Published online March 28, 2005

Selective Phenotyping to Accurately Map Quantitative Trait Loci

J.-L. Jannink*

ABSTRACT

Recombination events are necessary to map QTL accurately and some progeny result from gametes with a greater number of recombination events than others. This observation suggests that some progeny should be more useful for QTL mapping than others. The marker genotypes of the progeny should allow the number of recombination events they carry to be determined such that the most useful progeny could be phenotyped, in a procedure termed selective phenotyping. Two methods to select genotypes for their usefulness in mapping are described, one that maximizes the overall mapping information content in the selected progeny, and one that seeks to maximize both overall mapping information and the uniformity of its distribution across the genome. Simulations showed that both methods successfully decreased the mean squared error (MSE) for QTL position. Average MSEs were similar for the two methods and variability of MSE was slightly lower for the latter relative to the former method. Simulations indicated that a large fraction of the decrease in the MSE achievable by selective phenotyping could be obtained by genotyping twice the number of progeny than were ultimately phenotyped, though further decreases in the MSE were observed when up to 16 times more progeny were genotyped than phenotyped. The procedure appears to most improve the accuracy of QTL mapping for QTL of small effect or when available markers do not allow marker spacing below 10 cM.

Darvasi and Soller (1995) proposed the advanced intercross line (AIL) population to "overcome the lack of sufficient recombinational events in small chromosomal regions" for the purpose of fine mapping. An AIL population is generated from a standard F_2 mapping population by randomly mating that population for a desired number of generations. This random mating decreases the linkage disequilibrium between a QTL and all but the most tightly linked markers, allowing the QTL to be mapped more accurately. Darvasi and Soller (1995) showed that eight generations of random mating allowed the position of QTL to be mapped three to five times more accurately in an AIL than in the initial F_2 population, though the AIL required more densely spaced markers. In practice, AIL populations have been developed in mice (*Mus musculus* L.) (e.g., Iraqi et al., 2000) and in maize (*Zea mays* L.) (e.g., Lee et al., 2002). These species are amenable to the development of AIL because they are relatively easy to cross and have short generation times.

To increase the accuracy of QTL mapping in species that do not share these characteristics, I propose taking advantage of the fact that, from a statistical point of view, recombination is a random event. A consequence of this fact is that in any given sample of recombinant progeny, the number of recombination breakpoints will not be uniform among progeny but will follow a distribution. Identifying progeny with a high number of recombination breakpoints to form a mapping population could therefore allow QTL positions to be mapped more accurately. The identification of such progeny will require genotype information from DNA markers. The idea, therefore, is to genotype a large number of recombinant progeny but retain only the optimal set for phenotyping. For this idea to be feasible, one must assume that phenotyping costs, rather than genotyping costs, are a major experimental constraint limiting the ability of researchers to develop data sets sufficient to identify QTL. While this assumption would have been farfetched only a few years ago, biotechnological developments have decreased the cost of genotyping much more rapidly than the cost of phenotyping (e.g., Jobs et al., 2003). Second, for species that are long-lived, phenotyping costs will include the maintenance of the progeny until such time as the traits of interest (e.g., yield and quality of harvestable fruit) can be measured, potentially a substantial cost. On the other hand, genotyping could be performed at an early life stage, allowing many progeny to be culled and thereby avoiding the costs of raising them. Third, quite expensive phenotypes can be envisioned, including obtaining microarray expression data for an individual at a number of time points. Finally, if the mapping population becomes a resource used by many researchers, the cumulative effort invested in phenotyping may be increased many fold.

I call the use of DNA marker information to select an optimal set of progeny before phenotyping *selective phenotyping*, in reference to the complementary idea of selective genotyping also introduced by Darvasi and Soller (1992). Darvasi (1998) proposed selective phenotyping in relation to a single marker interval known to contain a QTL. In that context, progeny would be genotyped only at flanking markers, and only progeny recombinant in that interval would be phenotyped. Here, I extend this idea to the whole genome. The objective of this study was to evaluate the potential of selective phenotyping to improve the accuracy of QTL mapping in whole genome scans, and to explore different methods for selecting the optimal set of progeny.

MATERIALS AND METHODS

Selection of Optimal Recombinant Progeny

The two methods tested here for selecting optimal sets of progeny assume that there are no marker errors, that recombinants can be unambiguously identified, and that a progeny

Dep. of Agronomy, Iowa State Univ., 1208 Agronomy Hall, Ames, IA 50011-1010. Received 4 May 2004. Genomics, Molecular Genetics & Biotechnology. *Corresponding author (jjannink@iastate.edu).

Published in Crop Sci. 45:901–908 (2005).
doi:10.2135/cropsci2004.0278

677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: AIL, advanced intercross line; ANOVA, analysis of variance; cM, centimorgan; MCMC, Markov chain Monte Carlo; MSE, mean squared error; QTL, quantitative trait locus.

can have only a single observable recombination event in any given marker interval. Progeny types that satisfy this requirement include backcross, doubled-haploid, and recombinant inbred progeny. In the case of F_2 progeny, there is the possibility of observing two recombination events in a single interval if both gametes inherited by the F_2 progeny were recombinant. Extensions to deal with F_2 progeny should be straightforward but were not tested here. The rationale underlying these methods is that a recombination event represents a quantum of information for the mapping of QTL.

