Genet. Sel. Evol. 30 (1998) 453–468

Original article

Mapping QTL in outbred populations using selected samples

Mario L. Martinez, Natascha Vukasinovic*,
A.E. Freeman, Rohan L. Fernando

Department of Animal Science, Iowa State University, Ames, IA 50011, USA

(Received 22 December 1997; accepted 9 September 1998)

Abstract – A simulation study was carried out to investigate the influence of family selection and selective genotyping within selected families on the power and bias of estimation of genetic parameters in an outbred population with a half-sib family structure. Marker genotypes were determined only for sires that had offspring in the high and low phenotypic tails of the entire distribution of the trait of interest. Offspring of selected sires were genotyped. Within selected families, three different sampling schemes were considered: 1) offspring sampled from the tails of the distribution; 2) offspring randomly sampled; 3) all offspring of a selected sire analyzed. Control data consisted of randomly sampled offspring from randomly chosen sires. An interval mapping procedure based on the random model approach was applied to simulated data. The QTL location and the variance components were estimated using the maximum likelihood technique. Compared with the control data, selective genotyping of sires increased power of QTL detection, but also resulted in severely biased estimates for variance components, especially when the most extreme offspring of selected sires were sampled. Including phenotypic data from all individuals along with marker information obtained only on selected offspring provided improved estimates of the QTL parameters without loss in power. © Inra/Elsevier, Paris

QTL / family selection / selective genotyping / interval mapping

Résumé – Détection de QTLs dans une population non consanguine à partir d'un échantillon sélectionné. Une simulation a été réalisée de manière à analyser l'influence de la sélection familiale et du typage sélectif dans les familles sélectionnées, sur la qualité d'estimation des paramètres génétiques dans une population non consanguine ayant une structure de demi-frères. Les génotypes marqueurs ont été déterminés uniquement pour les pères dont la descendance s'est située aux extrémités haute ou basse de la distribution phénotypique pour le caractère étudié. La descendance

* Correspondence and reprints: Animal Breeding Group, Swiss Federal Institute of Technology, Clausiusstr. 50, 8092 Zurich, Switzerland
E-mail: vukasinovic@inw.agrl.ethz.ch

des pères sélectionnés a été génotypée. À l'intérieur des familles sélectionnées, trois schémas différents d'échantillonnage ont été considérés : (i) aux extrémités de la distribution (ii) au hasard (iii) échantillonnage exhaustif. Les données de contrôle étaient constituées de la descendance triée au hasard de pères triés au hasard. Une procédure de détection de QTL par intervalle basée sur l'approche du modèle aléatoire a été appliquée aux données simulées. La position du QTL et la valeur des composantes de variance ont été estimées en utilisant une technique de maximum de vraisemblance. Par rapport aux données de contrôle, le typage sélectif des pères a augmenté la puissance de détection des QTLs mais a entraîné des estimées de composantes de variance sévèrement biaisées, particulièrement quand la descendance extrême des pères sélectionnés a été échantillonnée. L'inclusion des données phénotypiques de tous les individus et non seulement ceux typés pour les marqueurs améliore la qualité d'estimation des paramètres QTL sans perte de puissance de détection de QTL.

QTL / sélection familiale / typage sélectif / détection de QTL par intervalle

1. INTRODUCTION

Selective genotyping is a method of quantitative trait locus (QTL) mapping in which the analysis of linkage between marker loci and a QTL affecting the trait of interest is carried out by genotyping only individuals from the high and low phenotypic tails of the entire distribution of the trait values in the population [2]. Individuals that deviate most from the population mean are considered to be most informative for linkage, because their genotypes can be inferred from their phenotypes more clearly than can those for average animals [7].

For a given power, selective genotyping can considerably reduce the number of individuals genotyped at the expense of an increase in the number of individuals phenotyped. Thus, the benefits of selective genotyping depend on whether the information on the trait is readily available or whether additional expensive testing is required. In a livestock population that is part of a breeding program, performance records are easily accessible for a large number of animals. By genotyping only extreme animals, the cost of linkage analysis can be considerably reduced.

An important aspect of using selected samples for QTL detection is to choose extreme sibs from parents with average phenotypic values, because such parents are more likely to be heterozygous for the QTL. If parents have similar extreme phenotypes (either high or low) they are probably homozygous for the QTL and, therefore, the linkage would be much more difficult to detect [12]. Sires with a large within family deviation are considered to be most informative for linkage. If a QTL with a reasonably large effect segregates in the population, phenotypic deviation between the extreme offspring will be due to the presence of the alternative QTL alleles in either tail of the distribution. Phenotypic differences among individuals that are due to a large polygenic or environmental deviation will be eliminated if the families that the individuals for genotyping are sampled from are large enough. Therefore, in livestock populations with usually large half-sib families, it would be useful to select sire families with most extreme offspring prior to genotyping to ensure sufficient within family genetic

variability necessary for successful detection of a putative QTL segregating in the population. However, very little research on this topic has been carried out to date.

