A Multi-Marker Model for Detecting Chromosomal Segments Displaying QTL Activity

François Rodolphe and Marianne Lefort

INRA, Centre de Recherches de Jouy en Josas, Laboratoire de Biométrie, 78350 Jouy en Josas, France

Manuscript received August 1, 1992
Accepted for publication April 13, 1993

ABSTRACT

A statistical method is presented for detecting quantitative trait loci (QTLs), based on the linear model. Unlike methods able to detect a few well separated QTLs and to estimate their effects and positions, this method considers the genome as a whole and enables the detection of chromosomal segments involved in the differences between two homozygous lines, and their backcross, doubled haploid, or F_2 progenies, for a quantitative trait. Genetic markers must be codominant, but missing markers are accepted, provided they are missing independently from the experiment. Asymptotic properties, which are of practical use, are developed. This method does not rely on strong genetic hypotheses, and thus does not permit any precise genetic analysis of the trait under study, but it does assess which regions of the genome are involved, whatever the complexity of the genetic determinism (number, effects and interactions among QTLs). Simultaneous use of several methods, including this one, should lead to better efficiency in QTL detection.

FOR a long time, geneticists and breeders have been interested in the genetic analysis of quantitative characters. However, traditional quantitative genetic methodology is based on the statistical properties of the total effect of all loci contributing to quantitative variation, and not their number, locations and individual effects. The recent use of codominant molecular markers opens new perspectives for the study of the genetic basis of quantitative variation; for it is possible to ascribe a fraction of the genetic basis of quantitative variation to several Mendelian loci (QTLs) having major effects. Such studies are based on the statistical associations between the "genotypic" variation of markers and the quantitative trait variation, among a population of segregating individuals where linkage disequilibrium is maximized (F_2, backcross, doubled haploid lines, recombinant inbred lines).

Different statistical methods have been developed for locating and estimating genetic effects of QTLs: the first ones use the information from each individual marker separately. Among methods considering single marker information, the simplest one is the linear model analysis (SOLLER and BRODY 1976; SOLLER, BRODY and GENIZI 1979): the comparison of phenotypic means at each marker leads to the estimation of the effect of different QTLs (EDWARDS, STUBER and WENDEL 1987; STUBER, EDWARDS and WENDEL 1987; TANKSLEY and HEWITT 1988). However, in this method both the recombination fraction and the QTL effect are confounded and the effects of QTLs are always underestimated. WELLER (1986) proposed a maximum likelihood (ML) approach, involving three unknown parameters; the estimates of the parameters were not good. LUO and KEARSEY (1989) used a more simple likelihood function with only one unknown parameter: the recombination fraction; they concluded from simulation studies with 500 F_2 that this approach led to accurate estimates of parameters as long as the heritability was not less than 0.10. However, DARVASI and WELLER (1991) showed that the last method led to results differing from those obtained with the "true" ML method (in a seven-parameter likelihood space), especially for a dominant QTL loosely linked to the genetic marker considered.

Recently, much more efficient methods considering pairs of successive markers flanking a putative QTL have been developed. The interval mapping method is based on maximum likelihood parameter estimation, and provides a test for QTL detection, based on the ratio of likelihood functions, under the following hypotheses: there is a QTL lying in the considered interval *vs.* there is no QTL in the interval (LANDER and GREEN 1987; LANDER and BOTSTEIN 1989) (LB). The method is indubitably the best in case of F_2 or backcross populations, assuming only one QTL and a Gaussian distribution for the considered character. Such a hypothesis may appear restrictive for the study of quantitative trait determinism, for which one can imagine the occurrence of several loci, each with small effects, dispersed all over the genome or grouped over specific chromosomal regions. In this case, when considering an interval flanked by two markers, the character distribution is not Gaussian but is a mixture of Gaussian laws; the situation gets even more complicated when QTLs belong to the same linkage

group. Other flanking marker methods were proposed by KNAPP (1991) (K), essentially based on first and second order moments comparison, and CARBONELL *et al.* (1992) (C). They are probably robust and better suit non-Gaussian distributions. However, their purpose is still detection and characterization of individual QTLs. The situation of several QTLs, each with a small effect and genetically linked, has not been addressed.

This last problem is considered in the present paper. We propose here a method based on the linear model, with the genome considered as a whole, which can be used in a very general genetic context. The purpose is to detect the regions displaying QTL effects, rather than to detect isolated QTLs and to estimate their effects. Simultaneous use of different methods, each optimal for a specific situation, is recommended to improve the efficiency for QTL searching as well as the understanding of the genetic basis of quantitative character variation.

STATISTICAL MODEL

Consider that n genotypes are obtained independently, from a cross between two homozygous lines A and B. The cross can be a hybrid autofecondation (F_2), a backcross (BC), a testcross (TC), or a doubled haploidization (DH). Formally TC and DH are not different from BC and will no longer be mentioned. These genotypes are characterized independently for a quantitative trait of interest (Y_i, $i = 1, \ldots n$), and for a set of m inherited codominant markers (M_i, $i = 1, \ldots n$).

We may then write: $Y_i = E[Y_i/M_i] + E_i$, where $E[Y_i/M_i]$ represents the conditional expectation of Y_i given M_i, E_i are independent random variables supposed to be identically distributed, with $E[E_i] = 0$; $V[E_i] = \sigma^2$. This assumption will be discussed later.

M_i belongs to the set of all possible configurations for the m markers. The conditional expectation, $E[Y_i/M_i]$, is a real valued function on this set, usually very large (with 2^m (BC) or 3^m (F_2) elements). In order to make it estimable, it must be considerably constrained; the following models will be considered here:

$$Y_i = \mu + \sum_{j=1}^{m} \begin{cases} +\alpha_j & \text{for } M_i(j) = A \\ -\alpha_j & \text{for } M_i(j) = B \end{cases} + E_i, \quad \text{for a BC.}$$

$$Y_i = \mu + \sum_{j=1}^{m} \begin{cases} +\alpha_j - \delta_j & \text{for } M_i(j) = AA \\ +\delta_j & \text{for } M_i(j) = AB \\ -\alpha_j - \delta_j & \text{for } M_i(j) = BB \end{cases} + E_i, \text{ for an } F_2.$$

$M_i(j)$ represents the value taken by the jth marker on individual i. Parameters α_j and δ_j are statistical effects associated with the markers, their interpretation in terms of additivity and dominance is obvious; no epistasis is assumed.

The previous models can be written following a matrix notation:

$$Y = R\Theta + E \quad E[E] = 0 \quad V[E] = \sigma^2 I$$

where, Θ is the vector of p unknown parameters:

$$\Theta = \{\mu, \alpha_1, \ldots, \alpha_m\}, \quad \text{for a BC;}$$

$$\Theta = \{\mu, \alpha_1, \ldots, \alpha_m, \delta_1, \ldots, \delta_m\}, \quad \text{for an } F_2$$

R is a $n \times p$ matrix filled with $\{-1, +1\}$ or $\{-1, 0, +1\}$, depending on the cross-design (BC or F_2). This matrix is defined by the segregation of the markers: it is random and usually uncontrolled. Estimation of Θ as well as tests are made conditional on R, hence we have here a linear fixed-effects model (SCHEFFÉ 1959; COURSOL 1980). The statistical analysis of these models is well known. The Gauss-Markov estimator for Θ is $\hat{\Theta}$ and the best estimator of σ^2 is $\hat{\sigma}^2$:

$$\hat{\Theta} = (R'R)^{-1}R'Y, \quad \hat{\Theta} \rightsquigarrow N(\Theta, \sigma^2[R'R]^{-1}) \text{ and,}$$

$$\hat{\sigma}^2 = \|Y - R\hat{\Theta}\|^2/(n - p).$$

Tests are available for testing the existence of any class of effects. However, it is interesting to place these tests in an asymptotic frame. If n is large enough, the distributions of several statistics are well approximated by their limits as n tends to infinity. Here, the limit distributions of estimators and test statistics possess interesting properties: as the number of individuals n must be large enough with respect to the number of parameters p, these properties are of practical use.

