Heredity **83** (1999) 347–353 *Received 5 February 1999, accepted 27 May 1999*

Robust QTL effect estimation using the Minimum Distance method

M. PÉREZ-ENCISO*† & M. A. TORO‡

†*Area de Producció Animal, Centre UdL-IRTA, 25198 Lleida, Spain and* ‡*Area de Biotecnología y Mejora Genética Animal, CIT-INIA, Carretera Coruña km 7, 28040 Madrid, Spain*

Robustness has received little attention in QTL studies. We compare Maximum Likelihood (ML) and the Minimum Distance (MD) methods when there exists data contamination caused by outliers. A backcross population of size (N) 200 and 500 and 0, 5 or 25 outliers was simulated. The mean and standard deviation of the first QTL genotype were set to 1. Four cases were considered: (i) $\mu_2 = 1$, $\sigma_2 = 1$; (ii) $\mu_2 = 1$, $\sigma_2 = 1.25$; (iii) $\mu_2 = 1.252$, $\sigma_2 = 1$; (iv) $\mu_2 = 1.282$, $\sigma_2 = 1.25$, where μ_2 and σ_2 are the mean and standard deviation of the second genotype. Either full or selective genotyping was considered. A Monte Carlo MD method is proposed to deal with missing genotypes. MD estimates were much more robust than ML estimates, especially with respect to scale parameter estimates, and with selective genotyping.

Keywords: maximum likelihood, minimum distance, outliers, quantitative trait loci.

Introduction

Mapping genes or genomic regions responsible for the variation in quantitative traits (QTLs) is now a feasible task for most economically important species because of the large number of DNA polymorphisms available scattered along the genome. The development of statistical methods to detect QTLs in experimental crosses and in outbred populations has run parallel to advances in molecular methodologies. Thus, a number of approaches like maximum likelihood (ML, Lander & Botstein, 1989), regression (Haley & Knott, 1992; Knott *et al*., 1996), the method of moments (Darvasi & Weller, 1992; Luo & Woolliams, 1993) or, more recently, nonparametric methods (Kruglyak & Lander, 1995; Coppieters *et al*., 1998) have been applied to estimate QTL effect and position. Estimation procedures have been usually compared by means of simulation where the trait was distributed as assumed under the analysis, typically a mixture of normal distributions with equal variances within QTL genotypes. ML and regression have been shown to perform similarly under a variety of situations (Haley & Knott, 1992; Knott *et al*., 1996), whereas the method of moments was more unreliable than ML in the heteroscedastic model, i.e. when the variances for each QTL genotype differ (Luo & Woolliams, 1993). Most of these methods have been evaluated solely in terms of bias and accuracy, whereas robustness has received little attention. In particular, the effect of extreme phenotypic observations, i.e. outliers, on QTL effect estimates has not been studied in detail, to our knowledge. The scarcity of studies on robust QTL estimation occurs despite the ample statistical theory and methods available (e.g. Staudte & Sheather, 1990). Nonetheless, Jansen & Stam (1994) presented a strategy to detect outliers within an ML estimation framework. It should be stressed that outliers may be caused by relatively common phenomena in animal or plant management like preferential treatment of a subgroup of individuals, or a disease causing a less than average performance. This issue is particularly relevant because if a large-effect gene is segregating the trait will not be normally distributed. Thus, the researcher may be misled whenever the data show departures from normality that may be caused simply by outliers. This is especially the case with segregation analysis.

In this work we explore the performance of the 'Minimum Distance' (MD) estimation method in the context of QTL effect estimation. It should be remarked that other robust approaches like robust regression (e.g. Haley & Knott, 1992) or robust maximum likelihood (e.g. MacLachlan & Basford, 1987) are available, but the Minimum Distance approach has been shown to be very efficient when there is data contamination because of, for instance, outliers and is especially powerful when applied to mixtures (Parr & Schucany, 1988; Cao *et al*., 1995). It was first proposed by Wolfowitz (1957) although its use at that time was limited by

*Correspondence. E-mail: miguel.perez@irta.es

© 1999 The Genetical Society of Great Britain.

computational constraints. Nowadays it has become a more popular tool in the context of robust statistics theory and practice. We also present a generalization of the MD methodology to deal with missing data that occur with, for example, selective genotyping. In this latter instance regression provides biased estimates even without outliers.

