

Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines

Changjian Jiang¹ & Zhao-Bang Zeng²

¹*CIMMYT Int. Lisboa 27, Apdo Postal 6-641, Mexico 06600 D.F. Mexico;* ²*Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA*

Received 6 December 1996 Accepted 14 April 1997

Key words: dominant markers, genetic mapping, Markov chain, missing data, quantitative trait loci

Abstract

Dominant phenotype of a genetic marker provides incomplete information about the marker genotype of an individual. A consequence of using this incomplete information for mapping quantitative trait loci (QTL) is that the inference of the genotype of a putative QTL flanked by a marker with dominant phenotype will depend on the genotype or phenotype of the next marker. This dependence can be extended further until a marker genotype is fully observed. A general algorithm is derived to calculate the probability distribution of the genotype of a putative QTL at a given genomic position, conditional on all observed marker phenotypes in the region with dominant and missing marker information for an individual. The algorithm is implemented for various populations stemming from two inbred lines in the context of mapping QTL. Simulation results show that if only a proportion of markers contain missing or dominant phenotypes, QTL mapping can be almost as efficient as if there were no missing information in the data. The efficiency of the analysis, however, may decrease substantially when a very large proportion of markers contain missing or dominant phenotypes and a genetic map has to be reconstructed first on the same data as well. So it is important to combine dominant markers with codominant markers in a QTL mapping study.

Introduction

Many PCR-based genetic markers behave like dominant markers. Unlike codominant markers such as restriction fragment length polymorphism (RFLP), which can generally show three band patterns in an F_2 population in electrophoretic gels, each representing one genotype of a probe, dominant markers such as random amplified polymorphic DNA (RAPD) can generally show only two patterns, presence or absence of a band, so that a heterozygote can have the same band pattern as one of the homozygotes. There has been some concern about the use of dominant markers in mapping quantitative trait loci (QTL) because of partial missing information. Although the problem of dominant markers can be avoided through experimental designs, such as using recombinant inbred lines and double haploids to remove heterozygote class or just using segregating markers in a backcross, it is common

to have dominant markers in other populations, such as F_2 .

When a marker is not fully informative, it is generally known that information from other markers in a linkage group can be used to recover some missing information. In their original paper on linkage map reconstruction, Lander and Green (1987) outlined a Markov chain method to recover missing information. The same idea has been used repeatedly in literature for linkage map reconstruction and for mapping QTL in various experimental designs. For example, Martinez and Curnow (1994) discussed missing marker problems in QTL mapping. Haley, Knott and Elsen (1994) used similar arguments to improve the calculation of conditional QTL genotype probabilities given marker phenotypes for mapping QTL from a cross between outbred populations. A similar method was used by Fulker, Cherney and Cardon (1994) for sib-pair interval mapping using multiple markers. In many of these analyses, however, the proposed methods enumerate

various possibilities and consider them one by one. This approach works effectively only in a few limited situations. Jansen (1996) used a Monte Carlo EM algorithm via Gibbs sampling to deal with missing and dominant markers. This is an approximate method and requires extensive computation, particularly when the number of markers with incomplete information is large.

In this paper, we derive a general algorithm to systematically deal with dominant and missing markers in F_2 and other populations derived from two inbred lines. In particular, we try to formulate the algorithm in a way that can efficiently calculate QTL genotype distribution given observed marker phenotypes. Our analysis is similar to the method of Lander and Green (1987) in spirit, but with derivation and sufficient details in the analysis.

An algorithm for F_2 population

Algorithm

Consider an F_2 population from a cross between two inbred lines, P_1 and P_2 . Suppose there are m markers on a chromosome whose map positions are known and arranged in the order of M_1, \dots, M_m . In this population, each marker or QTL has three possible genotypes. Let x_k denote the genotype of marker (or QTL) M_k for an individual, which takes a value 1, 0 or -1 if M_k is homozygote of P_1 type, heterozygote, or homozygote of P_2 type, respectively.

To facilitate the following discussion, we let z_k denote the phenotype of marker (or QTL) M_k for the same individual. When a marker is fully observed, the phenotype equals the genotype, i.e., $z_k = x_k = \{1\}, \{0\}$, or $\{-1\}$. When a marker is unobserved (i.e., missing), the genotype is unknown with $z_k = \{1, 0, -1\} = M$ for missing. When a marker is partially observed, the phenotype includes two possible genotypes. In this paper, a dominant phenotype represents homozygote of P_1 type or heterozygote (i.e., $z_k = \{1, 0\} = D$ for dominance or $z_k \neq -1$), and a recessive phenotype represents heterozygote or homozygote of P_2 type (i.e., $z_k = \{0, -1\} = R$ for recessive or $z_k \neq 1$). (Another subset $z_k = \{1, -1\}$ can also be included in analysis if necessary).

If M_k is a putative QTL whose genotype may or may not be observed (depending on the testing position), a very important analysis in QTL mapping is to calculate, for each individual, the conditional probab-

ity of x_k taking different values given observed marker phenotypes on the chromosome. We denote this probability as $P(x_k | z_1, \dots, z_m)$.

If M_{k-1} and M_{k+1} , the two flanking markers of M_k , are both fully observed for the individual, the conditional probability depends entirely on the phenotype of M_k , the genotypes of M_{k-1} and M_{k+1} , and the recombination frequencies between M_{k-1} and M_k and between M_k and M_{k+1} and is independent of other markers on the chromosome under the assumption of no crossing-over interference, i.e.,

$$\begin{aligned} P(x_k | z_1, \dots, x_{k-1}, z_k, x_{k+1}, \dots, z_m) \\ = P(x_k | x_{k-1}, z_k, x_{k+1}). \end{aligned}$$

However, if one (or both) of the flanking markers is unobserved or only partially observed, the genotype or phenotype at the next marker away from the flanking marker can provide some information about the genotype of the flanking marker and this in turn will improve the estimation of the probability distribution of the genotype at the testing position for the QTL for the individual. This dependence can be extended further in each direction until a marker locus for which the genotype is fully observed or to the terminal marker of the chromosome.

