More about quantitative trait locus mapping with diallel designs

AHMED REBAÏ* AND BRUNO GOFFINET
INRA Centre de Toulouse, Unit of Biometry and Artificial Intelligence, BP 27, 31326 Castanet-Tolosan, France

(*Received 25 March 1999 and in revised form 8 September 1999*)

Summary

We present a general regression-based method for mapping quantitative trait loci (QTL) by combining different populations derived from diallel designs. The model expresses, at any map position, the phenotypic value of each individual as a function of the specific-mean of the population to which the individual belongs, the additive and dominance effects of the alleles carried by the parents of that population and the probabilities of QTL genotypes conditional on those of neighbouring markers. Standard linear model procedures (ordinary or iteratively reweighted least-squares) are used for estimation and test of the parameters.

1. Introduction

Most methods for mapping quantitative trait loci (QTL) are designated to handle a single cross between two inbred or outbred parents. Combining information from multiple crosses has proved to be a more powerful approach (Rebaï & Goffinet, 1993; Muranty, 1996; Xu, 1996; Xie *et al.*, 1998; Xu, 1998); it increases the chance that polymorphic alleles are present in the parental gene pool and allows the estimation of QTL effects and positions over a larger set of lines. Using multiple families of crosses also increases the statistical inference space and may permit the detection of QTLs which are undetectable in a single-line cross, where the two parents could be fixed for the same allele at a particular QTL.

We have considered QTL mapping using multiple crossing diallel designs between a set of inbred lines (Rebaï & Goffinet 1993, 1996; Rebaï *et al.* 1994, 1997*a*, *b*), while Cockerham & Zeng (1996) considered the North Carolina design III, wherein the F2 from two inbred lines are backcrossed to both parental lines. Xu (1996) has developed methods for QTL mapping using four-way crosses. Muranty (1996) studied the power of different mating designs between outbreds for QTL detection using single-marker methods. Xu (1998) described and compared two strategies (fixed and random-model strategies) of combining data from multiple families of line crosses. Recently, Xie *et al.* (1998) have developed a new approach based on an identity by descent variance component method which allows QTL mapping by combining different line crosses.

Regression-mapping with multiple markers has proved to be a powerful and robust method for QTL mapping in classical populations derived from biparental crosses between inbred or outbred lines (see Haley & Knott, 1992; Haley *et al.*, 1994; Rebaï, 1997). Its simplicity of generalization and implementation and its computation efficiency relative to computer-intensive likelihood-based methods have made it an attractive approach to QTL mapping in complex designs.

In our previous work, we generalized and used this method to analyse multiple populations derived from complete half-diallel crosses among inbred lines. In this paper, we give a further generalization of the method for the analysis of incomplete diallels among inbred and outbred lines.

2. Model and methods

In this section we first consider a complete half-diallel cross between l parental inbred lines, where $p = l(1-1)/2$ F2 populations are derived from the F1

* Corresponding author: Centre of Biotechnology of Sfax, Laboratory of Plant Protection and Transformation, B.P. 'K', 3038 Sfax, Tunisia.

https://doi.org/10.1017/S0016672399004358 Published online by Cambridge University Press

hybrids. The generalization to other population types such as recombinant inbred lines, double haploids evaluated *per se* or by testcross is straightforward. We then describe in Section 3 the application of the method to incomplete diallel crosses and diallels of outbred lines.

(i) *The mapping model*

Consider that each parental line Li ($i = 1 \ldots l$) has a different Qi at the QTL, so that there are l homozygous genotypes and $l(l-1)$ 2 heterozygotes ($g = l(l+1)/2$ genotypes). For any individual k of the F2 population derived from the cross $Li \times Lj$ and having phenotypic value y_{ijk}, the model could be written, at any given map position, as:

$$\begin{aligned} y_{ijk} = \mu_{ij} &+ Pr(QiQi/GM)_k(2a_i) \\ &+ Pr(QjQj/GM)_k(2a_j) \\ &+ Pr(QiQj/GM)_k \times (a_i + a_j + d_{ij}) + e_{ijk}, \end{aligned} \quad (1)$$

