is exactly at a marker the likelihood contribution from the ith individual will simply be $f_0(y_i; \boldsymbol{\lambda})$ or $f_1(y_i; \boldsymbol{\lambda})$, according to the value of g_i, because the genotype is known exactly.)

The familiar LOD score method, as implemented in experimental populations, summarizes the evidence for the gene as follows:

$$LOD = \log_{10}(L(\hat{\boldsymbol{\lambda}})/L(\hat{\boldsymbol{\lambda}}_{H_0})),$$

where $\hat{\boldsymbol{\lambda}}_{H_0}$ is the constrained MLE under the null hypothesis that no gene is linked. The LOD score is then computed at each genetic location, with high scores (exceeding a threshold computed to control the false positive rate) used to identify regions likely to contain a trait gene. Note that the LOD score takes a somewhat different form in traditional two-point linkage analysis in humans (OTT 1991).

The following example illustrates LOD score mapping under misspecification.

Example: *The normal single-QTL assumed model.* Most QTL mapping studies use the assumed model

$$y_i = a + bg_i + \epsilon_i, \qquad (1)$$

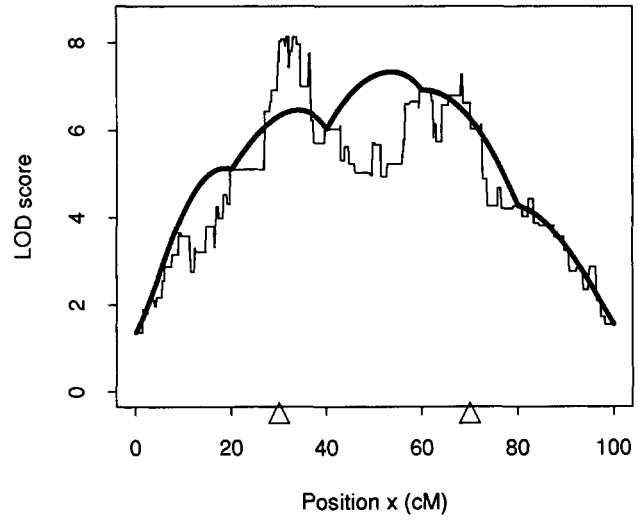

FIGURE 1.—The two-QTL true model with a QTL at 30 cM and a second QTL of somewhat smaller effect at 70 cM (true locations indicated by △). A normal single-QTL model is assumed and the LOD score for 100 simulated individuals is given for dense markers (thin curve) and markers at 20-cM intervals (bold curve).

where the ϵ_i are independent $N(0, \sigma^2)$ errors. Here $\boldsymbol{\lambda} = (a, b, \sigma^2)$, f_0 is $N(a, \sigma^2)$ and f_1 is $N(a + b, \sigma^2)$.

Now suppose that in fact there are two QTLs on the chromosome, with true model

$$y_i = 1 + u_i + .75\, v_i + \epsilon_i, \qquad (2)$$

where $\epsilon_i \sim N(0, 1)$ and u_i, v_i are the genotypes of the QTLs at 30 and 70 cM on a single chromosome of length 100 cM.

One hundred such backcross individuals were simulated for illustration. Figure 1 plots the LOD curve for these simulated individuals under two marker density scenarios: the dense marker case (thin curve) and the nondense marker case (bold curve), with markers at intervals of 20 cM. The dense-marker LOD has a distinct peak at 32 cM and a lesser peak at 68 cM, corresponding to the genes at 30 and 70 cM, respectively. In this example (as will be proven later) it appears that the dense-marker case provides a kind of robustness for location estimation, with the global maximum near one of the gene locations although the model is misspecified. Note that the dense-marker LOD is a step function, with jumps at the observed crossover locations. Examinations of such dense-marker LODs have been performed for breeding designs by DAVARSI *et al.* (1993) and KONG and WRIGHT (1994), and for human relative pairs by KRUGLYAK and LANDER (1995).

In contrast, the nondense LOD peaks at 53 cM, in an interval that contains *neither* of the trait genes. Several researchers have noted this phenomenon of a so-called "ghost" gene (HALEY and KNOTT 1992; MARTINEZ and CURNOW 1992), but the underlying reasons for it have not been extensively studied.

The approach: A fixed set of fully informative mark-

ers is typed for each individual, and a single-gene model is adopted for likelihood estimation. Two forms of model violation are considered:

1. the phenotype distributions are incorrectly specified,
2. more than one gene influences the trait.

HUBER (1967) has shown that under misspecification maximum likelihood estimates will, with increasing sample size, generally converge to the value that has greatest expected log-likelihood (see APPENDIX, Part 1). This result has been used in other linkage contexts (WILLIAMSON and AMOS 1990, 1995). We denote the expected log-likelihood maximized over $\boldsymbol{\lambda}$

$$M = \max_{\boldsymbol{\lambda}} E(\log L_i(\boldsymbol{\lambda})),$$

where L_i is an arbitrary individual i's contribution to the likelihood and "log" signifies natural logarithm. M may be computed at each putative gene location, forming an entire curve $M(x)$, which has the same shape as the expected LOD curve. If the maximum of M occurs at a true gene location, then we declare the likelihood procedure (*i.e.*, the LOD score) "robust" for localizing that gene.