Materials and methods

Theory

A minimum distance estimator of a parameter θ is a value that minimizes δ [F(.), F(.|θ)], where F(.|θ) is the distribution of interest, F(.) is an empirical distribution function obtained from the data, and δ [ ] is a distance measure (Titterington *et al*., 1985). The MD method thus comprises a variety of procedures, depending on the actual distance used. We used the Cramer–von Mises distance, as it has been shown to perform well in a variety of settings (Woodward *et al*., 1984; García-Dorado, 1997; García-Dorado & Marin, 1998):

$$\delta = \sum_{i=1}^{N} [\mathrm{F}(y_i|\theta) \quad (i \quad 0.5)/N]^2, \tag{1}$$

where N is the number of observations, and F($y_i|\theta$) is the value of the distribution function for the ith observation when the observations are ranked in ascending order. Now consider two inbred lines fixed for alternative allele markers (MM vs. mm) and QTLs (QQ vs. qq). In a backcross, assume that the trait of interest follows a normal distribution N(μ_1, σ_1) in homozygous individuals for the QTL (QQ), and that the distribution is N(μ_2, σ_2) in the heterozygous (Qq) individuals. The distribution function of the trait in individuals classified according to a linked marker is, assuming that haplotype QM is the nonrecombinant,

$$\mathrm{F}(y_i|\mathrm{M} = MM) = (1 \quad r)\Phi[(y_i \quad \mu_1)/\sigma_1] + r\Phi[(y_i \quad \mu_2)/\sigma_2] \tag{2a}$$

and

$$\mathrm{F}(y_i|\mathrm{M} = Mm) = r\Phi[(y_i \quad \mu_1)/\sigma_1] + (1 \quad r)\Phi[(y_i \quad \mu_2)/\sigma_2], \tag{2b}$$

where r is the recombination fraction between the marker and the QTL, and Φ[ ] is the standard normal distribution function. In order to apply eqn (1), individuals are classified and ranked within marker genotype, and the distance minimized is the sum of the distances within a marker class. If the QTL effect is estimated using flanking markers, then an expression similar to eqn (2a,b) is derived, taking into account that there are four marker classes in a backcross.

A common strategy for limiting molecular work is selective genotyping (Lander & Botstein, 1989), which consists of genotyping only the extreme individuals. The effect of a small number of aberrant data may be magnified with this strategy, and its consequences have not been explored. Although the ML estimation theory is well developed with missing data, i.e. the EM algorithm, the MD method has not been applied. Here we develop a computer intensive strategy, the Monte Carlo MD (MC-MD), which is similar to the MC-EM algorithm proposed by Wei & Tanner (1990). It consists of a double iteration loop. The MC-MD steps are as follows.

1 Initialize $\hat{\theta} = \{\mu_1^0, \mu_2^0, \sigma_1^0, \sigma_2^0\}$.
2 Do $j = 1, J$
  2.1 For each untyped individual with record y_i, compute the probability of having QTL genotype $G = QQ$ or Qq:

$$\mathrm{P}(G = QQ|y, \hat{\theta}) = \frac{\Phi[(y_i \quad \mu_1)/\sigma_1]}{\sum_{k=1,2} \Phi[(y_i \quad \mu_k)/\sigma_k]}$$

$$\mathrm{P}(G = Qq|y, \hat{\theta}) = 1 \quad \mathrm{P}(G = QQ|y, \hat{\theta}).$$

  2.2 Draw a random marker genotype given P($G = QQ|y, \hat{\theta}$), P($G = Qq|y, \hat{\theta}$) and distances between marker and QTL if an untyped individual.
  2.3 Rank individuals within the current marker class and estimate θ^j using a regular MD algorithm.
3 Update $\hat{\theta} = \sum_{j=1}^{J} \theta^j / J$.
4 Repeat from **2** until the distribution of $\hat{\theta}$ stabilizes.

