Genet. Res., Camb. (2009), **91**, *pp.* 85–99. © Cambridge University Press 2009
doi:10.1017/S0016672309000081 Printed in the United Kingdom

A study on the mapping of quantitative trait loci in advanced populations derived from two inbred lines

CHEN-HUNG KAO* AND MIAO-HUI ZENG
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China

(*Received 25 June 2008 and in revised form 2 October 2008 and 29 January 2009*)

Summary

In genetic and biological studies, the F_2 population is one of the most popular and commonly used experimental populations mainly because it can be readily produced and its genome structure possesses several niceties that allow for productive investigation. These niceties include the equivalence between the proportion of recombinants and recombination rates, the capability of providing a complete set of three genotypes for every locus and an analytically attractive first-order Markovian property. Recently, there has been growing interest in using the progeny populations from F_2 (advanced populations) because their genomes can be managed to meet specific purposes or can be used to enhance investigative studies. These advanced populations include recombinant inbred populations, advanced intercrossed populations, intermated recombinant inbred populations and immortalized F_2 populations. Due to an increased number of meiosis cycles, the genomes of these advanced populations no longer possess the Markovian property and are relatively more complicated and different from the F_2 genomes. Although issues related to quantitative trait locus (QTL) mapping using advanced populations have been well documented, still these advanced populations are often investigated in a manner similar to the way F_2 populations are studied using a first-order Markovian assumption. Therefore, more efforts are needed to address the complexities of these advanced populations in more details. In this article, we attempt to tackle these issues by first modifying current methods developed under this Markovian assumption to propose an *ad hoc* method (the Markovian method) and explore its possible problems. We then consider the specific genome structures present in the advanced populations without invoking this assumption to propose a more adequate method (the non-Markovian method) for QTL mapping. Further, some QTL mapping properties related to the confounding problems that result from ignoring epistasis and to mapping closely linked QTL are derived and investigated across the different populations. Simulations show that the non-Markovian method outperforms the Markovian method, especially in the advanced populations subject to selfing. The results presented here may give some clues to the use of advanced populations for more powerful and precise QTL mapping.

1. Introduction

Many quantitative trait loci (QTLs) detection experiments and statistical QTL mapping methods are conducted and developed on the basis of the backcross and F_2 populations. These two populations are popular mainly for economic reasons as they can be readily generated for use in experiments, thus saving time and money. Further, due to the fact that these populations undergo just a single cycle of meiosis, they have several significant features that make them attractive for general purpose genetic and biological studies (Lander & Botstein, 1989; Jansen, 1993; Zeng, 1994; Jiang & Zeng, 1997; Kao *et al.*, 1999; Xu, 2007). For example, the recombination rate between different loci is equivalent to the proportion of the recombinants and their genomes have a first-order Markovian property in the two populations. Also, the progeny

* Corresponding author: Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China. Tel: 886-2-2783-5611 ext. 418 Fax. 886-2-2783-1523. e-mail: chkao@stat.sinica.edu.tw

populations after F_2 (advanced populations) have been well devised and implemented in genetic studies. These advanced populations include recombinant inbred (RI) populations, advanced intercrossed (AI) populations, intermated recombinant inbred (IRI) populations and immortalized F_2 populations. For a review of these advanced populations, see e.g. Rockman & Kruglyak (2008).

These advanced populations have some very useful features in that their genomic structures allow investigators to achieve better performance in their studies. For example, the RI populations consist of nearly fixed genomes for multiple phenotyping and contain a specific genotype to increase the accuracy of assessment in studying quantitative traits (Lander & Botstein, 1989). Further, the AI populations can harbour more recombination events in a short chromosome segment for genetic fine mapping (Darvasi, 1998). Also, the IRI and RIX (recombinant inbred intercrosses) populations can be managed to have both the advantages of RI and AI populations (Liu *et al.*, 1996; Hua *et al.*, 2002; Winkler *et al.*, 2003; Zou *et al.*, 2005).

The derivation of the RI populations or AI populations is obtained by recurrently selfing (inbreeding) or randomly intermating the F_2 individuals for several generations. The IRI populations are derived by first producing AI populations, followed by repeated selfing. The immortalized F_2 populations are obtained by first producing RI populations, followed by a generation of random mating. As a generation advances beyond F_2, either by further selfing or intermating, the advanced populations must undergo multiple cycles of meiosis, so that the crossovers will accumulate and the proportions of recombinants will increase in the populations (Haldane & Waddington, 1931; Liu *et al.*, 1996; Darvasi, 1998; Winkler *et al.*, 2003). In the literature, it has been noted that the proportion of recombinants in RI populations can be twice that in the F_2 populations for closely linked loci, and that linkage is broken down even more rapidly by random intercrossing in the AI populations (Haldane & Wanddington, 1931; Darvasi, 1998). The increased number of recombinants provided by the advanced populations facilitates the construction of high-resolution genetic maps and detection of closely linked QTLs (Liu *et al.*, 1996; Darvasi, 1998). Further, cycles of inbreeding and/or random mating in a population will shape differences in the population genomic structures such as the homozygosity, genotypic frequencies and variance components (Weir, 1996). As such, different advanced populations produce different genomic structures to be used for different breeding and study purposes (Liu *et al.*, 1996; Hua *et al.*, 2002; Winkler *et al.*, 2003; Broman, 2005).

When using these advanced populations for QTL mapping, it should be noted that their genome structures no longer have a first-order Markovian property and have different genomic constitutions from that of the F_2 populations (Jiang & Zeng, 1997). So far, most of the current QTL mapping methods and related mapping properties are developed and investigated for the genomes of backcross and F_2 populations with the Markovian property (Lander & Botstein, 1989; Jansen, 1993; Churchill & Doerge, 1994; Zeng, 1994; Kao *et al.*, 1999; Kao & Zeng, 2002; Kao, 2004; Xu, 2007). Although issues related to using advanced populations in QTL mapping have been raised (Jiang & Zeng, 1997; Darvasi, 1998; Martin & Hospital, 2006), they are still investigated by invoking this Markovian assumption. It is therefore desirable to consider the specific structures of these advanced populations for QTL detection, so that their advantages can be utilized to enhance QTL resolution. In this paper, detailed analyses and discussions related to these advanced populations will be given. When samples are drawn from the advanced populations, statistical methods are developed by considering and ignoring their specific population genome structures (without and with a first-order Markovian assumption) and are compared for use with the multiple-QTL model for use in QTL mapping studies. In addition, the QTL mapping properties across different advanced populations are derived and discussed. Simulation studies are performed for purposes of evaluation and comparison. The results show that the proposed methods can improve the resolution of the genetic architecture of quantitative traits and serve as a tool for studying QTL mapping in various advanced populations derived from two inbred lines.

2. The genome structures of advanced populations

We refer an AI (RI) F_t population as an AI (RI) population from intercrossing (selfing) the F_2 individuals for $t-2$, $t>3$, generations. An IRI $F_{i:j}$ population is referred to as a population produced by first randomly intercrossing the F_2 individuals for $i-2$ generations, followed by j, $j \geqslant 1$, cycles of selfing, and an IF_2 population denotes an immortalized F_2 population.

(i) *Genome structure*

In an F_2 population, the genotypic frequencies of P_1 homozygote, heterozygote and P_2 homozygote are 1/4, 1/2 and 1/4, respectively, for one locus, and the heterozygosity H_t is 0·5. The genotypic distribution for any two pairwise loci, say A and B, is also well known and characterized (see, for example, Kao & Zeng, 1997), and it has a simple relationship with the recombination rate between them (r). For example, the genotypic frequency of genotype AB/AB is $(1-2r)^2/4$, and the other nine genotypic frequencies also have similar simple relationships with r (see, for example, Table 2 in Kao & Zeng, 1997). Also, the proportion of recombinants (R) between A and B is equivalent to the recombination rate, i.e. $R=r$, and the linkage parameter between A and B can be found to be $\lambda=1-2r$ in the population. Besides, a very important and nice feature for the F_2 population is that the F_2 genomes have a first-order Markovian structure under the Haldane map function. This allows that the distribution of the multiple genes can be obtained from the distributions of pairwise genes. For example, the probability distribution of three ordered genes, A, B and C, can be derived from the probability distributions of first pairwise genes, A and B, and the second pairwise genes, B and C, i.e. $P(\mathrm{ABC})=P(\mathrm{AB})\times P(\mathrm{BC})$.

The heterozygosity for one locus in the RI F_t and IRI $F_{i:j}$ populations are $\frac{1}{2^{t-1}}$ and $\frac{1}{2^{j+1}}$, which is decreasing with t and j increasing, as selfing will increase the homozygotes at the expense of heterozygotes, and it is expected to be $H_t=0{\cdot}5$ in the AI F_t and IF_2 population (RIX) populations for any t due to random mating. Also, during the process of further meiosis, crossovers will accumulate so that the proportion of recombinants will be increasing and becoming larger than the recombination rate ($R>r$), and the linkage disequilibrium coefficient will decrease. To generally formulate these genetic parameters, we adopt the notations in Haldane & Waddington (1931) to define C as the frequencies of AB/AB and ab/ab genotypes, D as the frequencies of Ab/Ab and aB/aB genotypes, E as the frequencies of AB/Ab, AB/aB, Ab/ab and aB/ab genotypes, F as the frequency of AB/ab genotype and G as the frequency of Ab/aB genoytype, respectively, for any two loci A and B, and they in terms of C, D, E, F, and G are

$$H=2E+F+G, \quad R=2D+2E+G \quad \text{and}$$

$$D_{\mathrm{AB}}=(C+E+\tfrac{1}{2}F)-\tfrac{1}{4},$$

respectively, in any advanced population. In the F_2 population, the frequencies C, D, E, F, G in terms of r are $C=(1-r)^2/4$, $D=(1-2r)/4$, $E=r(1-r)/2$, $F=(1-r)^2/2$ and $G=r^2/2$, which have simple relations with r, and $H=1/2$, $R=r$ and $D_{\mathrm{AB}}=(1-2r)/4$. In advanced populations, these values in terms of r become relatively complicated and will vary with different t, i and j, and they can be obtained without difficulty (Jennings, 1916; Robbins, 1918; Haldane & Waddington, 1931; Winkler *et al.*, 2003). The more important and challenging parts in this QTL mapping context under the framework of interval mapping procedure are to characterize the genotypic distributions of three loci for various advanced populations, whose genomes do not have a first-order Markovian property.