In the following RESULTS section we examine the function $M(x)$ to determine the effects of marker density and model misspecification on the estimation procedure. The theory is illustrated with QTL mapping examples. A common theme throughout the paper is that robustness depends more on the presence of reasonable flexibility in the assumed model than in the particular form of the true model. This fact may bring peace of mind to the researcher, who has control over the assumed model and only imperfect knowledge of the true model.

Additional notation: A consequence of the breeding design is that for each individual the chromosome will be composed of entire regions of genotype = 0 and genotype = 1, with the crossovers forming the boundaries between regions. We will use $g(x)$ to denote the genotype at location x. Let x^* represent the true location of the gene, with the phenotype following either the true distribution h_0 if $g(x^*) = 0$, or h_1 if $g(x^*) = 1$ (the notation required for multiple true genes will be introduced as necessary).

RESULTS

One gene per chromosome, dense markers

Here we consider the dense marker situation where a single gene lies on the chromosome under study. The main result of this section is covered in some detail, because it underlies much of the subsequent development. To illustrate the robustness in this case, consider the following contrived example:

Example: Suppose a researcher adopts the normal single-QTL assumed model (1), but does not wish to bother with maximum likelihood estimation of the phenotypic means and variances. The researcher decides to *fix* the assumed phenotype distributions as $f_0 = N(0, 1)$ and $f_1 = N(1, 1)$ when performing likelihood estimation of the gene location (Figure 2A). In other words, the researcher assumes a particular parameter value $\boldsymbol{\lambda}_1 = (a, b, \sigma^2) = (0, 1, 1)$, and will consider no other possible parameter values. Note that this is a far cry from the usual maximum likelihood approach in which the best-fitting means and variance are computed at each putative location. Now also suppose that the true phenotype distributions h_0 and h_1 are as plotted in Figure 2B, with the gene at 30 cM on a 100 cM chromosome with dense markers. Clearly the model is misspecified: the assumed distributions have a very different form than the true distributions. Nonetheless, the expected log-likelihood peaks at the true gene location x^* (Figure 2C), indicating that the researcher's location estimate is robust, even under misspecification.

To understand this phenomenon, we will examine the expected log-likelihood over the chromosome.

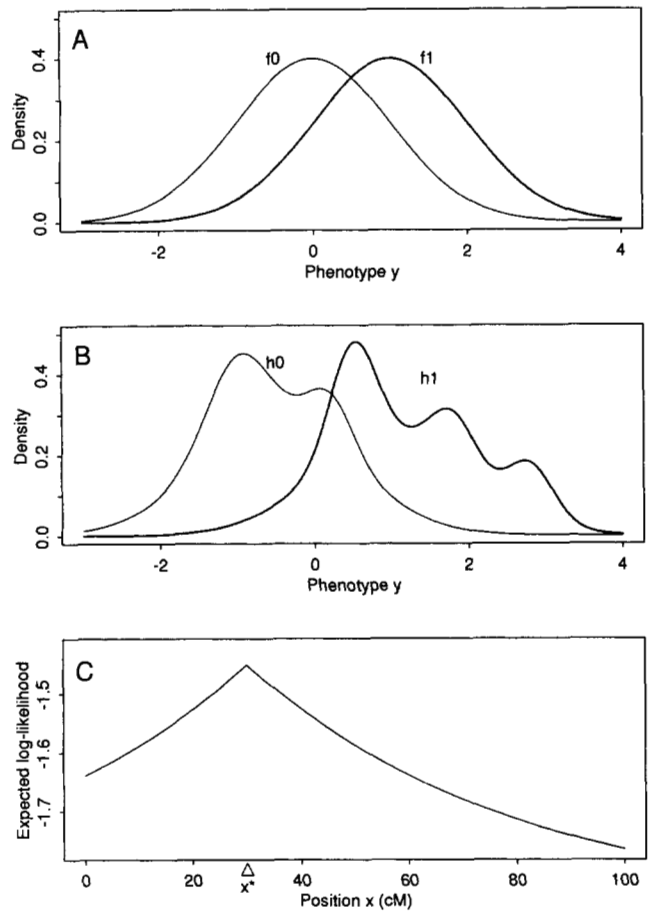

FIGURE 2.—Misspecification of the phenotype model. (A) The assumed distributions f_0 and f_1. (B) The true distributions h_0, h_1. (C) The expected log-likelihood across the chromosome when the markers are dense. Despite the misspecification, the function is maximized at exactly the true location $x^* = 30$ cM (indicated by △).

