Genet. Res., Camb. (2001), **78**, *pp.* 177–186. *With 7 figures.* © 2001 Cambridge University Press
DOI: 10.1017/S0016672301005225 Printed in the United Kingdom

Selection and subsequent analysis of sib pair data for QTL detection

DIMITRIOS G. CHATZIPLIS†, HENNING HAMANN AND CHRIS S. HALEY*
Roslin Institute (*Edinburgh*), *Roslin*, *Midlothian*, *EH25 9PS*, *UK*

(*Received 3 February 2000 and in revised form 26 June 2000 and 2 April 2001*)

Summary

Haseman and Elston (1972) developed a robust regression method for the detection of linkage between a marker and a quantitative trait locus (QTL) using sib pair data. The principle underlying this method is that the difference in phenotypes between pairs of sibs becomes larger as they share a decreasing number of alleles at a particular QTL identical by descent (IBD) from their parents. In this case, phenotypically very different sibs will also on average share a proportion of alleles IBD at any marker linked to the QTL that is lower than the expected value of 0·5. Thus, the deviation of the proportion of marker alleles IBD from the expected value in pairs of sibs selected to be phenotypically different (i.e. discordant) can provide a test for the presence of a QTL. A simple regression method for QTL detection in sib pairs selected for high phenotypic differences is presented here. The power of the analytical method was found to be greater than the power obtained using the standard analysis when samples of sib pairs with high phenotypic differences were used. However, the use of discordant sib pairs was found to be less powerful for QTL detection than alternative selective genotyping schemes based on the phenotypic values of the sibs except with intense selection, when its advantage was only marginal. The most effective selection scheme overall was the use of sib pairs from entire families selected on the basis of high within-family variance for the trait in question. There is little effect of selection on QTL position estimates, which are in good agreement with the simulated values. However, QTL variance estimates are biased to a greater or lesser degree, depending on the selection method.

1. Introduction

The mapping of loci affecting quantitative traits (quantitative trait loci or QTLs) in livestock provides a potential tool for genetic improvement through marker-assisted selection as well as a route to the ultimate cloning of the underlying genes. In many species, including some livestock, crosses between genetically divergent lines provide the basis for a QTL mapping study. However, such crosses are often not possible or practicable in livestock and QTL detection within a single outbred population may be the preferred option. QTL detection in outbred populations is problematical since markers and QTLs are segregating in the population and linkage phases will differ from family to family. One approach, which has been widely used for studies of QTLs in human populations, where these same problems exist, has been sib pair analysis (Haseman & Elston, 1972).

In sib pair analysis a relationship between the differences between the phenotypes of sib pairs and the number of alleles shared identical by descent (IBD) at a marker locus provides a means for identifying a QTL near that marker. A simple test of this association is the regression of squared phenotypic difference between a sib pair on the proportion of alleles the sib pair shares IBD at a marker locus. When a QTL is a recombination fraction θ from a marker, the expectation of the regression coefficient (β) is

$$\beta = -2(1-2\theta)^2\sigma_g^2,$$

where σ_g^2 is the additive variance explained by the QTL when there is no dominance variation present (Haseman & Elston, 1972). A significant negative

* Corresponding author. Department of Genetics and Biometry, Roslin Institute, Roslin, Midlothian EH25 9PS, UK.
Tel: +44(0)131 527 4462. Fax: +44(0)131 440 0434.
e-mail: chris.haley@bbsrc.ac.uk
† Present address: Ross Breeders Ltd, Newbridge, Midlothian EH28 8SZ, UK.

regression coefficient, as tested by a simple *t* statistic, indicates linkage between the marker and a QTL.

The attraction of sib pair analysis is its simplicity, and hence computational rapidity, and its relative robustness compared with more fully parameterized approaches. However, sib pair analysis may be less powerful than other more parameterized methods, especially in the case of small family sizes (small number of available sib pairs). Consequently, a large number of sib pairs is needed to achieve adequate power for the detection of QTLs (Blackwelder & Elston, 1982). In livestock populations, however, the large full-sib families that may be available can provide large numbers of sib pairs, making the method attractive for use in these populations (Götz & Ollivier, 1992). This is especially so since the non-independence of sib pairs in large sibships does not have a major adverse effect on power (Blackwelder & Elston, 1982; Götz & Ollivier, 1992; Wan *et al*., 1997; Chatziplis, 1998). Power for QTL detection is largely determined by the total number of sib pairs and not the size of the families from which they are sampled. Consequently, the number of individuals that need to be genotyped to achieve adequate power of detection is reduced when large families are used.

The use of samples of sib pairs selected on their phenotype has been suggested as a means to reduce the amount of genotyping for a given power of QTL detection (Darvasi & Soller, 1992; Mackinnon & Georges, 1992; Risch & Zhang, 1995). Phenotypically different sibs are expected to share a smaller (and similar sib pairs a greater) than average proportion of alleles IBD at markers linked to any segregating QTL. Hence, sib pairs can be selected for analysis based on their trait values and inspected to see whether the proportion of marker alleles inherited IBD deviates from expectation, for example using chi-square analysis (Eaves & Meyer, 1994). However, this analysis is unable to provide estimates of QTL position and variance explained by the QTL. The usual Haseman & Elston (1972) analysis loses power with data selected to contain only phenotypically different sib pairs, because of the very limited range of relationships between IBD status and phenotypic difference. In this paper, we develop a method for analysing data from selected full sib pairs and compare the power of detection and parameter estimates obtained from sib pair analysis using different selective genotyping methods in large full-sib families. In another paper (Chatziplis & Haley, 2000) alternative selective genotyping schemes in a joint analysis of full and half-sib data are compared for power of QTL detection.