Let M_i and M_l ($i \leq k \leq l$) be two most adjacent fully observed markers. If there is no fully observed marker in one or both directions, take $M_i = M_1$ or $M_l = M_m$ or both. The task is to calculate $P(x_k | z_i, \dots, z_l)$. By Bayes' theorem,

$$P(x_k | z_i, \dots, z_l) = \frac{P(x_k)P(z_i \dots z_l | x_k)}{\sum_{x_k} P(x_k)P(z_i \dots z_l | x_k)}. \quad (1)$$

Note that under the assumption of no crossing-over interference, for a given specific value of x_k

$$\begin{aligned} P(z_i \dots z_l | x_k) \\ = P(z_i \dots z_k | x_k, z_{k+1}, \dots, z_l)P(z_{k+1} \dots z_l | x_k) \\ = P(z_i \dots z_k | x_k)P(z_{k+1} \dots z_l | x_k). \end{aligned}$$

In (1), $P(x_k)$ is the unconditional or prior probability of x_k in a population. Let

$$\mathbf{q}_k = \{P(x_k)\}_{(3 \times 1)}$$

denote a row vector of the prior probability $P(x_k)$, i.e.,

$$\mathbf{q}'_k = [P(x_k = 1), P(x_k = 0), P(x_k = -1)]$$

where \prime denotes transposition. Similarly, let also

$$\begin{aligned}\mathbf{p}_k^R &= \{P(z_{k+1} \cdots z_l | x_k)\}_{(3 \times 1)} \\ \mathbf{p}_k^L &= \{P(z_i \cdots z_k | x_k)\}_{(3 \times 1)} \\ \mathbf{p}_k &= \{P(x_k | z_i, \cdots, z_l)\}_{(3 \times 1)}.\end{aligned}$$

Then, equation (1) can be expressed as

$$\mathbf{p}_k = \frac{\mathbf{q}_k \circ (\mathbf{p}_k^L \circ \mathbf{p}_k^R)}{\mathbf{q}'_k (\mathbf{p}_k^L \circ \mathbf{p}_k^R)} \quad (2)$$

where \circ denotes componentwise product of vectors.

For an F_2 population,

$$\mathbf{q}'_k = [\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]. \quad (3)$$

\mathbf{p}_k^R and \mathbf{p}_k^L can be calculated via a Markov chain process. To show how to calculate them, let us first consider some simple situations. If an individual has the dominant phenotype at M_{k+1} (i.e., $z_{k+1} = \{1, 0\}$),

$$\begin{aligned}P(z_{k+1} | x_k) &= \sum_{x_{k+1} \in z_{k+1}} P(x_{k+1} | x_k) \\ &= P(x_{k+1} = 1 | x_k) + P(x_{k+1} = 0 | x_k).\end{aligned}$$

If the individual also has the recessive phenotype at M_{k+2} (i.e. $z_{k+2} = \{0, -1\}$),

$$\begin{aligned}&P(z_{k+1} z_{k+2} | x_k) \\ &= \sum_{x_{k+1} \in z_{k+1}} \sum_{x_{k+2} \in z_{k+2}} P(x_{k+1} x_{k+2} | x_k) \\ &= \sum_{x_{k+1} \in z_{k+1}} \sum_{x_{k+2} \in z_{k+2}} P(x_{k+2} | x_{k+1}) P(x_{k+1} | x_k) \\ &= [P(x_{k+2} = 0 | x_{k+1} = 1) \\ &\quad + P(x_{k+2} = -1 | x_{k+1} = 1)] P(x_{k+1} = 1 | x_k) \\ &\quad + [P(x_{k+2} = 0 | x_{k+1} = 0) \\ &\quad + P(x_{k+2} = -1 | x_{k+1} = 0)] P(x_{k+1} = 0 | x_k).\end{aligned} \quad (4)$$

Then if we let

$$\begin{aligned}\mathbf{H}(r_k) &= \begin{bmatrix} P(x_{k+1} = 1 | x_k = 1) & P(x_{k+1} = 0 | x_k = 1) & P(x_{k+1} = -1 | x_k = 1) \\ P(x_{k+1} = 1 | x_k = 0) & P(x_{k+1} = 0 | x_k = 0) & P(x_{k+1} = -1 | x_k = 0) \\ P(x_{k+1} = 1 | x_k = -1) & P(x_{k+1} = 0 | x_k = -1) & P(x_{k+1} = -1 | x_k = -1) \end{bmatrix} \\ &= \begin{bmatrix} (1-r_k)^2 & 2r_k(1-r_k) & r_k^2 \\ r_k(1-r_k) & (1-r_k)^2 + r_k^2 & r_k(1-r_k) \\ r_k^2 & 2r_k(1-r_k) & (1-r_k)^2 \end{bmatrix} \end{aligned} \quad (5)$$

which denotes a transition probability matrix from M_k to M_{k+1} (and is also a transition probability matrix from M_{k+1} to M_k), where r_k is the recombination frequency between M_k and M_{k+1} , the equation (4) can be expressed in matrix form as

$$\mathbf{p}_k^R = \mathbf{H}_D(r_k) \mathbf{H}_R(r_{k+1}) \mathbf{c}$$

with $\mathbf{H}_D(r_k) = \mathbf{H}(r_k) \mathbf{I}_D$, $\mathbf{H}_R(r_{k+1}) = \mathbf{H}(r_{k+1}) \mathbf{I}_R$,

$$\mathbf{I}_D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \mathbf{I}_R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

The function of matrices \mathbf{I}_D and \mathbf{I}_R is to make appropriate column elements of \mathbf{H} zero.

Thus, in general,

$$\mathbf{p}_k^R = \mathbf{H}_{z_{k+1}}(r_k) \mathbf{H}_{z_{k+2}}(r_{k+1}) \cdots \mathbf{H}_{z_l}(r_{l-1}) \mathbf{c} \quad (6)$$

where $z_j = M, D, R, 1, 0$ or -1 , depending on the information content of the phenotype of marker M_j . As specified above, $\mathbf{H}_{z_j} = \mathbf{H} \mathbf{I}_{z_j}$ with $\mathbf{I}_M = \mathbf{I}$ the identity matrix, $\mathbf{I}_1, \mathbf{I}_0$ and \mathbf{I}_{-1} having 1 in one corresponding diagonal element and 0 everywhere else. The chain specified by (6) connects and sums together all relevant paths of joint probabilities of genotypes based on observed marker phenotypes on the right side of x_k . Similarly,

$$\mathbf{p}_k^L = \mathbf{I}_{z_k} \mathbf{H}_{z_{k-1}}(r_{k-1}) \mathbf{H}_{z_{k-2}}(r_{k-2}) \cdots \mathbf{H}_{z_i}(r_i) \mathbf{c}. \quad (7)$$

The function of \mathbf{I}_{z_k} is to make appropriate row elements of \mathbf{p}_k^L and thus \mathbf{p}_k as well zero. Similarly, if \mathbf{p}_k^R is defined to include z_k as well for some applications, i.e., $\mathbf{p}_k^R = \{P(z_k \cdots z_l | x_k)\}_{(3 \times 1)}$ (see below),

\mathbf{p}_k^R can also be calculated by (6) and then premultiplied by \mathbf{I}_{z_k} just like (7). Usually, the testing position M_k is between markers and z_k is missing. In this case, the individual has non-zero probability at all of the three genotypes. When the testing position M_k for a QTL is at a marker and the individual has the dominant phenotype at the position, $z_k = D$ and the individual has zero probability for $x_k = -1$.