3. Methods

(i) *Data structure*

Consider a sample of size n from an advanced population, such as AI, RI, IRI or IF_2 population, derived from two inbred lines. The n individuals are genotyped for markers (X_i, $i=1, 2, \ldots, n$) and phenotyped for traits (y_i's, $i=1, 2, \ldots, n$). When such a sample is used to detect QTL, two approaches under the framework of the interval mapping procedure are proposed here. The approach developed under the Markovian assumption will be hereinafter called the Markovian method, and the approach developed without the Markovian assumption will be hereinafter referred to as the non-Markovian method.

(ii) *Genetic model and variance components*

Consider that a trait is controlled by m QTLs, Q_1, Q_2, …, Q_m, and there are 3^m possible QTL genotypes. For any individual i, its QTL genotype belongs to one of the 3^m genotypes, and the corresponding genotypic values, G_i's, can be expressed as

$$\begin{aligned} G_i=\mu+\sum_{j=1}^{m} a_j x_{ij}^{*}+\sum_{j=1}^{m} d_j z_{ij}^{*}+\sum_{j<k} (i_{aa})_{jk}(x_{ij}^{*}x_{ik}^{*}) \\ +\sum_{j<k} (i_{ad})_{jk}(x_{ij}^{*}z_{ik}^{*})+\sum_{j<k} (i_{da})_{jk}(z_{ij}^{*}x_{ik}^{*}) \\ +\sum_{j<k} (i_{dd})_{jk}(z_{ij}^{*}z_{ik}^{*}), \end{aligned} \qquad (1)$$

where μ is the intercept, a_{j} and d_j are the additive and dominance effects of Q_j, $j=1, 2, \ldots, m$, and $(i_{aa})_{jk}$, $(i_{ad})_{jk}$, $(i_{da})_{jk}$ and $(i_{dd})_{jk}$ are additive × additive, additive × dominance, dominance × additive, and dominance × dominance interaction effects between Q_j and Q_k. The variables, x_{ij}^{*} and Z_{ij}^{*}, associated with a_j and d_j are coded as $(1,-1/2)$, $(0,1/2)$ and $(-1,-1/2)$ for genotypes Q_jQ_j, Q_jq_j and q_jq_j, respectively, according to Cockerham's model (Kao & Zeng, 2002). Under the genetic model (1), the genetic variances of a quantitative trait can be generally decomposed into $2m^2$ variances and $2m^4-m^2$ covariances. In practice, the variance component structure will be simpler in the advanced populations as some covariances vanish due to equal frequencies of the two alleles at any locus. Taking $m=2$ as an example, the genetic variance components are

$$\begin{aligned} V_G=2(C+D+E)(a_1^2+a_2^2)+\tfrac{1}{4}[1-(1-4E-2F-2G)^2] \\ \times(d_1^2+d_2^2)+2[C+D-2(C-D)^2]i_{aa}^2+4(C-D)a_1a_2 \\ +\tfrac{1}{2}(C+D+E)(i_{ad}^2+i_{da}^2)+\tfrac{1}{16}[1-(1-8E)^2]i_{dd}^2 \\ +\tfrac{1}{4}[1-8E-(1-4E-2F-2G)^2]d_1d_2 \\ -(C-D)(4E+2F+2G)(d_1i_{aa}+d_2i_{aa}) \\ +(E-C-D)(a_1i_{ad}+a_2i_{da})-(C-D)(a_2i_{ad}+a_1i_{da}) \\ +\tfrac{1}{2}(C-D)i_{ad}i_{da}-4E(C-D)i_{aa}i_{dd} \\ -E(1-4E-2F-2G)(d_1i_{dd}+d_2i_{dd}). \end{aligned} \qquad (2)$$

https://doi.org/10.1017/S0016672309000081 Published online by Cambridge University Press

The component structures allow us to investigate some QTL mapping properties. For example, the additive (dominance) variances are found to increase (decrease) in the RI or IRI population, showing that these populations may facilitate (hinder) the estimation of the additive (dominance) effects (Kao, 2006). Also, the possible confounding problems in QTL estimation may be identified from the covariances between genetic effects (Kao & Zeng, 2002; Kao, 2006). If the two-locus model is expressed as a model of 15 parameters to distinguish each allelic effect, the genetic variance becomes even more complicated (Weir & Cockerham, 1977).

(iii) *Markovian and non-Markovian methods*

With the genetic model in eqn (1), the statistical model to relate a quantitative trait value, y, to the genotypic value, G, contributed from the m QTLs at positions, p_1, p_2, …, and p_m can be written as

$$y_i = G_i + \varepsilon_i, \tag{3}$$

where ε_i is the environmental deviation and assumed to follow normal distribution with mean zero and variance σ^2. In QTL mapping, the QTLs are usually assumed be located in the intervals and need to be estimated, so that the 3^m genotypes, (x_{ij}^* and z_{ij}^*), may not be observed, and the model becomes a normal mixture model. For n individuals, the likelihood function for θ can be generally expressed as

$$L(\theta \mid \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{3^m} p_{ij} N(\mu_j, \sigma^2) \right], \tag{4}$$

where the mixing proportions, p_{ij}'s, $j = 1, 2, \ldots, 3^m$, are the conditional probabilities of the putative QTL genotypes given marker genotypes, and μ_j's, $j = 1, 2, \ldots, 3^m$, correspond to the genotypic values of the 3^m different QTL genotypes. Using the interval mapping procedure (Lander & Botstein, 1989), the conditional probabilities can be predetermined by successively and jointly using the flanking markers of the putative QTL; hence they need not to be estimated. The parameters θ involved in the statistical estimation of the normal mixture model are μ, σ^2, a_i's, d_i's, i_{aa}'s, i_{ad}'s, i_{da}'s and i_{dd}'s. Especially, it should be pointed out that the derivation of the conditional probabilities for each putative QTL using its flanking markers is not straightforward in the advanced populations as has been done for the F_2 and backcross populations (see below). When m putative QTLs are considered at a time, the joint conditional probability is approximated by the product of m individual conditional probabilities. In the following, we propose two QTL mapping methods for the advanced populations under eqn (3). The one using the conditional probabilities derived from a first-order Markovian assumption as the mixing proportions will be called the Markovian method hereafter, and the other using the conditional probabilities obtained without this assumption (by using the proposed transition equations) as mixing proportions will be called the non-Markovian method hereafter.

(iv) *Conditional probabilities of the putative QTL genotypes*

The interval mapping approach intends to compute the conditional probabilities of a putative QTL by using the information from its two flanking markers. Set M with alleles M and m, Q with alleles Q and q and N with alleles N and n, where Q is the putative QTL, and M and N are the flanking markers, and assume that r, r_1 and r_2 are the recombination rates between M and N, between M and Q and between Q and N. To derive the conditional probability of the QTL genotype within the flanking marker genotype, $P(\mathrm{Q} \mid \mathrm{M}, \mathrm{N}) = P(\mathrm{MQN})/P(\mathrm{MN})$, for a population, both the genotypic distributions of two and three genes under generations of selfing or/and random mating are needed. The genotypic distribution of two genes, $P(\mathrm{MN})$, under random mating and self has been very well known (Jennings, 1916; Robbins, 1918; Haldane & Waddington, 1931). For the F_2 population, the derivation of the genotypic distribution for three genes, $P(\mathrm{MQN})$, is simple and can be obtained by using the probabilities of two adjacent pairwise genes, $P(\mathrm{MQ})$ and $P(\mathrm{QN})$, as its genomes have a first-order Markovian property. That is, $P(\mathrm{MQN}) = P(\mathrm{M})P(\mathrm{Q} \mid \mathrm{M})P(\mathrm{N} \mid \mathrm{Q}, \mathrm{M}) = P(\mathrm{M})P(\mathrm{Q} \mid \mathrm{M})P(\mathrm{N} \mid \mathrm{Q})$, as $P(\mathrm{N} \mid \mathrm{Q}, \mathrm{M}) = P(\mathrm{N} \mid \mathrm{Q})$. However, for advanced populations, this Markovian property disappears so that the genotypic distribution of three genes cannot be obtained directly from the distributions of two genes, i.e. by simply replacing the recombination rates (r_1, r_2 and r) by frequencies of recombinants (R_1, R_2 and R) as suggested by Jiang & Zeng (1997) and Lynch & Walsh (1998). For example, it is suggested to approximate the two conditional gametic frequencies by $\Pr(Mqn \mid Mn) \approx R_1(1-R_2)/R$ and $\Pr(MQn \mid Mn) \approx (1-R_1)R_2/R$ in an advanced population. Such a replacing implicitly assumes that the genomes of the advanced populations still have a first-order Markovian property and, therefore, the obtained frequencies are approximate. Another obvious yet often unnoticed problem for this replacing is that the sum of the approximate probabilities may not be equal to one as the Haldane map function does not hold for the R ($R \neq R_1 + R_2 - 2R_1R_2$) in the advanced populations. Appropriate correction is needed when using these approximate probabilities. In this article, correction will be made by dividing the approximate probabilities by their sum. The derivation of

https://doi.org/10.1017/S0016672309000081 Published online by Cambridge University Press

the exact genotypic distribution for three genes needs more delicate considerations as provided below.