ASYMPTOTIC PROPERTIES

We suppose that no selection is made on the genotypes to be taken into account in the analysis. The observed individuals are obtained independently from the same cross-design; the R matrix is built up with independent equidistributed random rows (R_i, $i = 1, \ldots n$). Then:

$$\frac{1}{n}[R'R] = \frac{1}{n}\sum_{i=1}^{n}[R_i'R_i] \xrightarrow{\text{a.s.}} E[R_i'R_i] = U$$

(strong law of large numbers)

and

$$\sqrt{n}(\hat{\Theta}_n - \Theta) \xrightarrow{\mathcal{L}} N(0, \sigma^2 U^{-1}) \quad \text{(central limit theorem)}$$

where $\hat{\Theta}_n$ refers to the Gauss-Markov estimator of Θ in an experiment with n different individuals, and U^{-1} stands for the inverse of U. Notice that $\forall n$, $E^{-1}[R'R] \leqslant E[(R'R)^{-1}]$ (in the sense of inequalities between positive semidefinite matrices) hence the asymptotic equivalence above for the variance-covariance matrix of $\hat{\Theta}$ is, with finite n, a lower bound for the mean.

Whatever the cross-design, the matrix U is calculable from the genetic map. It is invertible, and for both

F_2 and BC, its inverse has a simple structure: it is block-diagonal and tridiagonal. Results for these two crosses are presented in APPENDIX A.

We get the following asymptotic result: the partition of all parameters, following the chromosomes and the type of effects (additivity or dominance), is orthogonal.

The correlation structure of $\hat{\Theta}$ is simple: an estimator $\hat{\theta}_j$, of an effect attached to a marker, is correlated only with the estimators of the effects of the same type, for flanking markers. We have:

BC:

$$V[\hat{\alpha}_j] = \frac{\sigma^2}{n} \frac{1 - e^{-4\Delta_{j-1,j+1}}}{(1 - e^{-4\Delta_{j-1,j}})(1 - e^{-4\Delta_{j,j+1}})}$$

$$= \frac{\sigma^2}{4n} \frac{\Delta_{j-1,j+1}}{\Delta_{j-1,j}\Delta_{j,j+1}} + o(\Delta_{j-1,j}, \Delta_{j,j+1})$$

and

$$C[\hat{\alpha}_j, \hat{\alpha}_{j+1}] = \frac{\sigma^2}{n} \frac{-e^{-2\Delta_{j,j+1}}}{1 - e^{-4\Delta_{j,j+1}}}$$

$$= -\frac{\sigma^2}{4n} \frac{1 - 2\Delta_{j,j+1}}{\Delta_{j,j+1}} + o(\Delta_{j,j+1}).$$

F_2, for additive effects:

$$V[\hat{\alpha}_j] = \frac{2\sigma^2}{n} \frac{1 - e^{-4\Delta_{j-1,j+1}}}{(1 - e^{-4\Delta_{j-1,j}})(1 - e^{-4\Delta_{j,j+1}})}$$

$$= \frac{\sigma^2}{2n} \frac{\Delta_{j-1,j+1}}{\Delta_{j-1,j}\Delta_{j,j+1}} + o(\Delta_{j-1,j}, \Delta_{j,j+1})$$

and

$$C[\hat{\alpha}_j, \hat{\alpha}_{j+1}] = \frac{2\sigma^2}{n} \frac{-e^{-2\Delta_{j,j+1}}}{1 - e^{-4\Delta_{j,j+1}}}$$

$$= -\frac{\sigma^2}{2n} \frac{1 - 2\Delta_{j,j+1}}{\Delta_{j,j+1}} + o(\Delta_{j,j+1}).$$

F_2, for dominant effects:

$$V[\hat{\delta}_j] = \frac{\sigma^2}{n} \frac{1 - e^{-8\Delta_{j-1,j+1}}}{(1 - e^{-8\Delta_{j-1,j}})(1 - e^{-8\Delta_{j,j+1}})}$$

$$= \frac{\sigma^2}{8n} \frac{\Delta_{j-1,j+1}}{\Delta_{j-1,j}\Delta_{j,j+1}} + o(\Delta_{j-1,j}, \Delta_{j,j+1})$$

and

$$C[\hat{\delta}_j, \hat{\delta}_{j+1}] = \frac{\sigma^2}{n} \frac{-e^{-4\Delta_{j,j+1}}}{1 - e^{-8\Delta_{j,j+1}}}$$

$$= -\frac{\sigma^2}{8n} \frac{1 - 4\Delta_{j,j+1}}{\Delta_{j,j+1}} + o(\Delta_{j,j+1}).$$

Where $\Delta_{j,k}$ is the genetic distance (in Morgans) between markers j and k. We have:

$$V[\hat{\alpha}_j] \text{ and } V[\hat{\delta}_j] \nearrow +\infty \quad \text{as} \quad n\Delta_{j-1,j} \searrow 0$$

$$\text{or} \quad n\Delta_{j,j+1} \searrow 0$$

$$C[\hat{\alpha}_j, \hat{\alpha}_{j+1}] \text{ and } C[\hat{\delta}_j, \hat{\delta}_{j+1}] \searrow -\infty \quad \text{as} \quad n\Delta_{j,j+1} \searrow 0$$

$$V[\hat{\alpha}_j] \rightsquigarrow \frac{\sigma^2}{n} \text{ (BC) } \frac{2\sigma^2}{n} \text{ (F_2) } V[\hat{\delta}_j] \rightsquigarrow \frac{\sigma^2}{n} \text{ (F_2)}$$

$$\text{as} \quad \Delta_{j-1,j} \nearrow \infty \text{ and } \Delta_{j,j+1} \nearrow \infty$$

$$C[\hat{\alpha}_j, \hat{\alpha}_{j+1}] \text{ and } C[\hat{\delta}_j, \hat{\delta}_{j+1}] \nearrow 0 \quad \text{as} \quad \Delta_{j,j+1} \nearrow \infty.$$

GENETIC INTERPRETATION

This statistical model is empirical; since there is no precise genetic model, a genetic interpretation is not unique. We shall now examine the genetic interpretation, in the asymptotic frame, assuming that the differences between the genotypes (parent lines and their progenies) are due to genes (QTLs) with no epistasis, distributed all along the genome (calculations are developed in APPENDIX B).

Effects of QTLs: For BC and F_2 crosses, the only QTLs which contribute to $\hat{\alpha}_j$ or $\hat{\delta}_j$ (additive or dominant effect in the linear model attached to the jth marker) are those located between markers $j - 1$ and $j + 1$. Roughly speaking, each effect of a QTL is shared between its flanking markers, in proportion with their probabilities of segregating with it, in case of recombination between them. In fact the modulus is slightly biased (underestimated) due to the possibility of multiple recombination between both flanking markers.

We can define, for the comparisons under study, two functions b and c along each chromosome representing the additive and dominant effects of the gene present at each locus belonging to this chromosome (see definition of b_q and c_q in APPENDIX B). These functions are very irregular, taking in particular the value 0 when both genes are identical, but there is no obvious reason, in general, to suppose the number of QTLs (in this sense all polymorphic loci) to be small. With this global linear model, we are estimating $\alpha_j = \int_{M_{j-1}}^{M_{j+1}} b\phi_j$ and $\delta_j = \int_{M_{j-1}}^{M_{j+1}} c\psi_j$, $(j = 1, \ldots m)$ a discrete approximation of the functions b and c. If we neglect double recombinations ϕ_j is identical to ψ_j and is the "chinese hat" piecewise linear function, taking the value 1 on marker j, 0 on its left and right neighbours, and outside this interval. Clearly, looking at the variance-covariance of $\hat{\Theta}$, there is a conflict between fineness (marker density) and precision (estimator variance) in this attempt to estimate the functions b and c.

Homoscedasticity: The assumption of homoscedasticity ($V[E_j] = \sigma^2$) is in this context artificial. If there are QTLs lying between two markers, recombinant

genotypes between these two markers will have a greater variance than nonrecombinant ones. However, our aim was to treat chromosome segments, not isolated QTLs. Without a detailed genetic model, it is impossible to solve this contradiction. Furthermore, the fixed effect linear model is sufficiently robust against heteroscedasticity.