2. Materials and methods

In the traditional Haseman & Elston (1972) regression method, one is looking for a relationship between differences between the phenotypes of sib pairs and the number of alleles shared IBD at a marker, as a means of identifying a QTL near that marker. The *t*-value of the regression coefficient is used as the test statistic and under the assumption of no dominance variation we have

$$\mathrm{E}(Y_j|\hat{\pi}_{j_m}) = [\sigma^2_e + 2(1-2\theta+2\theta^2)\sigma^2_g] - 2(1-2\theta)^2\sigma^2_g\hat{\pi}_{j_m}, \quad (1)$$

where Y_j is the squared phenotypic difference of the jth sib pair, $\hat{\pi}_{j_m}$ is the estimated proportion of alleles shared IBD at a marker locus m of the jth sib pair, σ^2_e is the residual variance due to environmental variance and covariance of full sibs and any other effect, σ^2_g is the variance due to the QTL and θ is the recombination rate between the marker and the QTL.

However, in a full-sib pair, the expected proportion of alleles at any locus IBD from parents is 0·5 (Amos & Elston, 1989). Phenotypically very different (similar) full-sibs on average share a proportion smaller (greater) than 0·5 of their genes IBD at a QTL which is influencing the phenotype in question as well as at linked markers.

With selected sib pairs, instead of regressing the squared phenotypic differences onto the estimated proportion of alleles IBD at a marker locus, one can invert the regression. In this analysis in practice, the proportion of alleles shared IBD at a marker locus of every sib pair in a selected sample is estimated and their squared phenotypic differences is calculated. Then the expected proportion shared IBD (0·5) is subtracted from the estimated proportion of alleles shared IBD of every sib pair in the selected sample. The adjusted value for every sib pair is then regressed onto the deviations of their squared phenotypic differences from the mean (squared) phenotypic difference of all the available sib pairs from which the sample has been drawn (see Appendix). Then:

$$\hat{\pi}_{j_m} - 0{\cdot}5 = (\mathrm{E}(\pi_m) - 0{\cdot}5) - \frac{2(1-2\theta)^2\sigma^2_g\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}}(Y_j - \mathrm{E}(Y)), \quad (2)$$

where $\mathrm{E}(\pi_m)$ is the expected proportion of alleles shared IBD, of selected sib pairs, at a marker locus m, $\mathrm{E}(Y)$ is the expected squared phenotypic differences of *all* sib pairs, $\sigma^2_{\hat{\pi}_{j_m}}$ is the variance of the proportion of alleles shared IBD at a marker locus m and $\sigma^2_{Y_j}$ is the variance of squared phenotypic differences.

The above equation (1) is of the general linear form:

$$Y = \alpha + \beta X,$$

where

$$\alpha = (E(\pi_{j_m}) - 0{\cdot}5)$$

https://doi.org/10.1017/S0016672301005225 Published online by Cambridge University Press

and

$$\beta = \frac{2(1-2\theta)^2\sigma_g^2\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}}.$$

Consequently, a significantly negative regression constant (α) provides evidence for the presence of a QTL linked to a marker when a sample of sib pairs with high phenotypic differences is examined. A significantly positive regression constant would indicate the presence of linkage when the sample of sib pairs examined has low phenotypic differences. In the absence of linkage, a zero regression constant is expected. The analysis potentially provides three test statistics for the detection of QTLs linked with markers: firstly, the commonly used *t*-values of the regression coefficient (Haseman & Elston, 1972); secondly, the *t*-value of the regression constant; and thirdly, the *F*-value of the regression analysis with two degrees of freedom, which is the joint test of both the regression coefficient and the regression constant. The three test statistics are compared for power of detection, as will be described in a later section.

The regression coefficient (β) can provide parameter estimates for the recombination rate and the variance due to the QTL in the same way as in the ordinary regression method (Hamann & Götz, 1995), since $\sigma^2_{\hat{\pi}_{j_m}}$ and $\sigma^2_{Y_j}$ can be calculated from the data (Amos *et al*., 1997; for fully informative markers, $\sigma^2_{\hat{\pi}_{j_m}} = 1/8$ (Amos & Elston, 1989)).

For example, if the markers with the highest *t*-values of their regression coefficients are known to be linked (θ_t), the three parameters (recombination rates between the markers and the QTL (θ_1, θ_2) and QTL variance (σ_g^2)) can be estimated from the regression coefficients of the two markers (β_1, β_2) and Haldane's map function (Hamann & Götz, 1995). For example by simultaneously solving the equations

$$\mathrm{E}(\beta_1) = \frac{2(1-2\theta_1)^2\sigma_g^2\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}}$$

$$\mathrm{E}(\beta_2) = \frac{2(1-2\theta_2)^2\sigma_g^2\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}}$$

$$\theta_t = \theta_1 + \theta_2 - 2\theta_1\theta_2,$$

solutions for θ_1, θ_2 and σ_g^2 can be obtained. Liang *et al*. (2000) give an alternative approach to estimate QTL position in selected samples.

(i) *Simulations*

For the comparison of the test statistics in the inverted regression (2) and the power comparisons of alternative genotyping schemes, data sets with one marker with relatively low information content (2 alleles) completely linked ($\theta = 0$) with a biallelic QTL were simulated. For the comparison of parameter estimates (QTL position and effect), data sets were simulated with a 100 cM chromosome with markers spaced at 10 cM intervals, each marker having 8 alleles at equal allelic frequencies (0·125).

For most studies, 100 independent families of family size 8 were simulated. This resulted in 800 progeny with 2800 sib pairs in total. To investigate the power of the methods with a smaller family size, the above simulations were repeated with 200 families of size 4. This resulted in 800 progeny with 1200 sib pairs in total.

