

A mixed model QTL analysis for a complex cross population consisting of a half diallel of two-way hybrids in *Arabidopsis thaliana*: analysis of simulated data

Maria-João Paulo · Martin Boer · Xueqing Huang · Maarten Koornneef · Fred van Eeuwijk

Received: 8 May 2007 / Accepted: 5 February 2008
© The Author(s) 2008

Abstract To improve QTL detection power for QTL main effects and interactions and QTL mapping resolution, new types of multi-founder crossing populations are created in plants and animals. Some recent examples are complex intercrossed populations in mice and *Arabidopsis thaliana*. For the latter, a set of eight accessions was intercrossed to produce four two-way hybrids that were subsequently intercrossed again in a half diallel fashion leading to six subpopulations of four-way hybrids, each subpopulation containing 100 individuals. Within each subpopulation, individuals were inbred for four generations via single seed descent. QTL mapping in the complex crosses requires new statistical tools. We present a first sketch of a QTL mapping methodology for the complex cross in *Arabidopsis* based on mixed model analyses. As experimental

data were not yet available, we illustrate our methodology on simulated but realistic data.

Keywords Complex cross · Four-way RILs · Hidden Markov models · Mixed models

Introduction to complex crosses

Arabidopsis thaliana is used extensively for the functional analysis of its genes. In addition to mutants, natural variation among accessions worldwide is a promising source of genetic variation (Alonso-Blanco and Koornneef 2000; Weigel and Nordborg 2005). *Arabidopsis* has a broad geographic distribution and it grows in very different environments, therefore phenotypic variation among accessions is expected to reflect genetic diversity underlying adaptation to specific conditions. *Arabidopsis* accessions display genetic variation for many morphological and physiological traits (Alonso-Blanco and Koornneef 2000; Koornneef et al. 2004).

To extend the possibilities for the study of the genetic basis of adaptation in *Arabidopsis* and following a similar initiative in mice (The Complex Trait Consortium 2004), recently a new type of complex cross population was created. A set of eight founder accessions was chosen to cover a wide genetic variation (Table 1). The eight founder accessions were pairwise crossed to produce four two-way

M.-J. Paulo · X. Huang · M. Koornneef
Max Planck Institute for Plant Breeding Research,
Carl-von-Linné-Weg 10, Cologne 50829, Germany

M.-J. Paulo
Laboratory of Plant Breeding, Wageningen UR,
P. O. Box 386, Wageningen, The Netherlands

M.-J. Paulo (✉) · M. Boer · F. van Eeuwijk
Biometris, Wageningen UR, P. O. Box 100, Wageningen,
The Netherlands
e-mail: joao.paulo@wur.nl

Table 1 The set of eight accessions with their country of origin

Name	Abbreviation	Country
Antwerp	An-1	Belgium
Cape Verde Islands	Cvi	Cape Verde Islands
C24	C24	Portugal
Columbia	Col	USA
Eriengsoda	Eri	Sweden
Kyoto	Kyo	Japan
Landsberg <i>erecta</i>	Ler	Poland
Shakdara	Sha	Tadjikistan

These accessions were used as founders in the complex crosses

hybrids. These four two-way hybrids were intercrossed in a half diallel fashion leading to six (sub)populations, that we will refer to as F1 populations, each subpopulation consisting of approximately 100 four-way hybrid individuals. The individual plants in the subpopulations were self fertilized and advanced to the F4 generation by single seed descent (Fig. 1).

Genotyping was done in single F4 plants using microsatellite markers with four to seven different

alleles among the eight founder parents. Phenotyping for traits such as flowering time, leaf number and leaf shape is underway for the F5 progeny of the genotyped F4 plants. In an F4 generation considerable homozygosity will have been achieved, but an F4 generation still contains enough residual heterozygosity to allow the generation of Heterogenous Inbred line Families (HIFs) as described by Tuinstra et al. (1997). This will permit the development of near-isogenic lines for specific regions in which Quantitative Trait Loci (QTLs) can be detected.

The *Arabidopsis* complex cross population is likely to cover a large part of the natural variation observed in *Arabidopsis thaliana*, while providing new genetic combinations not present in the founder accessions. It is likely that new allele combinations can be studied that were not observed in nature. The complex cross population is expected to produce phenotypic variation even beyond the phenotypic range covered by the founders. Furthermore, the combination of a very diverse population with large sample sizes is likely to result in increased power for QTL detection and epistatic interactions (Valdar et al. 2006).

