Statistical Issues in the Analysis of Quantitative Traits in Combined Crosses

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

Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706

Manuscript received August 14, 2000
Accepted for publication April 18, 2001

ABSTRACT

We consider some practical statistical issues in QTL analysis where several crosses originate in multiple inbred parents. Our results show that ignoring background polygenic variation in different crosses may lead to biased interval mapping estimates of QTL effects or loss of efficiency. Threshold and power approximations are derived by extending earlier results based on the Ornstein-Uhlenbeck diffusion process. The results are useful in the design and analysis of genome screen experiments. Several common designs are evaluated in terms of their power to detect QTL.

QUANTITATIVE trait analysis has many applications in plant and animal breeding and in human genetics. Mapping quantitative trait loci (QTL) that influence agriculturally important traits such as grain yield in rice or milk production in cows can help scientists produce specimens with more desirable qualities. Complex human diseases, like breast cancer and diabetes, are known to have genetic etiologies. Animal models may be useful in studying their origins.

Most existing statistical methods have been developed for experimental designs with a single cross from two inbred parents (Lander and Botstein 1989; Haley and Knott 1992; Zeng 1993, 1994; Jansen and Stam 1994). Doerge *et al.* (1997) provided a comprehensive review of methodologies for detecting and locating genes affecting quantitative traits in experimental breeding populations. However, quantitative traits are often influenced by several genes with large effects (major QTL) and many genes with relatively small effects (polygenes). In animal science, where outbred parental populations are available, the polygenic effect has been taken into consideration (Fernando and Grossman 1989). In horticulture, less attention has been paid to genes with small effects, perhaps because researchers are able to rely on simple crosses such as F_2 or backcross (BC).

The effects of polygenes on standard approaches to major QTL mapping are not well understood. With a single cross, the progeny have identical relationships given the QTL genotypes, resulting in a compound symmetry structure (Yandell 1997, Ch. 25). Thus, unbiased estimates of QTL effects are still obtained when the polygenic effect is ignored, even though the power to detect the QTL is influenced by the magnitude of the polygenic effect. The situation is more complicated with several crosses, since the correlations may not be the same among all individuals. To avoid this difficulty, researchers may analyze data for each cross separately and then compare and combine the results in some fashion. Hence, some power to detect the QTL may be lost and estimates of QTL effects may be less precise.

Recently, methods were proposed to analyze all crosses simultaneously. Bernardo (1994) used Wright's relationship matrix *A* to accommodate differential correlations when analyzing diallel crosses. However, when closely related crosses are from a small number of inbred lines, it is more reasonable to treat the polygenic effect as fixed. Rebai *et al.* (1994a) extended the regression method of Haley and Knott (1992) to several F_2's from a diallel design of multiple inbred lines with all effects fixed. Elston (1990) proposed models for discriminating among modes of inheritance, including one-locus, two-locus, polygenic, and mixed major locus/polygenic inheritance when considering the F_1 and the reciprocal backcrosses derived from two inbred lines. The polygenic effect is treated as fixed and different phenotypic means and variances are used for different crosses. However, flanking marker information is not utilized and an estimate of the QTL position is not provided.

In this article, we consider an arbitrary number of crosses from multiple inbred lines. While we were preparing this manuscript, Liu and Zeng (2000) proposed a fixed-effect model to analyze combined crosses from multiple inbred lines (with or without overlapping inbred lines). Our model includes both QTL and polygenic effects and is a special case of their heteroscedastic model in the sense that the fixed effect and the variance component identify the polygenic effect. For this reason, we refer readers to Liu and Zeng (2000) for the analysis of combined lines. Our focus is the practical implications of the polygenic effects for QTL mapping, specifically bias and efficiency. Furthermore, we calculate threshold values for controlling the genome-wise

Corresponding author: Fei Zou, Department of Statistics, 1210 W. Dayton St., Madison, WI 53706. E-mail: feizou@stat.wisc.edu

TABLE 1

Naive *vs.* polygenic model simulation

Theoretical				Estimated		
QTL	Polygene	Model[a]	Location	Additive	Dominant	Polygenic
0	5	N	—[b]	1.79 (0.85)	−0.99 (1.10)	—[c]
		P	—[b]	−0.036 (1.06)	0.088 (1.41)	4.29 (0.96)
0	10	N	—[b]	3.75 (1.10)	−1.70 (1.26)	—[c]
		P	—[b]	0.23 (1.27)	−0.03 (1.67)	8.59 (0.91)
5	0	N	29.02 (6.05)	5.03 (0.83)	0.12 (0.86)	—[c]
		P	29.84 (4.73)	5.02 (0.88)	0.14 (0.92)	0.07 (0.89)
5	5	N	29.93 (5.13)	6.6 (0.61)	−0.75 (0.91)	—[c]
		P	29.42 (5.07)	4.97 (0.70)	0.03 (0.96)	4.45 (0.84)
5	10	N	29.76 (5.85)	8.16 (0.77)	−1.41 (1.25)	—[c]
		P	29.60 (5.78)	4.94 (0.97)	0.04 (1.25)	8.82 (0.98)

The values in parentheses are the standard deviations from 100 simulations. The numbers in the first two columns are the theoretical QTL and polygenic effects.

[a] N and P stand for the naive model and our polygenic model, respectively.

[b] The QTL position is not estimated since most of the $\max_d\{2\ \mathrm{LR}(d)\}$'s are not significant among 100 simulations.

[c] The polygenic effect is nonestimable in the naive model.

type I error rate. Theoretical approximations were developed to address threshold and power (Lander and Botstein 1989; Dupuis and Siegmund 1999; Rebai *et al.* 1994b, 1995) in some standard designs. However, these methods are either impractical or inappropriate with combined crosses. Our general formulas are widely applicable and easy to implement.

