Behavior Genetics, Vol. 26, No. 2, 1996

Type I and Type II Error Rates for Quantitative Trait Loci (QTL) Mapping Studies Using Recombinant Inbred Mouse Strains

J. K. Belknap,[1,2] S. R. Mitchell,[1] L. A. O'Toole,[1] M. L. Helms,[1] and J. C. Crabbe[1]

Received 11 Sept. 1995—Final 17 Oct. 1995

Effective mapping strategies for quantitative traits must allow for the detection of the more important quantitative trait loci (QTLs) while minimizing false positives. Type I (false-positive) and Type II (false-negative) error rates were estimated from a computer simulation of QTL mapping in the BXD recombinant inbred (RI) set comprising 26 strains of mice, and comparisons made with theoretical predictions. The results are generally applicable to other RI sets when corrections are made for differing strain numbers and marker densities. Regardless of the number or magnitude of simulated QTLs contributing to the trait variance, the p value necessary to provide genome-wide .05 Type I error protection was found to be about $p = .0001$. To provide adequate protection against both Type I ($\alpha = .0001$) and Type II ($\beta = .2$) errors, a QTL would have to account for more than half of the between-strain (genetic) variance if the BXD or similar set was used alone. In contrast, a two-step mapping strategy was also considered, where RI strains are used as a preliminary screen for QTLs to be specifically tested (confirmed) in an F_2 (or other) population. In this case, QTLs accounting for ~16% of the between-strain variance could be detected with an 80% probability in the BXD set when $\alpha = 0.2$. To balance the competing goals of minimizing Type I and II errors, an economical strategy is to adopt a more stringent α initially for the RI screen, since this requires only a limited genome search in the F_2 of the RI-implicated regions (~10% of the F_2 genome when $p < .01$ in the RIs). If confirmed QTLs do not account in the aggregate for a sufficient proportion of the genetic variance, then a more relaxed α value can be used in the RI screen to increase the statistical power. This flexibility in setting RI α values is appropriate only when adequate protection against Type I errors comes from the F_2 (or other) confirmation test(s).

KEY WORDS: QTL mapping; recombinant inbred strains; C57BL/6; DBA/2; BXD.

INTRODUCTION

Loci that influence a continuously distributed (or quantitative) trait are known as quantitative trait loci, or QTLs. Typically, several QTLs influence a quantitative trait (polygenic inheritance), as well as multiple environmental influences. The pioneering QTL mapping efforts in laboratory rodents used large F_2 or backcross populations and the then-new marker technologies, i.e., RFLPs and SSLPs (microsatellites), to generate the large number of marker loci needed for a genome-wide search (Rise *et al.*, 1991; Jacob *et al.*, 1991; Hilbert *et al.*, 1991). More recently, recombinant inbred (RI) strains have been used for provisional QTL mapping

[1] Research Service (151W), VA Medical Center and Department of Medical Psychology, Oregon Health Sciences University, Portland, Oregon 97201.
[2] To whom correspondence should be addressed at Research Service (151W), VA Medical Center, Portland, Oregon 97201. Fax: (503) 721-7985; e-mail: belknajo@ohsu.edu.

0001-8244/96/0300-0149$09.50/0

(Plomin *et al.*, 1991). Unlike segregating populations, they require no new genotyping and they possess all the advantages of inbred strains, e.g., the accumulation of trait and marker information across time and laboratories on the same highly replicable genotypes (Bailey, 1981). This has led to a marked increase in the number of papers using RI strains of mice for QTL mapping purposes. The BXD RI set, derived by Taylor (1989) from the C57BL/6 (B6) and DBA/2 (D2) progenitor strains, has been the most frequently used for this purpose largely because it has the largest marker database (~1500), and among the largest number of strains (n = 26), and its progenitors are two of the most genetically diverse and commonly studied inbred mouse strains. However, most of the basic conclusions presented below apply to other RI sets when corrected for differing strain numbers and marker densities. The largest RI sets are much preferred, and in the mouse this would include, in addition to the BXD set, the LSXSS set, with 27 strains (DeFries *et al.*, 1989; Markel *et al.*, 1995), the AKXD set, with 25 strains (Taylor, 1989), and the pooled AXB, BXA sets, with 31 strains (Marshall *et al.*, 1992).

Whether RI or segregating populations are used, detecting and mapping QTLs typically require a hundred or more marker loci distributed throughout the genome (Tanksley, 1993). Due largely to the development of high-density marker maps (e.g., Dietrich *et al.*, 1992), QTL mapping can now be carried out for almost any heritable trait in the mouse. This powerful capability comes at a rather high statistical price—greatly inflated Type I errors (false positives) arising from the large number of markers needed for a genome-wide search (Lander and Schork, 1994). The standard remedy for this problem is to use more stringent α levels (reduce acceptable Type I error risk), but no consensus exists for RI strains as to what α level for individual markers (or the interval between markers) should be used to determine statistical significance. A general guideline for which there is much agreement is to use an α value for individual markers (α_S) that yields $p < .05$ protection against *even one* Type I error occurring anywhere in the genome, i.e., $\alpha_G = .05$ (Lander and Botstein, 1989). A genome-wide $\alpha_G = .05$ implies that there will be a 95% probability of no Type I errors in the entire marker set. Thus, the conventional $p < .05$ significance level is still operative, but it applies to *all* markers used in a genome-wide search (α_G), and *not* to individual markers (or intervals) examined singly (α_S). The relationship between α_G and α_S can be estimated by $\alpha_G/\alpha_S = k$, where k is the Bonferroni correction used to adjust α_S to attain a desired value of α_G (Rice, 1989; Belknap, 1992; Miller, 1981). The estimated α_S for individual markers that results in $\alpha_G = .05$ has been the subject of differences in opinion. Recently recommended α_S values for mapping searches using only RI strains have ranged from $p < .0006$ to $p < .00002$ (Belknap, 1992; Neumann, 1992; Manly, 1993; Lander and Schork, 1994).

Because the number of genotypes (strains) in an RI set is limited in the mouse (n = 26 for the BXD), using more stringent α levels to reduce Type I errors (false positives) has the unfortunate consequence of increasing the risk of Type II errors (β; or probability of false negatives). Type II errors are failures to detect and map QTLs when they are present. The statistical power is defined as $1 - \beta$, which is the probability of correctly detecting a QTL. Consideration must be given to both Type I (α) and II (β) errors because of this inevitable trade-off between the two sources of error when n is limited.