TESTING QTL ACTIVITY

Consequences of the orthogonality property: Consider a partition $(\Theta_1, \ldots \Theta_h)$ of Θ into h $(h \leq p)$ subsets of components, the related hypotheses H_{ω_k}:$\{\Theta_k = 0\}$ with their alternatives $H_{\omega_k^c}$: $\{\Theta_k \neq 0\}$, and the corresponding Fisher test statistics $F_k = (\hat{\Theta}_k' \hat{V}^{-1}[\hat{\Theta}_k]\hat{\Theta}_k)/m_k$ to test H_{ω_k} against $H_{\omega_k^c}$, $k = 1, \ldots h$; where $\hat{\Theta}_k$ is the restriction of $\hat{\Theta}$ to Θ_k, $\hat{V}^{-1}[\hat{\Theta}_k]$ the inverse of its variance-covariance matrix estimator, and m_k the dimension of Θ_k (here, the number of markers on the chromosome).

Orthogonality for such a partition of Θ (COURSOL 1980), means that the numerators of the Fisher test statistics F_k are independent, and that the distribution of F_k depends only on σ^2 and Θ_k. F_k is distributed as a Fisher-Snedecor $F(m_k, n - p)$ under H_{ω_k}, as an $F'(m_k, n - p)$ with noncentrality parameter $\Theta_k' V^{-1}[\hat{\Theta}_k]\Theta_k$ under $H_{\omega_k^c}$.

Asymptotically, testing H_{ω_k} against $H_{\omega_k^c}$, is purely testing the additive (respectively dominant) QTL activity on this chromosome without being influenced by the existence (or not) of a dominant (respectively additive) QTL activity on the same chromosome, nor by the existence of any QTL activity on other chromosomes. This would not be the case without orthogonality.

Power considerations: The power of a Fisher F test depends only on its noncentrality parameter (JOHNSON and KOTZ 1970; COURSOL 1980). Convergence of $V(\hat{\Theta})$ to a block-diagonal matrix (see APPENDIX A) leads finally to simple asymptotic equivalences for the noncentrality parameter of F_k:

$$n\Theta_k' A_k \Theta_k = n \sum_{j,j' \in \text{chromosome}} \alpha_j \alpha_{j'} e^{-2\Delta_{j,j'}}/\sigma^2 \quad \text{for a BC.}$$

$$\frac{n}{2}\Theta_k' A_k \Theta_k = n \sum_{j,j' \in \text{chromosome}} \alpha_j \alpha_{j'} e^{-2\Delta_{j,j'}}/2\sigma^2$$

for additive effects in an F_2.

$$n\Theta_k' B_k \Theta_k = n \sum_{j,j' \in \text{chromosome}} \delta_j \delta_{j'} e^{-4\Delta_{j,j'}}/\sigma^2$$

for dominant effects in an F_2.

More accuracy in detecting QTL activity: SCHEFFÉ's S method (SCHEFFÉ 1959; COURSOL 1980) provides a *simultaneous* test, S_k, for all linear combinations of effects belonging to a subset Θ_k, which is coherent with the corresponding Fisher test: if the F_k statistics is not significant at level α, no linear combination of the parameters belonging to Θ_k will be declared significantly non-zero by the S_k test at the same level; on the contrary, if the F_k statistics is significant at level α, there is at least a linear combination of the parameters belonging to Θ_k which is significantly non-zero in the S_k test at the same level. This test is based on the following theorem, which expresses that Θ_k belongs with probability $1 - \alpha$ to its confidence ellipsoïd (constructed under no particular hypothesis on Θ_k):

Let m_k be the number of parameters belonging to Θ_k (the number of markers on the chromosome). $\forall \alpha$, $P[\forall c \in \mathbb{R}^{m_k} : |c'\hat{\Theta}_k - c'\Theta_k| \leq (\hat{\sigma}^2_{c'\hat{\theta}_k} m_k \cdot f(\alpha; m_k, n - p))^{1/2}] = 1 - \alpha$ where $\hat{\sigma}^2_{c'\hat{\theta}_k} = c'\hat{V}[\hat{\Theta}_k]c = \hat{\sigma}^2 c'(R'R)_k^{-1} c$ is the variance estimator of $c'\hat{\Theta}_k$ estimator of $c'\Theta_k$ linear combination on Θ_k, $(R'R)_k^{-1}$ the $m_k \times m_k$ matrix restriction of $(R'R)^{-1}$ to rows and columns corresponding to Θ_k, and $f(\alpha; m_k, n - p)$ the α upper quantile of the Fisher-Snedecor distribution with m_k and $n - p$ degrees of freedom.

Thus we can decide, simultaneously for all linear combinations on Θ_k, and with a global level α, $c'\Theta_k \neq 0$ iff $|c'\hat{\theta}_k| > (\hat{\sigma}^2_{c'\hat{\theta}_k} m_k \cdot f(\alpha; m_k, n - p))^{1/2}$. Asymptotic orthogonality for the partition of Θ makes these S_k tests asymptotically independent.

This enables one, each time a type of effect (additive or dominant) has been detected as significant on a whole chromosome (by rejecting the corresponding H_{ω_k} null hypothesis) to identify, with a global controlled level, which linear combinations of effects are significantly different from zero, and thus, which segments appear to be responsible for the differences between both parent lines and within their recombination products for the trait under study.

Usually sums of additive (respectively dominant) effects of successive markers are tested, because they are easy to interpret, one is testing the total QTL activity of the segment covered by these markers. It can happen that no individual effect is significant, but that their sum is. More complex linear combinations of such parameters can also be of interest.

Suppose there are two QTLs on a segment, one with a positive effect, the other with a negative one; the total effect can be low and not significantly different from zero, but the difference (eventually weighted) between the effects attached to the flanking markers can be significantly non-zero. Of course, such combinations contribute little to the difference between the parent lines, but do contribute to differences between recombinant genotypes for these QTLs; they are therefore not easy to detect. Here, the variance of a linear combination with negative coefficients is greater than the variance of the linear combination with the same coefficients but all of the same sign, because of the negative covariances between effect estimators on neighbour markers, and

thus the power of the S test is less for a difference than for a sum.

MISSING DATA

Missing data are always encountered in such experiments. Unlike the methods based on sequential screening of intervals, it is impossible here to simply discard individuals for which there are missing markers on the examined segment. Even with a small proportion of missing markers, as we are considering here all markers simultaneously, the proportion of individuals for which there would be at least one missing marker, would be too high.

In our model, missing markers are unobserved terms of the R matrix. We can calculate the joint distribution of these missing terms, given the observed ones, which depends on a recombination model (we assume Haldane's recombination model, with no difference between male and female recombination).

Let M_{obs} be the set of observed markers. We define $R_{exp} = E[R/M_{obs}]$, the $n \times p$ matrix corresponding to R where missing terms are replaced by their conditional expectation given all known markers; it will be supposed to be full rank (calculation of R_{exp} is detailed in APPENDIX C). We define:

$$\tilde{\Theta} = (R'_{exp}R_{exp})^{-1}R'_{exp}Y.$$

The following result holds. If we assume that markers are missing at random (independently from their value and Y), given M_{obs}, $\tilde{\Theta}$ is an unbiased Gaussian estimator of Θ. $\tilde{\Theta}$ is consistent, and has the same asymptotic orthogonality properties as $\hat{\Theta}$ (for the partition of Θ following chromosomes and additive and dominance effects).

