DEVELOPMENT OF HIGHLY RECOMBINANT INBRED POPULATIONS FOR QUANTITATIVE-TRAIT LOCUS MAPPING

by

PRASHANTH BODDHIREDDY

M.S., Kansas State University, 2005

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics Program
Department of Plant Pathology
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2009

Abstract

The goal of quantitative-trait locus (QTL) mapping is to understand the genetic architecture of an organism by identifying the genes underlying quantitative traits. It targets gene numbers and locations, interaction with other genes and environments, and the sizes of gene effects on the traits. QTL mapping in plants is often done on a population of progeny derived from one or more designed, or controlled, crosses. These crosses are designed to exploit correlation among marker genotypes for the purposes of mapping QTL. Reducing correlations between markers can improve the precision of location and effect estimates by reducing multicollinearity. The purpose of this thesis is to propose an approach for developing experimental populations to reduce correlation by increasing recombination between markers in QTL mapping populations especially in selfing species.

QTL mapping resolution of recombinant inbred lines (RILs) is limited by the amount of recombination RILs experience during development. Intercrossing during line development can be used to counter this disadvantage, but requires additional generations and is difficult in self-pollinated species. In this thesis I propose a way of improving mapping resolution through recombination enrichment. This method is based on genotyping at each generation and advancing lines selected for high recombination and/or low heterozygosity. These lines developed are called SA-RILs (selectively advanced recombinant inbred lines). In simulations, the method yields lines that represent up to twice as many recombination events as RILs developed conventionally by selfing without selection, or the same amount but in three generations, without reduction in homozygosity. Compared to methods that require maintaining a large population for several generations and selecting lines only from the finished population, the method proposed here achieves up to 25% more recombination.

Although SA-RILs accumulate more recombination than conventional RILs and can be used as fine-mapping populations for selfing species, the effectiveness of the SA-RIL approach decreases with genome size and is most valuable only when applied either to small genomes or to defined regions of large genomes. Here I propose the development of QTL-focused SA-RILs (QSA-RILs), which are SA-RILs enriched for recombination in regions of a large genome

selected for evidence for the presence of a QTL. This evidence can be derived from QTL analysis in a subset of the population at the F_2 generation and/or from previous studies. In simulations QSA-RILs afford up to threefold increase in recombination and twofold increase in accuracy of QTL position estimate in comparison with RILs. The regional-selection method also shows potential for resolving QTL linked in repulsion.

One of the recent Bayesian methods for QTL mapping, the shrinkage Bayesian method (BayesA (Xu)), has been successfully used for estimating marker effects in the QTL mapping populations. Although the implementation of the BayesA (Xu) method for estimating main effects was described by the author, the equations for the posterior mean and variance, used in estimation of the effects, were not elaborated. Here I derive the equations used for the estimation of main effects for doubled-haploid and F_2 populations. I then extend these equations to estimate interaction effects in doubled-haploid populations. These derivations are helpful for an understanding of the intermediate steps leading to the equations described in the original paper introducing the shrinkage Bayesian method.

DEVELOPMENT OF HIGHLY RECOMBINANT INBRED POPULATIONS FOR QUANTITATIVE-TRAIT LOCUS MAPPING

by

PRASHANTH BODDHIREDDY

M.S., Kansas State University, 2005

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics Program
Department of Plant Pathology
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2009

Approved by:

Major Professor
James C. Nelson

Copyright

Abstract

The goal of quantitative-trait locus (QTL) mapping is to understand the genetic architecture of an organism by identifying the genes underlying quantitative traits. It targets gene numbers and locations, interaction with other genes and environments, and the sizes of gene effects on the traits. QTL mapping in plants is often done on a population of progeny derived from one or more designed, or controlled, crosses. These crosses are designed to exploit correlation among marker genotypes for the purposes of mapping QTL. Reducing correlations between markers can improve the precision of location and effect estimates by reducing multicollinearity. The purpose of this thesis is to propose an approach for developing experimental populations to reduce correlation by increasing recombination between markers in QTL mapping populations especially in selfing species.

QTL mapping resolution of recombinant inbred lines (RILs), is limited by the amount of recombination RILs experience during development. Intercrossing during line development can be used to counter this disadvantage, but requires additional generations and is difficult in self-pollinated species. In this thesis I propose a way of improving mapping resolution through recombination enrichment. This method is based on genotyping at each generation and advancing lines selected for high recombination and/or low heterozygosity. These lines developed are called SA-RILs (selectively advanced recombinant inbred lines). In simulations, the method yields lines that represent up to twice as many recombination events as RILs developed conventionally by selfing without selection, or the same amount but in three generations, without reduction in homozygosity. Compared to methods that require maintaining a large population for several generations and selecting lines only from the finished population, the method proposed here achieves up to 25% more recombination.

Although SA-RILs accumulate more recombination than conventional RILs and can be used as fine-mapping populations for selfing species, the effectiveness of the SA-RIL approach decreases with genome size and is most valuable only when applied either to small genomes or to defined regions of large genomes. Here I propose the development of QTL-focused SA-RILs (QSA-RILs), which are SA-RILs enriched for recombination in regions of a large genome selected for evidence for the presence of a QTL. This evidence can be derived from QTL

analysis in a subset of the population at the F_2 generation and/or from previous studies. In simulations QSA-RILs afford up to threefold increase in recombination and twofold increase in accuracy of QTL position estimate in comparison with RILs. The regional-selection method also shows potential for resolving QTL linked in repulsion.

One of the recent Bayesian methods for QTL mapping, the shrinkage Bayesian method (BayesA (Xu)), has been successfully used for estimating marker effects in the QTL mapping populations. Although the implementation of the BayesA (Xu) method for estimating main effects was described by the author, the equations for the posterior mean and variance, used in estimation of the effects, were not elaborated. Here I derive the equations used for the estimation of main effects for doubled-haploid and F_2 populations. I then extend these equations to estimate interaction effects in doubled-haploid populations. These derivations are helpful for an understanding of the intermediate steps leading to the equations described in the original paper introducing the shrinkage Bayesian method.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIL: Advanced intercross lines

ARI: Advanced recombinant inbred lines

BC: Backcross

CI: Confidence interval

CIM: Composite interval mapping

FP: False positive

IRI: Intercross recombinant inbred lines

LA: Linkage analysis

LDA: Linkage disequilibrium analysis

QSA-RIL: QTL-focused selectively advanced inbred lines

QTL: Quantitative-trait locus

RE: Recombinant enrichment

RIL: Recombinant inbred lines

RIX: Recombinant intercross lines

SA-RIL: Selectively advanced recombinant inbred lines

SIM: Simple interval mapping

SMA: Single marker analysis

TP: True positive

Acknowledgements

I would like to thank Dr. James C. Nelson for his constant guidance, advice and helpful discussions during my doctoral research. He has been very patient with me in regards to teaching the importance of the skills necessary in communicating research ideas. His constructive criticism has greatly improved the quality of my research. I am deeply indebted to Dr. Jean-Luc Jannink for his timely advice, valuable interaction and constant encouragement. His motivating comments in regards to my research work are invaluable. I would like to thank Dr. Mitch Tuinstra for his valuable comments on my thesis. I would like to express my appreciation to Dr. Haiyan Wang for serving in my committee.

I would like to dedicate this thesis to my parents Boddireddy Prakash Reddy and Boddireddy Rama Devi and my brother Boddireddy Srikanth Reddy.

CHAPTER 1 - QTL Mapping Methods and Mating Designs

Most agronomically important traits are quantitative and display continuous variation. Unlike qualitative traits such as coat color and blood type, for which phenotype (trait value) falls into discrete classes and genotype can be deduced from phenotype alone, quantitative traits such as grain yield, weight gain, and milk yield are influenced by one or more genes, gene interactions and environmental factors, so that genotype cannot be deduced reliably from phenotype.

It is desirable to identify the genes underlying quantitative traits, for improving the performance and/or productivity of plants and animals and for understanding the genetic architecture. This last includes gene numbers and locations, interaction with other genes and environments, and the sizes of gene effects on the traits. DNA sequence variants (*markers*) that show association with phenotype across individuals facilitate the study of genetic architecture. Due to limitations of technology to identify and genotype markers throughout the whole genome in many individuals, identifying *quantitative trait loci* (QTL) has been difficult. With availability in recent years of markers across the genome, the phenotypic variation among individuals can be partitioned into components, one of which is genetic variation as estimated from marker genotypic variation. This partitioning allows locating, or *mapping*, QTL.

In plants, calculating gene-level variation and identifying QTL location is done on a population of progeny derived from one or more designed, or controlled, crosses. These crosses are designed to exploit correlation among marker genotypes. Pairwise correlation between all markers is perfect during the initial generation of crosses and declines in subsequent generations of mating so that only markers near a QTL show correlation with it. The precision with which QTL can be located increases as this breakdown in correlation increases. The factor that increases the breakdown is the *recombination* that occurs during the mating. Recombination is a random event, and some plants experience more recombination than others. In this thesis, methods are proposed to increase the frequency of recombination by selection in a QTL mapping population.

In this chapter, the following concepts essential to understanding the process of QTL mapping are presented: association between genotypic and phenotypic variation, genetic variation and meiosis, recombination and genetic distances, mating designs, and statistical approaches for QTL mapping.

Genotypic and phenotypic variation and genetic map distances

Evaluating association between genetic and phenotypic variation

Genetics is the science of identifying predictive associations between gene-level, or DNA variation and gross phenotypic variation. Genetic variation arises from DNA sequence variation in *genes* encoding RNAs and proteins. Genes lie at positions known as *loci* on linear structures called *chromosomes*. Chromosomes come in pairs with one member inherited from the father and other from the mother. The same genes lie in the same order on both maternal and paternal chromosomes, but genes may not have identical DNA sequences. Each variant form of a gene is known as an *allele*.

A natural way to evaluate the contribution of genetic to phenotypic variation is to construct a linear regression model in which phenotype is modeled as a linear combination of effects of genes. In such a model the gene levels or *genotypes* are expressed as number of maternal (or paternal) alleles carried at a locus. Because in general, gene alleles are not known, DNA marker genotypes are used instead. If a marker is determined to be predictive of a phenotype, it may be inferred that a QTL is present nearby on the same chromosome as the marker.

Determining the subset of markers that best explain the variation of a trait is a *model-* or *variable-selection* process. Typically the model space is too large to evaluate all possible models to find the one closest to the underlying true model, and multicollinearity arising from correlation between markers makes the problem more challenging. The correlation between markers increases with decreasing distance and recombination between them.

Genetic variation and meiosis

Genetic variation, or variation of genotypes among progeny, arises from a process in which each chromosome inherited from a parent is formed by segmental mixing of the parent's own maternal and paternal chromosomes. The mixing occurs during *meiosis*, or reductive

division, in the reproductive cells of the organism, by a process called crossing over, illustrated schematically in Fig 1.1. The products of meiosis are thus a mosaic of grandparental chromosomes, and are packaged in sperm or egg cells (*gametes*) for subsequent union by fertilization. Gene order is conserved in crossing over. Because the locations of crossover events are random, even progeny derived from the same parents show genotypic variation.

