

A Haplotype-Based Method for QTL Mapping of F₁ Populations in Outbred Plant Species

Cuahtemoc Cervantes-Martinez and J. Steven Brown*

ABSTRACT

The integration of quantitative trait loci (QTL) analysis into breeding strategies rather than being seen as separated processes has been proposed to increase the power and accuracy of QTL detection and to allow the two activities to be joined. The main objective of this research is to develop a specific scheme for mapping QTL in actual breeding F₁ populations of outbred plant species with a high degree of accuracy. The proposed method groups populations by common founders and statistically associates founder-origin probabilities that trace the common founder haplotypes in a given region of the progeny genome with the phenotypic expression, using a linear model with a structured covariance matrix. The method was applied to computer simulated data sets, corresponding to five F₁ populations of 100 individuals each obtained from the crosses of a common founder with several other founders. We are currently using this scheme with cocoa (*Theobroma cacao* L.) crosses, using selected clones resistant to specific diseases to widen the genetic base of disease resistance. The results indicate that the position and effect of QTLs in the common founder, that explain each at least 14% of the phenotypic variance, can be estimated with good precision and accuracy. The theoretical assumptions on which this approach was developed render the method appropriate for outbred plant species that are highly heterozygous, which is often the case in tropical tree crops like cocoa, and have phenotypic traits that show few interlocus interaction effects.

ACCURATE QTL ANALYSES have been developed in recent years to detect and estimate the effects of quantitative trait loci in plant populations with different genetic structures. While high resolution QTL maps can be obtained from large populations of annual plant species developed from crossing inbred lines followed by self-fertilization for two or more generations, the analysis of quantitative trait loci is more difficult in outbred plant species. Some of the difficulties arise when heterozygous heterogeneous parents are crossed to develop a mapping population, in which parents are differentially informative at different loci. To be informative, a parent must be heterozygous both at marker loci and a linked QTL. Complications arise if parents have alleles in common at the QTL or marker loci, or if the parents share QTL alleles in different linkage phases with the marker loci (Jansen et al., 1998; Lynch and Walsh, 1998). In addition, the biological properties of some outbred species, like fruit trees and forest trees, impose limiting factors for mapping QTL. The number of generations per time unit and the progeny size per space unit are usually fewer than in annual species, resulting in lower power for QTL

detection. Luo (1993), Soller and Genizi (1978), and Weller et al. (1990) have confirmed that very large progeny sizes are needed to detect QTL with good statistical power in outbred populations for different designs. However, extremely large progeny sizes and designs requiring two or more generations are not generally feasible in practice for trees and some other outbred plant species. For example, the most recent QTL maps for yield components, vigor, resistance to *Phytophthora palmivora* (E.J. Butler) E.J. Butler, beans traits, and ovule number in *T. cacao* were estimated from F₁ populations ranging from 88 to 125 individuals. These were obtained from the cross of a highly homozygous clone (Catongo) with other heterozygous clones (DR1, S52, and IMC78) (Clement et al., 2003a, 2003b). Therefore, alternative accurate methods must be developed for mapping QTL given the conditions and genetic structure of outbred plant species. Beavis (1998) first proposed the integration of QTL analysis into cultivar development to increase the resolution of QTL detection, by integrating mapping analyses across the numerous and large populations typically used by maize (*Zea mays* L.) breeders.

A haplotypic method for QTL analysis in trees species using founder-origin probabilities that trace specific segments of the chromosomes in individual offspring as independent variables with phenotypic values as the dependent variable in a simple regression analysis has been proposed for one population using the granddaughter design (Reyes-Valdés and Williams, 2002). Their results were similar to those obtained by Haley et al. (1994) that used all marker information. This method requires, however, the information from three generations for QTL detection. In contrast, we suggest an approach that uses founder-origin probabilities in several F₁ populations obtained in a full-sib mating design and combines the F₁ populations with a selected common founder in a regression-based analysis, using a linear model with a structured covariance matrix (Searle, 1971; Littell et al., 1996). Jannink and Jansen (2001) and Jansen et al. (2003), assuming additive effects, showed that combining related breeding populations for QTL analysis increases the power and accuracy of detection, associated mainly with the increased progeny numbers in the combined analysis.

QTL mapping analyses based on linear regression models (Haley and Knott, 1992), such as the one we propose in this study, are approximate methods that generally give results similar to maximum likelihood methods (Lander and Botstein, 1989); however, they are computationally much less demanding and have greater flexibility to implement complex linear models (Piepho, 2000). The ob-

C. Cervantes-Martinez, University of Florida, C/O USDA-ARS, SHRS, 13601 Old Cutler Road, Miami, FL 33158, USA; S. Brown, USDA-ARS, SHRS, 13601 Old Cutler Road, Miami, FL 33158, USA. Received 6 Nov. 2003. *Corresponding author (miaj@ars-grin.gov).

Published in Crop Sci. 44:1572–1583 (2004).
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: QTL, quantitative trait loci; REML, restricted maximum likelihood.

jective of this paper is to explain our method in detail and apply it to the analysis of computer simulated data of five F_1 populations with a common founder in a manner similar to a currently used breeding scheme and structure of cocoa breeding populations (Clement et al., 2003a, 2003b).

METHODS

Haplotypic Conditional Probabilities

The founder-origin probabilities that trace specific haplotype segments of the genome of F_1 individuals to the haplotype of their founders are developed here as an extension of the model for marker-based selection in gene introgression showed by Reyes-Valdés (2000) and Reyes-Valdés and Williams (2002). Consider two diploid founders P_{CF} and P_{SF_i} of a population i (subscripts CF and SF_{*i*} stand for common founder and second founder, respectively) and two informative marker loci, A and B , in the genotypic array for founder P_{CF} : A_1B_1/A_2B_2 and for founder P_{SF_i} : $A_{i3}B_{i3}/A_{i4}B_{i4}$. The chromosome segments A_1B_1 and A_2B_2 are the first and second haplotypes of the founder P_{CF} , and the segments $A_{i3}B_{i3}$ and $A_{i4}B_{i4}$ are the first and second haplotypes of the founder P_{SF_i} . Let H_{1x} and H_{2x} be specific alleles of the locus at the map position x between marker loci A and B , in the first and second haplotype of the founder P_{CF} , and H_{i3x} and H_{i4x} be specific alleles in the first and second haplotype of the founder P_{SF_i} , in the same locus at a map position x , and let d_1 and d_2 be the map positions of marker loci A and B . The absolute map distances between markers is $|d_1 - d_2|$ and the distances between the locus located at position x and A and B are $|d_1 - x|$ and $|d_2 - x|$, respectively. The map distances are converted to recombination fractions using the inverse of the Haldane mapping function (Haldane, 1919) assuming no interference,

$$r = \frac{(1 - e^{-2|d_1-d_2|})}{2}, \quad r_{Ax} = \frac{(1 - e^{-2|d_1-x|})}{2},$$

and

$$r_{Bx} = \frac{(r - r_{Ax})}{(1 - 2r_{Ax})}. \quad [1]$$

The conditional probabilities that F_1 individuals have inherited specific founder alleles in a locus at position x , given that they have specific marker haplotypes, are shown in Table 1. These probabilities are shown when the flanking markers are linked in coupling or repulsion phase in the founders genotypes. To determine the founder-origin probabilities for each particular F_1 individual, the marker haplotypes that trace to either founder genome must be specified for every segment analyzed. Some difficulty arises when there are identical alleles in founders P_{CF} and P_{SF_i} for the marker loci under consideration. If founders P_{CF} and P_{SF_i} share one or two alleles for either locus A or locus B with equal or different linkage phase, only the F_1 individuals with homozygous genotypes for that locus are informative. When the founder P_{SF_i} shares one or two alleles with the founder P_{CF} for both loci, with the same or different linkage phase, only the doubly homozygous F_1 individuals are informative. The linkage phase of markers is estimated from data with linkage analysis methods for full-sib families obtained with the cross of heterozygous parents (Maliepaard et al., 1997; Wu et al., 2002). The haplotypic conditional probabilities are calculated considering only informative F_1 individuals with the equations stated in Table 1, for intervals delimited for informative flanking markers, implying

that the length of the interval and number of informative F_1 individuals may vary among intervals and founder haplotypes analyzed.

Marker loci at the ends of the linkage groups are considered in the analysis whether they are informative or not. If the marker loci at the ends are not informative, then founder-origin probabilities are calculated as follows: the founder-origin probabilities of the H_{1x} allele in the locus at the position x between a non-informative marker extreme and an informative marker locus with genotypes AA and B_1B_2 , respectively, are $\Pr(H_{1x}|B_1) = 1 - r_{Bx}$ and $\Pr(H_{1x}|B_2) = r_{Bx}$. The founder-origin probabilities for the allele H_{i3x} are obtained in an analogous manner, and the conditional probabilities for the alleles H_{2x} and H_{i4x} are the corresponding complementary probabilities. The founder-origin probabilities described above are calculated for all informative individuals in F_1 populations, in every map position between informative flanking markers. The phenotypic data and the founder-origin probabilities are used for the QTL analysis as outlined below.