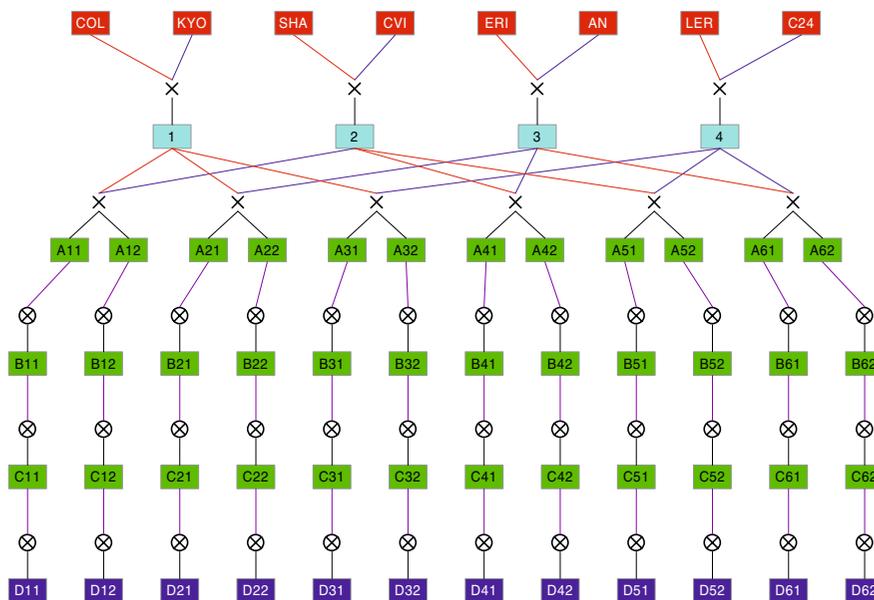


Fig. 1 Schematic illustration of the complex cross in *Arabidopsis*. The top row shows the eight accessions, the founders. The second row contains the four two-way hybrids. These two-way hybrids are again crossed with each other according to a half diallel scheme, resulting in six subpopulations of four-way hybrids, say F1 (A), each with 100

individuals (only two individuals per subpopulation are shown). All individual F1 plants are self fertilized and advanced to the F4 (D) generation by single seed descent. This figure was generated with the software package Pedimap (Roeland Voorrips, Plant Research International, Wageningen, The Netherlands)

Complex cross populations are becoming an increasingly accessible tool to plant breeders. For a recent example see Blanc et al. (2006). These crosses allow more powerful studies of the genetic basis of plant traits in more relevant genetic backgrounds (Charcosset et al. 2001; Darvasi and Soller 1995). The statistical analysis of complex crosses requires further development of existing methodology. In this paper we present a QTL analysis for the complex cross in *Arabidopsis*, where we will depart from a mixed model framework. The statistical methodology will be illustrated for simulated data.

Methodology and simulated data

Simulated data

We simulated 1,000 complex cross populations, each consisting of six four-way crosses, as illustrated in Fig. 1. First, the genotypes at marker loci for the eight founders were independently simulated, assuming they could take up one of four distinct values; AA, BB, CC, or DD, each with probability $\frac{1}{4}$. Then six four-way crosses, each cross forming a subpopulation, were simulated conditional on the founders' genotypes. The total number of F4 lines was 600 (100 per subpopulation). Figure 2 shows a pedigree for an F4 individual following from a particular four-way cross, namely [*Ler* × C24] × [*Cvi* × Sha], where this subpopulation is created from the cross between the two-way hybrids 2, [*Ler* × C24], and 4, [*Cvi* × Sha]. Generations starting at the offspring of two-way hybrids we will call F1, and we use the letter A for referring to such generations. Subsequently, F2 or B generations are obtained by selfing F1's, while F3's or C's and F4's or D's are generated by consecutive selfing adopting a single seed descent scheme. Figure 2 gives the pedigree for an F4 or D individual in the 5th subpopulation. This individual is given the label D51, with D from F4, 5 from the 5th subpopulation, and 1 as the individual identification number within this subpopulation. The individuals in this particular subpopulation have the labels D51, D52,...,D599, D5100.