TABLE 1

Genotypes, phenotype distributions and likelihood

Genotypes $g(x^*)$, $g(x)$	Probability	Likelihood at x	True distribution of y
{0, −0}	$(1-\theta)/2$	$f_0(y)$	h_0
{0, −01}	$\theta/2$	$f_1(y)$	h_0
{1, −00}	$\theta/2$	$f_0(y)$	h_1
{1, −01}	$(1-\theta)/2$	$f_1(y)$	h_1

Again assume that the researcher rigidly adopts the fixed phenotype distributions f_0 and f_1 for the assumed model, *i.e.*, the parameter $\boldsymbol{\lambda}$ is fixed rather than estimated. Recall that the likelihood contribution for a single individual i is $f_0(y_i)$ in regions of the genome where $g_i = 0$ and is $f_1(y_i)$ where $g_i = 1$. The likelihoods and true phenotype distributions are given below for the four possible pairs of genotypes at x^* (the true location) and another location x, where θ is the recombination fraction between the two loci (Table 1). Summing over the genotype possibilities and taking expectations over the phenotypes gives the expected log-likelihood at x

$$E[\log L_i] = \left\{\frac{1-\theta}{2} E_{h_0}[\log f_0(y)] + \frac{\theta}{2} E_{h_0}[\log f_1(y)] + \frac{\theta}{2} E_{h_1}[\log f_0(y)] + \frac{1-\theta}{2} E_{h_1}[\log f_1(y)]\right\}$$

$$= \frac{1}{2}\{E_{h_0}[\log f_0(y)] + E_{h_1}[\log f_1(y)]\} - \frac{\theta}{2} K, \quad (3)$$

where $K = E_{h_0}[\log (f_0(y)/f_1(y))] + E_{h_1}[\log (f_1(y)/f_0(y))]$ is a constant. Because θ is the recombination fraction between the (fixed) true location and the current putative location x, it is apparent that if $K > 0$, then the expectation decreases for putative locations moving away from x^*. In other words, if $K > 0$, then the maximum of the expected log-likelihood must be exactly at x^*, and robustness holds.

K is a form of distance measure between probability distributions, similar to that discussed in KONG and WRIGHT (1994). Roughly speaking, $K > 0$ if the distribution h_0 is "more similar" to f_0 than it is to f_1, and if h_1 is more similar to f_1 than f_0. It is intuitively reasonable that $K > 0$ in our example above, as one can see by inspection of Figure 2, A–C that h_0 has a similar mean and variance as f_0, and h_1 has a similar mean and variance as f_1.

To establish robustness generally, we must consider what happens when f_0 and f_1 are specified up to the parameter $\boldsymbol{\lambda}$, and where the likelihood is maximized over $\boldsymbol{\lambda}$ at each putative location x. It can be shown (APPENDIX, Part 2) that indeed robustness holds in most realistic situations, and that this robustness essentially does not depend on the true distributions h_0 and h_1.

Result 1: *Suppose a single gene lies on a chromosome having dense markers. If the assumed phenotype distributions f_0 and f_1 are of the same distributional form, with no restrictions on the values of $\boldsymbol{\lambda}$ to be considered in maximizing the likelihood, then the LOD score is robust for the gene.*

The term "distributional form" means that f_0 and f_1 are the same type of distribution, *e.g.*, both normal as in the normal single-QTL model (1). This will almost always be the case in gene mapping, whether the trait is quantitative or categorical. We emphasize that Result 1 holds *no matter what the true state of nature is* because of the maximization over $\boldsymbol{\lambda}$ at each location x. The only requirement is that f_0 and f_1 be of the same distributional form, and the researcher has utter control over this choice.

In view of the example above the intuition behind robustness may seem fairly simple: the maximization over $\boldsymbol{\lambda}$ will tend to choose estimates $\tilde{\boldsymbol{\lambda}}$ such that f_0 is similar to h_0 and f_1 is similar to h_1. Thus $K > 0$ and our overall LOD score is robust. This intuitive argument is largely correct and will suffice for those readers interested mainly in the results and implications. However, the generality of Result 1 stems from a subtle symmetry in the parameterization of f_0 and f_1, considered in greater detail in the APPENDIX, Part 2.

As a final note, we stress that the results of this section apply when one gene lies on the chromosome under study, regardless of the number of genes on the remaining chromosomes.

One gene per chromosome, nondense markers

The main result of the previous section was that the single-gene assumed model tends to be robust when the markers are dense and a single gene lies on the chromosome under study. In this section we consider a similar scenario, but with nondense markers. The main result is similar to that of the previous section, but only guarantees that the LOD score will tend to be maximized *near* the gene.

Result 2: *The markers are uniformly spaced on the chromosome under study, a single gene lies on the chromosome under study, and f_0 and f_1 are of the same distributional form. Then the LOD score is robust to within the marker interval containing the gene, or an adjacent interval.*

The proof is given in the APPENDIX, Part 3, and is similar to that of the previous section. It is clear from the proof that for most models the gene will tend to be mapped to within the correct interval, rather than an adjacent interval. However, within the correct interval, the misspecification will generally cause the estimate within the interval to be biased.

This result forms a logical basis for current practice in linkage mapping, whereby local peaks in the LOD score are attributed to the presence of genes at those locations. As long as no more than one gene lies on the chromosome under study, such a procedure will

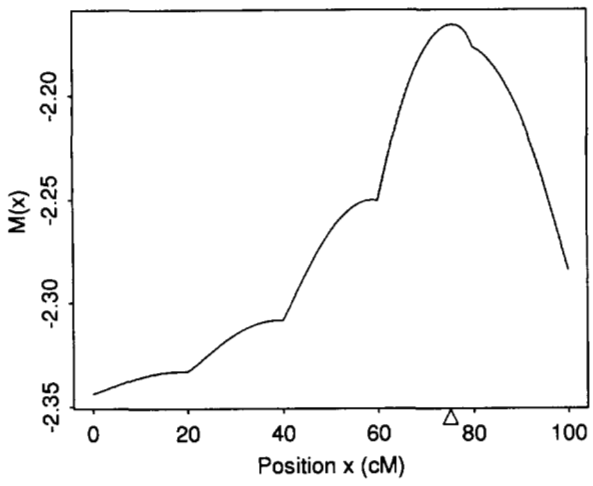

FIGURE 3.—$M(x)$ for a normal single-QTL assumed model under a two-QTL true model when only one of the true genes lies on the chromosome under study, at location 75 cM (indicated by △). Markers are present at 20-cM intervals. Despite the presence of an additional unlinked gene, $M(x)$ is maximized very near the true location.

not lead the researcher to an incorrect portion of the chromosome.

