Regression-based quantitative trait loci mapping: robust, efficient and effective

Sara A. Knott*

*School of Biological Sciences, Institute of Evolutionary Biology, University of Edinburgh,
West Mains Road, Edinburgh EH9 3JT, UK*

Regression has always been an important tool for quantitative geneticists. The use of maximum likelihood (ML) has been advocated for the detection of quantitative trait loci (QTL) through linkage with molecular markers, and this approach can be very effective. However, linear regression models have also been proposed which perform similarly to ML, while retaining the many beneficial features of regression and, hence, can be more tractable and versatile than ML in some circumstances. Here, the use of linear regression to detect QTL in structured outbred populations is reviewed and its perceived shortfalls are revisited. It is argued that the approach is valuable now and will remain so in the future.

Keywords: quantitative trait loci mapping; regression; structured outbred populations

1. HISTORY

The idea of using markers associated with a trait of interest, for example, to predict the performance of individuals in the trait, is not new. Initially, however, the markers used were not identified at the molecular level but rather through the phenotype, for example, coat colour or by the use of simple biochemical procedures such as blood groups. An early implementation in plants is presented by Sax (1923) and by Neimann-Sorensen & Robertson (1961) in livestock. These original studies used a single marker at a time, because there were few markers and information about the location of the markers in the genome was insufficient. The phenotypes could be analysed either by directly fitting the marker genotype in a classic analysis of variance (ANOVA) or linkage could be modelled through a maximum likelihood (ML) approach. Rebaï *et al.* (1995) showed that these two approaches are asymptotically equivalent in terms of power. The marker effect estimated in the ANOVA approach can be reparametrized as a combination of the effect and recombination distance from the marker, but these two parameters cannot be estimated separately. On the other hand, the ML approach uses information about the distribution of the phenotypes within the marker genotype and, theoretically, can distinguish quantitative trait loci (QTL) with large effect some distance from the marker from one close by with a smaller effect (equivalent models for ANOVA). There is not much information coming from the phenotypic distribution within marker genotype, however, and most evidence for QTL is obtained from differences between the marker means and, hence, the ability of ML to correctly locate the QTL is limited.

Even if the recombination fraction can be correctly estimated, only the distance from the marker is known, not the side to which it lies. In addition, the ML approach makes an assumption about the distribution of phenotypes (usually that residual errors are normally distributed), which makes it more sensitive to the actual distribution and may lead to spurious QTL detection (Knott & Haley 1992a,b).

Although other forms of regression have been considered (e.g. Kearsey & Hyne 1994; Henshall & Goddard 1999), the work presented and discussed here will be based on interval mapping (Lander & Botstein 1989).

(a) Development

A great advance in QTL mapping was the introduction of interval mapping for crosses between inbred lines by Lander & Botstein (1989). This enabled the location of the QTL to be estimated separately from its effect by using information from the markers flanking a location of interest. Interval mapping became feasible because of the improvements in molecular techniques, which enabled large numbers of reliable molecular markers to be identified. With sufficient markers available, a genetic map can be determined, which locates markers relative to each other using genetic distances. When using Haldane's mapping function and thus assuming there is no interference in recombination events, flanking markers give all the available information about recombination within the interval between them. Interval mapping compares two hypotheses: the null is that there are not QTL in the interval (or linked to it) and the alternative is that there is a QTL within the interval. The model for the alternative hypothesis includes parameters to describe the effect of the QTL as well as one to define its location (only one is required since the flanking marker locations are fixed and a mapping function is assumed). Lander & Botstein (1989) proposed their approach in an ML framework, optimizing the location through a grid search. That is,

* Author for correspondence (s.knott@ed.ac.uk).

One contribution of 16 to a Theme Issue 'Population genetics, quantitative genetics and animal improvement: papers in honour of William (Bill) Hill'.

at steps throughout the interval, a model including a QTL at that location can be fitted and the location with the highest test statistic (comparing the hypotheses of a QTL at this location versus no QTL) is the most likely location of the QTL. An alternative model that includes location as a parameter to be optimized could also be considered.

As generally implemented, there are effectively two stages to the analysis: (i) calculating the probability of individuals of interest being each genotype (in terms of the alleles from the two original lines) at given locations throughout the genome conditional on the marker data and assuming a mapping function and that the marker linkage map is known without error and (ii) the use of these probabilities in an analysis looking for an association between the phenotypes and the genotypes. The interval mapping likelihood is a mixture distribution where individuals have a probability of being each of the possible genotypes. Lander & Botstein (1989) illustrate the approach with a backcross example, but it is applicable to any segregating generation from a cross between inbred lines (e.g. F_2 , doubled haploids, recombinant inbreds). Ignoring fixed effects and covariates, the likelihood for the F_2 can be written as follows

$$L = \prod_{i=1}^N \sum_{j=1}^3 \Pr(Q_j|A_iB_i)f(y_i|m_j, \sigma^2),$$

where,

$$f(y_i|m_j, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - m_j)^2}{2\sigma^2}\right],$$

Q_j is the genotype at the QTL; A_i is the genotype at marker A for individual i and B_i likewise for marker B ; y_i is the phenotype for individual i ; m_j is the effect of QTL genotype j ; σ^2 is the residual variance (assumed to be equal for all QTL genotypes) and N is the number of F_2 individuals.