Two methods for selecting progeny for phenotyping were evaluated. In the maximum number of recombinations method (denoted maxRec below), marker scores are available for M progeny. For each progeny j, define $c_{ij} = 1$ if j is recombinant in marker interval i, and $c_{ij} = 0$ otherwise. The number of recombinant marker intervals for progeny j is counted, $c_{\bullet j} = \sum_i c_{ij}$, and the N progeny with the highest $c_{\bullet j}$ are selected for phenotyping and analysis. While the maxRec method seeks solely to maximize the overall mapping information available among the selected progeny, the uniform number of recombinations method (denoted uniRec) seeks to maximize both the mapping information and the uniformity of its distribution across the genome. The rationale for this approach follows from the assumption that researchers do not know a priori where QTL are located so that it is desirable to have mapping information evenly distributed. In the maxRec method, the distribution of that information depends directly on the distribution of recombination events occurring in that interval in the M original progeny. The uniRec method I propose here is a simple way to meet the two objectives of maximizing total information and its uniformity. Other more optimal methods may exist.

Define $d_{ij} = c_{ij}/m_i$, where m_i is the map distance in centimorgans between the markers flanking interval i. The variable d_{ij} represents the amount of mapping information for interval i conferred by progeny j on a per map unit basis. The uniRec method is iterative. First, the single progeny with the highest $d_{\bullet j} = \sum_i d_{ij}$ is selected to form the initial set S of selected progeny. The following steps are then iterated until N progeny are selected:

1. The amount of mapping information available in S for each marker interval i is calculated as the sum across individuals j in S of d_{ij}: $d_{i\bullet} = \sum_{j \in S} d_{ij}$;
2. The maximum mapping information in S across all marker intervals is calculated as $p = \max_i(d_{i\bullet})$;
3. For all unselected progeny $j (j \notin S)$ a score is calculated $t_j = \sum_i (p - d_{i\bullet}) c_{ij}$;
4. The progeny with the highest t_j is selected for inclusion in S. If several progeny have the same value for t_j, one of them is selected at random. Inclusion of a progeny into S changes the $d_{i\bullet}$ values so that the method must start again at step 1 to determine the next progeny to include.

The value $p - d_{i\bullet}$ represents the value added to S by including a progeny that is recombinant in interval i. If S already contains much mapping information for interval i ($d_{i\bullet}$ is high) then including a progeny that is recombinant for that interval adds less value than if S contains little mapping information for interval i ($d_{i\bullet}$ is low). The score t_j represents the sum across marker intervals of the value of including j in S.

Simulations of Mapping Populations and Analyses

The potential of the uniRec and maxRec methods to improve the accuracy of estimates of QTL position was tested by simulation. For simplicity, I used progeny from a backcross design, in which two inbred parents were crossed and their F_1 progeny was backcrossed to one of the parents to create segregating progeny. The simulated genome contained six chromosomes, each 100 cM long. Molecular marker data was simulated for these progeny assuming evenly spaced markers, with marker spacings of 10 cM on chromosomes 1, 2, and 3 and 20 cM on chromosomes 4, 5, and 6. Recombination across marker intervals was assumed independent, as in Haldane's mapping function (Haldane, 1919). Plant geneticists often use Kosambi's mapping function (Kosambi, 1944). Empirical evidence, however, is lacking for the Kosambi assumption of increasing positive crossover interference as marker spacing becomes smaller. Indeed, negative crossover interference over short distances has been observed (Esch and Weber, 2002; Rong et al., 2004; Sherman and Stack, 1995), a condition under which Haldane's function would fit better than Kosambi's. Given that neither Haldane's nor Kosambi's assumptions appear to be empirically justified, I chose Haldane's, which is the simpler of the two.

A single QTL was simulated at the center of the marker interval closest to the center of each chromosome (resulting in positions of 45 cM and 50 cM from the beginning of the chromosome for chromosomes with 10 cM and 20 cM marker spacings, respectively). The QTL explained 7, 11, and 17% of the phenotypic variance for chromosomes 1 and 4, 2 and 5, and 3 and 6, respectively. Thus, the simulation allowed for a full factorial of two marker spacings and three QTL effects. The total genetic variance over the six chromosomes was 70% of the phenotypic variance.

The uniRec and maxRec progeny selection methods were applied to obtain either 100 or 200 progeny for phenotyping and analysis. That is, the value N was set at either 100 or 200. The original number of genotyped progeny M was set anywhere from 100 to 3200 such that the selected fraction N/M of progeny ranged over five values: 1, 1/2, 1/4, 1/8, or 1/16. The selected fraction of 1 corresponds to no selection, that is, all genotyped progeny were also phenotyped and the data analyzed. The other selected fractions correspond to progressively more stringent selection of which progeny to phenotype, with a maximum of 16 times more progeny genotyped than phenotyped. In total there were 20 simulation settings: two selection methods times two levels for the number of progeny analyzed times five levels of selected fractions.

