A Nonparametric Approach for Mapping Quantitative Trait Loci

Leonid Kruglyak* and Eric S. Lander*,†

**Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142 and †Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*

Manuscript received August 9, 1994
Accepted for publication November 14, 1994

ABSTRACT

Genetic mapping of quantitative trait loci (QTLs) is performed typically by using a parametric approach, based on the assumption that the phenotype follows a normal distribution. Many traits of interest, however, are not normally distributed. In this paper, we present a nonparametric approach to QTL mapping applicable to any phenotypic distribution. The method is based on a statistic Z_W, which generalizes the nonparametric Wilcoxon rank-sum test to the situation of whole-genome search by interval mapping. We determine the appropriate significance level for the statistic Z_W, by showing that its asymptotic null distribution follows an Ornstein-Uhlenbeck process. These results provide a robust, distribution-free method for mapping QTLs.

MAPPING genetic loci responsible for quantitative traits (quantitative trait loci or QTLs) in plants and animals is an important problem with a broad range of applications. Physiological traits involved in human disease can be studied through the genetic dissection of quantitative modifier genes in experimental models such as mouse and rat. Such studies have been carried out for hypertension (HILBERT *et al.* 1991; JACOB *et al.* 1991), type I diabetes (TODD *et al.* 1991), epilepsy (RISE *et al.* 1991) and colon cancer (DIETRICH *et al.* 1993). Similarly, genetic mapping of QTLs influencing agriculturally important traits can lead to a better understanding of such traits as well as to novel improvement programs. Recent examples include mapping QTLs for fruit mass in tomato (PATERSON *et al.* 1988), grain yield in maize (STUBER *et al.* 1992) and fatness and growth in pigs (ANDERSSON *et al.* 1994). Although the notion of QTL mapping dates back to early in the century (SAX 1923), there has been a recent explosion of interest fueled by the development of genetic linkage maps based on DNA polymorphisms for a number of organisms including the mouse (DIETRICH *et al.* 1992, 1994), zebrafish (POSTLETHWAIT *et al.* 1994), pig (ROHRER *et al.* 1994), cattle (BISHOP *et al.* 1994), rice and maize (AHN and TANKSLEY 1993) and tomato and potato (TANKSLEY *et al.* 1992). Improvements in existing maps and development of new ones will continue to present new opportunities for QTL mapping.

Initially, QTL mapping was carried out by looking for associations between genotypes at individual markers and phenotypic traits of interest (SAX 1923; SOLLER and BRODY 1976). LANDER and BOTSTEIN (1989) introduced the somewhat more powerful approach of interval mapping, in which the presence of a QTL can be tested at every location in a genome by exploiting the full power of a complete genetic linkage map. Because genome-wide searches involve testing many hypotheses concerning the possible location of QTLs, a key issue is the proper statistical threshold to correct for such multiple testing. LANDER and BOTSTEIN (1989) derived the appropriate threshold to keep the false positive rate low, by relating the QTL statistic (the LOD score) to a known random process (the Ornstein-Uhlenbeck diffusion). The basic approach of interval mapping has been further generalized by a number of authors (*e.g.*, LUO and KEARSEY 1992; JANSEN 1993; MORENO-GONZALEZ 1993; RODOLPHE and LEFORT 1993; ZENG 1993; FULKER and CARDON 1994; HALEY *et al.* 1994; JANSEN and STAM 1994).

All these QTL mapping methods share a common assumption: that the phenotype follows a normal distribution with equal variance in both parental strains. Under this assumption, the presence of a QTL can be tested by a simple parametric test (a t-test in the case of a single marker test, the related LOD score in the case of interval mapping).

Many phenotypes of interest, however, are not normally distributed. Examples include counts generated by a Poisson process [such as number of tumors, which in many cases follows a negative binomial distribution (DRINKWATER and KLOTZ 1981)], truncated data (such as survival times in an experiment of limited duration), probabilities (such as chance of an epileptic seizure in a given trial), and qualitative data (such as severity grades assigned upon histological examination). Traditional QTL mapping methods cannot be directly applied in such cases. One approach is to attempt to find a mathematical transformation that will convert the trait into an approximately normal distribution with

Corresponding author: Eric S. Lander, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142.

equal variance in both parental strains (WRIGHT 1968). The problem is that this approach may not work: no appropriate transformation may be found, and, even if one is, the effect of outliers may still be too great.

An alternative approach is to apply nonparametric (or distribution free) methods to QTL mapping. In the case of a single marker, this is straightforward. Consider, for example, an (A × B) × A backcross. For any given marker, the backcross progeny can be classified according to genotype as either A (parental) or H (hybrid). One then can perform a nonparametric Wilcoxon rank-sum rank test (see, *e.g.*, LINDGREN 1968) to decide whether the distribution of the phenotype differs between the two groups. For tumor counts that follow a negative binomial distribution, this test was found in many cases to be more powerful than the *t*-test (DRINKWATER and KLOTZ 1981). Appropriate significance thresholds for the Wilcoxon rank-sum test may be found in (LINDGREN 1968) and in other standard statistics texts.