J is the number of missing marker genotypes per individual drawn that are used to compute $\hat{\theta}$, and Φ[ ] is the standard normal density function. As in other Monte Carlo methods, $\hat{\theta}$ does not provide a point estimate but rather samples are obtained from the distribution of $\hat{\theta}$, taking into account uncertainty in missing genotypes, once convergence has been attained. For a general study on convergence properties of Monte Carlo methods, the reader is referred to Tanner (1993). A preliminary study here showed that convergence was reached in a few iterations because observations were independently distributed.

Computer simulation

The MD and ML methods were compared by means of simulation. Backcross populations of size $N = 200$ and 500 individuals were simulated. The mean and standard

deviation of the first genotype were set to $\mu_1 = \sigma_1 = 1$. Four cases were considered: (i) $\mu_2 = 1$, $\sigma_2 = 1$; (ii) $\mu_2 = 1$, $\sigma_2 = 1.25$; (iii) $\mu_2 = 1.252$, $\sigma_2 = 1$; and (iv) $\mu_2 = 1.282$, $\sigma_2 = 1.25$. The values for μ_2 in cases (iii) and (iv) were chosen so that the difference between QTL means was 1.25 phenotypic standard deviations. Two markers 25 cM apart were simulated, and the QTL was simulated at position 15 cM within the marker bracket. Five outliers were simulated for a population size of 200, and five or 25 outliers with $N = 500$. An outlier was generated by taking a random number from a distribution $N(2\mu, 2\sigma)$ instead of the 'correct' $N(\mu, \sigma)$. As argued in the introduction, this way of simulating outliers mimics directional bias that may be caused by disease or preferential treatment. The likelihood (or the distance) were maximized (or minimized) using the E04JAF subroutine of NAG software (Numerical Algorithm Group, 1995) with full genotyping. This subroutine uses a quasi-Newton algorithm that allows constraints to be fixed on the variables, i.e. $\sigma_i > 0$. The EM algorithm as described in Lander & Botstein (1989) and Luo & Kearsey (1992) was implemented for ML and MC-MD as above with selective genotyping. The proportion genotyped in this case was 40%, the extreme 20% in each tail.

Two hundred and 100 replicates were run for each case with full and selective genotyping, respectively. Bias and empirical standard deviation (SD) of the parameter estimates and power were calculated. Significance thresholds were obtained by permutation (Churchill & Doerge, 1994), 1000 permutations of the data being obtained for *each* replicate. Power was calculated as the number of replicates where the statistics exceeded the significance value ($P = 0.05$). The tests were whether $\mu_1 = \mu_2$, and whether $\sigma_1 = \sigma_2$.

Results and discussion

In order to compare both methods in the most favourable setting for ML, simulations were run without outliers (Table 1). ML and MD methods provided unbiased estimates for both location and scale parameters. Standard deviations of MD and ML estimates for means were very similar, resulting in equal power of both methods to detect differences between QTL means. Standard deviations of ML estimates of scale parameters were slightly smaller than with MD and, consequently, power was slightly larger with ML than MD methodology in heteroscedastic models. In summary, when the 'true' model equals the model used in the analysis, ML procedure shows, overall, equal or better properties than its MD counterpart although the difference is not large.

Nonetheless, if the data simulated do not correspond exactly to the model assumed in the analysis, i.e. in all real data analyses, then ML may not be the best choice. Tables 2 and 3 show the results for 'contaminated' populations. The percentage of outliers was either 1%

Table 1 Mean estimates of location (μ) and scale (σ) parameters with Minimum Distance (MD) and Maximum Likelihood (ML) methodologies. Values in parentheses are the empirical standard deviations over 200 replicates. The power in detecting differences between QTL means ($\prod_\mu$) and standard deviations ($\prod_\sigma$) is also shown. No outliers, $\mu_1 = \sigma_1 = 1$; (a) population size 200; (b) population size 500

