J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Combined use of phenotypic and genotypic information in sampling animals for genotyping in detection of quantitative trait loci

S. Ansari-Mahyari[1,2,3] & P. Berg[1]

1 Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, University of Aarhus, Denmark
2 Agricultural Biotechnology Research Institute of Iran, Karaj, Iran
3 Department of Large Animal Sciences, Faculty of Life Sciences, University of Copenhagen, Denmark

Keywords
selective genotyping; phenotype and genotype information; QTL mapping.

Correspondence
Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, University of Aarhus, Research Centre Foulum, Blichers Allé 20, PO BOX 50, 8830 Tjele, Denmark.
Tel. +45 8999 1092; Fax: +45 8999 1300;
E-mail: saeidansari.mahyari@agrsci.dk

Received: 23 April 2007;
accepted: 18 October 2007

Summary

Conventional selective genotyping which is using the extreme phenotypes (EP) was compared with alternative criteria to find the most informative animals for genotyping with respects to mapping quantitative trait loci (QTL). Alternative sampling strategies were based on minimizing the sampling error of the estimated QTL effect (MinERR) and maximizing likelihood ratio test (MaxLRT) using both phenotypic and genotypic information. In comparison, animals were randomly genotyped either within or across families. One hundred data sets were simulated each with 30 half-sib families and 120 daughters per family. The strategies were compared in these datasets with respect to estimated effect and position of a QTL within a previously defined genomic region at genotyping 10, 20 or 30% of the animals. Combined linkage disequilibrium linkage analysis (LDLA) was applied in a variance component approach. Power to detect QTL was significantly higher for both MinERR and MaxLRT compared with EP and random genotyping methods (either across or within family), for all the proportions of genotyped animals. Power to detect significant QTL ($\alpha = 0.01$) with 20% genotyping for MinERR and MaxLRT was 80 and 75% of that obtained with complete genotyping compared with 70 and 38% genotyping for EP within and across families respectively. With 30% genotyping, the powers were 78, 83, 78 and 58% respectively. The estimated variance components were unbiased in EP strategies (within and across family), only when at least 30% was genotyped. To decrease the number of genotyped individuals either MinERR or MaxLRT could be considered. With 20% genotyping in MinERR, the estimated QTL variance components were not significant compared with complete genotype information but all studied strategies at 20% genotyping overestimated the QTL effect. Results showed that combining the phenotypic and genotypic information in selective genotyping (e.g. MinERR and MaxLRT) is better than using only the EPs and the combined methods can be considered as alternative approaches to decrease genotyping costs, with unbiased QTL effects, decreased sampling variance of the QTL variance component and also increased the power of QTL detection.

Introduction

Use of gene (marker) information of quantitative traits to increase genetic gain in breeding schemes has been argued as an alternative in dairy cattle selection. The cost of genotyping is generally high but phenotypes of some economically important traits are routinely recorded in dairy cattle, e.g. milk yield and milk contents. Selecting individuals for genotyping is an attempt to overcome this problem. Selective genotyping was firstly introduced by Lander & Botstein (1989) to increase power of detecting quantitative trait loci (QTL) with a small effect. This approach selects portion of individuals for genotyping, on the basis of the individuals' phenotype (generally those with extremely high or low phenotypic value). The advantage is that fewer individuals need to be genotyped for a given probability of identifying QTLs. Selective genotyping yields significant savings in expensive genotyping. Darvasi & Soller (1992) practiced this method for a single marker linked to a QTL to genotype only the individuals with potentially most informative observations. The extreme phenotype method (EP) considers individuals with the most EPs, without using available pedigree or marker genotype information. The general principle exploited is that most linkage information can be inferred by individuals with EP values and for a given number of individuals genotyped, this increases power to detect a QTL compared with random genotyping. However, due to selecting a subset of animals for genotyping, EP could cause a bias in QTL parameter estimations (Lander & Botstein 1989; Darvasi & Soller 1992).

Instead of phenotype information, Kinghorn (1997) presented an index to indicate the information content of genotype probabilities derived from a segregation analysis. According to this index, individuals with the least accurately known genotypes can be identified sequentially. The genotyping models could be used for identifying individual-by-individual (Kinghorn 1997, 1999) or group-by-group (Macrossan 2004). However, this method becomes more complicated when a large haplotype is considered.

Simulation studies proposed that selective methods (e.g. EPs) are using most informative individuals for detecting the QTLs (Muranty & Goffinet 1997; Martinez *et al.* 1998; Van Gestel *et al.* 2000) compared with random genotyping and the costs would be reduced. This is an advantage of genotyping the EPs when the traits are easily and routinely collected in human and animal genetics studies in large populations (Casas *et al.* 2000). Casu *et al.* (2003) used a daughter design (DD) combined with EP and showed that number of genotyped individuals are lower in DD when combined with EP (50 and 25% of the population) with a reasonable power for intermediate QTL effects than using combined EP with grand DD. The choice of what fraction to genotype depends on the relative cost of phenotyping and genotyping (Darvasi 1997). It has been demonstrated that in EP, the power of QTL detection is at least as great as random genotyping (Ronin *et al.* 1998; Bovenhuis & Spelman 2000). Stella & Boettcher (2004) used 10 different strategies in genotyping and concluded that all strategies had the same precise estimates of the QTL position but they were better than random sampling from the population.