SIMULATION STUDY OF BIAS AND EFFICIENCY

If one combines different crosses simultaneously but ignores the different relationships among individuals, substantial bias may result. In this section, we show the effect of polygenes on the QTL estimates. We examine two crosses, BC1 and F_2, from common inbred parents P1 and P2. Although the design is simple, it illustrates the key issues. The additive effect of a single major QTL is set to 0 (*i.e.*, no QTL) and 5, respectively, with no dominance effect. Five markers are located at 0, 20, 40, 60, and 80 cM. The major QTL is located at 30 cM. The environmental errors are identically distributed for BC1 and F_2 and are sampled from $N(0, 25)$. One hundred individuals from BC1 and F_2 are simulated without background polygenes or with 10 background polygenes. The 10 background polygenes are in coupling phase and have common additive effects (*i.e.*, allele substitution effect α_k, $k = 1, 2, \ldots, 10$) 1 or 2 (see Fernando *et al.* 1994). This leads to expected polygenic differences between F_2 and BC1 of 5 and 10, respectively. We fit the model using Liu and Zeng (2000), hereafter called "model P." In addition, we employed traditional interval mapping by ignoring the polygenic effects, hereafter called "model N." For each parameter combination, 100 simulated datasets were analyzed. The results are presented in Table 1.

We observe that when there are no polygenes both models consistently estimate the QTL effects. However, model P gives more accurate estimates than model N when there are polygenic effects. The bias of model N increases as the expected polygenic differences between BC1 and F_2 increase. In summary, our simulations indicate that when analyzing combined crosses, the polygenic model produces more precise and less biased estimates than the traditional interval mapping method.

THRESHOLD AND POWER CALCULATIONS

On the basis of the simulations in the above section, fitting combined crosses (Liu and Zeng 2000) has many advantages. Calculating thresholds and power is an important practical issue in the design and analysis of such studies. The usual pointwise significance level based on the chi-square approximation is inadequate because the entire genome is tested for the presence of a QTL. Lander and Botstein (1989) showed that with an infinitely dense map, the LOD score may be approximated in large samples by an Ornstein-Uhlenbeck diffusion process for BC. Dupuis and Siegmund (1999) derived a similar result for F_2. These approximations provide formulas for the threshold and power.

For more general models (Liu and Zeng 2000), no such approximation is available. Churchill and Doerge (1994) used a randomization idea to calculate the threshold. The approach is applicable for all designs, with a dense or sparse map. However, the method is computationally intensive. In addition, since the thresholds depend on the observed data, it is unclear how to compare various designs. Rebai *et al.* (1994b, 1995) gave an upper bound for the threshold for BC and F_2 based on Davies (1977, 1987). The calculation is formidable, even for an F_2 population, and is not exact. Piepho (2001) proposed an efficient numerical method to compute the thresholds in Rebai *et al.* (1994b, 1995) for general designs.

Our approach extends the Ornstein-Uhlenbeck large sample approximations. It is quite simple and practically useful. Calculating the threshold and power under different map distances can be accomplished with closed-form expressions arising from the Ornstein-Uhlenbeck setup. Simulations shown below indicate this works well with realistic sample sizes.

Two inbred strains: In this section, we consider combined crosses from two inbred parents (P1 and P2), BC1, F_2, and BC2. Our goal is to extend Dupuis and Siegmund (1999). Generalizing the results to other designs (Liu and Zeng 2000) is straightforward. In the sequel, we assume an equispaced marker map. Suppose

$$\begin{aligned} Y &= (y_{11}, y_{12}, \ldots y_{1n_1}; y_{21}, y_{22}, \ldots, y_{2n_2}; y_{31}, y_{32}, \ldots, y_{3n_3})' \\ &= Xb + e = (X_1\, X_2)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + e, \end{aligned}$$

where X is the design matrix and e is the random error. n_1, n_2, n_3 are the number of observations for BC1, F_2, and BC2, respectively, with n observations in total. The submatrix X_1 corresponds to the covariates identifying crosses or other measurements that do not involve the QTL effects and X_2 corresponds to the QTL effects. Suppose the allele from parent P1 is q and from P2 is Q. The possible QTL genotypes are qq, Qq, and QQ. Ignoring other covariate effects, we let

$$X_1 = \left\{\begin{bmatrix} 1 & 1 & 0 \\ . & . & . \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ . & . & . \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ . & . & . \\ 1 & 0 & 1 \end{bmatrix}\right\} \quad \text{and} \quad X_2 = \left\{\begin{bmatrix} x_{111} & x_{211} \\ \ldots & \ldots \\ x_{11n_1} & x_{21n_1} \\ x_{121} & x_{221} \\ \ldots & \ldots \\ x_{12n_2} & x_{22n_2} \\ x_{131} & x_{231} \\ \ldots & \ldots \\ x_{13n_3} & x_{23n_3} \end{bmatrix}\right\},$$

where $x_{1ki} = 1$ (or 0) if individual i in cross k has genotype Qq (or else), and $x_{2ki} = 1$ (or 0) if individual i in cross k has genotype QQ (or else). The random error e is normally distributed with mean 0 and $\mathrm{Var}(e_{ki}) = \sigma_k^2$, $i = 1, 2, \ldots, n_k$, $k = 1, 2, 3$. In general $\mathrm{Var}(e) = \sigma_3^2 G$ with $G^{-1} = \mathrm{diag}(\lambda_1, \ldots, \lambda_1; \lambda_2; \ldots, \lambda_2; 1, \ldots, 1)$, where $\lambda_k = \sigma_3^2/\sigma_k^2$ for $k = 1, 2$. In the following, we assume that λ_k is known. If λ_k is unknown, then consistent maximum-likelihood (ML) estimates may be substituted and the result still holds. Without loss of generality, assume that $\sigma_3^2 = 1$. At locus d, the hypothesis of no QTL effect is H_0: $b_2 = 0$ *vs.* H_1: $b_2 \neq 0$, or equivalently, H_0: $Hb = 0$ *vs.* H_1: $Hb \neq 0$, where