Note that $\mathbf{H}_M(r_{k,k+1}) = \mathbf{H}_M(r_k)\mathbf{H}_M(r_{k+1})$ with $r_{k,k+1} = r_k + r_{k+1} - 2r_k r_{k+1}$. Also $\mathbf{H}_D(r_{k,k+1}) = \mathbf{H}_M(r_k)\mathbf{H}_D(r_{k+1})$ and $\mathbf{H}_R(r_{k,k+1}) = \mathbf{H}_M(r_k)\mathbf{H}_R(r_{k+1})$. Thus, the operation of the chains can be shortened for intervals with missing markers.

In practice, \mathbf{p}_k^R and \mathbf{p}_k^L can be calculated first for all markers, which can then be used later to obtain the conditional probabilities of genotypes for any position covered by markers in mapping QTL. For example, assuming that $M_{k'}$ is a testing position for a QTL in an interval flanked by M_k and M_{k+1} , then

$$\mathbf{p}_{k'}^R = \mathbf{H}(r_{k'}^R)\mathbf{p}_{k+1}^R \quad \text{and} \quad \mathbf{p}_{k'}^L = \mathbf{H}(r_{k'}^L)\mathbf{p}_k^L$$

where $r_{k'}^R$ is the recombination frequency between $M_{k'}$ and M_{k+1} and $r_{k'}^L$ between M_k and $M_{k'}$. This, in our opinion, is the most efficient way to calculate the conditional probability distribution of QTL genotype given observed marker phenotypes. Unlike Haley, Knott and Elsen (1994), the calculation here is performed in a recursive manner through a Markov chain and puts no constraint on the number of markers with incomplete information in analysis. Also, the chain analysis directly gives the probability distribution of QTL genotype given the observed marker phenotype while the Monte Carlo EM algorithm of Jansen (1996) provides only an approximation with intensive computation.

Efficiency of the algorithm

Simulations were performed to investigate the effects of missing and dominant markers on mapping QTL. We simulated a chromosome of 80 cM in length with marker coverage at every 5 cM, 10 cM, or 20 cM (three marker coverages) for an F_2 population. The linkage map is assumed to be known. Five marker compositions were simulated and compared: (a) all markers are codominant with no missing marker data; (b) all markers are codominant with 15% random missing marker data; (c) markers are codominant and dominant in alternate order; (d) markers are codominant, dominant, and recessive in alternate order; and (e) markers are dominant and recessive in alternate order. One QTL was

considered and simulated at 47.5 cM position. Analysis was performed by using simple interval mapping (Lander & Botstein, 1989; Model III of Zeng, 1994) with the conditional probability of the putative QTL genotype at a testing position calculated by (2).

The threshold used in reporting the power of the test was chosen to be $\text{LOD} = 2.3$ for all the marker compositions. Although it is not strictly appropriate, this value was chosen merely for the convenience of comparison. The sample size is 150 and the replicates of simulation were 1000. Results are presented in Table 1.

Results show that both statistical power of QTL detection and precision of QTL estimation generally decrease as more markers become missing or partially missing as expected. The power, however, does not change significantly when the marker density is 5 cM. There is also relatively little difference on the estimated standard deviations (SD) of estimates of QTL additive and dominance effects for different marker compositions. The standard deviations of estimates of QTL position increase noticeably only for case (e) when the marker density is 5 cM and for cases (d) and (e) when 10 and 20 cM. Significant decrease on the proportion of QTL mapped to the correct interval occurs also mostly for cases (d) and (e). Overall, the effect of different marker compositions on the power and precision of QTL mapping is small. This basically reflects the efficiency of the algorithm in utilizing all available marker information to infer the probability of QTL genotype.

It is also interesting to compare case (c) of 5 cM interval with case (a) of 10 cM interval. The latter case approximately corresponds to the former case when the dominant marker data are not utilized in analysis. The results of case (c) of 5 cM are consistently closer to those of case (a) of 5 cM than those of case (a) of 10 cM. Similar results can also be found when we compare case (c) of 10 cM with case (a) of 20 cM. These results clearly show that efficient use of incomplete marker information can greatly improve the power and precision of QTL mapping.

Note that all simulations in this section assumed that the marker linkage map is known. However, in practice the marker linkage map may be unknown and needs also to be inferred from the data. The effect of dominant and missing markers on linkage map reconstruction is investigated below. The error of missing and partial missing data on linkage map reconstruction will carry over to the calculation of distribution of QTL genotype.

Table 1. Simulation results on estimated power, QTL additive effect (a), dominance effect (d), position (θ), and proportion of QTL mapped into the correct interval (p_{int}) for different marker compositions and densities with parameters $a = 0.5$, $d = 0.25$ and $\theta = 47.5$ cM

Parameter estimated	Marker composition ^a	Marker density (interval size)		
		5 cM	10 cM	20 cM
Power	(a)	0.95	0.91	0.85
	(b)	0.95	0.90	0.83
	(c)	0.94	0.91	0.84
	(d)	0.95	0.85	0.76
	(e)	0.93	0.90	0.75
a (SD)	(a)	0.51(0.11)	0.50(0.13)	0.48(0.13)
	(b)	0.51(0.11)	0.50(0.13)	0.49(0.13)
	(c)	0.51(0.11)	0.50(0.13)	0.49(0.13)
	(d)	0.51(0.11)	0.50(0.13)	0.48(0.14)
	(e)	0.51(0.12)	0.51(0.13)	0.49(0.14)
d (SD)	(a)	0.26(0.20)	0.25(0.19)	0.24(0.22)
	(b)	0.26(0.20)	0.25(0.20)	0.24(0.23)
	(c)	0.26(0.19)	0.25(0.19)	0.24(0.23)
	(d)	0.26(0.20)	0.24(0.21)	0.23(0.25)
	(e)	0.26(0.20)	0.27(0.21)	0.22(0.27)
θ (SD)	(a)	47.2(6.6)	48.0(7.3)	47.5(10.2)
	(b)	47.1(6.7)	48.0(7.6)	47.1(11.0)
	(c)	47.2(6.7)	48.2(7.3)	47.8(10.2)
	(d)	47.2(6.5)	46.9(8.7)	48.5(12.1)
	(e)	46.9(6.9)	48.5(8.1)	48.1(12.1)
p_{int}	(a)	0.62	0.65	0.72
	(b)	0.61	0.66	0.69
	(c)	0.62	0.64	0.71
	(d)	0.61	0.59	0.66
	(e)	0.61	0.64	0.64

^a Marker composition: (a) all markers are codominant; (b) markers are codominant with 15% data missing at random; (c) markers are codominant and dominant in alternate order; (d) markers are codominant, dominant, and recessive in alternate order; (e) markers are dominant and recessive in alternate order.