Recombination fraction, map distance and mapping functions

The basis for correlation among markers is the genetic distance between them. The information about distances between markers on a chromosome is needed for inferring expected genotypes of an unknown gene or QTL and for prediction of QTL location. The estimate of distance between two markers, computed in a population of plants derived from controlled crosses, is based on their *recombination fraction* (θ), which is the ratio of their *recombinants* to the total number of gametes observed following exactly one generation of meiosis. Consider two markers A and B with alleles A_1, A_2 and B_1, B_2 and a population of individuals with *genotype* A_1B_1/A_2B_2. Although the term *genotype* describes gene levels in the context of a single locus, it is also used to describe the configuration of alleles at one or more loci in an individual. Via meiosis, an individual provides one of two kinds of gametes to progeny: parental gametes A_1B_1 and A_2B_2 and recombinant gametes A_1B_2 and A_2B_1. The number of recombinants observed in the population gives an estimate of the number of crossovers between the markers. More crossover events are expected to occur when markers are far apart than when they are close to each other. Markers A and B are said to be *linked* when the recombination fraction between the markers A and B, θ_{AB} is less than 0.5. The linear arrangement of markers calculated from their recombination fractions is known as a *linkage* or *genetic map.*

Recombination fraction, θ, is underestimated as the physical distance between markers increases. This is because with increase in distance the chance of two crossovers within an interval increases. With respect to markers A and B, gametes that result from two (or any other even number of) crossover events between these markers are indistinguishable from parental gametes. Because of double crossovers, θ are not additive, *i.e.,* $\theta_{AC} = \theta_{AB} + \theta_{BC} - 2\theta_{AB}\theta_{BC}$. θ can be further biased because of meiotic recombination *interference.* This happens when a crossover event at a given location prevents other crossover events in close proximity.

It is desirable to display the genetic relationships between loci on a graphical map in which distances between loci, unlike their θs, are additive. Conversion of these θs to distances

may be done with any of several *mapping functions* (Sturtevant 1913, Haldane 1919, Kosambi 1944, Rao et al. 1977, Karlin 1984). Of these, Haldane's and Kosambi's are used most often. Haldane's function, expressed as $m_{AB} = -0.5 \log(1-2\theta_{AB})$, models the crossover events on chromosomes as a Poisson process, ignoring interference. The Kosambi function, $m_{AB} = -0.5 \tanh^{-1}(2\theta_{AB})$, takes interference into account. The Haldane mapping function has been used for conversion of recombination fraction into mapping distance in all the simulations conducted for this thesis.

Mating designs used for QTL mapping

QTL mapping populations are created from one or more controlled crosses. Mating designs set up correlation between markers and QTL. The populations used for QTL analysis can be classified into three categories based on the kind of crosses used to generate them and duration of generation time: inbred, outbred and natural populations. Though only inbred population designs are relevant to this thesis, other designs are described for completeness.

Inbred

These populations are derived from the progeny resulting from a cross of two inbred parents. The parents are individuals whose alleles are identical at most loci because of self pollination (in plants) or mating among close relatives, over several generations. Inbred designs are most appropriate for QTL mapping in plant species such as wheat, barley, maize, *Arabidopsis*, and *Brassica,* and animal species with short generation time, such as mouse. Several kinds of populations can be derived from F_1 plants. A backcross (BC_1) population is derived by crossing F_1 progeny to one of the parents. A selfed population (F_{n+1}) is derived by self-pollination of F_n progeny. A recombinant inbred line (RIL) population is produced by several generations of selfing or sib mating. An advanced intercross line (AIL_x) population is derived by random and sequential intercrossing for a few generations starting with the F_2 (Darvasi and Soller 1995), with the subscript x denoting the number of intermating generations. An intermated recombinant inbred line (IRI_x) is derived by selfing of AIL_x populations until a desired level of homozygosity is achieved (Liu et al. 1996) and these populations are also denoted as IRIP(i,j) to denote i intercrossing generations and j selfing generations (Lee et al. 2002, Sharopova et al. 2002). Other notations for IRIs include ARI (advanced recombinant inbred lines), IRIL (intercross recombinant inbred lines) and RIX (recombinant intercross lines).

Outbred

These designs are appropriate for domestic species such as cattle, where it is impractical to produce inbred lines. The mating scheme is designed to create three kinds of families: half-sib, full-sib, and granddaughter-based families (Weller et al., 1990). Each family in a half-sib population consists of progeny derived from the same father (or mother) and randomly chosen mothers (or fathers). Each family in a full-sib population consists of progeny derived from the same father and mother. Granddaughter-based families are similar to half-sib families except that the genotyping is done on sons of sires and phenotype is measured on the daughters of the sons (granddaughters).

Natural

In natural populations, where it is impractical to enforce a mating scheme, experimental design lies in prudent selection of a subset from an existing population. Sib-pair population (Haseman and Elston, 1972) data, as the name suggests, is a collection of sib genotype and phenotype data from several families. It is a commonly used design for QTL mapping in humans. The rationale for this design is that the variation of the trait value between sibs is expected to be inversely proportional to the number of alleles that are identical by inheritance for the genes controlling the trait. In case–control designs, QTL mapping populations consist of two sets of subjects, one of affected and the other of unaffected. The rationale for this design is that the differences in the frequencies of the genotypes between affected and unaffected sets suggest possible association between the disease and the genes causing it. Several extensions of case–control design have been suggested. One such extension is a father–mother–progeny trio-based population (Spielman et al. 1993). In this population, all the progeny are affected and markers are useful only when they are heterozygous in at least one of the parents.

QTL mapping methods

Quantitative trait models

The genetic variation observed in quantitative traits can be modeled in two ways. The *infinitesimal* model (Fisher 1918) assumes that a trait value is determined by an infinite number of unlinked loci, each with an infinitely small effect. The *finite loci* model (Shrimpton and Robertson 1998, Otto and Jones 2000, Hayes and Goddard 2001, Xu 2003) assumes that a

smaller number of loci have a large effect on a trait (major QTL), and other loci (minor QTL) have a small effect.

Statistical model of a trait

For a quantitative trait with one QTL with alleles *Q* and *q,* the trait value of an individual *i* can be modeled as

$$y_i = \mu + ax_i + bz_i + e_i, \tag{1.1}$$

where *a* is *additive* effect and *b* is *dominance* effect*,* which is the deviation from additivity, observed due to interaction of alleles *Q* and *q*. The variables x_i = {–1 for genotype *QQ* , +1 for genotype *qq* and 0 for genotype *Qq*} and z_i = {+1 for genotype *Qq* and 0 for genotypes *QQ* and *qq*} are computed from the QTL genotype and $e_i \sim N(0, \sigma^2)$ indicates Gaussian noise with variance σ^2. For a trait with several QTL, the model can be extended as

$$y_i = \mu + \sum_{j=1}^{N_Q} a_j x_{ij} + \sum_{j=1}^{N_Q} b_j z_{ij} + e_i, \tag{1.2}$$

where N_Q denotes the number of QTL. The additive and dominance effects are treated as fixed effects. For simplicity, interaction effects and other fixed and random effects are ignored here. A typical data set used for QTL mapping is shown in Table 1.1.

Several statistical approaches are available to estimate the parameters a_j, b_j, N_Q and the position of QTL. These approaches can be broadly classified into four categories: least-squares, likelihood, Bayesian, and semi- and non-parametric methods. In all of these approaches, neither position nor genotype of a QTL is known in advance. The parameter estimates are based on markers, of which some may show high correlation with the QTL. Three widely used QTL mapping analyses amenable to both likelihood and regression methods are described briefly here.

Single-marker analysis (SMA)

In SMA, the model used to estimate parameters is similar to the trait model for a single QTL as shown in equation 1.1. The indicator variables x_i and z_i are based on the marker genotype. The parameters are estimated separately at each marker position. SMA is useful when marker density is high and/or the linkage map is not known.

Simple interval mapping (SIM)

In simple interval mapping (Lander and Botstein 1989), the parameters are estimated separately at each candidate position, for instance at every 1 or 2 cM across the genetic map. The model used to estimate parameters is similar to (1.1). In the regression-based interval mapping method (Haley and Knott 1992), x_i and z_i are replaced with indicators based on expected QTL genotypes, E(x_i) and E(z_i) as shown in (1.3), which are estimated using markers flanking the candidate position. The inference made about QTL position is more accurate in SIM than in SMA, because of the inferred QTL genotypes in SIM. The variables E(x_i) and E(z_i) used in the model are:

$$E(x_i) = [-1,0,1]\begin{bmatrix} p(QQ \mid M_1, M_2, r_1, r_2) \\ p(Qq \mid M_1, M_2, r_1, r_2) \\ p(qq \mid M_1, M_2, r_1, r_2) \end{bmatrix}_i, E(z_i) = [0,1,0]\begin{bmatrix} p(QQ \mid M_1, M_2, r_1, r_2) \\ p(Qq \mid M_1, M_2, r_1, r_2) \\ p(qq \mid M_1, M_2, r_1, r_2) \end{bmatrix}_i, \quad (1.3)$$

where $p(x|y)$ denotes the conditional probability of x given y and M_1, M_2 denotes marker genotypes flanking the putative QTL and r_1 and r_2 the recombination fractions between marker M_1 and QTL and between QTL and M_2 respectively. The conditional probability of a QTL genotype given flanking marker genotypes can be computed as the ratio of the expected probability of three consecutive marker genotypes to the expected probabilities of the two flanking marker genotypes under the given mating design. The algebraic forms of the conditional probabilities of a QTL genotype given flanking marker genotypes are described in Haley and Knott (1992), Luo and Kearsey (1992) and Lynch and Walsh (1998).

Composite interval mapping (CIM)

CIM is an extension of SIM that incorporates into the model the effects of QTL elsewhere than at the putative QTL position, via markers known as cofactors (Jansen 1993; Zeng 1993, 1994). Cofactors are selected by multiple regression using well-known stepwise model-selection algorithms such as forward, backward, and forward–backward selection. In forward selection, for example, the model begins with no markers and adds, at each step, the marker that explains the highest phenotypic variation, evaluated as the *F* value. In subsequent steps, only markers not already included in the model are evaluated and the marker with the highest *F* value is included. The steps are repeated until the highest *F* value is smaller than a predefined *F* value, known as *F-to-enter*.

To compute a QTL evidence-score profile along the chromosome, a test statistic is computed at each candidate position and only cofactors that are at least *x* cM away from this position are included in the model. The window size, *x,* is chosen to exclude markers (close to the test position) whose genotypes are highly correlated with the genotypes of the putative QTL. In practice, the analyst chooses a window size of 10–30 cM. The statistical model for CIM is as follows:

$$y_i = \mu + aE(x_i) + \sum_{j=1}^{N_Q} b_j z_{ij} + e_i, \qquad (1.4)$$

where N_Q is the number of cofactors and b_j is the effect of cofactor marker j. When the same data are analyzed with CIM and SIM, inferences made about the location and additive and dominance effect estimates of a QTL are more precise with CIM than with SIM, because the cofactors fitted with CIM reduce the sampling variance of the test statistic by absorbing some residual genetic variance due to other QTL. The choice of the *F-to-enter,* which determines the number cofactors and the window size, is subjective. Consequently analysts could reach different conclusions about QTL estimates.

Threshold value for declaring a QTL

For declaring a QTL, a likelihood ratio statistic known as the LOD (logarithm of odds) is commonly used. LOD = $\log_{10}$[(likelihood of H_1)/(likelihood of H_0)], where H_0 and H_1 denote the hypotheses of absence and presence of a QTL around the test position on the genetic map. The permutation test proposed by Churchill and Doerge (1994) is used to find an appropriate threshold value for declaring a QTL significant. In a permutation test, an empirical distribution for the test statistic is created from permuted data sets. In each set, phenotypic records are randomly shuffled among plants, so that the association between the marker data and the phenotype data is disrupted. The maximum value of a test statistic, commonly the LOD score, among the values computed across all candidate positions, is recorded from each permuted data set. The resulting array of maximum statistics approximates their null-hypothesis distribution. The statistic corresponding to the $100(1 - \alpha)$ percentile of the values in the distribution is taken as the threshold value used for significance testing (for the presence of a QTL) at level α.