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

From the normal regression theory, the likelihood ratio statistic is

$$2\ \mathrm{LR}(d) = -n \log\left[1 - \frac{(H\hat{b})'\ (H(\tilde{X}(d)'\ \tilde{X}(d))^{-1}H')^{-1}(H\hat{b})}{\|\tilde{y} - \tilde{X}(d)\hat{b}\|^2 + (H\hat{b})'\ (H(\tilde{X}(d)'\ \tilde{X}(d))^{-1}H')^{-1}(H\hat{b})}\right],$$

where $\tilde{y} = G^{-1/2}y$ and $\tilde{X}(d) = G^{-1/2}X(d)$. Under H_0,

$$2\ \mathrm{LR}(d) \approx n(\hat{\delta}_1\ \hat{\delta}_2)(A_{22})^{-1}(\hat{\delta}_1\ \hat{\delta}_2)',$$

where $(\hat{\delta}_1\ \hat{\delta}_2)'$ is the maximum-likelihood estimate of $b_2 = (\delta_1\ \delta_2)'$ and A_{22} is given in (A2) in the appendix.

The distribution of $2\ \mathrm{LR}(d)$ depends on $\hat{\delta}_1$ and $\hat{\delta}_2$, which are correlated. Thus, we cannot directly apply Dupuis and Siegmund (1999) to derive the threshold and power. In the appendix, we propose an orthogonal transformation such that $2\ \mathrm{LR}(d)$ is partitioned into the sum of squares of two uncorrelated random variables Z_1 and Z_2. This makes the asymptotic distribution transparent. Letting $\hat{W}_1 = \xi_1\hat{\delta}_1$, $\hat{W}_2 = \eta_1\hat{\delta}_1 + \eta_2\hat{\delta}_2$ (see appendix for ξ_1 and η_i), $Z_i = \sqrt{n}\hat{W}_i$, $i = 1, 2$. It is shown in the appendix that Z_1 and Z_2 are asymptotically independent and distributed N (0, 1). Thus,

$$2\ \mathrm{LR}(d) \approx n(\hat{W}_1\ \hat{W}_2)\begin{pmatrix} \hat{W}_1 \\ \hat{W}_2 \end{pmatrix} = n(\hat{W}_1^2 + \hat{W}_2^2) = Z_1^2 + Z_2^2,$$

which depends on two uncorrelated normal variates and is asymptotically χ_2^2. Note that δ_i, $\hat{W}_i$, and Z_i, $i = 1, 2$ all depend on the locus d. In the sequel, when necessary, we use $\delta_1(d)$, $\delta_2(d)$, $\hat{W}_i(d)$, and $Z_i(d)$, $i = 1, 2$ to emphasize their dependence on d.

To demonstrate the Ornstein-Uhlenbeck equivalence, the covariances at different loci d_1 and d_2 are proved to be

$$\mathrm{Cov}(Z_1(d_1),\ Z_1(d_2)) = 1 - \beta_1 r + O(r^2)$$

and

$$\mathrm{Cov}(Z_2(d_1),\ Z_2(d_2)) = 1 - \beta_2 r + O(r^2).$$

This means that for large n, $Z_1(d)$ and $Z_2(d)$ are approximately independent Ornstein-Uhlenbeck processes with mean zero and covariance $1 - \beta_1 r + O(r^2)$ and $1 - \beta_2 r + O(r^2)$, respectively. Adapting the argument in Dupuis and Siegmund (1999), the tail distribution of 2 LR under the null hypothesis satisfies

$$\begin{aligned} &P(\max_d 2\ \mathrm{LR}(d) \geq a^2) \\ &\quad \approx 1 - \exp\left\{-\left[C + v(a(2\beta\Delta)^{1/2})\,a^2 L\left(\frac{\beta_1 + \beta_2}{2}\right)\right]\exp\left(-\frac{a^2}{2}\right)\right\}, \end{aligned} \tag{1}$$

where Δ = the distance between markers (in Morgans), C = the number of chromosomes, and L = total length of the genome (in Morgans). The definition of $v(x)$ can

TABLE 2

95 and 99% critical LR threshold values

Δ[a] (cM)	Model[b]	Empirical[c]	Sparse map[d]	Dense map[e]
10	a	10.83 (14.19)[f]	9.85 (13.31)	12.33 (16.11)
	b	10.86 (14.32)	9.85 (13.31)	12.33 (16.11)
	c	10.24 (13.51)	9.84 (13.31)	12.31 (16.09)
	d	10.49 (14.25)	9.89 (13.34)	12.45 (16.22)
5	a	11.01 (14.50)	10.50 (14.05)	12.33 (16.11)
	b	11.02 (14.51)	10.50 (14.05)	12.33 (16.11)
	c	10.8 (14.50)	10.49 (14.04)	12.31 (16.09)
	d	10.9 (14.20)	10.56 (14.10)	12.45 (16.22)
2	a	11.66 (15.50)	11.13 (14.76)	12.33 (16.11)
	b	11.66 (15.51)	11.13 (14.76)	12.33 (16.11)
	c	11.38 (15.3)	11.11 (14.75)	12.31 (16.09)
	d	11.45 (15.3)	11.20 (14.83)	12.45 (16.22)

[a] Δ, the marker interval length. The length of chromosome is 100 cM.