The set of progeny selected was then analyzed using a Bayesian mapping analysis (Satagopan et al., 1996; Sillanpää and Arjas, 1998). The analysis assumed the presence of a single QTL on each of the six chromosomes. The model of the phenotype for progeny j was:

$$y_j = \mu + \sum_{k=1}^{6} x_{kj}\alpha_j + \varepsilon_j$$

where μ is the population mean, x_{kj} is an indicator of the QTL genotype ($x_{kj} = 0$ for a heterozygote or $x_{kj} = 1$ for a homozygote recurrent parent) for QTL k and progeny j, α_k is the effect of QTL k, and ε_j is a residual distributed as $\varepsilon_j \sim N(0, \sigma^2)$. The Bayesian analysis was implemented using Markov-chain Monte Carlo (MCMC) to obtain posterior distributions of the parameters, with particular interest being paid to the posterior distributions of the QTL location parameters, λ_k. Briefly, the parameters μ, σ^2, and α_k were updated using a simple random-walk Metropolis–Hastings procedure (Gilks et al., 1996), as described in more detail in (Jannink and Wu, 2003), assuming positive improper uniform priors. The QTL location λ_k and progeny QTL genotypes x_{kj} were blocked and updated jointly as follows. The proposal distribution for a new location, $p(\lambda_k^* | \lambda_k)$ was uniform over the interval [max(0, $\lambda_k - b$), min(100,

$\lambda_k + b)$]. New QTL genotypes $x_k^* = \{x_{k1}^*,..., x_{kj}^*,..x_{kN}^*$ were then proposed from their full conditional distribution $p(x_k^*|y_j, \mu, \sigma^2, \boldsymbol{\alpha}, \mathbf{x}_{-\mathbf{k}}, \mathbf{m})$, where $\boldsymbol{\alpha}$ is the vector of all α_k, $\mathbf{x}_{-\mathbf{k}}$ is the matrix of all QTL genotypes except those at QTL k, and $\mathbf{m}$ is the matrix of marker genotypes. The joint update of λ_k^* and x_k^* was then accepted with probability

$$\min\left[1, \frac{p(\mathbf{y}|\theta^*)p(\mathbf{x_k^*}|\mathbf{m},\lambda_k^*)p(\lambda_k|\lambda_k^*)p(\mathbf{x_k}|\mathbf{y},\mu,\sigma^2,\boldsymbol{\alpha},\mathbf{x}_{-\mathbf{k}},\mathbf{m},\lambda_k)}{p(\mathbf{y}|\theta)p(\mathbf{x_k}|\mathbf{m},\lambda_k)p(\lambda_k^*|\lambda_k)p(\mathbf{x_k^*}|\mathbf{y},\mu,\sigma^2,\boldsymbol{\alpha},\mathbf{x}_{-\mathbf{k}},\mathbf{m},\lambda_k^*)}\right]$$

where $p(\mathbf{x_k}|\mathbf{m}, \lambda_k)$ is the prior distribution for QTL genotypes, which depends on the QTL location and the marker genotypes. This same update method is given by Yi and Xu (2001) in the context of QTL mapping under complex mating designs. The prior for QTL position was uniform over its chromosome. The Markov chain was burned in for 200 iterations and then run for 5000 iterations. New data sets and complete Markov chains were simulated 2500 times for each of the 20 simulation settings.

The measure of QTL mapping accuracy I chose was the posterior MSE relative to the simulated QTL position. Assuming a single QTL on a chromosome of 100 cM, the posterior MSE is calculated as $\text{MSE} = \int_0^{100}(\lambda - \Lambda)^2\pi(\lambda)d\lambda$ where Λ is the simulated QTL position, λ is the estimate of that position, and $\pi(\lambda)$ is the posterior density of λ. If there is no information in the data relative to the QTL position, then the posterior is the same as the prior [$\pi(\lambda) = p(\lambda)$]. In that case, for a QTL simulated at the center of a 100 cM chromosome, the posterior MSE would equal 833 $(\text{cM})^2$. The MSE was calculated in practice using the Markov chain samples as

$$\text{MSE} = \frac{1}{5000}\sum_{s=1}^{5000}(\hat{\lambda}_s - \Lambda)^2$$

where $\hat{\lambda}_s$ was the Markov chain sample for QTL location from iteration s. To ensure that 5000 MCMC iterations were enough to estimate the MSE, I also ran 125 simulations with Markov chains of 100 000 iterations for the uniRec method with N = 200 and a selected fraction of 0.5. I found no evidence that short runs were biased relative to long runs: mean MSE for the short runs was not significantly different from mean MSE for the long runs. Variance among short run MSE was also not significantly different from variance among long-run MSE (data not shown). To determine error in MSE estimates caused by MCMC sampling, I used the "batch means" method (Roberts, 1996), breaking down the long runs into 20 batches of 5000 iterations each. For estimating MSE from a simulated data set, MCMC sampling error as a fraction of the variance among estimates from independently simulated data ranged from less than 1% for QTL of large effect to about 9% for QTL of smaller effect. These small MCMC sampling errors suggested that chains of 5000 iterations were sufficient to adequately estimate the MSE.

Differences among selection methods and the effects of the number of progeny analyzed, the selected fraction, the marker spacing, and the QTL effects were assessed by a full factorial ANOVA. The selected fraction of 1 was not included in this ANOVA because maxRec and uniRec methods were identical for it. The factorial of five main effects led to 96 simple effects: 2 selection methods × 2 progeny numbers × 4 selected fractions × 2 marker spacings × 3 QTL effects. All effects were assumed to be fixed. A log transform was found to best normalize MSE values so that analyses were conducted and averages obtained on a log scale. Averages reported here were backtransformed from the log scale. Variances across replicate simulations in the MSE were calculated, resulting in a single variance for each of the 96 simple effects. Treatment effects on these variances were assessed by factorial ANOVA, but assuming that all four-way and five-way interactions could be folded into the error term. This led to an ANOVA model with 29 degrees of freedom in the error and 66 model degrees of freedom.