The null hypothesis constrains the three means to be equal. The null and alternative hypothesis likelihoods can be compared using a likelihood ratio test, which, for a given location, is expected to follow a chi-square distribution with degrees of freedom equal to the number of parameters used to explain the QTL effect (i.e. 2 for the F_2 population, for the additive and dominance effect).

A simpler approach reduces the analysis to a linear model and regresses phenotype onto the probability of being each QTL genotype at given locations. Hence, the model for the F_2 example can be written as follows

$$y_i = \sum_{j=1}^3 m_j \Pr(Q_j|A_iB_i) + e_i,$$

with e_i being the residual for individual i and other variables as previously described.

If the QTL genotypes are known, $\Pr(Q_j|A_iB_i)$ is equal to 1 or 0 and the regression above is equivalent to an ANOVA fitting the QTL genotypes (and to ML fitting normally distributed errors).

An F -ratio of the variance explained by the QTL (after fitting all other effects in the model) to the

residual MS can be used to test for the significance of the QTL effects at this location. For a given location, this test statistic is approximately F distributed with degrees of freedom given by the number of parameters to describe the QTL (2 in the F_2) and the residual degrees of freedom ($N-3$, for a model with no additional effects).

This linear approach was given by Martinez & Curnow (1992) for a backcross and by Haley & Knott (1992) for the F_2 generation. In the latter case, linear combinations of the QTL genotype probabilities at each location were fitted in order to estimate the additive and dominance effects of the QTL directly. The asymptotic equivalence of this approach with ML interval mapping has been shown both through simulation (Haley & Knott 1992) and through theoretical predictions of power (Rebaï *et al.* 1995). Note that the regression F -statistic can be approximately converted to a likelihood ratio as $LR = -\text{d.f.} \ln[RSS/RSS_r]$, where RSS and RSS_r are the residual SS fitting and omitting the QTL, respectively, and d.f. are the residual degrees of freedom associated with RSS_r . ($RSS_r = RSS + SSQ$, with SSQ being the SS owing to the QTL).

Whittaker *et al.* (1996) showed the equivalence of the regression interval mapping approach to multiple regression on the genotype at flanking markers. With this latter approach, it is easier to see what information is available and therefore how many parameters can be estimated and it can reduce computational effort markedly where markers are fully genotyped with no missing information.

The nature of the interval mapping approach, where multiple correlated intervals are being tested and the highest test statistic selected, means that we are outside a standard testing regime and standard distributions cannot be used. This problem has been considered theoretically (e.g. Lander & Botstein; 1989; Rebaï *et al.* 1994) and empirically by Churchill & Doerge (1994). Churchill & Doerge (1994) suggested the use of a permutation analysis to set the appropriate significance thresholds. This involves repeated random permutation of the phenotypes with the genotypes and reanalysis selecting the highest test statistic. The distribution of highest test statistics obtained in this way reflect the distribution obtained using the observed marker data and phenotypic distribution but with no association between marker and QTL. It is therefore appropriate for obtaining the relevant significance thresholds.

Appropriate confidence intervals (CI) for, in particular, the location of the detected QTL, but also the other parameters have been investigated. The standard errors of parameters obtained from the analysis with the QTL at the best location will be underestimates, because they do not take into account the uncertainty of the location estimate. For the estimate of location, it is clear from a test statistic profile that the distribution is not symmetric around the best location, and the presence of the peaks suggests that the surface around the maximum test statistic will give misleading results regarding the CI. Under asymptotic conditions, a standard LOD drop (i.e. moving along the chromosome from the highest test statistic until the LOD

score, which is $\log_{10}$ of the ratio of the alternative over the null hypothesis likelihoods (and, therefore, $LR=2\ln[10]LOD$), has dropped by 1) should give approximately a 97% CI and 99% is given by a two LOD drop. From simulation, however, the CI determined in this way has been found to be too short, especially for small effect QTL (van Ooijen 1992). Mangin *et al.* (1994) presented a theoretical approach to constructing CI, while Visscher *et al.* (1996a) suggested the use of the bootstrap (i.e. sampling the population with replacement). The bootstrap is generally found, by simulation, to give conservative estimates of the CI.

In summary, the use of regression for the detection of QTL in inbred line crosses has a number of advantages including speed of calculation, easy inclusion of additional explanatory variables (fixed effects and covariates), ease of prediction of power and other theoretical considerations, easy increased complexity of the QTL model (see later), robustness (lacks sensitivity to distribution of phenotypes), ease of application of sampling methods to obtain CI (e.g. using a bootstrap) and significance thresholds (using permutation).