μ_2	σ_2		$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\prod_\mu$	$\prod_\sigma$
(a)								
1	1	MD	0.999 (0.111)	1.000 (0.105)	1.012 (0.087)	1.011 (0.094)	0.03	0.06
		ML	1.000 (0.107)	1.003 (0.103)	0.997 (0.070)	1.006 (0.094)	0.05	0.05
1	1.25	MD	0.991 (0.100)	1.012 (0.142)	1.006 (0.094)	1.267 (0.107)	0.07	0.40
		ML	0.993 (0.106)	1.011 (0.138)	0.989 (0.075)	1.234 (0.087)	0.06	0.51
1.252	1	MD	0.998 (0.110)	1.254 (0.111)	1.007 (0.085)	1.015 (0.084)	0.33	0.03
		ML	0.997 (0.110)	1.257 (0.104)	0.985 (0.066)	0.998 (0.071)	0.37	0.03
1.282	1.25	MD	0.996 (0.113)	1.287 (0.131)	1.006 (0.087)	1.269 (0.116)	0.35	0.39
		ML	0.999 (0.109)	1.279 (0.124)	0.990 (0.067)	1.237 (0.089)	0.33	0.53
(b)								
1	1	MD	1.003 (0.076)	0.985 (0.063)	1.008 (0.058)	1.008 (0.058)	0.04	0.04
		ML	1.003 (0.075)	0.987 (0.061)	0.998 (0.047)	0.998 (0.045)	0.07	0.07
1	1.25	MD	1.001 (0.068)	0.991 (0.091)	1.001 (0.055)	1.258 (0.075)	0.06	0.76
		ML	1.002 (0.065)	0.993 (0.085)	0.992 (0.044)	1.248 (0.056)	0.04	0.95
1.252	1	MD	0.990 (0.063)	1.256 (0.065)	1.003 (0.057)	1.005 (0.059)	0.79	0.05
		ML	0.992 (0.059)	1.255 (0.063)	0.997 (0.043)	0.996 (0.047)	0.83	0.03
1.282	1.25	MD	0.994 (0.064)	1.284 (0.085)	1.007 (0.062)	1.251 (0.074)	0.72	0.73
		ML	0.996 (0.064)	1.278 (0.081)	0.996 (0.047)	1.242 (0.056)	0.73	0.91

Table 2 Mean estimates of location (μ) and scale (σ) parameters with Minimum Distance (MD) and Maximum Likelihood (ML) methodologies. Values in parentheses are the empirical standard deviations over 200 replicates. The power in detecting differences between QTL means ($\prod_\mu$) and standard deviations ($\prod_\sigma$) is also shown. Population size 200, 5 outliers, $\mu_1 = \sigma_1 = 1$

μ_2	σ_2		$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\prod_\mu$	$\prod_\sigma$
1	1	MD	1.032 (0.119)	1.038 (0.103)	1.038 (0.101)	1.051 (0.101)	0.05	0.07
		ML	1.068 (0.130)	1.082 (0.115)	1.107 (0.159)	1.124 (0.148)	0.05	0.03
1	1.25	MD	1.035 (0.113)	1.037 (0.134)	1.037 (0.093)	1.306 (0.116)	0.05	0.39
		ML	1.051 (0.119)	1.098 (0.138)	1.062 (0.133)	1.401 (0.145)	0.05	0.46
1.252	1	MD	1.028 (0.119)	1.300 (0.107)	1.035 (0.091)	1.056 (0.098)	0.37	0.05
		ML	1.062 (0.126)	1.355 (0.124)	1.094 (0.158)	1.182 (0.165)	0.36	0.04
1.282	1.25	MD	1.023 (0.114)	1.339 (0.138)	1.030 (0.088)	1.319 (0.118)	0.36	0.43
		ML	1.032 (0.122)	1.412 (0.148)	1.033 (0.126)	1.458 (0.165)	0.42	0.47

Table 3 Mean estimates of location (μ) and scale (σ) parameters with Minimum Distance (MD) and Maximum Likelihood (ML) methodologies. Values in parentheses are the empirical standard deviations over 200 replicates. The power in detecting differences between QTL means ($\prod_\mu$) and standard deviations ($\prod_\sigma$) is also shown. Population size 500, $\mu_1 = \sigma_1 = 1$; (a) 5 outliers; (b) 25 outliers