Linear Model Formulation

Consider a number q of F_1 populations with a common founder (P_{CF}) and a second nonidentical founder for every population (P_{SF_i} ; $i = 1, 2, \dots, q$). Let A and B be marker loci at a given map distance of the linkage group, and x a map position between markers A and B , with founder-origin probabilities $P_{H_{CFx}|AB_{ij}}$ and $P_{H_{SF_i x}|AB_{ij}}$ (Table 1) that an F_1 individual j from population i with marker haplotype AB , has inherited specific founder alleles from P_{CF} and P_{SF_i} , respectively, in a locus at position x . A basic linear model can be fit as follows:

$$y_{ij} = \mu_i + \alpha_{H_{CFx}} P_{H_{CFx}|AB_{ij}} + \alpha_{H_{CFx}} (1 - P_{H_{CFx}|AB_{ij}}) + \alpha_{H_{SF_i x}} P_{H_{SF_i x}|AB_{ij}} + \alpha_{H_{SF_i x}} (1 - P_{H_{SF_i x}|AB_{ij}}) + \varepsilon_{ij};$$

$$i = 1, 2, \dots, q; j = 1, 2, \dots, n_i, \quad [2]$$

where y_{ij} refers to the phenotypic values of individual j in population i ; μ is the mean of population i ; $\alpha_{H_{CFx}}$ and $\alpha_{H_{SF_i x}}$ are the parameters corresponding to the fixed effects of the allele at the first homologs (Jansen et al., 1998; Lynch and Walsh, 1998) of the founders P_{CF} and P_{SF_i} for the putative QTL in a locus at position x in population i ; $\alpha_{H_{CFx}}$ and $\alpha_{H_{SF_i x}}$ are the parameters corresponding to the fixed effects of the allele at the second homologs of the founders P_{CF} and P_{SF_i} ; and ε_{ij} is a random variable identically distributed with mean zero and variance σ^2_{ε} , that includes the background effect, the environmental variation, and the inadequacy of the model. The model [2] can be reparameterized as

$$y_{ij} = \mu_i^* + \alpha_{H_{CFx}}^* P_{H_{CFx}|AB_{ij}} + \alpha_{H_{SF_i x}}^* P_{H_{SF_i x}|AB_{ij}} + \varepsilon_{ij}. \quad [3]$$

Here, $\mu_i^* = \mu_i + \alpha_{H_{CFx}} + \alpha_{H_{SF_i x}}$; $\alpha_{H_{CFx}}^* = \alpha_{H_{CFx}} - \alpha_{H_{CFx}}$ and $\alpha_{H_{SF_i x}}^* = \alpha_{H_{SF_i x}} - \alpha_{H_{SF_i x}}$ are the allele-substitution fixed effects (Jansen et al., 1998) of QTL alleles of founders P_{CF} and P_{SF_i} in population i , respectively. A second model is also formulated to include the intralocus interaction among QTL alleles at the QTL loci, adding the dominance effects to the additive effects model [2], so that

$$y_{ij} = \mu_i + \alpha_{H_{CFx}} P_{H_{CFx}|AB_{ij}} + \alpha_{H_{CFx}} (1 - P_{H_{CFx}|AB_{ij}}) + \alpha_{H_{SF_i x}} P_{H_{SF_i x}|AB_{ij}} + \alpha_{H_{SF_i x}} (1 - P_{H_{SF_i x}|AB_{ij}}) + \delta_{H_{CF}H_{SF_i x}} P_{H_{CFx}|AB_{ij}} P_{H_{SF_i x}|AB_{ij}} + \delta_{H_{CF}H_{SF_i x}} P_{H_{CFx}|AB_{ij}} (1 - P_{H_{SF_i x}|AB_{ij}}) + \delta_{H_{CF}H_{SF_i x}} (1 - P_{H_{CFx}|AB_{ij}}) P_{H_{SF_i x}|AB_{ij}} + \delta_{H_{CF}H_{SF_i x}} (1 - P_{H_{CFx}|AB_{ij}}) (1 - P_{H_{SF_i x}|AB_{ij}}) + \varepsilon_{ij}, \quad [4]$$

Table 1. Founder-origin probabilities for each possible haplotypic state of F₁ progeny.

Founder <i>P_{CF}</i>	F ₁ progeny		Founder <i>P_{SF_i}</i>	F ₁ progeny	
	Marker haplotype from <i>P_{CF}</i>	$\Pr(H_{1x} A_kB_l)$		Marker haplotype from <i>P_{SF_i}</i>	$\Pr(H_{1x} A_{ik}B_{il})$
<i>A₁B₁/A₂B₂</i>	<i>A₁B₁</i>	$\frac{(1 - r_{Ax_{CF}})(1 - r_{Bx_{CF}})}{1 - r}$	<i>A₃B₃/A₄B₄</i>	<i>A₃B₃</i>	$\frac{(1 - r_{Ax_{SF_i}})(1 - r_{Bx_{SF_i}})}{1 - r}$
	<i>A₁B₂</i>	$\frac{(1 - r_{Ax_{CF}})r_{Bx_{CF}}}{r}$		<i>A₃B₄</i>	$\frac{(1 - r_{Ax_{SF_i}})r_{Bx_{SF_i}}}{r}$
	<i>A₂B₁</i>	$\frac{r_{Ax_{CF}}(1 - r_{Bx_{CF}})}{r}$		<i>A₄B₃</i>	$\frac{r_{Ax_{SF_i}}(1 - r_{Bx_{SF_i}})}{r}$
	<i>A₂B₃</i>	$\frac{r_{Ax_{CF}} r_{Bx_{CF}}}{1 - r}$		<i>A₄B₄</i>	$\frac{r_{Ax_{SF_i}} r_{Bx_{SF_i}}}{1 - r}$
<i>A₁B₂/A₂B₁</i>	<i>A₁B₁</i>	$\frac{r_{Ax_{CF}}(1 - r_{Bx_{CF}})}{r}$	<i>A₃B₄/A₄B₃</i>	<i>A₃B₃</i>	$\frac{r_{Ax_{SF_i}}(1 - r_{Bx_{SF_i}})}{r}$
	<i>A₁B₂</i>	$\frac{r_{Ax_{CF}} r_{Bx_{CF}}}{1 - r}$		<i>A₃B₄</i>	$\frac{r_{Ax_{SF_i}} r_{Bx_{SF_i}}}{1 - r}$
	<i>A₂B₁</i>	$\frac{(1 - r_{Ax_{CF}})(1 - r_{Bx_{CF}})}{1 - r}$		<i>A₄B₃</i>	$\frac{(1 - r_{Ax_{SF_i}})(1 - r_{Bx_{SF_i}})}{1 - r}$
	<i>A₂B₂</i>	$\frac{(1 - r_{Ax_{CF}})r_{Bx_{CF}}}{r}$		<i>A₄B₄</i>	$\frac{(1 - r_{Ax_{SF_i}})r_{Bx_{SF_i}}}{r}$

†, ‡ Parental common founder and second founder, respectively. *H_{1x}* and *H_{2x}* are the alleles in the first and second haplotype of the founder *P_{CF}*; *H_{1x}* and *H_{2x}* are the alleles in the first and second haplotype of the founder *P_{SF_i}*, in the locus at the map position *x*; *r*, *r_{Ax}* and *r_{Bx}* are the recombination fractions between two informative markers A and B, and between the locus at map position *x* and markers A and B. The conditional probabilities for the alleles *H_{1x}* and *H_{2x}* are given by $\Pr(H_{2x}|A_kB_l) = 1 - \Pr(H_{1x}|A_kB_l)$ and $\Pr(H_{1x}|A_{ik}B_{il}) = 1 - \Pr(H_{2x}|A_{ik}B_{il})$.

where the coefficients $\delta_{H_{CF},H_{SF_i},x}$, $\delta_{H_{CF},H_{SF_i},x}$, $\delta_{H_{CF},H_{SF_i},x}$ and $\delta_{H_{CF},H_{SF_i},x}$ represent the fixed dominance effects between the QTL alleles of the locus at position *x* of founders *P_{CF}* and *P_{SF_i}*. A reduction is achieved by setting the restriction on the dominance effects $\delta_i = \delta_{H_{CF},H_{SF_i},x} = -\delta_{H_{CF},H_{SF_i},x} = -\delta_{H_{CF},H_{SF_i},x} = \delta_{H_{CF},H_{SF_i},x}$, and using the allele-substitution effects in [4], resulting in

$$y_{ij} = \mu_i^* + \alpha_{H_{CF},x}^* P_{H_{CF},x|AB_{ij}} + \alpha_{H_{SF_i},x}^* P_{H_{SF_i},x|AB_{ij}} + \delta_i (2P_{H_{CF},x|AB_{ij}} - 1)(2P_{H_{SF_i},x|AB_{ij}} - 1) + \epsilon_{ij}. \quad [5]$$

Here, δ_i is dominance effect in population *i*.

The details of reparameterization of models [2] and [4] are shown in the APPENDIX. The models [3] and [5] are referred to here as the additive effects model, and the additive and dominance effects model, respectively.