The derivation of the genotypic frequencies of three genes for the advanced populations needs to consider two different types of mating systems: random mating and selfing. When mating is random, the frequency of a zygotic genotype is the product of two gametic frequencies in the previous population, and the focus is on deriving the transition equations for the frequencies of eight different gametic types from generation to generation. For example, in AI F$_t$, the probability of $\underline{MQN}$ ($\underline{mqn}$) gamete, $P_{1,t}$, can be generally obtained as

$$\begin{aligned} P_{1,t} = {} & [(1-r_1)(1-r_2)+1]P^2_{1,t-1} + r_1 r_2 P^2_{2,t-1} \\ & + [(1-r_1)r_2]P^2_{3,t-1} + [r_1(1-r_2)]P^2_{4,t-1} \\ & + [1+(1-r_1)(1-r_2)+r_1 r_2]P_{1,t-1}P_{2,t-1} \\ & + (2-r_1)P_{1,t-1}P_{3,t-1} + (2-r_2)P_{1,t}P_{4,t-1} \\ & + r_2 P_{2,t-1}P_{3,t-1} + r_1 P_{2,t}P_{4,t-1} \\ & + [r_1(1-r_2)+r_2(1-r_1)]P_{3,t-1}P_{4,t-1}, \end{aligned} \tag{5}$$

where $P_{2,t-1}$ is the frequency of $\underline{MqN}$ ($\underline{mQn}$) gamete, $P_{3,t-1}$ is the frequency of $\underline{MQn}$ ($\underline{mqN}$) gamete, and $P_{4,t-1}$ is the frequency of $\underline{mQN}$ ($\underline{Mqn}$) gamete in the previous population. An alternative iteration equation for $P_{1,t}$ can be derived by using Geiringer's formulation (1944). If the population is self-fertilized, the gametes of an individual are randomly mating within the individual and are not allowed to seminate the gametes from different individuals, and the focus is on deriving the transition equations for the frequencies of 36 different zygotes from generation to generation. For example, in RI F$_t$ population, the probability of $\frac{MQN}{MQN}$ zygote is

$$\begin{aligned} P_t\left(\frac{MQN}{MQN}\right) = {} & P_{t-1}\left(\frac{MQN}{MQN}\right) + \frac{1}{4}\left[P_{t-1}\left(\frac{MQN}{MqN}\right) + P_{t-1}\left(\frac{MQN}{MQn}\right) + P_{t-1}\left(\frac{MQN}{mQN}\right)\right] \\ & + \frac{(1-r_2)^2}{4}P_{t-1}\left(\frac{MQN}{Mqn}\right) + \frac{r_2^2}{4}P_{t-1}\left(\frac{MqN}{MQn}\right) + \frac{r_1^2 r_2^2}{4}P_{t-1}\left(\frac{MqN}{mQn}\right) \\ & + \frac{(1-r_1)^2}{4}P_{t-1}\left(\frac{MQN}{mqN}\right) + \frac{r_1^2}{4}P_{t-1}\left(\frac{MqN}{mQN}\right) + \frac{[r_1(1-r_2)^2]}{4}P_{t-1}\left(\frac{Mqn}{mQN}\right) \\ & + \frac{[(1-r_1)(1-r_2)+r_1 r_2]^2}{4}P_{t-1}\left(\frac{MQN}{mQn}\right) + \frac{[(1-r_1)r_2]^2}{4}P_{t-1}\left(\frac{mqN}{MQn}\right) \\ & + \frac{[r_1(1-r_2)+r_2(1-r_1)]^2}{4}P_{t-1}\left(\frac{MQn}{mQN}\right) + \frac{[(1-r_1)(1-r_2)]^2}{4}P_{t-1}\left(\frac{MQN}{mqn}\right). \end{aligned} \tag{6}$$

Similarly, the other transition equations for the three gamete frequencies under random mating and for the 35 zygote frequencies under selfing can be obtained (see Supplementary material). By jointly using these transition equations, it is sufficient to obtain the gamete or genotypic frequencies to calculate all conditional probabilities for various fixed and unfixed advanced populations subject to different cycles of random mating and/or self. Teuscher & Broman (2007) developed an alternative technique by solving a set of linear equations to obtain the unknown trigenic haplotype (gametic) probabilities for fixed RIL populations.

The differences between the conditional probabilities of QTL genotypes given marker genotypes obtained with and without a first-order Markovian assumption can be very significant and in turn can have a substantial impact on QTL mapping (see below). Numerical investigation of their differences for *QQ*, *Qq* and *qq* genotypes given the marker genotype *MN*/*MN* for the case of $r_1 = r_2 = 0{\cdot}1$ in AI F_t, RI F_t, IRI $F_{10,t}$ and RIX $F_{10,t}$ populations is shown in Figs 1 *a–d* for illustration. For AI F_t populations, the differences are generally very minor (the differences are within ~0·01; see Figure 1(a)). All three curves are below zero, implying that the probabilities of QTL genotypes are underestimated by the Markovian assumption. The differences between the conditional probabilities become more significant (between ~−0·06 and 0·07; see Figure 1(b)) in RI F_t populations as compared with those in the AI F_t populations. Such differences are increasing at the first few generations of selfing and become stable on proceeding further. For IRI $F_{10,t}$ populations, the differences are very significant (between ~−0·2 and 0·4) and increase as the selfing cycle increases. For RIX $F_{10,t}$ populations, the differences are greatly reduced by intercrossing. In general, persistent selfing tends to enlarge their differences, and continuous intercrossing eventually mitigates their differences. The method with the Markovian assumption also overestimates the frequency of *Qq* and underestimates the other two frequencies during selfing. The sums of the three conditional probabilities are about 0·962–0·980, 0·977–0·995, 0·976–0·991 and 0·964–0·980, respectively,

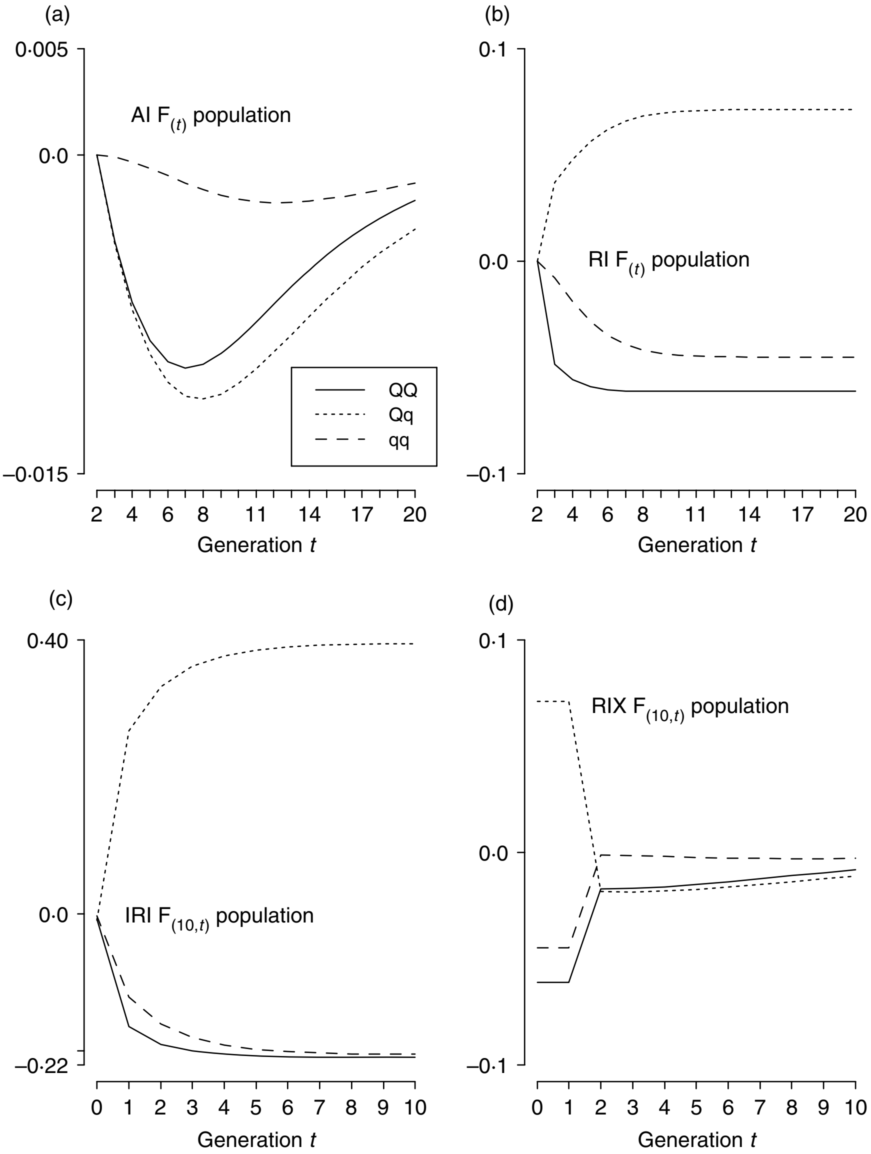

Fig. 1. The differences between the conditional probabilities of QQ, Qq and qq genotypes given the flanking marker genotype MN/MN obtained by using the Markovian and non-Markovian methods for the case of $r_1 = 0{\cdot}1$ and $r_2 = 0{\cdot}1$ in the AI, RI, IRI and RIX populations. The curve below zero implies that the probabilities of QTL genotypes are underestimated by using the Markovian method. (a) AI populations. (b) RI populations. (c) IRI $F_{10,t}$ populations. (d) RIX $F_{10,t}$ populations.

in the RI, AI, IRI and RIX populations. Figures 2*a–d* show the numerical differences in conditional probability for *QQ*, *Qq* and *qq* genotypes given the marker genotype *MN*/*Mn*. More significant differences are observed in the *Mn*/*Mn* class, and the sum of the conditional probabilities may be up to 1·125 (not shown). Therefore, it is important to compute the correct conditional probabilities of the putative QTL genotypes, as they serve as the mixing proportions of the normal mixture model in QTL mapping. The problem of using incorrect (approximate) conditional probabilities of QTL genotypes includes the loss of power and precision in QTL detection as mentioned by Martin & Hospital (2006) and shown in this paper (see the Simulation study section).