In addition to using more stringent α_S levels, there is a second way to control Type I error rates--independent experimental confirmation. The RI QTL mapping results are subjected to confirmation testing using other genetic models to determine which presumed QTLs can be independently supported (Johnson *et al.*, 1992; Neumann, 1992; Belknap, 1992). This could be carried out using other RI sets, recombinant congenic sets (Groot *et al.*, 1992), standard inbred (non-RI) strains (e.g., Goldman *et al.*, 1987), selectively bred lines originating from an F_2 (Belknap *et al.*, in preparation), congenic strains (Oliverio *et al.*, 1976; Bailey, 1981), or $BXDF_1$'s (Plomin *et al.*, 1995) or by linkage analysis in F_2 or backcross populations using the same progenitors as the RI set (e.g., Belknap *et al.*, 1995; Crabbe *et al.*, 1994a, b; Buck, 1995). This strategy is embodied in recently proposed two-step or multistep mapping approaches where all RI QTL results are subjected to confirmation testing to weed out Type I errors (Johnson *et al.*, 1992; Belknap, 1992; Crabbe *et al.*, 1994a, b; Buck, 1995; Plomin *et al.*, 1995; Markel *et al.*, 1996). When confirmation testing is done, less stringent α_S levels can be used for the initial RI

screen, e.g., $p < .05$, resulting in lower Type II error rates (failing to detect QTLs) compared to more stringent values. This is feasible when adequate protection against Type I errors comes from the confirmation test(s).

The purpose of this paper is to assess the risk of Type I and II errors in RI strains used for QTL mapping purposes, from both a theoretical and a computer simulation framework, and to compare the results with those in the literature. We also wished to compare error rates when no QTLs were simulated compared to simulations when one, two, or three QTLs contribute to the strain variation. The latter better approximates most QTL mapping studies of heritable traits. Finally, we wished to explore the implications of these findings for two- or multistep mapping strategies.

METHODS

Theoretical Relationships Between Type I and Type II Errors

The relationship among α_S, power $(1 - \beta)$, n (sample size), and QTL effect size is given by Lander and Botstein (1989) and Soller *et al.* (1976) as $n = (Z_\alpha + Z_\beta)^2/(s^2_{QTL}/s^2_{RES})$ for a backcross or F_2 population, where Z_α and Z_β are the normal variates for the desired values of α_S and β, s^2_{QTL} is the phenotypic variance due to (or explained by) a QTL, and s^2_{RES} is the residual (unexplained) phenotypic variance. The total phenotypic variance, s^2_P, is equal to $s^2_{QTL} + s^2_{RES}$. These expressions are useful for estimating the F_2 or backcross sample size needed to test adequately RI QTL results in two-step mapping studies, or when a segregating population is used alone. [These calculations presume that all phenotyped animals are also to be genotyped. If only the extreme ends of the phenotypic F_2 distribution are to be genotyped to save costs, the sample size of phenotyped animals will need to be somewhat greater to offset the usually small loss in power shown in Fig. 5 of Lander and Botstein (1989).]

The above expression can be adapted to RI strains by using strain means rather than individual mice, since each strain is a unique genotype, and using t rather than Z when the number of strains (n) is <30, as follows: $n - 2 = (t_\alpha + t_\beta)^2/(s^2_{QTL}/s^2_{RES})$, where s^2_{QTL} and s^2_{RES} are sample variances based on RI strain means for the phenotype, and $s^2_{QTL} + s^2_{RES} = s^2_B$, the total between-strain variance. Using $n - 2 = 24$ *df* for the BXD RI set, this equation can be rearranged to $(s^2_{QTL}/s^2_{RES}) = (t_\alpha + t_\beta)^2/24$. In addition to s^2_{QTL}/s^2_{RES}, the ratio of the QTL variance to the residual (error) variance, we also want to know the ratio of the QTL variance to the total variance between strains, or s^2_{QTL}/s^2_B. This is given by $s^2_{QTL}/s^2_B = s^2_{QTL}/(s^2_{QTL} + s^2_{RES}) = (s^2_{QTL}/s^2_{RES})/[1 + (s^2_{QTL}/s^2_{RES})]$. [The latter expression is useful in the next paragraph.] Since s^2_B estimates the additive genetic variance, V_A, in the RI set (or 2 V_A in the F_2 as noted below), the ratio s^2_{QTL}/s^2_B provides an estimate of the proportion of the additive genetic variance due to a QTL. (We ignore for the moment the effect of errors in strain mean estimation, which would result in $s^2_B > V_A$.) We refer to s^2_{QTL}/s^2_B as the QTL effect size in the discussion below.

As an example of the use of the above expressions, let us set $\alpha_S = .05$ (two-tailed, $t_\alpha = 2.06$, $df = 25$) and $\beta = .1$ (one-tailed, $t_\beta = 1.32$, $df = 25$), i.e., power = .9. We then have $s^2_{QTL}/s^2_{RES} = (t_\alpha + t_\beta)^2/24 = 3.38^2/24 = .47$. To get s^2_{QTL}/s^2_B, we calculate .47/(1 + .47) = .32 from the expression given in the prior paragraph. Thus, QTLs of this effect size (32% of s^2_B) can be expected to be detected correctly 90% of the time (i.e., power = .9) at $p < .05$. In general, this equation can be used to determine the relationship among α_S, power $(1 - \beta)$, QTL effect size (s^2_{QTL}/s^2_B), and n. This relationship is shown in Fig. 1 for several values of α_S ranging from .20 to .0001 with $n = 26$ strains and in Fig. 2 for $n = 31$ strains, as in the AXB, BXA set. For other n, a correction can be made as given in the figure legends.

While s^2_B provides a reasonable estimate of V_A in the RI set (or $2V_A$ in the F_2) for many practical purposes, it is often desirable to have a more accurate estimate of the additive genetic variance from RI data, which in turn permits more accurate narrow-sense heritability (h^2_{RI}) estimates. This can be obtained from a one-way ANOVA between and within RI strains, and calculating the components of variance, σ^2_B and σ^2_W (Sokal and Rohlf, 1981), as follows: $V_A = \sigma^2_B = (MS_B - MS_W)/n_W$, where MS_B is the mean square between strains, MS_W is the mean square within strains (which directly estimates σ^2_W), and n_W is the number of mice tested per strain. [If n_W is not equal for all strains, a correction is given by Sokal and Rohlf (1981).] From this, $h^2_{RI} = \sigma^2_B/(\sigma^2_B + \sigma^2_W)$. Estimates of h^2_{RI} can be used to estimate the heritability to be expected in an F_2 population derived from the same two progenitor

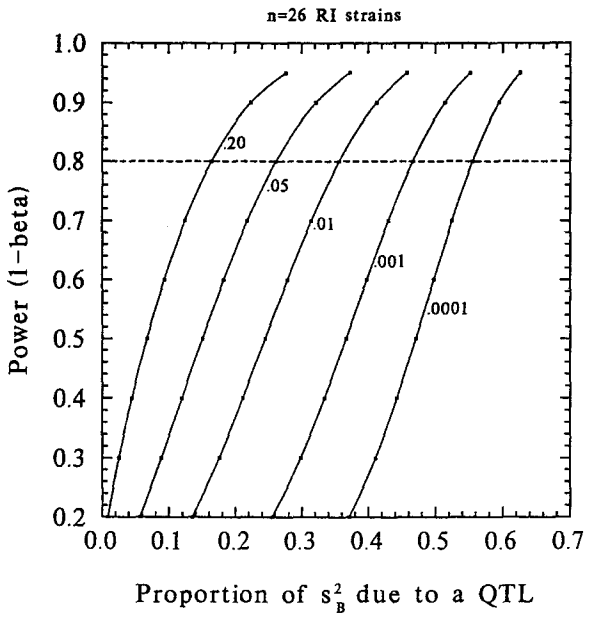

Fig. 1. Statistical power is shown plotted against QTL effect size expressed as a proportion of the between-strain variance, s^2_{QTL}/s^2_B, for five preset α_S levels from .20 to .0001. Plotted values were from the equations given under Methods based on $n = 26$ strains. For other n, this figure can be used if the X-axis values are multiplied by $(26/n)^{3/4}$. This correction is only approximate but is accurate to within $\pm 3\%$ if n is within ± 5 of 26, and s^2_{QTL}/s^2_B is not too large (<.4). More accurate values can be obtained by using the equations given under Methods.