Analyses of linear regression (Reyes-Valdés and Williams, 2002) can be implemented to estimate the models [3] and [5] for every F₁ population, where μ_i^* is the intercept, and $\alpha_{H_{CF},x}^*$, $\alpha_{H_{SF_i},x}^*$ and δ_i are the regression coefficients. The additive effects, and the additive and dominance effects linear models for the single population analyses are written by convenience in matrix notation as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \mathbf{e}_i = [\mathbf{X}_i \quad \mathbf{Z}_i] \begin{bmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{bmatrix} + \mathbf{e}_i = \mathbf{X}_i \mathbf{a}_i + \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_i; \quad i = 1, \dots, q, \quad [6]$$

where $\mathbf{y}_i^T = [y_{i1} \ y_{i2} \ \dots \ y_{in_i}]$ is the vector of phenotypic observations of the population *i*, *n_i* is the number of individuals in the population *i*; $\mathbf{X}_i = \mathbf{1}_i$ as a vector *n_i* × 1 of elements ones; $\mathbf{a}_i = [\mu_i^*]$ is the intercept for the population *i*; $\mathbf{Z}_i = [\mathbf{P}_{H_{CF},x|AB_i}, \mathbf{P}_{H_{SF_i},x|AB_i}]$ is the founder-origin probability matrix for the additive model for the population *i*, $\mathbf{P}_{H_{CF},x|AB_i} = [P_{H_{CF},x|AB_{i1}} \ P_{H_{CF},x|AB_{i2}} \ \dots$

$P_{H_{CF},x|AB_{in_i}}]$ is the vector containing the haplotypic conditional probabilities for the population *i* corresponding to the common founder *P_{CF}*, and $\mathbf{P}_{H_{SF_i},x|AB_i} = [P_{H_{SF_i},x|AB_{i1}} \ P_{H_{SF_i},x|AB_{i2}} \ \dots \ P_{H_{SF_i},x|AB_{in_i}}]$ is the vector containing the haplotypic conditional probabilities corresponding to the founder *P_{SF_i}* of the population *i*; $\mathbf{Z}_i = [\mathbf{P}_{H_{CF},x|AB_i}, \mathbf{P}_{H_{SF_i},x|AB_i}, \boldsymbol{\gamma}_i]$ is the founder-origin probability matrix for the additive and dominance effects model for the population *i*, $\boldsymbol{\gamma}_i^T = [\gamma_{i1} \ \gamma_{i2} \ \dots \ \gamma_{in_i}]$, $\gamma_{ij} = (2P_{H_{CF},x|AB_{ij}} - 1)(2P_{H_{SF_i},x|AB_{ij}} - 1)$; $\mathbf{b}_i^T = [\alpha_{H_{CF},x}^* \ \alpha_{H_{SF_i},x}^*]$ is the vector of the fixed parameters for the additive effects model, $\alpha_{H_{CF},x}^*$ is the allele-substitution parameter of the putative QTL alleles in the common founder *P_{CF}* and $\alpha_{H_{SF_i},x}^*$ is the allele-substitution parameter of the QTL alleles corresponding to the founder *P_{SF_i}* from population *i*; $\mathbf{h}_i^T = [\alpha_{H_{CF},x}^* \ \alpha_{H_{SF_i},x}^* \ \delta_i]$ is the vector of the fixed parameters for the additive and dominance effects model, δ_i is the dominance effect for population *i*; $\mathbf{e}_i^T = [\epsilon_{i1} \ \epsilon_{i2} \ \dots \ \epsilon_{in_i}]$ is the vector of random deviations for population *i*. The random vector \mathbf{e}_i is assumed to be normally distributed with $E(\mathbf{e}_i) = \mathbf{0}$ and $Var(\mathbf{e}_i) = \mathbf{R}_i = \mathbf{I}_i \sigma_i^2$. Here, \mathbf{I}_i denotes the identity matrix of order *n_i* and σ_i^2 is the component of the residual variance corresponding to population *i*. Alternatively, all populations can be analyzed simultaneously using a covariance model with a structured residual covariance matrix (Searle, 1971; Littell et al., 1996). The linear model for the combined analysis is represented in matrix notation by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = [\mathbf{X}_1 \quad \mathbf{Z}] \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} + \mathbf{e} = \mathbf{X}_1 \mathbf{a} + \mathbf{Z}\mathbf{b} + \mathbf{e}. \quad [7]$$

Where $\mathbf{y}^T = [\mathbf{y}_1^T \ \mathbf{y}_2^T \ \dots \ \mathbf{y}_q^T]$ is the vector of phenotypic observations of all populations; $\mathbf{X}_1 = \bigoplus_{i=1}^q \mathbf{X}_i$, \bigoplus denotes the matrix

direct sum; $\mathbf{a}^T = [\mathbf{a}_1^T \mathbf{a}_2^T \cdots \mathbf{a}_q^T]$ is the intercept vector; $\mathbf{Z} = [\mathbf{P}_{\text{H}_{CF}/\text{AB}} \oplus \mathbf{P}_{\text{H}_{SF}/\text{AB}_i}]$ is the founder-origin probability matrix for the additive effects model, with $\mathbf{P}_{\text{H}_{CF}/\text{AB}} = [\mathbf{P}_{\text{H}_{CF}/\text{AB}_1} \mathbf{P}_{\text{H}_{CF}/\text{AB}_2} \cdots \mathbf{P}_{\text{H}_{CF}/\text{AB}_q}]$; $\mathbf{Z} = [\mathbf{P}_{\text{H}_{CF}/\text{AB}_i} \oplus \mathbf{P}_{\text{H}_{SF}/\text{AB}_i} \oplus \mathbf{D}_i]$ is the founder-origin probability matrix for the additive and dominance effects model; $\mathbf{b}^T = [\alpha_{\text{H}_{CF}^x}^* \alpha_{\text{H}_{SF}^x}^* \cdots \alpha_{\text{H}_{SF}^x}^* \delta_1 \delta_2 \cdots \delta_q]$ is the vector of the fixed covariate parameters for the additive effects model, $\alpha_{\text{H}_{CF}^x}^*$ is the allele-substitution parameter of the putative QTL alleles in the common founder P_{CF} , and $\alpha_{\text{H}_{SF}^x}^*$ is the allele-substitution parameter of the QTL alleles corresponding to the founder P_{SF} of population i ; $\mathbf{b}^T = [\alpha_{\text{H}_{CF}^x}^* \alpha_{\text{H}_{SF}^x}^* \cdots \alpha_{\text{H}_{SF}^x}^* \delta_1 \delta_2 \cdots \delta_q]$ is the fixed covariate parameter vector for the additive and dominance effects model, δ_i is the dominance parameter for population i ; and $\mathbf{e}^T = [\mathbf{e}_1^T \mathbf{e}_2^T \cdots \mathbf{e}_q^T]$ is the vector of random deviations. The random vector \mathbf{e} is also assumed to be normally distributed with $E(\mathbf{e}) = \mathbf{0}$ and $\text{Var}(\mathbf{e}) = \text{Var}(\mathbf{y}) = \mathbf{R} = \bigoplus_{i=1}^q \mathbf{I}_i \sigma_i^2$, given that the vector \mathbf{b} only contains fixed-effect parameters.

It is assumed that the genetic background effects absorbed by the residual component of the model are independent among individuals within populations in model [6], and among populations in model [7]. This assumption might be unrealistic because individuals within populations are full-sibs and among populations are half-sibs. To control part of the genetic background by reducing the segregation variance generated by linked and unlinked QTLs, when the analysis is performed for a given position in the linkage map, appropriate markers outside of the interval analyzed can be fitted as cofactors in models [3] and [5]. The addition of marker cofactors to partially remove the background genetic effect has shown to increase the sensitivity and precision of QTL mapping (Jansen and Stam, 1994; Zeng, 1994). Since the number of observations in the combined analysis differs from the number of observations in the single population analyses, the markers associated with the interval analyzed for a given linkage group may differ between the combined analysis and the single population analyses. Therefore, the selection of cofactors sets should be done separately for each single population analysis, and for the combined analysis.

Cofactors are included in models [3] and [5] by adding the term

$$\sum_{k_1 \notin I} (v_{\text{H}_{CF}k_1} C_{\text{H}_{CF}k_1j}) + \sum_{k_2 \notin I} (v_{\text{H}_{SF}k_2} C_{\text{H}_{SF}k_2j}), \quad [8]$$

where I is the interval of the linkage group analyzed and delimited by two fully informative markers loci (A and B); $C_{\text{H}_{CF}k_1j}$ and $C_{\text{H}_{SF}k_2j}$ are the known coefficients for the k_1 th and k_2 th markers selected as cofactors of the common founder and second founder of individual j from population i , taking the value of 1 or 0 depending on the markers haplotype; $v_{\text{H}_{CF}k_1}$ and $v_{\text{H}_{SF}k_2}$ are the associated regression coefficients. The model in matrix notation [6] is modified to include the cofactors by redefining the following matrices

$$\mathbf{X}_{li} = [\mathbf{1}_i \mathbf{C}_{\text{H}_{CF}l_1} \mathbf{C}_{\text{H}_{CF}l_2} \cdots \mathbf{C}_{\text{H}_{CF}m_1} \mathbf{C}_{\text{H}_{SF}l_1} \mathbf{C}_{\text{H}_{SF}l_2} \cdots \mathbf{C}_{\text{H}_{SF}m_2}]$$

$$\mathbf{a}_i^T = [\mu_i^* v_{\text{H}_{CF}l_1} v_{\text{H}_{CF}l_2} \cdots v_{\text{H}_{CF}m_1} v_{\text{H}_{SF}l_1} v_{\text{H}_{SF}l_2} \cdots v_{\text{H}_{SF}m_2}]$$

Where $\mathbf{C}_{\text{H}_{CF}k_1}$ and $\mathbf{C}_{\text{H}_{SF}k_2}$ are $n_i \times 1$ vectors of known coefficients of the k_1 th and k_2 th common founder and second founder cofactors; m_1 and m_2 are number of marker cofactors considered from common founder and second founder haplotypes from population i . The model in [7] is modified by redefining the following matrices as

$$\mathbf{X}_{li} = [\mathbf{1}_i \mathbf{C}_{\text{H}_{SF}l_1} \mathbf{C}_{\text{H}_{SF}l_2} \cdots \mathbf{C}_{\text{H}_{SF}m_2}]$$

$$\mathbf{X}_1 = [\mathbf{C}_{\text{H}_{CF}1} \mathbf{C}_{\text{H}_{CF}2} \cdots \mathbf{C}_{\text{H}_{CF}m_1} \oplus_{i=1}^q \mathbf{X}_{li}]$$

$$\mathbf{a}_i^T = [\mu_i^* v_{\text{H}_{SF}l_1} v_{\text{H}_{SF}l_2} \cdots v_{\text{H}_{SF}m_2}]$$

$$\mathbf{a}^T = [v_{\text{H}_{CF}1} v_{\text{H}_{CF}2} \cdots v_{\text{H}_{CF}m_1} \mathbf{a}_1^T \mathbf{a}_2^T \cdots \mathbf{a}_q^T]$$

In this case, $\mathbf{C}_{\text{H}_{CF}k_1}$ is a $\sum_{i=1}^q n_i \times 1$ vector of known coefficients of the k_1 th common founder cofactor, and m_1 is the number of marker cofactors in the common founder haplotype considered in the combined model.