One disadvantage of EP is that linear model estimates of the QTL effects are conditional on the individuals with genotype information and this might cause a bias. One solution to solve this problem is to use a mixture model approach which was presented by Jansen *et al.* (1998) for the mapping of QTLs in an outbred population. Johnson *et al.* (1999) have simulated this approach in a half-sib family to demonstrate that estimates of the allelic effects of a QTL are unbiased not only for the main trait used to select individuals but also for a correlated trait when both traits were jointly analysed in a bivariate model. In this case, Markov chain Monte Carlo methods are appropriate for sampling missing data (for individuals without genotype information) and then all phenotype records in the population could be used for the mapping of QTL. In addition, Ronin *et al.* (1998) showed that it is possible to estimate unbiased parameters if all phenotype records for the trait under selection are included in the analysis.

All current traditional selective methods for QTL mapping experiments have used either phenotypic or genotypic information, even if a proportion of animals already have been genotyped. Therefore, QTL mapping studies are expected to be more powerful by utilizing both genotype and phenotype information in finding the most informative animals for genotyping.

The objective of the current study was to compare strategies for selective genotyping with respect to power of detecting a QTL in simulated data. New criteria based on both phenotypic and genotypic information are contrasted with traditional selective genotyping approaches.

Material and methods

Data simulation

Selective genotyping approaches were compared on simulated data and precision of the QTL mapping and power of detection related to the selective genotyping criteria were studied in a DD resembling, e.g. dairy cattle population.

Population structure and genetic model

The simulation of the parents (base population) was based on the method of Meuwissen & Goddard (2000), with discrete generation assumption to generate linkage disequilibrium. This method assumed that variation in a QTL is due to a mutation that occurred 100 generations ago due to random drift, given a specific effective population size. Therefore, according to this method and based on the genetic model described below, 100 generations of completely random mating were simulated with effective population size of 200. In generation 1, the genotypes were simulated for six marker loci equally spaced in a 50-cM chromosome segment and also an associated QTL placed in the midpoint between markers 3 and 4. A simple bi-allelic QTL and markers with five alleles were assumed. In the first generation, unique alleles were sampled and the frequencies of alleles in each marker were equal to 0.20. Initially in generation 1, the frequencies in the QTL alleles were the same and equal to 0.50. Marker and QTL genotypes (each animal with two unique alleles) were assigned according to Mendelian inheritance and allowed for recombination within the region. Animals of generation 100 were considered as unknown parents of generation 101. Genotypes of the progenies (generation 102) were sampled from the parental haplotypes (generation 101) allowing for recombination. A Poisson distribution, given the distance between markers or QTL, was used to determine the probability of an uneven number of crossovers. In generation 100, a mutant QTL allele was sampled at random with the requirement that the frequency (P_Q) was between 0.45 and 0.55, otherwise if P_Q was out of this range then a new LD population was simulated. Comparisons of the genotyping strategies were based on the generation 101.

Phenotype

One quantitative trait was simulated, so that the sum of QTL, polygenic and residual effects generated a trait (e.g. milk yield) with heritability of 0.25 and total phenotypic variance of one. The phenotype records were assigned to all the daughters in the analysis and not to sires. The variance due to the QTL was calculated as $V_{qtl}=2P_Q(1-P_Q)\delta^2$, where P_Q is the frequency of the favourable QTL allele in generation 100 and δ is the gene substitution effect. Proportion of QTL variance relative to total phenotypic variance was between 0.0620 and 0.0627 and therefore, allele substitution effect was 0.354. Polygenic effects were drawn from a normal distribution (ND), $ND(1/2a_s+1/2a_d,\ 1/2\sigma_a^2)$ where $a_s(a_d)$ is the polygenic breeding value of the sire (dam) and σ_a^2 is the polygenic variance (0.1874). For base animals, polygenic breeding values are sampled from $ND(0, \sigma_a^2)$. Residual effects are sampled from $ND(0,\sigma_e^2)$, where σ_e^2 is the residual variance (0.75).

Genotyping selection strategies

Four strategies for selective genotyping in a population which contains paternal half-sib family groups were defined based on a DD. Each strategy was used to select daughters for genotyping until 10, 20 and 30% of the 3600 daughters in the population were genotyped. All genotyping strategies were used for the daughters and it was assumed the sire's genotypes was available.

Random genotyping

Random selection of daughters for genotyping was used. The random strategy was practiced either randomly across family (RAN_A) or randomly within family (RAN_W).

Extreme phenotype genotyping

In this strategy, the daughters with EP were identified for genotyping either across sire families (EP_A) or within sire families (EP_W). EP_A strategy is known as conventional selective genotyping. The daughters with phenotypes in the highest and lowest percentages were selected for genotyping with equal proportions in each extreme.

Minimum error of QTL variance component (MinERR)

This approach uses all available phenotype and genotype information in selecting daughters for genotyping. Based on a mixed inheritance model with QTL as random effect (as it will be described later in Analysis), and via restricted maximum likelihood (REML), the asymptotic standard error of the estimated QTL variance component effect was computed, conditional on the QTL position. This standard error was derived from the second derivative of the likelihood function equation (3) and computed over all sire groups.