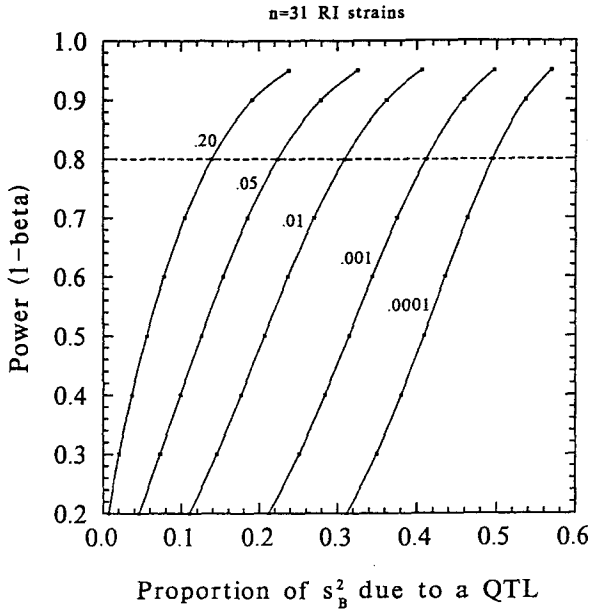

Fig. 2. Conditions are the same as noted in the legend to Fig. 1 except that $n = 31$ strains as in the AXB, BXA set. For other n, this figure can be used if the X-axis values are multiplied by $(31/n)^{3/4}$. This correction is only approximate but is accurate to within $\pm 3\%$ if n is within ± 6 of 31, and s^2_{QTL}/s^2_B is not too large (<.4). More accurate values can be obtained by using the equations given under Methods.

strains, as shown by Hegmann and Possidente (1981): $h^2_{F2} = 1/2\sigma^2_B/(1/2\sigma^2_B + \sigma^2_W)$, where σ^2_W and σ^2_B are estimated from a one-way ANOVA of RI data as described above. Roughly, h^2_{F2} will be about one-half of h^2_{RI} when the latter is not too large, i.e., <.4. The difference between h^2_{RI} and h^2_{F2} is due largely to the fact that RI strains are comprised only of homozygotes, which contribute much more to the additive genetic variance than do the usually intermediate-scoring heterozygotes comprising one-half of F_2 populations. For the same reason, s^2_{QTL} and V_A in an F_2 will also tend to be about half that seen in comparable RI data. In contrast, the total phenotypic variance (s^2_P) can be expected to be only slightly smaller in an F_2 compared to an RI population if V_A is not too large. The variance due to a QTL in an RI set, s^2_{QTL}, can also be estimated by $(M_{A1} - M_{A2})^2/4$, where $M_{A1} - M_{A2}$ is the difference in phenotypic means between the two homozygote classes at a QTL or very closely linked marker. One-half of this value, or $(M_{A1} - M_{A2})^2/8$, gives an estimate of s^2_{QTL} to be expected in a comparable F_2 population for the same QTL.

Simulation of Type I and II Errors in the BXD Set

We attempted a small-scale computer simulation to verify that the above theoretical relationships were generally appropriate for the BXD RI set. Simulation of Type I errors was carried out under two conditions, either assuming no QTLs anywhere in the genome, as assumed by Lander and Schork (1994), or by allowing one, two, or three simulated QTLs to contribute to the strain variation, each with varying effect sizes. For no QTLs, strain distributions for 26 strains were generated by sampling at random from the normal distribution with a mean of zero and a standard deviation of one. This was repeated 20 times, resulting in 20 random variables (traits or phenotypes), each with 26 randomly generated strain means. This simulates traits that either are not heritable or have QTLs so small as to be beyond detection.

Each simulated trait was subjected to QTL analysis, which involved the determination of the point biserial correlation coefficient between each trait (RI strain means) and each of the 1493 marker loci from the BXD marker set of Manly and Cudmore (1994) from the database of Dr. R. W. Elliott. For each marker, strains bearing the B6 allele were scored as zeros and the D2 allele as ones. While we used correlation coefficients to screen the genome for associations with each trait, the resulting p values are equivalent to those based on a t test of the difference in phenotypic values between the two genotypic classes (B6 or D2 homozygotes) for each marker. This, in turn, is equivalent in p values to those given by a regression of phenotype on gene dosage (0 or 2). This RI QTL analysis method has been recently described (e.g., Crabbe *et al.*, 1994a, b; Belknap *et al.*, 1995). The Systat statistical package (versions 5 and 6) was used for all statistical analyses.

Type I errors were tallied whenever at least one marker showed p values $<.05$, $<.01$, or $<.001$ (two-tailed) in a chromosome region. A Type I error (false-positive QTL) typically involved a string of closely linked marker loci showing p values exceeding one or more of the three preset α_S levels over a span of 1 to 30 cM of a chromosome. These could be unambiguously determined and counted >90% of the time. Occasionally, two apparent false-positive QTLs located close together overlapped to the point where it was difficult to ascertain whether one or two QTLs should be tallied from the bimodal distribution of p values along the length of a chromosome. In these instances, we used a decision rule employed previously (Crabbe *et al.*, 1994a, b). We first took the smallest p value (largest correlation) for each putative QTL as the best estimate of location. We counted two false-positive QTLs whenever the difference in peak p values was >10 cM and one or more of the intermediate markers showed $p > .25$. While other decision rules could have been used, the fact that less than 10% of false-positive QTLs required any decision rule implies that the final results are not likely to be appreciably affected by the decision rule choice. The centimorgan map locations were from Silver *et al.* (1994).

When simulated QTLs were present, we used the general linear model: $s^2_B = s^2_{QTL\ A} + s^2_{QTL\ B} + s^2_{QTL\ C} + s^2_{RES}$, where the between-strain variance is partitioned into the variances due to individual QTLs, and s^2_{RES}, the residual or unexplained variance due to undetected QTLs and environmental sources of strain mean variation, e.g., measurement error. s^2_{RES} is a normally distributed random variable ("error" term) uncorrelated with any of the QTL variances.