2. APPLICATION TO STRUCTURED OUTBRED POPULATIONS

There are a number of problems associated with analysing outbred rather than inbred based populations. A parent is only informative if both QTL and linked marker are heterozygous so that segregation in the progeny can be observed. Markers will not be fully informative, however, as they will be segregating in the parental generation(s). Any QTL will not be segregating in all families. The phase of linkage may vary across families; that is, the favourable allele at the QTL may be associated with different marker alleles in different families. This means that any association between marker and QTL is at the family level rather than population level (i.e. linkage disequilibrium between the QTL and linked marker is generated within families but may not exist across families). In addition, there will be additional variance between families that is caused by other loci affecting the trait of interest.

Within outbred populations, there are two distinct approaches. One is to look for QTL that explain differences between two populations, breeds or lines (analogous to the inbred situation described above). The other is to look for QTL explaining variation within a population.

Having developed and justified the approach for inbred line crosses, it is straightforward to extend it to structured outbred populations, such as crosses between outbred lines, where the QTL may be assumed to be fixed in the original lines, half-sib (HS) populations, HS and large full-sib (FS) populations. Regression has also been used for the analysis of sib-pairs (Haseman & Elston 1972); however, the nature of the model for these data where a variance component is estimated is very different and will not be considered further here.

The general, two-step strategy described above is still appropriate. The difference is that more markers

need to be considered simultaneously. Two markers are no longer sufficient to provide all available information (e.g. Knott & Haley 1992b).

(a) Three-generation crosses

Essentially, three-generation crosses are outbred versions of the inbred line crosses; that is, outbred lines that are extreme and divergent for phenotypes of interest are crossed to produce the equivalent of an F_1 generation. These outbred F_1 individuals are mated to produce an F_2 generation or backcrossed to one of the original outbred lines to give an outbred backcross. The aim is to detect genes that differ between the two outbred lines. The model for analysis is identical to the inbred version, once the inheritance of the marker alleles has been traced (Haley *et al.* 1994). The genetic model at the QTL assumes that the original lines are fixed for different alleles although genes can be segregating elsewhere. Hence, it is possible to combine information about the QTL across families. The assumption of fixation at the QTL can be tested by fitting a model that allows a different QTL effect in each family and the comparison of this with a model with the same QTL effect in each family (i.e. test for evidence of QTL by family interaction). If these models produce very different results (the interaction is significant), then it suggests that the original lines were not homogeneous when considering these QTL and the original model is not appropriate. If the QTL are not fixed in the original lines, the analysis can still be performed but there is a very quick loss of power as the allele frequencies become less extreme (Alfonso & Haley 1998). In addition, the interpretation of the estimated QTL effects changes; that is, the estimates for the additive and dominance effects are reduced by a factor of Δp and Δp^2 , respectively, where Δp is the difference in allele frequency between the lines (de Koning *et al.* 2002).

In the outbred F_2 cross, there is the potential of fitting an additional parameter to explain the effect of the QTL, because it is possible to discriminate between the two heterozygous genotypes depending on whether the increasing allele has been passed through the male or female parent. This is a line of origin effect and may be explained, for example, by imprinting. The first approach implemented was simply to estimate an effect for the difference in effect of the two heterozygous genotypes in addition to the additive and dominance effects at the QTL (Knott *et al.* 1998). A line of origin effect was detected if the test statistic for the best QTL with line of origin effect was a significant improvement over the test statistic with only Mendelian effect. In contrast to statements by Thomsen *et al.* (2004), male or female expression can be determined by considering the direction of the effect of the line of origin estimate compared with the additive estimate. Subsequently, de Koning *et al.* (2000) chose to estimate an effect of the difference in the two alleles inherited through the male parent and through the female parent. A significant difference in these two estimates is suggestive of a line of origin effect. de Koning *et al.* (2000) initially tested to see if the two estimates were significantly different from zero and reported

imprinting when the estimate from one sex of parent was significant whereas the other one was not. More recently, de Koning *et al.* (2002) investigated the behaviour of several testing regimes to determine whether there was evidence for line of origin effects and found that the proposed approaches inferred line of origin effects too frequently. Further investigation by Thomsen *et al.* (2004) proposed a more conservative test, by separating the detection of the imprinted QTL from the determination of the mode of expression and by imposing a more stringent significance threshold to account for the changes in location estimated by the different models.

There have been many examples of the use of regression for QTL mapping in outbred crosses, for example in pigs Knott *et al.* (1998) and in poultry Sewalem *et al.* (2002).

The same approach can also be used to analyse a three-generation pedigree consisting of a single large F_2 family (i.e. at maximum four grandparents and two F_1). With this structure, there is no longer the need to assume fixation of the QTL alleles in the grandparental generation, because up to four different alleles at any one location can be modelled. Rather than the additive and dominance model previously described, this is accomplished by fitting the genotypes explicitly or the effect from male and from the female parent can be estimated directly (i.e. the difference in the effect of the two alleles from each parent) and their interaction. This approach was used by Knott *et al.* (1997) for the analysis of an outbred pedigree of loblolly pine.