To explore the interaction between QTL effect size, marker spacing, and the selected fraction, a set of simulations with a similar genome of six chromosomes but with QTL of the same effect placed at the center of each chromosome was run. Each chromosome, however, had different marker spacings, with markers every 25, 20, 10, 5, 4, and 2 cM over the chromosomes. Analyses were run with 200 phenotyped progeny under no progeny selection versus a selected fraction of 1/16 under the uniRec method. The effects of the QTL were 11.6% of the phenotypic variance in one set of simulations, and 5% of the phenotypic variance in the other set of simulations, for respective total genetic variances of 70 and 30% of the phenotypic variance in these two sets of simulations.

RESULTS AND DISCUSSION

Increase in the Number of Recombinations Using maxRec and uniRec

Under the assumption that underlies Haldane's mapping function that recombination events are independent across the genome, the number of recombinations for a progeny is Poisson distributed with mean and variance equal to the map size of the genome in Morgans. The only way to identify all such recombinations, however, would be to have very tight marker coverage over the whole genome. Without such coverage, the probability that a double recombination will go undetected within a marker interval reduces the mean and variance for the number of recombinations identified. For any single marker interval, the number of recombination events follows a Bernoulli distribution with a success probability equal to the recombination frequency over the interval. If R is the number of recombinants detected for a single progeny, then the expectation and variance for R under Haldane's mapping function are

$$E(R) = \sum_i \frac{1}{2}(1 - e^{-2m_i}) = \sum_i f_i$$

and

$$\text{var}(R) = \sum_i f_i(1 - f_i)$$

where f_i is the recombination frequency in interval i. These sums do not follow a standard distribution though if all marker intervals in the genome are of the same size, the distribution will be binomial, with success probability equal to f_i and number of trials equal to the number of intervals. Again, if there are many small marker intervals, R will follow a Poisson distribution with mean and variance parameter equal to the genome size in Morgans. This relationship between the distribution of R and the Poisson distribution has a bearing on how well selective phenotyping will work with organisms of different map sizes. In particular, $E(R)$ and $\text{var}(R)$ will increase linearly with the map size of the organism studied. Gains obtained from selective phenotyping in the total number of recombinants, however, will increase linearly with the standard deviation of R,

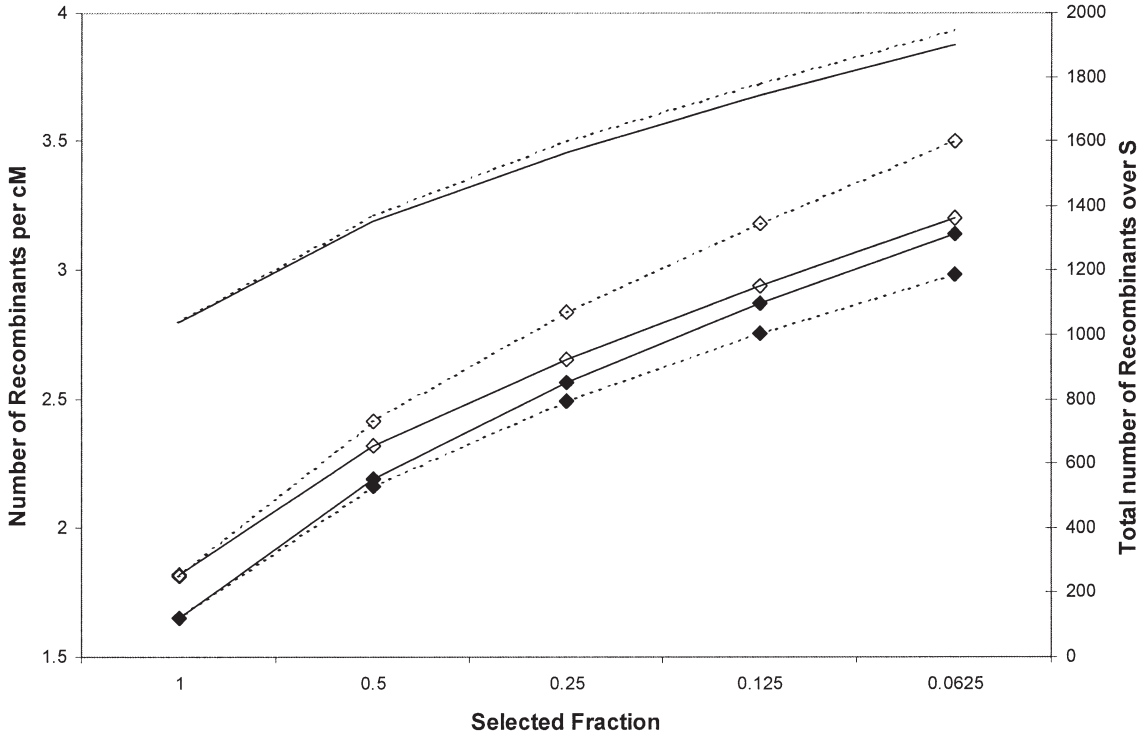

Fig. 1. Average number of recombinants per cM in 10- and 20-cM marker intervals (lines with symbols, left scale), and total number of recombinants over all selected genotypes (lines without symbols, right scale) as affected by the selected fraction of genotypes. Continuous lines are for the uniRec method and dotted lines are for the maxRec method of selection. Closed symbols are for 20-cM marker spacings and open symbols are for the 10-cM spacing. A selected fraction of 1 indicates no selection. Average number of recombinants for 10- and 20-cM intervals calculated using three chromosomes for each spacing.