We wished to examine both Type I and Type II error rates when single simulated QTLs were allowed to vary in effect size from 10 to 54% of s^2_B. For this purpose, our simulated QTLs were assigned the strain distribution patterns (SDPs) of the markers *Pep3*, *Scn2a*, and *Il2*, and scored as zeros and ones for the two possible genotypes for each marker. A random variable was added, s^2_{RES}, which was sampled from the normal distribution with a mean of zero and a standard deviation which varied so that any desired value of s^2_{QTL}/s^2_B could be attained. The introduced residual (error) term, s^2_{RES}, reduces the accuracy of predicting phenotype from QTL genotype, i.e., it reduces s^2_{QTL}/s^2_B of each QTL. For each combination of QTLs and effect sizes, five random samples (s^2_{RES}) were used, resulting in 45 new traits for analysis. The distributions were unimodal and approximated the t distribution with very small df. Previously, we have shown that bimodality appears in BXD RI strain mean data only when individual QTLs account for about two-thirds or more of s^2_B (Belknap *et al.*, 1993). The failure to observe any obvious bimodal distributions in this study is fully consistent with this earlier finding.

In addition to the above simulation of single QTLs with varying magnitudes, we also wished to simulate the simultaneous presence of two or three QTLs for each trait. This was done by adding together two or three simulated QTL SDPs in all possible combinations (AB, AC, BC, and ABC) plus random variation, s^2_{RES}, in varying degrees as described above. There were 15 replications (5 per simulated QTL) for each of the 4 possible QTL combinations and effect sizes, for a total of 60 new traits. These were also distributed unimodally approximating the t distribution with very small df.

From the Type I errors observed, it is possible to determine from the Bonferroni inequality the α_S for single markers that yields $\alpha_G = .05$ for a genome-wide search: $\alpha_S = \alpha_G/k$ or $k = \alpha_G/\alpha_S$, where k is the Bonferroni correction (Miller, 1981; Rice, 1989; Belknap, 1992). For example, if $\alpha_S = .0001$ and $\alpha_G = .05$, $k = 500$. [This equation is reasonably accurate when both α's are fairly small ($<.20$); otherwise, $\alpha_S = 1 - (1 - \alpha_G)^{1/k}$ is more

Table I. Mean Number of Type I Errors (N_G) Observed as a Function of Three Preset Alpha (α_S) Levels for No QTLs or One to Three Simulated QTLs as Described under Methods[a]

	$\alpha_S = .05$	$\alpha_S = .01$	$\alpha_S = .001$
No QTLs (random variation)			
$N_G \pm$ SE	19.4 ± .98 (20)	4.3 ± .48 (20)	.52 ± .18 (20)
α_S (p) estimate	1.3×10^{-4}	1.2×10^{-4}	1.0×10^{-4}
Estimated LOD	3.2	3.2	3.3
One to three QTLs			
$N_G \pm$ SE	19.0 ± .40 (105)	3.7 ± .21 (105)	.54 ± .06 (105)
α_S (p) estimate	1.3×10^{-4}	1.4×10^{-4}	1.0×10^{-4}
Estimated LOD	3.2	3.2	3.3

[a] Also given are the estimated α_S values yielding genome-wide .05 protection against even one false positive, i.e., $\alpha_G = .05$, from the Bonferroni inequality. These were estimated from N_G by calculations described under Methods. The number of simulated traits used to calculate each mean value is given in parentheses. All simulations used the BXD RI marker set of Manly and Cudmore (1994) and assumed no planned confirmation testing.

appropriate (Sokal and Rohlf, 1981).] Since the raw data from our simulations were numbers of Type I errors per genome-wide analysis, the above relationship can also be expressed in terms of the number of such errors rather than α, as follows: $k = N_G/N_S$, where N_G is the number of Type I (false-positive QTLs) actually observed in a genome-wide search using a preset α_S, and N_S is the expected number of errors for a single marker for the same α_S; thus $N_S = \alpha_S$. For example, if $N_G = 4$ false positives were observed in a genome-wide search with α_S set at .01 (i.e., $p < .01$), then the expected number of errors for a single marker, N_S, is .01. The Bonferroni correction is estimated by $N_G/N_S = 4/.01 = k = 400$. Having thus estimated k, α_S that will yield $\alpha_G = .05$ can be estimated by calculating $\alpha_S = \alpha_G/k = .05/400 = .00013$. As noted below, this α_S estimate is close to that found in our simulation studies.

Since LOD scores are asymptotically distributed as χ^2, they can be estimated from p values using the expression LOD = $1/2(\log_{10}e)\chi^2 = .2172\chi^2$ for an additive (df = 1) model (Lander and Botstein, 1989). Because the number of RI strains is limited, the LOD scores calculated from this equation are only approximate.

For all but the no-QTL condition, we also tabulated Type II errors (failures correctly to detect simulated QTLs) at each of the three preset α_S levels. Since we wished to express the data in terms of statistical power, we determined the proportion of all simulated QTLs that were correctly detected as a function of QTL effect size. For this purpose, we used the same SDPs to simulate QTLs as in our Type I error simulations.

While our simulated and theoretical QTLs were at a marker, somewhat lower power estimates would be expected when QTLs are between markers. However, due to the high density of markers used (~1-cM average density), this reduction should be small and almost negligible for the purposes of this study (Simpson, 1989; Darvasi *et al.*, 1993). However, in RI sets with few markers, interval mapping (Markel *et al.*, 1996) can be used to largely offset the loss in power when QTLs are a considerable distance from the nearest markers.

RESULTS

No-QTL Condition

The results when there were no simulated QTLs are given in Table I. The mean numbers of Type I errors observed throughout the genome (N_G) were about 19, 4, and .5 false-positive QTLs for the preset .05, .01, and .001 α_S values. From N_G, the three estimates of the Bonferroni correction (k) were in the 400 to 500 range, yielding α_S estimates giving $\alpha_G = .05$ ranging from 1.0 to 1.3×10^{-4}. The corresponding LOD score estimates (additive; df = 1) were in the 3.2 to 3.3 range. Thus, when there are no QTLs present, about $p < .0001$ (LOD = 3.3) is needed to provide .05 protection against even one false positive in the genome, or a 95%

probability that there will be no Type I errors. Type I errors (N_G) were distributed approximately as the Poisson distribution, with the mean about equal to the variance.

One to Three Simulated QTLs

The Type I error results are shown in Table I. They were similar to the no-QTL condition for all three preset α_S. Thus, the estimated α_S when α_G = .05 was again about $p < .0001$ or 1×10^{-4}. The number of Type I errors was not significantly related to the individual QTL effect size, or the proportion of s^2_B due to all QTLs (data not shown). However, the number of Type I errors was reduced as a function of the number of QTLs by about 15% overall when comparing no QTLs to three QTLs, a small but highly significant difference (data not shown). To some degree, this is to be expected since the presence of a QTL is likely to mask any false-positive QTLs in the same chromosome region.