For the first level of genotyping (10%), genotype information from 2% of the extremes (4% in total) in each sire family (within family) was a starting point for the genotyping cycles. Then the next daughters for genotyping at each iteration cycle were chosen to minimize the standard error of the QTL variance component until 10, 20 or 30 genotyping levels were achieved.

In each iteration, and given all phenotypic and a fraction of genotypic information, first maximum likelihood tests were computed for the positions in the middle of the marker intervals (5, 15, 25, 35 and 45 cM), to identify the most likely position of the QTL. Then the standard error of the QTL variance component was computed through solving the mixed model equations conditional on the identified QTL position for all candidates. The candidates were the potential daughters for genotyping in each step of genotyping. Because identical by descent (IBD) computation is time consuming in solving the equations (2 1/2 h for 2% genotyping and about 30 h for each replicate), only the extremes of each sire family were considered as candidates (10 offspring per sire family). Therefore, depending on available daughters per sire, up to 300 candidates were chosen at each iteration. To assign genotypes for the candidates, paternal haplotypes were sampled with recombination, and maternal haplotypes were sampled from marker allele frequencies. Parents were assumed to be unrelated, and no previous generations were taken into account.

The IBD matrix was computed as a function of available marker data and the position of a putative QTL on the chromosome. This process was repeated for all candidates, one at a time. In each genotyping step, 72 offspring from 300 candidates (24% of the candidates) with the largest reduction in standard error of QTL variance component were genotyped. Therefore in each step, 2% of the daughters were genotyped. This approach was repeated to find further daughters (next 2%) for genotyping until 10, 20 and 30% of the population was genotyped.

Maximizing likelihood ratio test

Another criterion to use all available phenotypic and genotypic information in selecting the daughters for genotyping was based on maximum increase in likelihood ratio test (MaxLRT). The daughters were selected for genotyping based on the largest LRT in potentially selected daughters for genotyping. Thus, given the most likely QTL position, all phenotypic and genotypic information (including the daughters genotyped previously), plus the candidates genotype sampled as described in previous criterion, the criterion MaxLRT was computed through a mixed model approach. In each iteration of genotyping cycles, 2% of the candidates with the highest LRT were selected for genotyping. This criterion was conditional on the most likely QTL position in the middle of the marker intervals (5, 15, 25, 35 and 45 cM). This process of selective genotyping was continued until 10, 20 and 30% of the population was genotyped.

Analysis

In total, 100 data sets were simulated and each strategy used the simulated data as described above. After selection of the daughters to be genotyped, a mixed inheritance model with QTL as random effect (as below), was fitted and the variance component of QTL effect was computed, conditional on the QTL position on the identified segment of the chromosome:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{q} + \mathbf{e} \qquad (1)$$

where, $\mathbf{y}$ is an $(N \times 1)$ vector of phenotypes of the daughters and N is number of daughter (genotyped and non-genotyped animals) phenotypic records; $\mathbf{X}$ is an incidence matrix for β, which reduces to a vector on N ones; β is a vector of fixed effects, which reduces to the overall mean here; $\mathbf{Z}$ is an $(N \times p)$ incidence matrix relating phenotypes to random sire effects; $\mathbf{u}$ is a $(p \times 1)$ vector of additive polygenic effect for all the sires (polygenic effect), $\mathbf{W}$ is an $(N \times q)$ incidence matrix relating phenotypes to random QTL effects, $\mathbf{q}$ is a $(q \times 1)$ vector of random additive effects due to the QTL (haplotype effect) and $\mathbf{e}$ $(N \times 1)$ contains the residuals. The phenotypic variance of the observations is as follows:

$$\mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_u^2 + \mathbf{W}\mathbf{G}_p\mathbf{W}'\sigma_q^2 + R \qquad (2)$$

where, $\mathbf{A}$ is the numerator relationship matrix based on additive genetic relationships and reduces to an identity matrix here, $\mathbf{G}_\mathrm{p}$ is the matrix containing the IBD probabilities of a putative QTL at five positions between the six markers and $R = \mathbf{I}\sigma_e^2$ ($\mathbf{I}$ is an identity matrix).

Gametic relationship matrix at each putative QTL position ($\mathbf{G}_\mathrm{p}$) was derived with assumption about historical population structure and clustering approach for LDLA analysis across sire family groups using the method described by Meuwissen & Goddard (2000, 2001). Here it is assumed that marker data are available for all sires and a fraction of the daughters (depending on the genotyping level). The

linkage information is based on the IBD probability between the parental and offspring haplotype given that both animals are genotyped. Analyses of LDLA and comparisons of genotype strategies were based on the animals at the generation 101 and their daughters at generation 102, in a DD approach. The residual log likelihood under multivariate normality in the model equation (1) was:

$$\log L(\mathbf{G}_p, \sigma_u^2, \sigma_q^2, \sigma_e^2) \propto -\frac{1}{2}(\log|R| + \log|\mathbf{A}\sigma_u^2| + \log|\mathbf{G}_P\sigma_q^2| + \log|\mathbf{C}| + \mathbf{y}'R^{-1}\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'R^{-1}\mathbf{y} - \hat{u}'\mathbf{Z}'R^{-1}\mathbf{y} - \hat{q}'\mathbf{W}'R^{-1}\mathbf{y}) \quad (3)$$

where, **C** is the coefficient matrix of the mixed model equations from model (1) and the rest are as explained in Equations (1) and (2). Given IBD matrix in a putative QTL position, the log L was maximized using the Newton-Raphson algorithm to obtain variance components of random effects in model equation (1).