Evaluating QTL mapping resolution

Mapping resolution, which is the precision of the QTL position estimate, is often quantified as the confidence interval (CI) of the QTL location. CI are useful in planning future experiments aimed at identifying the DNA sequence variation that explains the phenotypic variation. Several ways of computing CI have been proposed (Lander and Botstein 1989, Van Ooijen 1992, Darvasi 1993, Mangin et al. 1994, Visscher et al. 1996, Darvasi 1997, Visscher and Goddard 2003). The LOD drop-off method proposed by Lander and Botstein (1989) is often used. In this method, the location corresponding to a decrease in the LOD score of 1 or 2 units from a peak in the LOD score profile gives a 96.8–99.8% CI (Mangin et al. 1994). The CI found using this approach is biased in small- (<100) to medium-sized (<200) populations, and when the QTL effect is small (Van Ooijen 1992; Mangin et al. 1994). In simulations, the 95% CI is computed from the distribution of the QTL position estimates (the position corresponding to the maximum LOD score). Estimates of the QTL position in repeated simulations forms the distribution. The CI obtained through this method sometimes comprises an entire chromosome (Darvasi et al. 1993). A bootstrapping approach (for computing the CI) proposed by Visscher et al. (1996) requires computing a distribution for a QTL position. The distribution is produced from QTL analysis of several populations, where each population consists of N individuals sampled with replacement from the original N observations. A simple formula proposed by Darvasi and Soller (1997) for computing a 95% CI has the form

$$3000/(kN(a+d)^2,$$

where k is 1 for BC_1 and 2 for F_2, N is the population size, and a and d are estimated additive and dominance effects. This formula was based on simulation results, under the assumption that an infinite number of markers is scored. Visscher and Goddard (2003) proposed a similar formula,

$$CI(1-\beta) = 400kX_{(1-\beta)}/na^2,$$

where k takes the values 1 for a backcross and 2 for an F_2 design and $X_{(1-\beta)}$ denotes the threshold of a central χ^2 distribution with 1 degree of freedom corresponding to a cumulative density of (1-β).

The typical 95% CI for BC_1 and F_2 populations is 10–40 cM. The CI for RIL populations is two- to fourfold narrower than that in F_2 populations, and the CI for AIL populations is 5–10 times narrower, depending on the number of intercrossing generations.

Figures and Tables

Figure 1.1 Genotypes and gametes resulting from a cross

Panel *a* shows genotypes of two individuals, each represented as two strings. Each string represents alleles inherited together, a gamete, from the individual's parents. The symbol 'X' in blue denotes the operation of crossing the two individuals shown on the left and right-hand sides. Panel *b* shows in green the locations of crossovers between two strings: two crossovers on the left and one on the right. Panel *c* shows the gametes formed after meiosis. Panel *d* shows the genotype of two progeny resulting from the cross.

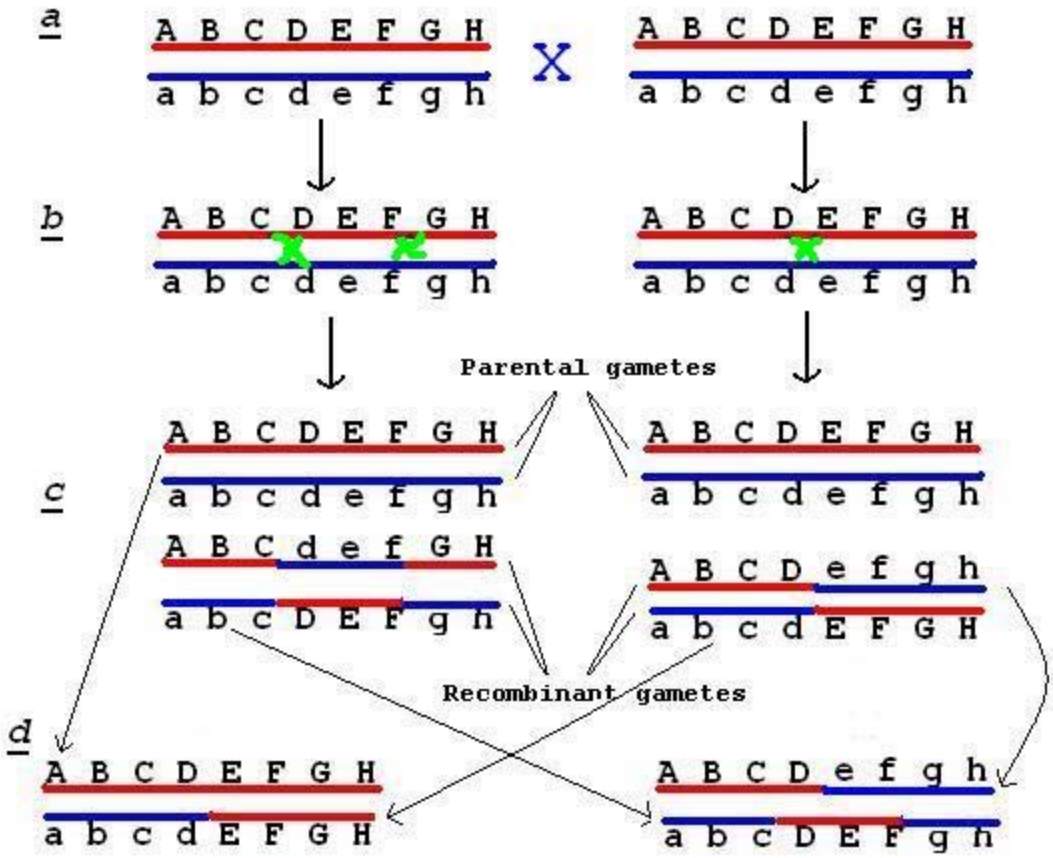


Figure 1.2 LOD-drop-off method and confidence interval (CI)

The X axis shows markers along a genetic map and the Y axis the LOD score. A two-LOD-decrease indicates 99.5% of the confidence interval for the location estimate.

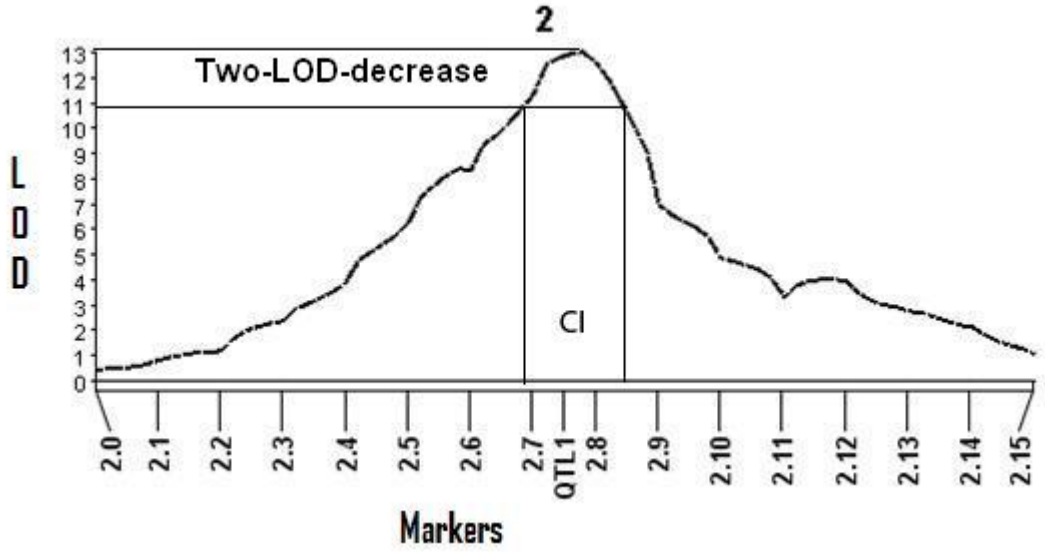


Table 1.1 A data set from an F2 population used for QTL mapping

Each row shows the trait and genotype record for an individual. Each number in the locus column indicates number of paternal (maternal) alleles.

Individual	**Trait**	**Chromosome 1**			**Chromosome 2**			**…..**	**Chromosome m**		
		Locus1	**Locus2**	**….**	**Locus1**	**Locus2**	**…**	**….**	**Locus1**	**Locus2**	**….**
Ind_1	3.5	0	1	..	2	1	..	..	0	0	..
Ind_2	5.5	2	1	..	0	1	..	..	1	0	..
Ind_3	6.6	1	0	..	2	2	..	..	0	0	..
.	.	..	..	..	..	..	..	..	..	..	..
Ind_{n-1}	5.5	0	2	..	1	1	..	..	0	0	..
Ind_n	4.7	1	1	..	2	1	..	..	1	1	..

CHAPTER 2 - Selective Advance for Accelerated Development of Recombinant Inbred QTL Mapping Populations

Abstract

QTL mapping resolution of RILs is limited by the amount of recombination they experience during development. Intercrossing during line development can be used to counter this disadvantage but requires additional generations and is difficult in self-pollinated species. It is desirable to improve mapping resolution for success of marker-assisted selection and map-based cloning. Here I propose a way of achieving this through recombination enrichment. This method is based on genotyping at each generation and advancing lines selected for high recombination and/or low heterozygosity. In simulations, the method yields lines that represent up to twice as many recombination events as RILs developed conventionally by selfing without selection, or the same amount but in three generations, without reduction in homozygosity. Compared to methods that require maintaining a large population for several generations and selecting lines only from the finished population, the method proposed here achieves up to 25% more recombination. The precision of QTL location is increased by up to 10% with negligible drop in QTL detection power at average marker intervals of 5 cM or less.

Introduction

High precision of QTL location estimates is desired in marker-assisted selection and positional cloning of QTL. This precision, also referred to as QTL map resolution, is low in biparentally derived populations such as backcross and F_2, with a 95% confidence interval of typically 10 to 30 cM for QTL location (Kearsey and Farquhar, 1998; Nadeau and Frankel, 2000; Dekkers and Hospital, 2002). For a given mapping population, QTL mapping resolution may be improved by statistical refinements (Zeng, 1993; Jansen and Stam, 1994; Jiang and Zeng, 1995; Korol et al., 1995, 1998) and increased marker density. But it is population size and mating design that determine the number of meioses and thereby the maximum number of effective

meiotic crossovers, or recombinations, that occur during line development. Higher recombination is obtained both by increasing population size and by increasing the number of generations beyond the F_2 in developing the finished population.

However, increasing population size to enhance mapping precision incurs higher phenotyping costs, especially for multiyear or multilocation studies and for expression QTL experiments, where microarray expression data are used as phenotypes (e.g., Hubner et al., 2005). Another way to improve QTL map resolution is to increase recombination per individual. This is conventionally achieved with multigeneration populations such as recombinant inbred lines (RILs), advanced intercross lines (AILs; Darvasi and Soller, 1995), and intermated recombinant inbreds (IRIs; Lee et al., 2002). The use of IRIs, which require intercrossing for a few generations and selfing to a desired level of homozygosity, improves QTL map resolution (Winkler et al., 2003). Five generations of intermating in maize (*Zea mays L.*) resulted in nearly fourfold improvement of QTL map resolution compared to F_2 populations (Lee et al., 2002). Balint-Kurti et al. (2007) resolved one of the QTL found in a RIL population (Carson, 1998; Carson et al., 2004) into two QTL of opposing effects using an IRI population derived from the same parents. The confidence intervals in the IRI population were 2.5 to 50 times narrower than those in the RILs. Intercross populations have also been used for fine mapping and positional cloning of genes in mice (Becanovic et al., 2006; Yu et al., 2006, Wang et al., 2003).