where μ_g is a cross-specific mean, a_i is the additive effect of allele Qi and d_{ij} the dominance effect between alleles Qi and Qj (so that genotypes $QiQi$ and $QiQj$ have genotypic values of $2a_i$ and $a_i + a_j + d_{ij}$, respectively), and e_{ijk} are the residuals assumed to be i.i.d with variance σ^2. $Pr(QiQi/GM)_k$ is the probability that individual k has genotype $QiQi$ for the putative QTL conditional on its observed genotype for the markers GM and is a function of the distances between markers and the position of the QTL. The expressions of these probabilities are easy to obtain for two flanking markers (e.g. Rebaï *et al.*, 1994). If l parents are involved, only $(l-1)$ additive parameters are estimable and a constraint on the a_i values should be used (e.g. $\Sigma_i a_i = 0$) while all dominance parameters are estimable. Thus the total number of estimable parameter is $q = (l-1)(l+2)/2$.

In matrix notation, model (1) could be expressed as:

$$\begin{aligned} \mathbf{Y} = \mathbf{X\beta} + \mathbf{e} &= \mathbf{X_o\beta_o} + \mathbf{X_q\beta_o} + \mathbf{e} \\ &= \mathbf{X_o\beta_o} + \mathbf{X_a\beta_a} + \mathbf{X_d\beta_d} + \mathbf{e}, \end{aligned} \quad (2)$$

where $\mathbf{Y}$ is the $n \times 1$ vector of phenotypic values of n individuals from p different populations, $\boldsymbol{\beta}_\mathbf{o}$ is a $p \times 1$ vector of cross-specific means, $\mathbf{X_o}$ is a $n \times p$ matrix whose (ij)th element is 0 or 1 according to whether or not individual i ($i = 1 \ldots n$) belongs to population j ($j = 1 \ldots p$), $\boldsymbol{\beta}_\mathbf{q}$ is a $q \times 1$ vector of QTL effects, $\mathbf{X_q}$ is a $n \times q$ matrix whose elements are linear combinations of the probabilities of QTL genotypes conditional on those of neighbouring markers and e is a $n \times 1$ vector of residuals with known variance matrix $\mathrm{Var}(e) = \sigma^2 \mathbf{I}$ ($\mathbf{I}$ being the identity matrix). $\boldsymbol{\beta}'_\mathbf{q}$ ($'$ stands for the transpose) could be partitioned as $[\boldsymbol{\beta}_\mathbf{a} | \boldsymbol{\beta}_\mathbf{d}]'$, where $\boldsymbol{\beta}_\mathbf{a}$ and $\boldsymbol{\beta}_\mathbf{d}$ are vectors of estimable additive and dominance parameters. X and β could thus be partitioned as:

$$\mathbf{X} = [\mathbf{X_o} | \mathbf{X_q}] = [\mathbf{X_o} | \mathbf{X_a} | \mathbf{X_d}]$$

and

$$\boldsymbol{\beta}' = [\boldsymbol{\beta}_\mathbf{o} | \boldsymbol{\beta}_\mathbf{q}]' = [\boldsymbol{\beta}_\mathbf{o} | \boldsymbol{\beta}_\mathbf{a} | \boldsymbol{\beta}_\mathbf{d}]'.$$

Note that only $\mathbf{X_q}$ needs to be considered at each genome position ($\mathbf{X_o}$ is built once).

(ii) *Construction of* $\mathbf{X_q}$

$\mathbf{X_q}$ could be obtained by a product of two matrices $\mathbf{X_q} = \mathbf{P} \times \mathbf{C}$, where $\mathbf{P}$ is the $n \times g$ probability matrix and $\mathbf{C}$ is a $g \times q$ matrix of constants expressing the constraints on the parameters. g is the number of all possible genotypes at the QTL ($g = l(l+1)/2$). The element of line i and column j of $\mathbf{P}$ is the probability that individual i has the jth genotype at the QTL (for a given position) conditional on its genotype at neighbouring marker loci.