The DMU program package (Madsen & Jensen 2002) was used for the estimation of variance components. This program uses average information restricted maximum likelihood (AI-REML) algorithm for estimation of (co)variance components in mixed models and the restricted likelihood is maximized with respect to variance components associated with the random effects (Sørensen *et al.* 2003). A likelihood ratio test (LRT) was calculated as follows:

$$LRT = -2\,(\log \text{likelihood}\,(H_0) - \log \text{likelihood}\,(H_1)) \quad (4)$$

where, log likelihood (H_1) is the likelihood for a model with a QTL-effect and it is calculated for each bracket. Log likelihood (H_0) is based on a model excluding the QTL-effect(s). The LRT-statistic has a chi-square distribution with one degree of freedom (because of one QTL). The LRT statistic does not take into account that multiple tests are performed along the chromosome, but our simulations with no-QTL effect ($\text{Phen}_{\text{no\_QTL}}$) will provide an estimate of chromosome-wise false positive rate. The LRT statistic thresholds for significant ($p < 0.05$), and highly significant ($p < 0.01$) QTL effects were calculated for each data set using a quick approximate method according to the method of Piepho (2001). Power of detection was counted as the number of QTL detected at a given genomewide significance level (α), based on LRT. Accuracy of the position was the number of data sets where the estimated QTL position was in the bracket containing the QTL.

Results

Power in QTL detection

The number of simulations where the test statistic based on LRT was significant (p-value <0.01 and 0.05) is presented in Table 1 and partly in Figures 1 and 2. Substantial differences can be observed between random genotyping strategy and other strategies for selective genotyping. The power of detection for EP_A is much smaller than what was obtained for EP_W. Number of QTL at a significance level of 0.01 with 20% genotyping is found for MinERR to be 80% of that obtained with complete genotyping, and in MaxLRT and EP_W the relative power was 75 and 70% respectively. With 30% genotyping, these figures changed to 78, 83 and 78% respectively. Similar differences in power of detection were achieved at a significance level of 0.05 (Figure 1) with MinERR, MaxLRT and EP_W being superior to EP_W and random genotyping.

The percentage of simulations with a significant QTL (p-values <0.05 or 0.01) for no-QTL effect ($\text{Phen}_{\text{no\_QTL}}$), i.e. false positives, is given in Table 2. At a significance level of 0.05, five of 100 simula-

Table 1 Power of QTL detection and accuracy of the position[1] in different selective genotyping strategies with two levels of error type-I

Strategy	Genotype (%)	Detection ($\alpha = 1\%$)	Detection ($\alpha = 5\%$)	Position ($\alpha = 1\%$)	Position ($\alpha = 5\%$)
All[2]	100	88	95	86	95
RAN_W	10	2	7	0	5
RAN_W	20	3	19	1	10
RAN_W	30	14	31	10	24
RAN_A	10	0	7	0	5
RAN_A	20	4	17	3	14
RAN_A	30	11	29	8	22
EP_W	10	38	51	33	47
EP_W	20	62	72	54	67
EP_W	30	69	78	64	75
EP_A	10	8	18	7	18
EP_A	20	33	43	26	38
EP_A	30	51	58	43	53
START[3]	4	8	22	7	14
MinERR	10	46	63	35	55
MinERR	20	70	80	63	75
MinERR	30	69	82	67	77
MaxLRT	10	45	57	37	48
MaxLRT	20	66	74	62	70
MaxLRT	30	73	83	69	80

[1]Accuracy of QTL position shows number of datasets out of 100 where the estimated QTL was in the marker bracket containing the QTL.
[2]All genotypes from the offspring were included.
[3]First 4% in the strategies MinERR and MaxLRT is extreme phenotyping as start point.

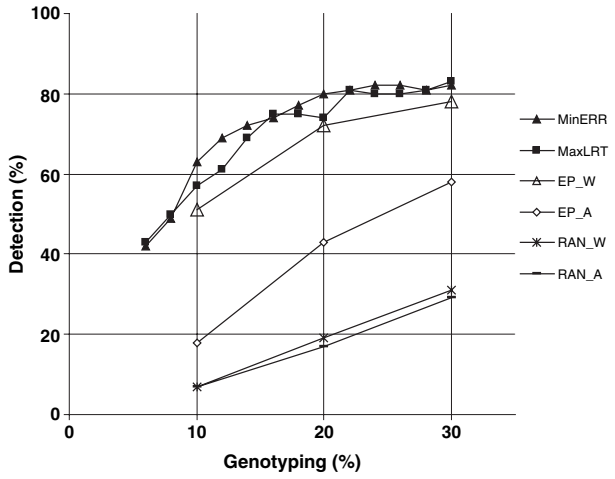

Figure 1 Power of quantitative trait loci (QTL) detection under different selective genotyping criteria with a significant QTL ($p < 0.05$).