In selfing species such as wheat (*Triticum aestivum L.*), barley (*Hordeum vulgare L.*), rice (*Oryza sativa L.*), *Arabidopsis*, and *Brassica,* intercrossing is laborious. Hundreds of plants must be manually pollinated over several generations, with each cross possibly subject to a low success rate, incompatible flowering times of parents, and introduction of male-sterility or expression of hybrid lethality genes. For these species, RILs are a common choice of QTL mapping design, but RILs are limited by the number of recombinations they accumulate, so that mapping resolution is limited.

A few methods have been proposed to select from a large genotyped population a subset that preserves the most information for QTL mapping. Jannink (2005) proposed two selection strategies: *maxRec* and *uniRec*. Both select lines with high recombination across the genome, but *uniRec* imposes greater genomewide uniformity of the distribution of recombination. Xu et al. (2005) proposed selection of recombination-enriched individuals based on an objective function that minimizes the sum of squared bin lengths. A bin was defined on a sample of individuals as a

marker interval within which there are no recombinations in any sampled individual and that is bounded on either side by a recombination in at least one individual or by the end of a linkage group. Jin et al. (2004) proposed selecting lines that are genotypically dissimilar using a criterion known as minimum moment aberration. Compared to randomly selected lines of the same population size, the selected set of individuals showed an increase in either the precision of QTL location (Jannink, 2005), the specificity (proportion of true positives [TPs] among detected QTL [Xu et al., 2005]), or the sensitivity (proportion of TPs among simulated QTL) for highly heritable traits (Jin et al., 2004).

Although the methods proposed by Jannink (2005) and by Xu et al. (2005) select a recombination-enriched subset, the application of these methods to RILs requires maintaining a large set of lines, of which only a subset is retained for phenotyping, for several (seven to eight) generations. With decreases in genotyping costs, it may well be less expensive to genotype and then discard a plant than to advance it for another generation. Here we extend the *maxRec* strategy to generate highly recombinant RILs by selecting highly recombinant individuals during the inbreeding process rather than in the finished population.

Materials and Methods

Simulation of RIL population

Let N be the number of RILs to be generated and p the number of progeny genotyped for each line at each generation. Recombinant inbred lines were simulated as follows: F1 plants were generated from the cross of two inbred parents and F_2 progeny were then generated by selfing. $N \times p$ F_2 plants were then genotyped and N of these were selected based on one of four selection strategies described below. From each of the N selected F_2 parents, p F_3 progeny were generated and genotyped, and one progeny from each F_3 family was selected for advancing to the next generation. This pedigree selection process was repeated to the desired degree of homozygosity. For comparing with other selection methods described below, this option will be called *FamSel*. The parameter levels used for simulation were as follows: heritability, 0.2, 0.4 and 0.6; family size, 2, 4, and 10; and marker spacing, 5, 10, and 20 cM. Two maps, one with a single chromosome and the other with seven chromosomes, were simulated. The length of each chromosome on either map was 150 cM. All simulations and QTL analysis were conducted with QGene (Joehanes and Nelson, 2008) on a computer cluster with four dual-core processors.

Selection strategy

Let a selection strategy be defined as the sequence of selection and advancing methods applied at every generation. Let *r* denote random selection of a progeny from a family; *R*, selection on maximum recombination; *H*, selection against heterozygosity; and *D*, the operation of making doubled haploids and selecting on recombination. Selection against heterozygosity was done based on a heterozygosity index (i.e., average number of heterozygotes per hundred markers). The individual with the lowest heterozygosity index was selected from each family. A strategy is described by a sequence of characters, each character indicating the selection method and its position indicating the generation of selection. For example, string *RrrRrrrr* denotes selection for recombination in generations 1 and 4 (after the F_1) and random selection in the rest. Conventional RIL development is described by *rrrrrrrr*. The strategies used for illustration here are *D*, *RD*, *RRD*, and *RRRHHH*, representing reduced-generation strategies, and *RRRrrrrr*, *RrRrRrRr*, and *RRRRRRRR*, representing early-, alternate-, and all-generation selection strategies, respectively. Recombinant inbred lines subjected to selection in at least one generation were called selectively advanced RILs (SA-RILs).

Other selection strategies

A strategy in which selection was done at every generation on recombination in the whole population without family constraints was denoted by *SimpleSel*, and a strategy in which selection was done only at the final generation in a large population was denoted by *FinSel*. In FinSel, if *N* plants were chosen, a population of size *Np* was maintained through all generations.

Chromosome map

On the one-chromosome map, one QTL was simulated at 54 cM. On the seven-chromosome map, two QTL on each of chromosomes 1 and 2, one QTL on each of chromosomes 3, 4, and 5, and no QTL on chromosomes 6 and 7 were simulated. All seven QTL were of equal effect, and together accounted for all of the genetic variation. The sign of the effect and the positions of the QTL are given in Table 2.1. The decision to simulate the same effect for all QTL was made for simplicity in computing specificity and sensitivity estimates. However, the relative effects of QTL on phenotypes were varied by varying heritability. Unless otherwise specified, the results and discussion presented are based on the seven-chromosome map.

Trait simulation

Simulation of trait values for the progeny in the final generation was according to the model $y_i = \sum_k^{N_Q} a_k q_{ik} + e_i$, where N_Q is the number of QTL, a_k is the additive effect of the kth QTL, q_{ik} = 1,–1, 0 for QTL genotypes *QQ*, *qq*, and *Qq*, respectively, and $e_i \sim N(0, \sigma^2)$. Since the variance explained by each QTL is a^2, the total variance explained by all QTL is $\sum a_k^2$, and the error variance for a given heritability h^2 was calculated as $\sigma^2 = \sum a_k^2[(1 - h^2)/h^2]$.

QTL analysis

For each parameter combination (here referred to as an *experiment*), 1000 populations of 300 lines were generated and subjected to QTL analysis. Analysis was done using regression-based composite interval mapping (Zeng, 1994) with cofactors selected by stepwise forward regression. No cofactors were used on the chromosome in which a QTL was evaluated. A likelihood ratio test statistic (LOD score) was calculated at 2-cM intervals on chromosomes. Separately for each experiment, one threshold value was calculated for declaring a QTL using permutation analysis (Churchill and Doerge, 1994), with 1000 permutations (one from each population) at $\alpha = 0.05$. A QTL peak was defined as a point whose LOD exceeded the threshold and the LODs at the adjacent points on either side.

Two types of intervals were evaluated for counting true positives (TPs) and false positives (FPs): a 20-cM constant interval and the *LOD-drop-2* interval. A *LOD-drop-2* (Lander and Botstein, 1989) QTL interval for a peak included all the points on either side of the QTL peak with LOD scores at most two units lower. Asymptotically a *LOD-drop-2* interval is equivalent to a confidence interval of 96.8 to 99.8% to contain a QTL (Mangin et al., 1994). A TP was recorded when the *LOD-drop-2* interval contained a true QTL and an FP when it did not. True QTL not detected were thus false negatives (FNs). Specificity was expressed as TP/(TP + FP) = TP/(number of QTL regions detected) and sensitivity as TP/(TP + FN) = TP/(number of simulated QTL). Deviation of a peak from a simulated QTL, d_i, was computed as absolute

difference between position of the *i*th QTL and the peak in the *LOD-drop-2* interval. The average deviation was $\bar{d} = \left(\sum_{i=1}^{\mathrm{TP}} d_i \right) \Big/ \mathrm{TP}$.

A genetic map generated from the SA-RIL population differs from the map on the basis of which recombination was simulated. To determine whether increased recombination influences the detection of QTL, we compared the results of QTL analyses based on updated map distances with results from the original map. These updated distances were computed from the recombination fractions observed at the final generation rather than those of the map used for generating the data. The marker order was the same for both maps. The updated QTL positions were used in computing specificity, sensitivity, and $\bar{d}$.

Calculation of recombination and heterozygosity index

A recombination is defined here as a change of parental allele between two adjacent marker loci. The recombination index for a RIL is the sum of all inferred recombinations along its genome. These were counted as shown in Table 2.2. To investigate the effect of different weighting schemes for selection for increased heterozygosity on recombination, we included in the recombination formula for heterozygotes a weight, *w*, taking values 1, 10, and 100. Note that with $w = 1$, the formula of Table 2.2 gives the expected number of recombinations occurring in a double-heterozygote interval during one meiosis. Increasing *w* should thus both slow the inbreeding process and lead to more opportunity for recombination to occur.

Calculation of standard errors of estimates

Standard errors of all parameter estimates except threshold (e.g., recombinations and specificity) were computed as the standard deviation of the estimate divided by the square root of the sample size (here 1000). For threshold, the standard error was computed using a jackknife approach as follows: let *T* denote the threshold computed from a set of 1000 ascendingly ordered LOD score statistics that were computed using permutation analysis (see description above). One thousand T_is ($i = \{1 \ldots 1000\}$) were then recorded, each from a set of 999 LOD score statistics formed by omission of the *i*th value. Since for $i = 1$ to 950, T_i was the 950th value and for $i = 951$ to 1000, T_i was the 949th value, this amounts to sampling only these two values at frequencies dictated by the desired threshold. For each T_i, p_i was calculated as $1000 * T - 999 * T_i$. The

standard error of the p_i values was computed as the standard deviation of the estimate divided by the square root of the sample size.

Results and Discussion

Effects of parameter changes

On recombination

Observed recombination increased with increased marker density for all selection strategies (Fig. 2.1a) due to the increased chance of detecting double-recombination events across a given interval with higher marker densities. A similar trend was also observed with increased family size (Fig. 2.1b). All selection strategies, with the exception of *D*, resulted in more recombination than *rrrrrrrr*, and completely homozygous lines with as many recombinations as in RILs could be generated in as few as two generations (*RD*). Recombination in selectively advanced RILs slightly increased when weight *w* was increased from 1 to 10, but decreased at higher levels (Fig. 2.1d). This level apparently strikes an optimum balance between maintaining double-heterozygote intervals to allow for continued recombination and allowing inbreeding to proceed.

The *SimpleSel* method achieved higher recombination count than *FamSel* and *FinSel* (Fig. 2.1e). However, our attempts to use *SimpleSel* resulted in low genotypic diversity (data not shown), with most of the selected progeny in the final generation descending from a few highly recombinant F_2 plants. *FamSel* achieves up to 25% more recombination than *FinSel* (Fig. 2.1e).

With decrease in map length, the recombination enrichment (RE) per map unit increases. This is because RE is inversely proportional to the square root of map length and directly proportional to selection intensity, as explained in Appendix A. Thus RE will be greatest in small genomes.

On amount of heterozygosity

Increasing marker density did not affect the heterozygosity index (Fig. 2.2a). Increasing the family size slightly decreased heterozygosity index (Fig. 2.1b), presumably by increasing the representation of doubly recombinant (*AAbb* and *aaBB*) homozygotes exposed to selection. The

reduction in number of heterozygotes by (approximately) half every generation was not affected as family size increased (Fig. 2.2c).