(b) *Half-sib design*

The HS design is related to the backcross, that is, we are interested in the segregation of alleles from one of the parents in their offspring. In the outbred context, however, we are hunting for genes causing variation within the population (from segregation within family). Essentially, it is a two-generation design, where offspring genotypes (and other relatives if available) are used to determine the two haplotypes in the common parent. Subsequently, these haplotypes are assumed to be known without error. Then, for each offspring, the probability of inheriting from one of these haplotypes is calculated at locations throughout the genome using flanking informative marker information. In the regression application (Knott *et al.* 1996), the evidence for QTL comes from mean differences between the offspring inheriting the allele from one haplotype and those inheriting from the other. In the outbred situation, we have to assume that the contribution of the mates of the common parent is equal over all common parents. The population cannot be assumed to be in complete linkage disequilibrium, and so we cannot assume that the same marker allele is always associated with the same QTL allele in all HS families and, hence, when it comes to the regression analysis we are interested in the QTL allele contrast nested within a HS family. The phenotypes are regressed onto the probabilities of inheriting from one haplotype of the common parent within families using the following model

$$y_{ij} = \mu + c_i + \alpha_i \Pr(Q_{ij}|H_i, M_{ij}) + e_{ij},$$

where y_{ij} is the phenotype of the j th individual from HS family i ; c_i is the effect of common parent i ; α_i is the substitution effect (allele from haplotype 1 – allele from haplotype 2) at given location for HS family i ; $\Pr(Q_{ij}|H_i, M_{ij})$ is the probability of the ij th individual inheriting from a given haplotype of the common parent given the two haplotypes of the common parent, H_i , and the marker genotypes of the individual ij , M_{ij} , and e_{ij} is the residual error.

The model differs from those previously described in that a substitution effect is estimated for each informative common parent in the dataset. For a single marker, a common parent is informative if heterozygous, while for multiple markers, the parent only requires to be heterozygous at one or more markers in the linkage group. Hence, in this model, the QTL effect is not assumed to be the same in each family. If QTL exist, however, they may be segregating in more than one HS family and accumulating evidence over HS families in the nested design may improve power to detect it.

A test for significance can be performed using a standard F -ratio (marker within HS family MS (with degrees of freedom equal to the number of informative common parents) compared with residual MS), but as with all interval mapping applications, the standard F distribution will not be appropriate because of the multiple testing and, hence, permutation (within families) or a theoretical approach can be used to obtain suitable distributions for the test statistic under the null hypothesis. There is a potential problem in that the residual variance is expected to be heterogeneous over families because of segregation of the QTL in the absence of informative markers. Regression is found to be robust to such violations of the model, which, in any case, is not expected to be large.

Weller *et al.* (1990) termed the above structure a ‘daughter design’ (based on an implementation in dairy cattle) and also suggested a procedure whereby the phenotypes are recorded in a third generation—the ‘granddaughter design’ (Weller *et al.* 1990). In this case, the expected QTL contrast is halved, but the error associated with the contrast is significantly reduced because a large number of grand-offspring can be phenotyped. This approach, of measuring phenotypes in the offspring of the individuals being genotyped, can also be implemented in other situations where the phenotyping is relatively cheap compared with the genotyping. An early example of the use of regression to analyse HS populations is Spelman *et al.* (1996).

(c) *Half- and full-sib families*

The HS approach described above was extended by Van Kaam *et al.* (1998) to enable analysis of outbred FS families. The same approach is used, except that haplotypes are generated for both parents and probabilities of offspring inheriting from a given haplotype from each of them calculated. A similar nested regression model is used with a substitution effect being estimated for both parents. Further testing of the two effects can determine whether only one of the parents may be segregating for the QTL. Within this structure, where substitution effects are estimated for

both parents, it is relatively straightforward to combine information from full and HS progeny.

3. PERCEIVED DISADVANTAGES AND PROBLEMS

(a) *Residual variance estimate*

There has been some confusion about the estimate of the residual variance from regression. When compared with a ML residual variance estimate or the simulated error variance, the value from regression may appear to be inflated. This is, of course, because the model is not accounting for segregation of the QTL within the marker genotype, and hence the variance will be inflated as a result. In the single-marker situation, it is easy to see that regression analysis is fitting a mean marker effect and the residual variance will include the component owing to segregation of the QTL within the marker genotype (i.e. assuming an additive effect QTL for an F_2 population with recombination fraction θ between the marker being fitted and the QTL, the inflation will be $2\theta(1-\theta)a^2$, where a is the additive effect of the QTL). If the QTL genotypes are known, then the variance would not be inflated. This point was made by Xu (1995) with the example of a backcross population. Of course, formulae can be given for different population structures.

For interval mapping, the QTL genotypes are more precisely known (i.e. probabilities are nearer to 1 and 0) and, hence, the contribution of the QTL effect to the residual variance estimate is less (as a result of unobservable double recombination events and uncertainty in the location of a known recombination). Although, as Xu (1995) states, we may be interested in the QTL effect in terms of the proportion of variance it explains, for small effect QTL the inflation will be negligible and, for larger ones, we could make some corrections. In any case, this is not a problem as long as we recall what we are estimating.