To extend this approach to a genome-wide QTL search, two key issues must be addressed: how to generalize the Wilcoxon rank-sum statistic to the region between markers and how to determine the appropriate significance threshold for genome-wide search that maintains a low false positive rate. In this paper, we solve these problems by defining a version of the Wilcoxon rank-sum statistic appropriate for interval mapping and showing that the statistic is asymptotically distributed as an Ornstein-Uhlenbeck diffusion in the case that no QTL is present, allowing the appropriate significance threshold to be derived from the extreme value properties of this process. These results provide the analog of the work of LANDER and BOTSTEIN (1989) for nonparametric QTL mapping.

RESULTS

Notation: We introduce the following notation. Consider a cross in which n progeny (labeled by i, $i = 1, \ldots, n$) are phenotyped for a quantitative trait and are genotyped at m genetic markers (labeled by j, $j = 1, \ldots, m$) across the genome. Let ϕ_i denote the phenotype of progeny i and let s_j denote the position of marker j in the genome. The function $g_i(s)$ will denote the genotype of progeny i at any location s in the genome, so that $g_i(s_j)$ denotes the genotype of progeny i at marker j. The complete genotype data is denoted by $DATA = [g_i(s_j)]_{i,j}$. Let $E[X|DATA]$ denote the expected value of a quantity X given the genotype data and let $\langle X \rangle$ denote the expected value of X in the absence of genotype data, *i.e.*, over all possible sets of genotypes.

Definition of nonparametric statistic for QTL mapping: We will initially consider an (A × B) × A backcross; a generalization to other crosses will be discussed below. The genotypes of the progeny are either A (parental) or H (hybrid). We define the function $x_i(s)$ to be $+1$ or -1 according to whether the genotype $g_i(s)$ is A or H.

For any location s in the genome, we define a nonparametric QTL statistic $Z_W(s)$. We first define an auxiliary statistic $Y_W(s)$ by

$$Y_W(s) = \sum_{i=1}^{n} [n + 1 - 2 \cdot rank(i)] E[x_i(s) | DATA], \quad (1)$$

where $rank(i)$ denotes the rank by phenotype of progeny i. At the location s_j of a marker, the value of $x_i(s_j)$ is known with certainty. For other locations s, the value of $x_i(s_j)$ is not directly observed, but its expectation $E[x_i(s_j) | DATA]$ can be easily computed based on the genotypes observed at the closest flanking markers (see APPENDIX A for an explicit formula). The statistic $Z_W(s)$ then is defined by

$$Z_W(s) = Y_W(s) / \sqrt{\langle Y_W(s)^2 \rangle}, \quad (2)$$

that is, by dividing Y_W by its standard deviation (see APPENDIX A for an explicit formula in terms of the recombination frequencies between s and the nearest flanking markers).

When s coincides with the location of a genetic marker, $Z_W(s)$ is easily seen to be equivalent (up to rescaling) to the Wilcoxon rank-sum test. More generally, for *any* location s in the genome unlinked to a QTL, $Z_W(s)$ is asymptotically distributed (for large n) as a standard normal variable with mean 0 and variance 1. (See APPENDIX A for details.) Thus, the significance level of $Z_W(s)$ at any single point s can be evaluated by a *t*-test.

If the phenotypes are discrete rather than continuous (*e.g.*, counts), tied phenotypic values may occur in the data set. Two methods of handling ties are available (KENDALL and STUART 1979). The first is to rank tied individuals at random. This approach has the benefit of simplicity, because no new theory is necessary and all the results above apply directly. It does ignore some information contained in the data. The second approach is to assign to each tied individual the average rank of those tied. This approach is somewhat more efficient, but the gain is slight (KENDALL and STUART 1979). It has the drawback that the variance of the test statistic now depends on the number and type of ties observed. We chose to use the first approach in view of its greater simplicity and only a slight loss of efficiency.

Appropriate threshold for genome-wide search: When searching an entire genome for QTLs, one cannot use the threshold appropriate for testing significance at a single point. Because many independently segregating markers are examined across the genome, $Z_W(s)$ is likely to show substantial deviation from 0 *some-*

where in the genome just by chance. One should select a threshold T such that the probability (under the null hypothesis) that $|Z_W(s)|$ exceeds T anywhere in the genome equals the desired false positive rate α.

For QTL mapping of normally distributed traits, LANDER and BOTSTEIN (1989) considered three cases. They are as follows: sparse-map case, in which consecutive markers are separated far enough to be considered independent; dense-map case, in which the spacing between consecutive markers approaches zero; and intermediate-map case, in which the spacing between consecutive markers is moderate and is not well approximated by either of the first two cases.

The same approach may be used when considering appropriate thresholds for $|Z_W(s)|$. In the sparse-map case, one studies only markers and treats each marker as independent; the value of $|Z_W|$ at each marker is assumed to be uncorrelated with the value at adjacent markers. The appropriate threshold is obtained from that for a single marker by a simple Bonferroni correction for multiple testing. If we test M markers, we set the nominal significance level for each test at α / M and choose the single t-test threshold for this significance level. Note that since the values at adjacent markers *are* positively correlated, this threshold is a conservative choice; the actual false positive rate is guaranteed to be less than α. In fact, the sparse-map approximation becomes too conservative as the marker spacing decreases; the threshold approaches infinity as the marker spacing approaches 0.