Furthermore, most of the experiments considering selective genotyping have been designed assuming a biallelic QTL and expecting an increased frequency of alternative QTL alleles in either tail of the distribution. This assumption is correct for experiments involving inbred line crosses or backcrosses, when the QTL alleles can be directly inferred from the marker alleles. This assumption, however, does not hold for outbred populations. In an outbred population, inbred lines are not easily available. Linkage phases are usually unknown as well as the number of genes affecting the trait and the number of alleles at the putative QTL. The genetic architecture and the exact mode of inheritance at the QTL are unknown. As a consequence, the allelic effects of genes cannot be estimated. In such situations, a robust method for linkage analysis, which does not require specification of the genetic model, is preferable. Goldgar [5] defined a random model for linkage analysis that has been proved to be robust against different genetic models and efficient for linkage analysis in outbred populations. Under the random model, QTL effects are assumed to be normally distributed, which leads to the estimation of the variance associated with the QTL (i.e. with a chromosomal region) instead of estimating QTL allelic effects.

The random model approach to QTL mapping in half-sib families is based on phenotypic similarity (or covariance) between genetically related individuals. This covariance can be defined as a function of the proportion of genes identical-by-descent (IBD) that two individuals share at the loci affecting the trait. The covariance between two relatives comprises the polygenic and the QTL component. The polygenic component consists of many genes with small effects. Thus, it is assumed that the average proportion of alleles IBD shared by two relatives equals the genetic relationship coefficient between them, i.e. 1/4 in half-sib families. On the other hand, the QTL component usually represents one major locus (QTL) with a large effect. Therefore, for the same kind of relationship, the proportion of alleles IBD shared by the relatives at the QTL differs from one pair of relatives to another. In half-sib families with one common parent the proportion of alleles IBD at the QTL ranges from 0 to 1/2. Because the QTL itself is unobservable, the proportion of alleles IBD at the QTL must be inferred from the available information on linked marker loci [6].

The greater the shared proportion of alleles IBD, the more similar are the phenotypes of the two relatives. With a larger deviation of the actual IBD proportion from the expected average value of 1/4, the power of separating the QTL from the polygenic component and the power of detecting a QTL become larger. Selective genotyping is expected to increase deviation of the IBD proportion from the average by changing the IBD proportion towards the maximum within the extreme groups, and towards zero between the extreme groups. Therefore, a QTL analysis under the random model should be more efficient if individuals for genotyping are sampled from the tails of the distribution.

The objectives of this paper have been defined as follows:

1) to examine efficiency of selection of sires, i.e. half-sib families prior to selective genotyping of the offspring;

2) to examine the impact of selective genotyping within selected families on power and estimation of QTL parameters using different sampling schemes;

3) to examine the efficiency of the random model approach for QTL mapping under selective genotyping, with information available on only genotyped individuals or on all phenotyped animals.

2. METHODS

2.1. Data simulation and analyses

Genetic and phenotypic data were generated by Monte-Carlo simulation techniques. Mapping QTL was considered within a 20 cM long chromosomal segment flanked by two markers, both with four equally frequent alleles. For simplicity, a QTL was simulated in the middle of the segment, i.e. at 10 cM. Five codominant alleles with equal frequency were assumed at the QTL.

Parents were generated by random allocation of genotypes at each locus assuming Hardy-Weinberg equilibrium. Parental linkage phases were assumed unknown. Progeny were generated assuming no interference, so that a recombination event between the first marker and the QTL did not affect the occurrence of a recombination event between the QTL and the second marker. The recombination fraction was calculated by the Haldane map function.

Phenotypic data for progeny were simulated as follows:

$$y_{ij} = \mu + q_{ij} + 1/2(s_i + d_{ij}) + \phi_{ij} + e_{ij}$$

where y_{ij} is the phenotypic value of the individual j in the half-sib family i; μ is the population mean; q_{ij} is the effect of the QTL genotype of individual j in family i; s_i is the sire's contribution to the polygenic value; d_{ij} is the dam's contribution to the polygenic value; ϕ_{ij} is the effect of Mendelian sampling on the polygenic value; and e_{ij} is the residual error.

The phenotypic value of the trait was assumed to be normally distributed with mean equal to zero and variance equal to one. Heritability of the trait was assumed to be 0.25. Allelic effect of the QTL was defined so that the additive variance of the QTL accounted for 40, 20 and 4 % of the genetic variance, i.e. 10, 5 and 1 % of the total phenotypic variance, so that the true values of QTL heritability (h_g^2) and polygenic heritability (h_a^2) were 0.10, 0.05 and 0.01 and 0.15, 0.20 and 0.24, respectively.

2.2. Sampling schemes

A typical dairy cattle population with prevailing half-sib family structure was assumed. The base population under the breeding program consisted of 500 sires used by an artificial insemination (AI) organization and an infinite number of females. Each sire was bred with 300 randomly chosen unrelated dams to produce one phenotyped offspring per mating. The selection of individuals for genotyping followed in two steps. In the first step sire families assumed to be most informative for QTL mapping were selected. In the second step offspring from selected families were chosen for genotyping and QTL analysis.