The estimation of the variance components for both single population QTL analyses and combined analysis can be performed by restricted maximum likelihood (REML) using a ridge-stabilized Newton-Raphson algorithm, which given conditions of regularity of the likelihood function and adequate starting values, produces a quadratic convergence. The best linear unbiased estimators of the fixed effects parameters are obtained by solving the mixed model equations (Searle et al., 1992; Littell et al., 1996). The analyses are performed at each 1 cM position in the linkage group, and the likelihood ratio statistic (LR) is calculated by obtaining the difference between the -2 times the REML log likelihood of the reduced model with no QTL consideration (l_0) and the full model with the QTL parameters (l_1). Full models are represented by [3], [5], [6], and [7]. The reduced models only include the parameter μ_i^* and the random deviation ϵ_{ij} for the population analyses, and the vectors $\mathbf{X}_i \mathbf{a}$ and \mathbf{e} for the combined analysis. The REML log likelihood function for the full model is described as follows

$$l_1 = \frac{n-p}{2} \log(2\pi) - \frac{1}{2} \log|\mathbf{R}_i| - \frac{1}{2} \log|\mathbf{X}_i^T \mathbf{R}_i^{-1} \mathbf{X}_i| - \frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_i)^T \mathbf{R}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_i), \quad [9]$$

for the single population analyses, where $p = 3 + m_{1i} + m_{2i}$ in model [3] and $p = 4 + m_{1i} + m_{2i}$ in model [5]. If no cofactors are used, then $m_{1i} = m_{2i} = 0$. The matrices \mathbf{X}_i and \mathbf{R}_i , and the vector \mathbf{b}_i are established for the model [6]. Likewise, the REML log likelihood function for the combined analysis is represented by

$$l_1 = \frac{n-p}{2} \log(2\pi) - \frac{1}{2} \log|\mathbf{R}| - \frac{1}{2} \log|\mathbf{X}^T \mathbf{R}^{-1} \mathbf{X}| - \frac{1}{2} (\mathbf{y} - \mathbf{X} \mathbf{b})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X} \mathbf{b}). \quad [10]$$

The matrices \mathbf{X} and \mathbf{R} , and the vector \mathbf{b} correspond to the model [7]. In this case, $p = m_1 + \sum_{i=1}^q m_{2i} + 2q + 1$ for the additive effects model, and $p = m_1 + \sum_{i=1}^q m_{2i} + 3q + 1$ for the additive and dominance effects model. Note that $m_1 + \sum_{i=1}^q m_{2i} = 0$ when cofactors are not considered in the model.

The REML log likelihood functions for the reduced models (l_0) for the single population analyses and the combined analysis are obtained substituting \mathbf{X}_i by \mathbf{X}_{li} and \mathbf{b}_i by \mathbf{a}_i in [9], and \mathbf{X} by \mathbf{X}_1 and \mathbf{b} by \mathbf{a} in [10], with $p = m_{1i} + m_{2i} + 1$ in [9] and $p = m_1 + \sum_{i=1}^q m_{2i} + q$ in [10]. -2 times the REML log likelihood functions are evaluated with the values of the fixed effects vector and covariance matrix that maximize [9] and [10] given by the Newton-Raphson algorithm; for example, the REML esti-

mates of \mathbf{R} and the generalized least squares estimates (GLS) of $\mathbf{\beta}$ in (10), which are denoted $\hat{\mathbf{R}}$ and $\hat{\mathbf{\beta}} = (\mathbf{X}^T \hat{\mathbf{R}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{R}}^{-1} \mathbf{y}$, respectively. The estimated variance-covariance matrices of the GLS estimate of $\mathbf{\beta}_i$ and $\mathbf{\beta}$ are $(\mathbf{X}_i^T \hat{\mathbf{R}}_i^{-1} \mathbf{X}_i)^{-1}$ and $(\mathbf{X}^T \hat{\mathbf{R}}^{-1} \mathbf{X})^{-1}$ (Searle, 1971; Searle et al., 1992; Littell et al., 1996).

The significance of the putative QTL can be obtained by the approximation of the likelihood ratio test to the χ^2 distribution (Self and Liang, 1987). The approximated thresholds are $\chi_{\alpha/M, 2q}^2$ and $\chi_{\alpha/M, 3q}^2$ for the additive effects model and additive and dominance effects model, respectively, where q is the number of populations ($q = 1$ for single population analyses) and M is the number of intervals in the genome. The overall significance level of α/M is discussed by Zeng (1994). The use of empirical thresholds based on the permutation test would be a more robust alternative (Churchill and Doerge, 1994), however, computationally more demanding.

Genetic Considerations

Let us assume first an F_1 population developed from the cross of the founder clones, P_{CF} and P_{SF_i} , a putative QTL in the locus at position x in the linkage group, with alleles H_{1x} and H_{2x} in the founder P_{CF} , and the H_{13x} and H_{14x} alleles in the founder P_{SF_i} . The regression coefficients of the phenotype on H_{1x} and H_{13x} founder-origin probabilities of the locus at position x for F_1 individuals in model [3] and [5], represent the fixed allele-substitution effect of H_{1x} by H_{2x} and H_{13x} by H_{14x} .

Since not all possible QTL allele combinations are obtained in the heterozygous progeny in F_1 populations from crossing QTL informative clones, there are some limitations in estimating dominance deviations. In fact, dominance can be estimated only by a lack of parallelism between the phenotypic values of the genotypic pairs H_1H_{13x} , H_2H_{13x} , and H_1H_{14x} , H_2H_{14x} . Since only these four genotypes are available for estimating both additive and dominance genetic effects, with three degrees of freedom, two independent parameters are used to estimate the allele-substitution effects ($\alpha_{H_{CF},x}^*$ and $\alpha_{H_{SF_i},x}^*$), leaving one degree of freedom for estimating a dominance parameter. Given this constraint, we must impose a restriction on the dominance effect such that $\delta_i = \delta_{H_{CF}H_{SF_i},x} = -\delta_{H_{CF}H_{SF_i},x} = -\delta_{H_{CF}H_{SF_i},x} = \delta_{H_{CF}H_{SF_i},x}$. A value $\delta_i \neq 0$ would indicate a lack of parallelism and imply the existence of dominance among QTL alleles at a locus. Conversely, however, a value of $\delta_i = 0$ would not definitively exclude the existence of dominance, as the dominance effect could be affecting the phenotype equally at all four genotypes, in which case it is not detected.

This principle can be extended to several F_1 populations that share a common founder. Such scenario is seen in the context of breeding populations of fruit trees species, like *T. cacao*, in which the common founder is a selected clone with some very desirable traits, but also with undesirable genetic

constitution for other traits. This clone, therefore, it is crossed to other selected clones with good complementary responses to other characteristics for further selection of the superior recombinants in the F_1 populations.

Theoretical calculations as well as computer simulation research have shown that under conditions of regularity, high resolution QTL mapping is dependant on large progeny sizes. The power of QTL detection (the probability of a true marker-trait association) is improved by decreasing the within-marker class variance, or residual variance, which come with increased sample size (Soller and Genizi, 1978; Lander and Botstein, 1989; Weller et al., 1990; Lynch and Walsh, 1998). In the most general view, the q F_1 populations can be considered a set of $n = n_1 + n_2 + \dots + n_q$ individuals containing the common founder P_{CF} haplotype, establishing a very suitable scenario in which to implement a haplotype-based approach for QTL analysis. The power of QTL detection in a combined analysis that includes all populations is expected to increase in a manner proportional to the number of populations included in the analysis, and hence increase in sample size, resulting in more accurate QTL maps. When the intralocus interaction (dominance) with the alleles of the second parent of every population is incorporated into the model as showed above, however, an additional assumption of independence from interlocus QTL allele effects (epistasis) must be made. Departure from this assumption might result in lower power of QTL detection and biased estimates of the allele-substitution and dominance parameters. Another consideration is that if the common founder is not QTL informative (QTL heterozygote), then $\alpha_{H_{CF},x} = \alpha_{H_{CF},x}$ and $\alpha_{H_{CF},x}^* = \alpha_{H_{CF},x} - \alpha_{H_{CF},x} = 0$ in the single population analysis and in the combined analysis, meaning that the utility of the proposed method relies on the assumption that the common founder is, indeed, heterozygous for the putative QTL.

Data Simulation and Statistical Analysis

Four data sets of five F_1 populations of 100 individuals obtained from the crosses of one heterozygous common founder clone (P_{CF}) with other five founder clones (P_{SF_1} , P_{SF_2} , P_{SF_3} , P_{SF_4} , and P_{SF_5}) were simulated with a SAS macro for Windows Version 9.0 (SAS Institute Inc., Cary, NC). The genome of every individual consisted of two chromosomes of 100 cM in length with one marker locus every 5 cM, and two QTLs, the first located at a position 59 cM distal from the beginning of the first chromosome and a the second located at a position 29 cM distal from the beginning of second chromosome. The genomes of the parental clones were simulated considering different percentages of homozygosity (20, 30, 35, 40, 45, and 50% for founders P_{CF} , P_{SF_1} , P_{SF_2} , P_{SF_3} , P_{SF_4} , and P_{SF_5} , respectively), represented by non-informative marker loci located randomly

Table 2. Genotypic values for the F_1 progeny of population i , with founders P_{SF_i} and P_{SF_i} ($i = 1, 2, 3, 4, 5$).