Extension to several experimental designs

General algorithm

We now generalize the above Markov chain analysis to many other commonly used experimental designs stemming from two inbred lines. We first outline a general algorithm and then specify it for different experimental designs.

Recently Fisch, Ragot, and Gay (1996) derived the conditional genotypic probability distribution of

a testing position given two fully observed flanking markers for F_t populations by selfing and backcrosses of F_t to parental lines by using a Markov chain to link the transition of crossing-over events in different generations. They, however, employed two intervals involving genotypes of three loci in their analysis and also did not analyze the dominant marker situation. As demonstrated above, the analysis can be performed only in one interval in different generations and multiple intervals can be linked by another Markov chain. This approach is particularly important when

we consider multiple intervals involving dominant markers.

In the above F_2 analysis, marker genotyping and trait phenotyping are assumed to be performed on the same individual. However, in some QTL mapping experiments or breeding programs, traits can be measured on some progeny of the individuals whose marker composition is genotyped (e.g., Stuber et al., 1992; Beavis et al., 1994). In this analysis, we also take this situation into account, as did Fisch, Ragot, and Gay (1996). Let $\{z\}^u$ denote a phenotypic observation of the set of markers M_i, \dots, M_l for an individual in population u , and x_k^v denote the putative QTL genotype at M_k in population v for the same individual (when $v = u$) or for the progeny of the individual (when $v \neq u$). Then

$$\begin{aligned} P(x_k^v | \{z\}^u) &= \sum_{x_k^u} P(x_k^v x_k^u | \{z\}^u) \\ &= \sum_{x_k^u} P(x_k^u | \{z\}^u) P(x_k^v | x_k^u, \{z\}^u) \\ &= \sum_{x_k^u} P(x_k^u | \{z\}^u) P(x_k^v | x_k^u) \quad (8) \end{aligned}$$

Let

$$\begin{aligned} \mathbf{p}_k^v &= \{P(x_k^v | \{z\}^u)\}_{(3 \times 1)} \\ \mathbf{q}_k^u &= \{P(x_k^u)\}_{(3 \times 1)} \\ \mathbf{p}_k^{Ru} \circ \mathbf{p}_k^{Lu} &= \{P(\{z\}^u | x_k^u)\}_{(3 \times 1)} \\ \mathbf{M}_{u \rightarrow v} &= \{P(x_k^v | x_k^u)\}_{(3 \times 3)}. \end{aligned}$$

Equation (8) can be expressed as

$$\mathbf{p}_k^v = \frac{\mathbf{M}'_{u \rightarrow v} [\mathbf{q}_k^u \circ (\mathbf{p}_k^{Ru} \circ \mathbf{p}_k^{Lu})]}{\mathbf{q}_k^{u'} (\mathbf{p}_k^{Ru} \circ \mathbf{p}_k^{Lu})}. \quad (9)$$

Now it remains to determine \mathbf{q}_k^u , \mathbf{p}_k^{Ru} , \mathbf{p}_k^{Lu} and $\mathbf{M}_{u \rightarrow v}$ for different experimental designs.

Selfed F_t

We first extend the analysis from F_2 to selfed F_t for an arbitrary t . When $u = v$, $\mathbf{M}_{u \rightarrow v} = \mathbf{I}$. For $u = F_t$, it is

known that

$$\mathbf{q}_k^{F_t'} = \left[\frac{1}{2} - \frac{1}{2^t}, \frac{1}{2^{t-1}}, \frac{1}{2} - \frac{1}{2^t} \right]. \quad (10)$$

To obtain \mathbf{p}_k^{Ru} and \mathbf{p}_k^{Lu} , we need to derive \mathbf{H} for $u = F_t$. For two loci, there are ten genotypes including two double heterozygotes. These ten genotypes are

$$\begin{aligned} &\left[\frac{AB}{AB}, \frac{AB}{Ab}, \frac{Ab}{Ab}, \frac{AB}{aB}, \frac{AB}{ab}, \frac{Ab}{aB}, \right. \\ &\left. \frac{Ab}{ab}, \frac{aB}{aB}, \frac{aB}{ab}, \frac{ab}{ab} \right]. \end{aligned}$$

Let \mathbf{p}'_{F_t} be a row vector of frequencies of these ten genotypes in an F_t population. For F_1 which is a cross from two inbred lines,

$$\mathbf{p}'_{F_1} = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0].$$

The transition probability matrix of these genotypes in two generations by selfing is

$$\mathbf{T}_S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 \\ \frac{w^2}{4} & \frac{rw}{2} & \frac{r^2}{4} & \frac{rw}{2} & \frac{w^2}{2} & \frac{r^2}{2} & \frac{rw}{2} & \frac{r^2}{4} & \frac{rw}{2} & \frac{w^2}{4} \\ \frac{r^2}{4} & \frac{rw}{2} & \frac{w^2}{4} & \frac{rw}{2} & \frac{r^2}{2} & \frac{w^2}{2} & \frac{rw}{2} & \frac{w^2}{4} & \frac{rw}{2} & \frac{r^2}{4} \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where $w = 1 - r$. Then,