2.2.1. Selection of families

Offspring of all sires were ranked according to their simulated phenotypes to choose sires whose progeny will be genotyped. Only sires with offspring within the top and the bottom 10 % of the entire distribution were considered for selection. The selection decision was based on the assumption that these sires are most likely to be heterozygous for the QTL affecting the trait. The selection criterion for sires was defined as

$$c = 1/n_1 + 1/n_2$$

where n_1 is the number of progeny in the top 10 % of the distribution and n_2 is the number of progeny in the bottom 10 % of the distribution. If a sire has a large number of daughters in both the top and the bottom 10 % of the distribution, both n_1 and n_2 will be large, and c will have a small value, closer to zero as n_1 and n_2 increase. Therefore, sires were ranked according to the value of c, assigning higher rank to those sires with a smaller value of c. Sires were selected starting from that with the smallest value of c, i.e. from the sire with the largest number of offspring equally distributed in the top and bottom 10 % of the entire distribution. Sampling continued until the number of sires needed for genotyping was reached.

2.2.2. Selection of individuals within selected families

Three different sampling schemes were applied to the progeny of the selected sires.

Scheme I: from each of the selected sires, the number of offspring needed for analysis were sampled starting from the tails of the distribution. Therefore, 50 % of the animals for genotyping had the lowest and 50 % the highest phenotypic values.

Scheme II: from each of the selected sires, the offspring needed for genotyping were randomly sampled from the entire family.

Scheme III: each sire from the base population was allowed to produce only the exact number of offspring needed for genotyping. Sires were selected according to the criterion c. No selection was applied to the offspring, i.e. all offspring of a selected sire were analyzed.

Note that not all of the offspring of the selected sires chosen for genotyping were necessarily within the top and bottom 10 % of the entire phenotypic distribution.

Control: in addition to the sampling schemes, control data were generated assuming no selection in either sires or offspring. These data were used as a comparison basis.

The number of genotyped offspring was held constant at 2 000. Number of families and number of offspring per family varied. For each sampling scheme, three different combinations were examined: 100 families of 20 offspring, 40 families of 50 offspring and 20 families of 100 offspring.

For scheme I, additional simulations were carried out assuming a base population consisting of 100 sires with 80 offspring each. Twenty sires were chosen for genotyping starting from the sire with the largest number of

offspring equally distributed in the top and the bottom 10 % of the phenotypic distribution. The proportions of offspring chosen for genotyping were 0.10, 0.25, 0.50 and 1.00. One half of the total number of the genotyped individuals was taken from either tail of the phenotypic distribution. But, in the analysis, all data were considered: typed and untyped offspring from the selected sires as well as all (untyped) offspring from the unselected sires. Thus, the sample size was equal for all analyses – 100 families with 80 offspring each.

2.3. Statistical analyses

Simulated data were analyzed using the following model:

$$y_{ij} = \mu + g_{ij} + a_{ij} + e_{ij} \quad [1]$$

where y_{ij} is the phenotypic trait value of the jth individual in the ith family assumed ideally precorrected for environmental fixed effects, μ is the population mean, g_{ij} is the additive genetic effect of the QTL with $g_{ij} \sim N(0, \sigma_g^2)$, a_{ij} is the additive effect of the polygenic component with $a_{ij} \sim N(0, \sigma_a^2)$, and e_{ij} is the random environmental variation with $e_{ij} \sim N(0, \sigma_e^2)$. Assuming linkage equilibrium, the variance of y_{ij} is

$$\mathrm{Var}(y_{ij}) = \sigma^2 = \sigma_g^2 + \sigma_a^2 + \sigma_e^2 \quad [2]$$

where σ^2 is the phenotypic variance, σ_g^2 is the variance associated with a QTL, σ_a^2 is the variance associated with genes other than the tested QTL (polygenic variance), and σ_e^2 is the environmental (residual) variance.

The expected value of the covariance between two non-inbred half-sibs within the family is

$$\mathrm{Cov}(y_{ij}, y_{ij'}) = \pi_q \sigma_g^2 + 1/4\, \sigma_a^2 \quad [3]$$

where π_q is the proportion of alleles identical-by-descent (IBD) shared by the half-sibs j and j' at the putative QTL. The coefficient of the polygenic variance is 1/4 because, by expectation, two non-inbred half-sibs share 1/4 alleles IBD.