We represent the *Arabidopsis thaliana* genome by five chromosomes of length 100 cM each, but simulated only one chromosome, carrying 11

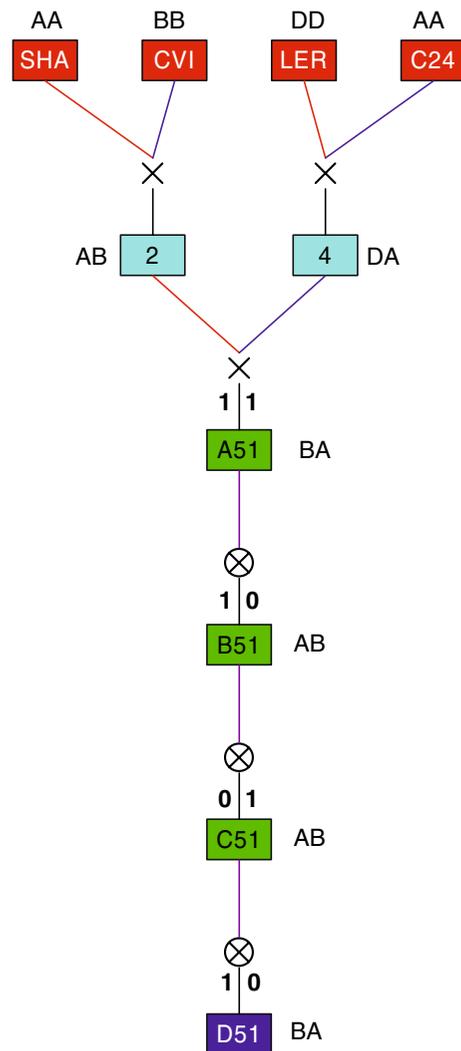


Fig. 2 Example of an inheritance vector $h = (1,1,1,0,0,1,1,0)$, at one particular locus (marker M6, see also Table 2), for the first individual of the F4 generation in the fifth subpopulation, D51

equidistant markers, with a distance of 10 cM between the markers. Thus, the marker on the left end of the chromosome, labeled M1, was simulated at 0 cM, and the marker at the other end of the chromosome, M11, was simulated at 100 cM. A single bi-allelic QTL was simulated at 75 cM, with additive effect $a = 0.4$. The random residual error was assumed to have a normal distribution with mean 0 and variance 1.0, which resulted in low heritabilities; between 0.10 and 0.13 for the subpopulations. We assumed that the founders Kyo, Sha, Cvi, Eri were homozygous for the allele with a positive

additive effect, and that the other four founders, Col, An, *Ler*, and C24, were homozygous for the allele with the negative additive effects.

Analysis of the simulated data

We need to develop a methodology for interval mapping that accounts for the relationships between the parents in our complex cross. For our complex cross, estimation of the probabilities for individual genotypes at genomic positions in-between markers, as needed for interval mapping, is not trivial because each F4 line has four founders and exhibits accumulated recombination over three generations. The genotypes for individual loci of these F4 lines are thus potentially more diverse than those of the individuals in standard bi-parental populations. Closely similar to the situation for our complex cross, Broman (2005) derived two-locus and three-locus haplotype probabilities for four-way and eight-way RILs assuming the RILs were fully inbred. As Broman (2005) assumed complete homozygosity and larger accumulated recombination fractions than valid for our F4 lines, his results are not completely pertinent to our complex cross containing a higher degree of heterozygosity.

A further complication is that founder alleles are not always distinguishable at marker locations; if any two founders have an identical genotype at a particular locus, say AA, then it is impossible to be sure about the origin of an A allele when observed in a descendant. We then know that the alleles in founder and descendant are identical in state, but we are not sure about the alleles being identical by descent.

As we work with relatively small pedigrees, we can use concepts proposed by Lander and Green (1987) for the estimation of genotypic probabilities at arbitrary genomic positions, where we will combine so-called inheritance vectors and Hidden Markov Models. For a recent example of a similar kind of approach in the context of QTL analysis for multiple environment data see Boer et al. (2007).