Statistical Power (1-Type II Error Rate)

The Type II errors (false negatives) were tallied for all conditions except when no simulated QTLs were present. The observed power was calculated as the proportion of all simulated QTLs that were correctly detected at preset α_S levels of .05, .01, and .001. The results are shown in Fig. 3 as a function of QTL effect size, s^2_{QTL}/s^2_B. Each point represents the observed power based on 15 QTLs at each of the three preset α_S values (for six of the points, 30 QTLs). From these results, the minimum size of a QTL that could be detected when power = .8 was .27, .38, and .50, respectively, for $p <$ .05, $p < .01$, and $p < .001$, as shown in Fig. 3. These values are close to the theoretical values of .26, .36, and .47 shown in Fig. 1.

Power was assessed using two end points. The first, shown in Fig. 3, required that the marker representing each QTL had to surpass one or more of the three preset α_S to qualify as detected. The second required that *any* marker within ±6 cM of the marker defining a QTL must exceed the preset α_S criteria. The latter increased the power by ~15% on average over the former, and the minimum QTL effect size that could be detected when power = .8 was reduced to .24, .34, and .47 for α_S = .05, .01 and .001, respectively. This occurred because the most significantly associated marker was not the marker representing the simulated QTL in about half of all cases but, rather, was a nearby marker a few centimorgans distant. Thus, increasing the "error" term not only reduces the power, but also causes an error in estimated QTL map location away from the "true" location of ±6 cM (90% confidence limits).

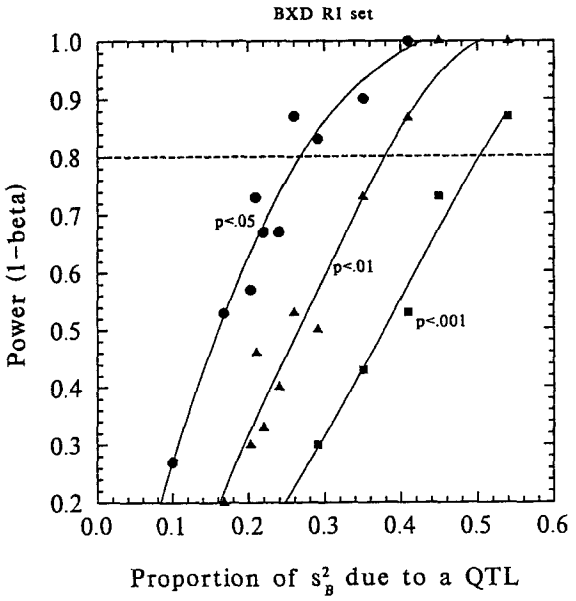

Fig. 3. Observed statistical power plotted against QTL effect size for three preset α_S levels from .05 to .001 in computer simulation studies in the BXD set (n = 26 RI strains) using 1493 markers. Each point represents the proportion of correctly detected QTLs (observed power) of 15 simulated QTLs tested, except for six points, where 30 QTLs were tested. The distance-weighted least-squares curves are also shown, which do not presuppose any functional relationship (Wilkinson, 1990).

DISCUSSION

When there were no simulated QTLs, the present simulation study indicates that a p value of ~1 $\times 10^{-4}$ (LOD ≈ 3.3) is needed to provide .05 protection against even one false-positive QTL in the genome when the BXD RI set is used alone. Roughly the same findings emerged when simulated QTLs were present. Strictly speaking, this estimated α_s is specific for the marker set we used in the simulations, which approximates a 1 to 2-cM density map. If we presume that the markers are distributed at random throughout the genome, the proportion of the genome (P) within d cM of at

least one of N markers is given by $P = 1 - e^{-2Nd/D}$ (Jacob *et al.*, 1991), where D is the map length of the entire genome, about 1600 cM in the mouse, d is the span in centimorgans on either side of each marker locus, and e is 2.71. From this expression, about 84% of the genome can be expected to be within 1 cM of a marker when there are 1493 marker loci. Since marker distributions cannot be assumed to be entirely random, these are rough but still useful estimates. In any case, in RI sets with fewer markers, the number of Type I errors would be lower (Lander and Botstein, 1989), and thus a less stringent α_S would be needed for genome-wide .05 protection. The opposite would be true for a marker set more dense than that used in our study, although the upper limit is probably about a 1-cM map in the BXD and other sets of similar strain numbers for reasons noted below.

Belknap (1992) estimated α_S when $\alpha_G = .05$ to be $p = .0006$ when the BXD set is used alone to map single-locus (qualitative) traits based on a set of empirical rules. While data are not presented here, we have carried out simulations for qualitative traits (in preparation), and the numbers of Type I errors were similar to those observed in the present study. There appears to be no fundamental difference in false-positive error rates between qualitative and quantitative traits despite the differing distributions encountered. Thus, it is clear that the earlier recommendation of Belknap (1992) is not stringent enough to provide .05 genome-wide protection with the current BXD marker set in the absence of confirmation testing.

Neumann (1992), based on a Bayesian analysis of RI strains, estimated that a p value of 2.5×10^{-4} (LOD $\approx$ 2.9) was needed to obtain a 95% probability of linkage. This analysis takes into account the very much smaller prior probability of linkage compared to nonlinkage between a QTL and a marker. The probability of linkage was calculated based on the binomial expansion, and an unlimited number of markers was tacitly assumed.

More recently, Lander and Schork (1994) recommended using $p < 2 \times 10^{-5}$ (LOD = 3.9) for significance in RI strain QTL studies, the most stringent level proposed thus far. This is largely because these authors made several assumptions that can be described as "worst case" in terms of generating more false positives. First, they assumed an unlimited number of markers, while our marker set, while large (~1500), approximates a 1- to 2-cM density map. Second, they assumed an infinite number of distinct genotypes due to recombination. In contrast, the number of RI strains (distinct genotypes) in any set is quite limited, and deriving new RI strains is not likely due to the cost and time required. The estimate of the map distance associated with 1 discordant strain of 26 (the minimum difference between two adjacent but distinct SDPs) is about 1.0 cM (Silver, 1985), which implies that there are probably only about 1600 distinct SDPs that are genetically possible of the 2^{26} (67 million) SDPs that are mathematically possible (Neumann, 1992). This suggests that even if infinite numbers of markers were available, the extent of SDP redundancy would be so high that only a 1-cM density map (equivalent to an ~4-cM map in an F_2 population) is ultimately attainable with 26 strains. Figure 4 of Lander and Botstein (1989) indicates that the difference between an infinite number of markers and a 4-cM marker density in an F_2 or backcross mouse population is about 0.4 LOD, or a 2.5-fold difference in p values. This consideration alone may account for most of the difference between the Lander and Schork (1994) estimated α_S yielding genome-wide .05 protection and those found in the present study. If the finite number of genotypes is taken into account, .4 LOD should be subtracted from the Lander and Schork (1994) estimate, yielding LOD = 3.5 ($p < 6 \times 10^{-5}$) for RI strains with strain numbers similar to the BXD set.