With k half-sibs in the ith family, the covariance matrix (V_i) among phenotypic values of the half-sibs (y_{ij}) is

$$V_i = \mathrm{Var} \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ik} \end{bmatrix} = \sigma^2\, \mathbf{C}_i \quad [4]$$

with

$$\mathbf{C}_i = \begin{bmatrix} 1 & r_{12} & \cdots & \cdots & r_{1k} \\ r_{21} & 1 & & & r_{2k} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & \cdots & 1 \end{bmatrix} \quad [5]$$

and

$$r_{jj'} = \pi_q \, h_g^2 + 1/4 \, h_a^2 \qquad [6]$$

where $h_g^2 = \sigma_g^2/\sigma^2$ and $h_a^2 = \sigma_a^2/\sigma^2$. π_q is the proportion of alleles IBD shared by the individuals j and j' at the QTL. π_q must be estimated using information on linked marker loci. Given the proportion of alleles IBD at two markers flanking the putative QTL, the proportion of alleles IBD at the QTL can be estimated using linear regression [3]:

$$\widehat{\pi}_q = \alpha + \beta_i \, \pi_1 + \beta_2 \, \pi_2 \qquad [7]$$

where π_1 and π_2 are IBD values for two flanking markers. For simplicity, marker genotypes were assumed known in both parents. The proportion of alleles IBD at marker loci shared by two half-sibs within a family was estimated using simulated marker genotypes of the offspring and their parents using the procedure described by Haseman and Elston [6] for the situation with known parental information, appropriately adjusted to fit the half-sib family structure [9]. For those samples in which only a part of the individuals were genotyped, but all phenotypes were included in the analysis, the same procedure was applied to calculate the proportion of IBD at marker loci shared by two typed half-sibs from a typed sire. The unknown proportions of IBD shared by two untyped half-sibs or by one typed and another untyped half-sib were replaced by their expected value of 0.25.

Assuming a multivariate normal distribution of the data (y_{ij}), we have a joint density function of the observations within a half-sib family:

$$f(y_i) = \frac{1}{(2\pi\sigma^2)^{k/2}|\mathbf{C}_i|^{1/2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(y_i - 1\mu)'\mathbf{C}_i^{-1}(y_i - 1\mu)\right\} \qquad [8]$$

where $y_i = [y_{i1}\, y_{i2}\, y_{i3} \ldots y_{ik}]'$ is a $k \times 1$ vector of observed phenotypic values for k half-sibs within the ith family, and 1 is a $k \times 1$ vector with all entries equal to one.

The overall log likelihood for N independent half-sib families is

$$L = \sum_{i=1}^{N} \log\left[f(y_i)\right] \qquad [9]$$

The maximum likelihood interval mapping procedure was applied to the generated data. The likelihood function was maximized with respect to h_g^2, h_a^2, and σ^2 for each testing position along the chromosomal segment using a simplex algorithm described by Xu and Atchley [11]. The chromosome was screened from the left to the right end in steps of 2 cM. For each position, the likelihood ratio test (LR) was computed as minus twice the difference in log likelihood between the null hypothesis ($h_g^2 = 0$) and the alternative hypothesis ($h_g^2 \neq 0$). The testing position with the highest LR was accepted as the most likely position of the QTL. Similarly, estimated variance components (h_g^2 and h_a^2) at the position with the highest likelihood ratio were accepted as maximum

likelihood estimates for these parameters. For each sampling scheme and each parameter combination, the simulation and analysis were repeated 100 times.

The power of QTL detection was obtained empirically by simulation. The empirical distribution of the LR test statistic under H_0 was generated by simulating and analyzing data in the same manner, but assuming no QTL in the entire segment. For each sampling scheme and each parameter combination, data simulation and estimation under H_0 were repeated 100 times. Each time the highest value of the LR was recorded. After 100 replicates, the obtained LR values were ordered, and the 95th value was chosen as an empirical 5 % significance threshold for this parameter combination. The power of QTL detection was then calculated as a percentage of replicates in which the maximum LR exceeded the corresponding threshold.

3. RESULTS AND DISCUSSION

3.1. Power, QTL position and variance components with selected samples

Power of detecting QTL by using different sampling schemes for different parameter combinations is given in *table I.*

Table I. Empirical power of QTL detection with selected samples.

NS	NO	Thr	QTL heritability		
			0.10	0.05	0.01
Scheme I	extreme				
100 (500)	20 (300)	0.00	*	*	*
40 (500)	50 (300)	0.00	*	*	*
20 (500)	100 (300)	0.00	*	*	*
Scheme II	random				
100 (500)	20 (300)	0.86	64	19	8
40 (500)	50 (300)	2.81	92	39	8
20 (500)	100 (300)	2.32	98	49	7
Scheme III	all				
100 (500)	20 (20)	3.48	50	20	0
40 (500)	50 (50)	3.12	86	54	7
20 (500)	100 (100)	4.86	99	69	11
Control					
100 (100)	20 (20)	3.37	48	25	13
40 (40)	50 (50)	3.77	61	28	12
20 (20)	100 (100)	2.49	92	46	14

NS: number of sires, in parentheses: number of sires in the base population; NO: number of offspring per sire, in parentheses: total number of offspring produced by each sire; Thr: empirical threshold, obtained as the 95th percentile of the ordered likelihood ratio test statistics under H_0; * parameter not estimable.