$$\mathbf{p}'_{F_t} = \mathbf{p}'_{F_1} \mathbf{T}_S^{t-1}. \quad (11)$$

The transition probability matrix equivalent to (5) for F_t is then

$$\mathbf{H}_{F_t} = \begin{bmatrix} \frac{\mathbf{p}_{F_t}[1]}{\mathbf{p}_{F_t}[1]+\mathbf{p}_{F_t}[2]+\mathbf{p}_{F_t}[3]} & \frac{\mathbf{p}_{F_t}[2]}{\mathbf{p}_{F_t}[1]+\mathbf{p}_{F_t}[2]+\mathbf{p}_{F_t}[3]} & \frac{\mathbf{p}_{F_t}[3]}{\mathbf{p}_{F_t}[1]+\mathbf{p}_{F_t}[2]+\mathbf{p}_{F_t}[3]} \\ \frac{\mathbf{p}_{F_t}[4]}{\mathbf{p}_{F_t}[4]+\mathbf{p}_{F_t}[5]+\mathbf{p}_{F_t}[6]+\mathbf{p}_{F_t}[7]} & \frac{\mathbf{p}_{F_t}[5]+\mathbf{p}_{F_t}[6]}{\mathbf{p}_{F_t}[4]+\mathbf{p}_{F_t}[5]+\mathbf{p}_{F_t}[6]+\mathbf{p}_{F_t}[7]} & \frac{\mathbf{p}_{F_t}[7]}{\mathbf{p}_{F_t}[4]+\mathbf{p}_{F_t}[5]+\mathbf{p}_{F_t}[6]+\mathbf{p}_{F_t}[7]} \\ \frac{\mathbf{p}_{F_t}[8]}{\mathbf{p}_{F_t}[8]+\mathbf{p}_{F_t}[9]+\mathbf{p}_{F_t}[10]} & \frac{\mathbf{p}_{F_t}[9]}{\mathbf{p}_{F_t}[8]+\mathbf{p}_{F_t}[9]+\mathbf{p}_{F_t}[10]} & \frac{\mathbf{p}_{F_t}[10]}{\mathbf{p}_{F_t}[8]+\mathbf{p}_{F_t}[9]+\mathbf{p}_{F_t}[10]} \end{bmatrix},$$

where $\mathbf{p}_{F_t}[i]$ is the i th element of \mathbf{p}_{F_t} . A general solution of \mathbf{p}_{F_t} in terms of r and t can be found in Bulmer (1985, pp. 33). By using this result, it is shown that

$$\begin{aligned} \mathbf{H}_{F_t}[1, 1] = \mathbf{H}_{F_t}[3, 3] &= \frac{1}{2^{t-1} - 1} \left[\frac{2^{t-1}}{1+2r} - 1 - \frac{2^{t-1}(\frac{1}{2} - r)^t}{1+2r} + 2^{t-2}[\frac{1}{2} - r(1-r)]^{t-1} \right] \\ \mathbf{H}_{F_t}[1, 2] = \mathbf{H}_{F_t}[3, 2] &= \frac{1}{2^{t-1} - 1} \left[1 - 2^{t-1}[\frac{1}{2} - r(1-r)]^{t-1} \right] \\ \mathbf{H}_{F_t}[1, 3] = \mathbf{H}_{F_t}[3, 1] &= \frac{1}{2^{t-1} - 1} \left[\frac{2^t r}{1+2r} - 1 + \frac{2^{t-1}(\frac{1}{2} - r)^t}{1+2r} + 2^{t-2}[\frac{1}{2} - r(1-r)]^{t-1} \right] \\ \mathbf{H}_{F_t}[2, 1] = \mathbf{H}_{F_t}[2, 3] &= \frac{1}{2} - 2^{t-2}[\frac{1}{2} - r(1-r)]^{t-1} \\ \mathbf{H}_{F_t}[2, 2] &= 2^{t-1}[\frac{1}{2} - r(1-r)]^{t-1} \end{aligned} \quad (12)$$

where $\mathbf{H}_{F_t}[i, j]$ is the i th row and j th column element of \mathbf{H}_{F_t} .

It is easy to check that when $t = \infty$ (recombinant inbred lines),

$$\mathbf{H}_{F_\infty} = \frac{1}{1+2r} \begin{bmatrix} 1 & 0 & 2r \\ 0 & 0 & 0 \\ 2r & 0 & 1 \end{bmatrix}$$

which is the classical result given by Haldane and Waddington (1931). When $t = 2$, \mathbf{H}_{F_2} reduces to (5) for a corresponding interval.

When we use the transition probability matrix (12) in (6) and (7) to obtain $\mathbf{p}_k^{RF_t}$ and $\mathbf{p}_k^{LF_t}$, we need to note that whereas $\mathbf{H}_{F_2}(r_k, k+1) = \mathbf{H}_{F_2}(r_k)\mathbf{H}_{F_2}(r_{k+1})$, $\mathbf{H}_{F_t}(r_k, k+1) \neq \mathbf{H}_{F_t}(r_k)\mathbf{H}_{F_t}(r_{k+1})$ for $t > 2$ even under the assumption of no crossing-over interference because of multiple generations of recombination. Similar results can be observed for the following backcross from selfed F_t and advanced backcross. However, generally $\mathbf{H}_{F_t}(r_k, k+1)$ is very close to $\mathbf{H}_{F_t}(r_k)\mathbf{H}_{F_t}(r_{k+1})$. For example,

$$\mathbf{H}_{F_\infty}(r_k, k+1) = \frac{1}{1+2r_{k,k+1}} \begin{bmatrix} 1 & 0 & 2r_{k,k+1} \\ 0 & 0 & 0 \\ 2r_{k,k+1} & 0 & 1 \end{bmatrix}$$

is very close to

$$\begin{aligned} \mathbf{H}_{F_\infty}(r_k)\mathbf{H}_{F_\infty}(r_{k+1}) &= \frac{1}{(1+2r_k)(1+2r_{k+1})} \\ &\times \begin{bmatrix} 1+4r_k r_{k+1} & 0 & 2r_k + 2r_{k+1} \\ 0 & 0 & 0 \\ 2r_k + 2r_{k+1} & 0 & 1+4r_k r_{k+1} \end{bmatrix} \end{aligned}$$

with $r_{k,k+1} = r_k + r_{k+1} - 2r_k r_{k+1}$ when $r_k r_{k+1}$ is very small. Given $\mathbf{p}_k^{RF_t}$ and $\mathbf{p}_k^{LF_t}$ approximated by (12), $\mathbf{p}_k^{F_t}$ is obtained by (9) with the prior probability vector (10).

When $v \neq u$ and $v = F_{\tilde{t}}$ for $\tilde{t} > t$, we need to derive $\mathbf{M}_{u \rightarrow v}$. Let

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 \end{bmatrix}$$

be the transition probability matrix between generations for three genotypes of a locus by selfing.

$$\mathbf{M}_{F_t \rightarrow F_t} = \mathbf{S}^{t-t} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} - \frac{1}{2^{t-t+1}} & \frac{1}{2^{t-t}} & \frac{1}{2} - \frac{1}{2^{t-t+1}} \\ 0 & 0 & 1 \end{bmatrix}.$$

Random mating F_t

The random mating F_t population is another quite commonly used experimental design for gene mapping. This design has an advantage in separating closely linked QTL. The effect of further random mating on the conditional genotypic distribution \mathbf{H} is equivalent to increasing the recombination fraction between linked markers. Thus, the transition probability matrix for F_t by random mating can be approximated by (5) with r replaced by

$$r^{(t)} = \frac{1}{2} - \frac{1}{2}(1-r)^{t-2}(1-2r) \quad (13)$$

(Falconer and Mackay 1996; Darvasi and Soller 1995). For random mating populations, it is usually necessary to have $u = v$. \mathbf{q}_k^u should remain the same as (3) for large random mating population.