In the dense-map case, one assumes that the genotype is known at every point in the genome, *i.e.*, that there are markers everywhere. This is the limiting case for maps of increasing density, and the threshold for this case is once again a conservative estimate, but one appropriate for denser maps. In the limiting case of an infinitely dense-map and large sample size, one can prove that the statistic Z_W follows an Ornstein-Uhlenbeck diffusion process (see APPENDIX B for details). The appropriate threshold then may be derived from the extreme value properties of this process. Specifically, we have the following result (which is analogous to Proposition 2 in LANDER and BOTSTEIN (1989) concerning the dense-map limit of the LOD score for parametric QTL mapping).

Proposition 1: For an organism with C chromosomes and genetic length G (measured in Morgans), the probability that $|Z_W|$ exceeds a high threshold T somewhere in the genome is $\alpha \approx (C + 2\rho GT^2) P[|Z| > T]$, where $\rho = 1$ is the rate of crossovers per Morgan and $P[|Z| > T] = (2/\sqrt{2\pi}) \int_T^\infty e^{-t^2/2} dt$ is the probability that a variable Z, distributed as a standard normal, exceeds T (this probability is simply the two-sided tail probability of the standard normal).

In short, the probability that $|Z_W|$ exceeds T *somewhere* in the genome is larger by a factor of $(C + 2GT^2)$ than the corresponding probability at any single point. For a given false positive rate α, this equation can be numerically solved for the appropriate threshold T. Thresholds for several genome sizes are plotted in Figure 1. For the mouse genome of 20 chromosomes and 1600 cM, the significance level $\alpha = 0.05$ corresponds to a threshold of 3.9. (This is equivalent to a LOD score of 3.3.)

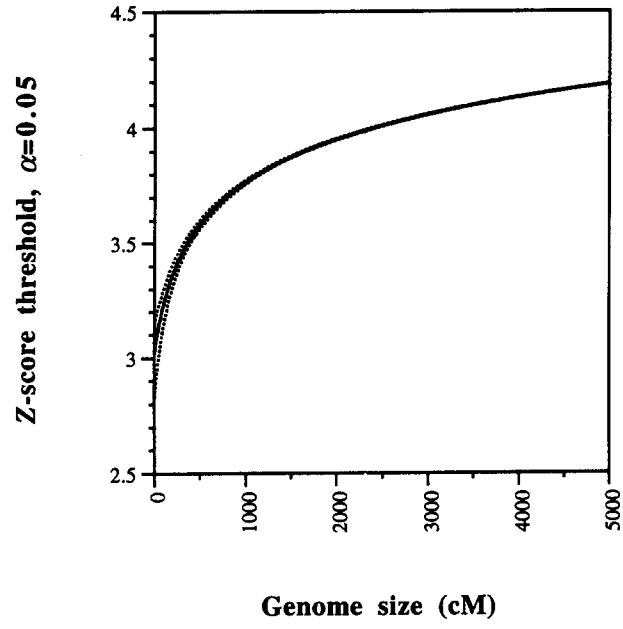

FIGURE 1.—Z-score thresholds as a function of genome size for a false positive rate of 0.05. The solid line is for a genome with 20 chromosomes; the dotted lines are for genomes with 10 and 30 chromosomes. It is clear that the number of chromosomes has very little effect on the threshold except for very small genomes.

In the intermediate-map case, one assumes that markers are spaced uniformly with intermarker distance Δ Morgans. An approximation for the intermediate-map case is available. The appropriate threshold T can be found from $\alpha \approx [C + 2\rho GT^2 v(2T\sqrt{\Delta})] P[|Z| > T]$, where $v(x)$ is a special function that can be approximated by $e^{-0.583}x$ for small x (SIEGMUND 1985; FEINGOLD *et al.* 1993). Note that as Δ decreases, v approaches 1 and the threshold reduces to the dense-map case. As Δ increases, v is better approximated by $2/x^2$, and the threshold reduces to the multiple testing sparse-map case. This indicates that the sparse-map approximation is "correct" in the sense that it interpolates correctly between the sparse-map and dense-map cases (D. SIEGMUND, personal communication).

Which threshold should be used in practice? We would recommend strongly that the dense-map threshold always be used (for both parametric and nonparametric QTL mapping) regardless of the actual density of the map used. Even if a particular study employs a

relatively sparse map, it must be anticipated that similar studies on the same trait will be carried out by other researchers. Because of the selection bias that only positive results tend to be reported, the scientific literature must regard all proposed linkages as if they were obtained by scanning the entire genome. Otherwise, published papers will contain an excess of false positive linkages. This practice is employed in human genetics, where all reported linkages must meet the LOD score threshold of 3 appropriate for whole-genome search.

Note also that all significance thresholds in this paper are computed for a two-sided test, in which large deviations of Z_W in *either* direction from 0 are considered significant. For a one-sided test, in which one searches only for positive values of Z_W, the appropriate thresholds may be obtained by dividing all estimates of false positive rates by 2. Such a test could, in principle, be used if the direction of a QTL's effect is known *a priori*. However, the fact that one parental strain shows larger phenotypic values does not guarantee that *all* QTLs segregating in that cross increase the phenotypic value in that strain; some may decrease it. Because both kinds of QTLs are of interest, we would recommend that the two-sided threshold be used in all cases.