and therefore with the square root of the map size of the organism. Finally, to predict gains possible in the number of recombinants *per map unit*, we need to divide the standard deviation of R by the organism's map size, yielding a ratio that will be inversely proportional to the square root of the map size. Consequently, it seems reasonable to predict that selective phenotyping will be most effective for organisms with small map sizes.

Of interest in QTL mapping is total number of recombinant intervals over the set S of progeny analyzed, $R_S = \Sigma_S R$. Generally R_S will also not follow a standard distribution, but since it is a sum over many independent random variables, it will be approximately normal. In the case of the simulations performed here, there were 30 intervals of 10 cM each, and 15 intervals of 20 cM each, such that $E(R) = 5.19$ and $\text{var}(R) = 4.54$. This expected number of detected recombinants of 5.19 is less than the genome size of 6 Morgans because double recombinants in some intervals would not be detected. Thus, if 200 progeny are obtained without selection, $E(R_S) = 200(5.19) = 1038$. Given that maxRec is a simple directional selection method for high numbers of recombinations, the number of recombinations in S can be predicted using standard selection theory (Falconer and Mackay, 1996). These predictions are 1378, 1580, 1740, and 1876 for the 0.5, 0.25, 0.125, and 0.0625 selected fractions, respectively. In practice, using maxRec and averaged over the 2500 replicate simulations with $N = 200$, I obtained 1372, 1600, 1781, and 1945 recombinations for those respective selected fractions (Fig. 1). The deviations from the predictions arise from differences between the actual distribution of R_S and the normal distribution used to obtain the predictions (standard errors for the observed averages were below 1).

As might be expected, the sets of progeny selected by the uniRec method had fewer recombinations than those selected by the maxRec method, though the difference was not great (Fig. 1). The uniRec method captured 94 to 95% of the increase in recombinations achieved by the maxRec method. The number of recombinations observed with $N = 200$ were 1352, 1565, 1744, and 1903 for 0.5, 0.25, 0.125, 0.0625 selected fractions, respectively. The uniRec method was, however, effective at equalizing the number of recombinations per centimorgan observed across intervals of different lengths (Fig. 1). That is, the differences between intervals with 10- and 20-cM marker spacings in recombinations per centimorgan decreased with increasing selection pressure (Fig. 1). For a given possible set of selected progeny, the uniRec method assesses the number of recombinants available in a marker interval on a per map unit basis. Consequently, relative to the maxRec method, it selected sets of progeny with more recombinants in longer (20 cM) intervals, and fewer recombinants in shorter (10 cM) intervals (Fig. 1). Also as expected, the two methods differed in terms of their effects on the distribution across the genome of recombinations in the selected sets of progeny. With the maxRec method, the variance across intervals and repeated simulations in the number of recombinations per centimorgan increased