The Lander and Schork "worst-case" assumptions are defensible by arguing that it is better to be biased on the side of being too stringent rather than not stringent enough. However, they did not take into consideration the negative consequences of more stringent α_S on Type II error rates (false negatives). Because their assumption of unlimited numbers of markers and recombinant genotypes guarantees that Type II errors will be essentially zero, they could well afford to ignore this concern. Unfortunately, Type II errors cannot be ignored with the limited genotype numbers available in existing mouse RI sets. Moreover, the limited number of RI genotypes clearly differentiates them from segregating populations, where nearly unlimited numbers of recombinant genotypes can be approximated by using very large sample sizes.

In addition to the use of stringent α_S per experiment, another strategy for dealing with Type I errors is to emphasize independent replication. A good example is the two-step mapping approach

we (and others) have recently employed for QTL mapping. RI strains are used to screen the genome in search of provisional QTLs as Step 1. This initial step requires no genotyping, since each RI strain has been genotyped for hundreds of markers throughout the genome as a result of the cumulative effort of many workers over many years (Silver *et al.*, 1994). Since the number of available RI strains is generally too small to map all but the largest QTLs unequivocally (see Fig. 1), the RI QTL results must be followed by a second step—confirmation testing in an independent test, usually in an F_2 intercross between the two RI progenitor strains. In Step 2, each F_2 mouse must be genotyped, but only for chromosome regions that Step 1 results indicate contain QTLs. In our experience, the limited genome search typically involves only about 10% of the genome when α_S is set at $p < .01$ in Step 1. This results in a large savings in genotyping cost and effort compared to the full genome search required when there is no preliminary RI screen. This approach is exemplified by several recent reports (Crabbe *et al.*, 1994a, b; Belknap *et al.*, 1995; Buck, 1995).

Since the two-step strategy requires only a partial genome search in the F_2 (or other) population, this implies a lower risk of Type I errors compared to a full genome search. Is it therefore appropriate to use a less stringent criterion for significance from that required for a full genome search? The same question can also be asked in cases where only a few markers are used in a genome-wide search (sparse map condition). In both cases, Lander and Schork (1994) argue that lowering the significance threshold is not appropriate on the grounds that today's limited search by one investigator can potentially lead to future full genome scans carried out by the genome research community using unlimited markers and genotypes. While we see some merit in this view, for example, the establishment of significance criteria that should remain stable in the future, there are particular circumstances where exceptions are warranted. We suggest that among these are (1) studies in RI sets where the number of genotypes is fixed and is unlikely to be increased, (2) testing for the presence of a QTL previously substantiated in other studies, or (3) testing hypothesized candidate genes of known map location.

Since Steps 1 and 2 are statistically independent experiments using the same progenitor strains, it could be argued that their LOD scores for a given marker (or interval) could be added to obtain an overall LOD score for both experiments, as is often done in the human literature (Ott, 1991). However, another useful approach, which is more conservative, is to use Fisher's (1958) method of combining p values from two (or more) independent experiments testing the same hypothesis, as articulated by Sokal and Rohlf (1981, p. 779). This approach takes advantage of the fact that $-2ln(p)$ is distributed as χ^2 with df = 2. The equation for combining p's is $-2\Sigma ln(p) = \chi^2$ with $df = 2t$, where t is the number of p values from separate experiments to be combined. For example, $p = .01$(RI) and $p = .001(F_2)$ yields $p = 1 \times 10^{-4}$ as the joint probability ($\chi^2 = 23$, $df = 4$). The two-step results, in turn, can be subjected to further confirmation testing using another population derived from the same two progenitor strains. For example, using Fisher's method, three such independent experiments (e.g., RI, F2, backcross) using the same trait, markers, and progenitors, each yielding $p = .01$, .001, and .01, respectively, when pooled yields $p = 1 \times 10^{-5}$ (LOD $\approx$ 4.2) in the aggregate. Even though no one experiment alone comes close to meeting the Neumann (1992), the Lander and Schork (1994), or the present estimated α_S for significance (and might be denied publication for this reason), all three experiments jointly would meet these criteria. The separate studies could conceivably involve the work of two or more laboratories aggregated over a considerable length of time. Replication is probably the best insurance available against Type I errors and should be given at least as much emphasis as stringent criteria for significance for single experiments.

Concerning statistical power (1-Type II error rates), the estimates from our simulation studies (Fig. 3) were in good agreement with those derived theoretically from equations given by Lander and Botstein (1989) and Soller *et al.* (1976), as shown in Fig. 1. However, in RI sets with few markers, the power would be less, but this reduction can be largely offset by the use of interval mapping methods such as those recently developed for RI strains by Fulker and colleagues (Markel *et al.*, 1996). In any case, it is clear that the power to detect QTLs when an RI set is used alone is adequate only for the larger QTL effect sizes. If we assume that α_S must be set at .0001 to obtain $\alpha_G = .05$, then we can reliably detect QTLs (power = .8) only if they account for more than half of the between-strain

variance in the larger mouse RI sets. Thus, the existing RI sets, when used alone, can map unequivocally only the largest QTLs. This limitation has prompted the use of two- or multistep mapping strategies referred to above, where the burden of protection against both Type I and Type II errors can be distributed over a larger number of genotypes than those in an RI set.

In the assessment of power in our simulations, quite often the most significantly associated marker was not that assigned to serve as our QTLs but, rather, was a few centimorgans distant. This was most likely to occur when the QTL effect size was small, i.e., the error variance was greatest. This indicates that the introduction of random error variance not only reduces the ability to detect the presence of a QTL, but also can change the estimated location of a detected QTL away from its "true" location. The 90% confidence limits in our simulation studies were ±6 cM. Furthermore, using the mostly significantly associated marker within 6 cM of the "true" QTL location increased the power by about 15% (not shown). This may better simulate most research efforts where the precise location of the QTL is not known beforehand.

While power is usually defined as the probability of correctly detecting a QTL, the power also estimates the expected proportion of s_B^2 accounted for by all correctly detected QTLs. For example, let us assume that we have four QTLs, each accounting for 25% of s_B^2. (This approximates 25% of V_A when there is negligible strain mean estimation error, or a correction is made for the latter using components of variance as described under Methods.) The power analysis (Fig. 1) indicates that QTLs of this effect size can be detected with a power of .78 at $\alpha_S = .05$. The expectation is that $.78 \times 4 = 3.1$ QTLs will be correctly detected (on average), accounting for $3.1 \times .25 = .78$ of s_B^2, which equals the power.