Example: One QTL on the chromosome under study, another QTL elsewhere on the genome. Here the assumed model is the single-QTL normal model, while the true model is

$$y_i = u_i + v_i + 4u_iv_i + \epsilon_i,$$

where $\epsilon \sim N(0, 1)$, u_i is the genotype for a QTL at 75 cM on the chromosome under study, and v_i is the genotype for a QTL on another chromosome. The QTLs interact, and the interaction term is intentionally strong in an attempt to heighten the degree of misspecification. The chromosome under study is of length 100 cM, with markers at intervals of 20 cM. The curve in Figure 3 plots $M(x)$ for this example. As predicted by Result 2, the function peaks in the correct interval. Indeed, it peaks at almost exactly the correct location at 75 cM.

Multiple genes per chromosome, dense markers

The results of the previous two sections suggest that single-gene assumed models have desirable robustness properties when a single gene is present on the chromosome under study.

In this section we consider the dense marker case with multiple genes per chromosome. For most of the assumed models in current use, the LOD score will still tend to be maximized at one of the gene locations. We begin with the following simple lemma. As always, the expectations are under the true model.

Lemma: On a chromosome with multiple genes, the difference in conditional phenotype means

$$D(x) = |E(y_i|g_i(x) = 1) - E(y_i|g_i(x) = 0)|$$

attains its maximum at exactly the location of one (or more) of the genes.

Proof of Lemma: We focus on a chromosome with multiple true genes. Suppose x is a location between two adjacent genes denoted x_1^* and x_2^*. Let θ_1 denote the recombination fraction between x_1^* and x, and θ_2 the recombination fraction between x and x_2^*. Define μ_{jk} as the average phenotype among individuals with genotype $= j$ at the first gene and genotype $= k$ at the second gene, or $\mu_{jk} = E(y_i|g_i(x_1^*) = j, g_i(x_2^*) = k)$. We have for a single individual

$$E(y_i|g_i(x) = 1) = E[E(y_i|g_i(x_1^*), g_i(x) = 1, g_1(x_2^*))] = \theta_1\theta_2\mu_{00} + \theta_1(1 - \theta_2)\mu_{01} + (1 - \theta_1)\theta_2\mu_{10} + (1 - \theta_1)(1 - \theta_2)\mu_{11}.$$

A similar expression may be derived for $E(y_i|g_i(x) = 0)$, giving

$$\{D(x)\}^2 = \{E(y_i|g_i(x) = 1) - E(y_i|g_i(x) = 0)\}^2 = \{(\mu_{11} - \mu_{00})(1 - \theta_1 - \theta_2) + (\mu_{10} - \mu_{01})(\theta_1 - \theta_2)\}^2.$$

The second derivative of the above expression with respect to θ_1 is positive. Therefore $D(x)$ is convex between x_1^* and x_2^* and achieves its maximum over the interval $[x_1^*, x_2^*]$ at either x_1^* or x_2^*, or perhaps at both locations. The above argument applies to any pair of adjacent genes, so that $D(x)$ must be maximized at a gene location.

Using the Lemma, the following result can be shown:

Result 3: *The markers are dense on the chromosome under study, and the assumed model is the normal single-QTL model (1) or the one-parameter exponential family single-gene model (discussed below). Then $M(x)$ is maximized at the location of one of the genes, and consequently the LOD score is robust for that gene.*

The one-parameter exponential family of models includes commonly used models for categorical traits such as Poisson and binomial distributions, including models for dichotomous traits such as disease status. The proof is given in the APPENDIX, Part 4. The proof proceeds by pointing out that for these assumed models $M(x)$ is maximized at the same location that maximizes $D(x)$, and thus robustness follows from the lemma.

Example: *The normal single-QTL assumed model:* The assumed model is the normal single-QTL model given earlier in (1). The true model is (repeating Equation 2),

$$y_i = 1 + u_i + .75v_i + \epsilon_i,$$

with $\epsilon_i \sim N(0, 1)$ and the QTL residing at 30 and 70 cM on a chromosome of length 100 cM. The thin curve in Figure 4 plots the dense-marker $M(x)$ for this example. Result 3 applies and $M(x)$ peaks at the QTL of greater effect at 30 cM. Here $M(x)$ represents the asymptotic shape of the dense-marker LOD, and may be compared with the small-sample results in Figure 1.