	Genotype			
	H_1H_{13x}	H_1H_{14x}	H_2H_{13x}	H_2H_{14x}
Expected frequencies	1/4	1/4	1/4	1/4
Genotypic value	$G_{1,i3} = \mu_i + \alpha_{H_{CF},x} + \alpha_{H_{SF_i},x} + \delta_i$	$G_{1,i4} = \mu_i + \alpha_{H_{CF},x} + \alpha_{H_{SF_i},x} - \delta_i$	$G_{2,i3} = \mu_i + \alpha_{H_{CF},x} + \alpha_{H_{SF_i},x} - \delta_i$	$G_{2,i4} = \mu_i + \alpha_{H_{CF},x} + \alpha_{H_{SF_i},x} + \delta_i$
Genotypic value† for simulation	$G_{1,i3} = \mu_i^* + \alpha_{H_{CF},x}^* + \alpha_{H_{SF_i},x}^* + \delta_i$	$G_{1,i4} = \mu_i^* + \alpha_{H_{CF},x}^* - \delta_i$	$G_{2,i3} = \mu_i^* + \alpha_{H_{SF_i},x}^* - \delta_i$	$G_{2,i4} = \mu_i^* + \delta_i$

† Reparameterized genotypic value. H_{1x} , H_{2x} are the alleles of the putative QTL of the common founder P_{CF} in a locus at position x , with effect $\alpha_{H_{CF},x}$ and $\alpha_{H_{CF},x}$, respectively, and H_{13x} and H_{14x} are the alleles of the second founder P_{SF_i} , with effects $\alpha_{H_{SF_i},x}$ and $\alpha_{H_{SF_i},x}$. The allele-substitution parameters meet the constrain $\alpha_{H_{CF},x}^* = \alpha_{H_{CF},x} - \alpha_{H_{CF},x}$ and $\alpha_{H_{SF_i},x}^* = \alpha_{H_{SF_i},x} - \alpha_{H_{SF_i},x}$. The reparameterized mean meets the constrain $\mu_i^* = \mu_i + \alpha_{H_{CF},x} + \alpha_{H_{SF_i},x}$.

through the chromosomes. A frequency of 5% of allele sharing between founder clones for different loci was also included. The recombination probabilities between homologs were obtained under the assumption of no interference among marker loci, and selecting the location in the chromosome at random for each recombination. The genotypic model for simulation corresponding to population i is described in the Table 2.

The four phenotypic data sets were simulated with additive QTL effects in both chromosomes, but with QTL dominance effects only for the first chromosome (Table 3), with very specific restrictions on the genetic parameter values, explained below. The allele-substitution coefficients for the first chromosome were set to

$$\begin{aligned} \alpha_{H_{SF_1}^*}^* &= \frac{1}{2}\alpha_{H_{CF^*}}^*, \alpha_{H_{SF_2}^*}^* = \frac{1}{3}\alpha_{H_{CF^*}}^*, \alpha_{H_{SF_3}^*}^* = \frac{1}{4}\alpha_{H_{CF^*}}^*, \\ \alpha_{H_{SF_4}^*}^* &= \frac{1}{5}\alpha_{H_{CF^*}}^* \text{ and } \alpha_{H_{SF_5}^*}^* = 0. \end{aligned} \quad [11]$$

Here, $\alpha_{H_{CF^*}}^*$ and $\alpha_{H_{SF_i}^*}^*$ refer to the allele-substitution of the putative QTL in the first chromosome of the common founder and the second founder of population i ($i = 1, 2, 3, 4, 5$). The dominance effects used for simulation were

$$\begin{aligned} \delta_1 &= \frac{1}{4}\alpha_{H_{CF^*}}^*, \delta_2 = \frac{1}{4}\alpha_{H_{CF^*}}^*, \delta_3 = \frac{1}{4}\alpha_{H_{CF^*}}^*, \\ \delta_4 &= \frac{1}{4}\alpha_{H_{CF^*}}^* \text{ and } \delta_5 = 0, \end{aligned} \quad [12]$$

for population 1 to 5, respectively. The restrictions on the allele-substitution effects for the second chromosome were

$$\alpha_{H_{CF^*}}^* = 0, \alpha_{H_{SF_1}^*}^* = 0, \alpha_{H_{SF_2}^*}^* = \alpha_{H_{SF_3}^*}^* = \alpha_{H_{SF_4}^*}^*$$

and

$$\alpha_{H_{SF_5}^*}^* = 0. \quad [13]$$

The residual component was obtained as a random observation from a normal distribution with mean zero and variance set to meet the restrictions in Table 3. The statistical analysis was performed with a SAS macro for Windows Version 9.0. The macro estimates the reduced and full models (6) and (7), the likelihood ratio test statistic, the covariance parameters and their standard error with the MIXED procedure (Littell

et al., 1996) at every 1 cM position of the linkage group. The analyses were performed for the additive model and for the additive and dominance model, the latter both with and without cofactors. The significance of the allele-substitution for the different founders and the dominance effects were tested with a t test. As candidate cofactors were considered all markers with the exception of the flanking markers of the interval to be analyzed for the putative QTL. Cofactors were selected for models (6) and (7) separately by multiple regression using the backward method ($\alpha = 0.05$).

RESULTS

Plots of the likelihood ratio (LR) test statistic against the chromosomal position obtained with the additive and dominance effects model are shown in Fig. 1. Analyses of the first simulated data set with QTLs of the smallest effects (Table 3) did not show particularly satisfactory results in terms of the QTL position estimates. However, analyses performed on the data sets of the latter three simulations with QTLs of larger magnitude showed that the QTL position can be estimated with precision by this approach, as described next.

The LR test statistic was larger when this method, based on founder-origin probabilities, was run on the combined population analysis than when it was based on single population analyses. This was expected since the difference in the number of parameters between the full and reduced models is not the same for the two types of analysis. The extra number of parameters fitted in the full model represents the expected value of the LR test statistic according with its asymptotic χ^2 distribution (Self and Liang, 1987). Therefore, we compared the test statistic of both methods to χ^2 thresholds, defined by the significance level and the degrees of freedom determined by the extra number of parameters fitted in the full model. Thresholds using the χ^2 approximation are $\chi_{0.05/40,3}^2 = \chi_{0.00125,3}^2 \approx 15.8$ and $\chi_{0.05/40,15}^2 = \chi_{0.00125,15}^2 \approx 37.04$ for the single population analyses and the combined analyses, respectively. The likelihood ratio had larger values in chromosome 1 than in chromosome 2 over the chromosome segment containing the putative QTL, as expected, since chromosome 1 has the larger QTL. Analyses based on single populations showed larger values of the test statistic than the respective threshold when performed in the chromosome segment surrounding the putative QTL. The likelihood ratio obtained with the combined analyses was larger than the threshold for the most of chromosome 1. The absolute maximum of the test statistic in the combined analyses was generally in a segment very close to the true QTL position in the linkage group, while the absolute maxima of the likelihood ratio test statistic were slightly distant from this position for the individual population analyses.

Likelihood ratio plots from analyses using the additive effects model only (results not shown) were very similar to the plots in Fig. 1, but with slightly lower values of the test statistic for chromosome 1. Figure 2 shows the plots of the likelihood ratio versus the chromosome position for the analyses from the model with both additive and dominance effects and with cofactors. The number of cofactors included after the backward

Table 3. Genetic parameters used for data simulation of the five populations using the model with additive and dominance effects, for two chromosomes of 100 cM in length each.

	Populations				
	1	2	3	4	5
Mean	50	52	54	56	58
Phenotypic variance	100	108.6	116.6	125.4	134.6
CV†	0.2	0.2	0.2	0.2	0.2
Variance explained by the QTL, %					
	Chromosome 1				
Sim.‡ 1	10	7.8	6.7	6.0	4.9
Sim. 2	20	15.7	13.4	11.9	9.9
Sim. 3	40	31.3	26.8	23.8	19.8
Sim. 4	50	39.2	33.5	29.8	24.8
	Chromosome 2				
Sim. 1	0	6.2	5.7	5.3	0
Sim. 2	0	12.3	11.4	10.6	0
Sim. 3	0	24.6	22.9	21.3	0
Sim. 4	0	30.8	28.6	26.6	0

† Coefficient of variation.

‡ Simulated data set.