[b] Sampled distributions of models a–d corresponding to BC1, F_2, and BC2 populations, respectively, are a, ($N(1, 1)$, $N(0, 1)$, $N(-1, 1)$); b, ($N(0, 1)$, $N(0, 1)$, $N(0, 1)$); c, ($N(1, 1)$, $N(0, 2)$, $N(-1, 1)$); d, ($N(1, 2)$, $N(0, 2)$, $N(-1, 1)$).

[c] The empirical thresholds are based on 5000 replications.

[d] Sparse map calculation of the theoretical threshold based on (1).

[e] Dense map calculation using (1) with $v = 1$ (*i.e.*, $\Delta = 0$).

[f] Thresholds at 95% (99%).

be found in Siegmund (1985). For dense maps, $v = 1$. Similarly, the power is given by

$$P(\max_d 2\ \mathrm{LR} \geq a^2) \approx 1 - \Phi(a - \omega^*) + \phi(a - \omega^*)$$

$$\times \left[\frac{1}{2\omega^*} + \frac{2\sqrt{a}v(a\{2\Delta\beta\}^{1/2})}{\omega^{*3/2}} - \frac{\sqrt{a}v(a\{2\Delta\beta\}^{1/2})^2}{\omega^{*1/2}(a + \omega^*)}\right]$$

when a QTL is located at a marker locus. Here

$$v = v(a\{2\Delta\beta\}^{1/2}),$$

$$\omega^* = \sqrt{n \log\left[1 + \frac{\xi_1^2\delta_1^2 + (\eta_1\delta_1 + \eta_2\delta_2)^2}{\sigma_{\mathrm{BC}}^2}\right]},$$

$$\omega_1^* = \omega^* \frac{\xi_1\delta_1}{\sqrt{\xi_1^2\delta_1^2 + (\eta_1\delta_1 + \eta_2\delta_2)^2}},$$

$$\omega_2^* = \omega^* \frac{\eta_1\delta_1 + \eta_2\delta_2}{\sqrt{\xi_1^2\delta_1^2 + (\eta_1\delta_1 + \eta_2\delta_2)^2}},$$

$$\beta = \frac{\beta_1\omega_1^{*2} + \beta_2\omega_2^{*2}}{\omega^{*2}}.$$

For a QTL between markers, the noncentrality parameters ω_1^* and ω_2^* are $\omega_1^* \exp(-\beta_1\Delta_1)$ and $\omega_2^* \exp(-\beta_2\Delta_1)$, respectively, where the distance between the QTL and the marker is Δ_1. In the case of an F_2 population, the formulas above reduce to those in Dupuis and Siegmund (1999).

General Results: The derivations above can be generalized to more complicated models, including those in Liu and Zeng (2000). Advanced crosses, such as F_x, $x \geq 2$, and models with covariates are also possible. Our framework can be modified for a wide variety of designs.

As before, let the model be

$$Y = Xb + e = (X_1\ X_2)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + e,$$

where X_1 is an $n \times p$ submatrix not involving the QTL effects, X_2 is an $n \times m$ matrix corresponding to the m QTL effects, and e is the random error. Following the procedure in the appendix, we first compute A_{22} using (A2) and then derive the orthogonal transformation matrix P from (A3). Note that both A_{22} and P involve only the design matrix X and not β or the correlation parameters. Next, 2 LR(d) can be partitioned into the sum of squares of m asymptotically independent $N(0, 1)$ random variables $Z_1, \ldots, Z_m$, where $Z = (Z_1, \ldots, Z_m)' = P\hat{b}_2$. To calculate $\mathrm{Cov}(Z_j(d_1), Z_j(d_2))$, $j = 1, 2, \ldots, m$, we find D in (A4) on the basis of the specific designs. It is straightforward to establish

$$\mathrm{Cov}(Z_j(d_1), Z_j(d_2)) = 1 - \beta_j r + O(r^2), \quad j = 1, 2, \ldots, m,$$

where β_j is the jth diagonal element of $-PA_{22}DA'_{22}P'$. Now, the tail distribution of 2 LR under the null hypothesis is approximately

$$P(\max_d 2\ \mathrm{LR}(d) \geq \alpha^2) \approx 1 - \exp\{-C[1 - \chi_m^2(\alpha^2)]$$

$$- v(a\{2\beta\Delta\}^{1/2})\beta L 2^{(2-m)/2}$$

$$\times [\Gamma(m/2)]^{-1} a^m \exp(-a^2/2)\},$$

TABLE 3

Simulation of chromosome-wise type I error

Δ (cM)	Model	$\alpha = 0.05$	$\alpha = 0.01$
10	a	0.078	0.016
	b	0.077	0.018
	c	0.060	0.012
	d	0.064	0.015
5	a	0.065	0.013
	b	0.063	0.012
	c	0.056	0.012
	d	0.057	0.011
2	a	0.062	0.014
	b	0.062	0.014
	c	0.056	0.013
	d	0.056	0.012

Models a–d are the same as in Table 2. The type I errors are calculated on the sparse map thresholds of the fourth column of Table 2.

