Genet. Res., Camb. (1997), **69**, *pp.* 69–74

Comparison of methods for regression interval mapping in QTL analysis with non-normal traits

AHMED REBAÏ
INRA, Unité de Biométrie et d'Intelligence Artificielle, Auzeville B.P. 27, 31326 Castanet-Tolosan, France

(*Received 28 July 1996 and in revised form 11 September 1996*)

Summary

We compare the powers of three methods for the QTL analysis of non-normally distributed traits. We describe the nonparametric and the logistic regression approaches recently proposed in the literature and study the properties of the standard regression interval mapping method when the trait is not normally distributed. It is shown that the standard approach is robust against non-normality and behaves quite well for both continuous and discrete characters. The loss of power compared with the nonparametric or the logistic approach is generally minor. Moreover, the least squares estimation procedure of the regression interval mapping is not affected by departure from normality. The use of other approaches could be restricted to extreme cases where the trait distribution is very skewed.

1. Introduction

The problem of mapping quantitative trait loci (QTL) using high-density marker maps has been extensively studied with respect to both methodological and experimental aspects. Since the pioneering paper of Lander & Botstein (1989) many powerful statistical methods for detecting, locating and characterizing QTL have been proposed and their properties investigated (e.g. Haley & Knott, 1992; Haley *et al*., 1994; Jansen & Stam, 1994; Rebaï *et al*., 1994*a*; Zeng, 1994).

However, the problems of QTL interval mapping with non-normally distributed traits have only recently been addressed (Kruglyak & Lander, 1995; Hackett & Weller, 1995; henceforth abbreviated KL and HW, respectively). Two alternative methods were proposed: the nonparametric approach (KL) for continuous and categorial data and the logistic regression approach for ordinal data (HW). Unfortunately, the comparison between these approaches and standard interval mapping has not been discussed in depth. The aim of this paper is to identify the advantages and drawbacks of these two methods and to investigate the behaviour of the standard regression based interval mapping (abbreviated RIM) approach (Haley & Knott, 1992) when the data are not normally distributed.

2. The nonparametric approach

(i) *The method*

KL used a generalization of the Wilcoxon rank-sum test which they performed at every genome position. They established the asymptotic null distribution of the test process, thus providing an approximation to find the appropriate threshold. The major difficulty in generalizing the rank test for a genuine-wide QTL search is the occurrence of ties (individuals with the same phenotypic score). KL chose simply to rank tied individuals at random. Such an approximation has the merit of simplicity and needs no new theory because it does not affect the asymptotic distribution of the test and does not lead to a significant loss of efficiency compared with the method of average ranking of tied individuals (Kendall & Stuart, 1979). However, when the number of ties is high, which is the case for categorial traits, that approach is expected to be less powerful than the average-rank method for tie breaking (which is to attribute to each of the tied observations the average rank of those tied). Unfortunately this removes the feature of rank order tests causing the variance of sums of ranks to depend on the number and extent of the ties observed.

The nonparametric approach, although applicable to all distributions (distribution-free), would bring no or a minor advantage over the standard interval mapping regression approach when the data are normal or close to normality. The authors stated that

https://doi.org/10.1017/S0016672396002558 Published online by Cambridge University Press

the loss of efficiency, relative to the *t*-test, would be minor (4%) when the distribution is normal while the gain could be substantial for very skewed distributions (e.g. exponential). The major disadvantage is that the QTL effect could not be estimated with the non-parametric approach. KL recommended that both parametric and nonparametric approaches be used, especially when there is evidence for non-normality. If the results differ, then the experiment should be regarded with caution; otherwise the normal theory standard RIM could be used for the detection and estimation of QTL effects.

(ii) *Validity of the RIM approach*

The RIM approach could be applied to a wide range of non-normal distributions without loss of efficiency. It is known that least squares estimation in regression theory and its optimum properties do not involve the assumption of normality of errors. Thus all estimates remain valid, as do their variances, in the face of non-normality. Least squares is distribution-free to this extent (Kendall & Stuart, 1979). The normality assumption is required for hypothesis testing and particularly to establish the sampling distribution of test statistics. To find the F distribution for the statistic used for testing the presence of a QTL at each genome position, we need the independence of its numerator and denominator, which holds exactly only for normal populations. If we are sampling from a non-normal population, the central limit theorem nevertheless assures us that the regression coefficient will be asymptotically normally distributed. In fact, normality of the regression coefficients rather than that of the observations is required.