(v) *Maximum likelihood estimation*

In parameter estimation, it is straightforward to treat the normal mixture model in eqn (4) as an incomplete-data problem by regarding the trait, **Y**, and markers, **X**, as observed data and the putative QTLs, x^*_{ij}'s and z^*_{ij}'s, as missing data, then the EM algorithm (Dempster *et al.*, 1977) can be readily implemented to obtain their maximum likelihood estimates (MLEs). Alternatively, the marker genotypes and the unknown QTL genotypes can be treated as the observed state

https://doi.org/10.1017/S0016672309000081 Published online by Cambridge University Press

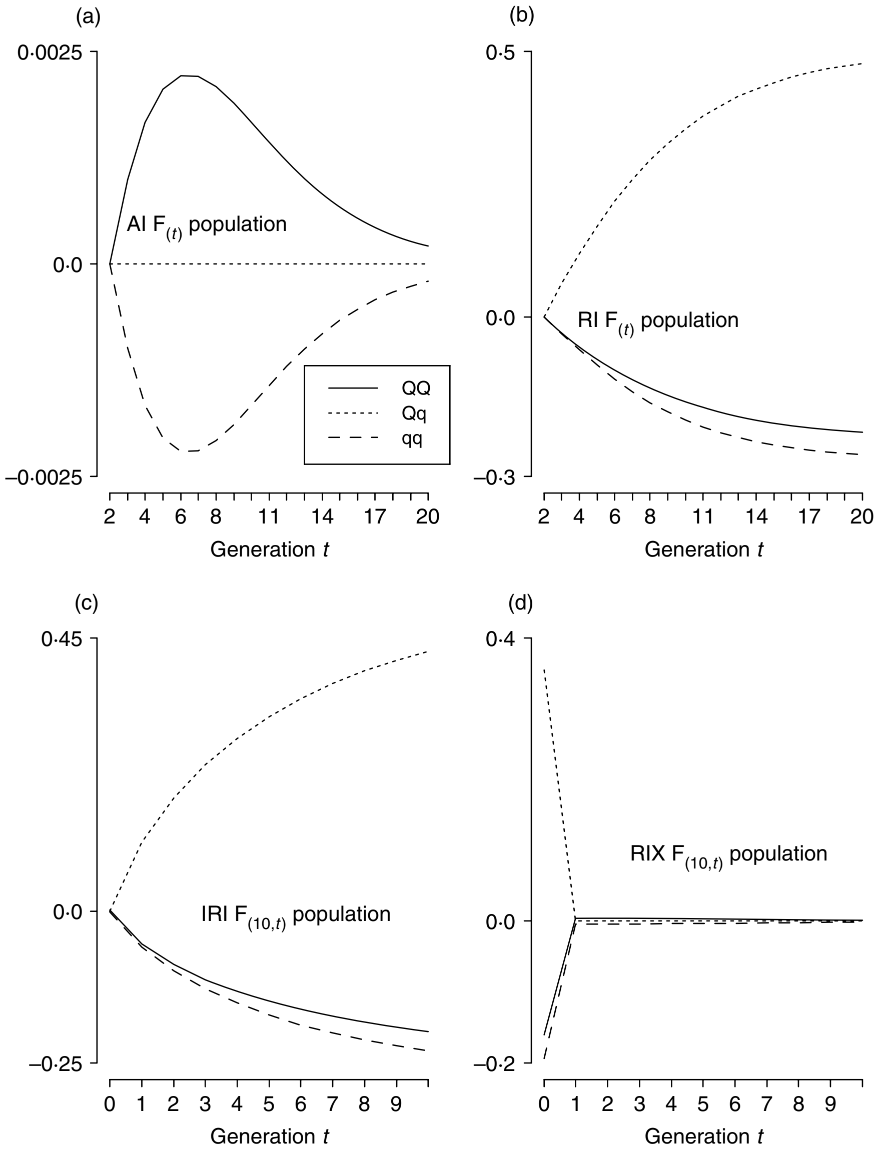

Fig. 2. The differences between the conditional probabilities of QQ, Qq and qq genotypes given the flanking marker genotype MN/Mn obtained by using the Markovian and non-Markovian methods for the case of $r_1=0{\cdot}1$ and $r_2=0{\cdot}1$ in the AI, RI, IRI and RIX populations. The curve below zero implies that the probabilities of QTL genotypes are underestimated by using the Markovian method. (a) AI populations. (b) RI populations. (c) IRI $F_{10,t}$ populations. (d) RIX $F_{10,t}$ populations.

and hidden state in the set-up of the hidden Markov model (HMM; Koski, 2001) under the Markovian assumption along the genome. The EM algorithm is an iterated procedure and, in each iteration, it consists of an expectation step (E-step), followed by a maximization step (M-step). When applying the EM algorithm, the general formulae devised by Kao & Zeng (1997) can be implemented to obtain the MLE applied here. The E-step is to compute the posterior probabilities of 3^m QTL genotypes. In M-step, the coded variables associated with the m QTLs in all the 3^m possible genotypic values are assigned to the elements of genetic design matrix. The E- and M-steps are iterated until convergence, and the converged values are the MLEs.

(vi) *QTL mapping properties*

To investigate and explore QTL mapping properties across populations, without loss of generality, assume that the quantitative trait is affected by the two linked epistatic QTLs, Q_A and Q_B, with complete effects. We consider the scenarios of using Q_A only and of using both Q_A and Q_B in the quantitative trait analysis. If the quantitative trait is regressed on Q_A only, the regression coefficient for the additive effect of Q_A is

$$a_A = a_1 + \frac{C-D}{C+D+E}a_2 + \frac{E-(C+D)}{2(C+D+E)}i_{ad} - \frac{C-D}{2(C+D+E)}i_{da}, \qquad (7)$$

https://doi.org/10.1017/S0016672309000081 Published online by Cambridge University Press

in an advanced population. Similarly, the regression coefficient for the dominance effect of Q_A, d_A, and the partial regression coefficient for the additive (dominance) effect of Q_A given the additive (dominance) effect of Q_B, $a_{A.B_a}$ ($d_{A.B_a}$), can be derived and their components are shown in Table 1. By analysing the coefficients, it is possible to decompose the regression coefficient into components and to trace the changes of these components for identifying the confounding problems as the population advances. Taking Eqn (7) as an example, under selfing, the coefficient associated with a_2 (i_{da}) is positive (negative) and decreasing (increasing) from $1-2r(-(1-2r)/2)$ to $\frac{1-2r}{1+2r}\left(-\frac{1-2r}{2(1+2r)}\right)$, and the coefficient associated with i_{ad} is negative and decreasing from $-(1-2r)^2/2$ to $-1/2$, as generation proceeds (t increases). For $t\to\infty$ under self, $a_A=a_1+\frac{1-2r}{1+2r}a_2-\frac{i_{ad}}{2}-\frac{1-2r}{2(1+2r)}i_{da}$. If mating is random, the coefficient can be generally expressed as $a_A=a_1+(1-2r)\,(1-r)^t a_2-\frac{1}{2}(1-2r)^2(1-r)^{2t}i_{ad}-\frac{1}{2}(1-2r)\,(1-r)^t a_2$. The coefficients associated with a_2, i_{ad} and i_{da} approach to zero as $t\to\infty$. Such analyses make it possible to clearly identify how the different genotypes and effects play a role in the confounding problem across populations. In general, the confounding problem generally becomes less severe as the generation proceeds under random mating. Under selfing, the confounding of i_{ad} becomes more severe and the confounding of i_{da} becomes less severe in the estimation of additive effects of Q_A as generation proceeds. The confounding of the i_{dd} becomes more severe, and i_{aa} will be always confounded in the estimation of the dominance effects as generation proceeds by selfing.