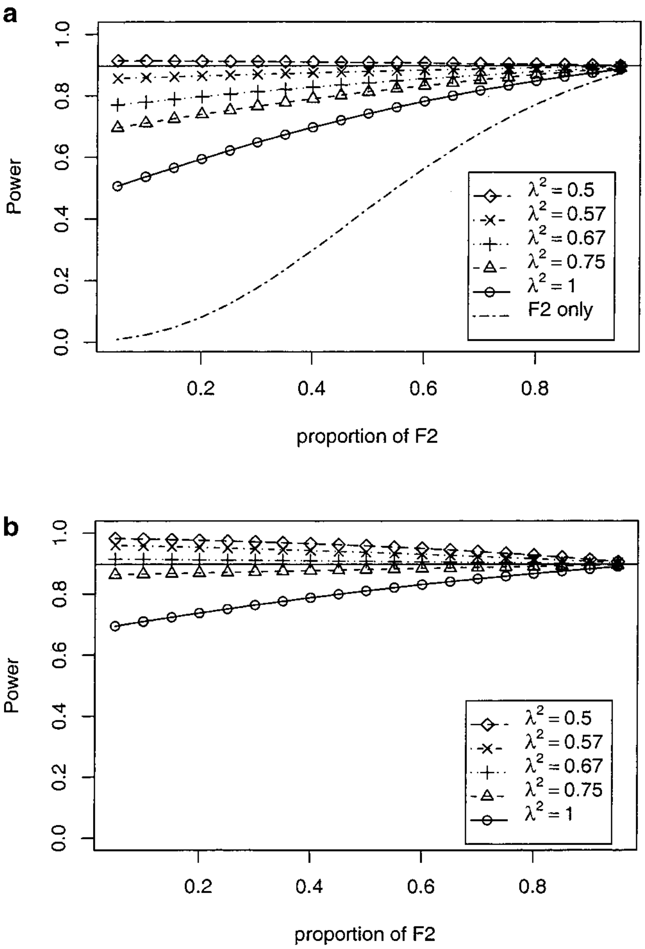

FIGURE 1.—(a) Additive QTL model. (b) Dominant QTL model. Power curves for equal BC1 and BC2 ratios under dense maps. The horizontal line is the power calculated from DUPUIS and SIEGMUND's (1999) formula (*i.e.*, the proportion of F_2 is 1). The lowest curve in a is the power when only available F_2's are used by discarding BCs. Thus power is reduced because there are fewer individuals.

where χ^2_m is a χ^2 random variable with m degrees of freedom and $\beta = (\beta_1 + \beta_2 + \ldots + \beta_m)/m$.

The formula for power may also be obtained. However, it is quite complicated and is omitted here.

SIMULATION STUDY OF THRESHOLDS AND POWER

We investigated the performance of (1) with different marker distances and different polygenic backgrounds. Thresholds for the log-likelihood were based on interval mapping with combined BC1, F_2, and BC2 crosses. $n_1 = n_2 = n_3 = 100$, giving 300 observations in total and chromosome length = 100 cM. The marker interval lengths are set at 10, 5, and 2 cM, respectively. Different polygenic effects are sampled, as reflected by models a–d (see legend of Table 2 for details; Table 3). The approximations from (1) with $v(a\{2\beta\Delta\}^{1/2})$ are always smaller than the empirical thresholds derived in the simulations. However, as the interval length decreases, our approximations are more similar to the empirical thresholds. In general, the dense map assumption ($v = 1$) produces conservative thresholds. Since more markers are likely to be typed around promising loci (LANDER and KRUGLYAK 1995), the stringent thresholds based on a dense map should be used even with a sparse map. Also, the approximations provide conservative control of the genome-wise type I error rate. Note that (1) gives upper and lower bounds for the threshold with $v = 1$ (assuming a dense map) and with $v(a\{2\beta\Delta\}^{1/2})$ (using the true map distances), respectively.

Next, we evaluate the power with different proportions of BC1, F_2, and BC2. The power is calculated for dominant ($\delta_1 = \delta_2$) and additive ($\delta_2 = 2\delta_1$) models. We compare our results with those of DUPUIS and SIEGMUND (1999) for the dominant model. We use the same values of the noncentrality parameter. In theory, as the proportion of F_2 approaches 1, our power approximation should agree with DUPUIS and SIEGMUND (1999). Other noncentrality parameters in DUPUIS and SIEGMUND (1999) show the same pattern and are omitted. For the additive model, the comparisons are qualitatively similar.

Figures 1 and 2 exhibit the power curves. When the polygenes are in linkage equilibrium and have only additive effects, the phenotypic variation due to polygenes and environment satisfies $\sigma^2_{BC1} = \sigma^2_{BC2} = \sigma^2_P + \sigma^2_e$ and $\sigma_{F2} = 2\sigma^2_P + \sigma^2_e$, respectively, where σ^2_P is the total polygenic variation in the BC population and σ^2_P is the environmental variation. For this reason, we take $\lambda_1 \equiv 1$ and choose $\lambda_2^{-1} = 1, 0.75, 0.67, 0.57, 0.5$, which correspond to $\sigma^2_P = 0, \sigma^2_e/2, \sigma^2_e, 3\sigma^2_e$, or $\sigma^2_P \gg \sigma^2_e$, respectively. We also evaluate the power by using F_2's only, which quantifies the loss in power when discarding data from the BCs (see Figure 1a).

In Figure 1, the proportions of BC1 and BC2 are assumed equal. When the QTL is dominant, power is