(b) *Selective genotyping*

When Lander & Botstein (1989) considered the use of selective genotyping (the practice where only individuals of extreme phenotype are selected for genotyping) to increase power for a given amount of genotyping, they stated that 'standard programs for linear regression cannot be used when only extreme progeny have been genotyped'. This is not the case as others have since demonstrated (Bovenhuis & Spelman 2000). Using regression and omitting ungenotyped individuals from the analysis will not bias the estimation of the QTL location or cause spurious QTL to be detected for the trait on which the selection was performed. Incorporating ungenotyped individuals with their expected genotypic values does not bias the location estimate, but inflates the test statistic compared with omitting these individuals from the analysis. Hence, appropriate thresholds are required to prevent false detection of QTL in this case. Whether ungenotyped individuals are incorporated or not, the estimate of the additive effect will be overestimated by approximately $(1 + xi)$, where x is the ordinate of the normal distribution, which has a proportion p , the selected proportion, higher and i is the expected mean of the

selected high individuals (Darvasi & Soller 1992). For correlated traits, it is possible to generate evidence for QTL when none are present. Bovenhuis & Spelman (2000) showed that the marker contrast between the offspring inheriting one allele or the other from a parent will be inflated on a given trait when selective genotyping is performed on a correlated trait (phenotypic/genetic level). The inflation of the estimated effect is approximately $raxi$ (assuming the phenotypic and genetic correlations between traits are the same, r , that the residual variance is equal to 1 for both traits and that the QTL is additive with a being the effect on the selected trait). This is a simplification of the formula given by Bovenhuis & Spelman (2000) for correlated traits in a HS population and illustrates that there will be no inflation in the estimate in the absence of selection (when $x=0$) or when the traits are uncorrelated (when $r=0$).

The results using ML are expected to be similar, except the QTL effect is not expected to be inflated as long as the information on which selection was performed is incorporated in the analysis. This means that when analysing a single trait that is either genetically (other than the QTL) or environmentally correlated to the trait on which selection was performed, at the location of QTL affecting the selected trait, the estimated effect of a putative QTL on the correlated trait under analysis will be inflated. This can be overcome, however, by analysing with a joint likelihood for selected and correlated traits simultaneously (e.g. as proposed by Johnson *et al.* 1999 for outbred populations or as implemented in QTLcartographer (Basten *et al.* 2002 for inbred lines).

(c) *Binary traits and other special phenotypes*

Within the regression framework, it is relatively straightforward to model binary traits using a generalized linear model with probit or logit links. Similarly, non-normal distributions can be accommodated, either by transformation or by the use of alternative error distributions. In practice, however, as long as the threshold or distribution is not too extreme, a normal approximation will not give misleading results or cause great loss of efficiency (Visscher *et al.* 1996b; Rebai 1997).

The estimation procedure does not require the assumption of normality of errors, only for hypothesis testing. Even for hypothesis testing, however, the central limit theorem assures that the regression coefficient will be asymptotically normally distributed enabling standard distributions to be used. For QTL mapping, the multiple testing regimes means that we are outside the standard distributions and the use of permutation avoids these assumptions altogether.

4. CURRENT STATUS

(a) *Multiple trait QTL models*

Many researchers have considered multiple trait QTL analyses. These have been approached in a number of ways, either to decompose the traits into a number of linear combinations that can be analysed separately (e.g. Weller *et al.* 1996; Mangin *et al.* 1998 consider principal component analysis, while Gilbert & Le Roy

2003 also consider discriminate analysis), or to analyse simultaneously (e.g. Jiang & Zeng 1995 and Ronin *et al.* 1995 propose an extension of the mixture model ML approach described above). As appealing as it may be to remain within the framework of the well established, single-trait analyses, transformation to a number of independent traits generates problems of interpretation. Transformation to independence is usually based on phenotypic correlations between the traits. Therefore, simple interpretations, for example that QTL affecting different linear combinations of the traits must be different QTL, do not necessarily hold because correlations between traits due to individual QTL will not be the same as the phenotypic ones. Jiang & Zeng (1995) showed that if the true model is a pleiotropic QTL, then analysing the multiple affected traits simultaneously by fitting a pleiotropic QTL increases power and improves the resolution to map the QTL.

The simple single-trait regression analysis can, within a standard structure, be extended to multiple traits (Wu *et al.* 1999, Knott & Haley 2000, Hackett *et al.* 2001). The simplest model of interest is whether a single location has an effect on more than one trait, i.e. the search for a pleiotropic QTL. The procedure is identical to the single-trait regression analysis, although having found the best location for the QTL, additional tests can be performed to determine the traits being affected. An additional interest is to break down the genetic correlation into the part caused by pleiotropic QTL and the part caused by linked QTL, which can be achieved through both ML and regression.