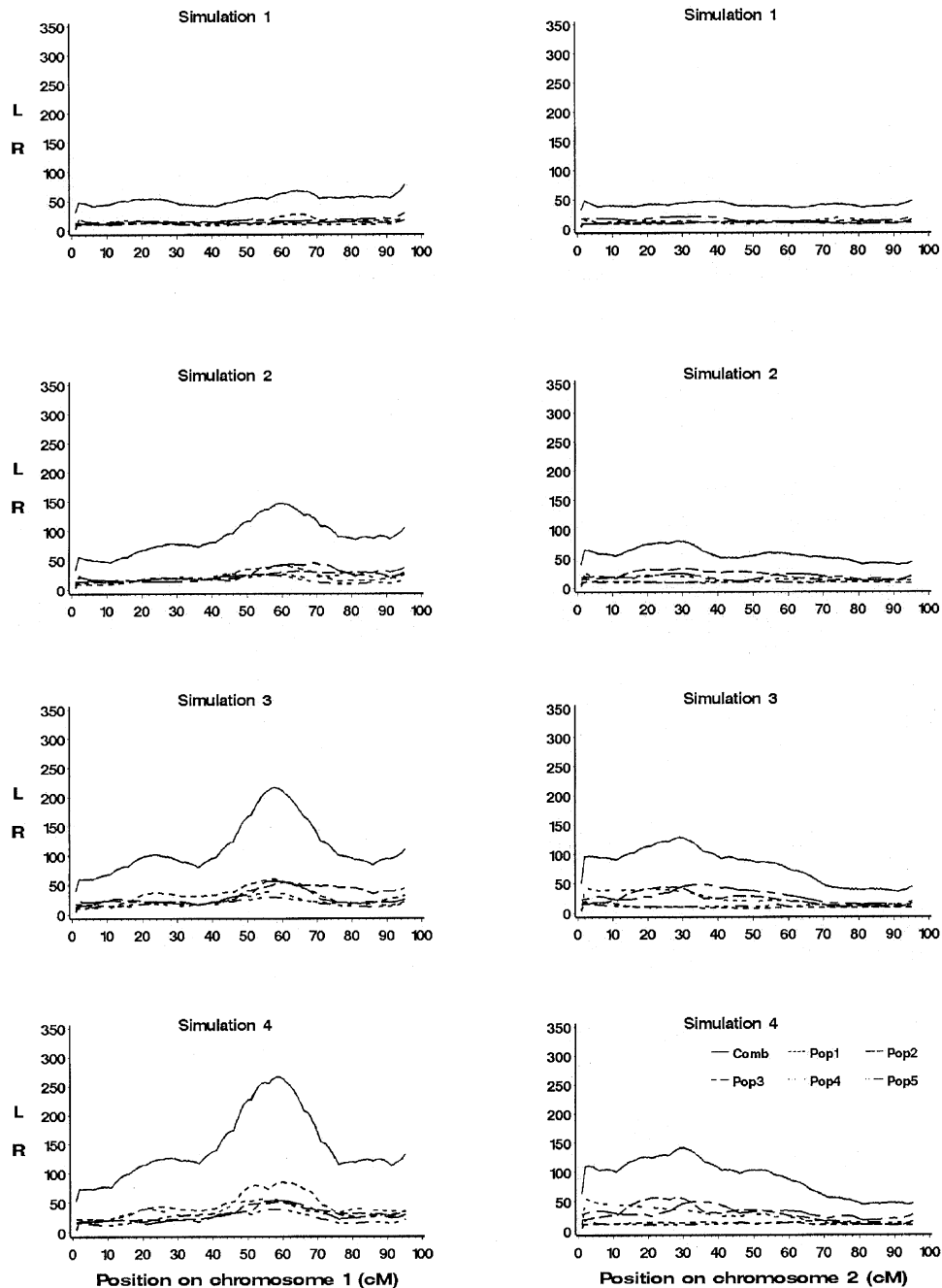


Fig. 1. Likelihood ratio curves vs. genome position of the haplotype-based QTL analyses for two linkage groups of 100 cM in length each, of the simulated five F_1 populations (Pop1, . . . , Pop5) of 100 individuals each. The single population analyses and the combined analyses (Comb) were performed under the additive and dominance effects model.

elimination was variable, from none to 10 for the single populations analyses and from 1 to 11 for the combined analyses. Improvement was achieved when marker cofactors were added to the model when estimating the QTL position in chromosome 1 for simulations 2, 3, and 4, effectively removing substantial residual variance. While the absolute maximum of the test statistic curves tended to be over intervals close to the true QTL position, the inflexion points flanking the maximum peaks covered segments larger than 30 cM in length when analyses were performed without cofactors (Fig. 1). These segments were narrowed to approximately 10 cM

in length when cofactors were added to the analyses (Fig. 2). The confidence intervals based on the two-LOD rule (Van Ooijen, 1992) are shown in Table 4. The cofactor model seriously misestimated the QTL position in the first simulated data set with the smallest QTL effects. However, the misestimation was no larger than 1 cM for QTL position in the other three simulations with QTL alleles of larger magnitude. Shorter confidence intervals containing true QTL positions were obtained with the model that included cofactors for the QTL in chromosome 1 in simulations 2, 3, and 4, no larger than 6 cM.

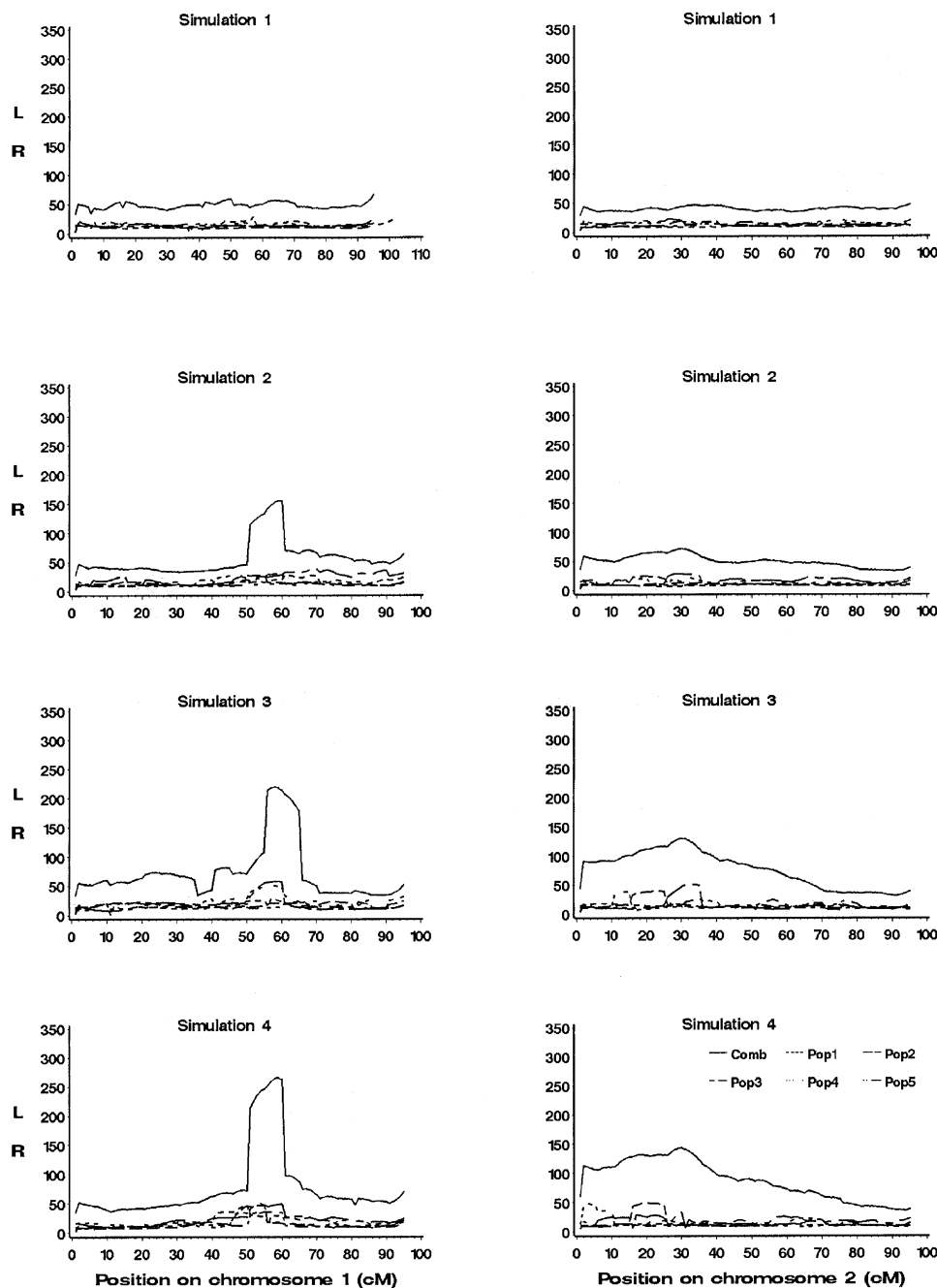


Fig. 2. Likelihood ratio curves vs. genome position of the haplotype-based QTL analyses for two linkage groups of 100 cM in length each, of the simulated five F_1 populations (Pop 1, . . . , Pop5) of 100 individuals each. The single population analyses and the combined analyses (Comb) were performed under the additive and dominance effects model. Cofactors were included in the analyses with a window of 5 cM.

The estimated effect for the allele-substitution values of QTLs contained in the founder clones using the additive and dominance effects model are shown in Table 5. The estimates were obtained with the combined analyses using cofactors, and correspond to the likelihood ratio curves showed in Fig. 2. In the first simulated data set with small QTL values, the estimated effects of the QTL alleles in both chromosomes were less accurate, nor did the analysis find the correct position of the QTLs.

For the QTL alleles on chromosome 1, the allele substitution in all four simulations (Table 5) was significantly different from 0 only for the common founder

and the second founder of the first population. For the latter three simulations, even though there is an upward bias in the point estimates of the allele-substitution effects of these founders, a 95% confidence interval contains the true parameter value. Zeng (1993) showed that the partial regression coefficients are biased estimates of QTL effects. In this research, the allele-substitution effects of QTL alleles in the common founder were upwardly biased from approximately 15, 5, and 7% for QTLs that explain on average 14, 28, and 35% of the phenotypic variance across populations (Table 3). Larger bias was observed for the point estimates of the allele-

Table 4. Estimated position of the putative QTL and its confidence interval (in parentheses) for the analyses performed using the model with additive and dominance effects.