An application of nonparametric QTL interval mapping: To illustrate the use of nonparametric QTL mapping, we applied the method to data from a recently reported study of quantitative modifiers of intestinal neoplasias in mice. Mice carrying the *Min* mutation in the *Apc* gene on chromosome *18* develop intestinal tumors, but the tumor number is strongly influenced by genetic background (MOSER *et al.* 1992). By studying the distribution of tumor number in (AKR × B6-Min) × B6 backcross progeny, DIETRICH *et al.* (1993) found evidence for a major QTL on chromosome *4*. The QTL locus, named *Mom-1* (for Mo difier of M in-1), was subsequently confirmed in two additional crosses.

In the original AKR backcross, DIETRICH *et al.* (1993) genotyped 110 progeny for 75 markers throughout the mouse genome. Because tumor number was not normally distributed with constant variance, they applied a square-root transformation to obtain a phenotype with a more nearly normal distribution. Applying traditional parametric QTL mapping to this transformed phenotype, they found a LOD score of 4.7 for the *Mom-1* locus on chromosome *4*. No other significant LOD scores were found.

We reanalyzed the same data set using the nonparametric statistic Z_W. The *Mom-1* region on chromosome *4* shows a peak of $Z_W = 4.33$, which is well above the threshold of 3.9 appropriate for the mouse genome (Figure 2). The dense-map significance level of this score is $P < 0.01$. (*N.B.*, as noted above, we apply the dense-map threshold and significance level even though the marker spacing is ~20 cM). This confirms

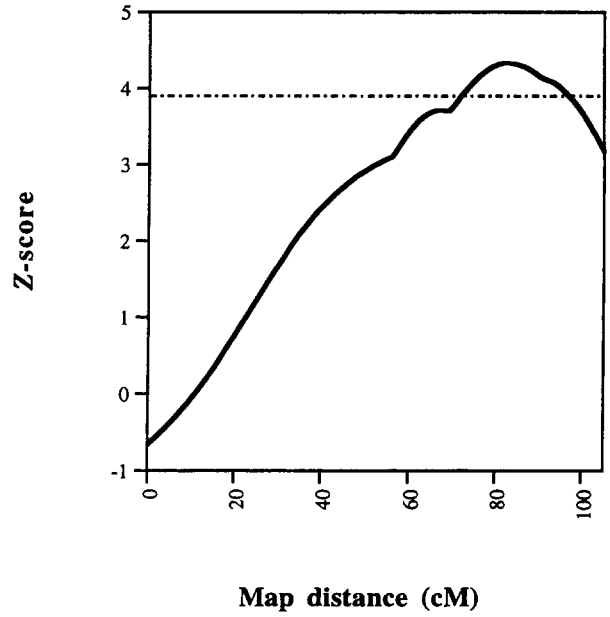

FIGURE 2.—Z-scores on chromosome 4 in the *Mom-1* cross. The dotted line is at $Z = 3.9$, the $\alpha = 0.05$ threshold for the mouse genome. Map distance is from the first marker. Note that the Z-score exceeds the threshold at the location of the *Mom-1* locus.

that the detection of the *Mom-1* locus is not sensitive to the assumptions of normality made by DIETRICH *et al.* (1993).

The importance of using the correct threshold is illustrated by Figure 3, which shows the Z_W-scores for chromosome *2*. The peak score approaches 3.0, which would be highly significant in a single marker test ($P < 0.003$; two-sided test). However, the chance that the absolute value of the Z_W-score exceeds 3.0 *somewhere* in the genome is estimated by Proposition 1 to be ~80%. In fact, because in this case the false positive rate is high, a slightly better approximation shows the chance to be ~60% (see APPENDIX B). No other chromosomes show high Z_W scores.

Generalization to other crosses: Above, we have considered an (A × B) × A backcross. We briefly comment on how to generalize these results to other crosses, such as an (A × B) F2 intercross. To do so, one can redefine the function $x_i(s)$ by $x_i(s) = f[g_i(s)]$, where f is any function that maps genotypes to real numbers or, more generally, to any real normed vector space. We assume that f is chosen so that $\langle x_i(s) \rangle = 0$, where the average is taken over the distribution of possible genotypes. If f takes vector values, every component must satisfy this condition. The statistic $Z_W(s)$ can again be defined by Equations *1* and *2*, with the proviso that one uses the magnitude of $Z_W(s)$ if f takes values in a vector space. The relevant expected values and variances may be computed using the following results:

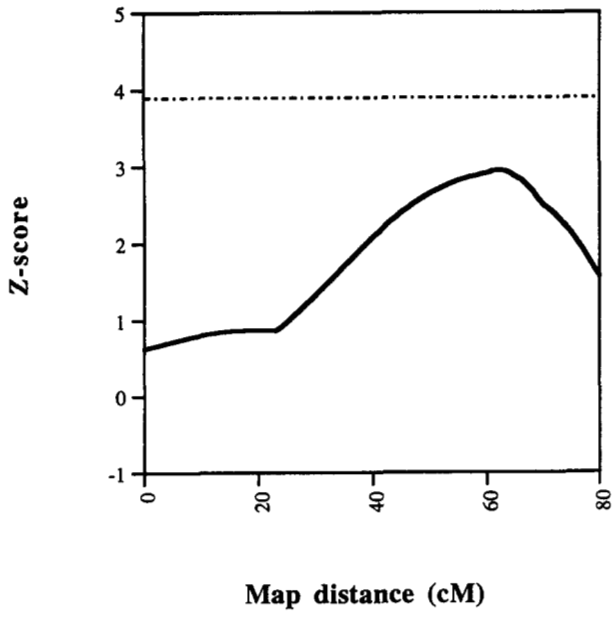

FIGURE 3.—*Z*-scores on chromosome *2* in the *Mom-1* cross. The dotted line is at $Z = 3.9$, the $\alpha = 0.05$ threshold for the mouse genome. Map distance is from the first marker. Chromosome *2* does not contain a QTL affecting tumor number, and the *Z*-score is below threshold. The peak value of Z is nearly 3.0, well in excess of the single-test threshold of 2.0.

$$E[x(s) \mid DATA] = \sum_{g(s) \in G} P[g(s) \mid g_L, g_R] f[g(s)],$$

$$\langle E[x(s) \mid DATA] \rangle = \sum_{g_L, g_R, g(s) \in G} P[g_L, g_R, g(s)] f[g(s)],$$

$$\langle E[x(s) \mid DATA]^2 \rangle = \sum_{g_L, g_R \in G} P(g_L, g_R) \times \left(\sum_{g(s) \in G} P[g(s) \mid g_L, g_R) f[g(s)] \right)^2,$$

where g_L and g_R are the genotypes at the left and right markers flanking s.

In an F2 intercross, there are three possible genotypes: homozygous A, homozygous B and heterozygous H. One can test different hypotheses about the presence of a QTL by using different functions f. To test for a QTL with an additive effect, an appropriate choice is $f(\mathrm{A}) = 1, f(\mathrm{H}) = 0, f(\mathrm{B}) = -1$ (we call the resulting statistic Z_A for future reference). To test whether progeny with genotype A differ from the other progeny (testing for the presence of a QTL showing recessive or dominant inheritance), one can lump together progeny with genotypes B and H by using the function $f(\mathrm{A}) = 3, f(\mathrm{B}) = f(\mathrm{H}) = -1$ (call this statistic Z_R). Alternatively, one can test for dominance (progeny with genotype H differ from the average of those with genotypes A and B) by using $f(\mathrm{H}) = 1, f(\mathrm{A}) = f(\mathrm{B}) = -1$ (call this statistic Z_D). As above, the resulting statistics will be asymptotically normal. Proposition 1 applies, with the only difference being the value of ρ. It can be shown that $\rho = 1$ for Z_A, $\rho = 4/3$ for Z_R and $\rho = 2$ for Z_D. The resulting dense-map thresholds for the mouse genome are 3.9, 4.0 and 4.1, respectively (see APPENDIX B for details; see also DUPUIS 1994).

When the existence of a dominance component is suspected but its magnitude is unknown, it may be desirable to test for the presence of a QTL with both additive and dominance components. We define a two-component statistic (Z_A, Z_D), using Z_A and Z_D from the preceding paragraph, and the joint statistic $X_W = \sqrt{Z_A^2 + Z_D^2}$. Apart from the square root, the resulting statistic corresponds to a generalization of the Wilcoxon rank-sum test to three subgroups (Section 31.71 of KENDALL and STUART 1979). X_W^2 is asymptotically distributed as chi squared with two degrees of freedom. Hence the false positive rate for X_W may be determined from the probability that a rough chi-squared process exceeds a high threshold. The relevant result is found in (ALDOUS 1989, Section I19) and gives $\alpha \approx (C + 2\rho G T^2) P[X > T]$, where $P[X > T] = e^{-T^2/2}$ is the probability that a variable X, the square of which is distributed as chi square with two degrees of freedom, exceeds T. (cf. APPENDIX A of FEINGOLD *et al.* 1993 for the one-sided test and DUPUIS 1994 for another derivation of this test). Here $\rho = 1.5$ is the average of the values for the additive and dominance components. This result is completely analogous to Proposition 1. The corresponding thresholds may be determined as before, and are plotted in Figure 4; the threshold for the mouse genome is 4.44.

A number of related approaches exist for testing whether three (or more) genotypic classes differ in distribution of phenotypes. Among these are the Kruskal-Wallis test and the Jonckheere-Terpstra test; see KENDALL and STUART (1979) for details.

DISCUSSION

QTL mapping is an important approach for dissecting the genetic factors affecting traits of physiological and agronomic importance. Rapid progress in the development of DNA-based genetic linkage maps has made QTL mapping a practical and widely used approach. Parametric interval mapping allows efficient detection and localization of QTLs for normally distributed traits, while keeping the false positive rate low through the use of appropriate thresholds (LANDER and BOTSTEIN 1989). However, many traits of interest are not normally distributed, and the use of standard QTL interval mapping for such traits may lead to low power or unacceptably high false positive rates.