Conditional moments of $\tilde{\Theta}$:

$$E[\tilde{\Theta}/M_{obs}] = (R'_{exp}R_{exp})^{-1}R'_{exp}E[Y/M_{obs}] = \Theta.$$

$$\begin{aligned} V[\tilde{\Theta}/M_{obs}] &= (R'_{exp}R_{exp})^{-1}R'_{exp}E[(R-R_{exp})\Theta\Theta'(R-R_{exp})' \\ &\quad + EE'/M_{obs}]R_{exp}(R'_{exp}R_{exp})^{-1} \\ &= \sigma^2(R'_{exp}R_{exp})^{-1} + (R'_{exp}R_{exp})^{-1}R'_{exp}E[(R-R_{exp}) \\ &\quad \Theta\Theta'(R-R_{exp})'/M_{obs}]R_{exp}(R'_{exp}R_{exp})^{-1}. \end{aligned}$$

The term (i,i') of $D = E[(R - R_{exp})\Theta\Theta'(R - R_{exp})'/M_{obs}]$ is:

$$E\Big[\sum_j \sum_{j'} (r_{i,j} - r_{expi,j})\theta_j\theta_{j'}(r_{i',j'} - r_{expi',j'})/M_{obs}\Big] = \sum_j \sum_{j'} \theta_j\theta_{j'}C[r_{i,j}r_{i',j'}/M_{obs}].$$

As the different genotypes are obtained independently, and markers missing independently from their value and from each other, $C[r_{i,j}r_{i',j'}/M_{obs}] = 0$ for $i \neq i'$; D is diagonal, positive and easily computed with the genetic map. Finally, we get

$$\begin{aligned} V[\tilde{\Theta}/M_{obs}] &= \sigma^2(R'_{exp}R_{exp})^{-1} \\ &\quad + (R'_{exp}R_{exp})^{-1}R'_{exp}D(\text{map}, \Theta)R_{exp}(R'_{exp}R_{exp})^{-1}. \end{aligned}$$

Notice that $E[(R'R)^{-1}/M_{obs}] \leq (R'_{exp}R_{exp})^{-1}$ (inequality between positive semidefinite matrices:

$$\begin{aligned} &\sigma^2((R'_{exp}R_{exp})^{-1} - E[(R'R)^{-1}/M_{obs}]) \\ &+ (R'_{exp}R_{exp})^{-1}R'_{exp}DR_{exp}(R'_{exp}R_{exp})^{-1} \end{aligned}$$

represents (in terms of variance) the expected cost of missing markers given M_{obs}.

Asymptotic properties of $\tilde{\Theta}$: Asymptotic properties of $\tilde{\Theta}$, rely on the following remark: the term (j, j') of $R'_{exp}DR_{exp}$ is $\sum_k\sum_l\theta_k\theta_l\sum_i r_{expi,j}r_{expi,j'}C[r_{i,k}r_{i,l}/M_{obs}]$,

$$\frac{1}{n}\sum_i r_{expi,j}r_{expi,j'}C[r_{i,k}r_{i,l}/M_{obs}] \xrightarrow{\text{a.s.}} E[r_{expi,j}r_{expi,j'}C[r_{i,k}r_{i,l}/M_{obs}]].$$

Thus $1/n\ R'_{exp}DR_{exp}$ converges to some positive matrix W function of the map, Θ, and the distribution of missing markers. Also $n(R'_{exp}R_{exp})^{-1}$ converges to some positive definite matrix V function of the map and the distribution of missing markers. Hence,

$$\sqrt{n}(\tilde{\Theta}_n - \Theta) \xrightarrow{\mathcal{L}} N(0,\ VWV + \sigma^2V)$$

where $\tilde{\Theta}_n$ refers to the estimator of Θ, with missing markers, in an experiment with n different individuals. It is shown in APPENDIX C, that W and V have the same block-diagonal structure as $E[R'R]$, thus the asymptotic orthogonality properties observed without missing data are kept (but the tridiagonal structure of $E^{-1}[R'R]$ is lost).

Estimation of σ^2 and $V[\tilde{\Theta}/M_{obs}]$: $\|Y - R_{exp}\tilde{\Theta}\|^2 = \|M((R - R_{exp})\Theta + E)\|^2$, where $M = (I - R_{exp}(R'_{exp}R_{exp})^{-1}R'_{exp})$. This is the squared norm of the projection of Y on an $(n - p)$ dimensional subspace of $\mathbb{R}^n$, hence, given R, it is distributed as a $\sigma^2\chi'^2_{n-p}$ with non-centrality parameter $\|M(R - R_{exp})\Theta\|^2/\sigma^2$. Thus $E[\|Y - R_{exp}\tilde{\Theta}\|^2/M_{obs}] = (n - p)\sigma^2 + \Theta'E[(R - R_{exp})'M(R - R_{exp})/M_{obs}]\Theta$. This last term equals: $\sum_{j,j'}\theta_j\theta_{j'}\sum_i m_{i,i}C[r_{i,j}r_{i,j'}/M_{obs}]$; it can be computed as a function of R_{exp}, the map and Θ. From this formula, we can define an estimator of σ^2,

$$\tilde{\sigma}^2 = (\|Y - R_{exp}\tilde{\Theta}\|^2 - \sum_{j,j'}\tilde{\theta}_j\tilde{\theta}_{j'}\sum_i m_{i,i}C[r_{i,j}r_{i,j'}/M_{obs}])/(n - p)$$

whose bias is in $o(1/n)$ and is thus consistent.

$V[\tilde{\Theta}/M_{obs}]$ is estimable, replacing σ^2 and Θ by $\tilde{\sigma}^2$ and $\tilde{\Theta}$ in its expression. This estimator, $\tilde{V}[\tilde{\Theta}/M_{obs}]$, is biased, but its bias is again in $o(1/n)$ and thus, it is consistent. Practical computations are developed in APPENDIX C.

Tests with missing data: Given M_{obs}, $\tilde{\Theta}_k \rightsquigarrow N(\Theta_k, V[\tilde{\Theta}_k/M_{obs}])$. Hence, $(\tilde{\Theta}_k - \Theta_k)'V^{-1}[\tilde{\Theta}_k/M_{obs}]$

$(\tilde{\Theta}_k - \Theta_k)$ is distributed as a $\chi^2_{m_k}$, where k is an index for the orthogonal partition of Θ and $V^{-1}[\tilde{\Theta}_k/M_{obs}]$ the inverse of the variance-covariance matrix of $\tilde{\Theta}_k$. Because of the block-diagonal structure of $VWV + \sigma^2 V$, these χ^2 variables, corresponding to the orthogonal partition of Θ, are asymptotically independent.

We can use this decomposition to test the different hypotheses H_{ω_k}:$\{\Theta_k = 0\}$ against their alternatives $H_{\omega_k^c}$: $\{\Theta_k \neq 0\}$: asymptotically, $\tilde{F}_k = (\tilde{\Theta}_k' \tilde{V}^{-1}[\tilde{\Theta}_k/M_{obs}]\tilde{\Theta}_k)/m_k$, are independent stochastic variables given M_{obs}, $\tilde{F}_k$ distributed as an $F(m_k, n - p)$ under H_{ω_k}, as an $F'(m_k, n - p)$ under $H_{\omega_k^c}$ with non-centrality parameter $\Theta_k' V^{-1}[\tilde{\Theta}_k/M_{obs}]\Theta_k$.

Notice that: $\Theta_k'(E[V^{-1}[\hat{\Theta}_k]/M_{obs}] - V^{-1}[\tilde{\Theta}_k/M_{obs}])$ $\Theta_k \geq 0$. It represents the asymptotic expected loss of power due to missing markers.

Extension of the S test to this situation can be as follows. If H_{ω_k} is rejected, we can test simultaneously all linear combinations of Θ_k, deciding, $\forall c \in \mathbb{R}^{m_k}$, $c'\Theta_k \neq 0$ iff: $|c'\tilde{\Theta}_k| > (c'\tilde{V}[\tilde{\Theta}_k/M_{obs}]cm_k \cdot f(\alpha; m_k, n - p))^{1/2}$. These $\tilde{S}_k$ tests are asymptotically of level α and independent.

DISCUSSION

The major interest of this method is the possibility to make global tests, recovering the whole or a part of a chromosome. In doing so, we are able to detect the existence of a set of QTLs, all belonging to the same linkage group and having small individual, or opposite effects. Of course, our method does not give precise information referring to the number, locations or effects of these QTLs, but it enables the user to select genomic segments of interest with a relatively high security with respect to the genetic determinism. As its asymptotic statistical properties are explicitly known, this method constitutes also a good tool for the design of an experiment.

The common characteristic of all existing two-marker methods is their sequential nature. This is the reason why they are not so well suited for the detection of linked QTLs. But the advantage of such a procedure is enormous in case of isolated QTLs; the power of their tests increases until a limit with the number of markers. On the contrary with our global method, the estimation precision and the power of the tests decreases with the density of the map (the proximity between markers belonging to the same chromosome, not their total number) as it can be seen on the asymptotic structure of the variance-covariance matrix of $\hat{\Theta}$.

If there is only one QTL, it is clear that sequential methods are always much better. If there are several QTLs, even belonging to different linkage groups, the comparison is not evident. We expect to reduce residual variance, by taking into account at the same time the effects of all QTLs. In such a situation, this method should be used with a reasonable amount of markers, eventually sampling available markers, but keeping markers on all linkage groups.