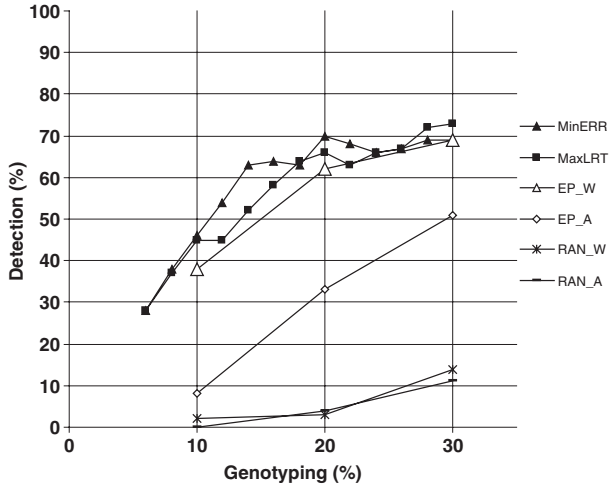

Figure 2 Power of quantitative trait loci (QTL) detection under different selective genotyping criteria with a highly significant QTL ($p < 0.01$).

tions are expected to show a significant QTL, and one of 100 simulations are expected to show a significant QTL at a significance level of 0.01. In almost all situations explored, the observed number was close to the expected number of false positive (1 or 5%) in Table 2. With low level of genotyping in EP methods (especially in the start point for MaxLRT and MinERR which is based on EP within family) the expected false numbers were exceeded. However, this effect disappeared with increasing the genotyped proportion. Fewer false positives were found for the 20 and 30% of genotyping in MinERR, when compared with MaxLRT and the EP strategies, and were close to the all genotyped progeny strategy.

Table 2 Frequency of false positives (out of 100) in QTL detection in different selective genotyping strategies with two levels of error type-I

Strategy	Genotype (%)	Detection ($\alpha = 1\%$)	Detection ($\alpha = 5\%$)
All[1]	100	1	1
RAN_W	10	0	0
RAN_W	20	0	0
RAN_W	30	0	1
RAN_A	10	2	4
RAN_A	20	0	1
RAN_A	30	0	1
EP_W	10	3	4
EP_W	20	1	4
EP_W	30	0	4
EP_A	10	1	5
EP_A	20	1	2
EP_A	30	2	4
START[2]	4	5	7
MinERR	10	4	8
MinERR	20	0	1
MinERR	30	0	1
MaxLRT	10	1	5
MaxLRT	20	0	5
MaxLRT	30	0	5

[1]All genotypes from the offspring were included.
[2]First 4% in the strategies MinERR and MaxLRT is extreme phenotyping as start point.

Accuracy of QTL position

The estimation of the correct location of the QTL is very important for the design of subsequent fine mapping experiments. Table 1 gives the number of datasets (of 100) where the estimated QTL position was in the bracket containing the QTL (25 cM). In total, the proportion of datasets with a significant QTL after 30% genotyping at correct location (simulated QTL position) ranged from 8 to 69 with $\alpha = 0.01$, and from 22 to 80 with $\alpha = 0.05$. The lowest accuracy in detecting the correct location was with random genotyping (both across and within sire family) and EP_A, in contrast with the other genotyping strategies. MaxLRT, MinERR and EP_W had a higher accuracy of QTL location (both at type I error levels of 1 and 5%) after 20% genotyping. Using a stringent threshold, e.g. 0.01, the correct position was found less frequently than for the 0.05 threshold. The strategies EP_W and EP_A (Figure 2) showed a large effect of sampling equally from all families when selecting EPs, whereas there were only a minor effect for random genotyping at all levels of selective genotyping. The means and standard error of the QTL position estimates (in cM) and statistic test criterion can be found in Table 3. The mean of position

Table 3 Means (±SE) of QTL position estimates in cM, and the test statistic criterion (likelihood ratio), based on 100 replicates in different selective genotyping strategies

Strategy	Genotype(%)	Position	Likelihood ratio test
True[1]		25.0	
All[2]	100	25.5 (±0.64)	24.7
RAN_W	10	23.4 (±1.46)	2.4
RAN_W	20	23.6 (±1.37)	3.7
RAN_W	30	26.0 (±1.23)	5.7
RAN_A	10	21.0 (±1.42)	2.3
RAN_A	20	23.0 (±1.42)	4.0
RAN_A	30	25.7 (±1.15)	5.7
EP_W	10	24.9 (±1.10)	10.5
EP_W	20	24.4 (±0.97)	16.1
EP_W	30	24.7 (±0.88)	19.3
EP_A	10	23.8 (±1.35)	4.7
EP_A	20	23.4 (±1.02)	8.3
EP_A	30	22.8 (±1.04)	12.0
START	4	22.5 (±1.44)	5.1
MinERR	10	24.3 (±1.17)	12.6
MinERR	20	24.2 (±0.97)	17.5
MinERR	30	24.3 (±0.86)	19.0
MaxLRT	10	23.3 (±1.12)	12.0
MaxLRT	20	23.9 (±0.90)	18.3
MaxLRT	30	25.0 (±0.84)	19.7

[1]True value of the position used in simulation.
[2]All genotypes from the offspring were included.

estimates over 100 replicates are approximately similar for the strategies used in this study and close to the true value position (with a standard error of 1–2.5 cM), except RAN_A at 10% genotyping. With regard to the standard error of the position estimates, MaxLRT, MinERR and EP_W are slightly superior at 20 and 30% genotyping (<1 cM). Likewise, the MinERR, MaxLRT and EP_W strategies have a higher LRT statistic compared with the others.