If we assume that the minimum acceptable proportion of s_B^2 due to all detected (and mapped) QTLs is 80% (i.e., power = .8), an RI set with 26 strains is useful as a preliminary screen only for QTLs accounting for ≥26% of s_B^2 when $\alpha_S = .05$ (from Fig. 1). In the case of an F_2 or a backcross, the power can be increased by increasing the number of genotypes tested, but this is not an option with RI sets when all available strains are routinely tested. In this case, power can be increased in the RIs only by using a less stringent α_S value, say, .20. Flexibility in setting α_S for Step 1 is possible only when the confirmation test(s), or Step 2, is (are) capable of providing adequate protection against Type I errors. When segregating populations are used for confirmation testing, the sample sizes necessary to provide adequate Type I error protection can be estimated from the equations given under Methods. Steps 1 and 2 together should be designed to meet the Lander and Schork (1994) significance level of $p < .0001$ for an F_2 or backcross (additive; df = 1).

With an adequate confirmation test(s), primacy can then be given to minimizing Type II errors (increasing power) in the RI preliminary screen to avoid missing important QTLs. For example, with $\alpha_S = .20$ and 26 strains, the minimum QTL effect size detectable at power = .8 is 16% of s_B^2 (Fig. 1). The major undesirable consequences of relaxing α_S in Step 1 are that a larger portion of the genome must be searched in the confirmation test (Step 2) and a more stringent Step 2 α_S will be needed to offset the relaxation in Step 1 when using Fisher's method or adding LOD scores. When Step 1 $\alpha_S = .2$, essentially the entire burden of Type I error protection falls on Step 2.

We very roughly estimate that about 40% of the genome would need to be searched in the F_2 (Step 2) when the Step 1 $\alpha_S = .20$, compared to about 10% when the Step 1 $\alpha_S = .01$. The 40% estimate comes from adding the 20% of markers likely to be "significant" at this α_S level due to chance, plus an additional 20% for the "real" QTLs and the portions of the genome needed to flank each putative QTL with markers. Further relaxation of α_S is possible, but at some point the advantage of reduced genotyping effort diminishes to where it no longer offsets the cost and effort of testing (phenotyping) the RI strains. At this point there would be little or no cost savings compared to a full genome search in an F_2 without an RI screen. We surmise that $\alpha_S = .20$ is probably not far from this practical lower limit of relaxation.

An economical strategy for two-step mapping efforts is to start with a more stringent RI α_S level, e.g., $p < .01$, to minimize the portion of the F_2 genome to be searched, and to resort to less stringent α_S values only when the aggregate of detected and confirmed QTLs falls short of an acceptable proportion of the genetic variance, e.g., two-thirds of s_B^2 in the RI strains or two-thirds of V_A in the F_2. [Multiple regression with adjusted R^2 can be used

to assess this since it corrects for correlations among the markers that can inflate aggregate QTL variances (Phillips *et al.*, 1994).] This two-step strategy with flexible α_S yields a high probability of mapping the larger QTLs (when power >.9) but will likely miss most of the smaller QTLs when the power < 0.5 at less stringent α_S levels. If we assume that α_S = .20 is as relaxed a value as is cost-effective, and .8 is the lowest power acceptable, then the practical lower limit in QTL effect size is 16% of s_B^2 with 26 strains (Fig. 1). For the other large RI sets, this value would be similar. If it is desired reliably to detect and map QTLs much smaller than this, the RI preliminary screen would not have sufficient power. In this case, a full genome search in a very large segregating population would be needed (e.g., Berrettini *et al.*, 1994; Flint *et al.*, 1995). However, in cases where we are interested primarily in QTLs with effects larger than the above-noted minimum, the low reliability of detecting those with even smaller effects may be of little concern.

As noted above, the limitation imposed by the available number of RI strains (genotypes) in a set is sufficiently great that confirmation tests are required for virtually all but the largest QTLs. This limitation is somewhat mitigated by two considerations. First, RI strains are homozygous at all loci, which are more informative for QTL mapping than the usually intermediate-scoring heterozygotes comprising half of F_2 populations. Thus, 26 RI strains are equivalent to about 50 F_2 intercross animals for any given values of α and β. Second, by testing several animals per strain, multiple measurements on what is the same genotype can be made to assess more accurately the phenotype associated with each genotype. This is possible because each genotype (strain) is replicable by simply breeding more members of the same strain. In an F_2, in contrast, each genotype is represented by only a single mouse that cannot be replicated. Thus, the accuracy of predicting phenotype from genotype, which is essential for efficient QTL mapping, is potentially much greater in the RIs (Lander and Botstein, 1989). Another advantage of replicable genotypes is that an unlimited number of phenotypes can be assessed on the same genotypes, so that direct comparisons can be made among them at both the genetic correlation and QTL levels (Crabbe *et al.*, 1994a). In contrast, F_2 populations tend to be "single-use" studies of only one phenotype. At best, only a limited number of traits can be assessed on each F_2 mouse, and then only if the multiple measurements per mouse are not confounding.

RI strains have been invaluable for mapping loci associated with single-locus Mendelian (qualitative) traits, a purpose for which they were originally developed (Bailey, 1981; Taylor, 1978). Recently, they have also been shown to be valuable as a tool for mapping loci associated with quantitative traits, particularly as part of two- or multistep mapping strategies. However, their proper use in QTL mapping requires an appreciation of error risk, which this paper has shown are considerable but not insurmountable concerns. Since the goal of QTL mapping is to detect (and map) QTLs with minimal errors, an appropriate balance between Type I and Type II error risk is critically important in the design of two- or multistep mapping experiments.

ACKNOWLEDGMENTS

This work was supported by NIDA Contract 271-90-7405, NIH Grants AA06243, AA08621, AA10760, and DA05228, and two VA Merit Review Projects from the Department of Veterans Affairs.

REFERENCES

Bailey, D. W. (1981). Recombinant inbred strains and bilineal congenic strains. In Foster, H. L., Small, J. D., and Fox, J. G. (eds.), *The Mouse in Biomedical Research, Vol. 1*, Academic Press, New York, pp. 223–239.

Belknap, J. K. (1992). Empirical estimates of Bonferroni corrections for use in chromosome mapping studies with the BXD recombinant inbred strains. *Behav. Genet.* **22**:677–684.

Belknap, J. K., Metten, P. A., Helms, M. L., O'Toole, L. A., Angeli-Gade, S., Crabbe, J. C., and Phillips, T. J. (1993). Quantitative trait loci (QTL) applications to substances of abuse: Physical dependence studies with nitrous oxide and ethanol. *Behav. Genet.* **23**:211–220.

Belknap, J. K., Mogil, J. S., Helms, M. L., Richards, S. P., O'Toole, L. A., Bergeson, S. E., and Buck, K. J. (1995). Localization to chromosome 10 of a locus influencing morphine-induced analgesia in crosses derived from C57BL/6 and DBA/2 mice. *Life Sci. (Pharmacol. Lett.)* **57**:PL117–PL124.

Berrettini, W. H., Ferraro, T. N., Alexander, R. C., Buchberg, A. M., and Vogel, W. H. (1994). Quantitative trait loci mapping of three loci controlling morphine preference using inbred mouse strains. *Nature Genet.* **7**:54–58.

Buck, K. J. (1995). Strategies for mapping and identifying quantitative trait loci specifying behavioral responses to alcohol. *Alc. Clin. Exp. Res.* **19**:795–801.