This remark opens a new perspective. It is possible to analyse the observations with this multimarker method, keeping markers on all chromosomes but one, and then to use a sequential method on the residuals, screening with a higher density of markers the chromosome left in the first analysis. This can be done successively on all chromosomes. Independence of chromosome segregation makes this iterative two-step procedure asymptotically correct.

Anyway, the limits of our method will only be precised by extensive simulations and real data studies.

Comparison between results obtained by our method and sequential methods can give an interesting insight especially on the number of QTLs; but if the purpose is a precise dissection of the genetic determinism of the trait under study, this can not be achieved with only one experiment: the search for QTLs must be an interactive process. The multimarker model enables the experimenter to focus on the interesting segments detected. New experiments must be made, with more markers on these segments and more individuals in order to get more recombinant genotypes between these narrower markers. These new experiments can be analyzed with more complex models (involving possibly epistatic effects); certainly, the analysis of such costly experiments is much improved, using different models and methods adapted to different situations and purposes.

In order to reduce environmental variance, one can collect the data in a field experiment. For instance, the products of a cross can be grown in blocks with a reference constituted by the parents or the F_1, and the differences between individual performances and that of the reference analyzed. More sophisticated designs are possible when the studied genotypes can be repeated (doubled haploids). Such data are not independent raw measures, but estimations of the genetic values given by the statistical analysis of this field experiment; they are therefore not independent, but have a known covariance structure. It is very easy to take this structure into account in the linear model, whereas this is generally impossible in the LB method. The reason is that the log-likelihood, at each interval, contains a sum on the elephantine set of all possible joint genotypes at the putative QTL lying in this interval, and is therefore of no practical use. Of course one can disregard the covariance structure; consequences depend on this structure itself; in some cases (with heavily unbalanced experimental designs), they could be important.

The study of F_n lines, derived from BC or F_2 pop-

ulations and then selfed for $(n - 2)$ generations, is possible. But genetic interpretation will be modified; even without interference in the recombination at each meiosis, there is an "apparent" one in the F_n lines with $n > 2$. Sequential methods are in this case not so well justified, since, the genotype at a putative QTL, given the state of its two flanking markers, is no longer independent from other markers linked with it.

LITERATURE CITED

CARBONELL, E. A., T. M. GERIG, E. BALANSARD and M. J. ASINS, 1992 Interval mapping in the analysis of nonadditive quantitative trait loci. Biometrics **48:** 305–315.

COURSOL, J., 1980 Technique statistique des modèles linéaires. 1. Aspects théoriques. Les cours du C.I.M.P.A.

DARVASI, A., and J. I. WELLER, 1991 On the use of the moments method of estimation to obtain approximate maximum likelihood estimates of linkage between a genetic marker and a quantitative locus. Heredity **68:** 43–46.

EDWARDS, M. D., C. W. STUBER and J. F. WENDEL, 1987 Molecular-marker-facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of genes action. Genetics **116:** 113–125.

JOHNSON, N. L., and S. KOTZ, 1970 Distributions in statistics: continuous univariate distributions. John Wiley, New York.

KNAPP, S. J., 1991 Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred and doubled haploid progeny. Theor. Appl. Genet. **81:** 333–338.

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

LANDER, E. S., and P. GREEN, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

LUO, Z. W., and J. M. KEARSEY, 1989 Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. Heredity **63:** 401–408.

PATERSON, A. H., J. W. DE VERNA, B. LANINI and S. D. TANKSLEY, 1990 Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecific cross of tomato. Genetics **124:** 735–742.

SCHEFFÉ, H., 1959 The analysis of variance. John Wiley, New York.

SOLLER, M., and T. BRODY, 1976 On the power of experimental designs for the detection of linkage marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. **47:** 35–59.

SOLLER, M., T. BRODY and A. GENIZI, 1979 The expected distribution of marker-linked quantitative effects in crosses between inbred lines. Heredity **43:** 179–190.

STUBER, C. W., M. D. EDWARDS and J. F. WENDEL, 1987 Molecular marker-facilitated investigations of quantitative trait loci in maize. II. Factors influencing yield and its component traits. Crop Sci. **27:** 639–648.

TANKSLEY, S. D., and J. HEWITT, 1988 Use of molecular markers for soluble solid content in tomato–a re-examination. Theor. Appl. Genet. **75:** 811–823.

WELLER, J. I., 1986 Maximum likelihood techniques for the mapping and the analysis of quantitative trait loci with the aid of genetic markers using approximate likelihood methods. Biometrics **42:** 627–640.

Communicating editor: T. F. C. MACKAY

APPENDIX

A. Asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$

In this appendix, $r_{i,j}$ represents the generic term of matrix R (ith row and jth column).

$\Delta_{jj'}$ represents the genetic distance between markers j and j'; it is expressed in Morgans and refers to Haldane's recombination model (without interference or sex linked effect).

Backcross: Remember that $r_{i,j} = +1$ if $M_i(j)$ equals A, and -1 if it equals B.

We have

$$r_{i,1} \equiv 1, \text{ and, for } j \neq 1, \begin{cases} P[r_{i,j} = +1] = \tfrac{1}{2} \\ P[r_{i,j} = -1] = \tfrac{1}{2} \end{cases}$$

$$E[r_{i,1}^2] = 1; \quad E[r_{i,1}r_{i,j}] = 0; \quad E[r_{i,j}^2] = 1$$

For $j \neq 1, j' \neq 1, j \neq j'$,

$$[r_{i,j}r_{i,j'}] = \begin{cases} +1, & \text{when markers are inherited from the same parent} \\ -1, & \text{when markers are inherited from different parents} \end{cases}$$

Considering that the recombination probability between two markers j and j' is $\rho_{jj'}$, this is the probability for the 2 markers to be inherited from different parents.

So,

$$E[r_{i,j}r_{i,j'}] = 1 - 2\rho_{jj'}.$$

We have in all cases: $E[r_{i,j}r_{i,j'}] = e^{-2\Delta_{jj'}}$.

So,

$$U = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & A_1 & 0 & \dots & 0 \\ 0 & 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \vdots & A_i & \vdots \\ 0 & 0 & 0 & \dots & A_k \end{bmatrix} \begin{matrix} 1^{\text{st}} \text{ chromosome} \\ 2^{\text{nd}} \text{ chromosome} \\ i^{\text{th}} \text{ chromosome} \\ k^{\text{th}} \text{ chromosome} \end{matrix}$$

where $A_i = [e^{-2\Delta_{jj'}}]$ (for the ith chromosome).

F_2: Remember that if $M_i(j)$ equals AA (respectively AB, BB), $r_{i,j} = +1$ (respectively 0, -1) for an additive effect, and $r_{i,j} = -1$ (respectively $+1$, -1) for a dominance effect.

$$r_{i,1} \equiv 1, \quad \mathrm{E}[r_{i,1}^2] = 1.$$

Considering j as an index for additive effects,

$$P[r_{i,j} = +1] = \tfrac{1}{4} \qquad E[r_{i,1}r_{i,j}] = 0$$

$$P[r_{i,j} = 0] = \tfrac{1}{2} \qquad E[r_{i,j}^2] = \tfrac{1}{2}.$$

$$P[r_{i,j} = -1] = \tfrac{1}{4}$$

Considering j as an index for dominance effects,

$$P[r_{i,j} = +1] = \tfrac{1}{2} \quad E[r_{i,1}r_{i,j}] = 0$$

$$P[r_{i,j} = -1] = \tfrac{1}{2} \quad E[r_{i,j}^2] = 1.$$

Considering j and j' as indexes for additive effects,

$$P[r_{i,j}r_{i,j'}=+1]=P[M_i(j)/M_i(j')=AA/AA \text{ or } BB/BB]$$
$$=(1-\rho_{jj'})^2/2$$
$$P[r_{i,j}r_{i,j'}=0] \quad =P[M_i(j)=AB \text{ or } M_i(j')=AB]$$
$$P[r_{i,j}r_{i,j'}=-1]=P[M_i(j)/M_i(j')=AA/BB \text{ or } BB/AA]$$
$$=\rho_{jj'}^2/2.$$