Backcross from selfed F_t

When a selfed F_t population is backcrossed to P_1 or P_2 or test crossed to another inbred line, we need to infer only the gametic type produced by F_t . Let \mathbf{g}'_{F_t} be a row vector of frequencies of the four gametic types [AB , Ab , aB , ab] produced by F_t . Given (11),

$$\mathbf{g}'_{F_t} = \mathbf{p}'_{F_t} \mathbf{C} \quad (14)$$

where

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{w}{2} & \frac{r}{2} & \frac{r}{2} & \frac{w}{2} \\ \frac{r}{2} & \frac{w}{2} & \frac{w}{2} & \frac{r}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and $w = 1 - r$. Because there are only two genotypes in a backcross at a locus, the transition probability matrix equivalent to (12) is

$$\begin{aligned} \mathbf{G}_{F_t} &= \begin{bmatrix} \frac{\mathbf{g}'_{F_t}[1]}{\mathbf{g}'_{F_t}[1]+\mathbf{g}'_{F_t}[2]} & \frac{\mathbf{g}'_{F_t}[2]}{\mathbf{g}'_{F_t}[1]+\mathbf{g}'_{F_t}[2]} \\ \frac{\mathbf{g}'_{F_t}[3]}{\mathbf{g}'_{F_t}[3]+\mathbf{g}'_{F_t}[4]} & \frac{\mathbf{g}'_{F_t}[4]}{\mathbf{g}'_{F_t}[3]+\mathbf{g}'_{F_t}[4]} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1-(\frac{1}{2}-r)^t}{1+2r} + (\frac{1}{2}-r)^t & \frac{2r+(\frac{1}{2}-r)^t}{1+2r} - (\frac{1}{2}-r)^t \\ \frac{2r+(\frac{1}{2}-r)^t}{1+2r} - (\frac{1}{2}-r)^t & \frac{1-(\frac{1}{2}-r)^t}{1+2r} + (\frac{1}{2}-r)^t \end{bmatrix}. \end{aligned} \quad (15)$$

The definition of the two genotypes, however, depends on whether the backcrossed parental population is P_1 or P_2 . Assuming $u = v$, \mathbf{p}_k^u is calculated from (9) using \mathbf{G}_{F_t} in the place of \mathbf{H} in (6) and (7) and with $\mathbf{q}_k^{u'} = [\frac{1}{2}, \frac{1}{2}]$. Since missing and dominant (or recessive) phenotypes are equivalent in backcrosses as far as information provided, consecutive intervals with missing information can be merged and this calculation becomes equivalent to that by Martinez and Curnow (1994) as the backcross population is derived from F_1 .

Advanced backcross BC_t

Advanced backcross is an experimental design that crosses a backcross to a recurrent parental line recursively. This design is proposed by Tanksley and Nelson (1996) as a method of combining QTL mapping with QTL transfer from unadapted germplasm into an elite inbred line.

Let the recurrent parent have a genotype ab/ab in two loci, and let \mathbf{g}'_{BC_t} be a row vector of frequencies of the four gametic types [AB , Ab , aB , ab] produced by an advanced backcross at generation t denoted as BC_t . We know that

$$\mathbf{g}'_{BC_t} = [(1-r)/2, r/2, r/2, (1-r)/2].$$

Because in each generation the recurrent parent always contributes a gamete aa , as in (11) we have

$$\mathbf{g}'_{BC_t} = \mathbf{g}'_{BC_1} \mathbf{T}_{BC}^{t-1} \quad (16)$$

with

$$\mathbf{T}_C = \begin{bmatrix} \frac{w}{2} & \frac{r}{2} & \frac{r}{2} & \frac{w}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Thus the transition probability matrix equivalent to (15) can be obtained with some manipulation as

$$\mathbf{G}_{BC_t} = \begin{bmatrix} (1-r)^t & 1-(1-r)^t \\ \frac{1}{2^t-1}[1-(1-r)^t] & \frac{1}{2^t-1}[(1-r)^t + 2^t - 2] \end{bmatrix}. \quad (17)$$

All other calculations will be the same as above except $\mathbf{q}_k^{BC_t} = [\frac{1}{2^t}, 1 - \frac{1}{2^t}]$.

Backcross from random mating F_t

For the backcrosses from random mating F_t , the transition probability matrix is given by

$$\mathbf{G}_{F_t} = \begin{bmatrix} 1-r^{(t+1)} & r^{(t+1)} \\ r^{(t+1)} & 1-r^{(t+1)} \end{bmatrix} \quad (18)$$

with $r^{(t+1)}$ defined by (13). All other specifications are the same as for the backcrosses from selfed F_t .

Design III

Another special design which has been used extensively in QTL mapping in plants is Design III (e.g., Stuber et al., 1992; Xiao et al., 1995). This design was originally introduced by Comstock and Robinson (1952) for estimating the average degree of dominance for quantitative trait loci. In this design, F_2 or F_t individuals were each backcrossed to both original parental inbred lines P_1 and P_2 . Adapted for QTL mapping, markers are usually genotyped on (selfed) F_t individuals and quantitative traits are measured on the individuals of backcrosses $B_{t1} = F_t \times P_1$ and $B_{t2} = F_t \times P_2$.

Thus the design involves $u = F_t$ and $v = B_{t1}$ or B_{t2} , or more generally $v = B_{\tilde{t}1}$ or $B_{\tilde{t}2}$ for $\tilde{t} \geq t$. Let

$$\mathbf{C}_1 = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{C}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}.$$

$\mathbf{M}_{F_t \rightarrow B_{\tilde{t}1}} = \mathbf{S}^{\tilde{t}-t} \mathbf{C}_1$ and $\mathbf{M}_{F_t \rightarrow B_{\tilde{t}2}} = \mathbf{S}^{\tilde{t}-t} \mathbf{C}_2$. Other terms are the same as for selfed F_t .