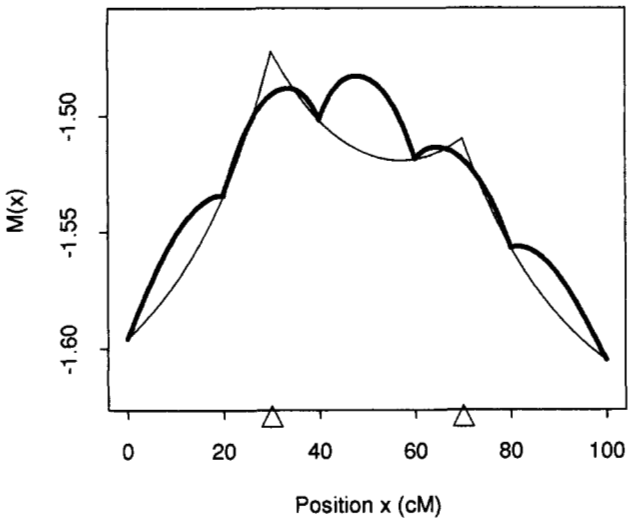

FIGURE 4.—$M(x)$ for a normal single-QTL assumed model under a two-QTL true model when both of the genes lie on the chromosome under study. This scenario was originally depicted in Figure 1. With dense markers (thin curve), $M(x)$ peaks at exactly 30 cM, the location of the QTL of stronger effect. With nondense markers at 20-cM intervals, $M(x)$ peaks at 47 cM in an incorrect interval (bold curve). Note the similarity in shape between the LODs in Figure 1 and the limiting forms depicted here.

Multiple genes per chromosome, nondense markers

We conclude the results on a cautionary note, pointing out that when the markers are not dense and multiple genes reside on a chromosome, the LOD score can be maximized in an *incorrect* interval on the chromosome under study. This was the case with the nondense map of the first example in the Introduction, in which apparent evidence was observed for a ghost gene. This bias was not merely the result of an unlucky simulation, but persists for large sample sizes, as demonstrated in Figure 4. The bold curve plots $M(x)$ for the nondense case with two QTL on the same chromosome and markers at 20-cM intervals. More detailed analytic explorations of the two-QTL true model can be found in WRIGHT and KONG (1995).

DISCUSSION

In a genome-wide search for trait genes, multiple high peaks in the LOD score on different chromosomes are often taken as evidence for separate genes at those locations. The results obtained in this paper provide considerable justification for this practice, despite the apparent contradiction in using a single-gene assumed model to map multiple genes. The extreme generality of the results is particularly useful, as more sophisticated models (*e.g.*, generalized linear models) for phenotypes become more commonly used in gene mapping. Interestingly, gene interactions (epistasis) played little role in the results.

TABLE 2

Summary of robustness results

No. genes on chromosome	Marker density	Robust?
1	Dense	Yes
1	Nondense	Yes, to within correct interval
>1	Dense	Yes, for one of the genes
>1	Nondense	No

In experimental crosses it is sometimes cost-effective to perform selective genotyping of the progeny with extreme phenotypes (LANDER and BOTSTEIN 1989), and it is important to note that the results here were derived assuming no such selection of the progeny. For the assumed models commonly used in practice, the results are summarized in Table 2.

It is important to keep in mind that this paper focuses only on the robustness of single-gene assumed models for estimating gene locations. Such a model may not be very efficient (*e.g.*, have much power to detect linkage) compared to a more realistic (*e.g.*, polygenic) assumed model. Even a plausible model, however, is unlikely to ever be the "true" state of nature, and it is of interest to understand how seriously the researcher can be led astray.

The results have been developed here for backcross populations, but an extension of some of the results to F_2 intercrosses is straightforward. For intercross data the assumed model specifies three phenotype distributions, f_0, f_1 and f_2, where as before the subscript indicates the number of alleles inherited identical by descent from the P1 population at the true gene. In general if a single gene is present on the chromosome under study, then robustness will hold, as long as f_0, f_1 and f_2 again have the same distributional form without restrictions on the parameter values. However, strict dominance assumptions can cause this condition to fail, and the robustness under these restrictions deserves further exploration.

A direct connection exists between mapping in experimental crosses and mapping using human relative pairs, where the genotype is recorded as the number of alleles shared identical-by-descent by the two relatives. For many types of relative pairs, the IBD status follows a transition pattern similar to that in experimental crosses (see Table 1 in KRUGLYAK and LANDER 1995). For quantitative trait mapping, pairs of relatives are often genotyped without prior selection based on phenotype, and for this type of design the robustness results of this paper also hold.

The authors thank the editor and referees for valuable suggestions. This research was supported in part by National Institutes of Health (NIH) grant R01-GM-46800, based on the first author's thesis at the University of Chicago Department of Statistics. Revisions supported by NIH grant 5 G12 RR-08124-04, at the BioStatistical Laboratory, University of Texas, El Paso.

LITERATURE CITED

CHURCHILL, G. A., and R. W. DOERGE, 1994 *Mapping Quantitative Trait Loci in Experimental Populations.* Cornell University Biometrics Technical Report BU-1238-M, Ithaca, NY.

CLERGET-DARPOUX, F., C. BONAITI-PELLIÉ and J. HOCHEZ, 1986 Effects of misspecifying genetic parameters in LOD score analysis. Biometrics **42:** 393–399.