The parameter with most influence on power was family size. For the fixed number of genotyped progeny (2 000), considerably higher power was obtained with larger families and a smaller number of families than with smaller family size and a larger number of families. For all sampling schemes, regardless of the size of QTL effect, the highest power was obtained with 20 families with 100 progeny each – almost twice as high as for the reverse combination with 100 families and 20 progeny each. This is explained by the increased number of half-sib pairs within a family. In general, for N families with n half-sibs each, the total number of half-sib pairs is $N\frac{n(n-1)}{2}$. As n increases while nN remains constant, the number of half-sib pairs also increases, and this results in an increased amount of information used in the analysis.

The proportion of variance explained by the QTL was another factor that influenced power of QTL detection. Generally, higher power was obtained with a larger QTL. With a small QTL ($h_g^2 = 0.05$ and 0.01) power was very low and ranged between 0 and 14 %, depending on the sampling schemes and family size.

For scheme I, in which the most extreme offspring of the selected sires were sampled, the power of QTL detection could not be calculated. In obtaining the empirical threshold value for scheme I, the LR was zero for all positions in all 100 replicates, i.e. likelihood failed to maximize through the entire chromosomal segment. Therefore, the advantage of using selected samples can be seen only from schemes II and III. A relatively large QTL ($h_g^2 = 0.10$) can be detected with higher power than in the situation when the sires are not selected. Also, a QTL with small effects ($h_g^2 = 0.05$) can be detected with higher power if the half-sib families are large enough. Only for a very small QTL ($h_g^2 = 0.01$) does the selection of sires seem not to be advantageous.

Mean estimates of QTL position with the corresponding among replicates standard deviations are given in *table II.*

Under scheme I, for some parameter combinations with $h_g^2 = 0.05$ and $h_g^2 = 0.01$, the position of the QTL was not estimable, because the likelihood failed to maximize through the entire segment. For other parameter combinations, the position of the QTL was poorly estimated and biased downwards with low QTL heritability and smaller family size. The estimates improved with increased QTL heritability and family size.

For scheme II the estimates for QTL position ranged between approximately 7 and 11 cM. Similar estimates were obtained for scheme III, except for the parameter combinations with a sample size of 100 families of 20 offspring and $h_g^2 = 0.05$ and 0.01. The estimates of the QTL position for the parameter combinations with a low QTL heritability tend to take values on the left-hand side of the chromosome, especially when low QTL heritability was accompanied by small family size. This downward bias was not expected, because QTL was simulated centrally. The unexpected results might be due to the properties of the simplex algorithm used to maximize the likelihood function. With a low QTL heritability, the simplex algorithm was apparently unable to continue maximization of the likelihood function after reaching a local maximum.

The among replicate standard deviations of the estimates for the QTL position were large with low QTL heritability and smaller family size, because the individual estimates largely vary from one replicate to the other. The

Table II. Estimates of the QTL position (in cM) averaged over 100 replicates with corresponding among replicates standard deviations (in parentheses).

NS	NO	QTL heritability		
		0.10	0.05	0.01
Scheme I	extremes			
100 (500)	20 (300)	5.96 (8.46)	*	*
40 (500)	50 (300)	10.30 (2.44)	3.94 (7.08)	*
20 (500)	100 (300)	10.26(2.61)	8.64 (5.41)	2.34 (2.12)
Scheme II	random			
100 (500)	20 (300)	9.78 (7.27)	10.40 (7.96)	7.34 (9.06)
40 (500)	50 (300)	10.14 (5.11)	8.28 (7.69)	8.18 (8.82)
20 (500)	100 (300)	10.19 (4.09)	10.67 (4.62)	8.83 (6.60)
Scheme III	all			
100 (500)	20 (20)	7.03 (8.76)	3.14 (7.07)	1.45 (4.82)
40 (500)	50 (50)	9.96 (5.95)	9.70 (6.90)	10.09 (9.08)
20 (500)	100 (100)	9.62 (4.20)	10.04 (5.16)	9.01 (6.90)
Control				
100 (100)	20 (20)	8.98 (6.72)	10.75 (8.09)	6.54 (8.53)
40 (40)	50 (50)	9.18 (5.89)	9.47 (7.30)	6.24 (8.23)
20 (20)	100 (100)	9.48 (4.39)	10.40 (5.84)	9.48 (6.93)

NS, NO, *: see *table I.*

estimates were more accurate, i.e. had smaller among replicate standard deviations as the family size and the QTL heritability increased. Compared with the control, the estimates for QTL position with selected samples were biased with smaller family size and lower QTL heritability.