As emphasized by Cockerham and Zeng (1996), when both backcrosses are available, QTL mapping analysis should be performed on both backcrosses *simultaneously*.

Marker linkage map reconstruction

In the above analysis, the marker linkage map is assumed to be known. For some organisms such as mouse, this may be safely assumed as so many markers have already been mapped in the mouse genome (e.g., Dietrich et al., 1994). For many other organisms, however, a marker linkage map may have to be reconstructed from the same data for mapping QTL. In this case, the inference of linkage map can be affected by missing marker information and this would in turn affect the inference of distribution of QTL genotype. Here, we investigate the effect of dominant and missing markers on the marker linkage map reconstruction. The algorithm for map reconstruction consists of two parts. One is the search for the best gene order and the other is the evaluation of a given gene order. In this paper, we concentrate our discussion on the likelihood evaluation of gene orders with dominant and missing markers and avoid the discussion of finding optimum ways to search for the best order. We will also not discuss issues involved in the test of marker linkage orders.

Let z_{j1}, \dots, z_{jm} be the phenotypic observations of markers M_1, \dots, M_m for the j th individual in an F_2 population. The markers are under the consideration for a linkage group. The likelihood of the observations depends on the model of linkage events and marker linkage order. Under the assumption of no crossing-over interference and for the order M_1, M_2, \dots, M_m , the likelihood is defined with our matrix notation as

$$\begin{aligned} L &= \prod_{j=1}^n P(z_{j1}, \dots, z_{jm}) \\ &= \prod_{j=1}^n P(x_{j1}) P(z_{j1}, \dots, z_{jm} | x_{j1}) \\ &= \prod_{j=1}^n \mathbf{q}'_1 \mathbf{I}_{z_{j1}} \mathbf{H}_{z_{j2}}(r_1) \mathbf{H}_{z_{j3}}(r_2) \cdots \mathbf{H}_{z_{jm}}(r_{m-1}) \mathbf{c} \end{aligned} \quad (19)$$

where n is the sample size. This likelihood can be used as a criteria for the comparison of different linkage orders.

The likelihood is a function of recombination frequencies between adjacent markers for a given marker linkage order. When there is no missing data in a sample, the recombination frequency can be estimated for each pair of markers independently, and is usually

estimated for all pair-wise of markers before the evaluation of likelihood. However, when some markers are missing or partially missing, estimation of some recombination frequencies depends on marker linkage order and has to be performed for each given order.

For a given linkage order such as M_1, M_2, \dots, M_m , the recombination frequency can be estimated by an EM algorithm and repeatedly updated for each adjacent marker interval in sequence until convergence. Each update will depend on estimates of recombination frequencies at some or all intervals. To update the estimate of recombination frequency for the k th interval, we need to calculate for each individual

$$\begin{aligned} & P(x_k x_{k+1} | z_i, \dots, z_l) \\ &= \frac{P(x_k x_{k+1}) P(z_i \dots z_l | x_k, x_{k+1})}{\sum_{x_k, x_{k+1}} P(x_k x_{k+1}) P(z_i \dots z_l | x_k, x_{k+1})} \\ &= \frac{P(x_k) P(x_{k+1} | x_k) P(z_i \dots z_k | x_k) P(z_{k+1} \dots z_l | x_{k+1})}{\sum_{x_k, x_{k+1}} P(x_k) P(x_{k+1} | x_k) P(z_i \dots z_k | x_k) P(z_{k+1} \dots z_l | x_{k+1})} \end{aligned} \quad (20)$$

which can be expressed in matrix form as

$$\mathbf{A}_k = \frac{[(\mathbf{q}_k \circ \mathbf{p}_k^L) \mathbf{p}_{k+1}^{R'}] \circ \mathbf{H}(r_k)}{(\mathbf{q}_k \circ \mathbf{p}_k^L)' \mathbf{H}(r_k) \mathbf{p}_{k+1}^R}, \quad (21)$$

where \mathbf{A}_k is a 3×3 matrix. Define a_{jcd} as the c th row and d th column element of \mathbf{A}_k for individual j . The recombination frequency for the k th interval can be updated as

$$\begin{aligned} \tilde{r}_k &= \frac{1}{2n} \sum_{j=1}^n \left[a_{j12} + 2a_{j13} + a_{j21} \right. \\ &\quad \left. + \frac{r_k^2}{(1-r_k)^2 + r_k^2} 2a_{j22} + a_{j23} + 2a_{j31} + a_{j32} \right], \end{aligned} \quad (22)$$

where r_k is the estimate in the previous round (see also Jansen & Stam, 1994). We performed a simulation study to examine the effect of dominant and missing markers on linkage map reconstruction. A chromosome of 60 cM is simulated for an F_2 population. We considered the same types of marker densities and compositions as in Table 1 in the comparison. For the marker densities 10 cM and 20 cM, an exhaustive search was conducted for the best-supported linkage order by likelihood. For the marker density 5 cM (with 13 markers), however, the number of possible linkage orders is too many (3.1 billion) to do exhaustive search. In this case,

each marker was added to the map in sequence starting with three well-separated markers. In each addition at least 30 previous orders with the highest likelihoods remained in consideration, and only the orders that differed from the highest order by LOD score 6.0 were removed from the consideration. The sample size was still 150 and the number of replicates was 1000 for each marker density and composition. The results are given in Table 2.

The results indicate that case (b) with 15% data missing has very little effect on the linkage map reconstruction for the parameters considered as compared to case (a). The proportion of correct linkage order

decreases only slightly for case (c), but very substantially for case (d). Apparently, the chance to obtain the correct linkage order for case (e) is low particularly for marker density 5 cM. However, the proportion of intervals with correct flanking markers remains reasonably high. This is good for marker-assisted selection based on QTL mapping. Even though the linkage order is incorrect, the QTL can still have a reasonably high chance to be mapped in the correct marker interval and be selected through flanking markers. Table 2 also shows that when the linkage order is incorrectly inferred, the length of map is usually enlarged (the mean estimated interval size is increased). This implies that, when a large proportion of markers are dominant, without well-separated codominant markers estimation of genome size can be unreliable. Unreliable inference of linkage map can significantly affect QTL mapping, particularly on the precision of estimation of QTL location.