Table 1. *The components of the regression coefficient and partial regression coefficient*

	a_1	d_1	a_2	d_2	i_{aa}	i_{ad}	i_{da}	i_{dd}
a_A	1		$\frac{C-D}{C+D+E}$			$\frac{E-(C+D)}{2(C+D+E)}$	$-\frac{C-D}{2(C+D+E)}$	
d_A		1		$\frac{1-8E-(1-4E-2F-2G)^2}{1-(1-4E-2F-2G)^2}$	$-\frac{4(C-D)(4E+2F+2G)}{1-(1-4E-2F-2G)^2}$			$-\frac{4E(1-4E-2F-2G)}{1-(1-4E-2F-2G)^2}$
a_{A,B_a}	1					$\frac{E^2-4CD}{2(2C+E)(2D+E)}$	$-\frac{E(C-D)}{(2C+E)(2D+E)}$	
d_{A,B_d}		1			$-\frac{2(C-D)(4E+2F+2G)}{[1-(1-4E-2F-2G)^2]-4E}$			$-\frac{2E(1-4E-2F-2G)}{[1-(1-4E-2F-2G)^2]-4E}$

Assume that the quantitative trait is controlled by two QTLs, Q_A and Q_B. a_1 and d_1 (a_2 and d_2) are the additive and dominance effects of Q_A (Q_B). i_{aa}, i_{ad}, i_{da} and i_{dd} are their epistatic effects.

a_A (d_A) is the regression coefficient for the additive (dominance) effect of Q_A, and $a_{A.B_a}$ ($d_{A.B_d}$) is the partial regression coefficient for the additive (dominance) effect of Q_A given the additive (dominance) effect of Q_B.

(vii) *Power of separating closely linked QTL*

To simplify the discussion, we first consider that two linked QTLs with additive effects, a_1 and a_2, only are located at known markers; then the QTL mapping model in eqn (3) reduces to a regression model fitting two correlated variables, x_{i1}^* and x_{i2}^*. As derived above, the correlation between x_{i1}^* and x_{i2}^* is equivalent to the linkage parameter between the two QTLs, $\lambda=(C-D)/(C+D+E)$, which can be interpreted as a measure of the difference between the recombinant (D) and non-recombinant proportions (C) in a population. We can expect that the linkage parameters will decrease for farther genes or in later populations as there are more recombinants and less non-recombinants in either case. In a statistical modelling, fitting correlated variables into the model will raise the problems of collinearity, e.g. inflated variances of $\hat{a}_1$ and $\hat{a}_2$, in estimation and testing (Marquardt, 1970), leading to the difficulty in obtaining simultaneously significant tests for QTL effects (successful separation of linked QTLs). For example, in the AI F_t population (under the process of random mating), $C+D+E=1/4$ and $C-D=(1–2r')/4$, where $r'=[1-(1-2r)$

$(1-r)^{t-2}]/2$, so that $\lambda=1-2r'$ is decreasing with t, and the decreasing rate of λ is $1-r$ for each generation of random mating. Under self, λ is also decreasing, but with a much lower rate. In RIL, $\lambda=(1-2r)/(1+2r)$, which is smaller than $(1-2r)$ in the F_2. In general, the linkage parameter is decreasing and the collinearity problem can be eased in the advanced population. As a consequence, the separation of closely linked QTLs can be more powerful by using the sample from the advanced population, especially from the population subject to several cycles of random mating.

4. Simulation studies

Simulations were conducted to evaluate the performances of the non-Markovian and Markovian methods, to validate the derived mapping properties and to compare relative efficiencies of using different advanced populations in QTL mapping. A large set of fixed and unfixed populations, including RI, AI, IRI and IF_2 populations, was simulated as they are very popular in biological studies (Lee *et al*., 2002; Rockman & Kruglyak, 2008). For RI and AI populations, F_3, F_4, F_5 and F_{10} populations were simulated. For IRI and RIX populations, IRI $F_{5:1}$, $F_{5:3}$ and IF_2 populations were simulated. For each population, two linked epistatic QTLs, Q_A ans Q_B, with complete effects $a_1=2$, $d_1=2$, $a_2=2$, $d_2=2$, $i_{aa}=2$, $i_{da}=2$ and $i_{dd}=2$ are considered, and the heritability is assumed to be 0·05 (defined in the F_2 population under the Cockerham model by Kao & Zeng, 2002). With such parameter settings, the total genetic variance and environmental variance are 6·32 and 120·88, respectively, and the genetic variances contributed by the marginal effects and epistatic effects and genetic covariance are 3, 2·227 and $-1{\cdot}865$, respectively. The positions of the two QTLs were assumed to be 30 cM apart and located at 25 and 55 cM along one 100 cM chromosome. Two marker maps are considered. The first map assumes 11 equally spaced markers (the sparse map hereinafter), and the second map assumes 19 markers placed at 0, 10, 15, 20, 24, 27, 30, 35, 40, 45, 50, 54, 57, 60, 65, 70, 80, 90 and 100 cM (the dense map hereinafter). The sample size is 1000 and the number of simulated replicates is 100 for each setting. The applied mapping models are all two-QTL models with different fixed numbers of effects. Except for RI F_{10} population (RIL), the mapping models applied to QTL detection include the eight-effect (complete-effect) model, the five-effect model (with a_1, d_1, a_2, d_2 and i_{aa}) and the four-effect model (with a_1, d_1, a_2 and d_2). For RIL, the three-effect model with epistasis (with a_1, a_2 and i_{aa}) and the two-effect model without epistasis (with a_1 and a_2) are applied to the analysis as RIL has very few heterozygotes and low power to detect dominance components. These models are applied to a two-dimensional grid search on the chromosome for QTL. At the positions with maximum value of the likelihood function, we test the significance of the first (second) QTL given the second (first) QTL by testing its main and epistatic effects jointly. For example, given the second (first) QTL, the hypothesis H_0: $a_1=d_1=i_{aa}=i_{ad}=i_{da}=i_{dd}=0$ (H_0: $a_2=d_2=i_{aa}=i_{ad}=i_{da}=i_{dd}=0$) is tested for the existence of the first (second) QTL at the positions if the complete-effect model is used. Similarly, if the five-effect (four-effect) model is used, the hypothesis H0: $a_1=d_1=i_{aa}=0$ (H_0: $a_1=d_1=0$) is tested for the existence of the first QTL given the second QTL. If both the LRT statistics are larger than the specified critical values at 5% level, a successful detection of the two QTLs (separation of the two linked QTLs) is declared at the tested positions, and the corresponding estimated effects are reported as the MLE of the effects. In QTL mapping, the issue of determining the critical value for declaring QTL detection has been very complicated, and several methods have been suggested to determine the critical value (see for a review, Zou & Zeng, 2008). Here, the critical values are evaluated using the quick method of Piepho (2001) as this method can handle a wide variety of experimental designs, such as the AI, RI, IRI and IF_2 populations considered here.

The non-Markovian method obviously performs better than the Markovian method in the populations subject to self, such as RI and IRI populations. For AI and RIX populations, the two methods have similar powers, but the non-Markovian method provides more precise and accurate estimates for the positions and effects. To condense tables, only the results under the sparse map are tabulated in Tables 2–4, and those under the dense map are not tabulated, but expounded in the context. Table 2 shows the QTL mapping results under the sparse map in the RI populations. For the case of the sparse (dense) map, by applying the complete-effect model to QTL detection, the powers of separation in the RI F_3, F_4 and F_5 populations are 0·39 (0·18), 0·23 (0·05) and 0·10 (0·11), respectively, by the non-Markovian approach, and they are 0·29 (0·19), 0·16 (0·03) and 0·04 (0·13), respectively, by the Markovian approach. The complete-effect model becomes less powerful in the later RI populations due to loss of heterozygotes. When epistasis is completely ignored by applying the four-effect model to the analysis, the powers are lower than those by the complete-effect models. The powers by the non-Markovian method are 0·09 (0·00), 0·02 (0·03) and 0·04 (0·11) for the three populations, respectively, and they are 0·10 (0·01), 0·03 (0·04) and 0·07 (0·14), respectively, by the Markovian method under the sparse (dense) map. When applying the five-effect model by considering i_{aa} to QTL detection, the powers by the non-Markovian method are 0·55 (0·67), 0·73 (0·84) and 0·68 (0·98), respectively, and they are

https://doi.org/10.1017/S0016672309000081 Published online by Cambridge University Press

Table 2. *Simulation results of using different mapping models of the Markovian and non-Markovian methods under the sparse marker map in the RI populations*