Chromosome	Simulation	QTL position†	
		Without cofactors	With cofactors
cM			
1	1	62 (44, 95)	95 (93, 95)
	2	60 (53, 68)	60 (56, 61)
	3	59 (55, 63)	58 (55, 61)
	4	59 (53, 62)	59 (56, 61)
2	1	39 (2, 68)	95 (86, 95)
	2	29 (22, 33)	30 (17, 34)
	3	30 (26, 34)	30 (26, 33)
	4	30 (27, 34)	30 (26, 33)

† The estimated QTL position was obtained as the point in the linkage map at which the likelihood ratio statistic showed a maximum value. The confidence intervals were obtained using the equality $LOD = LR / (2 \ln 10)$ (Lynch and Walsh, 1998) and the two-LOD rule. The true QTL position is $x = 59$ cM in chromosome 1 and $x = 29$ in chromosome 2.

substitution effects for QTL alleles in the second parent of the first population with larger standard errors than in the common founder. The average effect of allele substitution in chromosome 1 of founders 3 and 4 were not significant. These were QTLs of smaller effect than of those from the common founder and the second founder of the first population, as given by (11). Seventy-five percent of the estimates of the allele-substitution effects in chromosome 2 of founders 2, 3, and 4 were significant

with a relatively good approximation to the true value, as expected with QTL alleles of stronger magnitude.

The estimates of dominance effects are shown in Table 6. Most of the significant dominance coefficient estimates were observed for chromosome 1 of the third and fourth simulations, in which the dominance effect of the QTL had a larger magnitude. As in the case of the allele-substitution effect, there was a tendency among the significant coefficients to overestimate the true parameter value. There were two false positives, one in the allele-substitution effect and one in the dominance coefficient, both on chromosome 2 (Table 5 and 6).

DISCUSSION

This study shows the use of breeding populations of outbred plant species to map QTLs based on founder-origin probabilities (Reyes-Valdés, 2000; Reyes-Valdés and Williams, 2002) that trace specific haplotypes from the founders to their progeny. Figures 1 and 2 showed that the test statistic values are increased when a set of F_1 populations with one common founder are considered in the analysis, and the absolute peaks of the curves were within approximately 1 cM of either side of the real QTL position in the linkage group for quantitative trait loci that explain, on average, a minimum of 14% of the phenotypic variation. Although the allele-substitution effect of mild and strong QTLs in the common founder was often overestimated (Table 5), a 95% confidence interval contains the real value of the parameter.

The most recent QTL map for cocoa was developed by Clement et al. (2003a, 2003b), from the crosses of

Table 5. Allele-substitution effects of the parental founder clones corresponding to the five simulated populations. The true parameter values are shown on the first row of every combination of chromosome and simulation number, the estimated position and effect below, and the standard error of the estimated effect (in parentheses).

Chromosome	Simulation	QTL position†	Founder clones					
			P_{CF}	P_{SF_1}	P_{SF_2}	P_{SF_3}	P_{SF_4}	P_{SF_5}
cM								
1	1	95	5.2	2.6	1.7	1.3	1.0	0
			11.4*	19.4*	-0.2	0.6	0.5	-5.8
	2	60	(5.4)	(8.0)	(8.0)	(9.1)	(9.0)	(9.7)
			7.3	3.6	2.4	1.8	1.5	0
	3	58	8.4**	7.3**	-4.1	0.9	-1.2	-17.2
			(0.9)	(2.0)	(3.7)	(2.7)	(3.0)	(6.8)
	2	59	10.3	5.2	3.4	2.6	2.1	0
			10.8**	7.0**	9.8**	0.1	0.7	2.7
	2	59	(0.8)	(1.7)	(2.2)	(5.2)	(4.0)	(2.7)
			11.5	5.8	3.8	2.9	2.3	0
	2	59	12.3**	8.3**	6.0**	-1.6	-5.3	-4.1
			(0.8)	(2.0)	(1.7)	(2.5)	(2.7)	(3.1)
2	1	95	0	0	5.2	5.2	5.2	0
			2.6*	1.2	-0.4	4.1	2.0	0.4
	2	30	(1.2)	(1.9)	(9.6)	(11.0)	(2.1)	(2.2)
			0	0	7.3	7.3	7.3	0
	3	30	0.2	0.9	9.6**	7.1**	4.8*	-0.8
			(0.8)	(1.6)	(2.0)	(1.9)	(2.0)	(2.0)
	3	30	0	0	10.3	10.3	10.3	0
			-0.8	0.9	10.9**	11.1*	10.7**	-0.6
	4	30	(1.8)	(1.6)	(1.8)	(1.6)	(1.9)	(2.3)
			0	0	12.3	12.3	12.3	0
	4	30	-1.7	1.1	12.1**	9.8**	10.2**	1.6
			(1.8)	(1.7)	(1.6)	(1.6)	(1.7)	(2.1)

* Significance at a probability of 0.05.

** Significance at a probability of 0.01.

† The estimated QTL position was obtained as the point in the linkage map at which the likelihood ratio statistic showed a maximum value. The true QTL position is $x = 59$ cM in chromosome 1 and $x = 29$ in chromosome 2. The standard errors were obtained as the squared root of the appropriate elements of the diagonal of the matrix $(X^T R^{-1} X)^{-1}$. P_{CF} , P_{SF_i} stand for common founder and second founder i (with $i = 1, 2, 3, 4, 5$), respectively.

Table 6. Dominance effects for the five simulated populations. The true parameter values are shown on the first row of every combination of chromosome and simulation number, the estimated position and effect below, and the standard error of the estimated effect (in parentheses).

Chromosome	Simulation	QTL position†	Population					
			1	2	3	4	5	
1	1	95	1.3	1	0.9	0.7	0	
			11.1*	1.6	-2.6	-0.3	-2.1	
		2	60	(4.8)	(4.9)	(5.6)	(5.5)	(5.6)
				1.8	1.5	1.2	1.0	0
		3	58	1.5	1.3	3.7*	1.8	-0.7
				(0.8)	(1.1)	(1.0)	(1.2)	(1.0)
		4	59	2.6	2.1	1.7	1.5	0
				2.0**	0.8	2.4*	4.0**	1.4
				(0.7)	(1.0)	(1.0)	(1.2)	(1.3)
				2.9	2.3	1.9	1.6	0
				3.1**	1.2	4.2**	1.4	0.9
				(0.8)	(0.9)	(1.0)	(1.1)	(1.1)
2	1	95	0	0	0	0	0	
			-0.4	0.7	-0.1	1.4	1.4	
		2	30	(1.0)	(1.4)	(1.5)	(1.0)	(1.1)
				0	0	0	0	0
		3	30	-0.7	-0.2	0.2	1.9	-0.9
				(0.8)	(1.0)	(0.9)	(1.0)	(1.0)
				0	0	0	0	0
				1.1	-1.4	-0.7	0.6	0.5
				(0.8)	(0.9)	(0.8)	(1.0)	(1.2)
				0	0	0	0	0
				-0.9	-0.5	1.7*	0.9	0.2
				(0.9)	(0.8)	(0.8)	(0.8)	(1.1)

* Significance at a probability of 0.05.

** Significance at a probability of 0.01.

† The estimated QTL position was obtained as the point in the linkage map at which the likelihood ratio statistic showed a maximum value. The true QTL position is $x = 59$ cM in chromosome 1 and $x = 29$ in chromosome 2. The standard errors were obtained as the squared root of the appropriate elements of the diagonal of the matrix $(X^T R^{-1} X)^{-1}$.

three female parental clones DR1, S52, and IMC78 and the male parental clone Catongo. DR1 and S52 are Trinitario genotypes and IMC78 is an upper Amazon Forastero, with heterozygosity estimates of 37, 27, and 27%, respectively; Catongo is a lower Amazon Forastero clone with a highly homozygous genotype. Each population was analyzed individually using an approximation to a testcross. The number of individuals of the populations developed from the female parents DR1, S52 and IMC78, were 96, 94, and 125 for yield components, vigor, and resistance to *P. palmivora*, and 95, 88, and 124 for bean traits and ovule number, respectively. QTL analyses using a backcross model with a sample size of 100, and assuming an informative marker linked 5 cM from the putative QTL, would be able to detect a QTL whose segregation accounts approximately a minimum of 23% of total variance with a power of detection (probability of detecting a true association) of 90% and a significance level of 5% (Lynch and Walsh, 1998). A Half-Sib Design in which every marker informative male parent is crossed to 100 female parents and a single offspring is scored from each mating, would have a power of detection of 44% with a significance level of 5% and assuming a linked informative marker 5 cM away from the QTL that explains 14% of the phenotypic variance. However, power is increased to 75% if three half-sib families of 100 offspring each are used to perform the analysis, and to 90% if five half-sib families of 100 offspring each are used for the calculations (Lynch and Walsh, 1998). The results that we presented using the haplotypic-based method give clear evidence

that combining F_1 populations to perform association analysis would increase the likelihood ratio peaks over the region where the QTLs are located, as discussed by Jansen et al. (2003) for multiple related $F_{2,3}$ populations, yielding more accurate and precise QTL maps than single population analyses, especially for mild and strong QTLs contained in the common founder. However, more extensive simulation research should be done to test the proposed method that would include different number of populations and population sizes, unequal sizes among populations, and multiple QTLs in a linkage group. This method was designed to be implemented with fully informative codominant markers, such as restriction fragment length polymorphism (RFLP) and simple sequence repeats (SSR). Less-than fully informative markers or dominant markers cannot be used with the haplotypic method as described above. Research should be done on the implementation of partially informative markers to estimate QTL as an extension of the approach outlined in this study.