So,

$$E[r_{i,j}r_{i,j'}] = (1 - 2\rho_{jj'})/2.$$

Considering j and j' as indexes for dominance effects,

$$P[r_{i,j}r_{i,j'}=+1]=P[M_i(j)/M_i(j')=AA/AA \text{ or } BB/BB$$
$$\text{or } AA/BB \text{ or } BB/AA \text{ or } AB/AB]$$
$$=(1-2\rho_{jj'}+2\rho_{jj'}^2)$$
$$P[r_{i,j}r_{i,j'}=-1]=P[M_i(j)/M_i(j')=AB/AA \text{ or } AB/BB$$
$$\text{or } AA/AB \text{ or } BB/AB]$$
$$=2\rho_{jj'}(1-\rho_{jj'})$$

So,

$$E[r_{i,j}r_{i,j'}] = (1 - 2\rho_{jj'})^2$$

Considering j and j' as indexes for additive and dominance effects, we have the following table for $r_{i,j}r_{i,j'}$ values:

$M_i(j')$ / $M_i(j)$	AA	AB	BB
AA	−1	+1	−1
AB	0	0	0
BB	+1	−1	+1

and we know that:

$$P[M_i(j)/M_i(j')=AA/AA]=P[M_i(j)/M_i(j')=BB/BB]$$
$$P[M_i(j)/M_i(j')=BB/AA]=P[M_i(j)/M_i(j')=AA/BB]$$
$$P[M_i(j)/M_i(j')=AA/AB]=P[M_i(j)/M_i(j')=BB/AB]$$

So,

$$E[r_{i,j}r_{i,j'}] = 0.$$

We have in all cases: $(1 - 2\rho_{jj'})/2 = \frac{1}{2}e^{-2\Delta_{jj'}}$; $(1 - 2\rho_{jj'})^2 = e^{-4\Delta_{jj'}}$.

So,

$$U = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \cdots \\ 0 & \frac{1}{2}A_1 & 0 & \cdots & 0 & 0 & \cdots \\ 0 & 0 & \frac{1}{2}A_2 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & B_1 & 0 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & B_2 & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \end{bmatrix} \begin{matrix} \\ 1\text{st chromosome additive effects} \\ 2\text{nd chromosome additive effects} \\ \vdots \\ 1\text{st chromosome dominance effects} \\ 2\text{nd chromosome dominance effects} \\ \\ \end{matrix}$$

with $A_i = [e^{-2\Delta_{jj'}}]$ and $B_i = [e^{-4\Delta_{jj'}}]$ (for the ith chromosome).

Explicit inverses: These matrices have explicit inverses; in fact, considering that A and B have the following structure:

$$\begin{pmatrix} 1 & a_1 & a_1a_2 & \cdots & a_1a_2 \cdots a_{k-1} \\ a_1 & 1 & a_2 & \cdots & a_2 \cdots a_{k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_1a_2 \cdots a_{k-1} & a_2a_2 \ldots a_{k-1} & a_3 \cdots a_{k-1} & & 1 \end{pmatrix}$$

Their inverses have a tri-diagonal structure:

$$\begin{bmatrix} \frac{1}{1-a_1^2} & \frac{-a_1}{1-a_1^2} & 0 & \cdots & 0 \\ \frac{-a_1}{1-a_1^2} & \frac{1-a_1^2a_2^2}{(1-a_1^2)(1-a_2^2)} & \frac{-a_2}{1-a_2^2} & \cdots & 0 \\ 0 & \frac{-a_2}{1-a_2^2} & \frac{1-a_2^2a_3^2}{(1-a_2^2)(1-a_3^2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{1-a_{k-1}^2} \end{bmatrix}$$

B. Connexion between effects attached to the markers and true QTL effects

We suppose here, there are no missing data. Remember:

$$Y = R\Theta + E \quad \text{and} \quad \hat{\Theta} = (R'R)^{-1}R'Y$$
$$n[R'R]^{-1} \xrightarrow{a.s.} U^{-1}$$

In this appendix, $E[y/m_j = a]$ represents the conditional expectation of the performance of an individual, given that its jth marker equals a.

Suppose the existence of a set of genes (QTLs) distributed all over the genome, with no epistasy and additive effects b_q and dominant effects c_q, whose definition is as follows.

Let G_i be the genotype of individual i at all these

QTLs, then, we define these effects through the formulas:

$$E[Y_i/G_i] = \mu + \sum_{q=1}^{q_{max}} \begin{cases} +b_q & \text{for } G_i(q) = A \\ -b_q & \text{for } G_i(q) = B' \end{cases} \quad \text{for a BC.}$$

$$E[Y_i/G_i] = \mu + \sum_{q=1}^{q_{max}} \begin{cases} +b_q - c_q & \text{for } G_i(q) = AA \\ +c_q & \text{for } G_i(q) = AB, \\ -b_q - c_q & \text{for } G_i(q) = BB. \end{cases} \quad \text{for an } F_2.$$

This is not the classical definition of additive and dominant effects in quantitative genetics, since these effects are corrected for linkage, and, in this sense, "absolute effects" (but relative to the cross under study).

Backcross:

$$\frac{1}{n}[R'Y] \xrightarrow{\text{a.s.}} \begin{bmatrix} \mu \\ \cdots \\ \tfrac{1}{2}(\mathrm{E}[y/m_j = A] - E[y/m_j = B]) \\ \cdots \end{bmatrix}.$$

Then,

$$\hat{\alpha}_j = \frac{-a_l}{1 - a_l^2} z_{j-1} + \frac{1 - a_l^2 a_r^2}{(1 - a_l^2)(1 - a_r^2)} z_j + \frac{-a_r}{1 - a_r^2} z_{j+1}$$

where $a_l = e^{-2\Delta_{j-1,j}}$, $a_r = e^{-2\Delta_{j,j+1}}$, and $z_j = \frac{1}{2}(E[y/m_j = A] - E[y/m_j = B])$.

$$E[y/m_j = A] = \sum_q (b_q p_{q,j} - b_q(1 - p_{q,j})) = \sum_q b_q(2p_{q,j} - 1)$$

$$E[y/m_j = B] = \sum_q -b_q(2p_{q,j} - 1)$$

with $p_{q,j} = P[q^{\text{th}}qtl = A/m_j = A] = \frac{1}{2}(1 + e^{-2\Delta_{q,j}})$, where $\Delta_{q,j}$ is the genetic distance between $q^{\text{th}}qtl$ and j^{th} marker.

Thus, $z_j = \sum_q b_q e^{-2\Delta_{q,j}}$. From this formula, it is evident that QTLs located on other chromosomes do not contribute to $\hat{\alpha}_j$. Now consider the subset of all QTLs located left from marker $j - 1$; their contribution to $\hat{\alpha}_j$ is:

$$\sum_q b_q e^{-2\Delta_{q,j}} \left(\frac{-1}{1 - a_l^2} + \frac{1 - a_l^2 a_r^2}{(1 - a_l^2)(1 - a_r^2)} + \frac{-a_r^2}{1 - a_r^2} \right) = 0.$$

Now consider the subset of all QTLs located between marker $j - 1$ and marker j; their contribution to $\hat{\alpha}_j$ is:

$$\sum_q b_q \left(\frac{-a_l}{1 - a_l^2} e^{-2\Delta_{q,j-1}} + \frac{1 - a_l^2 a_r^2}{(1 - a_l^2)(1 - a_r^2)} e^{-2\Delta_{q,j}} + \frac{-a_r}{1 - a_r^2} e^{-2\Delta_{q,j+1}} \right) = \sum_q b_q e^{-2\Delta_{q,j}} \frac{1 - e^{-4\Delta_{q,j-1}}}{1 - e^{-4\Delta_{j-1,j}}},$$

which can be written $\int_{M_{j-1}}^{M_j} b\phi_j$ and whose first order approximation is: $\sum_q b_q(\Delta_{q,j-1}/\Delta_{j-1,j})$, which is also the first order approximation of $\sum_q b_q P[q\text{th}qtl = A/m_{j-1} = B \text{ and } m_j = A]$.