μ_2	σ_2		$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\prod_\mu$	$\prod_\sigma$
(a)								
1	1	MD	1.027 (0.071)	1.025 (0.069)	1.035 (0.061)	1.036 (0.065)	0.03	0.04
		ML	1.068 (0.082)	1.061 (0.079)	1.118 (0.112)	1.112 (0.102)	0.04	0.04
1	1.25	MD	1.029 (0.070)	1.025 (0.089)	1.032 (0.062)	1.296 (0.076)	0.06	0.74
		ML	1.042 (0.081)	1.085 (0.090)	1.057 (0.096)	1.399 (0.104)	0.06	0.76
1.252	1	MD	1.011 (0.065)	1.257 (0.068)	1.019 (0.061)	1.019 (0.061)	0.68	0.07
		ML	1.029 (0.065)	1.281 (0.067)	1.055 (0.080)	1.077 (0.084)	0.67	0.08
1.282	1.25	MD	1.003 (0.070)	1.297 (0.085)	1.017 (0.064)	1.275 (0.077)	0.73	0.71
		ML	1.003 (0.069)	1.331 (0.080)	1.010 (0.063)	1.350 (0.082)	0.79	0.82
(b)								
1	1	MD	1.058 (0.067)	1.068 (0.066)	1.073 (0.064)	1.069 (0.078)	0.06	0.02
		ML	1.130 (0.087)	1.155 (0.091)	1.213 (0.155)	1.258 (0.144)	0.02	0.03
1	1.25	MD	1.060 (0.067)	1.072 (0.091)	1.068 (0.064)	1.332 (0.079)	0.06	0.68
		ML	1.084 (0.090)	1.199 (0.105)	1.122 (0.135)	1.554 (0.135)	0.08	0.60
1.252	1	MD	1.053 (0.070)	1.333 (0.069)	1.072 (0.065)	1.077 (0.061)	0.75	0.04
		ML	1.109 (0.109)	1.461 (0.109)	1.176 (0.171)	1.363 (0.176)	0.55	0.04
1.282	1.25	MD	1.059 (0.069)	1.368 (0.084)	1.066 (0.063)	1.344 (0.083)	0.75	0.69
		ML	1.061 (0.081)	1.532 (0.101)	1.073 (0.110)	1.633 (0.125)	0.80	0.72

or 5% ($N = 500$) and 2.5% ($N = 200$). Contamination affected estimation by causing a bias and by increasing the empirical standard deviation of the estimates, thus augmenting the estimation error. However, the extent of these phenomena differed between methods. MD methodology was much more robust than its ML counterpart. Even with 1% of outliers and $N = 500$, ML estimates were quite sensitive. For instance, for $\mu_2 = 1$ and $\sigma_2 = 1.25$, the empirical SD of $\hat{\sigma}_2$ was almost doubled with ML compared to the case without outliers, whereas the SD remained constant with MD methods. Increasing the number of outliers resulted in a larger SD of ML, but only marginally in MD estimation. Similarly, the SD of estimates increased more rapidly for ML than MD methods as the percentage of outliers increased. Outliers affected both location and scale parameter estimates. However, scale parameter estimates were more affected than location estimates by the presence of outliers, especially with ML. Bias was about 50 and 100% larger for scale than location estimates.

Contamination affects power negatively by increasing the SD of estimates, but bias tends to augment spuriously the differences between genotypes, which favours power, especially in ML estimation. All in all, power was little affected by outliers with respect to location parameters with both ML and MD methods. In terms of comparing standard deviations, power decreased markedly

in contaminated populations with ML but only moderately with MD estimation. Power depended basically on population size, and it was barely affected by increasing the number of outliers.

Setting significance thresholds is not a straightforward issue in genome searches. Non-normal, or unknown, distributions make this issue even more complex. For instance, it is not clear how to set significance thresholds by analytical or simulation methods if the nature and the percentage of outliers is not known, as is the case in analysing real data. We have used data permutation because of its flexibility (Churchill & Doerge, 1994). Given the limited number of replicates run, power when $\mu_1 = \mu_2$, or $\sigma_1 = \sigma_2$ was very close to the *a priori* set significance level (5%), showing the adequacy of the permutation strategy. It can be seen that permutation behaved equally well whether a heteroscedastic model was simulated or not, and irrespective of the presence of outliers.