COX, D. R., and D. V. HINKLEY, 1974 *Theoretical Statistics.* Chapman and Hall, London.

DAVARSI, A., A. WEINREB, V. MNIKE, J. I. WELLER and M. SOLLER, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134:** 943–951.

HALEY, C. S., and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Genetics **69:** 315–324.

HODGE, S. E., and R. C. ELSTON, 1994 Lods, wrods, and mods: the interpretation of lod scores calculated under different models. Genet. Epidemiol. **11:** 329–342.

HUBER, P. J., 1967 The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Vol. 1, 221–233.

JACOB, H. J., K. LINDPAINTNER, S. E. LINCOLN, K. KUSUMI, R. K. BUNKER, *et al.,* 1991 Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. Cell **67:** 213–224.

JANSEN, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **35:** 205–211.

KENDALL, M., A. STUART and J. K. ORD, 1987 *The Advanced Theory of Statistics,* Vol. 1. Griffin, London.

KONG, A., and F. WRIGHT, 1994 Asymptotic theory for gene mapping. Proc. Natl. Acad. Sci. USA **91:** 9705–9709.

KRUGLYAK, L., and E. S. LANDER, 1995 High-resolution genetic mapping of complex traits. American Journal of Human Genetics **56:** 1212–1223.

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

MARTINEZ, O., and R. N. CURNOW, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theoret. Appl. Genet. **85:** 480–488.

MCCULLAGH, P., and J. A. NELDER, 1989 *Generalized Linear Models,* Ed. 2. University Press, Cambridge.

OTT, J., 1991 *Analysis of Human Genetic Linkage.* Johns Hopkins University Press, Baltimore, MD.

PATERSON, A. H., E. S. LANDER, J. HEWITT, S. PETERSON, and S. D. TANKSLEY, 1988 Resolution of quantitative traits into Mendelian factors using a complete RFLP linkage map. Nature **335:** 721–726.

RISCH, N., and L. GIUFFRA, 1992 Model misspecification and multipoint linkage analysis. Human Hered. **42:** 77–92.

VIELAND, V. J., S. E. HODGE, and D. A. GREENBERG, 1992 Adequacy of single-locus approximations for linkage analyses of oligogenic traits. Genet. Epidemiol. **9:** 45–59.

WILLIAMSON, J. A., and C. I. AMOS, 1990 On the asymptotic behavior of the estimate of the recombination fraction under the null hypothesis of no linkage when the model is misspecified. Genet. Epidemiol. **7:** 309–318.

WILLIAMSON, J. A., and C. I. AMOS, 1995 Guess LOD approach: sufficient conditions for robustness. Genet. Epidemiol. **12:** 163–176.

WRIGHT, F. A., and A. KONG, 1995 *Robustness of Single-Gene Models: Backcross and Doubled Haploid Populations.* Technical Report No. 369, Department of Statistics, The University of Chicago, IL.

Communicating editor: W. J. EWENS

APPENDIX

Part 1: HUBER (1967) provided results on the convergence of maximum likelihood estimates under misspecification. The conditions required are quite weak and will apply in most realistic situations. The two most relevant assumptions are that the expected log-likelihood exists and that the expected log-likelihood varies for different parameter values. The latter is considered in WRIGHT and KONG (1995) as an identifiability condition, along with an example of its failure.

The existence of the log-likelihood is difficult to state generally in terms of the f and h distributions, a difficulty also encountered by HUBER (1967). For discrete distributions, the log-likelihood will be unbounded if an individual's phenotype is observed that is incompatible with the corresponding f distribution. Thus, it is recommended (quite sensibly) that f_0 and f_1 be parameterized to have positive mass at all conceivable phenotype values. For continuous phenotypes, difficulty will arise if f_0 and f_1 have tails that are too short (*i.e.,* go to zero too quickly) compared to the true densities h. For biological data, such cases will likely be rare. Nonetheless, some data trimming and examination of influential phenotype observations may be reasonable when performing analysis.

Part 2: The argument of the proof is outlined below, followed by the (much shorter) proof. The researcher in the example was fortunate enough to fix a parameter value $\boldsymbol{\lambda}_1$ that gives robust estimates of the gene location. We will say that $\boldsymbol{\lambda}_1$ is robust because of this property. Now suppose that another researcher had chosen to map the same gene by fixing the assumed distributions as $f_0 = N(1, 1)$, $f_1 = N(0, 1)$, or equivalently had fixed a single parameter value $\boldsymbol{\lambda}_2 = (a, b, \sigma^2) = (1, -1, 1)$. Note the second researcher chose exactly the *opposite* assumed distributions as the first researcher. We will say that parameters $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are "complementary" because they specify an exact exchange of the assumed phenotype distributions. If K_1 (the K value corresponding to $\boldsymbol{\lambda}_1$) is greater than zero, then K_2 (corresponding to $\boldsymbol{\lambda}_2$) must be less than zero, because the switching of the assumed phenotype distributions (for any true distributions h_0 and h_1) implies $K_1 = -K_2$. In other words, $\boldsymbol{\lambda}_2$ is not robust at all, and in fact has its minimum expected log-likelihood at the true gene location. Manipulation of (3) shows that if $\boldsymbol{\lambda}_1$ is robust then $E[\log L_i(\boldsymbol{\lambda}_1)] > E[\log L_i(\boldsymbol{\lambda}_2)]$ for all locations on the chromosome.