Variance components

A sire model (Equation 1) was applied to the simulated data to partition total variance into QTL, polygenic and residual components. Estimation of these parameters required an iterative solving of more than 7280 equations (depending on the strategy) at each QTL position (five points) and also for each candidate (300 candidates per each 2% of animals to be genotyped, as described in Material and methods). Thus, it was computationally intensive and time consuming and therefore, given all available genotyped individuals in each iteration of genotyping, the most likely position of QTL with the highest likelihood ratio value was used to estimate the variance components and their standard errors. The resulting estimates are shown in Table 4, as the average over 100 data sets. Gametic QTL variance was estimated as half of the total QTL variance (bi-allelic QTL was assumed). Polygenic variance (from sire model) is a quarter of the total additive genetic variance used in the simulation. Therefore, 3/4 of the additive genetic variance is a part of the estimated residual variance component.

Corresponding QTL variance component estimates for random genotyping (RAN_W and RAN_A) and extreme genotyping across families were significantly higher than the true QTL variance and estimates obtained from complete genotyping. With increasing genotype information (from 10 to 30%) in different putative QTL positions, the estimated QTL variance component decreased but the estimated polygenic variance increased. In contrast with genetic effects, the residual component remained relatively stable at all levels of genotyping regardless of the strategy used. However, the variance components due to the QTL by RAN_A, RAN_W and EP_A were still overestimates after 30% genotyping.

All genetic variance components were biased with EP_A, MinERR and MaxLRT at 10% genotyping. As it was expected, by increasing the proportion of genotyping in these strategies, the estimated QTL effect approaches the true value. In contrast to the QTL effects, the polygenic variances were underestimated at the low levels of genotyping. This quantifies overestimation of QTL effects induced by selection of the extremes or random samples. With 20% genotyping, the QTL variance component was not significant between the MinERR strategy and complete genotype information but MaxLRT and random genotyping methods were significantly different from that with complete genotyping. Variance component estimates in the MaxLRT strategy have been close to the estimated values of MinERR. Standard error of the components showed a similar trend. The standard error of the QTL variance component was always lowest in MinERR strategy and highest in random genotyping.

Discussion

The present study proposes selective methods to identify more informative individuals for genotyping given all available marker, pedigree and phenotype information in a DD half-sib family, for example dairy cows. Whereas detection and positioning of QTL utilize the linkage between marker and phenotype information, the EP method (currently used in

Table 4 Averaged QTL, polygenic (sire) and residual variance components (±SE) over all QTL positions for different levels of genotyping and different strategies

Strategy	Genotype(%)	QTL effect[1]	Sire effect	Residual
True[2]		0.0313	0.0469	0.8901
All[3]	100	0.0335 (±0.00134)	0.0444 (±0.00176)	0.9052 (±0.0024)
RAN_W	10	0.0494 (±0.00407)	0.0370 (±0.00225)	0.8890 (±0.0046)
RAN_W	20	0.0426 (±0.00286)	0.0398 (±0.00198)	0.8960 (±0.0035)
RAN_W	30	0.0383 (±0.00228)	0.0414 (±0.00187)	0.9003 (±0.0031)
RAN_A	10	0.0492 (±0.00399)	0.0366 (±0.00222)	0.8893 (±0.0045)
RAN_A	20	0.0423 (±0.00286)	0.0399 (±0.00199)	0.8961 (±0.0035)
RAN_A	30	0.0379 (±0.00227)	0.0420 (±0.00190)	0.9006 (±0.0031)
EP_W	10	0.0480 (±0.00242)	0.0389 (±0.00194)	0.8879 (±0.0032)
EP_W	20	0.0389 (±0.00165)	0.0426 (±0.00185)	0.8960 (±0.0027)
EP_W	30	0.0335 (±0.00134)	0.0444 (±0.00180)	0.9010 (±0.0025)
EP_A	10	0.0404 (±0.00295)	0.0413 (±0.00209)	0.8973 (±0.0036)
EP_A	20	0.0384 (±0.00215)	0.0419 (±0.00190)	0.8988 (±0.0029)
EP_A	30	0.0374 (±0.001801)	0.0427 (±0.00183)	0.8995 (±0.0027)
START	4	0.0587 (±0.00400)	0.0340 (±0.00226)	0.8785 (±0.0045)
MinERR	10	0.0473 (±0.00222)	0.0389 (±0.00190)	0.8882 (±0.0031)
MinERR	20	0.0352 (±0.00149)	0.0437 (±0.00182)	0.8994 (±0.0026)
MinERR	30	0.0288 (±0.00121)	0.0461 (±0.00180)	0.9057 (±0.0024)
MaxLRT	10	0.0458 (±0.00225)	0.0395 (±0.00192)	0.8898 (±0.0031)
MaxLRT	20	0.0364 (±0.00154)	0.0435 (±0.00184)	0.8980 (±0.0026)
MaxLRT	30	0.0306 (±0.00128)	0.0453 (±0.00180)	0.9038 (±0.0025)

[1]Variance component of the QTL is gametic QTL effect.
[2]True value of the position used in simulation.
[3]All genotypes from the offspring were included.