In all cases the phenotypic values of a quantitative trait with only additive effect were determined by: (*a*) the biallelic QTL, (*b*) a polygenic effect created by 10 additional loci independent of each other and of the QTL (i.e. unlinked) and (*c*) an environmental component.

The total phenotypic variance (σ_g^2) of the trait was 1·0, with a trait heritability of 0·4 and variance explained by the QTL 28% of the total phenotypic variance ($0{\cdot}28\sigma_p^2$). The parents were mated at random to produce the offspring generation used in the analysis. Linkage equilibrium was assumed.

The simulated data were generated and analysed using programs written in FORTRAN 77, supplemented with routines from the NAG library (Numerical Algorithms Group 1990) for the random number generator (RAN2) and for simple linear regression (G02CAF).

(ii) *Sib pair selection*

Several strategies of selection of sib pairs were simulated as outlined below.

(a) *Most different sib pairs* (*MD*). First, using the distribution of the phenotypic differences between pairs, a percentage of the most different sib pairs (MD) were selected. Thus the phenotypic differences between all pairs of sibs were calculated and ranked and the desired proportion of sib pairs with the largest differences were selected.

(b) *Discordant sib pairs* (*D*). Second, using the distribution of the phenotypes of all sibs, sib pairs that had one sibling at one end of the distribution, and the other sibling at the other end, were selected. Thus equal proportions of individuals were selected from the two ends of the phenotypic distribution and sib pairs with an individual in either tail were used. This means that different families were represented in the sample with different numbers of sib pairs. Risch & Zhang (1995) have also used discordant sib pairs for sampling from families of size 2.

The above selection schemes using sib pairs with high phenotypic differences (D and MD) were used for the comparison of the three test statistics of the inverted regression (2).

(c) *Concordant and discordant sib pairs* (*CD*). Individuals in the population were ranked according

https://doi.org/10.1017/S0016672301005225 Published online by Cambridge University Press

to their phenotype. An equal number of individuals were selected from the upper and lower tails of the distribution. Sib pairs were used which had both siblings in the same (upper or lower) tail or one sibling in the upper tail and the other in the bottom tail of the phenotypic distribution of sibs. This selection means that different families were represented in the sample with different numbers of sib pairs. The truncation points of selection for the two tails were kept equal for both tails of the distribution. Note that with the lowest intensity of selection (50%) all individuals were selected and all sib pairs were used in the analysis. Gu *et al.* (1996) have also used similar methods for concordant and discordant sampling.

(d) *Within-family variance* (*WFV*). The within-family phenotypic variance was calculated for each full-sib family. Families with the highest within-family variance were selected for the analysis. All the possible sib pairs from selected families were included in the analysis. It should be noted that in the case of families of size 2, this selection method is the same as selecting discordant sib-pairs.

(e) *Random sample* (*RS*). Entire full-sib families were selected at random and all their sib pairs were used in the analysis.

The discordant selection (D), using the inverted regression method of analysis, was compared for power of detection with three additional selection schemes (CD, WFV and RS) using the standard Haseman & Elston (1972) analytical method.

For all methods, selection was applied to a population of fixed size, with proportions selected being varied to get the desired number of sib pairs. Note that with these selection schemes, the number of sib pairs selected vary from family to family and some sibs are involved in more than one sib pair.

(ii) *Analyses*

The power of the three test statistics (*t*-value of the regression constant, *t*-value of the regression coefficient and *F*-value of the regression) of the inverted regression method (2) under alternative methods of selection (MD and D) were compared in simulated data with a QTL completely linked to a single marker. The proportion of alleles shared IBD at the markers was estimated according to the algorithm described by Haseman & Elston (1972).

For the power comparisons of alternative selective genotyping schemes, two different methods of analysis were used depending on how the sib pairs had been selected. For the concordant-discordant (CD), within-family variance (WFV) and random selection (RS) schemes the standard Haseman & Elston (1972) analysis (1) was used. In this analysis, the squared difference in phenotypic scores between sib pairs is regressed onto the estimated proportion of alleles IBD at a marker. Linkage of a QTL to the marker is expected to cause a negative slope in this regression, and hence the *t*-value of the regression coefficient provides a test statistic for the presence of a linked QTL.

With selection of discordant sib pairs (scheme D) we used the inverted regression (2) as described above. This analysis was not used in the other selective genotyping schemes because, when selecting for both high and low phenotypic differences (CD) or for high within-family variance (WFV), the above test would not be as powerful as the traditional test of Haseman & Elston (1972). Since the mean expected proportion of alleles shared IBD at marker locus would be 0·5 if both high and low phenotypically different sib pairs were included in the sample, a test of significance for the regression constant would be inappropriate. Consequently, in these selection schemes (CD, WFV, RS), a significant negative *t*-value of the regression coefficient is used for the detection of linkage. Again the comparisons were made using simulated data with a QTL completely linked to a single marker.

In order to investigate more generally the power of the method and the parameter estimates obtained, analysis of data simulated with a 100 cM chromosome with equally spaced markers was used.

The parameter estimates for position and effect can be obtained from the estimated regression coefficients of markers flanking the QTL (Hamann & Götz, 1995) for the traditional Haseman & Elston (1972) regression method as described previously. In all cases the proportion of alleles shared IBD in any marker locus was estimated using a FORTRAN 77 program developed by the authors which was based on the algorithm described by Haseman & Elston (1972).

(iii) *Significance thresholds*

The empirical thresholds, at the 5% level of significance, were obtained for every level of selection (sample selection intensity) and for the different selection methods by simulations under the null hypothesis ($\sigma^2_{\mathrm{QTL}} = 0$) using 1000 replicates.