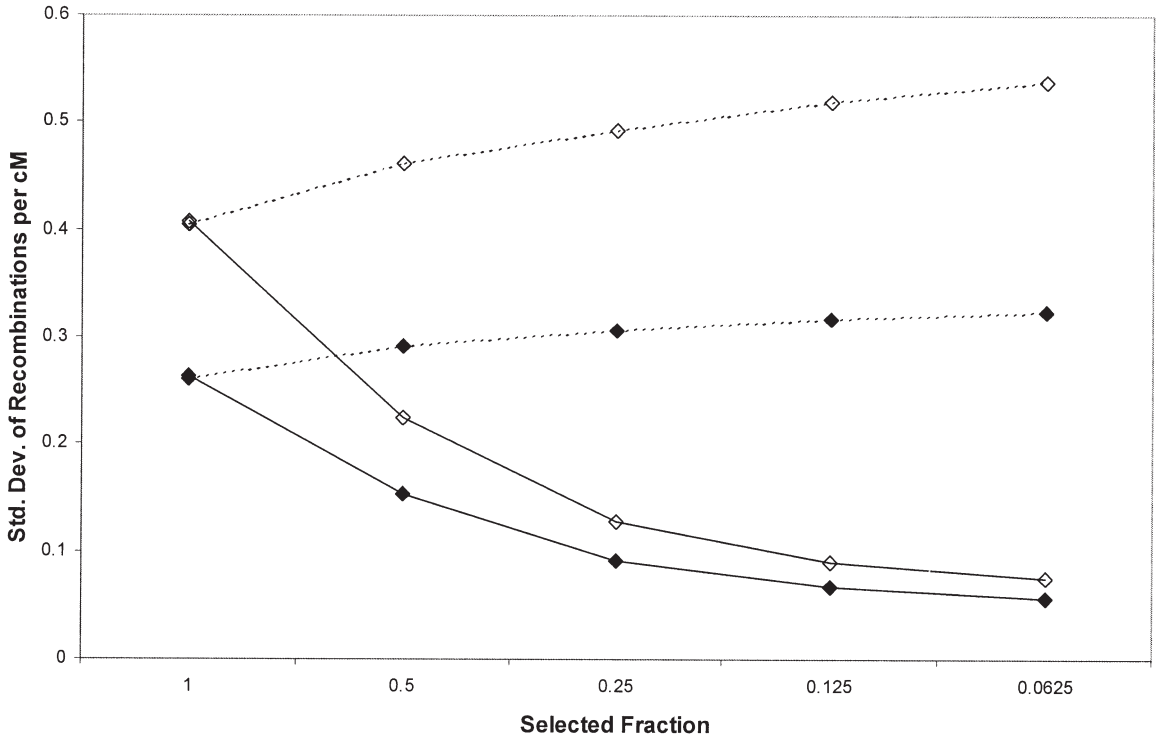

Fig. 2. Standard deviation across repeated simulations of the total number of recombinants per centimorgan among the selected genotypes in 10- and 20-cM marker intervals, as affected by the selected fraction of genotypes. Lines and symbols are as for Fig. 1.

with increasing selection pressure. In contrast, with the uniRec method that variance decreased with increasing selection pressure (Fig. 2). The cause of the increased variance with the maxRec method was simply that the method increased the number of recombinations present but did nothing to counteract the variance from increasing with the mean number. The uniRec method however was progressively more effective at counteracting increasing variance with increasing mean number of recombinations as the selection pressure went up. With the uniRec method, for $N = 200$ and selected fraction = 0.0625, 97% of all 10-cM marker intervals had between 31 and 33 recombinations, cumulated across the 200 progeny, while 98% of all 20-cM marker intervals had between 61 and 65 recombinations, cumulated across the 200 progeny. The comparable percentages with the maxRec method were 22 and 30% showing that the uniRec method achieved greater uniformity in recombination events both across intervals within a single simulation and across simulations.

Increase in the Accuracy of QTL Position Estimation Using maxRec and uniRec

Essentially no difference was observed between the maxRec and uniRec methods of selection in their effect on the MSE for QTL position. Back-transformed averages across all other treatments of MSE were 34.8 and 34.6 for the maxRec and uniRec methods, respectively. Given the level of replication, this difference was significant at $P = 0.07$, but is not of practical importance. All other main effects, selected fraction, number of progeny analyzed, marker spacing, and QTL effect, were highly significant, whereas few interaction effects approached significance. The sums of squares explained by the most significant interaction effect (the progeny number × marker spacing × QTL effect interaction) was one sixth that of the sums of squares explained by the least significant main effect (the selected fraction effect). I therefore discuss only the main effects here. Increasing the selection intensity (reducing the selected fraction) decreased the MSE for QTL position. Back-transformed averages across all other treatments of MSE were 38.9, 34.8, 33.4, and 32.0 for selected fractions of 0.5, 0.25, 0.125, and 0.0625, respectively, as compared to an MSE of 49.6 for nonselected progeny. Apparently, most of the gain from selecting progeny with greater numbers of recombination events could be obtained with a selected fraction of 0.5, in other words, by scoring markers on twice the number of individuals than would eventually be phenotyped and analyzed. Differences in MSE determined by other factors were 65.4 versus 18.4 for families of 100 versus 200 progeny; 48.4 versus 24.8 for marker spacings of 20 versus 10 cM; and 68.7, 34.1, and 17.6 for QTL explaining 7, 11, and 17% of the phenotypic variance, respectively.

The discussion above of the effect of selecting progeny with high numbers of recombinants differs somewhat from the discussion given in Darvasi and Soller (1995), in that they decreased marker spacing with increases in recombination frequency due to additional rounds of random mating. This decrease in marker spacing ensured that the recombination frequency across marker intervals remained constant despite increases in overall recombination frequency. Using selective pheno-