The QTL analysis of the complex cross population consists of two steps. In the first step, we use inheritance vectors and Hidden Markov Models to calculate the expected number of alleles derived from the inbred founders for a dense grid of evaluation points along the genome, taking as input information the marker positions and marker scores of the

founders and of the individuals in the subpopulations. This can be done independently for each subpopulation. In the second step we analyze the whole complex cross population with a mixed model for QTL detection, using the expected number of alleles of the founders as genetic predictors, and the phenotype as observed variable.

Inheritance vectors

Inheritance vectors can be used to define how the DNA of the founders is transmitted through a pedigree (Lander and Green 1987). In the complex cross described above we distinguished four generations, F1 to F4. A diploid individual inherits for a particular locus two alleles from the previous generation. These alleles come from two gametes that can be ordered, for example, the first gamete could correspond to the mother of the individual, while the second gamete corresponds to the father. In the F1 generation the transmission of alleles can be represented by a vector with two elements each taking a value 0 or 1. The first element of the vector represents the origin of the first gamete, and the second element represents the origin of the second gamete. They take value 0 if the gamete is a copy of the parent's first gamete (mother), or value 1 if the gamete is a copy of the parent's second gamete (father). For an example see Fig. 2, where for the first individual of the F1 in the fifth subpopulation, A51, we consider a particular position, marker M6. The inheritance vector for A51 has the value (1,1), because A51 inherited from its first parent the second gamete (allele B from hybrid 2), and A51 inherited from its second parent the second gamete (allele A from hybrid 4). This idea can be extended over several consecutive generations, so that the inheritance mechanism for an individual in the F4 generation can be represented by a binary vector of length eight. In Fig. 2, the inheritance vector of the first individual in the F4 of the fifth subpopulation, the individual with label D51, is (1,1,1,0,0,1,1,0), yielding genotype BA. The total number of possible inheritance vectors in the F4 generation is $2^8 = 256$. A useful quantity in this respect is $N_k(h)$, the number of alleles transmitted to a genotyped individual from founder k given the inheritance vector h . Notice that for the offspring of a four-way cross in a

diploid organism, choosing $k = 1, 2, 3, 4$; $N_1(h) + N_2(h) + N_3(h) + N_4(h) = 2$.

Hidden Markov models

The probability of each inheritance vector, h , can be calculated using all available marker information and the recombination frequencies between markers and the locus of interest. In mathematical notation, for an individual i with marker scores \mathbf{M}_i we define $\gamma_i(h)$ as the conditional probability of inheritance vector h , or $\gamma_i(h) = P(H = h | \mathbf{M}_i)$, with H the random variable representing the possible inheritance vectors and $\sum \gamma_i(h) = 1$. These conditional probabilities are a function of position on the genome, but to simplify the notation we assume that this position is given. The probabilities $\gamma_i(h)$ can be calculated by formulating the problem as a Hidden Markov Model (HMM). Basically, the calculation consists of the following steps. First, the transition probabilities between markers and the putative QTL position are calculated. Secondly, the QTL probabilities are calculated, conditional on all the marker data available on the left side of the QTL position. In the third step, the conditional QTL probabilities are calculated using all the markers on the right side. In the final step the left- and right-conditional probabilities are combined to calculate the QTL probabilities conditional on all the marker data. For further details and efficient algorithms to solve HMM see e.g. Lander and Green (1987), Rabiner (1989), and Bishop (2006).

QTL mapping

To analyse each subpopulation separately we use the following linear mixed model for the detection of QTLs, where we will underline random variables:

$$y_i = \mu + \sum_{k=1}^4 x_{ik} \underline{\beta}_k + \underline{\varepsilon}_i, \quad \underline{\beta}_k \sim \mathcal{N}(0, \sigma_{\beta}^2), \quad \underline{\varepsilon}_i \sim \mathcal{N}(0, \sigma_{\varepsilon}^2),$$

where y_i is the observed phenotype for individual i , μ is the overall mean, x_{ik} is the expected number of alleles derived from founder k , given marker information, $\underline{\beta}_k$ is a random effect for founder k , and $\underline{\varepsilon}_i$ is the residual error. The variance component for the founder effect is denoted by σ_{β}^2 , the residual variance by σ_{ε}^2 .