We now have the necessary components to complete the argument, in the more typical scenario where the likelihood is maximized over $\boldsymbol{\lambda}$ at each location x. We have the following:

1. Each parameter value $\boldsymbol{\lambda}_1$ has a complementary $\boldsymbol{\lambda}_2$. This is why we required that f_0 and f_1 be of the same distributional form without restriction on the possible choice of $\boldsymbol{\lambda}$, so that an exchange of the two distributions is permitted. This property is stated in WRIGHT and KONG (1995) as a *symmetry* condition for f_0 and f_1.

2. If $K_2 < 0$, then $K_1 > 0$. Thus if a parameter is nonrobust, its complementary parameter is robust, and the robust parameter gives a higher expected log-likelihood over the entire chromosome.

These facts ensure that only robust $\boldsymbol{\lambda}$'s need be considered, because these dominate the others in expected log-likelihood. For a robust $\boldsymbol{\lambda}$, the function $E[\log L_i(\boldsymbol{\lambda})]$ is maximized at x^*, and it is simple to show that, as a consequence, the global maximum over all x and $\boldsymbol{\lambda}$ occurs at location x^*.

Proof: Fix $x \neq x^*$. For any $\boldsymbol{\lambda}_1$ there is a complementary parameter value $\boldsymbol{\lambda}_2$, and assume without loss of generality that $E[\log L_i(\boldsymbol{\lambda}_1)] > E[\log L_i(\boldsymbol{\lambda}_2)]$ at location x. Rewriting $E[\log L_i(\boldsymbol{\lambda}_1)] - E[\log L_i(\boldsymbol{\lambda}_2)] > 0$ gives, after some manipulation of (3),

$$(\tfrac{1}{2} - \theta)K_1 > 0,$$

where θ is the recombination fraction between x and x^*, and thus $K_1 > 0$. This further implies $E[\log L_i(\boldsymbol{\lambda}_1)]$ is maximized at x^*. $\boldsymbol{\lambda}_1$ was chosen arbitrarily, so $M(x^*) > M(x)$.

Part 3: The markers are equally spaced across the chromosome, with adjacent markers lying map distance δ apart. For a single individual (the subscript i is suppressed) and at location x (not in the interval containing x^*), define $p_k^\delta(x) = P(g(x) = 1 \mid W = k; \delta)$, where W indexes the joint state of the two markers flanking location x as follows: $W = 1$, if the left and right marker genotypes are, respectively, 0 and 0; $W = 2$, if the left and right marker genotypes are 0, 1; $W = 3$, if the left and right marker genotypes are 1, 0; $W = 4$, if the left and right marker genotypes are 1, 1. Let x_A be the marker lying nearest the true gene location x^*. Let θ be the recombination fraction between x^* and x_A, and let θ_δ be the common recombination fraction between adjacent markers. At location x the expected log-likelihood can be written

$$E[\log L_i(\boldsymbol{\lambda})] = \tfrac{1}{2}\{E_{h_0}[X_1 + X_2] + E_{h_1}[X_3 + X_4] - \theta K\},$$

where

$$K = \{E_{h_0}[X_1 + X_2 - X_3 - X_4] - E_{h_1}[X_1 + X_2 - X_3 - X_4]\},$$

and where

$$X_1 = \log(f_0(y; \boldsymbol{\lambda})(1 - p_1^\delta(x)) + f_1(y; \boldsymbol{\lambda})p_1^\delta(x))(1 - \theta_\delta),$$

$$X_2 = \log(f_0(y; \boldsymbol{\lambda})(1 - p_2^\delta(x)) + f_1(y; \boldsymbol{\lambda})p_2^\delta(x))\theta_\delta,$$

$$X_3 = \log(f_0(y; \boldsymbol{\lambda})(1 - p_3^\delta(x)) + f_1(y; \boldsymbol{\lambda})p_3^\delta(x))\theta_\delta,$$

$$X_4 = \log(f_0(y; \boldsymbol{\lambda})(1 - p_4^\delta(x)) + f_1(y; \boldsymbol{\lambda})p_4^\delta(x))(1 - \theta_\delta).$$

Note the similarity in form between the expression for $E[\log L_i(\boldsymbol{\lambda})]$ above and the expression used in the proof of Result 1. Here, however, we have a term K that depends on both the location x and the parameter $\boldsymbol{\lambda}$. In general, $M(x)$ is not monotonically decreasing away from x^*. However, we may compare "analogous" locations, that is, different locations that have the same position relative to their flanking markers. Applying essentially the same argument as in Result 1 (covered in more detail in WRIGHT and KONG 1995), one can show that $M(x_B) > M(x_C)$, if x_B and x_C are analogous locations and x_C is farther from the gene. The global maximum in this case cannot then occur more than one marker interval away from x^*.

Careful examination of $E[\log L_i(\boldsymbol{\lambda})]$ in the interval containing x^* is much more difficult and the lengthy details are not reported here. These results can be roughly summarized as follows: if there are $\boldsymbol{\lambda}$ values such that *both* $E_{h_0}[\log(f_0(y; \boldsymbol{\lambda})/f_1(y; \boldsymbol{\lambda}))] > 0$ and $E_{h_1}[\log(f_1(y; \boldsymbol{\lambda})/f_0(y; \boldsymbol{\lambda}))] > 0$, then $M(x)$ will tend to be maximized in the correct interval. This condition will, of course hold unless one or both of h_0 and h_1 are severely misspecified.