For linear regression models, $Y = X\beta + \epsilon$, Sen & Srivatsava (1990, ch. 5) showed that, if the error vector ϵ satisfies the Gauss–Markov conditions ($E(\epsilon) = 0$ and $E(\epsilon\epsilon') = \sigma^2 I$) then the test statistic for testing the hypothesis $C\beta = 0$, where C is a full rank matrix, converges asymptotically (as $n \to \infty$) to the standard F distribution if

$$\max(h_{ii}) \to 0,$$

where h_{ii} are the diagonal elements of the matrix $X(X'X)^{-1}X'$. As a rough rule of thumb, $\max(h_{ii}) < 0{\cdot}2$ may be taken as small enough for most applications if the original distribution of the ϵ values is not excessively long-tailed or J-shaped. This provides a simple criterion for studying the validity of the normal theory approximation when applying the standard RIM approach to the QTL analysis of non-normal data. Thus since the β_j values are linear combinations of y_i values, the central limit theorem guarantees, under certain conditions, that β_j values and the test statistic have approximately the expected distributions even if the y_i values are not normally distributed. The test significance threshold could then be obtained by available methods, either by approximations (e.g. Rebaï *et al*., 1994*b*) or by permutation (Churchill & Doerge, 1994), which have the advantage of taking into account the actual distribution of the trait.

(iii) *Simulation results*

We investigated the properties of the RIM approach by simulation. In the case of a backcross population and for a chromosome of length 1 M and six equally spaced markers, we simulated three samples of 100, 200 and 500 individuals from an exponential distribution with mean 1 and a QTL with an allelic substitution effect of $a = 0{\cdot}35$ (explaining almost 11% of the phenotypic variance) located at the middle of the chromosome. The exponential distribution has considerable importance and has had widespread use in the analysis of data in which the response variable is a lifetime. It is asymmetric and L-shaped with a standardized skewness coefficient equal to 2. The values of residuals were generated using the RANEXP random function of the SAS software (SAS, 1990) and the QTL analysis by RIM was done using programs developed under SAS/IML (1985). We also considered two other distributions: a standardized gamma distribution with scale and location parameters 1 and 2, respectively, and a Poisson distribution with parameter 1. Both are left skewed and have a long right tail (see e.g. McCullagh & Nelder, 1983 pp. 129, 151) and they were simulated using the SAS functions RANGAM and RANPOI. We computed the average (over all replications) QTL position, QTL effect and its standard deviation. The linear model test, distributed as a chi-squared with one degree of freedom (Rebaï *et al*., 1994*b*), was performed every 1 cM and the QTL position determined as the position with maximum test value. The empirical power was calculated as the ratio of the number of times the test exceeds the threshold to the total number of replications (1000 here). The threshold value was set to 6·9, which corresponds to a 5% per chromosome significance level (Rebaï *et al*., 1994*b*).

Table 1 shows that the estimates of the QTL effect for the exponential distributions are close to the exact values, especially when the population size increases, illustrating the optimal properties of least squares estimation. The standard deviation of the estimate is approximately inversely proportional to the population size. For all population sizes, all replications and all genome positions, the values of $\max(h_{ii})$ were found to be smaller than 0·04. These values are small enough and the sample size large enough that the distribution of the test statistic could be relatively well approximated by the usual theoretical distribution. We also computed the empirical power for a normally distributed trait with the same QTL effect and found that it is only slightly superior to that with an exponential distribution. The difference between

https://doi.org/10.1017/S0016672396002558 Published online by Cambridge University Press

Table 1. *Simulation results for an exponentially distributed trait using the RIM approach*[a]

Population size	QTL position[b]	QTL effect[c]	Empirical power (%)[d]	Range max (h_{ii})
100	49·8	0·326 (0·216)	56·5	0·02–0·037
200	49·7	0·347 (0·124)	85·3	0·01–0·017
500	50·0	0·351 (0·064)	99·8	0·004–0·006

[a] In simulation of backcross progeny (1000 runs) the QTL was assumed to lie in the middle of the chromosome (100 cM in length) and its additive effect was set to 0·35.
[b] In centimorgans from the leftmost marker of the chromosome.
[c] Estimate of the allele substitution effect and its standard deviation (in brackets).
[d] The ratio of the number of ties the test exceeds the threshold at the 5% level of significance (6·9) to the total number of replications.

powers is respectively 1·5%, 0·7% and 0·2% for 100, 200 and 500 individuals. Simulation results (not shown) for the gamma and Poisson distributions are similar to those obtained for the exponential; for the 100 individuals case, estimated powers for the two distributions were close to 57%. Position and effect of the QTL are well estimated; the QTL effects were 0·331 and 0·336 for the gamma and Poisson distributions, respectively. All these results indicate that the RIM approach could be used for QTL analysis of non-normal continuous traits without a major loss in power, at least as a first scan. If max (h_{ii}) values are found to be too large (say more than 0·2) then the nonparametric approach could be a good alternative. This approach was implemented in the software MapMaker/QTL (version 1.9), where nonparametric QTL mapping could be easily achieved using the 'np scan' command.