3. Results

(i) *Comparison of test statistics*

The power obtained from the three test statistics (the *t*-value of the regression coefficient and the regression constant and the *F*-value of the regression) of the single marker analysis are presented in Figs. 1 and 2 for the D and MD methods, respectively. As is apparent from Fig. 1 (D), the *t*-value of the regression constant provided a more powerful test statistic for selected samples than the other two methods. The power obtained from the *t*-value of the regression

https://doi.org/10.1017/S0016672301005225 Published online by Cambridge University Press

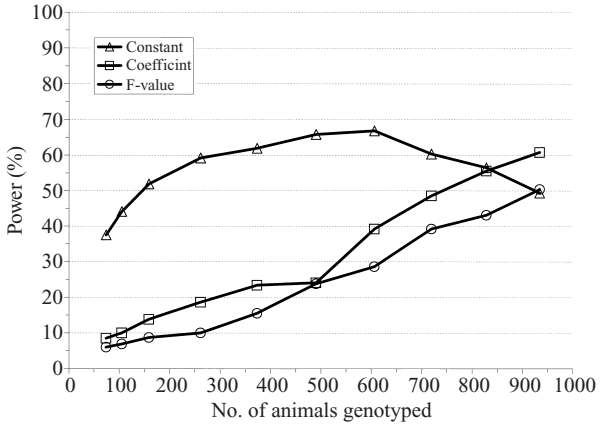

Fig. 1. Power obtained from the three test statistics in different sample sizes selected for discordant sib pairs (D). The graph shows power obtained from the *t*-values of the regression constant (triangles), the *t*-values of the regression coefficient (squares) and the *F*-values of the regression (circles). Data were simulated with one QTL of large effect ($\sigma^2_P = 1$, $h^2 = 0{\cdot}4$, $\sigma^2_{QTL} = 0{\cdot}28\sigma^2_p$) completely linked to a biallelic marker. The power was obtained from 1000 replicates and expressed as the percentage of replicates exceeding the simulated empirical threshold in each case. The population size simulated and hence the maximum number of animals genotyped was 1000 (800 progeny and 200 parents; family size 8).

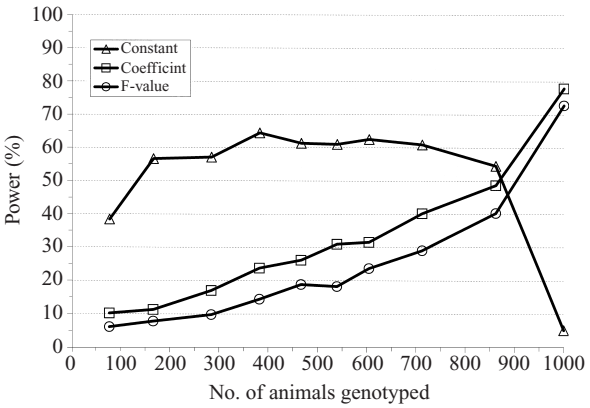

Fig. 2. Power obtained from the three test statistics in different sample sizes selected for the most different sib pairs (MD). Symbols, simulated population size and parameters are as in Fig. 1.

coefficient and *F*-value of the regression increased with sample size. The power obtained from the regression constant decreased with both selection schemes as the proportion of the population selected approached 1.

Similarly in Fig. 2 (MD), the *t*-value of the regression constant provided a more powerful test in selected samples than the other two test statistics. The power using the *t*-value of the regression coefficient and *F*-value of the regression increased with sample size. The power based on use of the regression constant dropped to the 5% threshold when all sib pairs were included. The difference between the MD

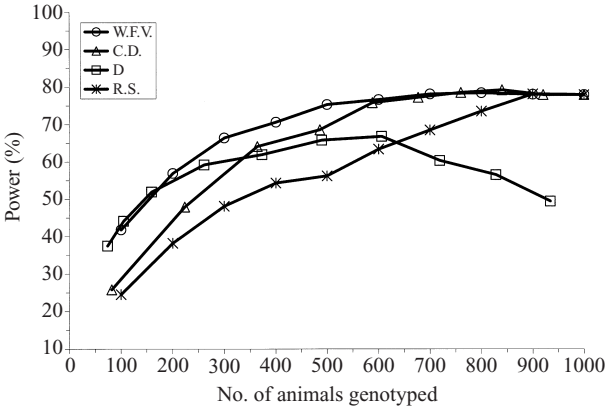

Fig. 3. Power of alternative selective genotyping schemes at different selection intensities. Simulated population size and parameters are as in Fig. 1.

and D approaches in this respect is that in the MD approach all sib pairs are included in the analysis, so the overall proportion of alleles IBD should be at the expectation of 0·5, other than deviations due to sampling. In the D approach, even when all sibs in both tails are potentially included in the analysis, only sib pairs with a sib in each tail (i.e. above and below the mean in this case) are actually included in the analysis.

Overall, the MD and D approaches produced quite similar results. The maximum power of the MD approach was maintained over a greater range of selection than the D approach. On the other hand, the maximum power achieved was greater for the D approach. The D approach was chosen for the further comparisons with other selection schemes.