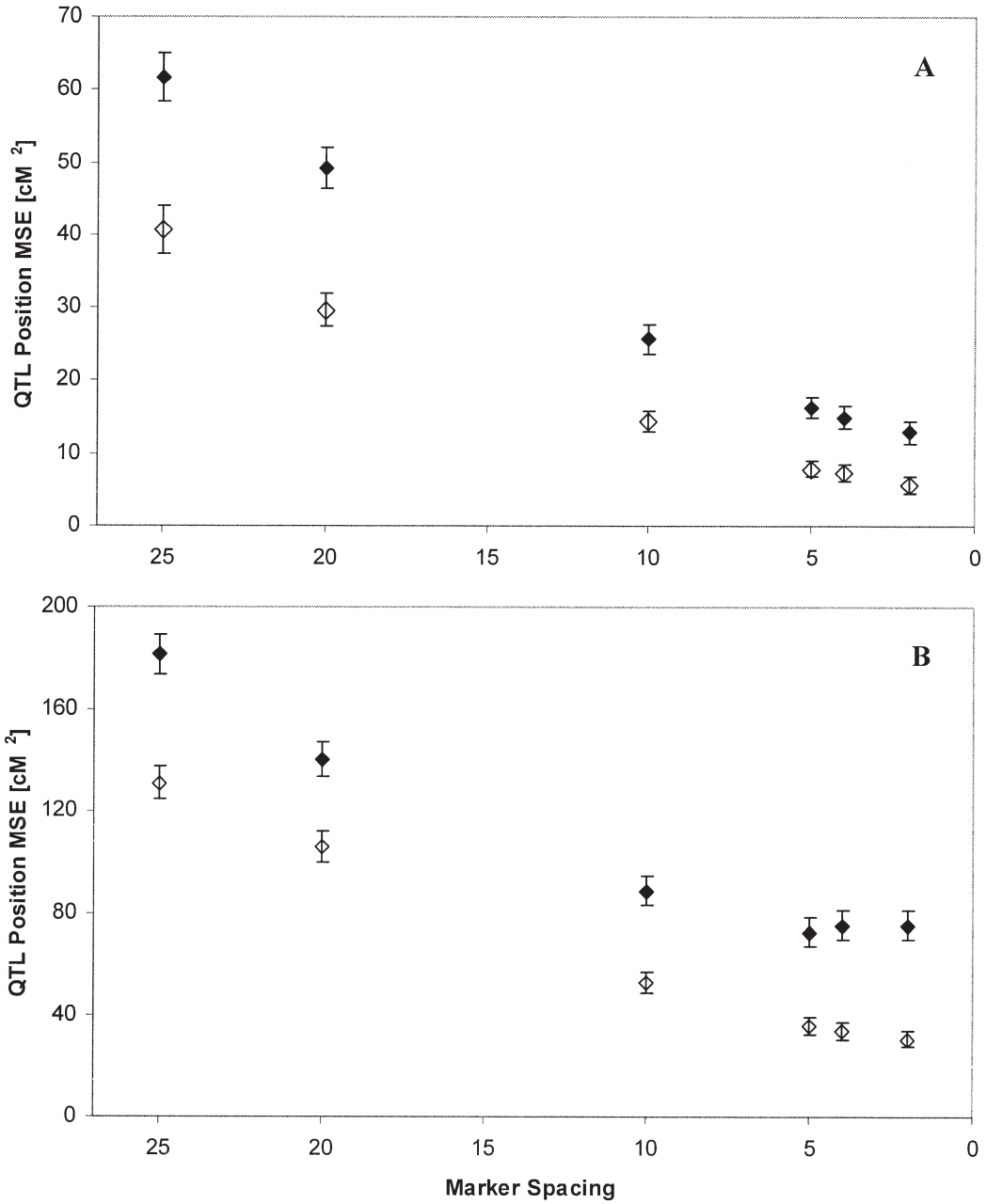

Fig. 3. Average QTL position MSE as affected by marker spacing. Closed symbols are for no selection (selected fraction of 1) and open symbols are for a selected fraction of 1/16 under the uniRec method. Whiskers give the 95% confidence interval for the average. Figure 3A is for QTL explaining 11.6% of the phenotypic variance; Fig. 3B is for QTL explaining 5% of the phenotypic variance.

typing, and at the highest selection intensity examined here, the recombination frequency approaches twice that expected under random selection. Thus, it may be appropriate to compare the MSE for a marker spacing of 20 cM but without selection to the MSE for a marker spacing of 10 cM and selective phenotyping with a selected fraction of 0.0625. The averages for those treatments were 66.3 versus 22.6, respectively, indicating that selective phenotyping reduced the QTL position MSE by 66%.

Further simulations provided information on the value of selective phenotyping relative to decreasing marker spacing for the purpose of accurately mapping QTL. For relatively large effect QTL (11.6% of the phenotypic variance), decreasing marker spacing all the way to 2 cM decreased average QTL position MSE without selection and for a selected fraction of 0.0625 cM (Fig. 3A). Consequently, the average QTL position MSE without selection with 2-cM marker spacing (12.9 cM^2) was statistically not significantly different from the average QTL position MSE under selective phenotyping with 10-cM marker spacing (14.3 cM^2). This result means that, if markers are available for high-density genotyping and if QTL of large effect are of primary interest, selective phenotyping may not provide gains in the accuracy of QTL mapping. In contrast, for QTL of smaller effect (5% of the phenotypic variance), decreasing marker spacing below 5 cM did not decrease QTL position MSE

without selection, but did decrease QTL position MSE for selective phenotyping (Fig. 3B). Darvasi et al. (1993), using maximum likelihood QTL mapping, also found that, in the absence of advanced intercrossing, the mapping accuracy of smaller-effect QTL benefited less from decreasing marker spacing than did the mapping accuracy of larger-effect QTL. In the case of selective phenotyping, the consequence of this phenomenon was that the MSE obtained under selective phenotyping at a marker spacing of 10 cM (53.0 cM^2) could not be matched without selection, where the lowest observed MSE was 74.3 cM^2. Furthermore, with decreasing marker spacing, the MSE under selective phenotyping declined further, to a value of 30.7 cM^2 for the 2-cM marker spacing (Fig. 3B). Selective phenotyping therefore appears to be more useful when marker resources are not available for dense marker mapping and when researchers also seek to accurately map QTL of small effect.

Effect of maxRec and uniRec on Variance of QTL Position Accuracy

Variance in the QTL position MSE accuracy across repeated independent simulations was significantly lower for the uniRec than for the maxRec method of selection, but the difference was again hardly of practical importance. The variance of the log-transformed MSE across repeated simulations was 0.900 for the maxRec method versus 0.876 for the uniRec method. In practice, this meant that MSE between 14.1 and 85.6 were within one standard deviation of the mean for the maxRec method, while MSE between 14.4 and 83.0 were within one standard deviation of the mean for the uniRec method. Thus, while the uniRec method did show lower extreme MSE than the maxRec method, the difference was not great. Variation in the MSE across repeated simulations apparently had very little to do with the uniformity of the distribution of recombination frequencies. The uniRec method was much more effective than the maxRec method at keeping this distribution uniform (Fig. 2), but it did not benefit from that ability in terms of reduced variability of the accuracy of QTL mapping. Presumably, the MSE depends more on the joint random sample of the phenotype with the marker genotypes than on the number of recombinations in the marker interval where the QTL is located.