With all individuals genotyped (i.e. no selective genotyping), the regression approach is not very different from ML and has similar advantages as noted for the single-trait analyses. It is also easy to break down the test statistic into contributions from the individual traits. As noted previously, with selective genotyping, the regression approach may yield misleading results when analysing a correlated trait to the one on which selection was performed. This can be overcome if analysed jointly by ML including the data on which selection was operated.

Currently, however, multiple trait approaches, whether implemented through regression or ML, suffer in their implementation. Further development is required to determine the most efficient way to select the traits to work with. Additionally, in practice, results are observed that seem intuitively incorrect. For example, single-trait analyses give evidence for all traits in one region of a linkage group, but when the traits are analysed together, the best location moves some distance away where there was no evidence for QTL from the individual-trait analyses (S. Knott, unpublished work). The question is whether these results are caused by true QTL, chance correlations in the data (owing to relatively small sample sizes), an inadequate model or a problem with the test being used to select the best location in a linkage group. Further investigation into the behaviour of the multi-trait test statistics is required.

(b) *Multiple gene models*

One advantage of least-squares models is that it is relatively easy to include more explanatory variables,

and unlike ML, this does not dramatically increase calculation speed. As found by several researchers (e.g. Zeng 1993; Goffinet & Mangin 1998) for multiple linked QTL, the most reliable approach is to directly perform a grid search fitting the QTL simultaneously (as the best model may be a summary of several QTL and, therefore, once fitted in the model, evidence for the individual QTL may not be found). Additional QTL can simply be added into the regression model by including the genotype probabilities at different locations as additional variables. Hence, two-dimensional searches, for example, are easy to perform. The correct procedure for testing for multiple QTL, whether in ML or regression framework, is unclear. Should the two QTL model be compared with the best one-QTL model or either of the locations found for the two-QTL model (section-union test)?

Both Jansen (1993) and Zeng (1993) highlighted the benefits of including cofactors in the model to account for QTL segregating elsewhere in the genome. Even in the ML framework, these cofactors have been proposed as linear covariates, simply by including the genotype probabilities at the locations selected. The choice of method of selecting the locations to be included differs between researchers, however. Some have proposed a preliminary analysis, considering marker locations and using stepwise regression to determine a small subset of markers that each explain a significant amount of the variance (e.g. as implemented in QTL Cartographer; Basten *et al.* 2002); others simply fit QTL locations once detected (e.g. Brockmann *et al.* 2004).

With the fitting of multiple QTL comes the possibility of fitting interactions between the QTL. Carlborg & Andersson (2002) proposed the use of genetic algorithms for multiple, possibly epistatic, QTL searches with the use of randomization to determine significance thresholds for each test in the process. Subsequently, Ljungberg *et al.* (2004) used an efficient global optimization routine to increase computation speed and, hence, enable models including more QTL to be fitted. Although theoretically feasible in the ML framework, the problem is one of optimizing the likelihood with many parameters. This analysis becomes feasible with a quick approach such as regression as a large volume of parameter space can be searched.

From the results of Carlborg *et al.* (2000), performing multiple QTL searches including epistasis is a more efficient procedure to detect QTL, as some QTL are not detected by their direct effect but by their interactions with others. Hence, previous approaches of hunting for QTL and then searching for interactions between them will not find all QTL.

(c) *More complicated QTL models*

Another benefit of regression is the easy ability to complicate the QTL model, for example, considering interactions with sex, such that QTL may have a different effect in one sex compared to the other. This has enabled the combined analysis of data from several similar experiments in order to determine whether the same or different QTL were segregating in the different experimental populations (e.g. Walling *et al.* 2000).

(d) Implementation

Two Web-based software packages have implemented the regression approaches described here. For outbred populations, they can be found in QTL EXPRESS (<http://qtl.cap.ed.ac.uk>; Seaton *et al.* 2002) and for inbred line derived populations, in WEBQTL (<http://www.genenetwork.org>; Wang *et al.* 2003), which has primarily been set up as a resource for the analysis of mouse recombinant inbred lines.

5. FUTURE

In the development of QTL mapping, regression has obviously played an important role, and for many applications, should still be the approach of choice, because it offers greater speed and flexibility compared with ML. Molecular techniques have developed dramatically with the effect that experiments are now possible with many more individuals sampled, many more traits recorded (including expression data on thousands of transcripts) and very dense marker maps. With this in mind, is the simple linkage mapping using regression still going to be a useful tool in the future?

Increasing sample size has only beneficial effects for QTL mapping, irrespective of the approach, and can easily be accommodated with the regression framework. With more traits, there is a lot more, potentially useful, information to help determine networks or pathways of genes involved. At present, however, it is unclear how best to analyse them. Multiple trait regression can easily be applied to many traits, but it is not obvious that this is the most efficient approach, as the interpretation could be virtually impossible. Preliminary analyses or other prior information will be required to investigate and combine or select traits as required. Increasing the number of markers may produce problems for current algorithms to determine the inheritance of alleles but this is true for both ML and regression applications.