(ii) *Comparison of selective genotyping schemes*

The power of detection of a QTL of large effect by linkage to a single marker with two alleles under the four selection schemes is shown in Fig. 3. Selecting the sample of sib pairs from families with high within-family variance was either the most powerful selection method for a given sample size or was close to the most powerful method. With this selection scheme, only 40–50% of individuals need to be genotyped to achieve power similar to that obtained when all individuals were genotyped. At high selection intensities (10–20% of the population genotyped), the power of the selection schemes using only sib pairs with high phenotypic differences (D, discordant sib pairs) was marginally the greatest, although selection based on the within-family variance (WFV) gave very similar power. At moderate selection intensities (30–50% of the population genotyped), selection on the within-family variance was better than any other method. At low selection intensities (> 50% of the population genotyped) the within-family variance and concordant–discordant selection (CD) schemes gave

https://doi.org/10.1017/S0016672301005225 Published online by Cambridge University Press

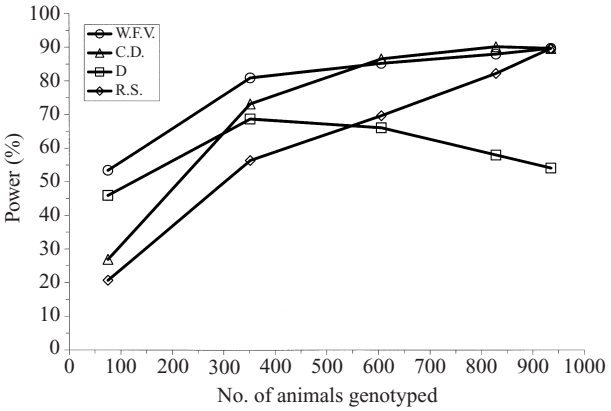

Fig. 4. Power of alternative selective genotyping schemes at different selection intensities. Simulated population size and parameters are as in Fig. 1 except that the QTL was placed on a 100 cM chromosome carrying 11 markers.

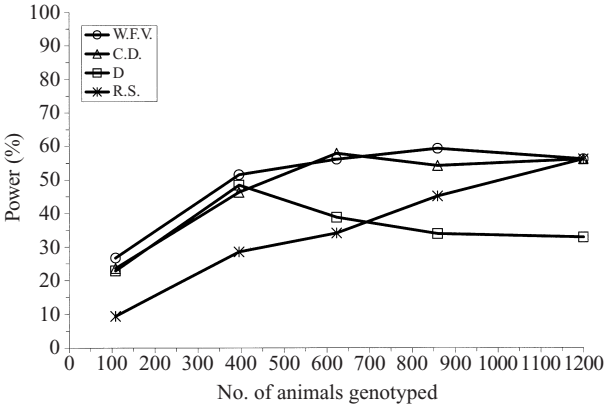

Fig. 5. Power of alternative selective genotyping schemes at different selection intensities. Simulated parameters are as in Fig. 1. The population size simulated and hence the maximum number of animals genotyped was 1000 (800 progeny and 400 parents; family size 4).

very similar power and the power of the discordant scheme declined. Similar results were obtained from analysis of data with a QTL of large effect on a 100 cM chromosome (Fig. 4).

Results obtained with families of size 4 using a 100 cM chromosome are shown in Fig. 5. With families of size 4 the total number of offspring is the same as in families of size 8, but the maximum number of sib pairs drops from 2800 to 1200, with a commensurate drop in the power achievable. As previously, the highest power overall was obtained when sib pairs from selected families with high within-family variance were used and approaching maximum power could be attained with only half the population being genotyped.

(iii) *Parameter estimates*

The parameter estimates (position and QTL variance based on data simulated with the 100 cM chromosome) are presented as the mean of all 1000 replicates, whether or not the QTL detected in a particular replicate was deemed significant. This was done in order to eliminate selection bias, which can result if only significant replicates are selected.

When the larger family size was used (8 sibs per family), the position estimates were generally near the simulated values with selected samples down to 60% of the total number of animals (Table 1). There was a general tendency for position estimates to be biased towards the centre of the chromosome. With the most intense selection, the mean estimated position was outside the correct interval, except in the case of samples selected for high within-family variance. When samples of discordant sib pairs were used the position estimates were not improved when more individuals are genotyped (decreased selection in-

Table 1. *Mean QTL position estimates in centimorgans and QTL variance estimates as the proportion of the total phenotypic variance, over 1000 replications with increasing sample size*

	Sample size									
	~ 100		~ 400		~ 600		~ 800		~ 1000	
Selection method	QTL position	QTL variance	QTL position	QTL variance	QTL position	QTL variance	QTL position	QTL variance	QTL position	QTL variance
WFV	29·8 (0·7)	0·394 (0·005)	28·3 (0·5)	0·341 (0·004)	26·6 (0·4)	0·331 (0·004)	26·2 (0·4)	0·322 (0·003)	26·3 (0·4)	0·310 (0·003)
CD	36·6 (0·9)	0·673 (0·005)	27·4 (0·6)	0·436 (0·004)	26·2 (0·4)	0·395 (0·004)	26·1 (0·4)	0·349 (0·004)	26·3 (0·4)	0·310 (0·003)
D	31·5 (0·9)	0·380 (0·020)	29·6 (0·8)	0·247 (0·007)	28·1 (0·7)	0·243 (0·005)	29·3 (0·7)	0·232 (0·004)	31·0 (0·7)	0·224 (0·004)
RS	37·0 (1·0)	0·478 (0·008)	29·4 (0·6)	0·346 (0·004)	27·6 (0·6)	0·327 (0·004)	26·8 (0·4)	0·317 (0·003)	26·3 (0·4)	0·310 (0·003)

The standard errors are given in parentheses. The simulated position was 25 cM and the QTL variance was $0{\cdot}28\sigma_P^2$ (family size 8).

https://doi.org/10.1017/S0016672301005225 Published online by Cambridge University Press

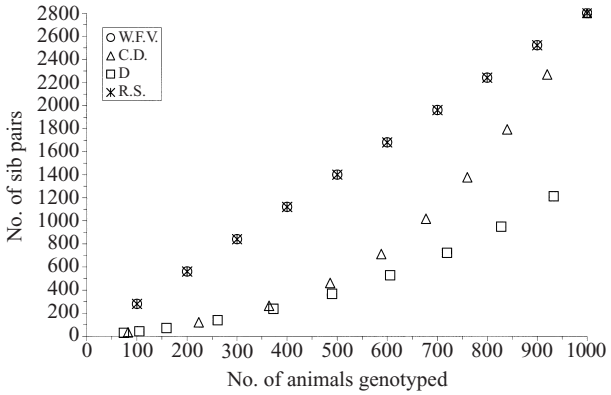

Fig. 6. The number of sib pairs generated in each selective genotyping scheme. Simulated parameters are as in Fig. 1. The numbers shown represent the mean of 1000 replicates. The population size simulated and hence the maximum number of animals genotyped was 1000 (800 progeny and 200 parents), resulting in a maximum of 2800 sib pairs (family size 8).