A further interesting aspect is convergence of maximization algorithms in heteroscedastic models. Luo & Woolliams (1993) reported that, under the heteroscedastic model, the log-likelihood might be unbounded and thus the ML estimates may not exist. We have not, apparently, encountered this problem: ML provided 'reasonable' estimates irrespective of whether σ_1 was equal or not to σ_2 whenever no outliers were simulated. Different algorithms, e.g. simplex or quasi-Newton, provided identical results.

Performance of EM-ML and MC-MD was compared with selective genotyping in the large population ($N = 500$). The 40% extreme individuals were genotyped and zero or 25 outliers were considered. After some exploratory analysis the total number of iterations was 30, and the number of times that missing marker genotypes were simulated per iteration in MC-MD (J) was also set to 30. The parameters reported are the mean of the last 10 iterations (Table 4). In general, it was found that the algorithm was rather stable, and quite independent of J. Selective genotyping did provide unbiased estimates of location parameters with only a marginal increase in error estimation compared to full genotyping (compare with Table 2b) in noncontaminated populations. Scale parameter estimates were, interestingly, slightly biased downwards in the heteroscedastic situations. Here differences in the SD of estimates were larger with MD than with ML, suggesting again that ML behaves better than MD when the model of analysis corresponds to that of simulation. And again the conclusion is reversed when there is data contamination.

Selective genotyping led to biased estimates in a contaminated population (Table 4b). It is quite clear that the stochastic MD method proposed is much more robust than ML using a standard EM algorithm. Even when $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, ML produced a larger bias than MD for both μ and σ estimates, and the bias increased if $\mu_1 \neq \mu_2$ and $\sigma_1 \neq \sigma_2$. Empirical SDs with ML were twice those of MD for scale parameters, and were

Table 4 Mean estimates of location (μ) and scale (σ) parameters with Minimum Distance (MD) and Maximum Likelihood (ML) methodologies. Values in parentheses are the empirical standard deviations over 100 replicates. Population size 500, $\mu_1 = \sigma_1 = 1$; (a) no outliers; (b) 25 outliers. Only the extreme 40% distribution is genotyped

μ_2	σ_2		$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
(a)						
1	1	MD	0.996 (0.075)	1.010 (0.070)	0.999 (0.053)	1.005 (0.049)
		ML	0.997 (0.077)	1.009 (0.066)	0.995 (0.036)	0.995 (0.036)
1	1.25	MD	1.009 (0.083)	0.997 (0.086)	1.034 (0.054)	1.203 (0.065)
		ML	1.007 (0.086)	1.001 (0.078)	1.065 (0.044)	1.172 (0.040)
1.252	1	MD	1.002 (0.071)	1.257 (0.078)	0.999 (0.056)	1.002 (0.050)
		ML	1.002 (0.067)	1.257 (0.076)	0.996 (0.038)	0.994 (0.040)
1.282	1.25	MD	0.994 (0.074)	1.274 (0.075)	1.042 (0.058)	1.198 (0.063)
		ML	0.984 (0.077)	1.267 (0.074)	1.071 (0.041)	1.167 (0.045)
(b)						
1	1	MD	1.067 (0.078)	1.078 (0.075)	1.062 (0.057)	1.067 (0.057)
		ML	1.137 (0.113)	1.153 (0.103)	1.224 (0.130)	1.230 (0.127)
1	1.25	MD	1.076 (0.077)	1.075 (0.086)	1.101 (0.056)	1.269 (0.074)
		ML	1.134 (0.101)	1.161 (0.088)	1.275 (0.124)	1.416 (0.106)
1.252	1	MD	1.075 (0.086)	1.331 (0.067)	1.058 (0.064)	1.078 (0.060)
		ML	1.105 (0.130)	1.456 (0.093)	1.151 (0.133)	1.369 (0.141)
1.282	1.25	MD	1.061 (0.071)	1.367 (0.101)	1.107 (0.062)	1.277 (0.069)
		ML	1.072 (0.107)	1.482 (0.101)	1.204 (0.099)	1.535 (0.108)

about 20% larger for location parameter estimates. Overall, selective genotyping caused almost no increase in SDs of the estimates for MD, but it did have a more noticeable effect on ML estimates. Power with selective genotyping could not be calculated using permutation because of the prohibitive amount of CPU required in MC-MD although it can be conjectured that MD and ML patterns should be similar to that with full genotyping (Table 3b).