Currently, the regression methods discussed above for structured populations are a useful first step, providing information about important QTL, if not a final analysis in the process to identify causal loci. With a high level of single nucleotide polymorphisms, association studies in general outbreeding populations may be feasible, removing the necessity to set up experimental populations. These analyses may well use regression but not in the context of linkage. Large population-based studies, however, may not pick up the variants we are interested in (e.g. if these are rare), in which case, family-based linkage studies may be more appropriate. In addition, in the near future, such large-scale genotyping will still be too expensive for some species and applications, and the structured linkage approach will provide a more viable investigation.

Alternative approaches that can deal with a wide range of population structures have been proposed. Bayesian approaches that negate the requirement for testing (and therefore determining significance thresholds) and can compare models with, for example, different numbers of QTL are appealing. Although flexible, the major drawback of these approaches is the long computation time and the requirement for

specialized software. ML variance component estimation methods enable complex pedigrees to be analysed with phenotypes recorded over multiple generations. If the data are essentially structured, however, it is not clear that these offer an improvement over the regression analysis.

There are many advantages to the use of regression interval mapping to detect and characterize QTL in structured outbred populations, which mean that it will still be a useful tool in the future. It is relatively fast to implement, asymptotically equivalent to ML interval mapping, easy to generalize to experimental populations and crossing designs commonly used, easy to include cofactors, it has good robustness properties against non-normality and the estimation procedure is distribution free (Rebai 1997). The future will bring many traits and complex models; however, regression-based QTL mapping will continue to play a role, given its status as a simple framework in which to fit complicated models.

I would like to thank Peter Visscher and Chris Haley for useful discussions and The Royal Society for their financial support.

REFERENCES

- Alfonso, L. & Haley, C. S. 1998 Power of different F2 schemes for QTL detection in livestock. *Anim. Sci.* **66**, 1–8.
- Basten, C. J., Weir, B. S. & Zeng, Z.-B. 2002 *QTL Cartographer: a reference manual and tutorial for QTL mapping*, v. 1.16. Raleigh, NC: Department of Statistics, North Carolina State University. (See <http://statgen.ncsu.edu/qtlcart>.)
- 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.
- Brockmann, G. A., Karatayli, E., Haley, C. S., Renne, U., Rottmann, O. J. & Karle, S. 2004 QTLs for precomposition and postweaning body weight and body in selected mice. *Mamm. Genome* **15**, 593–609.
- Carlborg, Ö. & Andersson, L. 2002 Use of randomization testing to detect multiple epistatic QTL. *Genet. Res.* **79**, 175–184.
- Carlborg, Ö., Kerje, S., Schütz, K., Jacobsson, L., Jensen, P. & Andersson, L. 2000 A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res.* **13**, 413–421.
- Churchill, G. A. & Doerge, R. W. 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.
- 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.
- de Koning, D. J., Rattink, A. P., Harlizius, B., van Arendonk, J. A. M., Brascamp, E. W. & Groenen, M. A. M. 2000 Genome-wide scan for body composition in pigs revealed important role of imprinting. *Proc. Natl Acad. Sci. USA* **97**, 7947–7950.
- de Koning, D. J., Bovenhuis, H. & van Arendonk, J. A. M. 2002 On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. *Genetics* **161**, 931–938.
- Gilbert, H. & Le Roy, P. 2003 Comparison of three multitrait methods for QTL detection. *Genet. Sel. Evol.* **35**, 281–304.