tensity). Moreover, when discordant sib pairs were selected (D) the maximum number of selected animals was always less than the overall maximum (i.e. 1000 animals) (Fig. 6). Therefore, even with the least intense selection, some sib pairs were not used in the analysis – something that resulted in more biased parameter estimates. With the smaller family size, where the overall power of QTL detection was lower, the position estimates were biased towards the centre of the chromosome and in many cases the mean estimated QTL position was outside the correct interval (Table 2).

The QTL variance estimates were biased upwards for all except the discordant selection scheme, where the bias was in the downwards direction for all but the most intense selection (Table 1). The upward bias in the estimated QTL variance was greatest with high selection intensities (and thus lower power for detection of the QTL). The bias in QTL position and variance estimates increased when smaller family sizes (4 sibs/family) were used (Table 2).

4. Discussion

The deviation of the proportion of alleles shared IBD at a marker locus from the expected mean (0·5) can provide the means of identifying a QTL linked to that marker in selected sib pairs. The analysis applied in this study was regression of the proportion of alleles IBD at a marker locus on the squared phenotypic differences between sibs. This method provides a test for the presence of a QTL via the test of significance of the regression constant. Others have successfully used a test based on chi-square for testing the deviation of the proportion of alleles IBD at a marker locus from the expected mean (Eaves & Meyer, 1994) and calculated size of sample requirements for a given power using a mean test (Risch & Zhang, 1995, 1996). However, these tests cannot provide means for parameter estimation. The method proposed here can be used to detect linkage and estimate the parameters (QTL position and variance) simultaneously, and furthermore it can be used in an interval mapping context (Fulker & Cardon, 1994), giving a small increase in power (on average 2%; data not shown).

Selective genotyping of sib pairs can reduce the number of animals that need to be genotyped (and hence the cost of genotyping) for a given power of detection of a QTL (Cardon & Fulker, 1994). Using a selected sample of discordant sib pairs, the deviation of the proportion of alleles shared IBD at a marker locus from the expected value of 0·5 allows detection of linkage between a marker and a QTL. However, the standard Haseman & Elston (1972) analysis of a sample of sib pairs selected from families with high within-family variance was a more powerful approach

Table 2. *Mean QTL position estimates in centimorgans and QTL variance estimates as a proportion of the total phenotypic variance, over 1000 replications with increasing sample size*

	Sample size									
	~ 100		~ 400		~ 600		~ 800		~ 1000	
Selection method	QTL position	QTL variance	QTL position	QTL variance	QTL position	QTL variance	QTL position	QTL variance	QTL position	QTL variance
WFV	33·7 (0·7)	0·433 (0·006)	30·1 (0·6)	0·363 (0·005)	29·1 (0·6)	0·354 (0·004)	29·3 (0·6)	0·342 (0·003)	29·9 (0·6)	0·319 (0·003)
CD	35·4 (0·9)	0·556 (0·008)	30·8 (0·7)	0·446 (0·005)	29·1 (0·6)	0·418 (0·005)	28·7 (0·6)	0·383 (0·004)	29·9 (0·6)	0·319 (0·003)
D	36·3 (1·1)	0·385 (0·019)	32·2 (0·9)	0·232 (0·008)	32·3 (0·9)	0·219 (0·006)	32·8 (0·9)	0·207 (0·006)	34·0 (0·8)	0·201 (0·006)
RS	40·7 (1·1)	0·678 (0·009)	34·3 (0·7)	0·395 (0·005)	32·5 (0·7)	0·359 (0·004)	31·3 (0·6)	0·338 (0·003)	29·9 (0·6)	0·319 (0·003)

The standard errors are given in parentheses. The simulated position was 25 cM and the QTL variance was $0{\cdot}28\sigma_P^2$ (family size 4).

https://doi.org/10.1017/S0016672301005225 Published online by Cambridge University Press

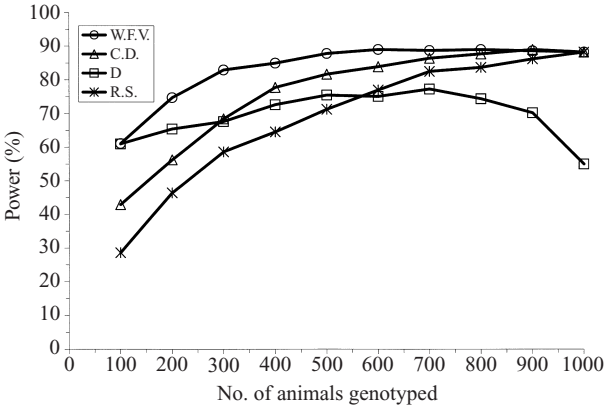

Fig. 7. Power of alternative selective genotyping schemes at different selection intensities. Data simulated were contained a *dominant* QTL of large effect ($\sigma^2_P = 1$, $h^2 = 0{\cdot}4$, $\sigma^2_{QTL} = 0{\cdot}28\sigma^2_P$) completely linked to a biallelic marker. Simulated population size is as in Fig. 1.

overall, except at high selection intensities. When the selection intensity was relatively low, the power of analyses of discordant sib pairs decreased dramatically. This decrease is associated with the declining deviation from expectation of the proportion of alleles IBD as less phenotypically divergent sib pairs are included in the sample. The power obtained using the standard test of the regression coefficient continued to increase with increasing sample size, when the samples were selected for high within-family variance (WFV), for concordant–discordant (CD) sib pairs or randomly (RS). With selection on concordant–discordant sib pairs or within-family variance the power reaches a maximum at a certain sample size and remains almost stable as the sample size further increases. This result shows that selection has been effective at selecting sib pairs that are most informative about linkage. Presumably, sib pairs added later in the analysis come from families in which the QTL was not segregating.