Crabbe, J. C., Belknap, J. K., and Buck, K. J. (1994a). Genetic animal models of alcohol and drug abuse. *Science* **264**: 1715–1723.

Crabbe, J. C., Belknap, J. K., Buck, K. J., and Metten, P. (1994b). Use of recombinant inbred strains for studying genetic determinants of responses to alcohol. *Alcohol Alcohol.* **S2**:69–73.

Crabbe, J. C., Buck, K. J., Metten, P., and Belknap, J. K. (1995). Strategies for identifying genes underlying drug abuse susceptibility. In Lee, T. N. H. (ed.), *Molecular Approaches to Drug Abuse Research, Vol. III*, NIDA Res. Monogr., USGPO, Washington, DC.

Darvasi, A., Weinreb, A., Minke, V. Weller, J. I., and Soller, M. (1993). Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134**:943–951.

DeFries, J. C., Wilson, J. R., Erwin, V. G., and Petersen, D. R. (1989). LSXSS recombinant inbred strains of mice: Initial characterization. *Alc. Clin. Exp. Res.* **13**:196–200.

Dietrich, W., Katz, H., Lincoln, S. E., Shin, H.-S., Friedman, J., Dracopoli, N. C., and Lander, E. S. (1992). A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131**:423–447.

Fisher, R. A. (1958). *Statistical Methods for Research Workers*, 13th ed., Hafner, New York, pp. 99–101.

Flint, J., Corley, R., DeFries, J. C., Fulker, D. W., Gray, J. A., Miller, S., and Collins, A. C. (1995). A simple genetic basis for a complex psychological trait in laboratory mice. *Science* **269**:1432–1435.

Goldman, D., Lister, R. G., and Crabbe, J. C. (1987). Mapping of a putative genetic locus determining ethanol intake in the mouse. *Brain Res.* **420**:220–226.

Groot, P. C., Moen, C. J. A., Dietrich, W., Stoye, J. P., Lander, E. S., and DeMant, P. (1992). The recombinant congenic strains for analysis of multigenic traits: Genetic composition. *FASEB J.* **6**:2826–2835.

Hegmann, J., and Possidente, B. (1981). Estimating genetic correlations from inbred strains. *Behav. Genet.* **11**:103–114.

Hilbert, P., Lindpainter, K., Beckmann, J. S., Serikawa, T., Soubrier, F., Dubay, C. S., Cartwright, P., DeGouyon, B., Julier, C., Takahasi, S., Vincent, M., Ganten, D., Georges, M., and Lathrop, G. M. (1991). Chromosomal mapping of two genetic loci associated with blood-pressure regulation in hereditary hypertensive rats. *Nature* **353**:521–529.

Jacob, H. J., Lindpaintner, K., Lincoln, S. E., Kusumi, K., Bunker, R. K., Mao, Y.-P., Ganten, D., Dzau, V. J., and Lander, E. S. (1991). Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. *Cell* **67**:213–224.

Johnson, T. E., DeFries, J. C., and Markel, P. (1992). Mapping quantitative trait loci for behavioral traits in the mouse. *Behav. Genet.* **22**:635–653.

Lander, E. S., and Schork, N. J. (1994). Genetic dissection of complex traits. *Science* **265**:2037–2048.

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

Manly, K. E. (1993). A Macintosh program for storage and analysis of experimental genetic mapping data. *Mammal. Genome* **4**:303–313.

Manly, K. E., and Cudmore, R. (1994). *Map Manager; A Program for Genetic Mapping (v. 2.6)*, Roswell Park Cancer Institute, Buffalo, NY.

Markel, P. D., Fulker, D. W., Bennett, B., Corley, R. P., DeFries, J. C., Erwin, V. G., and Johnson, T. E. (1996). Quantitative trait loci for ethanol sensitivity in the LSXSS recombinant inbred strains: Interval mapping. *Behav. Genet.* (in press).

Marshall, J. D., Mu, J.-L., Cheah, Y.-C., Nesbitt, M. N., Frankel, W. N., and Paigen, B. (1992). The AXB and BXA set of recombinant inbred mouse strains. *Mammal. Genome* **3**:669–680.

Miller, R. G., Jr. (1981). *Simultaneous Statistical Inference*, McGraw–Hill, New York.

Neumann, P. E. (1992). Inference in linkage analysis of multifactorial traits using recombinant inbred strains of mice. *Behav. Genet.* **22**:665–676.

Oliverio, A., and Eleftheriou, B. E. (1976). Motor activity and alcohol: A genetic investigation in the mouse. *Physiol. Behav.* **16**:577–581.

Ott, J. (1991). *Analysis of Human Genetic Linkage*, Johns Hopkins Press, Baltimore, MD.

Phillips, T. J., Crabbe, J. C., Metten, P., and Belknap, J. K. (1994). Localization of genes affecting alcohol drinking in mice. *Alc. Clin Exp. Res.* **18**:931–941.

Plomin, R., McClearn, G. E., and Gora-Maslak, G. (1991). Use of recombinant inbred strains to detect quantitative trait loci associated with behavior. *Behav. Genet.* **21**:99–116.

Plomin, R., Rodriguez, L. A., Blizard, D. A., Jones, B. C., and McClearn, G. E. (1995). Alcohol acceptance, preference, and sensitivity in mice. III. Using F_1 crosses between BXD recombinant inbred strains to replicate BXD-nominated quantitative trait loci (QTL). *Alc. Clin. Exp. Res.* (in press).

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution* **43**:223–225.

Rise, M. T., Frankel, W. N., Coffin, J. M., and Seyfried, T. N. (1991). Genes for epilepsy mapped in the mouse. *Science* **253**:669–673.

Silver, J. (1985). Confidence limits for estimates of gene linkage based on analysis of recombinant inbred strains. *J. Hered.* **76**:436–440.

Silver, L. M., Nadeau, J. H., and Goodfellow, P. N. (1994). Encyclopedia of the mouse genome IV. *Mammal. Genome* **5**:S1–S295 (Special issue).

Simpson, S. P. (1989). Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theor. Appl. Genet.* 77:815–819.

Sokal, R. R., and Rohlf, F. J. (1981). *Biometry*, Freeman, San Francisco.

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

Tanksley, S. D. (1993). Mapping polygenes. *Annu. Rev. Genet.* **27**:205–233.

Taylor, B. A. (1978). Recombinant inbred strains: Use in gene mapping. In Morse, H. C. (ed.), *Origins of Inbred Mice*, Academic Press, New York, pp. 423–438.

Taylor, B. A. (1989). Recombinant inbred strains. In Lyon, M. F., and Searle, A. G. (eds.), *Genetic Variants and Strains of the Laboratory Mouse*, 2nd ed., Oxford University Press, Oxford, pp. 773–789.

Wilkinson, L. (1990). *Systat: The System for Statistics*, Systat, Inc., Evanston, IL.

Edited by David Fulker