We may also calculate the contribution to $\hat{\alpha}_{j-1} + \hat{\alpha}_j$ of all QTLs located between marker $j - 1$ and marker j, which is:

$$\sum_q b_q(e^{-2\Delta_{q,j-1}} + e^{-2\Delta_{q,j}})/(1 + e^{-2\Delta_{j,j-1}}) = \sum_q b_q(1 - 2P[q^{\text{th}}qtl = B/m_{j-1} = A \text{ and } m_j = A])$$

whose first order approximation (in $\Delta_{j,j-1}$) is $\sum_q b_q$.

$$\mathbf{F_2}: \frac{1}{n}[R'Y] \xrightarrow{\text{a.s.}} \left[\begin{array}{c} \mu \\ \cdots \\ \tfrac{1}{4}(\mathrm{E}[y/m_j = AA] - E[y/m_j = BB]) \\ \cdots \\ \tfrac{1}{2}(E[y/m_j = AB] - E[y/m_j = AA \text{ or } BB]) \\ \cdots \end{array} \right] \begin{array}{l} \} \text{ additive effect} \\ \} \text{ dominant effect} \end{array}.$$

Then obviously, for additive terms, we have exactly the same results as for a backcross.

For a dominant effect, we have:

$$\hat{\delta}_j = \frac{-a_l^2}{1 - a_l^4} zd_{j-1} + \frac{1 - a_l^4 a_r^4}{(1 - a_l^4)(1 - a_r^4)} zd_j + \frac{-a_r^2}{1 - a_r^4} zd_{j+1}$$

where, $a_l = e^{-2\Delta_{j-1,j}}$, $a_r = e^{-2\Delta_{j,j+1}}$,
and $zd_j = \frac{1}{2}(E[y/m_j = AB] - E[y/m_j = AA \text{ or } BB])$.

$$E[y/m_j = AB] = \sum_q (c_q g_{q,j} - c_q(1 - g_{q,j})) = \sum_q c_q(2g_{q,j} - 1)$$

$$E[y/m_j = AA \text{ or } BB] = \sum_q -c_q(2g_{q,j} - 1)$$

with $g_{q,j} = P[q\text{th}qtl = AB/mj = AB] = P[q\text{th}qtl = AA \text{ or } BB/mj = AA] = P[q\text{th}qtl = BB \text{ or } AA/m_j = BB] = p_{q,j}^2 + (1 - p_{q,j})^2$.

Thus, $zd_j = \sum_q c_q(1 - 2p_{q,j})^2 = \sum_q c_q e^{-4\Delta_{q,j}}$. From this formula, it is evident that QTLs located on other chromosomes do not contribute to $\hat{\delta}_j$. Also, the same argument as for the backcross shows that QTLs located left from marker $j - 1$ do not contribute to $\hat{\delta}_j$.

Now consider the subset of all QTLs located between marker $j - 1$ and marker j; their contribution to $\hat{\delta}_j$ is:

$$\sum_q c_q\left(\frac{-a_l^2}{1 - a_l^4} e^{-4\Delta_{q,j-1}} + \frac{1 - a_l^4 a_r^4}{(1 - a_l^4)(1 - a_r^4)} e^{-4\Delta_{q,j}} + \frac{-a_r^2}{1 - a_r^4} e^{-4\Delta_{q,j+1}}\right) = \sum_q c_q e^{-4\Delta_{q,j}} \frac{1 - e^{-8\Delta_{q,j-1}}}{1 - e^{-8\Delta_{j-1,j}}},$$

which can be written $\int_{M_{j-1}}^{M_j} c\psi_j$ and whose first order approximation is: $\sum_q c_q(\Delta_{q,j-1}/\Delta_{j-1,j})$.

C. Variance-covariance matrix of $\tilde{\Theta}$ and its asymptotic structure

Let

$$v_{j,j'}^{-} = E[r_{\text{expi},j} r_{\text{expi},j'}]$$

$$w_{j,j'} = \sum_k \sum_l \theta_k \theta_l E[r_{\text{expi},j} r_{\text{expi},j'} C[r_{i,k} r_{i,l}/M_{\text{obs}}]]$$

be the (j, j') terms of V^{-1} and W, respectively.

Common features: Suppose that the j^{th} marker is missing on individual i ($M_i(j)$ missing). Let ρ_l and ρ_r be the recombination probabilities with the nearest left and right observed markers on the same individual, $M_i(l)$ and $M_i(r)$, and ρ_t the recombination probability between them (special cases where either $M_i(l)$ or $M_i(r)$ do not exist are obtained putting ρ_l or ρ_r and $\rho_t = \frac{1}{2}$ in the formulas, and if both are missing, the individual is not informative for this chromosome, and has to be discarded).

We have the following conditional distribution for $M_i(j)$ given $M_i(l)$ and $M_i(r)$:

missing: $M_i(j)$ / observed: $M_i(l), M_i(r)$	A	B
A, A	$\frac{(1-\rho_l)(1-\rho_r)}{1-\rho_t} = \bar{w}_{j1}$	$\frac{\rho_l\rho_r}{1-\rho_t} = 1 - \bar{w}_{j1}$
A, B	$\frac{(1-\rho_l)\rho_r}{\rho_t} = \bar{w}_{j2}$	$\frac{\rho_l(1-\rho_r)}{\rho_t} = 1 - \bar{w}_{j2}$
B, A	$\frac{\rho_l(1-\rho_r)}{\rho_t} = 1 - \bar{w}_{j2}$	$\frac{(1-\rho_l)\rho_r}{\rho_t} = \bar{w}_{j2}$
B, B	$\frac{\rho_l\rho_r}{1-\rho_t} = 1 - \bar{w}_{j1}$	$\frac{(1-\rho_l)(1-\rho_r)}{1-\rho_t} = \bar{w}_{j1}$.

Now consider two missing markers on the same individual, $M_i(j)$ and $M_i(k)$, their conditional distribution given M_{obs}, depends only on their neighbour (left and right) observed markers. If $M_i(j)$ and $M_i(k)$ belong to different chromosomes, or if they belong to the same chromosome, and if there is an observed marker between them, then they are independent given M_{obs}.

So, we only have to consider the case where the two missing markers $M_i(j)$ and $M_i(k)$ are not separated by any observed marker. Let $M_i(l)$ and $M_i(r)$ be their nearest observed flanking markers, ρ_l, ρ_b, ρ_r, the recombination probabilities between these four markers (from left to right in this order), and ρ_t the recombination probability between $M_i(l)$ and $M_i(r)$ ($\rho_t = \rho_l + \rho_b + \rho_r - 2(\rho_l\rho_b + \rho_b\rho_r + \rho_r\rho_l) + 4\rho_l\rho_b\rho_r$), (special cases where either $M_i(l)$ or $M_i(r)$ do not exist, are obtained putting ρ_l or ρ_r and $\rho_t = \frac{1}{2}$ in the formulas, and if both are missing, the individual is not informative for this chromosome, and has to be discarded).

We have the following joint conditional distribution for $M_i(j)$ and $M_i(k)$ given $M_i(l)$ and $M_i(r)$:

missing: $M_i(j)$, $M_i(k)$ / observed: $M_i(l)$, $M_i(r)$	A, A	A, B	B, A	B, B
A, A	$\frac{(1-\rho_l)(1-\rho_b)(1-\rho_r)}{(1-\rho_t)} = \pi_{11}$	$\frac{(1-\rho_l)(1-\rho_l)\rho_b\rho_r}{(1-\rho_t} = \pi_{12}$	π_{42}	π_{41}
A, B	$\frac{(1-\rho_l)(1-\rho_b)\rho_r}{\rho_t} = \pi_{21}$	$\frac{(1-\rho_l)\rho_b(1-\rho_r)}{\rho_t} = \pi_{22}$	π_{32}	π_{31}
B, A	$\frac{'_l(1-\rho_b)(1-\rho_r)}{\rho_t} = \pi_{31}$	$\frac{\rho_l\rho_b\rho_r}{\rho_t} = \pi_{32}$	π_{22}	τ_{21}
B, B	$\frac{\rho_l(1-\rho_b)\rho_r}{(1-\rho_t)} = \pi_{41}$	$\frac{\rho_l\rho_b(1-\rho_r)}{(1-\rho_t)} = \pi_{42}$	π_{12}	π_{11}

Backcross:

Computation of D:

$$P[M_i(l) = a_l,\ M_i(r) = a_r] = P[M_i(l) = a_l^*,\ M_i(r) = a_r^*],\quad \text{and}$$

$$P[M_i(j) = a_1,\ M_i(k) = a_2/M_i(l) = a_l,\ M_i(r) = a_r] = P[M_i(j) = a_1^*,\ M_i(k) = a_2^*/M_i(l) = a_l^*,\ M_i(r) = a_r^*]$$

where $a \in \{A, B\}$, and $A^* = B$, and $B^* = A$.