Selective genotyping on the basis of the WFV of the trait seems to be as effective even in cases of a rare QTL allele ($P = 0{\cdot}2$) (data not shown; Chatziplis, 1998) or dominant QTL (Fig. 7). In the case of a dominant QTL the power of detection was increased (9%) and the proposed selective genotyping method (WFV) slightly increases its advantage over the other selection methods (Fig. 7). In cases of a rare QTL allele, although the power and the expected proportion of informative matings are reduced, the WFV selection scheme still proved to be effective. Similar effects to those observed with a rare QTL allele should be expected in cases with a recessive QTL allele.

The results obtained in this study are in agreement with previous studies in human populations of small family size, which showed that sib pair methods using selected samples can achieve substantial power for QTL detection and decrease the number of individuals genotyped (Fulker & Cardon, 1994; Eaves & Meyer, 1994; Risch & Zhang, 1995, 1996; Guo *et al*., 1996). In human populations with small family sizes the use of concordant and discordant sib pairs seem to be the more powerful selection approach (Guo *et al*., 1996). Alternative analytical methods to the Haseman & Elston (1972) regression method for such samples have been suggested (EDAC test; Guo *et al*., 1996). However, which analytical method is the most powerful is an area of future research.

A comparison of the efficiency of selective genotyping schemes under different population structures and different models can only be suggestive of which analysis and selective genotyping methods are the most suitable in specific cases. It is the authors' belief that there is not an overall 'best' analytical or selective genotyping method. The analytical and selective genotyping method of choice depends on many parameters, both known (population structure, capital investment, etc.) and unknown (QTL mode of action and allelic frequency). The most objective way to determine the most powerful and cost-effective design for different scenarios would be the use of a simulation study under a variety of analytical and selective genotyping methods and tailored to the population of interest. These conclusions are underlined by the results of Allison *et al*. (1998), which show that in some circumstances where QTL of both large and small effect are segregating, power to detect the QTL of smaller effect can be reduced by some selective genotyping schemes.

Nevertheless, selection based upon phenotypic scores can be used to target genotyping effort in an efficient way. Typically, the best selection scheme in this study gives power for QTL detection approaching the maximum possible through genotyping only half of the individuals in the sample (Fig. 3). The major difference between this study and earlier ones is that it focuses on families larger than two sibs. With only two sibs in the family, selection on within-family variance is equivalent to selection of divergent sib pairs.

The best method overall in terms of power was selection based on the within-family variance, and this may in part be due to the number of sib pairs that are included in the analysis. With selection on within-family variance, the number of sib pairs increases linearly with the number of families and number of individuals genotyped. With selection of concordant–discordant sib pairs, the total number of sib pairs increases in an exponential manner (Fig. 6). Thus, with genotypes on half of the individuals, for example, selection on the within-family variance provides around 3 times the number of sib pairs in the analysis compared with the concordant–discordant selection scheme.

The results above and the relative power of studies based on 4 sibs per family versus 8 sibs per family

https://doi.org/10.1017/S0016672301005225 Published online by Cambridge University Press

emphasize the particular value of large families due to the number of sib pairs they contain. The expected *t*-value of the regression increases proportionally to the number of sib pairs whether they come from large families or small (Blackwelder & Elston, 1982). A family of 8 siblings requires 10 genotypes to generate information on 28 sib pairs. To obtain similar information from families of 2 siblings requires 112 genotypes (28 families and 4 genotypes per family). Such calculations suggest that genotyping families of size 8 is more than 11 times more efficient than genotyping families of size 2! Thus when selecting families for genotyping one should bear in mind the value of large families and balance the within-family variance against family size. Potentially one could develop an index that could be used to judge the value of genotyping a family and that incorporates both the size and phenotypic variability of a family. Defining the balance between variability and size will depend upon such factors as the distribution of QTL effects in the population, and is an area for future research.

Blackwelder & Elston (1982) and others (Amos & Elston, 1989; Götz & Ollivier, 1992) indicated that neither power nor type I errors are affected by treating all the comparisons in a sibship as independent. The above results are supported by Wan *et al*. (1997), who showed clearly that up to family size 5 there is no negative effect on power of detection. Our own simulations support these conclusions, with both the mean *t*-value in the presence of a linked QTL and the significance threshold being unaffected by the size of family from which the sib pairs were derived (Chatziplis, 1998). However, this same simulation study showed that the variance of the *t*-value over replicates was greater, both in the presence and absence (simulations under the null hypothesis) of a QTL, when the sib pairs were drawn from large families. Despite this reassuring conclusion, the use of a simulated threshold is recommended in order to provide a robust significance threshold.

Selective genotyping for a single trait is considered here. If more than one correlated trait is considered, some decrease in the selection intensity of the samples may secure sufficient power of detection for all traits. When traits are uncorrelated the applicability of any selective genotyping scheme will be somewhat reduced. This could be an interesting area for further study.