The results described in this paper extend the power of interval mapping to any quantitative trait regardless of its distribution through the use of nonparametric

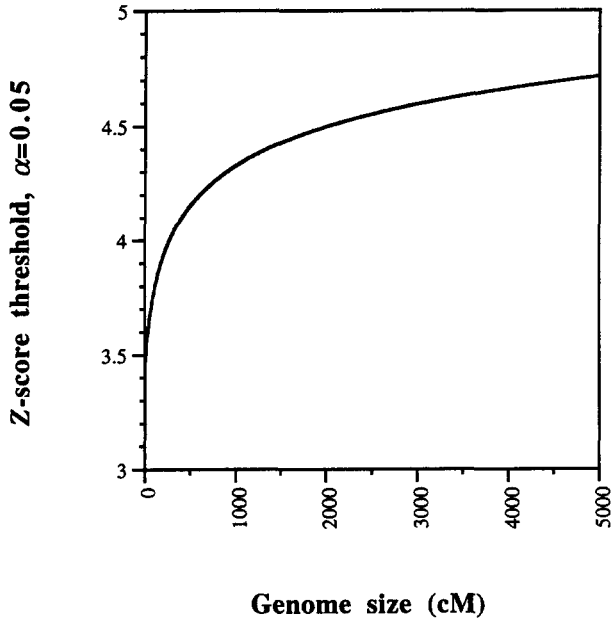

FIGURE 4.—Z-score thresholds as a function of genome size for an F2 intercross when both additive and dominance components are searched for. The false positive rate is set to 0.05, and the genome contains 20 chromosomes.

methods. The basic statistic Z_W is a generalization of the Wilcoxon rank-sum statistic to the situation of interval mapping. The Wilcoxon test has been extensively studied. It is known that the efficiency of this test relative to the t-test is 96% for shift alternatives if the distribution is normal and is never less than 86% for any distribution (LINDGREN 1968). The efficiency is defined as the ratio of the sample sizes required for the two tests to achieve the same (asymptotic) power; thus, the Wilcoxon test requires 1.04 times the sample of the t-test if the distribution is normal and at most 1.16 times the sample of the t-test for any distribution. The loss of efficiency in the case of normally distributed trait is thus slight and is offset by the robustness of the nonparametric method. Moreover, the nonparametric test can be much more powerful in certain cases. For an exponential distribution, the t-test requires three times the sample of the Wilcoxon test to achieve significance (KENDALL and STUART 1979).

Some other differences between the parametric and the nonparametric approaches should be mentioned. The parametric method provides a direct estimate of the phenotypic effect of the QTL, whereas the nonparametric method simply tests for the presence of a QTL. Also, the statistic for the parametric method, the LOD score, is proportional to the square of a normal variable, whereas the statistic Z_W for the nonparametric method is a standard normal variable. To convert Z_W to an "equivalent" LOD score LOD_W, one could use the formula $LOD_W = {}^1/_2(\log_{10}e)(Z_W)^2$. In general, we would recommend that both parametric and nonparametric QTL mapping should be used, especially when there is evidence of nonnormality. If the results differ between the two approaches, the experiment should be interpreted with considerable caution.

The nonparametric approach described here has been incorporated in the QTL mapping package MAPMAKER/QTL (version 2), available from the authors. With this modification, the package will allow robust mapping of QTLs without concern about the precise distribution of the trait.

We thank NORMAN DRINKWATER for calling our attention to the question of nonparametric statistics for QTL mapping and for comments on the manuscript. We thank WILLIAM DIETRICH for sharing the mapping data from the *Mom-1* study. We thank DAVID SIEGMUND for comments on the manuscript. This work was supported in part by a grant from the National Institutes of Health (HG-00098) to E. S. L.

LITERATURE CITED

AHN, S., and S. D. TANKSLEY, 1993 Comparative linkage maps of the rice and maize genomes. Proc. Natl. Acad. Sci. USA **90:** 7980–7984.

ALDOUS, D., 1989 *Probability Approximations via the Poisson Clumping Heuristic.* Spinger-Verlag, New York.

ANDERSSON, L., C. S. HALEY, H. ELLEGREN, S. A. KNOTT, M. JOHANSSON *et al.*, 1994 Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science **263:** 1771–1774.

BISHOP, M. D., S. M. KAPPES, J. W. KEELE, R. T. STONE, S. L. SUNDEN *et al.*, 1994 A genetic linkage map for cattle. Genetics **136:** 619–639.

DIETRICH, W. F., H. KATZ, S. E. LINCOLN, H. SHIN, H. FRIEDMAN *et al.*, 1992 A genetic map of the mouse suitable for typing intraspecific crosses. Genetics **131:** 423–447.

DIETRICH, W. F., E. S. LANDER, J. S. SMITH, A. R. MOSER, K. A. GOULD *et al.*, 1993 Genetic identification of Mom-1, a major modifier locus affecting Min-induced intestinal neoplasia in the mouse. Cell **75:** 631–639.

DIETRICH, W. F., J. C. MILLER, R. G. STEEN, M. MERCHANT, D. DAMRON *et al.*, 1994 A genetic map of the mouse with 4,006 simple sequence length polymorphisms. Nat. Genet. **7:** 220–245.

DRINKWATER, N. R., and KLOTZ, J. H., 1981 Statistical methods for the analysis of tumor multiplicity data. Cancer Res. **41:** 113–119.

DUPUIS, J., 1994 Statistical problems associated with mapping complex and quantitative traits from genomic mismatch scanning data. Technical Report No. 2, Department of Statistics, Stanford University.