Marker cofactors were successfully used in the models of this research to control genetic background (Jansen and Stam, 1994; Zeng, 1994). Another possibility to explore to control genetic background for more precise QTL estimation is the use of a structured covariance matrix that would include both components of genetic variance and covariance among individuals (Gianola et al., 2003; Lund et al., 2003; Piepho, 2000), given that individuals of the same population are full-sibs and individuals from different populations with one common founder are half-sibs. Thus, the $Cov(y_{ij}, y_{ij'}) = d_1$ and

$Cov(y_{ij}, y_{ij'}) = d_2$ are the covariances within and among populations due to additive and nonadditive gene action, including possible correlations caused by the environment. Precise prior estimates of the narrow sense heritability for the analyzed trait would allow separation of the additive component of variance and covariance of nonadditive components, otherwise they will be pooled together because of unreplicated F_1 progeny. We used a ridge-stabilized Newton-Raphson algorithm in this research that generally converges with few iterations and makes available asymptotic sample variances for the estimated parameters, however, requires matrix inversion in each iteration, making it highly demanding of computational resources. A covariance QTL analysis with a structured variance covariance-matrix as described above could also be performed with a derivative free algorithm that would require more rounds to reach convergence but would be computationally feasible since matrix inversion is not required (Meyer, 1989; Searle et al., 1992).

The haplotype-based approach proposed in this study requires founders that are both marker informative and QTL informative, requiring crosses between founders with heterozygous QTL genotypes with relatively closely linked heterozygous markers. Such a set of conditions indicates that this method is better suited for highly heterozygous outbred plant species, as is the case of many out crossing perennial plants with high genetic load (Williams, 1998), and with mainly additive gene action, as has been shown for cocoa clones (Clement et al., 2003a). In addition, combining populations by a common founder for QTL analysis implies the assumption of no interaction of the putative QTL with the genetic background. In the use of this approach in full-sib F_1 populations, every founder appears in more than one cross, allowing the combining of populations by common founders for QTL analysis (each population will be in more than one combination), increasing not only the accuracy and precision of the QTL position and effect estimates, but also the number of the putative QTLs given the increase in the number of parents considered.

APPENDIX

Reparameterization of Models [2] and [4]

The allele effects are reparameterized in terms of the allele-substitution effects as follows

$$\begin{aligned} \mu_i + \alpha_{H_{CF}^x} P_{H_{CF}^x/AB_{ij}} + \alpha_{H_{CF}^x} (1 - P_{H_{CF}^x/AB_{ij}}) \\ + \alpha_{H_{SF}^x} P_{H_{SF}^x/AB_{ij}} + \alpha_{H_{SF}^x} (1 - P_{H_{SF}^x/AB_{ij}}) \\ = \mu_i + \alpha_{H_{CF}^x} P_{H_{CF}^x/AB_{ij}} + \alpha_{H_{CF}^x} - \alpha_{H_{CF}^x} P_{H_{CF}^x/AB_{ij}} \\ + \alpha_{H_{SF}^x} P_{H_{SF}^x/AB_{ij}} + \alpha_{H_{SF}^x} - \alpha_{H_{SF}^x} P_{H_{SF}^x/AB_{ij}} \\ = (\mu_i + \alpha_{H_{CF}^x} + \alpha_{H_{SF}^x}) + (\alpha_{H_{CF}^x} - \alpha_{H_{CF}^x}) \\ P_{H_{CF}^x/AB_{ij}} + (\alpha_{H_{SF}^x} - \alpha_{H_{SF}^x}) P_{H_{SF}^x/AB_{ij}} \\ = \mu_i^* + \alpha_{H_{CF}^x}^* P_{H_{CF}^x/AB_{ij}} + \alpha_{H_{SF}^x}^* P_{H_{SF}^x/AB_{ij}} \end{aligned}$$

The reduction in the dominance parameters is achieved as follows

$$\begin{aligned} \delta_{H_{CF}^x H_{SF}^x} P_{H_{CF}^x/AB_{ij}} P_{H_{SF}^x/AB_{ij}} + \delta_{H_{CF}^x H_{SF}^x} P_{H_{CF}^x/AB_{ij}} P_{H_{SF}^x/AB_{ij}} + \\ \delta_{H_{CF}^x H_{SF}^x} P_{H_{CF}^x/AB_{ij}} + P_{H_{SF}^x/AB_{ij}} + \delta_{H_{CF}^x H_{SF}^x} P_{H_{CF}^x/AB_{ij}} P_{H_{SF}^x/AB_{ij}} = \\ \delta_x (P_{H_{CF}^x/AB_{ij}} P_{H_{SF}^x/AB_{ij}} - P_{H_{CF}^x/AB_{ij}} P_{H_{SF}^x/AB_{ij}} - \\ P_{H_{CF}^x/AB_{ij}} P_{H_{SF}^x/AB_{ij}} + P_{H_{CF}^x/AB_{ij}} P_{H_{SF}^x/AB_{ij}}) = \\ \delta_x [P_{H_{CF}^x/AB_{ij}} - (1 - P_{H_{CF}^x/AB_{ij}})] [P_{H_{SF}^x/AB_{ij}} - (1 - P_{H_{SF}^x/AB_{ij}})] = \\ \delta_x (2P_{H_{CF}^x/AB_{ij}} - 1)(2P_{H_{SF}^x/AB_{ij}} - 1). \end{aligned}$$

REFERENCES

- Beavis, W.D. 1998. QTL analyses: Power, precision and accuracy. p. 145–161. In A.H. Paterson (ed.) Molecular dissection of complex traits. CRC Press, Boca Raton, FL.
- Clement, D., A.M. Risterucci, J.C. Motamayor, J. N'Goran, and C. Lanaud. 2003a. Mapping quantitative trait loci for bean traits and ovule number in *Theobroma cacao* L. *Genome* 46:103–111.
- Clement, D., A.M. Risterucci, J.C. Motamayor, J. N'Goran, and C. Lanaud. 2003b. Mapping QTL for yield components, vigor, and resistance to *Phytophthora palmivora* in *Theobroma cacao* L. *Genome* 46:103–111.
- Churchill, G.A., and R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* 138:963–971.
- Gianola, D., M. Perez-Enciso, and M.A. Toro. 2003. On marker-assisted prediction of genetic value: Beyond the ridge. *Genetics* 163:347–365.
- Haldane, J.B.S. 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* 8:299–309.
- Haley, C.S., and S.A. Knott. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69:315–324.
- Haley, C.S., S.A. Knott, and J.M. Elsen. 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* 136:1195–1207.
- Jannink, J.L., and R.C. Jansen. 2001. Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* 157:445–454.
- Jansen, R.C., and P. Stam. 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* 138:871–881.
- Jansen, R.C., D.L. Johnson, and J.A.M. Van Arendonk. 1998. A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. *Genetics* 148:391–399.
- Jansen, R.C., J.L. Jannink, and W.D. Beavis. 2003. Mapping quantitative trait loci in plant breeding populations: Use of parental haplotype sharing. *Crop Sci.* 43:829–834.
- Lander, E.S., and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185–199.
- Littell, R.C., G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. 1996. SAS System for mixed models. SAS Institute Inc., Cary, NC.
- Lund, M.S., P. Sørensen, B. Guldbrandtsen, and D.A. Sorensen. 2003. Multitrait fine mapping of quantitative trait loci using combined linkage disequilibria and linkage analysis. *Genetics* 163:405–410.
- Luo, Z.W. 1993. The power of two experimental designs for detecting linkage between a marker locus and a locus affecting a quantitative character in a segregating population. *Genet. Sel. Evol.* 25:249–261.
- Lynch, M., and B. Walsh. 1998. *Genetics and analysis of quantitative traits*. Sinauer Associates, Inc. Sunderland, MA.
- Maliapaard, C., J. Jansen, and J.W. Van Ooijen. 1997. Linkage analysis in a full-sib family of an outbreeding plant species: Overview and consequences for applications. *Genet. Res.* 70:237–250.
- Meyer, K. 1989. Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genet. Sel. Evol.* 21:317–340.
- Piepho, H.P. 2000. A mixed-model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. *Genetics* 156:2043–2050.

- Reyes-Valdés, M.H. 2000. A model for marker-based selection in gene introgression breeding programs. *Crop Sci.* 40:91–98.
- Reyes-Valdés, M.H., and C.G. Williams. 2002. A haplotypic approach to founder-origin probabilities and outbred QTL analysis. *Genet. Res.* 80:231–236.
- Searle, S.R. 1971. *Linear models*. John Wiley and Sons, New York.
- Searle, S.R., G. Casella, and C.E. McCulloch. 1992. *Variance components*. John Wiley and Sons, New York.
- Self, G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Statist. Assoc.* 82:605–610.
- Soller, M., and A. Genizi. 1978. The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics* 34:47–55.
- Van Ooijen, J.W. 1992. Accuracy of mapping quantitative trait loci in autogamous species. *Theor. Appl. Genet.* 84:803–811.
- Weller, J.I., Y. Kashi, and M. Soller. 1990. Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *J. Dairy Sci.* 73:2525–2537.
- Williams, C.G. 1998. QTL mapping in outbred pedigrees. p. 81–94. *In* A.H. Paterson (ed.) *Molecular dissection of complex traits*. CRC Press, Boca Raton, FL.
- Wu, R., C.-X. Ma, I. Painter, and Z.-B. Zeng. 2002. Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. *Theor. Popul. Biol.* 61:349–363.
- Zeng, Z.-B. 1993. Theoretical basis for separation of multiple linked gene effects in mapping of quantitative trait loci. *Proc. Natl. Acad. Sci. USA* 90:10972–10976.
- Zeng, Z.-B. 1994. Precision mapping of quantitative trait loci. *Genetics* 136:1457–1468.