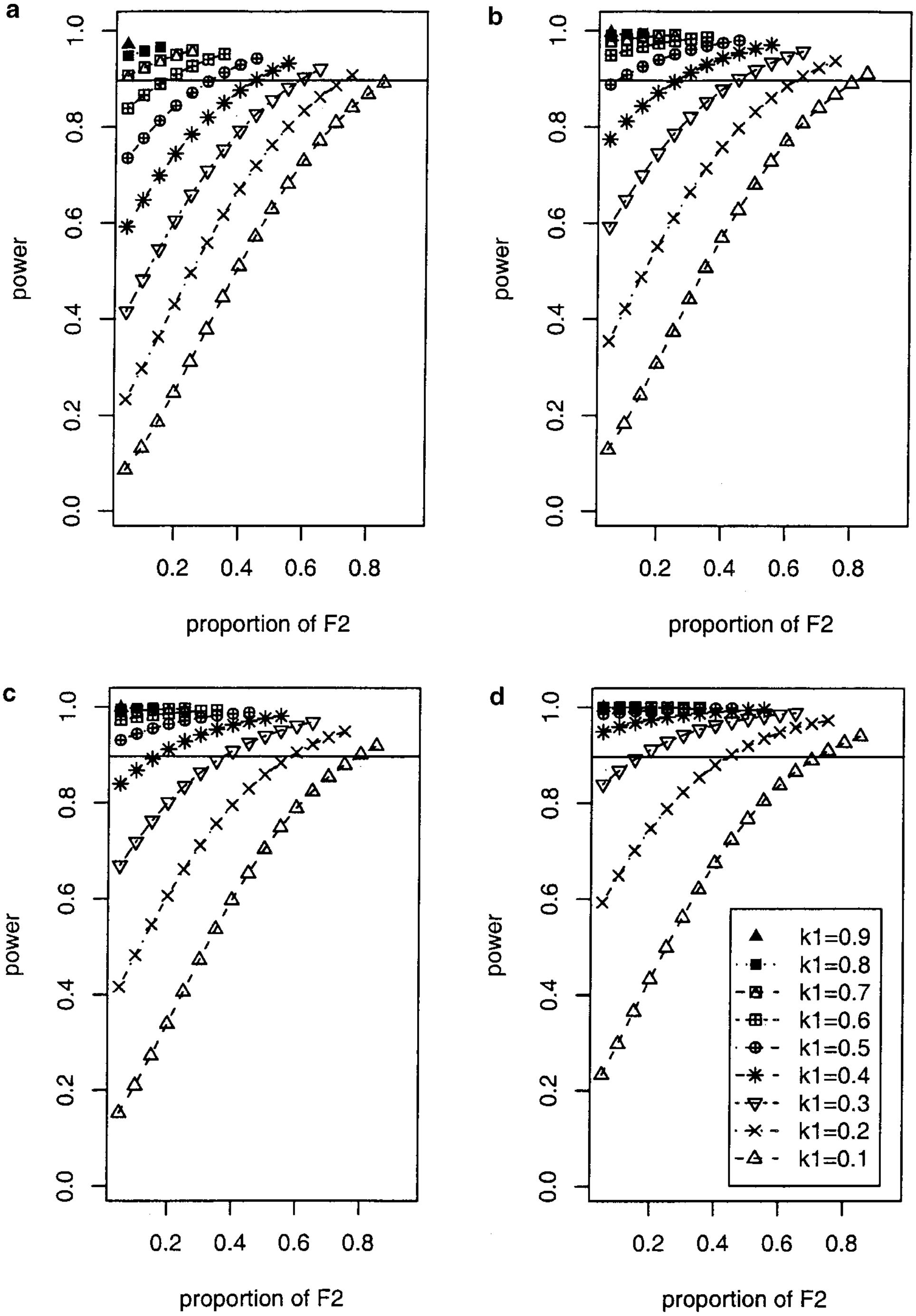

FIGURE 2.—Power curves for different ratios of BC1, BC2, and F_2 with dominant QTL under dense maps. Dominant model with (a) $\lambda^2 = 1$, (b) $\lambda^2 = 0.75$, (c) $\lambda^2 = 0.67$, (d) $\lambda^2 = 0.5$. a–d use the same graphical symbols. k1 is the sample proportion of BC1 and $\lambda^2 = \sigma^2_{BC2}/\sigma^2_{F2}$.

gained by using BC populations unless there is no polygenic effect (*i.e.*, λ_2 is close to 1). The larger the polygenic effects, the greater is the gain with BCs. However, when the QTL is additive, F_2's tend to have more information for detecting a QTL than do BCs, unless $\sigma_P \gg \sigma_e$ (*i.e.*, the polygenic effects are very large). Note that when the proportion of F_2 approaches 1, our results again match those of Dupuis and Siegmund.

In Figure 2, we allow the proportions of BC1 and BC2 to be unequal with a dominant QTL. In this case, BC1 is more powerful than F_2, which is expected. When the QTL is additive, both BC1 and BC2 individuals have identical contributions in detecting the QTL, so only the total proportion of BC1 and BC2 influences the power, as shown in Figure 1a.

CONCLUSION

In this article, we addressed some important practical issues in the analysis of closely related crosses derived from multiple inbred lines when both QTL and polygenes influence a trait. We showed that biased and inefficient estimates of the QTL effects may occur if the polygenic effect is ignored. We derived simple and general approximations for the threshold and power to detect a QTL, allowing different designs to be compared.

Based on our power calculations, we find that the F_2 population is more robust in detecting QTL than the two backcross populations. This confirms Liu and Zeng (2000). Thus if the goal is to detect the QTL, then using a large F_2 population is highly recommended. However, scientists may not be able to produce enough F_2 individuals or may for other reasons use different crosses. In this situation, analyzing all the data simultaneously is preferred. This strategy improves the power to detect major QTL. In addition, this is an opportunity to detect potential polygenic effects. The derivation of the threshold approximation is easily extended to other designs beyond the combination of BC1, F_2, and BC2. However, to our knowledge, the theoretical computation of thresholds involving multiple QTL is an open problem.

We thank anonymous reviewers for their critical reading of this manuscript. This research was supported in part by the U.S. Department of Agriculture Hatch project through the University of Wisconsin, College of Agricultural and Life Sciences.

LITERATURE CITED

Bernardo, R., 1994 Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. **34:** 20–25.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Davies, R. B., 1977 Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika **64:** 247–254.

Davies, R. B., 1987 Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika **74:** 33–43.

Doerge, R. W., Z-B. Zeng and B. S. Weir, 1997 Statistical issues in the search for genes affecting quantitative traits in experimental populations. Stat. Sci. **12:** 195–219.

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

Elston, R. C., 1990 Models for discrimination between statistical alternative modes of inheritance, pp. 41–55 in *Advances in Statistical Methods for Genetic Improvement for Livestock*, edited by D. Gianola and K. Hammond. Springer-Verlag, New York.