(iii) *Computation of the probability matrix*, $\mathbf{P}$

At any given genome position, elements of $\mathbf{P}$ are $p(ij) = Pr(GQ(i) = j/GM(i))$, where $GQ(i)$ is the QTL genotype (j denotes the jth possible QTL genotype) and $GM(i)$ is the vector of genotypes of individual i for all markers of the chromosome. For computation of $p(ij)$ only a subset of markers is used (those which provide information on the QTL genotype). The number of markers used in $GM(i)$ would vary among individuals and among positions for the same individual. The first step in computing $p(ij)$ is thus to extract the vector of informative markers from $GM(i)$. Let us denote this vector by $gm(i)$; $gm(i)$ thus contains genotypes of markers which are informative for individual i (not missing). If markers are co-dominant then only the two informative markers (having no missing data) flanking the QTL position are useful. If a dominant marker is encountered (at left or right of the genome position under study) other markers are included in the analysis. Limiting the number of informative markers to be used on each side of the QTL position to three seems to be a good compromise, as little gain of information is expected from using additional markers. Two closely linked (less than 5 cM apart) dominant markers linked in repulsion (on the same side of the QTL) have the same information content as a single co-dominant marker (Plomion *et al.*, 1995).

(iv) *Estimation and test of QTL effects*

Applying model (2) at each genome position, parameter estimates and tests are computed using

https://doi.org/10.1017/S0016672399004358 Published online by Cambridge University Press

standard linear model procedures (ordinary least-squares and *F*-test). Note that taking $\mathrm{Var}(e) = \sigma^2 R$, where R is a diagonal matrix whose elements are functions of QTL probabilities and QTL parameters, increases the accuracy of estimates when the QTL effects are large and/or marker density is weak (Xu, 1998). In this case, an iteratively reweighted least-squares algorithm should be used.

Given estimates of a_i and d_{ij}, parameters, additive and dominance QTL variances, denoted respectively as σ_a^2 and σ_d^2, could be estimated by the expressions (see also Xu, 1998):

$$\sigma_a^2 = \boldsymbol{\beta}_\mathbf{a}' \mathbf{X}_\mathbf{a}' (\mathbf{X}_{\mathbf{oa}} \mathbf{X}_{\mathbf{oa}}^- - \mathbf{X}_\mathbf{o} \mathbf{X}_\mathbf{o}^-) \mathbf{X}_\mathbf{a} \boldsymbol{\beta}_\mathbf{a}$$

and

$$\sigma_d^2 = \boldsymbol{\beta}_\mathbf{d}' \mathbf{X}_\mathbf{d}' (\mathbf{X}\mathbf{X}^- - \mathbf{X}_{\mathbf{oa}} \mathbf{X}_{\mathbf{oa}}^-) \mathbf{X}_\mathbf{d} \boldsymbol{\beta}_\mathbf{d},$$

where $\mathbf{X}^- = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. For complete half-diallel crosses with l parents these expressions simplify to $\sigma_a^2 = (\Sigma_i a_i^2)/(l-1)$ and $\sigma_d^2 = 2(\Sigma d_{ij}^2)/l(l-1)$. One can calculate an estimate of the variance due to the QTL as $\sigma_q^2 = \sigma_a^2 + \sigma_d^2$ and thus the part of total variance explained by the QTL as $r^2 = \sigma_q^2/(\sigma_q^2 + \sigma^2)$. A dominance ratio at the QTL could also be calculated as σ_d/σ_a. In all the above expressions, parameters are replaced by their estimated values and the estimates so obtained are asymptotically unbiased.

In our model it is assumed *a priori* that the QTL is fully informative (having l alleles, one for each parent), but it is possible to have a rough idea *a posteriori* of the actual number of alleles segregating for the QTL. This can be achieved by pairwise comparisons of additive effects at the QTL (using *t*-tests) since the sampling variances of these effects could be obtained from the estimated covariance matrix of the parameters.