3. The logistic approach

(i) *The method*

The logistic approach described by HW is appropriate for discrete-valued traits measured on a finite number of categories. A typical example of such traits is disease resistance in crop species, scored on a nominal scale varying from unaffected to dead. For these ordinal traits, QTL detection could be carried out using the generalized linear model approach (McCullagh & Nelder, 1983) and particularly the logistic regression. In this model, it is assumed that the observed categories are derived from the restriction of an underlying continuous variable (generally normal) to fixed unknown thresholds (see Gianola, 1982). If the trait has k categories (denoted 1 to k) then there are $k-1$ thresholds $\alpha_{i_{(i=1\ldots k-1)}}$ and the probability that an individual j belongs to category i $(i=1\ldots k)$, knowing its genotype for the flanking markers, is (in the backcross case):

$$\Pr(Y_j = i \mid x_j) = F(\alpha_i + \beta x_j) - F(\alpha_{i-1} + \beta x_j), \quad \text{for} \quad 1 \leqslant i \leqslant k,$$

where $F(x) = 1/(1+e^{-x})$ is the *logit* link function, with $\alpha_0 = -\infty$ and $\alpha_k = +\infty$, β is the QTL allele substitution effect of the QTL on the ordinal scale and x_j is the coefficient of the QTL effect in the expression of the expected phenotypic value of individual j (which is equal to the difference between the conditional probabilities of being of genotype QQ and Qq at the QTL, conditional on the flanking marker genotypes). The total parameter vector is $\theta = (\alpha_1, \ldots, \alpha_{k-1}, \beta)'$. In F_2 populations, β would be a vector of two components corresponding to additive and dominance parameters of the QTL.

For simultaneously estimating the QTL effect and position, HW used an EM algorithm. Here we propose to fit the logistic model, to estimate and test θ at every position of the genome (every 1 cM). The QTL analysis is achieved using an iteratively weighted least squares algorithm (e.g. Green, 1984) for estimation and a Wald test (see Appendix for details). In the interpretation of such models the parameter of interest is neither β nor the QTL effect measured on the unobserved normal scale, neither of which have any direct biological meaning, but rather the probability of being in a given category conditional on the QTL genotype, that is $\Pr(Y = i \mid \text{QTL genotype})$ for each i (see HW).

(ii) *Validity of the RIM approach*

HW compared the values of category probabilities obtained by the logistic and the normal mixture models. They found that the advantage of the logistic approach when using flanking markers is slight, especially when the mode of the trait is in a central category and the number of categories is not too small (more than four or five). For a binary trait the category probabilities obtained with a normal mixture model seem to be biased. The QTL effect could be estimated using the logistic models on both ordinal (β) and underlying normal scales ($a = \sqrt{3}\beta/\pi$; see HW) but these values could not be directly compared with that obtained by the RIM model applied on the

https://doi.org/10.1017/S0016672396002558 Published online by Cambridge University Press

Table 2. *Simulation results for an ordinal trait with the RIM and the logistic approach*[a]

Approach	Population size	Position	QTL effect β	Empirical power (%)	Range max (h_{ii})
Trait with four categories					
RIM	100	49·5	0·288 (0·228)	45·1	0·020–0·034
Logistic	100	51·3	0·552 (0·427)	41·0	—
RIM	200	49·9	0·300 (0·125)	74·1	0·010–0·015
Logistic	200	49·6	0·573 (0·264)	72·0	—
Trait with two categories					
RIM	100	50·1	0·110 (0·102)	35·8	0·020–0·031
Logistic	100	49·9	0·556 (0·503)	30·3	—
RIM	200	49·3	0·113 (0·069)	58·8	0·010–0·016
Logistic	200	51·0	0·561 (0·340)	58·2	—

[a] Same notation and comments as in Table 1.