Population	Method	Power	P1 = 25	P2 = 55	LRT_1	LRT_2	$\mu = 0$	$a_1 = 2$	$d_1 = 2$	$a_2 = 2$	$d_2 = 2$	$i_{aa} = 2$	$i_{ad} = 2$	$i_{da} = 2$	$i_{dd} = 2$	$\sigma^2 = 120·9$
RI F_3	8e[a]	39 %	18·5	61·5	24·4	25·6	0·176	1·808	1·584	1·835	1·637	1·428	1·472	1·385	1·987	119·9
			(4·5)	(4·5)	(9·1)	(9·2)	(0·454)	(0·537)	(1·083)	(0·608)	(1·160)	(0·602)	(1·303)	(1·481)	(2·505)	(6·2)
	8a	29 %	19·7	64·7	23·3	23·8	0·137	1·746	1·503	1·713	1·516	1·344	1·360	1·197	1·838	120·7
			(3·6)	(3·3)	(8·7)	(9·0)	(0·436)	(0·468)	(1·004)	(0·565)	(1·082)	(0·592)	(1·154)	(1·350)	(2·385)	(6·2)
	5e	55 %	25·3	54·8	21·3	22·1	−0·064	1·246	1·630	1·245	1·683	2·155				120·3
	5a	57 %	26·0	55·9	21·3	22·4	−0·111	1·193	1·687	1·291	1·704	2·191				120·5
	4e	9 %	25·8	55·8	10·2	10·4	0·302	1·154	0·767	1·362	0·765					122·0
	4a	10 %	25·8	56·5	9·9	10·8	0·259	1·308	0·624	1·195	0·778					122·2
RI F_4	8e	23 %	15·8	64·1	24·3	23·5	0·2056	1·689	1·499	1·609	1·554	1·461	1·150	1·147	1·301	121·5
			(2·8)	(2·6)	(10·3)	(9·1)	(0·814)	(0·781)	(1·834)	(0·795)	(1·899)	(0·564)	(1·592)	(1·891)	(4·322)	(6·6)
	8a	16 %	18·0	67·2	22·9	21·9	−0·037	1·580	1·317	1·497	1·177	1·424	0·987	0·947	1·350	122·8
			(2·0)	(2·1)	(9·3)	(9·00)	(0·656)	(0·644)	(1·430)	(0·669)	(1·538)	(0·529)	(1·450)	(1·633)	(3·507)	(6·2)
	5e	73 %	25·4	54·9	26·2	25·6	−0·230	1·118	1·631	1·095	1·591	2·266				120·8
	5a	74 %	27·1	56·7	26·00	25·8	−0·217	1·098	1·588	1·082	1·508	2·172				121·1
	4e	2 %	24·6	53·5	9·7	9·9	0·197	1·106	0·795	1·128	0·820					123·3
	4a	3 %	26·6	53·6	10·4	10·6	0·022	1·145	0·121	1·010	0·937					123·3
RI F_5	8e	10 %	15·0	64·7	19·5	19·3	0·154	1·483	1·663	1·638	1·200	1·276	0·777	1·217	1·469	121·5
			(1·4)	(2·2)	(8·5)	(8·7)	(1·245)	(1·107)	(2·714)	(1·015)	(2·740)	(0·446)	(2·342)	(2·026)	(6·013)	(6·4)
	8a	4 %	17·8	67·7	17·9	18·9	−0·073	1·400	1·360	1·351	1·142	1·146	0·685	0·501	1·647	121·2
			(2·0)	(2·6)	(7·7)	(10·9)	(0·861)	(0·966)	(1·756)	(0·836)	(1·941)	(0·454)	(2·241)	(1·353)	(3·619)	(7·5)
	5e	68 %	25·2	54·00	24·8	25·8	−0·168	1·083	1·452	1·107	1·376	2·050				119·7
	5a	67 %	27·3	56·7	25·1	25·9	0·348	1·039	1·267	1·115	1·355	2·015				120·2
	4e	4 %	24·6	54·2	9·4	10·3	−0·028	1·005	0·894	1·222	0·317					122·4
	4a	7 %	27·5	55·4	10·4	11·5	−0·429	1·068	−0·182	1·066	0·288					122·3
RI $F_{10(RIL)}$	3e	85 %	24·6	54·9	26·0	25·6	−1·455	1·055		1·000		1·988				120·0
	3a	83 %	27·6	57·6	25·9	25·4	−1·469	1·031		0·978		1·922				121·1
	2e	3 %	25·0	55·3	8·1	7·4	−0·721	0·955		1·148						123·3
	2a	2 %	25·6	57·1	7·8	7·5	−0·724	1·049		1·024						123·7

A total of 100 replicates, each with sample size 1000, were analysed with two linked epistatic QTLs, Q_A and Q_B. The heritability is 0·05 in the F_2 population. The critical values are determined by Piepho's method. P1 (P2): position of Q_A (Q_B). For reducing the text, standard deviations (SD; numbers in parentheses) are only shown for the complete-effect mode. The reduced models usually show similar or larger SD. SD are smaller in RIL as compared with the RI F3.

[a] 8e/8a indicates the eight-effect model with the non-Markovian/Markovian method.

Table 3. *Simulation results of using different mapping models of the Markovian and non-Markovian methods under the sparse marker map in the different AI populations*

Population	Method	Power	P1 = 25	P2 = 55	LRT_1	LRT_2	μ	$a_1=2$	$d_1=2$	$a_2=2$	$d_2=2$	$i_{aa}=2$	$i_{ad}=2$	$i_{da}=2$	$i_{dd}=2$	$\sigma^2=120·9$
AI F_3	8e[a]	61 %	24·8	56·2	31·3	29·9	0·075	2·088	1·886	1·927	1·829	1·724	1·919	1·999	2·467	119·6
			(4·9)	(5·3)	(10·8)	(10·5)	(0·458)	(0·749)	(1·149)	(0·806)	(1·081)	(1·435)	(1·889)	(1·894)	(2·278)	(6·4)
	8a	59 %	26·0	57·9	31·1	29·9	0·085	2·071	1·887	1·919	1·794	1·697	1·881	1·929	2·337	119·9
			(5·0)	(5·3)	(10·8)	(10·5)	(0·451)	(0·742)	(1·110)	(0·792)	(1·010)	(1·351)	(1·851)	(1·871)	(2·221)	(6·4)
	5e	41 %	25·3	55·3	20·2	19·4	−0·116	1·654	2·235	1·559	2·379	2·757				121·5
	5a	39 %	26·6	56·8	20·3	19·4	−0·100	1·656	2·181	1·538	2·313	2·680				121·7
	4e	11 %	33·9	56·3	13·3	12·3	0·530	1·674	1·240	1·604	1·200					123·0
	4a	11 %	25·0	57·4	13·1	12·2	0·531	1·651	1·216	1·608	1·191					123·1
AI F_4	8e	62 %	24·0	56·9	31·4	32·7	−0·089	1·806	2·021	1·922	1·762	2·017	1·806	1·813	2·029	120·4
			(3·8)	(4·4)	(10·1)	(9·9)	(0·422)	(0·565)	(1·205)	(0·659)	(1·063)	(0·675)	(1·539)	(1·605)	(2·661)	(7·0)
	8a	59 %	24·7	60·0	30·3	32·1	−0·049	1·769	1·874	1·917	1·651	1·864	1·619	1·692	1·780	121·1
			(4·4)	(4·7)	(10·3)	(10·0)	(0·423)	(0·593)	(1·134)	(0·672)	(0·998)	(0·633)	(1·434)	(1·478)	(2·417)	(7·0)
	5e	63 %	24·8	55·8	23·7	24·9	−0·046	1·188	1·864	1·346	1·593	2·145				122·2
	5a	37 %	27·3	57·0	18·7	21·0	−0·036	1·664	1·843	1·810	1·915	2·049				122·1
	4e	9 %	24·4	54·5	11·2	12·7	0·216	1·052	1·072	1·453	0·904					124·2
	4a	8 %	25·1	55·7	10·6	12·8	0·219	1·149	1·010	1·343	0·948					124·4
AI F_5	8e	47 %	25·8	55·2	27·5	29·1	0·003	1·995	1·794	2·046	1·826	1·742	1·690	2·020	1·863	119·3
	8a	46 %	28·1	57·3	27·4	29·0	−0·017	1·963	1·712	1·974	1·722	1·702	1·532	1·887	1·632	120·4
	5e	40 %	24·8	54·8	18·8	21·1	−0·027	1·699	1·950	1·842	2·077	2·174				121·6
	5a	41 %	27·3	57·0	18·7	21·0	−0·036	1·664	1·841	1·810	1·915	2·049				122·1
	4e	25 %	24·1	56·0	13·5	14·9	0·259	1·788	1·314	1·845	1·399					123·0
	4a	17 %	26·3	57·5	13·3	14·9	0·248	1·748	1·247	1·812	1·289					123·4
AI F_{10}	8e	6 %	25·2	55·0	18·6	19·5	−0·024	1·505	1·259	1·563	1·051	1·244	0·623	1·013	0·473	115·7
	8a	7 %	27·0	57·1	18·1	19·1	−0·049	1·354	0·949	1·432	0·778	1·011	0·421	0·570	−0·012	120·4
	5e	9 %	25·4	55·1	12·8	13·8	0·001	1·431	1·286	1·662	0·959	1·372				120·3
	5a	11 %	27·0	57·4	12·6	13·6	−0·045	1·392	0·988	1·480	0·764	1·157				122·5
	4e	12 %	25·3	55·0	10·0	10·7	0·002	1·532	1·259	1·714	0·857					121·2
	4a	5 %	26·2	56·6	9·8	10·6	−0·025	1·472	1·013	1·533	0·716					123·1

For reducing the text, SD (numbers in parentheses) are only shown for the complete-effect mode in AI F3 and F4 populations. SD for the reduced models are usually similar or larger. The SDs of AI F5 have similar size in positions and main effects and larger size in epistatic effects as compared with those in AI F3. The estimates in AI F10 have a much larger SD. A total of 100 replicates, each with sample size 1000, were analysed with two linked epistatic QTLs, Q_A and Q_B. The heritability is 0·05 in the F_2 population. The critical values are determined by Piepho's method.

[a] 8e/8a indicates the eight-effect model with the non-Markovian/Markovian method.