3. Application to incomplete diallels and outbred crosses

(i) *Application to diallels of inbred lines with missing crosses*

When the number of parents used in the diallel is large deriving and evaluating all possible crosses could become unmanageable. One can then use either several disconnected small diallels or a partial diallel by choosing some crosses of particular interest. In this latter case the number of populations p is less than $l(l-1)/2$ and some of the genotypes are not observed. One can show that if $p < l-1$ and there is no common parent between families then the design could no longer be considered as a diallel but QTL mapping can still be done with a different model by combining information from all populations (see for example Xu, 1998); however, with such an approach QTL effects are estimated using within-family information and would be similar to those obtained at the same map position from each family analysed separately (simulation results not shown).

If $p \geqslant l-1$ then there are least two populations with a common parent. In this case, $(l-1)$ additive parameters of the QTL are estimable (with the usual constraint $\Sigma a_i = 0$) but only p dominance parameters are estimable (one for each observed heterozygous genotype). However, this condition does not necessarily imply that the QTL analysis using the method described above gives parameter estimates which are more accurate (having a smaller sampling variance) than the estimates obtained from an analysis where the populations are considered as independent ones. One can show that this is true if and only if the two-way cross classification represented by the diallel is connected in the sense of connectedness in N-ways cross classifications (Weeks & Williams, 1964). Using the algorithm proposed by Weeks & Williams (1964) it is possible to check the connectedness of the diallel design and, if it is not connected, to find all connected subsets. These sub-diallels could therefore be analysed independently by the method described above.

Note that the sampling error of the additive effect of any QTL allele depends on the number of replications of the parent carrying this allele. Parents involved in more crosses will have their allele effects estimated with better precision (simulation results not shown).

(ii) *Application to outbred populations*

Full-sib families derived from diallel crosses among outbred lines assuming known linkage phases of the markers could be analysed using the model described above with some modifications. In fact, for a full-sib family, there are four types of markers depending on the number of alleles differing between the two parents; markers could be segregating for four, three or two alleles. In this latter case, a cross between two parents could be of three different types – $aa \times ab$ or $ab \times aa$ (backcross-like) or $ab \times ab$ (F2-like) – where a and b designate two different alleles. For three and four alleles, the cross is of type $ab \times ac$ and $ab \times cd$, respectively and there are four different genotypes in a full-sib family from such crosses. To calculate the probabilities of possible QTL genotypes in outbred crosses we also need information on the linkage phase of the markers which allows the haplotypes of the parents to be deduced. These linkage phases could be inferred from the marker data.

In the QTL mapping model we assume *a priori* that the QTL is fully informative among the parents (each outbred parent has two different alleles which are also different from those of other parents). Let us denote as *bi* and *ci* the alleles carried by the parent *Li*. With

https://doi.org/10.1017/S0016672399004358 Published online by Cambridge University Press

l parents and for a complete half-diallel, the number of possible genotypes for the QTL is thus $g = 4p = 2l(l-1)$. The estimable QTL parameters are: l additive parameters (one for each parent Li) and $l(l-1)/2$ dominance parameters (one for each population). We denote as a_i the additive effect of alleles of parent Li so that alleles bi and ci contribute to the genotypic value with $+a_i$ and $-a_i$, respectively. In a full-sib family from the cross $Li \times Lj$ the genotypic values of the four possible genotypes are thus:

$bibj$: $a_i + a_j + d_{ij}$ $bicj$: $a_i - a_j - d_{ij}$

$cibj$: $-a_i + a_j - d_{ij}$ $cicj$: $-a_i - a_j + d_{ij}$.

In this case, matrix **P** will be constructed at the within-family level: the first four columns of **P** are relative to the first family, the second four columns to the second family, etc. For the computation of the probabilities of possible genotypes at the QTL, markers are used sequentially until a fully informative marker is encountered or maximum information is reached. Note that these probabilities depend on both marker genotypes and linkage phases. For an incomplete diallel cross with p families ($p < l(l-1)/2$), the number of possible genotypes at the QTL is $4p$ and the number of estimable parameters is $(l+p)$ (l additive and p dominance parameters) whatever the number of observed families. There is no necessary condition on the values of l and p for the diallel to be analysed.

(iii) *Implementation*

The algorithms and statistical procedures used for the computation of matrices **P** and **X** (including numerical expressions of matrices **C**) and for parameter test and estimation are described in detail in Rebaï (1996). The programs for analysis of diallel data are now available in the MultiCrossQTL software (Rebaï *et al*., 1997*b*).