In conclusion, the MC-MD method proposed alleviates to a large extent the bias caused by contamination in selectively genotyped populations. Higher computing costs of MC-MD than EM-ML are fully justified in this instance. However, these are only preliminary results on MC-MD and further studies are needed in order to evaluate its convergence and statistical properties in a more general framework. These results confirm the expectation that selective genotyping may be a risk-prone strategy with outliers, because almost certainly these will be included in the genotyped pool and its weight in the resulting estimates will be larger than if the whole population is genotyped. A further disadvantage of selective genotyping is an increased error in determining the QTL position (Pérez-Enciso, 1998).

General discussion and conclusion

In this work we have focused on phenotypic outliers that can be caused by extreme environmental factors like disease, and we have not studied the effect of incorrect genotyping or wrong pedigree information. If marker information is wrong but compatible with parent genotypes, a bias will occur, and the QTL effect will be underestimated. The effect of wrong marker information on QTL estimation is limited, however, and will cause a bias smaller than 2% in backcrosses unless the percentage of errors is large, e.g. greater than 10% (Pérez-Enciso, 1998).

The MD-estimate is the value of the parameter that makes the model closest to the sampling information, which seems a very reasonable strategy when the model assumed in the analysis does not represent the 'true' model, and it provides an intuitively appealing interpretation of the MD-estimates. Some minimum distance estimation methods have especially good properties in mixture distribution problems (Titterington *et al*., 1985). An additional advantage of MD methodology is its robustness. The literature shows that MD is normally more robust than ML when the real distribution does not pertain to the assumed parametric distribution, or when the actual distribution is 'contaminated' as, for example, when there exist outliers (Woodward *et al*., 1984; Parr & Schucany, 1988). This is because MD methods do not give so much weight to extreme data as ML does. García-Dorado (1997) illustrates how the MD method can lead to more sensible estimates of mutation effects than ML. In this work, we have shown that this methodology has clear advantages in some instances that may be encountered in QTL analysis.

Nonetheless, location and scale parameter estimates are not equally sensitive to contamination. Dispersion parameter estimates are much more sensitive to outliers than location parameters, where MD methods show comparatively a more robust behaviour. This is because scale parameters depend to a larger extent on the square differences, which are magnified by outliers. As a result, contamination may result in erroneously concluding that variances are heterogeneous if, for some reason, the proportion of outliers differs between QTL genotypes.

Choice of the distance in MD methodology is somewhat arbitrary, and it may severely affect the estimates. Nonetheless, some distances have a more clear interpretation. For instance, the Kullback–Leibler distance (Kullback & Leibler, 1951) is equivalent to the ML criterion. All distance measures tend to produce asymptotically normally distributed estimators. The distance measure has a most critical impact in small samples. Here we used the Cramer–von Mises distance as it is one of the most widely used (Parr & Schucany, 1988; Garcia-Dorado, 1997). We also tried other measures, like the Kolmogorov–Smirnov measure, but Cramer–von Mises gave identical or better results.

Typically, MD consists of comparing distribution functions, but other alternative MD methods based on density rather than distribution functions have been developed by Cao *et al*. (1995). In this strategy, the distance between a density function and a nonparametric density estimator is minimized. This strategy consumes more CPU time than standard approaches. It requires specifying an appropriate smoothing and evaluating the chosen kernel estimator. According to Cao *et al*. (1995), MD density-based methods are specially suited for testing whether the assumed density belongs to a given parametric family. In the context of QTL studies, this may be relevant to detecting how many QTL genotypes are segregating in a given population, or to detecting departures from normality within QTL genotypes. This issue merits further attention. The possible presence of influential points but which are not outliers is a further aspect of robustness which has not been dealt with here. Jansen & Stam (1994) considered the changes in the weighted sum of squared residuals as a means to check for the presence of outliers. A comparison of changes in parameter estimates vs. changes in the sum of squared residuals obtained when a given observation is deleted may provide a means of identifying influential observations that are not outliers.