Discussion

Due to the widespread use of PCR technology in genotyping, dominant markers are commonly available for mapping analysis in plants and animals. Although a dominant marker is not as informative as a codominant marker in populations like F_2 , their abundance and

Table 2. Simulation results of linkage map reconstruction (the five marker compositions are the same as in Table 1)

Interval size (cM)	Marker compositions	Proportion of correct order	Proportion of intervals with correct flanking markers	Mean number of break-up point	Estimated mean interval size (SD)
5	(a)	1.00	1.00	0.00	5.0(0.41)
	(b)	0.998	1.00	0.002	5.0(0.42)
	(c)	0.98	0.997	0.032	5.0(0.42)
	(d)	0.90	0.985	0.18	5.0(0.43)
	(e)	0.12	0.70	3.54	5.2(0.86)
10	(a)	1.00	1.00	0.00	10.1(0.88)
	(b)	1.00	1.00	0.00	10.1(0.96)
	(c)	0.998	1.00	0.002	10.1(0.95)
	(d)	0.98	0.99	0.04	10.2(0.98)
	(e)	0.43	0.83	1.03	10.1(1.49)
20	(a)	1.00	1.00	0.00	20.1(2.01)
	(b)	1.00	1.00	0.00	20.2(2.30)
	(c)	0.998	0.999	0.002	20.2(2.53)
	(d)	0.89	0.96	0.13	21.2(4.69)
	(e)	0.41	0.77	0.69	21.4(5.48)

accessibility can be a big advantage for mapping analysis. For some organisms it can be hard to find sufficient number of codominant markers for mapping analysis and dominant markers may have to be used. For example, wheat is a hexaploid, with homoeologous groups of chromosomes, that contains a high proportion of highly repetitive DNA. Genomic variation is also relatively low among varieties in wheat. This makes it hard to find a large number of appropriate probes for RFLP. However, PCR based markers are rich and abundant. Thus, it is important and often necessary to incorporate those markers in mapping analysis.

We derived in this paper a general algorithm to incorporate dominant and missing markers in QTL mapping analysis and also in marker linkage map reconstruction. The method uses Markov chains to link multiple intervals and multiple generations in the calculation of conditional probability of a genotype of an individual at a specific genomic position given the observed relevant marker data of the individual and related individuals. The idea is the same as Lander and Green (1987). This analysis, of course, depends on the assumption of no crossing-over interference. When there is negative crossing-over interference, which is the usual case, the probability of double or triple crossing-over events will be lower than that considered in the algorithm. However, those probabilities are in the lower magnitudes in the analysis, and

the effect of the assumption of no crossing-over interference is likely to be small for the algorithm.

The method is developed for various populations stemming from two inbred lines. The idea and algorithm can also be extended to other types of experimental designs and to pedigrees. With multiple individuals related to an individual in consideration in a pedigree, the marker information of all those individuals can contribute to the inference of the genotype of the individual concerned. It can also be used to calculate the parental genotype distribution at a genomic position given observed marker information of the individual, its offspring, and perhaps its parents as well at multiple markers. This algorithm can be constructed by following the approach taken in this paper and will be proven to be an efficient way in utilizing all relevant marker information in pedigree data analysis.

Our simulation results also point out that it is important to have some codominant markers in a linkage group, although data with only dominant markers can still be analyzed. Those codominant markers can provide accurate anchors in various positions of the genome and can increase the accuracy of the analysis substantially. It is important to combine dominant and codominant markers in mapping analysis, especially when the marker linkage map has to be reconstructed from the data before being used for mapping QTL.

Acknowledgments

This study was supported in part by grant GM 45344 from National Institutes of Health, and a joint project of CIMMYT with University of Hohenheim in Germany funded by Eiselen foundation. We thank Dr. D. Horsington and Dr. González-de-león for many valuable discussions.

References

- Beavis, W.D., O.S. Smith, D. Grant & R. Fincher, 1994. Identification of quantitative trait loci using a small sample of topcrossed and F₄ progeny from maize. *Crop Sci.* 34: 882–896.
- Bulmer, M.G., 1985. *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.
- Cockerham, C.C. & Z.-B. Zeng, 1996. Design III with marker loci. *Genetics* 143: 1437–1456.
- Comstock, R.E. & H.F. Robinson, 1952. Estimation of average dominance of genes, pp. 495–516 in *Heterosis*, edited by J.W. Gowen. Iowa State College Press, Ames, Iowa.
- Darvasi, A. & M. Soller, 1995. Advanced intercross lines, an experimental population to interval mapping of quantitative trait loci. *Genetics* 141: 1199–1207.
- Dietrich, W.F., J.C. Miller, R.G. Steen, M. Merchant, D. Damron, R. Nahf, A. Gross, D.C. Joyce, M. Wessel, R.D. Dredge, A. Marquis, L.D. Stein, N. Goodman, D.C. Page & E. S. Lander, 1994. A genetic map of the mouse with 4,006 simple sequence length polymorphisms. *Nature Genetics* 7: 220–245.
- Falconer, D.S. & T.F.C. Mackay, 1996. *Introduction to Quantitative Genetics*. Edn. 4. Addison Wesley Longman, Harlow, Essex.
- Fisch, R.D., M. Ragot & G. Gay, 1996. A generalization of the mixture model in the mapping of quantitative trait loci for progeny from a bi-parental cross of inbred lines. *Genetics* 143: 571–577.
- Fulker, D.W., S. Cherny & L.R. Cardon, 1995. Multipoint interval mapping of quantitative trait loci using sib pairs, *Am. J. Hum. Genet.* 56: 1224–1233.
- Haldane, J.B.S. & C.H. Waddington, 1931. Inbreeding and linkage. *Genetics* 16: 357–374.
- Haley, C.S., S.A. Knott & J.M. Elsen, 1994. Mapping quantitative trait loci between outbred lines using least squares. *Genetics* 136: 1195–1207.
- Jansen, R.C., 1996. A general Monte Carlo method for mapping multiple quantitative trait loci. *Genetics* 142: 305–311.
- Jansen, R.C. & P. Stam, 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* 136: 1447–1455.
- Lander, E.S. & D. Botstein, 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185–199.
- Lander, E.S. & P. Green, 1987. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci.* 84: 2363–2367.
- Martinez, O. & R.N. Curnow, 1994. Missing markers when estimating quantitative trait loci using regression mapping. *Heredity* 73: 198–206.
- Stuber, C.W., S.E. Lincoln, D.W. Wolff, T. Helentjaris & E.S. Lander, 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Genetics* 132: 823–839.
- Tanksley, S.D. & J.C. Nelson, 1996. Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theor. Appl. Genet.* 92: 191–203.
- Xiao, J., J. Li, L. Yuan & S.D. Tanksley, 1995. Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers. *Genetics* 140: 745–754.
- Zeng, Z.-B., 1994. Precision mapping of quantitative trait loci. *Genetics* 136: 1457–1468.