On detection threshold

As expected, threshold values decreased with marker density (Fig. 2.3a) because of fewer tests at lower marker density. With an increase in recombination we expected an increase in threshold because of increase in independence among tests, owing to decreased correlation among neighboring marker genotypes leading to a higher effective number of tests. This trend was not observed (Fig. 2.3b) for unknown reasons.

On confidence interval and average deviation of a peak from true QTL position

Confidence interval width, computed as *LOD-drop-2* width, decreased considerably with marker interval size (Fig. 2.3c). This decrease resulted in fewer TPs at 5 cM than at 10 cM (Fig. 2.3d), leading us to doubt the reliability of *LOD-drop-2* interval widths derived from CIM LOD curves. Since TP behaved as expected when a constant interval width of 20 cM was used. Specificity and sensitivity calculations were based on 20-cM rather than *LOD-drop-2* intervals.

The average deviation of peaks from the true QTL position decreased with increased recombination at higher marker density (Fig. 2.3e) and was also slightly less than the value without selection.

On specificity

In general, specificity rose with recombination. This result is expected, since as recombination increases, crossovers break linkage blocks, tending to isolate QTL. Capturing the resulting finer-scale QTL signals requires an average marker separation of 5 cM or less. Specificity increased with increase in family size (Fig. 2.4b). With increase in heritability, variation in the trait explained by the QTL increased and the LOD score at and around the QTL was high. The chance of detecting TPs was high and consequently specificity increased with increased heritability (Fig. 2.4c).

On sensitivity

SA-RILs offered sensitivity equal to or slightly lower than conventional RILs at higher marker density (here 5 cM). However, sensitivity decreased slightly with increasing recombination at lower marker density and low heritability (Fig. 2.5a, 2.5b), as Xu et al. (2005)

also observed. This is likely to be for the same reason—isolation of QTL to smaller regions—as given above for the decline in specificity, and suggests the same remedy: increasing marker density. With increased family size, sensitivity decreased due to increased recombination. With increase in heritability, trait variation explained by QTL increased, resulting in improved sensitivity (Fig. 2.5b).

Effect of map updating on specificity and sensitivity in SA-RILs

Specificity and sensitivity decreased when updated map distances were used for QTL mapping. In the absence of map updating, QTL genotype expectations were based on marker recombination fractions in the map from which populations were simulated, but in the updated map (with QTL positions established by linkage analysis of the final data including the known simulated QTL genotypes), on fractions observed in the selected sample. Since recombination increased up to 1.5 and 3 times in 10- and 1.5-Morgan maps, respectively, the distances between markers in the updated maps increased accordingly, resulting in specificity and sensitivity decreases as shown in Fig. 2.4d and 2.5d.

Other phenomena that might influence QTL mapping in SA-RILs

Three kinds of nonindependence of recombination might affect QTL mapping in SA-RILs. This nonindependence may hold even in the absence of crossover interference in a single meiosis. Pseudo-crossing-over is nonindependence of recombinants from interval to interval in multigeneration experimental populations such as RILs and AILs (Martin and Hospital, 2006). Pseudo-negative crossing-over, a special case of pseudo-crossing-over, is clustering of double crossovers due to recombination being limited to some genome segments remaining heterozygous over generations of inbreeding, while other segments are rapidly fixed. These phenomena might affect the accuracy of QTL genotype probability estimates, but both also occur in RILs. Pseudo-interference is nonindependence of recombination in samples subjected to selection (Xu et al., 2005). In these samples, variation in the distance between recombination events is lower than in random samples, such that selection mimics the effect of true crossover interference. Pseudo-interference was negligible when marker spacing was low ($\leq$5 cM) and map lengths were large ($\geq$5 M) (Xu et al., 2005). In our simulations the map length was 10.5 M.

The present study focused on the advantages of SA-RILs over RILs and did not examine genetic phenomena common to both.

Cost and benefits of SA-RILs

Savings in cost associated with additional generations

Our results show that with selective advance an equal or greater degree of recombination and of allele fixation can be produced in fewer generations than required for conventionally developed RILs. For example, strategy *RRD* results in homozygous lines carrying more recombination than RILs produced by eight generations of unselected selfing. The attractiveness of the SA-RIL approach will depend on the costs of genotyping and haploid-doubling operations and the urgency of population development.

Amount of genotyping

If all markers are genotyped, the number of genotyping reactions run per marker for any strategy is the product of the family size p and the number of R, D, and H characters in the strategy string. If, in contrast, it is possible and economical to limit genotyping in any line to the markers that were heterozygous in its parent line, the number of reactions per marker is less than $2p$ that for conventional RILs; that is, $p\{1 + 1/2 + 1/4 + \ldots\}$, because of the decline of heterozygosity by half at every generation. About 10% higher marker density will be needed to exploit the higher resolution afforded by the design, a requirement not unique to SA-RILs but applying as well to AILs, association-mapping panels, and any other recombination-enriching design.

Possible effects of violation of random-segregation assumption

Selective-phenotyping approaches lead, by definition, to nonrandom genotypic distributions, violating the assumption of random segregation required for estimation of genotype frequencies at QTL test positions. But systematic distortion in the population is unlikely, owing to the independent descent of the lines from different F_2 progenitors. Another violation of random segregation might arise from genetic control of recombination, in which case certain alleles affecting recombination might be preferentially selected. This case will resemble segregation distortion that can arise from any kind of natural selection during line development.

Unless QTL for the trait under study lie near such loci, or in regions affected by nonuniform recombination enhancement by such loci, we suggest that this effect too is negligible, although our simulations did not model it. Other selective-phenotyping approaches (e.g., the *maxRec* and bin-length minimizing methods described above) would also suffer from these problems. High marker density and single- or multiple-marker rather than interval analyses—an approach akin to association mapping—may render them negligible.

Possible extensions to the SA-RIL method

The objective function used to increase recombination in our simulations was similar to that in the *maxRec* method, but within a family. The large gap between the curves of *FamSel* and *SimpleSel* methods shown in Fig. 2.1e indicates the scope for further recombination enrichment in selected samples. Selection methods (described in the Introduction) that use other objective functions might also increase the effectiveness of a population for QTL mapping.

Although recombination accumulating in AILs is not limited by inbreeding as in RILs, the selective-advance approach could also be applied during generation of AILs to increase recombination further. The focus would then shift to selecting pairs of lines for intermating that would lead to maximizing recombination in the final generation.

Conclusions

I propose a method of improving mapping resolution in selfing species through recombination enrichment during development of RILs. The method is based on genotyping at each generation and advancing lines selected for high recombination. Recombination may be enriched by two to three times in 10.5- and 1.5-Morgan maps, respectively. Compared to methods that require maintaining a large population for several generations and selecting on the finished population, the method proposed here achieves up to 25% more recombination. Strategies such as *RD* and *RRD* yield in two and three generations populations with the same amount of recombination as conventional RILs. Specificity of SA-RILs is 5 to 20% more than conventional RILs. Sensitivity is as good as or slightly lower than conventional RILs when the marker density is high (5 cM; for a 10.5 Morgan map). Quantitative trait locus mapping can be done with any desired QTL software, while recombinational selection of lines to be advanced should be manageable with simple scripts. In studies that require expensive phenotyping (e.g.,

expression QTL studies) but enjoy low genotyping costs, SA-RILs may be a cost-effective resource for QTL mapping in selfing species.

Figures and Tables

Figure 2.1 Average number of recombinations in SA-RILs

The population size is 300. Strategies are sorted by recombination at 10 cM. Recombinations for (a) several marker spacings and selection strategies, family size of four; (b) several family sizes and selection strategies, marker spacing of 10 cM. Figures 1c and 1d show recombinations of selectively advanced RILs (SA-RILs) across generations with strategy *RRRRRRRR*. Markers are separated by 10 cM and the length of the map is 10.5 M. Figure 1c shows recombinations for several family sizes and Fig. 1d for several weights *w*; family size of 4. Figure 1e shows recombination count for different selection methods for maps of length 10.5 morgans. *SimpleSel* selects the best subset (without family constraints); *FamSel* selects the best progeny from each family. *FinSel* selects the progeny in the final generation, from a large population maintained for several generations. *NoSel* denotes no selection.

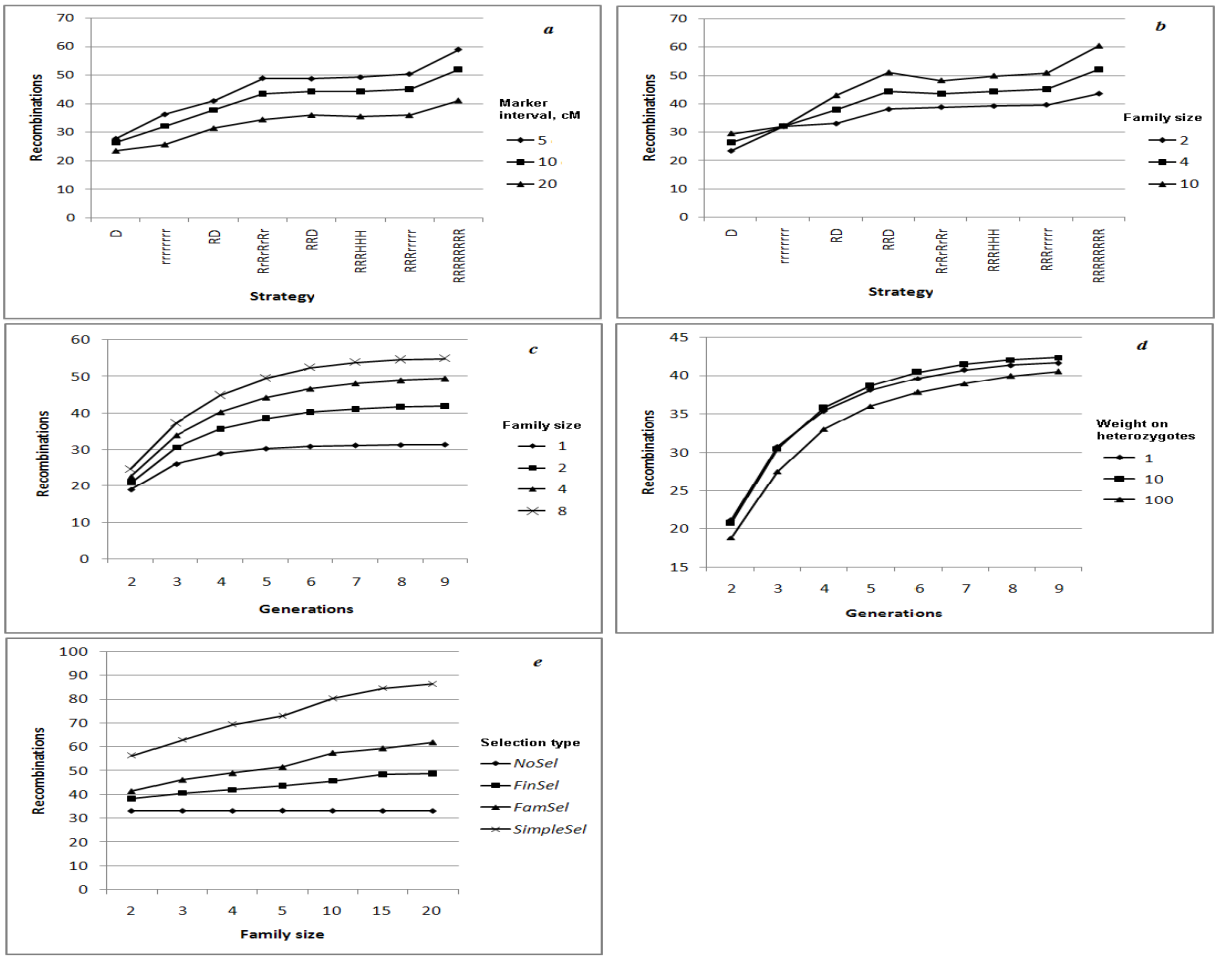


Figure 2.2 Heterozygosity in SA-RILs

The population size is 300. Heterozygosity index is shown for (a) several marker spacings and selection strategies with family size of four; (b) several family sizes and selection strategies with marker spacing of 10 cM. Number of heterozygous loci in selectively advanced RILs with strategy *RRRRRRRR* across generations with marker spacing of 10 cM and map length 10.5 M for (c) several family sizes and (d) several weights *w* with family size of four. Standard error bars are drawn in the upward direction.