The estimates for QTL heritability (h_g^2), polygenic heritability (h_a^2), total heritability (h_t^2) and phenotypic variance (σ^2), are given in *table III.* The true values of QTL heritability were 0.10, 0.05 and 0.01 with the corresponding polygenic heritability of 0.15, 0.20 and 0.24, respectively. With scheme I, the estimated σ^2 ranged from 2.5 to 5.0. The σ^2 in the sample was, thus, drastically increased compared with the simulated value of 1.0 in the base population prior to selection. The increased σ^2 was due to sampling individuals from the tails of the distribution. The increase in σ^2, however, was not accompanied by an equivalent increase in the estimated genetic variance. Moreover, the two components of the genetic variance were not equally affected. In general, the estimates for h_g^2 were closer to the simulated values and only slightly biased. But, the estimates for h_a^2 and, therefore, the estimates for h_t^2 expressed as a sum of h_g^2 and h_a^2, were severely underestimated. For parameter combinations in which the likelihood failed to maximize, the estimated values for h_a^2 were equal to zero in all replicates.

In scheme II, the estimated σ^2 was only slightly above the simulated value of 1.0. The estimates for h_g^2 were slightly underestimated for simulated QTL

Table III. Estimates for QTL heritability (h_g^2), polygenic heritability (h_a^2), total heritability (h_t^2), and phenotypic variance (σ^2), averaged over 100 replicates.

NS	NO	QTL heritability											
		0.10				0.05				0.01			
		h_g^2	h_a^2	h_t^2	σ^2	h_g^2	h_a^2	h_t^2	σ^2	h_g^2	h_a^2	h_t^2	σ^2
Scheme I	extreme												
100 (500)	20 (300)	0.037	0[1]	0.037	5.00	0	0	0	4.96	0	0	0	4.94
40 (500)	50 (300)	0.182	0	0.182	3.70	0.023	0	0.023	3.63	0	0	0	3.61
20 (500)	100 (300)	0.132	0	0.132	2.63	0.052	0	0.052	2.60	0.000	0	0.00	2.59
Scheme II	random												
100(500)	20 (300)	0.081	0.005	0.086	1.03	0.047	0.062	0.062	1.02	0.028	0.037	0.056	1.02
40 (500)	50 (300)	0.085	0.002	0.087	1.07	0.044	0.053	0.053	1.06	0.019	0.032	0.051	1.05
20 (500)	100 (300)	0.082	0.000	0.083	1.09	0.044	0.049	0.049	1.07	0.019	0.030	0.049	1.07
Scheme III	all												
100 (500)	20 (20)	0.026	0	0.026	1.27	0.010	0.000	0.010	1.25	0.004	0.001	0.005	1.42
40 (500)	50 (50)	0.075	0.004	0.079	1.23	0.050	0.013	0.063	1.23	0.019	0.028	0.047	1.23
20 (500)	100 (100)	0.075	0.005	0.080	1.19	0.045	0.012	0.058	1.18	0.016	0.036	0.052	1.17
Control													
100 (100)	20 (20)	0.092	0.165	0.256	1.00	0.059	0.187	0.246	1.00	0.041	0.218	0.259	1.00
40 (40)	50 (50)	0.073	0.172	0.245	1.00	0.048	0.186	0.234	1.00	0.023	0.219	0.242	1.00
20 (20)	100 (100)	0.062	0.174	0.236	1.00	0.031	0.211	0.242	1.00	0.018	0.193	0.211	1.00

NS, NO, *: see *table I.* h_g^2: QTL heritability; h_a^2: polygenic heritability; h_t^2: total heritability; σ^2: phenotypic variance. [1] 0 without decimals means that all estimates were equal to zero.

heritabilities of 0.10 and 0.05, and slightly overestimated for the simulated QTL heritability of 0.01. However, severe bias was observed for the estimates of h_a^2, and, consequently, the estimates of h_t^2 were biased downwards.

In scheme III, the estimated σ^2 was somewhat overestimated. The mean estimates ranged from 1.17 to 1.42 for the simulated value of 1.0. The estimates for h_g^2 were close to the simulated values except for the parameter combinations with a sample size of 100 families of 20 offspring. In this sampling scheme as well, severe bias in h_a^2 and h_t^2 was observed.

With the control data, considerably less biased estimates for h_g^2, h_a^2, h_t^2 and σ^2 were obtained for all parameter combinations.

3.2. Accounting for selection

The results presented show the advantage of selective genotyping over random samples in giving increased power to detect a QTL. On the other hand, the estimates of QTL position, and, especially, variance components, are grossly biased. This large downward bias is probably due to the method of analysis, which ignores selection. In all three schemes, the selection favors progeny of those sires with the largest number of offspring falling into the top and bottom 10 % of the entire distribution. Therefore, when the most extreme offspring of the selected sires are sampled (scheme I), or even when the offspring for genotyping are randomly sampled from the entire family (schemes II and III), the continuity of normal distribution of data that existed before selection is broken. The assumption of normality required for maximum likelihood estimation is violated, which results in biased estimates or inability to maximize the likelihood function. It is known that standard likelihood methods cannot produce proper results if only selected offspring or offspring from selected sires are genotyped [7]. Thus, an analysis by maximum likelihood techniques must account for truncated selection. This involves maximizing likelihood separately for individuals in the top and in the bottom tail of the distribution [2].