4. Discussion

The QTL mapping method described here is a general approach for combining data from different crosses. It could quite easily be adapted to other mating designs or for independent populations (having no parents in common) of different types. Data from different experiments and locations may be combined as well.

The generalization of our approach for multiple QTL mapping with cofactors (Jansen & Stam, 1994) is theoretically easy but may encounter some difficulty in practice, especially in the choice of marker cofactors. In fact, if fully informative markers are not available one may be obliged to use markers which will be non-informative in some crosses as cofactors. In this case, for each non-informative marker in a given cross the nearest informative marker will be used to compute the probability of marker genotypes, thereby complicating the analysis. An alternative is to use different sets of marker cofactors for different families, but this could increase substantially the number of model parameters, especially when the number of families is large, and may result in a significant loss in power.

Finally, it would be interesting to compare the performance of our fixed-model strategy for diallel analysis with the IBD-based random model approach of Xie *et al*. (1998) and, in particular, to study their relative power as the number of parents increases. Xu (1998) has shown that the random model approach is computationally superior to the fixed model when the number of families is large, but the two strategies perform equally well. We think that there could be a critical number of parents above which a random approach may become more powerful, but this will be addressed elsewhere.

References

Cockerham, C. C. & Zeng, Z.-B. (1996). Design III with marker loci. *Genetics* **143**, 1437–1456.

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.

Muranty, H. (1996). Power of tests for quantitative trait loci detection using full-sib families in different schemes. *Heredity* **76**, 156–165.

Plomion, C., Liu, B.-H., & O'Malley, A. E. (1995). Genetic analysis using trans-dominant markers in F2 family. *Theoretical and Applied Genetics* **93**, 1083–1089.

Rebaï, A. (1996). *Rapport de spécification du logicel de cartographie de QTL chez les plantes: MCQTL*. Technical report. Institut National de la Recherche Agronomique, Unité de Biométrie et Intelligence Artificielle, Toulouse.

Rebaï, A. (1997). Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. *Genetical Research* **69**, 69–74.

Rebaï, A. & Goffinet, B. (1993). Power of tests for QTL detection using replicated progenies derived from a diallel cross. *Theoretical and Applied Genetics* **86**, 1014–1022.

Rebaï, A. & Goffinet, B. (1996). Correction: power of tests for QTL detection using replicated progenies derived from a diallel cross. *Theoretical and Applied Genetics* **92**, 128–129.

Rebaï, A., Goffinet, B., Mangin, B. & Perret, D. (1994). Detecting QTL with diallel schemes. In *Biometrics in Plant Breeding: Applications of Molecular Markers. Proceedings of the Ninth Eucarpia Meeting, section Biometrics in Plant Breeding* (Jansen, J. & van Ooijen, J. W., editors), pp. 170–177, Wageningen, The Netherlands.

Rebaï, A., Blanchard, P., Perret, D. & Vincourt, P. (1997*a*). Mapping quantitative trait loci controlling silking date in a diallel cross among four lines of maize. *Theoretical and Applied Genetics* **95**, 451–459.

Rebaï, A., Jourjon, M. F., Goffinet, B. & Mangin, B. (1997*b*). QTL mapping in multiple crossing designs using MulticrossQTL. In *Advances in Biometrical Genetics*,

https://doi.org/10.1017/S0016672399004358 Published online by Cambridge University Press

Proceedings of the Tenth Eucarpia Meeting, *section Biometrics in Plant Breeding*, (Krajewskiand, P. & Kaczmarek, Z., editors), pp. 225–229, Poznan, Poland.

Weeks, D. L. & Williams, D. R. (1964). A note on the determination of connectedness in an N-way cross classification. *Technometrics* **6**, 319–324.

Xie, C., Gessler, D. G. & Xu, S. (1998). Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. *Genetics* **149**, 1139–1146.

Xu, S. (1996). Mapping quantitative trait loci using four-way crosses. *Genetical Research* **68**, 175–181.

Xu, S. (1998). Mapping quantitative trait loci using multiple families of line crosses. *Genetics* **148**, 517–524.