- Goffinet, B. & Mangin, B. 1998 Comparing methods to detect more than one QTL on a chromosome. *Theor. Appl. Genet.* **96**, 628–633.
- Hackett, C. A., Meyer, R. C. & Thomas, W. T. B. 2001 Multi-trait QTL mapping in barley using multivariate regression. *Genet. Res.* **77**, 95–106.
- Haley, C. S. & Knott, S. A. 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.
- Haley, C. S., Knott, S. A. & Elsen, J. M. 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195–1207.
- Haseman, J. K. & Elston, R. C. 1972 The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3–19.
- Henshall, J. M. & Goddard, M. E. 1999 Multiple trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics* **151**, 885–894.
- Jansen, R. C. 1993 Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.
- Jiang, C. & Zeng, Z. B. 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111–1127.
- Johnson, D. L., Jansen, R. C. & van Arendonk, J. A. M. 1999 Mapping a quantitative trait locus in a selectively genotyped population using a mixture model approach. *Genet. Res.* **73**, 75–83.
- Kearsey, M. J. & Hyne, V. 1994 QTL analysis—a simple marker-regression approach. *Theor. Appl. Genet.* **89**, 698–702.
- Knott, S. A. & Haley, C. S. 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genet. Res.* **60**, 139–151.
- Knott, S. A. & Haley, C. S. 1992 Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics* **132**, 1211–1222.
- Knott, S. A. & Haley, C. S. 2000 Multitrait least squares for quantitative trait loci detection. *Genetics* **156**, 899–911.
- Knott, S. A., Elsen, J. M. & Haley, C. S. 1996 Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theor. Appl. Genet.* **93**, 71–80.
- Knott, S. A., Neale, D. B., Sewell, M. M. & Haley, C. S. 1997 Multiple marker mapping of quantitative trait loci in an outbred pedigree of loblolly pine. *Theor. Appl. Genet.* **94**, 810–820.
- Knott, S. A. *et al.* 1998 Multiple marker mapping of quantitative trait loci in an cross between outbred wild boar and large white pigs. *Genetics* **149**, 1069–1080.
- Lander, E. S. & Botstein, D. 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.
- Ljungberg, K., Holmgren, S. & Carlborg, Ö. 2004 Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. *Bioinformatics* **20**, 1887–1895.
- Mangin, B., Goffinet, B. & Rebaï, A. 1994 Constructing confidence intervals for QTL location. *Genetics* **138**, 1301–1308.
- Mangin, B., Thoquet, P. & Grimsley, N. 1998 Pleiotropic QTL analysis. *Biometrics* **54**, 88–99.
- Martinez, O. & Curnow, R. N. 1992 Estimating the locations and sizes of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet.* **85**, 480–488.
- Neimann-Sorensen, A. & Robertson, A. 1961 The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agric. Scand.* **11**, 163–196.
- Rebaï, A. 1997 Comparison of methods of regression interval mapping in QTL analysis with non-normal traits. *Genet. Res.* **69**, 69–74.
- Rebaï, A., Goffinet, B. & Mangin, B. 1994 Approximate thresholds of interval mapping tests for QTL detection. *Genetics* **138**, 235–240.
- Rebaï, A., Goffinet, B. & Mangin, B. 1995 Comparing powers of different methods for QTL detection. *Biometrics* **51**, 87–99.
- Ronin, Y. I., Kirzhner, V. M. & Korol, A. B. 1995 Linkage between loci of quantitative traits and marker loci: multi-trait analysis with a single marker. *Theor. Appl. Genet.* **90**, 776–786.
- Sax, K. 1923 The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* **8**, 552–560.
- Seaton, G., Haley, C. S., Knott, S. A., Kearsey, M. & Visscher, P. M. 2002 QTL Express: mapping quantitative trait loci in of simple and complex pedigrees. *Bioinformatics* **18**, 339–340.
- Sewalem, A., Morrice, D. M., Law, A., Windsor, D., Haley, C. S., Ikeobi, C. O. N., Burt, D. W. & Hocking, P. M. 2002 Mapping of quantitative trait loci for body weight at three, six and nine weeks of age in a broiler layer cross. *Poult. Sci.* **81**, 1775–1781.
- Spelman, R., Coppieters, W., Karim, L., van Arendonk, J. A. M. & Bovenhuis, H. 1996 Quantitative trait loci analysis for five milk production traits on chromosome 6 in the Dutch Holstein–Friesian population. *Genetics* **144**, 1799–1808.
- Thomsen, H., Lee, H. K., Rothschild, M. F., Malek, M. & Dekkers, J. C. M. 2004 Characterization of quantitative trait loci for growth and meat quality in a cross between commercial breeds of swine. *J. Anim. Sci.* **82**, 2213–2228.
- van Kaam, J. B. C. H. M., van Arendonk, J. A. M., Groenen, M. A. M., Bovenhuis, H., Vereijken, A. L. J. *et al.* 1998 Whole genome scan for quantitative trait loci affecting body weight in chickens using a three generation design. *Livest. Prod. Sci.* **54**, 133–150.
- van Ooijen, J. W. 1992 Accuracy of mapping quantitative trait loci in autogamous species. *Theor. Appl. Genet.* **84**, 803–811.
- Visscher, P. M., Thompson, R. & Haley, C. S. 1996 Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013–1020.
- Visscher, P. M., Haley, C. S. & Knott, S. A. 1996 Mapping QTLs for binary traits in backcross and F2 populations. *Genet. Res.* **68**, 55–63.
- Walling, G. A. *et al.* 2000 Combined analyses of data from quantitative trait loci mapping studies: chromosome 4 effects on porcine growth and fatness. *Genetics* **155**, 1369–1378.
- Wang, J., Williams, R. W. & Manly, K. F. 2003 WebQTL web-based complex trait analysis. *Neuroinformatics* **1**, 299–308.
- Weller, J. I., Kashi, Y. & Soller, M. 1990 Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *J. Dairy Sci.* **73**, 2525–2537.
- Weller, J. I., Wiggans, G. R., VanRaden, P. M. & Ron, M. 1996 Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. *Theor. Appl. Genet.* **92**, 998–1002.
- Wu, W.-R., Li, W.-M., Tang, D.-Z., Lu, H.-R. & Worland, A. J. 1999 Time-related mapping of quantitative trait loci underlying tiller number in rice. *Genetics* **151**, 297–303.
- Whittaker, J. C., Thompson, R. & Visscher, P. M. 1996 On the mapping of QTL by regression of phenotype on marker-type. *Heredity* **77**, 23–32.
- Xu, S. 1995 A comment on the simple regression method for interval mapping. *Genetics* **141**, 1657–1659.