FEINGOLD, E., P. O. BROWN and D. SIEGMUND, 1993 Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am. J. Hum. Genet. **53:** 234–251.

FULKER, D. W., and L. R. CARDON, 1994 A sib-pair approach to interval mapping of quantitative trait loci. Am. J. Hum. Genet. **54:** 1092–1103.

HALEY, C. S., S. A. KNOTT and J. M. ELSEN, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics **138:** 1195–1207.

HILBERT, P., K. LINDPAINTNER, J. S. BECKMANN, T. SERIKAWA, F. SOUBRIER *et al.*, 1991 Chromosomal mapping of two genetic loci associated with blood-pressure regulation in hereditary hypertensive rats. Nature **353:** 521–529.

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 **135:** 205–211.

JANSEN, R. C., and P. STAM, 1994 High resolution of quantitative traits into multiple loci via nterval mapping. Genetics, **136:** 1447–1455.

KENDALL, M., and A. STUART, 1979 *The Advanced Theory of Statistics* Ed. 4. Charles Griffin, London.

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

LEADBETTER, M. R., G. LINDGREN and H. Rootzen, 1983 *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, New York.

LINDGREN, B. W., 1968 *Statistical Theory* Ed. 2. Macmillan, New York.

LUO, Z. W., and M. J. Kearsey, 1992 Interval mapping of quantitative trait loci in an F_2 population. Heredity, **69:** 236–242.

MORENO-GONZALEZ, J., 1993 Efficiency of generations for estimating marker-associated QTL effects by multiple regression. Genetics **135:** 223–231.

MOSER, A. R., W. F. DOVE, K. A. ROTH and J. I. Gordon, 1992 The Min (multiple intestinal neoplasia) mutation: its effect on gut epithelial cell differentiation and interaction with a modifier system. J. Cell Biol. **116:** 1517–1526.

PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN *et al.*, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature **335:** 721–726.

POSTLETHWAIT, J. H., S. L. JOHNSON, C. N. MIDSON, W. S. TALBOT, M. GATES *et al.*, 1994 A genetic linkage map for the zebrafish. Science **264:** 699–703.

RISE, M. L., W. N. FRANKEL, J. M. COFFIN and T. N. Seyfried, 1991 Genes for epilepsy mapped in the mouse. Science **253:** 669–673.

RODOLPHE, F., and M. Lefort, 1993 A multimarker model for detecting chromosomal segments displaying QTL activity. Genetics **134:** 1277–1288.

ROHRER, G. A., L. J. ALEXANDER, J. W. KEELE, T. P. SMITH and C. W. Beattie, 1994 A microsatellite linkage map of the porcine genome. Genetics **136:** 231–245.

SAX, K., 1923 The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. Genetics **8:** 552–560.

SIEGMUND, D., 1985 *Sequential Analysis: Tests and Confidence Intervals.* Springer, New York.

SOLLER, M., and T. Brody, 1976 On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. **47:** 35–39.

STUBER, C. W., S. E. LINCOLN, D. W. WOLFF, T. HELENTJARIS and E. S. Lander, 1992 Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics **132:** 823–839.

TANKSLEY, S. D., M. W. GANAL, J. P. PRINCE, M. C. DE VICENTE, M. W. BONIERBALE *et al.*, 1992 High density molecular linkage maps of the tomato and potato genomes. Genetics **132:** 1141–1160.

TODD, J. A., T. J. AITMAN, R. J. CORNALL, S. GHOSH, J. R. HALL *et al.*, 1991 Genetic analysis of autoimmune type 1 diabetes mellitus in mice. Nature **351:** 542–547.

WRIGHT, S., 1968 *Evolution and the Genetics of Populations, Vol. 1, Genetic and Biometric Foundations.* University of Chicago Press, Chicago.

ZENG, Z. B., 1993 Theoretical basis for separation of multiple linked gene effects in mappping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972–10976.

Communicating editor: D. BOTSTEIN

APPENDIX A

We show here that, under the null hypothesis that s is not linked to a QTL, $Z_W(s)$ has mean 0 and variance 1 and is asymptotically normal as $n \to \infty$. That $\langle Z_W(s) \rangle = \langle Y(s) \rangle = 0$ follows from the definition of $Y_W(s)$ and the fact that $rank(i)$ has expectation $(n + 1)/2$ and is independent of $E[x_i(s) \mid DATA]$. That $\langle Z_W(s)^2 \rangle = 1$ follows immediately from the Equation 2, in which $Z_W(s)$ is defined to be $Y_W(s)$ divided by its standard deviation. That $Z_W(s)$ is asymptotically normal as $n \to \infty$ follows from the Central Limit Theorem, because $Z_W(s)$ is a sum of independent random variables with finite variances and $Z_W(s)$ has fixed mean and variance as $n \to \infty$ (see, *e.g.*, LINDGREN 1968, section 2.5.2).