Fernando, R. L., and M. Grossman, 1989 Marker-assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21:** 467–477.

Fernando, R. L., C. Stricker and R. C. Elston, 1994 The finite polygenic mixed-model: an alternative formulation for the mixed-model of inheritance. Theor. Appl. Genet. **88:** 573–580.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait in line crosses using flanking markers. Heredity **69:** 315–324.

Jansen, R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple quantitative trait in line crosses using flanking markers. Heredity **69:** 315–324.

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

Lander, E. S., and L. Kruglyak, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. **11:** 241–247.

Liu, Y., and Z-B. Zeng, 2000 A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. Genet. Res. **75:** 345–355.

Piepho, H. P., 2001 A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics **157:** 425–432.

Rebai, A., B. Goffinet, B. Mangin and D. Perret, 1994a Detecting QTLs with diallel schemes, pp. 170–177 in *Biometrics in Plant Breeding: Applications of Molecular Markers. 9th Meeting of the EUCARPIA*, edited by J. W. van Ooijen and Jansen. CPRO-DLO, Wageningen, The Netherlands.

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

Rebai, A., B. Goffinet and B. Mangin, 1995 Comparing power of different methods for QTL detection. Biometrics **51:** 87–99.

Siegmund, D., 1985 *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York.

Yandell, B. S., 1997 *Practical Data Analysis for Designed Experiments.* Chapman & Hall/CRC Press, London/Cleveland.

Zeng, Z-B., 1983 Theoretical basis of separation of multiple link gene effects on mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972–10976.

Zeng, Z-B., 1994 Precision mapping of quantitative traits loci. Genetics **136:** 1457–1468.

Communicating editor: Z-B. Zeng

APPENDIX

In this section, for combined crosses from two inbred parents, we prove that the likelihood ratio 2 LR(d) can be partitioned into the sum of the squares of two asymptotically independent Ornstein-Uhlenbeck processes through an orthogonal transformation. Define

$$B = \lim_{n\to\infty} \frac{\tilde{X}(d)'\tilde{X}(d)}{n} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} \tilde{X}_1\tilde{X}_1 & \tilde{X}_1\tilde{X}_2 \\ \tilde{X}_2\tilde{X}_1 & \tilde{X}_2\tilde{X}_2 \end{pmatrix}. \quad \text{(A1)}$$

Note that B does not depend on locus d. Let

$$A = B^{-} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad \text{with } A_{22} = \begin{pmatrix} a_1 & a_3 \\ a_3 & a_2 \end{pmatrix}$$

$$C = A_{22}^{-1} = \begin{pmatrix} c_1 & c_3 \\ c_3 & c_2 \end{pmatrix}. \quad \text{(A2)}$$

Define

$$P = \begin{pmatrix} \xi_1 & 0 \\ \eta_1 & \eta_2 \end{pmatrix}; \quad \text{then } PA_{22}P' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{(A3)}$$

where $\xi_1 = \sqrt{c_1 - c_3^2/c_2}$, and $\eta_1 = c_3/\sqrt{c_2}$, $\eta_2 = \sqrt{c_2}$.

Making the orthogonal transformation

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = P\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix}$$

gives

$$\begin{aligned} \lim_{n\to\infty} \mathrm{Var}(Z_1) &= \lim_{n\to\infty} n\,\mathrm{Cov}\left\{(\xi_1, 0)\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix}, (\xi_1, 0)\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix}\right\} \\ &= \lim_{n\to\infty} n(\xi_1, 0)\mathrm{Var}\left\{\begin{bmatrix}\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix}\end{bmatrix}\right\}\begin{pmatrix} \hat{\xi}_1 \\ 0 \end{pmatrix} \\ &= (\xi_1, 0)(HAH')\begin{pmatrix} \hat{\xi}_1 \\ 0 \end{pmatrix} = 1. \end{aligned}$$

For the same reason,

$$\lim_{n\to\infty} \mathrm{Var}(Z_2) = (\eta_1, \eta_2)(HAH')\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = 1$$

and

$$\lim_{n\to\infty} \mathrm{Cov}(Z_1, Z_2) = (\xi_1, 0)(HAH')\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = 0.$$

This indicates that Z_1 and Z_2 are approximately independent $N(0, 1)$. Furthermore,

$$\begin{aligned} 2\ \mathrm{LR}(d) &\approx n(\hat\delta_1\ \hat\delta_2)(A_{22})^{-1}(\acute\delta_1\ \hat\delta_2)' \\ &= (\hat\delta_1\,\hat\delta_2)\begin{pmatrix}\xi_1 & \eta_1\\ 0 & \eta_2\end{pmatrix}\left(\begin{pmatrix}\xi_1 & 0\\ \eta_1 & \eta_2\end{pmatrix}A_{22}\begin{pmatrix}\xi_1 & \eta_1\\ 0 & \eta_2\end{pmatrix}\right)^{-1}\begin{pmatrix}\xi_1 & 0\\ \eta_1 & \eta_2\end{pmatrix}\begin{pmatrix}\hat\delta_1\\ \hat\delta_2\end{pmatrix} \\ &= (\acute\delta_1\,\hat\delta_2)\begin{pmatrix}\xi_1 & \eta_1\\ 0 & \eta_2\end{pmatrix}\begin{pmatrix}\xi_1 & 0\\ \eta_1 & \eta_2\end{pmatrix}\begin{pmatrix}\hat\delta_1\\ \hat\delta_2\end{pmatrix} \\ &= n(\hat W_1\ \hat W_2)\begin{pmatrix}\hat W_1\\ \hat W_1\end{pmatrix} = n(\hat W_1^2 + \hat W_2^2) = Z_1^2 + Z_2^2. \end{aligned}$$

Now, let $\tilde X(d_1)$ and $\tilde X(d_2)$ be the corresponding transformed incidence matrices at loci d_1 and d_2. Note that the first and second columns of $\tilde X$ depend only on the proportions of BC1, F_2, and BC2 and not on d_1 and d_2. The third and fourth columns depend on d_1 and d_2. Also $x_{1ki}(d) = 0, 1,$ or 0 and $x_{2ki}(d) = 0, 0,$ or 1 if individual i in cross K's genotype at locus d is qq, Qq, or QQ, respectively.