selective genotyping) only considers phenotypes. MinERR and MaxLRT in this study can be used to map QTL more accurately compared with EP and to decrease genotyping costs in QTL mapping experiments. The major restriction in QTL mapping and detection is the costs of collecting and typing of marker data and, therefore, sampling approaches for genotyping have been designed to provide considerable cost saving, particularly when the phenotype is routinely collected. Various strategies for selective genotyping have been proposed for human and animal QTL detection experiments using kinships (Cardon & Fulker 1994; Stella & Boettcher 2004). The principle underlying these methods is that the difference in phenotypes between pairs of sibs (within family) becomes larger as they share a decreasing number of QTL alleles transmitted from their parents (Chatziplis *et al.* 2001), and therefore in such families a QTL is more likely to be segregating. The current study showed that combining phenotype information with marker information would provide more accurate detection of QTL compared with only using the EPs. Additionally, to decrease the number of individuals in genotyping, MinERR (or MaxLRT) could be considered. MinERR was an alternative approach to decrease genotyping costs and needed to genotype only 20–25% of animals to achieve unbiased parameters compared with complete genotyping. The power with 20% genotyping for MinERR and MaxLRT was 80 and 75% of that obtained with complete genotyping compared with 70 and 38% with EP within and across families respectively. Higher power to detect QTL in the strategies based on both phenotypic and genotypic information (MinERR and MaxLRT) compared with only considering the phenotypic information (EP methods) and also random approach at the same level of genotyping is due to selection of more informative individuals for genotyping. In MinERR and MaxLRT, daughters from all sires were potential candidates for genotyping but some sire's daughters were never included in genotyping. Therefore, increased power is also due to selection of daughters for genotyping from segregating sires. Ultimately, this means more genotyped daughters from segregating sires to contribute information to contrasts between marker haplotypes. However in EP_W strategy, power was not usually significantly different compared with MinERR and MaxLRT. In this study, it was assumed that frequency of the mutant QTL allele was between 0.45 and 0.55. Therefore, the sires were more likely to be heterozygous compared with lower frequencies for favourable QTL allele. Using lower frequency could affect the within-family genotyping strategies and decrease the power of detection as heterozygous sires (informative sires) will be rarer

than in the current study. Besides in MinERR and MaxLRT, the candidates were genotyped as a group (2% at a time) to make starting more realistic. Therefore, if it can be possible to genotype the candidates one by one, it is expected that the power of MinERR and MaxLRT would be even larger than the powers obtained here.

QTL parameters indicated that application of EP selection strategies is superior to random genotyping of individuals, as also shown in previous experiments (e.g. Stella & Boettcher 2004). With increasing level of genotyping, both EP across and EP within family, the EP approach has significantly increased power of QTL detection and accuracy of QTL position. This shows that EP methods in the current study discriminate between the presence and absence of a segregating QTL in the families. Sampling induces a correlation between estimated residual and QTL effects. This correlation increased when the extreme daughters of all sires were selected for example in EP_A. It indicates that in the extremely selective genotyping approaches there is a real probability that an erroneous conclusion can be drawn about the location. In the EP methods, estimated effects are biased upwards when genotyping the extremes within the population or low proportion (<20%), even if all available phenotype information is used. This phenomenon might be due to the positive correlation between residual effects and the QTL effect in the individuals sampled for genotyping that magnifies the allelic effect. MinERR and MaxLRT obtain a power of detecting a QTL of approximately 70% at genotyping 20–30% compared with a power of 88% with complete genotyping. These two criteria showed increased power compared with random genotyping and genotyping the extremes across families, and marginally increased power compared with genotyping extremes within family. Power of a selective genotyping approach to detect a QTL can be affected by several factors such as the number of genotyped animals, the effect of segregating QTL (δ) and heritability of the trait. Stella & Boettcher (2004) showed that EP strategy was more precise than random genotyping when the heritability and QTL variance of the trait were low. However, lower power is expected in EP genotyping because in these situations (low heritability and QTL variance), the phenotypes provide relatively little information about the genotypes and EP method is based on the phenotype records. The combined strategies in this study could also use genotypic information and, therefore, they might be more useful for selective genotyping of a low heritable trait.

Estimated variance component from the MaxLRT and MinERR strategy differed only slightly. The QTL position from the MaxLRT strategy was more accurate compared with the MinERR method and therefore MaxLRT could be used to fine map a QTL. However, with using the EP strategy within families at 30% genotyping, estimated variance components (QTL and polygenic effects) were not significantly different from those obtained with complete genotype information. Besides, as proportion genotyped increased (10 to 30%) in both MinERR and MaxLRT, standard error of the estimated QTL variance component decreased, compared with other selective strategies in this study.

The genetic variance explained by QTL in this experiment ($1/4\sigma_G^2$) was in the range given by Druet *et al.* (2006). They reported that the proportion of genetic variance explained by a QTL was up to 36.0% for dairy traits in a Holstein population. The variance of QTL is a function of allele frequencies and the QTL substitution effect. The proportion of segregating sires (heterozygous sires) for QTL also is a function of QTL allele frequencies and therefore directly influences the QTL variance. Decreasing the variance of QTL by using lower frequency of the favourable QTL allele (e.g. $P_Q < 0.3$) will affect the efficiency of assessing which sires are heterozygous. If some sires are not informative, it is not possible to determine whether they are heterozygous. Therefore, the estimated proportion of heterozygous sires will be underestimated.