Table 3. *Simulation results for an ordinal trait with the RIM and the logistic approach for two values of the QTL effect and a population size of 100*

Approach	Simulated QTL effect	Position	QTL effect β[a]	Empirical power (%)
Trait with four categories				
RIM	0·2	49·6	0·156 (0·214)	19·1
Logistic	0·2	50·9	0·341 (0·416)	19·0
RIM	0·5	50·2	0·426 (0·188)	77·2
Logistic	0·5	49·5	0·856 (0·337)	78·0
Trait with two categories				
RIM	0·2	50·8	0·059 (0·106)	15·2
Logistic	0·2	51·1	0·227 (0·501)	13·1
RIM	0·5	50·6	0·166 (0·0832)	56·0
Logistic	0·5	49·2	0·687 (0·546)	52·9

[a] Same notation as in Table 1.

observed categorial data. By virtue of the same arguments developed earlier in this paper, the RIM model could also be applied and would give quite good results. In the following section we investigate this by comparing powers of the QTL detection tests used in both logistic and RIM approaches.

(iii) *Simulation results*

The simulations were carried out similarly to those previously described. We first generate a normally distributed trait using the RANNOR function of SAS and a QTL with effect $a = 0{\cdot}35$. Simulations with 100 individuals were also carried out with two other QTL effects: $a = 0{\cdot}2$ and $a = 0{\cdot}5$. Using this underlying distribution we derived two ordinal traits. The first trait has four categories and is derived using thresholds: 0, 0·5 and 1·5. This gives a left skewed distribution with approximately half of the expected observations in category 1, 19% in category 2, 24% in category 3 and 7% in category 4. The second trait is binary (two categories) derived by using a threshold of 0·5, which gives, in expectation, almost 70% of the individuals in category 1 and 30% in category 2. Values of QTL parameters and power were calculated as previously described. For the logistic approach, the test statistic (Appendix) also has a χ^2_1 distribution and the same significance threshold was used for both RIM and logistic tests.

The results for $a = 0{\cdot}35$ are given in Table 2. We see that the estimates of the QTL effect with the RIM method are close to the actual value of 0·35 for the first trait but not for the second. However, this is not our concern and we are not expecting them to be close to 0·35 because they are not measured on the same scale. We also notice that the QTL effect obtained by the logistic model and transformed onto the normal underlying scale by $a = \sqrt{3}\beta/\pi$ is about 0·32, which is close to the actual value of 0·35. This reflects the good convergence of the algorithm proposed. But the most interesting result is that the powers of the RIM and the logistic approaches are not significantly different. Moreover, the RIM seems to be more powerful when the population size is small. For both methods the power is increased when the population size increases, but more slowly for the RIM approach. For the

https://doi.org/10.1017/S0016672396002558 Published online by Cambridge University Press

binary trait the power is 10–15% less than that for the first trait (four categories), which itself is about 10% less than that obtained for the continuous trait (Table 1). This is intuitive because less variation is observable when the number of categories is smaller. Results in Table 3 confirm the conclusions already drawn and particularly that the powers of the RIM and logistic approaches are not significantly different.

4. Discussion

For traits with an exponential distribution, the use of nonparametric interval mapping test for QTL detection seems to give equal or less power than the standard interval mapping regression approach. This property would also hold for a wide family of continuous distributions, especially when the number of individuals is sufficiently large (more than 100, say) and the distribution is not very skewed. The nonparametric method of KL remains a good alternative to RIM and the combined use of both approaches for the analysis of experimental data is itself informative. It allows assessment of the validity of the RIM model and estimation of the QTL effect.

The logistic method is concerned with the QTL analysis of categorial traits. Its use needs more complicated models and iterative estimation procedures (EM or iteratively weighted least squares). Moreover, HW showed that the advantage of the logistic model over likelihood-based interval mapping is generally slight, and this advantage decreases as the number of categories increases, especially if the distribution of the trait is approximately symmetric. Our results showed that the RIM model gives at least the same power as the logistic approach and would be more powerful when the trait distribution is symmetric. Nevertheless, the logistic model could be a good alternative when one is interested in achieving an accurate estimation of category probabilities (e.g. for use in breeding experiments).

The advantages of the RIM approach are that it is asymptotically equivalent to likelihood-based interval mapping (Rebaï *et al*., 1995), easy to generalize to the experimental populations and crossing designs commonly used (e.g. Rebaï *et al*., 1994*a*) and could be quite simply adapted to include covariates to model the effects of multiple QTL (Jansen & Stam, 1994). In this paper was have shown, by simulation, that the RIM tests also have good robustness properties against non-normality, provided that the population size is sufficiently large (100 or more). This, combined with the fact that the least squares estimation procedure is distribution-free, makes the RIM an attractive approach for QTL mapping purposes.