The genetic predictors x_{ik} can be calculated as:

$$x_{ik} = \sum_h \gamma_i(h) \cdot N_k(h).$$

We used a residual maximum likelihood ratio test to test the significance of the founder effect variance component. We approximated the P -value by using a mixture of half χ_0^2 and half χ_1^2 (Self and Liang 1987; Stram and Lee 1994).

To analyse the complete complex cross consisting of six subpopulations, the following mixed model can be used:

$$y_i = \sum_{c=1}^6 z_{ic} \mu_c + \sum_{k=1}^8 x_{ik} \underline{\beta}_k + \underline{\varepsilon}_i,$$

where z_{ic} is an indicator variable, which is equal to one if individual i is a member of subpopulation c , and otherwise zero. The mean for subpopulation c is described by the parameter μ_c . Note that in this model we assume that the founder effects $\underline{\beta}_k$ will be equal for all subpopulations, which means that we assume that there are no QTL by genetic background effects.

We ran 1,000 simulations assuming no QTL to estimate the distribution of the P -values under the null hypothesis. These yield an estimate of $T = 2.6$ for the 1% chromosome-wide significance threshold of $-\log_{10}(P\text{-value})$ for the single cross analysis, and $T = 2.8$ for the combined analysis. The 1% chromosome-wide significance threshold corresponds to a 5% genomewide threshold for five chromosomes of length 100 cM each.

Some results and discussion

Example of the calculation of genetic predictors

The marker information for the four founders and for one inbred line (D51) is given in Table 2 for one particular simulated four-way cross. Five of the markers contain unambiguous information regarding the origin of the alleles of the F4-line, which means that for these positions it is known from which founders the alleles are inherited. For most of these markers this is immediately clear from Table 2. For marker M6, the situation is a little bit more complicated (Fig. 2). Both Sha and C24 have genotype AA,

Table 2 Marker scores for chromosome 1 for the four founders and for one four-way recombinant inbred line in the F4 generation, D51

ID	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
Sha	CC	AA	AA	DD	DD	AA	CC	CC	CC	DD	CC
Cvi	CC	CC	AA	DD	BB	BB	DD	AA	BB	CC	BB
Ler	BB	DD	CC	AA	AA	DD	DD	CC	BB	AA	DD
C24	BB	DD	BB	DD	BB	AA	BB	CC	CC	DD	DD
D51	CC	AA	AA	DD	BB	AB	DD	AA	BB	CC	BB

The marker scores shown in bold indicate scores for which the origin of the locus is uniquely identified. For example, for the second marker, located at 10 cM on chromosome 1, recombinant inbred line D51 is identical by descent with founder Sha

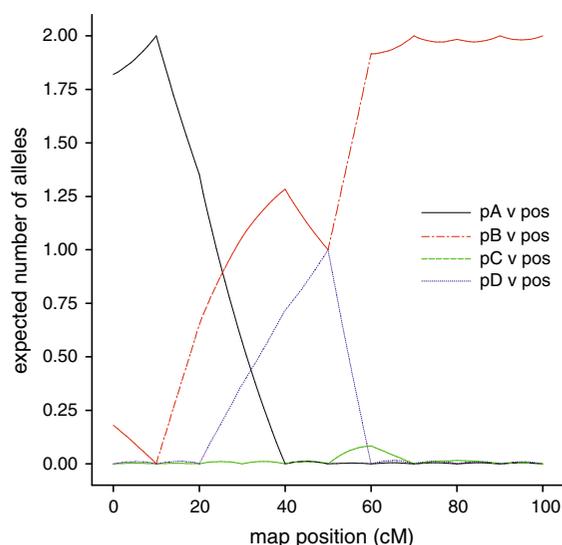


Fig. 3 Expected number of alleles for each founder as function of the position on the first chromosome, for one individual (D51). The curves are calculated using the HMM algorithm and using the marker scores as given in Table 2

so this seems to imply that allele A could have been inherited from each of these two founders. However, because individual D51 has the B-allele, derived from the founder Cvi, this implies that the A-allele cannot be inherited from Sha, and thus the A-allele is inherited from founder C24.