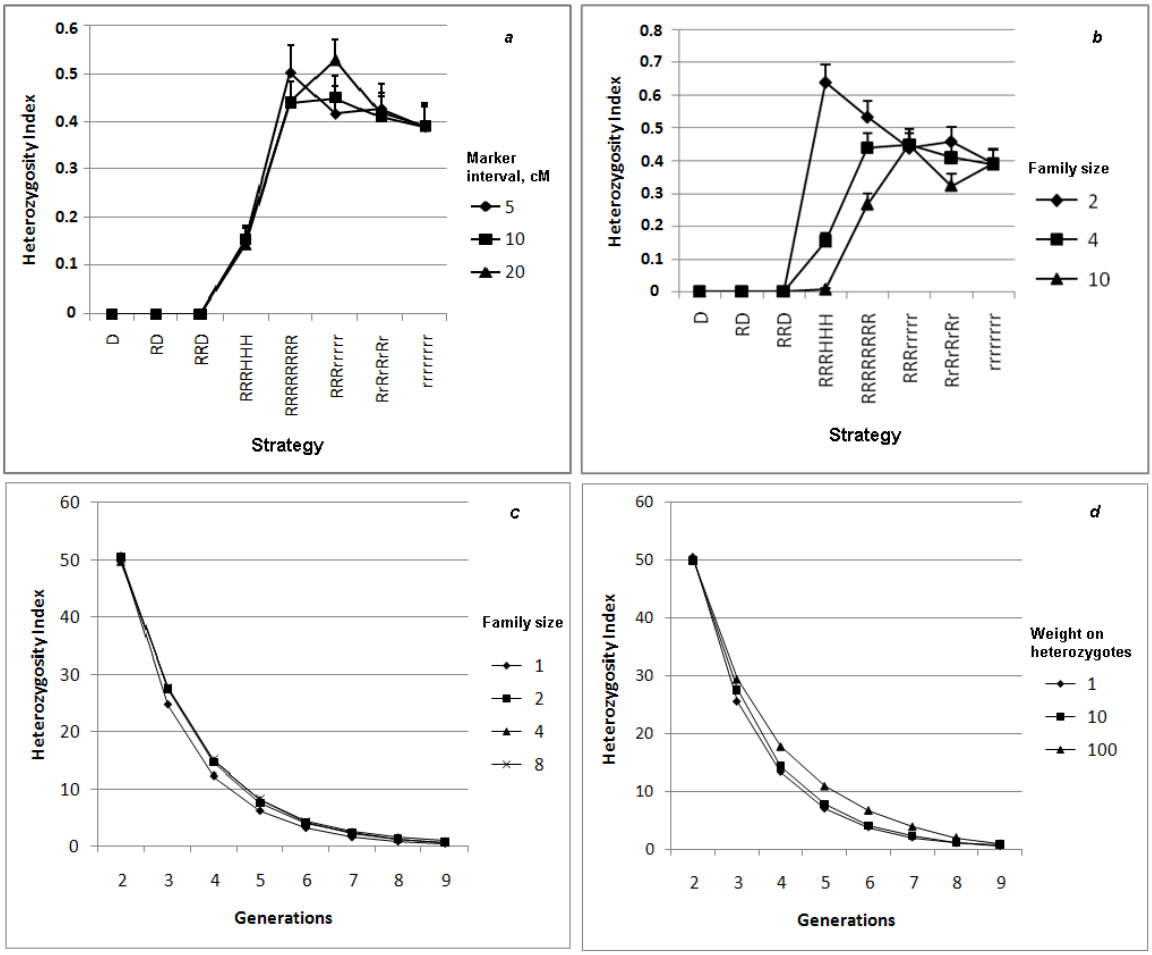


Figure 2.3 Acceptance threshold, LOD-drop-two interval widths, TP and average deviation of peaks from simulated QTL for several selection strategies in SA-RILs

Strategies are sorted by recombination at 10 cM. The threshold for (a) several marker densities and (b) several family sizes. The standard error bar is shown in the upward direction. Panels (c) and (d) show 95% CI interval widths and true positives detected in 95% CI widths computed from LOD-drop-2 intervals at several marker densities. (e) Average deviation of peaks from simulated quantitative trait loci (QTL) at several marker densities.

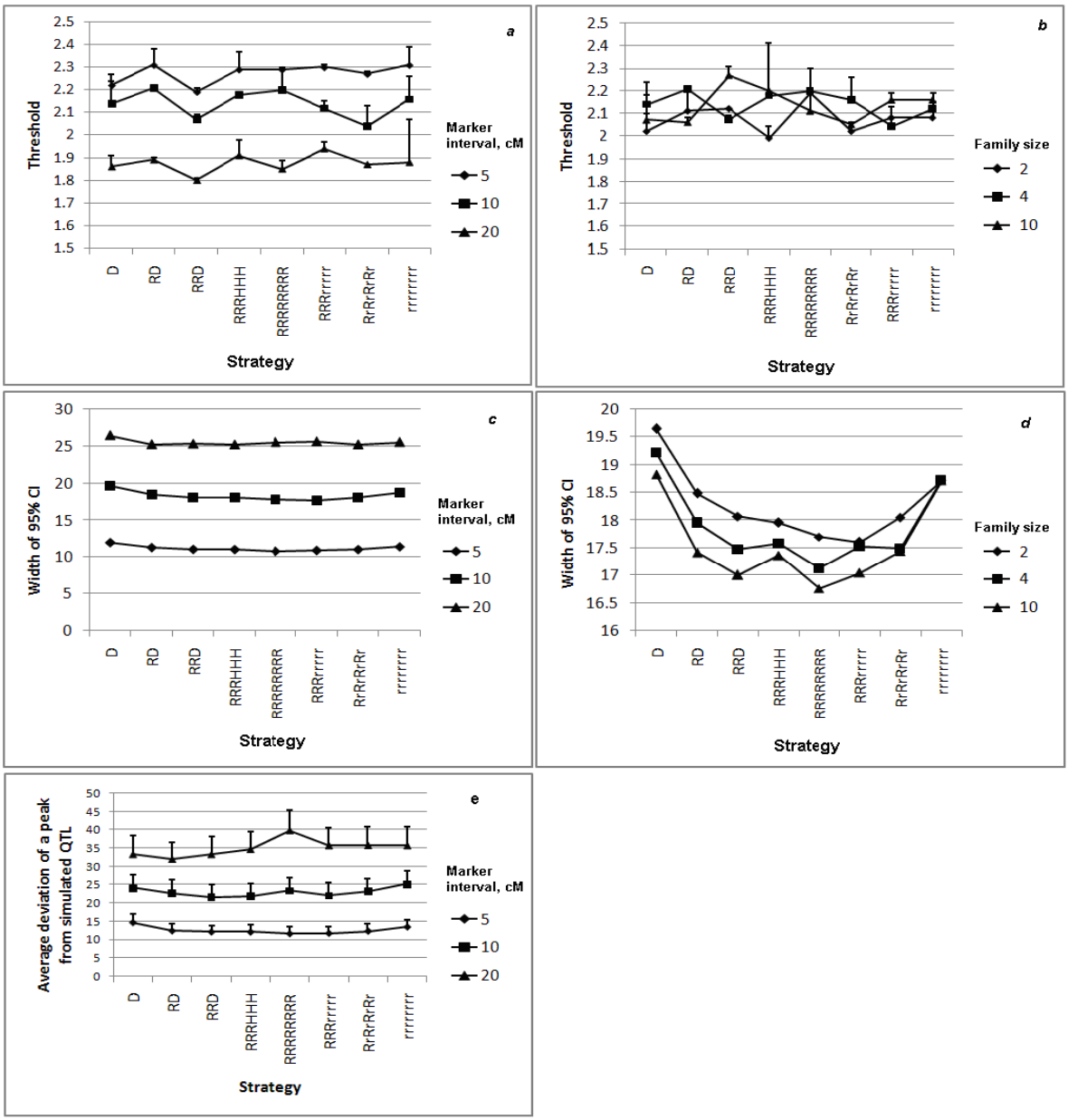


Figure 2.4 QTL detection specificity for several marker spacings and selection strategies used in SA-RILs

Population size is 300 and map length is 10.5 M. Specificity is computed using a constant interval width of 20 cM. Specificity is shown for (a) several marker spacings and selection strategies with family size 4 and heritability 0.4; (b) several family sizes and selection strategies with marker spacing 10 cM and heritability 0.4; (c) several heritabilities and selection strategies with marker spacing 10 cM and family size 4. (d) Specificity computed using a map computed from the recombination fractions observed in the final SA-RIL population.

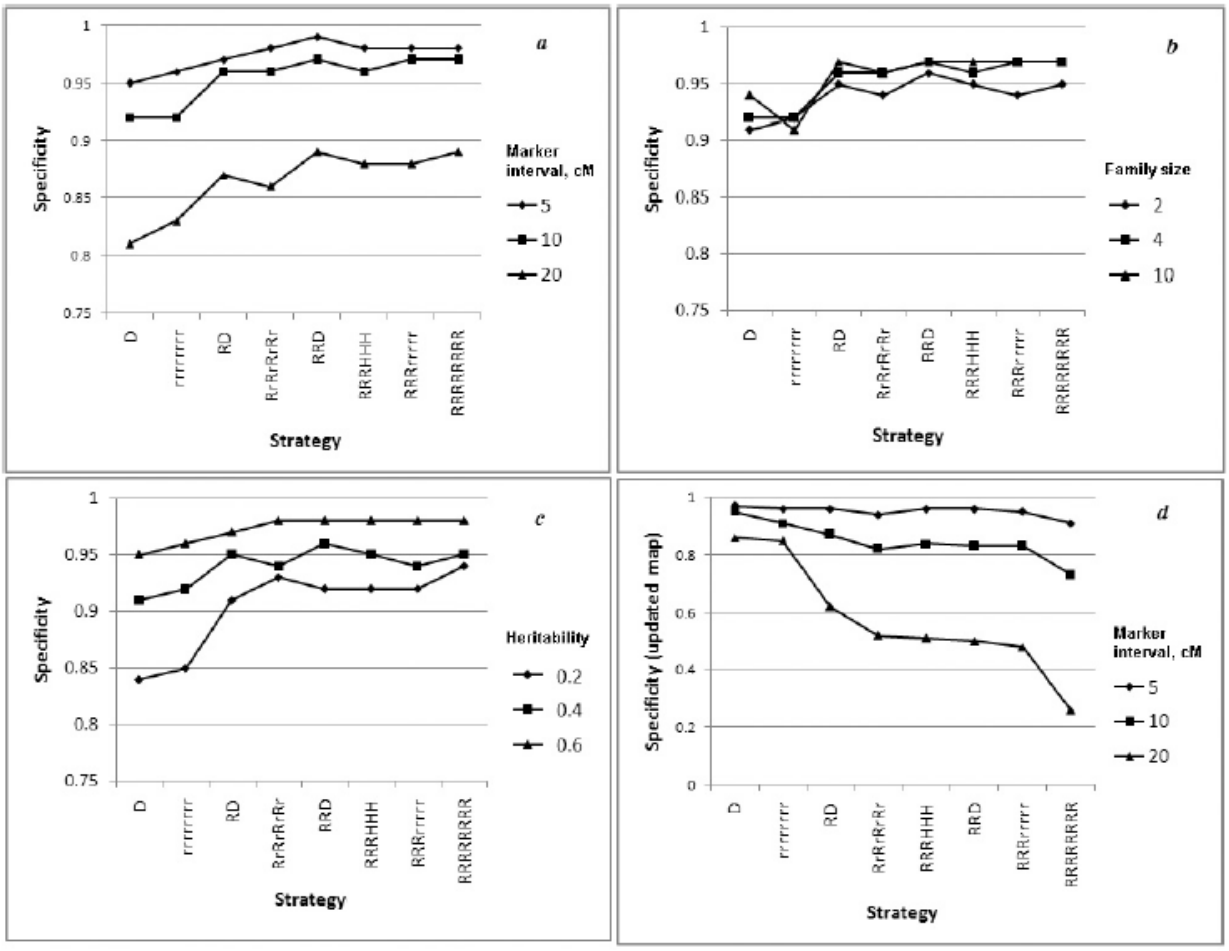


Figure 2.5 QTL detection sensitivity for several marker spacings and selection strategies in SA-RILs