Further work is, however, needed to investigate the effect of other problems related to non-normality of the trait on the performances of the RIM and alternative approaches. One of these is when the error terms are no longer independent and have unequal variances (Weller & Whyler, 1992; Korol *et al*., 1996). The properties of the RIM method and its optimality could be studied in more depth by further simulations and/or resampling methods (such as the bootstrap) to check the limits of the validity of the asymptotic theory.

Appendix

For each individual j we define the multinomial variable $Z_j = (Z_{1j}, \ldots, Z_{kj})$ such that $Z_{ij} = 1$ if $Y_j = i$ and 0 otherwise. Let $P_j = (p_{1j}, \ldots, p_{kj})$ where $p_{ij} = \Pr(Y_j = i)$. The expected value of Z_j is $E(Z_j) = P_j$. Let $V_j = \mathrm{cov}(Z_j)$ be the $k \times k$ variance–covariance matrix of Z_j and D_j the $k \times p$ matrix (p being the number of estimable parameters) of partial derivatives of P_j with respect to θ. The estimating equations for the regression parameters are:

$$\sum_j D_j' W_j (Z_j - P_j) = 0,$$

with $W_j = w_j V_j^-$, w_j is the weight of the j^{th} observation and V_j^- is a generalized inverse of V_j. We took for V_j^- the inverse of the diagonal matrix having vector P_j as a diagonal. The estimates are obtained iteratively as

$$\hat{\theta}_{m+1} = \hat{\theta}_m + \Big(\sum_j (\hat{D}_j' \hat{W}_j \hat{D}_j)\Big)^{-1} \sum_j \hat{D}_j' \hat{W}_j (Z_j - \hat{P}_j),$$

with the hat over D, W and P indicating that these matrices are evaluated at $\hat{\theta}_m$. The expression after the plus sign is the step size. To improve the convergence of the algorithm one can evaluate the log-likelihood (which is equal to $\sum_j w_j \log(\hat{p}_j)$) at $\hat{\theta}_{m+1}$. If its value is less than that evaluated at $\hat{\theta}_m$ then $\hat{\theta}_{m+1}$ is recomputed using half the step size. The estimated variance–covariance matrix of $\hat{\theta}$ could be obtained after convergence as $\big(\sum_j (\hat{D}_j' \hat{W}_j \hat{D}_j)\big)^{-1}$.

The algorithm described above converged in five to ten iterations in all the computations done in this study.

At each position we compute the test statistic $T = \hat{\beta}^2 / \mathrm{var}(\hat{\beta})$, which is asymptotically distributed as a χ_1^2. The threshold for a genome-wide scan could then be obtained by approximate methods (Rebaï *et al*., 1994*b*) or permutation tests (Churchill & Doerge, 1994).

References

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Gianola, D. (1982). Theory and analysis of threshold characters. *Journal of Animal Science* **54**, 1079–1096.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society*, *Series B* **46**, 149–192.

https://doi.org/10.1017/S0016672396002558 Published online by Cambridge University Press

Hackett, C. A. & Weller, J. I. (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**, 1252–1263.

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.

Jansen, R. C. & Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447–1455.

Kendall, M. & Stuart, A. (1979). *The Advanced Theory of Statistics*, 4th edn. London: Charles Griffin.

Korol, A. B., Ronin, Y. I., Tadmor, Y., Bar-Zur, A., Kirzhner, V. M. & Nevo, E. (1996). Estimating variance effect of QTL: an important prospect to increase the resolution power of interval mapping. *Genetical Research* **67**, 187–194.

Kruglyak, L. & Lander, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

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

McCullagh, P. & Nelder, J. A. (1983). *Generalized Linear Models*. London: Chapman and Hall.

Rebaï, A., Goffinet, B., Mangin, B. & Perret, D. (1994*a*). QTL detection with diallel schemes. In *Proceedings of the Ninth Eucarpia Meeting. Molecular Markers in Plant Breeding*, Wageningen, pp. 170–177.

Rebaï, A., Goffinet, B. & Mangin, B. (1994*b*). 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.

SAS Institute (1990). *SAS Language Reference*, version 6. Cary, NC: SAS Institute.

SAS Institute (1985). *SAS/IML User's Guide*, version 5. Cary, NC: SAS Institute.

Sen, A. & Srivastava, M. (1990). *Regression Analysis: Theory, Methods and Applications*. Berlin: Springer-Verlag.

Weller, J. I. & Whyler, A. (1992). Power of different sampling strategies to detect quantitative trait loci variance effects. *Theoretical and Applied Genetics* **83**, 582–588.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.