Table 4. *Simulation results of using different mapping models of the Markovian and non-Markovian methods under the sparse marker map in F_2, IF_2 and IRI populations*

Population	Method	Power	P1=25	P2=55	LRT_1	LRT_2	$\mu=0$	$a_1=2$	$d_1=2$	$a_2=2$	$d_2=2$	$i_{aa}=2$	$i_{ad}=2$	$i_{da}=2$	$i_{dd}=2$	$\sigma^2=120{\cdot}9$
F_2	8e[a]	42 %	20·0	59·8	26·3	26·7	0·273	1·801	1·691	1·974	1·616	1·427	1·842	1·565	2·054	121·4
			(4·8)	(4·7)	(9·8)	(10·2)	(0·434)	(0·935)	(0·935)	(0·962)	(1·025)	(1·236)	(1·820)	(1·788)	(2·249)	(17·8)
	5e	21 %	24·6	55·4	27·4	17·9	−0·137	1·478	2·403	1·405	2·289	2·928				122·5
	4e	3 %	24·0	57·5	10·5	10·9	0·715	1·512	1·248	1·431	0·942					123·7
IF_2	8e	6 %	23·9	54·9	19·9	19·8	−0·081	1·241	0·766	1·351	0·805	0·969	0·750	0·245	0·602	121·2
			(7·9)	(7·4)	(6·6)	(7·1)	(0·433)	(1·028)	(1·381)	(0·907)	(1·512)	(1·897)	(2·962)	(2·973)	(5·381)	(6·4)
	8a	5 %	23·4	54·5	19·6	19·6	−0·066	1·158	0·682	1·294	0·709	0·898	0·725	0·314	0·527	122·5
			(7·9)	(7·2)	(6·7)	(6·9)	(0·427)	(0·987)	(1·271)	(0·846)	(1·314)	(1·743)	(2·552)	(2·733)	(4·527)	(6·0)
	5e	5 %	24·2	54·5	12·8	12·4	−0·101	1·356	0·885	1·439	0·932	1·146				124·2
	5a	5 %	24·1	54·1	12·8	12·3	−0·101	1·283	0·822	1·404	0·868	1·071				124·8
	4e	2 %	24·4	54·7	10·0	9·3	−0·067	1·377	0·870	1·497	0·862					124·9
	4a	3 %	23·5	54·2	10·0	9·3	−0·076	1·292	0·813	1·433	0·779					125·4
IRI $F_{5:1}$	8e	13 %	17·6	63·0	20·4	21·0	−0·384	1·228	0·003	1·199	0·834	1·490	0·342	1·195	−1·607	119·3
			(4·0)	(4·0)	(8·6)	(8·8)	(2·069)	(1·678)	(3·609)	(1·530)	(3·022)	(0·473)	(3·590)	(3·315)	(5·005)	(6·8)
	8a	1 %	15·2	63·2	21·5	21·0	−0·229	2·043	2·073	1·076	1·384	1·317	1·964	−0·453	3·686	121·0
			(4·2)	(4·2)	(13·4)	(13·4)	(1·097)	(0·211)	(1·667)	(0·970)	(0·744)	(0·636)	(0·610)	(2·154)	(4·884)	(8·9)
	5e	77 %	24·7	55·3	26·0	26·6	−0·177	0·995	1·468	1·066	1·264	1·965				118·5
	5a	76 %	23·9	54·8	25·8	26·3	−0·578	0·946	1·165	1·052	0·729	1·824				119·8
	4e	1 %	23·6	54·3	8·6	9·8	−0·824	0·964	0·593	1·093	−0·820					121·6
	4a	5 %	52·8	53·4	10·6	10·9	−1·700	0·979	−1·126	0·995	−1·365					121·2
IRI $F_{5:3}$	5e	93 %	24·7	55·1	28·6	29·9	−1·670	1·050	−0·152	1·192	0·086	2·482				117·4
	5a	92 %	24·0	54·4	28·9	29·4	−1·697	0·963	−0·940	1·509	0·774	2·061				119·0
	4e	3 %	24·0	58·0	8·3	8·3	−1·464	1·035	−0·825	1·232	−0·561					122·5
	4a	4 %	27·0	51·7	11·6	10·7	−4·500	0·918	−4·273	1·038	−4·085					120·0

For reducing the text, SDs (numbers in parentheses) are only shown for the complete-effect mode. SDs for the reduced models are usually similar or larger. A total of 100 replicates, each with sample size 1000, were analysed with two linked epistatic QTLs, Q_A and Q_B. The heritability is 0·05 in the F_2 population. The critical values are determined by Piepho's method. P1 (P2): position of Q_A (Q_B).

[a] 8e/8a indicates the eight-effect model with the non-Markovian/Markovian method.

0·57 (0·66), 0·74 (0·86) and 0·67 (0·99), respectively, by the Markovian method under the sparse (dense) map. In parameter estimation, for all models, the estimates of positions and effects obtained by the non-Markovian method have a better precision as compared with those by the Markovian method. For example, in the RI F_4 population under the sparse map, the means of the estimated Q_A and Q_B positions for the five-effect model are 25·36 (SD 5·94) and 54·90 (SD 5·74), respectively, by the non-Markovian method, and they are 26·08 (SD 5·98) and 56·70 (SD 5·67), respectively, by the Markovian method. The five-effect model by taking i_{aa} into account tends to be more powerful and precise than the other two models, and this model becomes more powerful in the later RI populations. For the RI F_{10} population (RIL), when using the three-effect model, the powers of the non-Markovian (Markovian) method are 93% (94%) and 98% (97%) in the two maps. When using the two-effect model, the powers reduce dramatically to 5% (5%) and 8% (7%), respectively. This shows that the power to detect QTL can be greatly enhanced by taking i_{aa} into account in RIL. Confounding problems occur in the estimation of the effects if epistatic effects are not completely taken into account. For example, the means of the estimated a_1, a_2 and i_{aa} by the non-Markovian method are 1·031 (SD 0·388), 1·032 (SD 0·402) and 1·965 (SD 0·375), respectively (the predicted values by Table 1 are 1, 1 and 2) for RIL, under the dense map. It is interesting to compare these results with those in the F_2 population. The powers in the F_2 population are 0·42 (0·36), 0·21 (0·43) and 0·03 (0·05) for the complete-effect, five-effect and four-effect models, respectively, under the sparse (dense) map (Table 4). The more powerful performance of using the RI populations occurs only for the five-effect model and does not occur for the other two models.

Table 3 presents the QTL mapping results for AI populations under the sparse maps. Under the sparse (dense) map, when the complete-effect model is considered, the detecting powers by the non-Markovian method are 0·61 (0·61), 0·62 (0·52), 0·47 (0·79) and 0·06 (0·65), respectively, in the AI F_3, F_4, F_5 and F_{10} populations, and they are 0·59 (0·65), 0·59 (0·51), 0·46 (0·79) and 0·07 (0·70), respectively, by the Markovian method. When epistasis is ignored by using the four-effect model, the powers are reducing to 0·11 (0·11), 0·09 (0·08), 0·25 (0·17) and 0·12 (0·05), respectively, by the non-Markovian (Markovian) method. If the five-effect model is considered under sparse map, the powers by the non-Markovian (Markovian) method are 0·41 (0·39), 0·63 (0·37), 0·40 (0·41) and 0·09 (0·11), respectively, in the four populations. An increasing trend in power can be observed in the case of the dense map (not shown). However, such an increasing trend does not occur in the sparse map (Table 3). Also, by taking epistasis into account, the power can be much improved and the confounding problem can be avoided, and the means of the estimated effects are all very close to the true given parameters. Besides, the QTL positions are estimated with better precision in the AI populations as compared with those estimated in the RI populations. Among all the settings, the most powerful experimental population for QTL detection is the AI F_3 (AI F_5) population under the sparse (dense) map. The AI F_{10} population is not the optimal design under either map, as the powers are about 0·05–0·12 and about 0·45–0·70, respectively, in the two maps. It is expected that a much denser map is required to ensure more powerful QTL detection in the AI F_{10} population (see the Discussion section). When comparing the results of the AI and F_2 populations (Table 4), the AI populations show more powerful results than the F_2 population in all cases under the dense map.

Table 4 shows the QTL mapping results in the F_2, IF_2, IRI $F_{5:1}$ and IRI $F_{5:3}$ populations under the sparse maps. The QTL mapping results are better under the dense map in these later advanced populations as compared with those under the sparse map. For example, in the IF_2 population, the powers under the dense map are 0·59 (0·58), 0·54 (0·52) and 0·35 (0·35) by the non-Markovian (Markovian) method for the complete-effect, five-effect and four-effect models (not shown), respectively. Under the sparse map, they are 0·06 (0·05), 0·05 (0·05) and 0·02 (0·03), respectively. The estimated positions and effects are also found to be more precise in the dense map. For example, under the complete-effect model, the estimated effects of a_1, d_1, a_2 and d_2 by the non-Markovian method are 1·919 (SD 0·657), 1·689 (SD 1·010), 1·757 (SD 0·670) and 1·677 (SD 0·912), respectively, in the dense map (not shown), and they are 1·241 (SD 1·028), 0·766 (SD 1·381), 1·351 (SD 0·907) and 0·850 (SD 1·512), respectively, in the sparse map. Similar situations were also found in the IRI $F_{5:1}$ and IRI $F_{5:3}$ populations. Besides, the complete-effect model is not appropriate for the IRI populations, and the three-effect and five-effect models are more appropriate for these two populations. For example, the powers in the IRI $F_{5:1}$ population are 0·13 (0·01) and 0·07 (0·01) by the complete-effect model of the non-Markovian (Markovian) method in the two different maps, and the powers become 0·77 (0·76) and 0·92 (0·93) by the five-effect model, respectively. Also, taking the additive-by-additive effect into account can greatly benefit the QTL detection. A similar trend can be observed for the RIL.