For BC1,

$$(x_{1ki}(d_1), x_{1ki}(d_2)) = \begin{cases}(0,0) & \text{with probability } (1-r)/2\\ (0,1) \text{ or } (1,0) & \text{with probability } r\\ (1,1) & \text{with probability } (1-r)/2\end{cases}$$

$$(x_{1ki}(d_1), x_{2ki}(d_2)) = \begin{cases}(0,0) & \text{with probability } 1/2\\ (1,0) & \text{with probability } 1/2,\end{cases}$$

where r is the recombination fraction between loci d_1 and d_2. Enumerating the probabilities of $(x_{1ki}(d_1), x_{1ki}(d_2))$, $(x_{1ki}(d_1), x_{2ki}(d_2))$, and $(x_{2ki}(d_1), x_{2ki}(d_2))$ for BC1, F_2, and BC2, and using the fact that $\tilde X(d) = G^{-1/2}X(d)$, we obtain

$$\begin{aligned} &\lim_{n\to\infty}\frac{\tilde X'(d_1)\tilde X(d_2)}{n} \\ &= \begin{pmatrix}\tilde X_1'\tilde X_1 & \tilde X_1'\tilde X_2\\ \tilde X_2'\tilde X_1 & \tilde X_2'\tilde X_2 + \begin{pmatrix}-\frac{\lambda_1k_1 + 2\lambda_2k_2 + k_3}{2} & \frac{\lambda_2k_2 + k_3}{2}\\ \frac{\lambda_2k_2 + k_3}{2} & -\frac{\lambda_2k_2 + k_3}{2}\end{pmatrix}r + O(r^2)\end{pmatrix} \\ &= B + \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & D\end{pmatrix}r + O(r^2), \end{aligned} \tag{A4}$$

with

$$D = \begin{pmatrix}-\frac{\lambda_1k_1 + 2\lambda_2k_2 + k_3}{2} & \frac{\lambda_2k_2 + k_3}{2}\\ \frac{\lambda_2k_2 + k_3}{2} & -\frac{\lambda_2k_2 + k_3}{2}\end{pmatrix}.$$

Thus,

$$\begin{aligned} \mathrm{Cov}(\hat b(d_1), \hat b(d_2)) &= \mathrm{Cov}[(\tilde X(d_1)'\tilde X(d_1))^{-1}\tilde X(d_1)'y, \\ &\qquad (\tilde X(d_2)'\tilde X(d_2))^{-1}\tilde X(d_2)'y] \\ &\approx \frac{A\ \mathrm{Cov}(\tilde X(d_1)'y, \tilde X(d_2)'y)A}{n^2} \\ &= \frac{1}{n^2}A\tilde X(d_1)'\tilde X(d_2)A \\ &= \frac{A}{n} + \frac{A\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & D\end{pmatrix}A}{n}r + \frac{O(r^2)}{n}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{Cov}(Z_1(d_1), Z_1(d_2)) &= n\ \mathrm{Cov}(\hat W_1(d_1), \hat W_1(d_2)) \\ &= n(\xi_1, 0)H\ \mathrm{Cov}(\hat b(d_1), \hat b(d_2))H'\begin{pmatrix}\xi_1\\ 0\end{pmatrix} \\ &= 1 + \left[a_1^2\left(-\frac{\lambda_1k_1 + 2\lambda_2k_2 + k_3}{2}\right) + 2a_1a_3\frac{\lambda_2k_2 + k_3}{2}\right. \\ &\qquad \left. + a_3^2\left(-\frac{\lambda_2k_2 + k_3}{2}\right)\right]\xi_1^2r + O(r^2) \\ &= 1 - \beta_1r + O(r^2) \end{aligned}$$

and

$$\begin{aligned} \mathrm{Cov}(Z_2(d_1), Z_2(d_2)) &= n\ \mathrm{Cov}(\hat W_2(d_1), \hat W_2(d_2)) \\ &= n(\eta_1, \eta_2)H\ \mathrm{Cov}(\hat b(d_1), \hat b(d_2))H'\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} \\ &= 1 + \left[a_1^2\left(-\frac{\lambda_1k_1 + 2\lambda_2k_2 + k_3}{2}\right) + 2a_1a_3\frac{\lambda_2k_2 + k_3}{2}\right. \\ &\qquad \left. + a_3^2\left(-\frac{\lambda_2k_2 + k_3}{2}\right)\right]\eta_1^2r \\ &\quad + \left[a_3^2\left(-\frac{\lambda_1k_1 + 2\lambda_2k_2 + k_3}{2}\right) + 2a_3a_2\frac{\lambda_2k_2 + k_3}{2}\right. \\ &\qquad \left. + a_2^2\left(-\frac{\lambda_2k_2 + k_3}{2}\right)\right]\eta_2^2r \\ &\quad + 2\left[a_1a_3\left(-\frac{\lambda_1k_1 + 2\lambda_2k_2 + k_3}{2}\right) + (a_3^2 + a_1a_2)\frac{\lambda_2k_2 + k_3}{2}\right. \\ &\qquad \left. + a_3a_2\left(-\frac{\lambda_2k_2 + k_3}{2}\right)\right]\eta_1\eta_2r + O(r^2) \\ &= 1 - \beta_2r + O(r^2). \end{aligned}$$