MinERR and MaxLRT methods do not come without disadvantages. These strategies have higher computational requirements than EP and random genotyping. The most time-consuming part was the likelihood maximization, because MinERR and MaxLRT strategies are based on the standard error of estimated variance component of QTL effect and the maximum likelihood function of the model respectively. However, if there is no possibility to use these strategies, genotyping based on the phenotypes and using the extremes within family can be suggested to obtain a powerful approach in QTL detection with unbiased estimates with at least 30% genotyping.

The current study considered a single trait and one QTL on a specific chromosome segment. If either more than one correlated trait or QTL is considered, some decrease in the selection intensity of the samples may secure sufficient power of detection for all traits. The reverse way can be possible when traits are uncorrelated. This could be an interesting area for further study.

Conclusion

MinERR and MaxLRT, can be used as approaches in selecting individuals for genotyping, and the results from simulation in this study showed that unbiased gene substitution effects could be estimated with only 20% of the animals genotyped. QTL position parameters from MaxLRT strategy were more accurate compared with MinERR method and therefore it should be preferred. MinERR method decreased the level of genotyping (up to 22% compared with 30% in extreme phenotyping) to increase the power and unbiased estimation of QTL parameters.

Acknowledgements

This research was supported by a grant from Agriculture Research and Education Organization of the Ministry of Jahad-e-Agriculture, Iran. The authors are grateful to Peter Sørensen and Hauke Thomsen for using the IBD program, and referees for many helpful suggestions.

References

Bovenhuis H., Spelman R.J. (2000) Selective genotyping to detect quantitative trait loci for multiple traits in outbred populations. *J. Dairy Sci.*, **83**, 173–180.

Cardon L.R., Fulker D.W. (1994) The power of interval mapping of quantitative trait loci, using selected sib pairs. *Am. J. Hum. Genet.*, **55**, 825–833.

Casas E., Shackelford S.D., Keele J.W., Stone R.T., Kappes S.M., Koohmaraie M. (2000) Quantitative trait loci affecting growth and carcass composition of cattle segregating alternate forms of myostatin. *J. Anim. Sci.*, **78**, 560–569.

Casu S., Carta A., Elsen J.M. (2003) Strategies to optimize QTL detection designs in dairy sheep populations: The example of the Sarda breed. *Options Mediterraneennes*, **A-55**, 19–54.

Chatziplis D.G., Hamann H., Haley C.S. (2001) Selection and subsequent analysis of sib pair data for QTL detection. *Genet. Res.*, **78**, 177–186.

Darvasi A. (1997) The effect of selective genotyping on QTL mapping accuracy. *Mamm. Genome*, **8**, 67–68.

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

Druet, T., Fritz S., Boichard D., Colleau J.J. (2006) Estimation of genetic parameters for quantitative trait loci for dairy traits in the French Holstein population. *J.Dairy Sci.*, **89**, 4070–4076.

Jansen R.C., Johnson D.L., J.A.M. van Arendonk (1998) Mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. *Genetics*, **148**, 391–399.

Johnson D.L., Jansen R.C., J.A.M. van Arendonk (1999) Mapping quantitative trait loci in a selectively genotyped outbred population using a mixture model approach. *Genet. Res.*, **73**, 75–83.

Kinghorn B.P. (1997) An index of information content for genotype probabilities derived from segregation analysis. *Genetics*, **145**, 479–483.

Kinghorn B.P. (1999) Use of segregation analysis to reduce genotyping costs. *J. Anim. Breed. Genet.*, **116**, 175–180.

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

Macrossan P.E. (2004) Strategies to Minimise DNA Testing Costs for Research and Development Programs Involving Pedigreed Populations. PhD Thesis. University of New England, Australia.

Madsen P., Jensen J. (2002) A User's Guide to DMU. A Package for Analysing Multivariate Mixed Models. Version 6, release 4.4. Danish Institute of Agricultural Sciences, Tjele, Denmark.

Martinez M.L., Vukasinovic N., Freeman A.E., Fernando R.L. (1998) Mapping QTL in outbred populations using selected samples. *Genet. Sel. Evol.*, **30**, 453–468.

Meuwissen T.H.E., Goddard M.E. (2000) Fine scale mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, **155**, 421–430.

Meuwissen T.H.E., Goddard M.E. (2001) Prediction of identity by descent probabilities from marker-haplotypes. *Gen. Sel. Evol.*, **33**, 605–634.

Muranty H., Goffinet B. (1997) Selective genotyping for location and estimation of the effect of a quantitative trait locus. *Biometrics*, **53**, 629–643.

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

Ronin Y.I., Korol A.B., Weller J.I. (1998) Selective genotyping to detect quantitative trait loci affecting multiple traits: interval mapping analysis. *Theor. Appl. Genet.*, **97**, 1169–1178.

Sørensen P., Lund M.S., Guldbrandtsen B., Jensen J., Sorensen D. (2003) A comparison of bivariate and univariate QTL mapping in livestock populations. *Genet. Sel. Evol.*, **35**, 605–622.

Stella A., Boettcher P.J. (2004) Optimal designs for linkage disequilibrium mapping and candidate gene association tests in livestock populations. *Genetics*, **166**, 341–350.

Van Gestel S., Houwing-Duistermaat J.J., Adolfsson R., van Duijn C.M., Van Broeckhoven C. (2000) Power of selective genotyping in genetic association analyses of quantitative traits. *Behav. Genet.*, **30**, 141–146.