The expectation $E[x_i(s) \mid DATA]$ can be easily calculated for any position s given the genotypes at the closest flanking markers. For a fixed individual i, let $g(s)$ denote the genotype at position s, let g_L denote the genotype at the closest informative marker on the left and let g_R denote the genotype at the closest informative marker on the right. Let θ denote the recombination fraction between the two flanking markers, and let θ_1 and θ_2 denote the recombination fractions between position s and the left and right flanking markers, respectively. Note that s and θ determine θ_1 and θ_2 and that $\theta = \theta_1(1 - \theta_2) + \theta_2(1 - \theta_1)$. It is straightforward to calculate the probability distribution of $g(s)$ given g_L and g_R as follows:

g_L	g_R	$g(s)$	$P[g(s) \mid g_L g_R]$
A	H	A	$\theta_2(1 - \theta_1)/\theta$
A	H	H	$\theta_1(1 - \theta_2)/\theta$
A	A	A	$(1 - \theta_1)(1 - \theta_2)/(1 - \theta)$
A	A	H	$\theta_1\theta_2/(1 - \theta)$
H	A	A	$\theta_1(1 - \theta_2)/\theta$
H	A	H	$\theta_2(1 - \theta_1)/\theta$
H	H	A	$\theta_1\theta_2/(1 - \theta)$
H	H	H	$(1 - \theta_1)(1 - \theta_2)/(1 - \theta)$

From this table, it easy to compute the expected value of $x_i(s)$ given any value of the data (g_L, g_R). Moreover, it is easily calculated that $\langle E[x_i(s) \mid DATA] \rangle = 0$ and $\langle E[x_i(s) \mid DATA]^2 \rangle = v(\theta, \theta_1, \theta_2)$, where

$$v(\theta, \theta_1, \theta_2) = \left\{ \frac{(\theta_1 - \theta_2)^2}{\theta} + \frac{[1 - (\theta_1 + \theta_2)]^2}{1 - \theta} \right\}$$

Finally, the variance of $Y(s)$ is given by

$$\langle Y_W(s)^2 \rangle = \sum_i (n + 1 - 2 \cdot rank(i))^2 \times \langle E[x_i(s) \mid DATA]^2 \rangle$$

$$= \sum_i (n + 1 - 2i)^2 v(\theta, \theta_1, \theta_2)$$

$$= \left(\frac{n^3 - n}{3} \right) v(\theta, \theta_1, \theta_2)$$

APPENDIX B

We wish to show that $Z_W(s)$ is asymptotically distributed as a standard Ornstein-Uhlenbeck diffusion process. This process is an example of a class of stochastic

processes known as Gaussian processes. A stochastic process $Y(s)$ is Gaussian if for each $k = 1, 2, \cdots$ and for each $x_1 < x_2 < \cdots x_k$, the random variables $Y(x_1), Y(x_2), \ldots, Y(x_k)$ are jointly normally distributed. In particular, $Y(x)$ must be normally distributed at any point x. The Ornstein-Uhlenbeck process is a Gaussian process with the properties that its mean is $\langle Y(t) \rangle = 0$ and its correlation function is $\langle Y(t) Y(s) \rangle = e^{-\beta|s-t|}$. It has already been established in APPENDIX A above that $Z_W(s)$ is normally distributed with mean 0. The correlation function of $Z_W(s)$ satisfies the Ornstein-Uhlenbeck conditions with $\beta = 2$ because $\langle x_i(s), x_i(t) \rangle = e^{-2|s-t|}$ under Haldane's mapping function (cf. APPENDIX A3 of LANDER and BOTSTEIN 1989). That $Z_W(t_1), Z_W(t_2), \ldots, Z_W(t_k)$ is multivariate normal follows from the first two statements together with the fact that, for $t_1 < t_2 < t_3$, the genotypes $g_i(t_1)$ and $g_i(t_3)$ are conditionally independent given $g_i(t_2)$ (assuming no crossover interference). For the other one degree of freedom statistics discussed in the section on other crosses, the correlation functions still satisfy the Ornstein-Uhlenbeck condition, but now with the different values of β as discussed in the text.

The theory of extreme value distribution of the Ornstein-Uhlenbeck process has been well studied (LEADBETTER *et al.* 1983). Specifically, the probability that Z_W exceeds a high value T over an interval of length L can be approximated by $\alpha \approx (1 + \beta L T^2) P[Z > T]$, where $P[Z > T]$ is the tail probability of the unit normal (cf. FEINGOLD *et al.* 1993). Because maximum values of Z_W on different chromosomes are independent, Proposition 1 follows by summing the (small) false positive probabilities over the C chromosomes. We use $P[|Z| > T]$ because we assume a two-sided test, as discussed above.

A somewhat better approximation may be used to estimate the overall false positive probability when it is no longer much less than 1 but the probabilities for individual chromosomes are still much less than 1. This situation arises if a low threshold is chosen. Let α_i denote the above approximation for the ith chromosome. Then the overall false positive rate $\alpha \approx 1 - \prod(1 - \alpha_i) \approx 1 - \prod \exp(-\alpha_i) = 1 - \exp(-\Sigma\alpha_i)$. Note that $\Sigma\alpha_i$ is just the result in Proposition 1, which is recovered when α is small. For example, in the mouse genome for a threshold of 3.0, $\Sigma\alpha_i = 0.83$, whereas $1 - \exp(-\Sigma\alpha_i) = 0.56$. For a threshold of 3.9, $\Sigma\alpha_i = 0.049$, whereas $1 - \exp(-\Sigma\alpha_i) = 0.047$.