Appendix

The regression of the proportion of alleles shared IBD at marker locus between sib pairs on their squared phenotypic differences is known to be (Haseman & Elston, 1972):

$$\mathrm{E}(Y_j|\hat{\pi}_{j_m}) = [\sigma_e^2 + 2(1-2\theta+2\theta^2)\sigma_g^2] - 2(1-2\theta)^2\sigma_g^2\hat{\pi}_{j_m} \quad \text{(A1)}$$

with

$$\alpha = \sigma_e^2 + 2(1-2\theta+2\theta^2)\sigma_g^2$$

$$b_{Y_j\hat{\pi}_{jm}} = -2(1-2\theta)^2\sigma_g^2$$

$$r = -2(1-2\theta)^2\sigma_g^2\sqrt{\frac{\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}}}.$$

Since if $y = a + b_{yx}{}^*x$ then $r^2 = b_{xy}{}^*b_{yx}$ (Snedecor & Cochran, 1989):

$$b_{\hat{\pi}_{j_m}Y_j} = \frac{r^2}{b_{Y_j\hat{\pi}_{j_m}}} = \frac{2(1-2\theta)^2\sigma_g^2\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}} \quad \text{(A2)}$$

Given $\mathrm{E}(\pi|\bar{Y}) = \bar{\pi}$ then:

$$\alpha_{\hat{\pi}_{j_m}Y_j} = \mathrm{E}(\pi_m) + \frac{2(1-2\theta)^2\sigma_g^2\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}}\mathrm{E}(Y). \quad \text{(A3)}$$

From (A2) and (A3), the inverted Haseman & Elston (1972) regression would be:

$$\mathrm{E}(\hat{\pi}_{j_m}|Y_j) = \mathrm{E}(\pi_m) + \frac{2(1-2\theta)^2\sigma_g^2\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}}\mathrm{E}(Y) - \frac{2(1-2\theta)^2\sigma_g^2\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}}Y_j. \quad \text{(A4)}$$

Subtracting from both sides of the equation the expected mean proportion of alleles shared IBD at any locus (0·5), equation (A4) can be written as:

$$\hat{\pi}_{j_m} - 0{\cdot}5 = (\mathrm{E}(\pi_m) - 0{\cdot}5) - \frac{2(1-2\theta)^2\sigma_g^2\sigma^2_{\hat{\pi}_{j_m}}}{\sigma^2_{Y_j}}(Y_j - \mathrm{E}(Y)). \quad \text{(A5)}$$

We are grateful for support from State Scholarship Foundation of Greece (D.G.C.), the Biotechnology and Biological Sciences Research Council and the Ministry of Agriculture, Fisheries and Food in United Kingdom (C.S.H.) and the Marker Assisted Selection Consortium of the British Pig Industry (H.H.).

References

Allison, D. B., Heo, M., Schork, N. J., Wong, S. L. & Elston, R. C. (1998). Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. *Human Heredity* **48**, 97–107.

Amos, C. I. & Elston, R. C. (1989). Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genetic Epidemiology* **6**, 435–449.

Amos, C. I., Krushkal, J., Thiel, T. J., Young, A., Zhu, K., Boerwinkle, E. & De Andrade, M. (1997). Comparison of model-free linkage mapping strategies for the study of a complex trait. *Genetic Epidemiology* **14**, 743–748.

Blackwelder, C. W. & Elston, C. R. (1982). Power and robustness of sib pair linkage tests and extension to larger sibships. *Communications in Statistical and Theoretical Methods* **11**, 449–484.

Cardon, L. R. & Fulker, D. W. (1994). The power of interval mapping of quantitative trait loci, using selected sib pairs. *American Journal of Human Genetics* **55**, 825–833.

https://doi.org/10.1017/S0016672301005225 Published online by Cambridge University Press

Chatziplis, D. G. (1998). The use of selective genotyping in the detection of quantitative trait loci (QTL) by sib pair analysis. PhD thesis, University of Edinburgh, UK.

Chatziplis, D. G. & Haley, C. S. (2000). Selective genotyping for QTL detection using sib pair analysis in outbred populations with hierarchical structures. *Genetics*, *Selection*, *Evolution* **32**, 547–560.

Darvasi, A. & Soller, M. (1992). Selective genotyping for determination of linkage between marker locus and a quantitative trait locus. *Theoretical and Applied Genetics* **85**, 353–359.

Eaves, L. & Meyer, J. (1994). Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. *Behavior Genetics* **24**, 443–455.

Fulker, D. W. & Cardon, L. R. (1994). A sib pair approach to interval mapping of quantitative trait loci. *American Journal of Human Genetics* **54**, 1092–1103.

Götz, U. K. & Ollivier, L. (1992). Theoretical aspects of applying sib pair linkage test to livestock species. *Genetics*, *Selection*, *Evolution* **24**, 29–42.

Gu, C., Todorov, A. & Rao, D. C. (1996). Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. *Genetic Epidemiology* **13**, 513–533.

Hamann, H. & Götz, K. U. (1995). Use of sib pair linkage methods for the estimation of the genetic variance at a quantitative trait locus. *Genetics*, *Selection*, *Evolution* **27**, 97–110.

Haseman, J. K. & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2**, 3–19.

Liang, K.-Y., Huang, C.-Y. & Beaty, T. H. (2000). A unified sampling approach for multipoint analysis of qualitative and quantitative traits in sib pairs. *American Journal of Human Genetics* **66**, 1631–1641.

Mackinnon, M. J. & Georges, M. A. J. (1992). The effects of selection on linkage analysis for quantitative traits. *Genetics* **132**, 1177–1185.

Risch, N. & Zhang, H. (1995). Extreme discordant sib pairs for mapping QTL in humans. *Science* **268**, 1584–1589.

Risch, N. & Zhang, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *American Journal of Human Genetics* **58**, 836–843.

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical Methods*. Ames, Iowa: Iowa State University Press.

Wan, Y., Cohen, J. & Guerra, R. (1997). A permutation test for the robust sib pair linkage method. *Annals of Human Genetics* **61**, 79–87.