Population size is 300 and map length is 10.5 M. Panels show sensitivity for (a) several family sizes and selection strategies with marker spacing of 10 cM and heritability 0.5; (b) several marker spacings and selection strategies at a heritability of 0.5 and family size 4; (c) several heritabilities and selection strategies at marker spacing of 10 cM and family size 4; (d) sensitivity computed using a map calculated from the observed recombination frequencies in the final generation of the SA-RIL population.

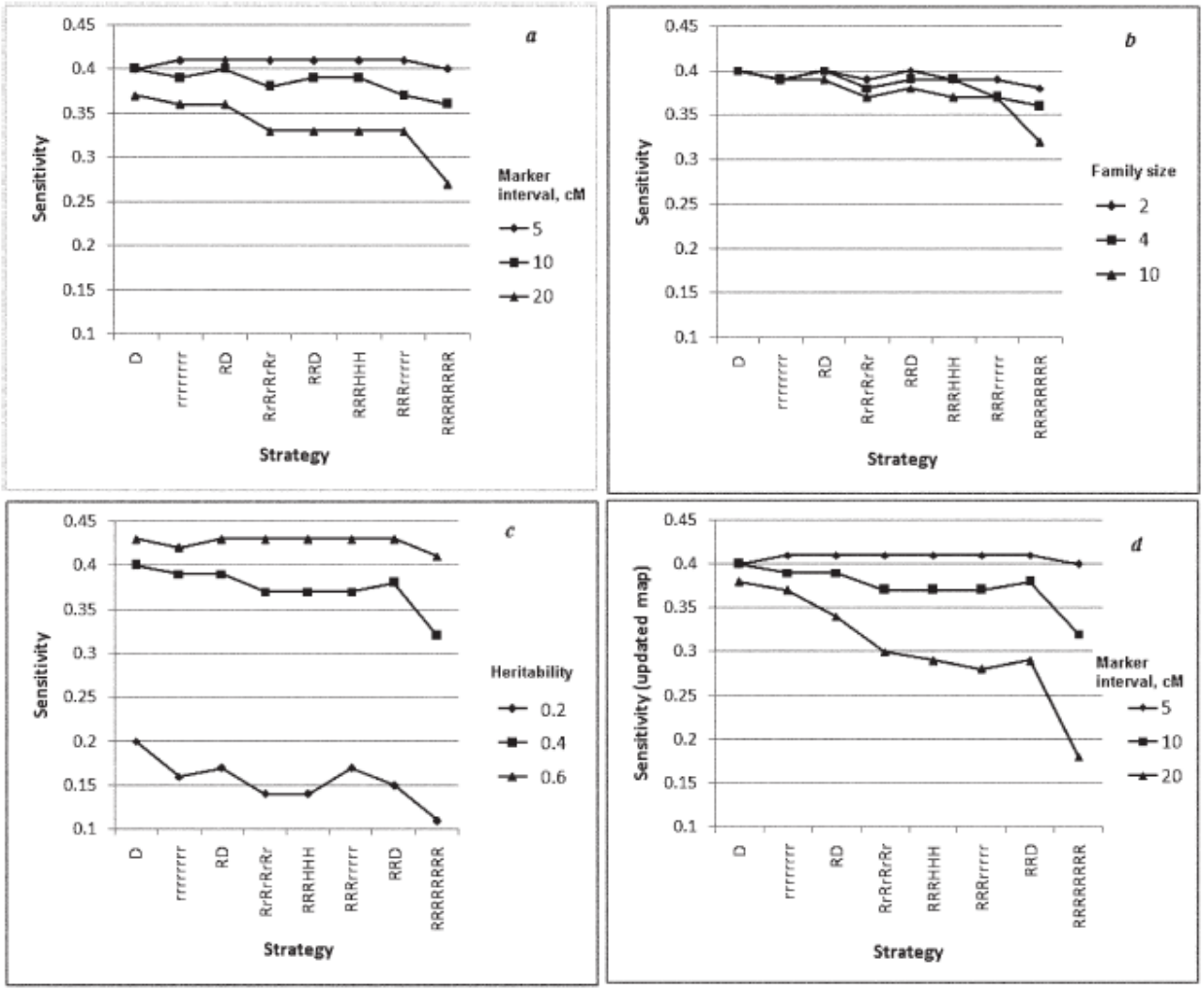


Figure 2.6 The distribution of recombination numbers for several map lengths

The expected number of recombinations follows a Poisson distribution with mean $E(R) = nc$ and standard deviation $\sqrt{nc}$, where n and c represent number of map intervals and probability of recombination in an interval. X axis shows the length of the genetic map expressed in recombinations.

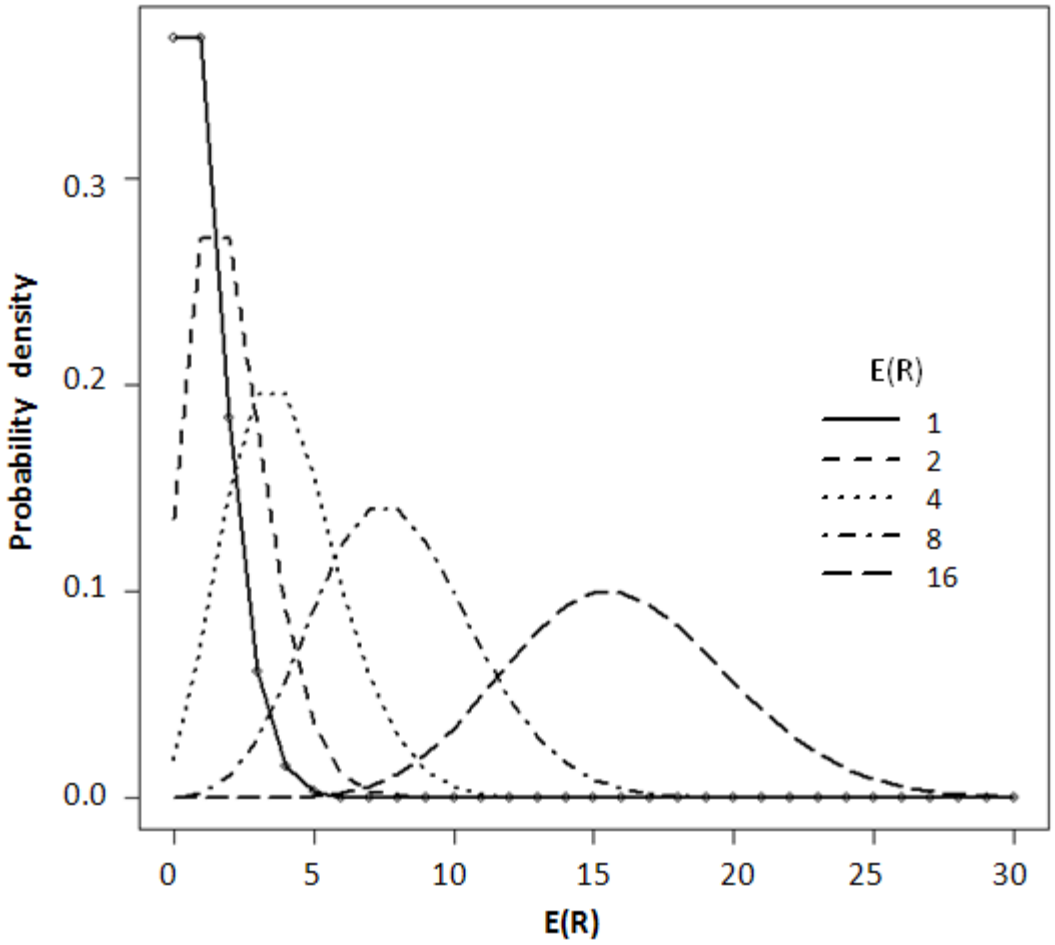


Table 2.1 Quantitative trait locus (QTL) positions and effect signs used in all simulations

Chromosome	QTL positions	Effect signs
	cM	
1	23.4, 113.5	+, +
2	23.4, 113.5	+,−
3, 4	63.5	+
5	63.5	−
6, 7	no QTL	

Table 2.2 Calculation of recombination between adjacent markers A_1 and B_1 with alleles {*A*, *a*} and {*B*, *b*}, for different two-locus genotypes

The symbol c represents the recombination fraction between markers A_1 and B_1. Weight w is used to intensify selection on heterozygosity.

Two-locus genotype	No. of recombinations
AABB or *aabb*	0
aaBB or *AAbb*	2
AABb or *aaBb*	1
AaBB or *Aabb*	1
AaBb	$2wc^2/[c^2 + (1 - c)^2]$

CHAPTER 3 - QTL-Focused Selectively Advanced Recombinant Inbred Lines

Although selectively advanced recombinant inbred lines (SA-RILs) have accumulated more recombination than conventional RILs, their effectiveness for recombination enrichment decreases with genome size. The SA-RIL approach is most valuable when applied either to small genomes or to defined regions of large genomes. Here I propose development of QSA-RILs, which are SA-RILs enriched for recombination in regions of a large genome selected for evidence for the presence of a QTL. This evidence can be derived from QTL analysis of a subset of the population at the F_2 generation and/or from previous studies. In simulations QSA-RILs afford up to threefold increase in recombination and twofold increase in accuracy of QTL position estimate in comparison with RILs. The regional-selection method also shows potential for resolving QTL linked in repulsion.

Introduction

Fine mapping, defined as a process of increasing the precision of QTL location estimates, is desirable for identifying genes underlying an observed phenotype. The two main approaches used for fine mapping are linkage disequilibrium analysis (LDA) and linkage analysis (LA). LDA can resolve a QTL position to less than 1 cM by exploiting historical recombination in a set of lines not derived from a single biparental cross (Nadeau and Frankel 2000). LA exploits recombination accumulated since the initial parental cross in experimental populations such as recombinant inbred lines (RILs) and advanced inbred lines (AILs; Darvasi and Soller, 1995). In maize AILs that had undergone four intermating generations, LA resolved a QTL position to less than 3 cM (Balint-Kurti et al. 2007). However, in selfing species where intermating is tedious, RILs are the only advanced-generation experimental populations conventionally used for fine mapping by LA.