5. Discussion

The genome structures of the advanced populations can be very different from each other and are no

https://doi.org/10.1017/S0016672309000081 Published online by Cambridge University Press

longer similar to that of the F_2 population as mentioned before. This paper tries to distinguish between the genome structures of different populations to deal with the issues of QTL mapping. When using the advanced populations for QTL mapping, we propose the Markovian and non-Markovian methods to map for QTL. Some important properties and issues in QTL mapping, such as mapping closely linked QTLs, confounding problems of ignoring epistasis and the choice of different mapping models, are also derived and discussed across different populations. Theoretically, the non-Markovian method have better performances than the Markovian method, as the more accurate mixing proportions can be used in statistical modelling as discussed. In fact, analytical and simulation studies show that the non-Markovian method does perform better than the Markovian method in the advanced populations, especially in populations subject to the selfing process. The advanced populations can be also designed to be more powerful than the F_2 population in QTL detection. Besides, the issues considered here are under the assumption of large sample size with no selection. In practice, selection and drift may play a role between generations, and it will cause unequal allele frequencies and potential segregation distortion. As suggested by Teuscher & Broman (2007), the solution to these problems is the use of a dense marker set with which the actual recombination breakpoints can be precisely mapped. The results presented here can give some clues to the use of advanced population for better investigation in genetical and biological studies.

The quality of QTL mapping relies on precisely deriving the conditional probabilities of putative QTL genotypes given marker genotypes and on applying appropriate statistical methods to link the quantitative traits with the putative QTL. When deriving the conditional genotypic distribution of a putative QTL, ideally, we would like to use the information from all the linked markers (as many linked markers as possible) to obtain it. This, however, is very challenging as the characterization of the genotypic distribution of many genes is not an easy task. The approach of interval mapping avoids this and proposes to use its two flanking markers instead in derivation, so that its task reduces to characterizing the genotypic distribution for three genes. Such an approach is optimal in capturing the QTL information for the genomes with a first-order Markovian property, but not for the genomes without this property. However, for the latter case, we believe that the closest marker pair may have already captured most of the information about QTLs. When multiple putative QTLs are considered in the advanced population, the joint conditional probability distribution used here is approximate and obtained by using conditional independence property as we are still not sure currently how to derive the exact conditional distribution for an arbitrary number of putative QTLs. In addition, when applying statistical models to detect QTLs, the specific genome structures of advanced populations have to be taken into account in modelling to benefit QTL detection. For example, in the RI or IRI populations, there are larger additive genetic variances (smaller dominance variances) and higher homozygosity (lower heterozygosity), and the applied models should consider that fitting the components involving additive effects into the model can benefit QTL detection and that fitting the components involving dominance effects into the model may deter QTL detection.

One of the most precious features in the advanced populations is that they can generate more recombinants to improve the QTL resolution. From the viewpoint of statistical modelling, such an improvement is to take advantage of more recombinants in a population to alleviate the collinearity problem in modelling-linked putative QTLs (to disassociate the linkage disequilibrium between linked putative QTLs), so that QTL mapping can be more powerful and precise (see the subsection 'Power of separating closely linked QTLs'); nevertheless, more recombinants also reduce the linkage disequilibrium between markers and QTLs to blur the information about the unobservable putative QTL. Therefore, to expect improved QTL mapping results in the advanced population, a denser marker map around the linked QTL region is required to ensure that the linkage disequilibrium is strong enough in the construction. In a marker interval with given width, the linkage disequilibrium between markers and putative QTLs is strongest in the F_2 population, and it becomes gradually weaker as generation advances. Taking a putative QTL Q in the middle of a 10 cM marker interval flanked by markers, A and B, as an example, the trigenic linkage disequilibrium defined as $D_{AQB}=P_{AQB}-P_A P_Q P_B$ (Wright, 1980) is 0·329 in the F_2 population, and it becomes 0·309 (0·300), 0·286 (0·292), 0·260 (0·290) and 0·111 (0·288) in the AI (RI) F_3, F_4, F_5 and F_{10} population, respectively. It shows that the linkage disequilibrium is declining more rapidly under random mating. In general, once the designed populations, such as IF_2, IRI $F_{5:1}$ and AI F_{10} populations, have undergone some generations of random mating, they usually require a much denser marker map to obtain improved results. Therefore, the marker density should be considered as a major factor not only in the comparison between the two proposed methods, but also in the issue of using advanced populations to improve QTL mapping results (see also the 'Simulation studies' section). Besides, the issues of trade-off between generation number and marker density and of extension to more than two founders (Mott *et al.*, 2000; Broman, 2005) are

https://doi.org/10.1017/S0016672309000081 Published online by Cambridge University Press

interesting and worthy of pursuing in the future. Together with the (F_u/F_v, $v \geqslant u$) designs (Fisch *et al.*, 1996; Kao, 2006) and the strategy of replicated trials (Hua *et al.*, 2002), it is very much possible for us to design experimental populations to recover or remove those undetected or ghost QTLs (Lander & Botstein, 1989) in the F_2 population for high-resolution QTL mapping.

The authors are grateful to two anonymous reviewers for helpful comments. This work was supported by grant numbers NSC97-2118-M-001–008 from the National Science Council, Taiwan, Republic of China.

References

Broman, K. W. (2005). The genomes of recombinant inbred lines. *Genetics* **169**, 1133–1146.

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 967–971.

Darvasi, A. (1998). Experimental strategies for the genetic dissection of complex traits in animal models. *Nature Genetics* **18**, 19–24.

Dempster, A. P., Larid, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**, 1–38.

Fisch, R. D., Ragot, M. & Gay, G. (1996). A generalization of the mixture model in the mapping of quantitative trait loci for progeny from a biparental cross of inbred lines. *Genetics* **143**, 571–577.

Geiringer, H. (1944). On the probability theory of linkage in Mendelian heredity. *The Annals of Mathematical Statistics* **15**, 25–57.

Haldane, J. B. S. & Waddington, C. H. (1931). Inbreeding and linkage. *Genetics* **16**, 357–374.

Hua, J. P., Xing, Y. Z., Xu, C. G., Sun, X. L., Yu, S. B. & Zhang, Q. (2002). Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. *Genetics* **162**, 1885–1895.

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

Jennings, H. S. (1916). The numerical results of diverse systems of breeding. *Genetics* **1**, 53–89.

Jiang, C.-J. & Zeng, Z.-B. (1997). Mapping quantitative trait loci with dominant and missing markers in various populations from inbred lines. *Genetica* **101**, 47–85.

Kao, C.-H. (2004). Multiple interval mapping for quantitative trait loci controlling endosperm traits. *Genetics* **167**, 1987–2002.

Kao, C.-H. (2006). Mapping quantitative trait loci using the experimental designs of recombinant inbred population. *Genetics* **174**, 1373–1386.

Kao, C.-H. & Zeng, Z.-B. (1997). General formulas for obtaining the MLE and the asymptotic variance–covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 359–371.

Kao, C.-H. & Zeng, Z.-B. (2002). Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* **160**, 1243–1261.

Kao, C.-H., Zeng, Z.-B. & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Koski, T. (2001). *Hidden Markov Models for Bioinformatics*. Boston, MA: Kluwer Academic Publishers.

Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lee, M., Sharopova, N., Beavis, W. D., Grant, D., Katt, M., Blair, D. & Hallauer, A. (2002). Expanding the genetic map of maize with the intermated B73 Mo17 (IBM) population. *Plant Molecular Biology* **48**, 453–461.

Liu, S.-C., Kowalski, S. P., Lan, T.-H., Feldmann, K. A. & Paterson, A. H. (1996). Genome-wide high-resolution mapping by recurrent intermating using *Arabidopsis thaliana* as a model. *Genetics* **142**, 247–258.

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* **12**, 591–612.

Martin, O. C. & Hospital, F. (2006). Two- and three-locus tests for linkage analysis using recombinant inbred lines. *Genetics* **173**, 451–459.

Mott, R., Talbot, C. J., Turri, M. G., Collins, A. C. & Flint, J. (2000). From the cover: a method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences USA* **97**, 12648–12654.

Piepho, H. P. (2001). A quick method for computing approximate threshold for quantitative trait loci detection. *Genetics* **157**, 425–432.

Robbins, R. B. (1918). Some applications of mathematics to breeding problems III. *Genetics* **3**, 375–389.

Rockman, M. L. & Kruglyak, L. (2008). Breeding designs for recombinant inbred advanced intercross lines. *Genetics* **179**, 1069–1078.

Teuscher, F. & Broman, K. W. (2007). Haplotype probabilities for multiple-strain recombinant inbred lines. *Genetics* **175**, 1267–1274.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates.

Weir, B. S. & Cockerham, C. C. (1977). Two-locus theory in quantitative genetics. *Proceedings of the International Conference on Quantitative Genetics* (ed. E. Pollak, O. Kempthorne & T. B. Bailey), pp. 247–269. Ames, IA, USA: Iowa State University.

Winkler, C. R., Jensen, N. M., Cooper, M., Podlich, D. W. & Smith, O. S. (2003). On the determination of recombination rates in intermated recombinant inbred populations. *Genetics* **164**, 741–745.

Wright, S. (1980). Genic and organismic selection. *Evolution* **34**, 825–843.

Xu, S. (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**, 513–521.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

Zou, F., Gelfond, J. A. L., Airey, D. C., Lu, L., Manly, K. F., Williams, R. W. & Threadgill, D. W. (2005). Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. *Genetics* **170**, 1299–1311.

Zou, W. & Zeng, Z.-B. (2008). Statistical methods for mapping multiple QTL. *International Journal of Plant Genomics*, Article ID 286561, doi: 10.1155/2008/286561.

https://doi.org/10.1017/S0016672309000081 Published online by Cambridge University Press