Minimum distance methods do not come without disadvantages. It is not clear how to take into account a correlated structure in the data (e.g. genetic relationships, common environment), because MD strategies are based upon the assumption of independence and identical distributions for each observation. We have also found that, in some instances, MD statistics may be unstable along a genome search when changing the marker interval. This problem can be alleviated by using alternatives to the genome scan. We (M. Pérez-Enciso & L. Varona, unpubl. obs.) have studied a strategy where the whole genome is partitioned into segments and the effect of each segment is analysed simultaneously by using a multiple regression/ANOVA approach. Among the issues that should be explored in more detail are the behaviour of the MD approach with more than one QTL and with non-normal distributions, as well as the properties of the MC-MD approach.

Acknowledgements

We thank Luis Varona and Antonio Cuevas for their comments. This work is funded by CICYT grant AGF96–2510.

References

CAO, R., CUEVAS, A. AND FRAIMAN, R. 1995. Minimum distance density-based estimation. *Comput. Stat. Data Analysis*, **20**, 611–631.

CHURCHILL, G. A. AND DOERGE, R. W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

COPPIETERS, W., KVASZ, A., FARNIR, F., ARRANZ, J. J., GRISART, B., MACKINNON, M. AND GEORGES, M. 1998. A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: application to milk production in a granddaughter design. *Genetics*, **149**, 1547–1555.

DARVASI, A. AND WELLER, J. I. 1992. On the use of the moments method of estimation to obtain approximate likelihood estimates of linkage between a genetic marker and a quantitative locus. *Heredity*, **68**, 43–46.

GARCÍA-DORADO, A. 1997. The rate and effects distribution of viability mutation in *Drosophila*: minimum distance estimation. *Evolution*, **51**, 1130–1139.

GARCÍA-DORADO, A. AND MARIN, J. M. 1998. Minimum distance estimation of mutational parameters for quantitative traits. *Biometrics*, **54**, 1097–1114.

HALEY, C. S. AND KNOTT, S. A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315–324.

JANSEN, R. C. AND STAM, P. 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.

KNOTT, S. A., ELSEN, J. M. AND HALEY, C. S. 1996. Methods for multiple mapping of quantitative trait loci in half sib populations. *Theor. Appl. Genet.*, **93**, 71–80.

KRUGLYAK, L. AND LANDER, E. S. 1995. A nonparametric approach for mapping quantitative trait loci. *Genetics*, **139**, 1421–1428.

KULLBACK, S. AND LEIBLER, R. A. 1951. On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

LANDER, E. S. AND BOTSTEIN, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

LUO, Z. W. AND KEARSEY, M. J. 1992. Interval mapping of quantitative trait loci in an F_2 population. *Heredity*, **69**, 236–242.

LUO, Z. W. AND WOOLLIAMS, J. A. 1993. Estimation of genetic parameters using linkage between a marker gene and a locus underlying a quantitative character in F_2 populations. *Heredity*, **70**, 245–253.

MACLACHLAN, G. AND BASFORD, K. 1987. *Mixture Models*. Marcel Dekkers, New York.

NUMERICAL ALGORITHM GROUP. 1995. NAG Fortran Library Manual, Mark 17. Oxford.

PARR, W. C. AND SCHUCANY, W. R. 1988. Minimum distance and robust estimation. *J. Am. Stat. Ass.*, **75**, 616–624.

PÉREZ-ENCISO, M. 1998. Sequential bulked typing: a rapid approach for detecting QTLs. *Theor. Appl. Genet.*, **96**, 551–557.

STAUDTE, R. G. AND SHEATHER, S. J. 1990. *Robust Estimation and Testing*. John Wiley, New York.

TANNER, M. 1993. *Tools for Statistical Inference*. Springer, Berlin.

TITTERINGTON, D. M., SMITH, A. F. M. AND MAKOV, U. E. 1985. *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York.

WEI, G. C. AND TANNER, M. A. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Stat. Ass.*, **85**, 699–704.

WOLFOWITZ, J. 1957. The minimum distance method. *Ann. Math. Stat.*, **28**, 75–84.

WOODWARD, W. A., PARR, W. C., SCHUCANY, W. R. AND LINDSEY, H. 1984. A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *J. Am. Stat. Ass.*, **79**, 590–598.