CONCLUSIONS

Continued declines in the cost of genotyping progeny are opening new opportunities in the realm of marker-assisted selection and QTL mapping. The research reported here explores the possibility that, among segregating progeny, some may carry greater information allowing for the localization of QTL than others. In particular, recombination events are necessary to map QTL accurately, and some individuals result from gametes with a greater number of recombination events than others. Two methods to use DNA marker genotypes to select individuals with high numbers of recombination events were outlined and explored through simulation. While the uniRec method ensured greater uniformity over the genome in the number of recombination events among selected genotypes than the maxRec method, the MSE for QTL position under the two methods was similar, and variability in that MSE across repeated simulations was only very slightly lower for the uniRec than the maxRec method. Simulations indicated that selective phenotyping will most improve the accuracy of QTL mapping for QTL of small effect or when available markers do not allow marker spacing below 10 cM. These qualitative conclusions were obtained in simulations of a relatively small genetic map (600 cM), but will presumably hold for organisms with larger maps. It seems likely, however, that selective phenotyping will be most effective for organisms with small map sizes, and further simulations should be conducted to determine the impact of map size on the gains in QTL mapping accuracy obtained from selective phenotyping. Finally, the simulations here have assumed that the same marker data would be used to select progeny and to perform QTL mapping. An option to reduce genotyping costs would be to use fairly large marker spacings in the selection phase and then genotype selected progeny further to obtain small marker spacings for the QTL mapping phase. Simulation could also be used to determine optimal numbers of markers for the selection and QTL mapping phases to locate QTL with maximal accuracy given fixed genotyping costs.

ACKNOWLEDGMENTS

Software used in the analyses presented in this research was developed under USDA-NRI, CSREES Project Award No. 2001-35301-10848, and supported by Hatch Act and State of Iowa. This report benefited from the suggestions of three anonymous reviewers.

REFERENCES

Darvasi, A. 1998. Experimental strategies for the genetic dissection of complex traits in animal models. Nat. Genet. 18:19–24.

Darvasi, A., and M. Soller. 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet. 85:353–359.

Darvasi, A., and M. Soller. 1995. Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141:1199–1207.

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

Esch, E., and W.E. Weber. 2002. Investigation of crossover interference in barley (*Hordeum vulgare* L.) using the coefficient of coincidence. Theor. Appl. Genet. 104:786–796.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Sci. and Tech., Harlow, UK.

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter. 1996. Markov chain Monte Carlo in practice. Chapman and Hall, London.

Haldane, J.B.S. 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. J. Genet. 8:299–309.

Iraqi, F., S.J. Clapcott, P. Kumari, C.S. Haley, S.J. Kemp, and A.J. Teale. 2000. Fine mapping of trypanosomiasis resistance loci in murine advanced intercross lines. Mamm. Genome 11:645–648.

Jannink, J.-L., and X.-L. Wu. 2003. Estimating allelic number and identity in state of QTL in interconnected families. Genet. Res. 81:133–144.

Jobs, M., W.M. Howell, L. Stromqvist, T. Mayr, and A.J. Brookes. 2003. DASH-2: Flexible, low-cost, and high-throughput SNP geno-

typing by dynamic allele-specific hybridization on membrane arrays. Genome Res. 13:916–924.
Kosambi, D.D. 1944. The estimation of map distances from recombination values. Ann. Eugen. 12:172–175.
Lee, M., N. Sharopova, W.D. Beavis, D. Grant, M. Katt, D. Blair, and A. Hallauer. 2002. Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. Plant Mol. Biol. 48:453–461.
Roberts, G.O. 1996. Markov chain concepts related to sampling algorithms. p. 45–58. *In* W.R. Gilks et al. (ed.) Markov chain Monte Carlo in practice. Chapman and Hall, London.
Rong, J.K., C. Abbey, J.E. Bowers, C.L. Brubaker, C. Chang, P.W. Chee, T.A. Delmonte, X.L. Ding, J.J. Garza, B.S. Marler, C.H. Park, G.J. Pierce, K.M. Rainey, V.K. Rastogi, S.R. Schulze, N.L. Trolinder, J.F. Wendel, T.A. Wilkins, T.D. Williams Coplin, R.A. Wing, R.J. Wright, X.P. Zhao, L.H. Zhu, and A.H. Paterson. 2004. A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). Genetics 166:389–417.
Satagopan, J.M., B.S. Yandell, M.A. Newton, and T.C. Osborn. 1996. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144:805–816.
Sherman, J.D., and S.M. Stack. 1995. Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (*Lycopersicon esculentum*). Genetics 141:683–708.
Sillanpää, M.J., and E. Arjas. 1998. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148:1373–1388.
Yi, N., and S. Xu. 2001. Bayesian mapping of quantitative trait loci under complicated mating designs. Genetics 157:1759–1771.