Part 4: The normal single-QTL assumed model: The assumed model is as given in (1). At a locus x, let σ_x^2 be the limiting value of the maximum likelihood estimate for σ^2. It can be shown that

$$\sigma_x^2 = \tfrac{1}{2}[\mathrm{Var}(y_i \mid g_i(x) = 0) + \mathrm{Var}(y_i \mid g_i(x) = 1)],$$

and that $M(x)$ is maximized where this value is minimized. A simple variance decomposition gives

$$\mathrm{Var}(y_i) = \sigma_x^2 + \tfrac{1}{4}\{E(y_i \mid g_i(x) = 1) - E(y_i \mid g_i(x) = 0)\}^2 = \sigma_x^2 + \tfrac{1}{4}\{D(x)\}^2$$

and because $\mathrm{Var}(y_i)$ is constant the result follows from the Lemma.

The one-parameter exponential family fitted model: We consider here the case where the assumed model has $\boldsymbol{\lambda} = [\psi_0, \psi_1]$ and the distributions can be written

$$f_0(y; \boldsymbol{\lambda}) = \exp\left(\frac{y\psi_0 - b(\psi_0)}{a} + c(y)\right),$$

$$f_1(y; \boldsymbol{\lambda}) = \exp\left(\frac{y\psi_1 - b(\psi_1)}{a} + c(y)\right),$$

for known constant $a > 0$ and known functions $b(\psi)$ and $c(y)$. Many commonly used distributions are members of the exponential family (COX and HINKLEY 1974; MCCULLAGH and NELDER, 1989), and may be continuous or discrete. Examples include Poisson and Binomial models (and hence binary traits such as disease), as well as Gamma models with known dispersion parameter.

Let J = the true number of trait genes on the chromosome under study. Define x_j^* = ordered map location of the jth gene, $j = \{1, \ldots, J\}$. With respect to

the chromosome under study, the true conditional phenotype distributions then depend on the *joint* genotype of the J genes $\{g_i(x_1^*), \ldots, g_i(x_J^*)\}$. For a backcross individual there are 2^J such possible joint genotypes.

At a locus x, an examination of the derivatives of $E[\log L_i(\boldsymbol{\lambda})]$ with respect to ψ_0 and ψ_1 reveals that for fixed x the limiting nuisance parameter estimates ψ_{0x} and ψ_{1x} satisfy

$$b'(\psi_{0x}) = E(y_i|g_i(x) = 0), \quad b'(\psi_{1x}) = E(y_i|g_i(x) = 1),$$

giving

$$M(x) = \frac{1}{2a}\{\psi_{0x}b'(\psi_{0x}) - b(\psi_{0x}) + \psi_{1x}b'(\psi_{1x}) - b(\psi_{1x})\} + E(c(y)).$$

Define $e_0(x) = E(y_i|g_i(x) = 0)$ and $e_1(x) = E(y_i|g_i(x) = 1)$. Using the fact that $e_0(x) + e_1(x) = 2E(y_i)$ implies that

$$\frac{\partial}{\partial e_0(x)} s(x) = -\frac{\partial}{\partial e_1(x)} s(x)$$

for a smooth function $s(x)$. Then

$$\frac{\partial}{\partial e_0(x)} M(x) = \frac{1}{2a}\left\{\frac{\partial}{\partial e_0(x)}[\psi_{0x}b'(\psi_{0x}) - b(\psi_{0x})] - \frac{\partial}{\partial e_1(x)}[\psi_{1x}b'(\psi_{1x}) - b(\psi_{1x})]\right\},$$

and

$$\frac{\partial}{\partial e_0(x)}[\psi_{0x}b'(\psi_{0x}) - b(\psi_{0x})] = \frac{\partial}{\partial \psi_{0x}}[\psi_{0x}b'(\psi_{0x}) - b(\psi_{0x})]\frac{\partial}{\partial e_0(x)}\psi_{0x}$$

$$= [\psi_{0x}b''(\psi_{0x}) + b'(\psi_{0x}) - b'(\psi_{0x})]\frac{\partial}{\partial e_0(x)} b'^{-1}(e_0(x))$$

$$= [\psi_{0x}b''(\psi_{0x})]\frac{1}{b''(\psi_{0x})} = \psi_{0x}.$$

So, finally

$$\frac{\partial}{\partial e_0(x)} M(x) = \frac{1}{2a}\{\psi_{0x} - \psi_{1x}\}.$$

A property of the exponential family implies that $e_0(x)$ is strictly increasing in ψ_{0x} (KENDALL *et al.* 1987), and similarly with $e_1(x)$ and ψ_{1x}. Then if, for example, $e_0(x) < e_1(x)$, it follows that $\psi_{0x} < \psi_{1x}$ and $M(x)$ decreases as $e_0(x)$ increases. Thus $M(x)$ is maximized at the location where $D(x) = (e_1(x) - e_0(x))^2$ is maximized.