Remember that $r_{i,j} = +1$ if $M_i(j)$, equals A, and -1 if it equals B. The two first conditional moments for missing terms are immediately obtained from the preceding tables.

$$E[r_{i,j}/M_i(l)M_i(r) = AA] = -E[r_{i,j}/M_i(l)M_i(r) = BB] = 2\bar{w}_{j1} - 1$$

$$V[r_{i,j}/M_i(l)M_i(r) = AA] = +V[r_{i,j}/M_i(l)M_i(r) = BB] = 4\bar{w}_{j1}(1 - \bar{w}_{j1})$$

$$C[r_{i,j}r_{i,j'}/M_i(l)M_i(r) = AA] = +C[r_{i,j}r_{i,j'}/M_i(l)M_i(r) = BB]$$
$$= (\pi_{11} + \pi_{41}) - (\pi_{12} + \pi_{42}) - (\pi_{11} - \pi_{41})^2 + (\pi_{12} - \pi_{42})^2$$

$$E[r_{i,j}/M_i(l)M_i(r) = AB] = -E[r_{i,j}/M_i(l)M_i(r) = BA] = 2\bar{w}_{j2} - 1$$

$$V[r_{i,j}/M_i(l)M_i(r) = AB] = +V[r_{i,j}/M_i(l)M_i(r) = BA] = 4\bar{w}_{j2}(1 - \bar{w}_{j2})$$

$$C[r_{i,j}r_{i,j'}/M_i(l)M_i(r) = AB] = +C[r_{i,j}r_{i,j'}/M_i(l)M_i(r) = BA]$$
$$= (\pi_{21} + \pi_{31}) - (\pi_{22} + \pi_{32}) - (\pi_{21} - \pi_{31})^2 + (\pi_{22} - \pi_{32})^2$$

Structure of V and W: If j and j' are indexes for missing terms corresponding to different chromosomes, informative markers for j and j' belong to different chromosomes, and are therefore independent. Thus, conditional expectations are independent. $v_{j,j'} = E[r_{\text{expi},j}]E[r_{\text{expi},j'}] = 0$. Also, $C[r_{i,k}r_{i,l}/M_{\text{obs}}] = 0$ if k and l are indexes for terms corresponding to different chromosomes, and $E[r_{\text{expi},j}r_{\text{expi},j'}C[r_{i,k}r_{i,l}/M_{\text{obs}}]]$ can be non zero only if j and j' are indexes for terms corresponding to the same chromosome (as k and k'). Thus V and W do conserve the block-diagonal structure chromosome by chromosome.

F_2:

Computation of D:

$$P[M_i(l) = a_l,\ M_i(r) = a_r] = P[M_i(l) = a_l^*,\ M_i(r) = a_r^*],$$

and

$$P[M_i(j) = a_j,\ M_i(k) = a_k/M_i(l) = a_l,\ M_i(r) = a_r] = P[M_i(j) = a_j^*,\ M_i(k) = a_k^*/M_i(l) = a_l^*,\ M_i(r) = a_r^*]$$

where $a_x \in \{(AA), (AB), (BB)\}$, and $(AA)^* = (BB)$, $(BB)^* = (AA)$, and $(AB)^* = (AB)$.

Remember that if $M_i(j)$, equals AA (respectively AB, BB), $r_{i,j} = +1$ (respectively 0, -1) for an additive effect, and $r_{i,j} = -1$ (respectively $+1$, -1) for a dominance effect ($ra_{i,j}$ and $rd_{i,j}$ stand for $_{i,j}$ terms following they correspond to additive or dominance effects). Therefore

$$E[ra_{i,j}/M_i(l) = a_l,\ M_i(r) = a_r] = -E[ra_{i,j}/M_i(l) = a_l^*,\ M_i(r) = a_r^*]$$

$$V[ra_{i,j}/M_i(l) = a_l,\ M_i(r) = a_r] = +V[ra_{i,j}/M_i(l) = a_l^*,\ M_i(r) = a_r^*]$$

$$C[ra_{i,j}ra_{i,j'}/M_i(l) = a_l,\ M_i(r) = a_r] = +C[ra_{i,j}ra_{i,j'}/M_i(l) = a_l^*,\ M_i(r) = a_r^*]$$

$$E[rd_{i,j}/M_i(l) = a_l,\ M_i(r) = a_r] = +E[rd_{i,j}/M_i(l) = a_l^*,\ M_i(r) = a_r^*]$$

$$V[rd_{i,j}/M_i(l) = a_l,\ M_i(r) = a_r] = +V[rd_{i,j}M_i(l) = a_l^*,\ M_i(r) = a_r^*]$$

$$C[rd_{i,j}rd_{i,j'}/M_i(l) = a_l,\ M_i(r) = a_r] = +C[rd_{i,j}rd_{i,j'}/M_i(l) = a_l^*,\ M_i(r) = a_r^*]$$

$$C[ra_{i,j}rd_{i,j'}/M_i(l) = a_l,\ M_i(r) = a_r] = -C[ra_{i,j}rd_{i,j'}/M_i(l) = a_l^*,\ M_i(r) = a_r^*]$$

$$C[rd_{i,j}ra_{i,j'}/M_i(l) = a_l,\ M_i(r) = a_r] = -C[rd_{i,j}ra_{i,j'}/M_i(l) = a_l^*,\ M_i(r) = a_r^*].$$

As an example of the contents of R_{exp}, the two first conditional moments for missing terms are detailed below for $M_i(l) = AA$, $M_i(r) = AA$.

$$E[ra_{i,j}/M_i(l) = AA,\ M_i(r) = AA] = 2\bar{w}_{j1} - 1$$

$$V[ra_{i,j}/M_i(l) = AA,\ M_i(r) = AA] = 2\bar{w}_{j1}(1 - \bar{w}_{j1})$$

$$C[ra_{i,j}ra_{i,j'}/M_i(l) = AA,\ M_i(r) = AA] = (\pi_{11}^2 + \pi_{41}^2) - (\pi_{12}^2 + \pi_{42}^2) - (2\bar{w}_{j1} - 1)(2\bar{w}_{j'1} - 1)$$

$$E[rd_{i,j}/M_i(l) = AA,\ M_i(r) = AA] = -(2\bar{w}_{j1} - 1)^2$$

$$V[rd_{i,j}/M_i(l) = AA,\ M_i(r) = AA] = 1 - (2\bar{w}_{j1} - 1)^4$$

$$C[rd_{i,j}rd_{i,j'}/M_i(l) = AA,\ M_i(r) = AA] = ((\pi_{11} + \pi_{41}) - (\pi_{12} + \pi_{42}))^2 - (2\bar{w}_{j1} - 1)^2(2\bar{w}_{j'1} - 1)^2$$

$$C[ra_{i,j}rd_{i,j'}/M_i(l) = AA,\ M_i(r) = AA] = (\pi_{41} - \pi_{42})^2 - (\pi_{11} - \pi_{12})^2 + (2\bar{w}_{j1} - 1)(2\bar{w}_{j'1} - 1)^2$$

$$C[rd_{i,j}ra_{i,j'}/M_i(l) = AA,\ M_i(r) = AA] = (\pi_{41} - \pi_{12})^2 - (\pi_{11} - \pi_{42})^2 + (2\bar{w}_{j1} - 1)^2(2\bar{w}_{j'1} - 1).$$

Structure of V and W: If j and j' are indexes for terms corresponding to different chromosomes, the same argument as for the backcross proves that $v_{j,j'}^{-} = 0$ and $w_{j,j'} = 0$. Now, consider j and j', two indexes for terms corresponding to different classes of effects (additive and dominant) on the same chromosome, without observed marker between them. The antisymmetry for an additive effect, the invariance for a dominant effect, of the conditional expectations, and the invariance of the probabilities, through the * transformation, make $v_{j,j'}^{-} = 0$. Also, $C[r_{i,k}r_{i,l}/M_{\text{obs}}]$ is antisymmetric, and $r_{\text{expi},j}r_{\text{expi},j'}$ invariant through the * transformation, therefore, $w_{j,j'} = 0$. Thus, V and W do conserve the block-diagonal structure chromosome by chromosome, and, within each chromosome, between additive and dominance effects.