For the ambiguous markers and evaluation points between markers we can calculate the expected number of alleles originating from each founder, or the genetic predictors, by using the HMM algorithm. Figure 3 shows the genetic predictors along the first chromosome, and calculated at evaluation points

1 cM apart. As can be seen from this figure, there is a high probability that inbred line D51 is identical by descent with founder Cvi between 60 and 100 cM. The reason that this probability is so high is because markers M8, M10, and M11, located at 70, 90, and 100 cM, respectively, are all identical by descent with founder Cvi, and the probability of a double cross over between two markers is relatively small.

QTL mapping for one simulated complex cross

First QTL mapping was performed for each subpopulation separately. The QTL was detected only in one subpopulation, namely $[Ler \times C24] \times [Cvi \times Sha]$. The QTL profile is shown in Fig. 4. Although the test statistic's peak is little above the significance threshold, its maximum is found at 75 cM, which corresponds to the simulated position. Thus, this example seems to indicate that a good prediction can be made for QTL position, even for a relatively sparse dense map. Further increase of marker density will only slightly increase the power to detect QTLs, and the precision of the estimated QTL location, in a subpopulation of 100 individuals.

The results of QTL mapping for the whole complex cross population are shown in Fig. 5. As expected the

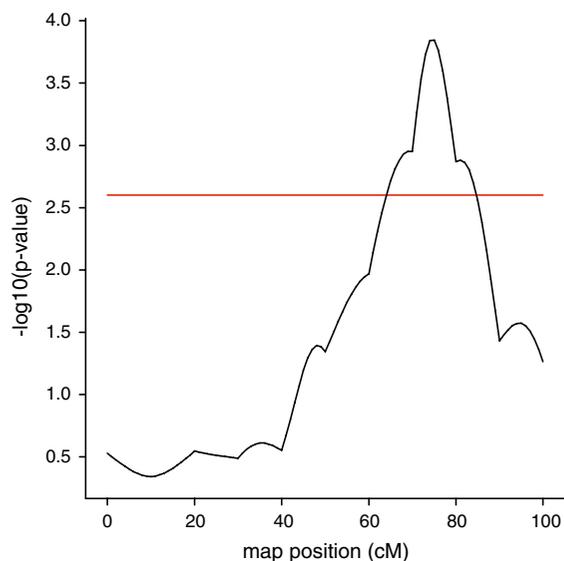


Fig. 4 QTL profile, given as $-\log_{10}(P\text{-value})$, along the first chromosome for the subpopulation $[Ler \times C24] \times [Cvi \times Sha]$, consisting of 100 four-way F4 lines. One single QTL was simulated, at 75 cM. The horizontal line is the 5% genomewide significance threshold

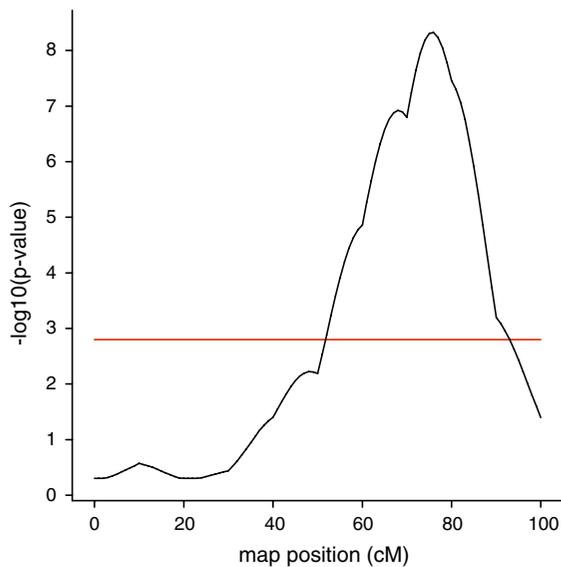


Fig. 5 QTL profile, given as $-\log_{10}(P\text{-value})$, along the first chromosome for the complex cross population consisting of six subpopulations, making a total of 600 four-way F4 lines. One single QTL was simulated, at 75 cM. The horizontal line is the 5% genomewide significance threshold

power of QTL detection improves substantially by increasing the number of subpopulations.