For the selection and the sampling schemes presented in this study, however, the method described by Darvasi and Soller [2] cannot be applied, because the truncation point cannot be unambiguously determined. Some of the genotyped offspring of the selected sires may not have extreme phenotypes, because the truncation point is not distinct, especially in sampling schemes II and III, where the offspring are randomly sampled or the whole family is analyzed. To account properly for this form of selection, missing data methods should be used [8]. According to Lander and Botstein [7], the correct results will be obtained by maximum likelihood techniques if the phenotypes are recorded for all animals and genotypes for untyped animals are simply entered as missing. Therefore, a part of the analysis was repeated with inclusion of all data available on typed and untyped individuals. The proportion of alleles IBD at marker loci for untyped animals was replaced by its expected average value of 0.25, as described in the Methods of the paper.

3.3. Power, QTL position and variance components with selected genotypes and all phenotypes

The results from the simulation for power, QTL position and heritabilities are given in *tables IV–VI*, respectively.

As expected, power to detect a QTL is higher when more individuals are genotyped (*table IV*). Compared with the situation when only 10 % of the population with the most extreme phenotypes are genotyped, the power is nearly doubled when complete offspring information is available. However, an increase in proportion of genotyped individuals above 25 % does not result in a corresponding increase in power, especially when the QTL accounts for a greater part of the genetic variance. With a smaller QTL effect, the selection of animals with extreme phenotypes is primarily based on polygenic and environmental effects, so that detection of the QTL definitely requires more genotyping.

Including all data in the analysis allowed for correct estimation of QTL position regardless of the proportion of untyped animals (*table V*). Mean estimates for QTL position range from 6 to 11 cM and are similar for all parameter combinations. This result was obtained even for the parameter combinations with a QTL heritability of 0.01. Clearly, the estimates are more accurate with larger proportions of genotyped animals, but this improvement in accuracy is not large enough to justify the costs of genotyping more individuals.

The estimates for QTL heritability (h_g^2), polygenic heritability (h_a^2), total heritability (h_t^2) and phenotypic variance (σ^2) are given in *table VI*. The

Table IV. Power of QTL detection with selected samples including all available data[1].

No. and proportion of genotyped offspring/sire	Threshold	QTL heritability		
		0.10	0.05	0.01
8 (0.10)	8.71	49	25	8
20 (0.25)	7.18	72	30	14
40 (0.50)	4.59	86	45	16
80 (1.00)	3.37	88	50	17

[1] Analysis based on 20 sires selected for genotyping out of 100; each half-sibship consists of 80 non-inbred individuals.

Table V. Estimates of the QTL position (in cM) obtained using selected samples including all available data[1] averaged over 100 replicates, with corresponding among replicates standard deviations (in parentheses).

No. and proportion of genotyped offspring/sire	QTL heritability		
	0.10	0.05	0.01
8 (0.10)	11.06 (7.93)	9.70 (8.36)	6.58 (8.92)
20 (0.25)	9.19 (5.15)	9.74 (7.78)	8.69 (8.91)
40 (0.50)	9.98 (5.01)	9.09 (7.19)	8.24 (8.89)
80 (1.00)	10.50 (4.81)	10.12 (6.60)	8.16 (8.64)

[1] See *table IV*.

Table VI. Estimates for QTL heritability, polygenic heritability, total heritability and phenotypic variance obtained using selected samples including all available data[1] and averaged over 100 replicates.

No. and proportion of genotyped offspring/sire	QTL heritability											
	0.10				0.05				0.01			
	h_g^2	h_a^2	h_t^2	σ^2	h_g^2	h_a^2	h_t^2	σ^2	h_g^2	h_a^2	h_t^2	σ^2
8 (0.10)	0.202	0.057	0.259	1.00	0.147	0.109	0.256	1.00	0.083	0.169	0.252	1.00
20 (0.25)	0.159	0.092	0.251	1.00	0.086	0.160	0.246	1.00	0.049	0.197	0.247	1.00
40 (0.50)	0.101	0.140	0.248	1.00	0.060	0.180	0.245	1.00	0.032	0.213	0.246	1.00
80 (1.00)	0.075	0.173	0.247	1.00	0.042	0.206	0.249	1.00	0.027	0.219	0.246	1.00

[1] See *table IV*. h_g^2, h_a^2, h_t^2, σ^2: see *table III*.

estimates for h_t^2 are very close to the simulated value of 0.25 for all parameter combinations. The mean estimates for h_g^2 are, however, mostly biased upwards. The bias is negatively proportional to the number of genotyped animals, and relatively higher as the QTL heritability decreases. Consequently, the mean estimates of h_a^2 are biased downwards. Nevertheless, the sum of $h_g^2 + h_a^2$ is conserved at ~ 0.25, which indicates a successful partitioning of overall genetic and residual variance.

Confounding between h_g^2 and h_a^2 is considered to be a general frailty of the sib-pair approach [4]. This problem has been addressed in several previous studies [1, 9]. Confounding between h_g^2 and h_a^2 can be regarded as independent of the experimental design used and, therefore, not primarily caused by selective genotyping. The power of separating h_g^2 and h_a^2, however, depends on the deviation of π_q from the average, i.e. from 0.25 in the case of the half-sib design [10]. When the data contain a greater proportion of missing marker genotypes, the proportion of alleles IBD at marker loci shared by two half-sibs is replaced by 0.25, and the estimated π_q is, consequently, closer to 0.25. Thus, when fewer animals are genotyped, the separation of h_g^2 and h_a^2 becomes more difficult. This can clearly be seen from the results presented in *table VI*.

Although this paper does not consider simulation studies for sampling schemes II and III with all data included, it is expected that similar results would have been obtained for both randomly sampled offspring and the entire families of the selected sires.

4. CONCLUSIONS

The results presented of the simulation study show that selective genotyping within selected families is advantageous compared with the conventional design based on random samples, because it results in increased power for a given number of individuals genotyped, or, in other words, reduces the number of individuals that need to be genotyped for a given power. This is due to the increased signal of QTL by selection, because over 80 % of the information used in linkage analysis comes from the top and the bottom 20 % of the distribution [2]. From the practical aspect, the method of selection considered in this study is even more efficient than the standard selective genotyping, because selection of extreme individuals is mainly based on sires, whose information is readily available or at least easier to obtain. Because the selection of candidates for genotyping is based on the entire distribution of progeny phenotypic values, it is not necessary to raise and measure any extra individual only for the sake of QTL analysis. In some instances, sires chosen for genotyping can be used more extensively to assure more intensive selection of extreme individuals and an additional increase in power. This is, however, not indispensable, because even an analysis of randomly sampled progeny of a selected sire results in a higher power than in a design without any selection.

To enable proper estimation of QTL parameters – QTL position and variance – when using selected samples, it is necessary to account for selection. The most convenient approach is to include phenotypic data for all individuals and marker data for selected ones, whereas marker data for unselected individuals can simply be entered as missing. The πs for genotyped individuals will then

be calculated in the usual manner, whereas the πs for all other individuals will be replaced by their expected average value of 1/4 for half-sibs. Such an analysis will give correct estimates for the QTL position and genetic variance. The separation of the QTL variance from the polygenic variance will be, however, affected by the proportion of untyped individuals. This is a known difficulty of the sib-pair approach. This problem might be solved if more sophisticated methods for QTL mapping were used. For practical applications the model of analysis described in this paper can be easily extended to include fixed effects or an additional random effect (e.g. a second QTL). The model can be also adjusted to handle general pedigrees and in this way take into account the relationships among animals.

ACKNOWLEDGMENTS

The authors want to express their appreciation to Dr Shizong Xu for providing programs and invaluable suggestions. The authors also thank the EMBRAPA and CNPq, Brazil (MLM) and the Swiss National Foundation, Switzerland (NV), for financial support, and the ISU Computational Center for providing resources. This is Journal Paper No. J-17473 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 3146, and supported by Hatch Act and State of Iowa funds.

REFERENCES

[1] Amos C.I., Robust variance-components approach for assessing genetic linkage in pedigrees, Am. J. Hum. Genet. 54 (1994) 535–543.

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

[3] Fulker D.W., Cardon L.R., A sib-pair approach to interval mapping of quantitative trait loci, Am. J. Hum. Genet. 54 (1994) 1092–1103.

[4] Gessler D.G.D., Xu S., Using the expectation or the distribution of the identity by descent for mapping quantitative trait loci under the random model, Am. J. Hum. Genet. 59 (1996) 1382–1390.

[5] Goldgar D.E., Multipoint analysis of human quantitative genetic variation, Am. J. Hum. Genet. 47 (1990) 957–967.

[6] Haseman J.M., Elston R.C., The investigation of linkage between a quantitative trait and a marker locus, Behav. Genet. 2 (1972) 3–19.

[7] Lander E.S., Botstein D., Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, Genetics 121 (1989) 185–199.

[8] Little R.J.A., Rubin D.B., Statistical Analysis with Missing Data, John Wiley, New York, 1987.

[9] Martinez M.L., Vukasinovic N., Freeman A.E., Estimating QTL location and QTL variance in half-sib families under the random model with missing parental genotypes, J. Anim. Breed. Genet. 115 (1998) 165–180.

[10] Xu S., Computation of the full likelihood function for estimating variance at a quantitative trait locus, Genetics 144 (1996) 1951–1960.

[11] Xu S., Atchley W.R., A random model approach to interval mapping of quantitative trait loci, Genetics 141 (1995) 1189–1197.

[12] Zhang H., Risch N., Mapping quantitative-trait loci in humans by use of extreme concordant sib pairs: selected samples by parental genotypes, Am. J. Hum. Genet. 59 (1996) 951–957.