Power of QTL mapping

Using multiple simulations we can estimate the efficiency of the analysis of single subpopulations and of the analysis of the complete pedigree. Table 3 shows the results for 1,000 simulations. We can see

from Table 3 that the power to detect the QTL varies between 18% and 45% in the subpopulations of 100 individuals. As might be expected, the combined analysis of all the subpopulations will highly increase the power to detect the QTL, and it will also improve the accuracy of the estimation of the QTL position (Li et al. 2005). In principle, a similar accuracy and power can be obtained by using a population of 600 RILs obtained from a biparental cross. However, such an approach has several disadvantages. First of all, there is a risk that the QTL will not segregate, if the two parents are identical by state for that particular locus. In the case of a single QTL this risk can be reduced by choosing two contrasting parents. However, in general the number of QTLs is unknown, so choosing two contrasting parents cannot guarantee that all the QTLs will segregate. Another disadvantage of a biparental cross is that we can only estimate the contrast between the two parents, while in the complex cross we can estimate the allelic effects of the eight founders.

In the real data set several other complications can be expected to occur. First, multiple QTLs on the same chromosome will further complicate the QTL analysis, and will possibly lead to the detection of ghost QTLs (see e.g. Lynch and Walsh 1998). Other complicated aspects are dominance effects and epistasis. However, combining the analysis of all subpopulations in the complex cross is expected to increase the power to detect QTLs and provide also the possibility to search for QTL by genetic background effects.

Table 3 Power study using 1,000 simulated complex crosses

(Sub)population	Heritability	Power	Position (cM)
[Col × Kyo] × [Sha × Cvi]	0.10	0.24	74.74 (11.08)
[Col × Kyo] × [Eri × An]	0.13	0.45	73.58 (11.20)
[Col × Kyo] × [Ler × C24]	0.10	0.23	72.04 (16.07)
[Sha × Cvi] × [Eri × An]	0.10	0.28	74.51 (13.66)
[Sha × Cvi] × [Ler × C24]	0.12	0.28	72.59 (14.33)
[Eri × An] × [Ler × C24]	0.10	0.18	71.53 (19.16)
Complex cross	0.13	0.99	74.79 (4.56)

The QTL position was simulated at 75 cM. Each subpopulation consists of 100 individuals, resulting in a complex cross of 600 individuals in total. The power is calculated as the fraction of simulations for which the maximum peak is above the 5% genomewide significance threshold. The estimated position of the QTL is given by the mean position (with standard deviation between parentheses) across all simulations for which the peak was significant

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Alonso-Blanco C, Koornneef M (2000) Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. *Trends Plant Sci* 5:22–29
- Bishop CM (2006) *Pattern recognition and Machine Learning*. Springer
- Blanc G, Charcosset A, Mangin B et al (2006) Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. *Theor Appl Genet* 113:206–224
- Boer M, Wright D, Feng L, Podlich D et al (2007) A mixed-model Quantitative Trait Loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. *Genetics* 177:1801–1813
- Broman K (2005) The genomes of recombinant inbred lines. *Genetics* 169:1133–1146
- Charcosset A, Mangin B, Moreau L et al (2001) Heterosis in maize investigated using connected RIL populations. *Quantitative genetics and breeding methods: the way ahead*. INRA, Paris, France, pp 89–98
- Darvasi A, Soller M (1995) Advanced intercross lines: an experimental population for fine genetic mapping. *Genetics* 141:1199–1207
- Koornneef M, Alonso-Blanco C, Vreugdenhil D (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*. *Ann Rev Plant Biol* 55:141–172
- Lander ES, Green P (1987) Construction of multilocus linkage maps in human. *Proc Natl Acad Sci USA* 84:2363–2367
- Li R, Lyons MA, Wittenburg H et al (2005) Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. *Genetics* 169:1699–1709
- Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, Massachusetts
- Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proc IEEE* 77:257–286
- Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J Am Statist Ass* 82:605–610
- Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics* 50:1171–1177
- The Complex Trait Consortium (2004) The collaborative cross a community resource for the genetic analysis of complex traits. *Nat Genet* 36:1133–1137
- Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. *Theor Appl Genet* 95:1005–1011
- Weigel D, Nordborg M (2005) Natural variation in *Arabidopsis*. How do we find the causal genes? *Plant Physiol* 138:567–568
- Valdar W, Mott R, Flint J (2006) Simulating the collaborative cross: power of QTL detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics* 172:1783–1797