QTL mapping resolution is limited in RILs by the number of recombinations achievable in the generations between the initial parental cross and genotype fixation. Recombination

enrichment allows the introduction of additional markers in physical proximity to a QTL, shrinking the physical distance relative to the genetic distance.

Several selective phenotyping methods have been proposed with the aim of developing recombination-enriched (RE) QTL-mapping populations. In an interval method proposed by Darvasi (1998), phenotypes were collected only on individuals showing recombination in the QTL-containing interval. Whole-genome methods proposed by Xu et al. (2005), Jannink (2005) and Boddhireddy et al. (2009) increased recombination throughout the genome. Jannink (2005) proposed two selection strategies: both select lines with high recombination across the genome, but one of the strategies enforces greater genome-wide uniformity of the distribution of recombination. Xu et al. (2005) proposed selection of RE individuals based on an objective function that minimizes the sum of squared bin lengths. Boddhireddy et al. (2009) proposed selective advance for accelerated development of recombinant inbred QTL mapping populations (SA-RILs), which are RILs that have been selected for recombination during generation advance of RIL development. The last method improved the precision of QTL location estimates by 5-20% in SA-RILs in comparison to conventional RILs.

Whole-genome recombination-enrichment approaches are valuable only for species of small genome size, because recombination enrichment is inversely proportional to the square root of map length (Xu et al., 2005; Jannink, 2005; Boddhireddy et al., 2009). To obtain RE RILs in species with large genomes, we may restrict enrichment to regions believed to contain QTL. This evidence may be obtained from QTL analysis of previous studies, of a subset of the population, and/or of earlier RIL generations.

Here I propose a RIL population-development method for simultaneously enriching recombination across several selected regions of the genome. Enrichment is achieved during RIL development via application of any of several schemes for selecting lines to advance to the next generation. I refer to these RILs as QSA-RILs (QTL-focused SA-RILs). Using simulations, I compute comparison metrics for QSA-RIL populations developed with varying selection schemes, in order to assess the merits of these populations for recombination enrichment and QTL mapping.

Methods

Population simulation

Let N be the number of RILs to be generated and p the number of progeny genotyped for each line at each generation. RILs were simulated as follows: F_1 plants were generated from the cross of two inbred parents and F_2 progeny were then generated by selfing. N x p F_2 plants were then genotyped and from each family the plant showing maximum recombination index was advanced. Recombination index was calculated as a sum of weighted recombination values across all intervals, where the weight applied to each interval was varied according to the population type simulated (see below). From each of the N selected F_2 parents, p F_3 progeny were generated and genotyped, and the process of selection and advance was repeated until the desired number of generations. Numbers of progeny, or family sizes, simulated were 1, 2, 4 and 10. For each experiment, consisting of a parameter combination of family size, population type and distance of a QTL from the nearest marker, 500 populations of 300 lines were simulated.

QTL analysis

Simple interval mapping (SIM; Haley and Knott, 1992) was used for QTL analysis of F_2 and QSA-RIL populations. A likelihood ratio test (LRT) statistic (LOD score) was calculated at every 1-cM across the genetic map. A relaxed threshold of LOD = 2.0 for F_2 populations and a threshold of LOD = 3.0 for QSA-RILs was used to declare presence of QTL peaks. A peak was defined as a point on the LOD-score profile across the chromosome whose LOD exceeded the threshold and the LODs at the adjacent test positions on either side. Each LOD score was standardized by division with the maximum LOD score on the same chromosome. All QTL analyses were conducted with QGene (Joehanes and Nelson, 2008).

Population type

Six population types were simulated. Each type, P_t, was defined by a weight, w_{ti}, applied to the recombination count, r_i, in marker interval i, where r_i was computed as shown in Table 1. The recombination index of an individual in population type t was $\sum r_i w_{ti}$.

P_1: $w_{1i} = 0$ for all i. This population type represented conventional RILs. The recombination index of all individuals in this population was equal to zero and selection was thus random.

P_2: $w_{2i} = 1$ for all i. This population type represented SA-RILs in which the recombination index of an individual was simply the sum of recombination values across all the intervals of the genome.

P_3: $w_{3i} = 1$ for all intervals overlapping selected regions, each of length L and containing a simulated QTL in its center. Thus for this population type, recombination was evaluated only in selected map intervals.

P_4: $w_{4i} = 1$ for all intervals overlapping a region defined by a LOD score peak computed from QTL analysis of an F_2 population rather than by a simulated QTL. The number of selected regions equaled the number of QTL peaks detected and might or might not be the same as those selected in P_3.

P_5: w_{5i} = mean of standardized LOD scores of all candidate positions in interval i. For this population type recombination across all intervals was weighted according to the standardized LOD score. Standardization was done so that the largest peak on one chromosome did not influence the weights around the largest peaks on other chromosomes.

P_6: $w_{6i} = w_{5i}$ for all intervals overlapping any region defined by a QTL peak as in P_4. For this population type recombination across selected intervals was subjected to standardized-LOD-score weighting.

Map simulation

Fifteen 100-cM chromosomes were simulated. A marker was simulated at every 5 cM. Five QTL were simulated. One QTL per chromosome on the first five chromosomes at d cM from the positions shown in Table 3.1, where d was 0, 1.0, or 2.5 cM.

Trait simulation

Progeny trait values were simulated according to the model $y_i = \sum_{k}^{N_Q} a_k q_{ik} + e_i$

where N_Q is the number of QTL, a_k is the additive effect of the k^{th} QTL, $q_{ik} = 1, -1, 0$ for QTL genotypes QQ and qq and Qq respectively, and $e_i \sim N(0, \sigma^2)$. Since the variance explained by each QTL is a^2, the total variance explained by all QTL is $\sum a_k^2$, and the error variance for a given heritability h^2 was calculated as $\sigma^2 = \sum a_k^2((1 - h^2) / h^2)$. The variation explained by each QTL in RILs (or SA-RILs) was 20%, 16%, 12%, 8%, or 4% and the heritability of the trait was 0.60. Heritability used to simulate traits of an F_2 population was computed as

$$h^2_{f_2} = \frac{h^2_r}{2 - h^2_r},$$

where h^2_r is the heritability in the RILs and $h^2_{f_2}$ that in the F_2 generation.

Comparison metrics

Three measures, computed for each simulated QTL, were used for comparing the six population types: true positives detected, proximity of a peak to the QTL, and recombination within *L* cM of a QTL-containing interval. A true positive (TP) was recorded when there was a peak within the QTL-containing interval; that is, within *L/2* cM of a simulated QTL. Proximity was calculated as the absolute distance between the position of the simulated QTL and the TP QTL peak.

Results

Recombination

As expected, with increase in family size, estimated recombination increased (Fig. 3.1). Also as expected, recombination around QTL was higher for population types that were selected for recombination in targeted regions of the genetic map than for those selected for recombination across the entire map. In comparison to recombination accumulating in conventional RILs within *L*/2 cM of a simulated QTL (referred to as a QTL region), recombination was twofold higher in populations selected for recombination across the entire map (P_2, P_5) and threefold higher in populations selected for recombination only in QTL regions (P_3) (Fig. 3.1). The recombination enrichment ranged from twofold to threefold in QTL regions (depending on the amount of variation explained by the QTL) in populations selected for recombination around LOD-score peaks (P_4, P_6).

Proximity estimate

As recombination increased around the QTL, the precision of the position estimate, calculated from the deviation of a QTL peak from a simulated QTL, increased. The increase was largest for QTL of small effect (Fig. 3.2). The precision obtained in QSA-RILs with population type selected for recombination around LOD score peaks with a family size of ten was equal to the precision obtained using a conventional RIL population of twice the size. With increasing

distance of the simulated QTL from the flanking marker, the accuracy of the position estimate decreased (Figs. 3.2 and 3.3).

True positives

When a simulated QTL was close to a marker, true positives increased as recombination increased. The number of true positives detected was low for QTL of small effect (Fig. 3.4).

Effect of F_2 QTL mapping results on recombination

Using a low threshold for declaring a QTL in analysis of the F_2 mapping population increased the length of the region involved in recombination selection in population types P_4, P_5, and P_6 where selection was based on the F_2 LOD score profile. The length of the region involved in selection is the QTL region length, L, times the positives discovered in the F_2 mapping population. Using a low threshold increased true positives along with false positives, increasing the length of the region involved in selection. The higher the length of the selected region, the less was the accumulated recombination. This is because recombination accumulation is inversely proportional to the square root of the length of the region involved in selection.

Discussion

In simulations, QSA-RILs achieve three times more recombination around candidate QTL than RILs. QSA-RILs combine useful features of both AILs (in producing maximum recombination) and RILs (in providing, via replication, reduced environmental variance in comparison to AILs). The latter feature is useful in fine mapping, when the trait of interest is influenced by a small number of QTL with low heritability (Darvasi and Soller, 1995).

Comparison of recombination in QSA-RILs and AILs

QSA-RILs are comparable to AILs with six or more intercrossings, with respect to recombination in QTL regions of interest. The comparison uses the equation $x = 2r_x/r$, (Darvasi and Soller, 1995), to compute x, the number of generations of intercrossing in AILs required to produce the same recombination as in RILs, QSA-RIL_2 and QSA-RIL_3. Here QSA-RIL_k with k = 2 and 3 denote QSA-RILs with two and three times the recombination of RILs and $r_{x,}$and r denote recombination between two neighboring loci in AIL_x and in the F_2. RILs, with twice the recombination of an F_2, are approximately equivalent to AIL_4. Similarly QSA-RIL_2 and QSA-

RIL_3, with four and six times more recombination than an F_2, are approximately equivalent to AIL_8 and AIL_{12}.

For separating closely linked QTL, QSA-RILs are better suited than F_2 and RILs, because the separation between linked QTL increases in proportion to the recombination enrichment between the loci. For example two QTL separated by 5 cM in F_2 would be separated by 10 cM in RILs (equivalent to AIL_4) and up to 26 cM in QSA-RILs (equivalent to AIL_{10}).

Violation of random-segregation assumption

Because of selection, QSA-RILs, like SA-RILs, violate the assumption of random segregation required for estimation of genotype probabilities from flanking markers at a QTL test position. In this study, SIM was employed as the computationally least demanding way to compute proximity estimates, the metric we chose for comparing population types. However, in practice, single- or multiple-marker rather than interval analysis should be used for QTL mapping in QSA-RILs.

Cost effectiveness of QSA-RIL development

The necessity to genotype several plants in each generation makes QSA-RIL more expensive than conventional RIL development. In order to evaluate the cost effectiveness of QSA-RILs development, I derive two equations: one describing genotyping costs, and the other describing overall cost benefit considering phenotyping costs and improvement in precision of QTL location estimate.

Cost of genotyping

Let p be the number of progeny used for selection; n, the number of lines genotyped; x, the cost of genotyping per marker; m, the number of markers; k, the proportion of total markers lying in the selected region. The cost of genotyping for a population type not selected for recombination is xmn. For other population types it is $xmnp\ (1 + k(½ + ¼ + ...)) \approx xmnp(1 + k)$, where k is 1 for population types P_2 and P_5 and is often a small fraction for population types P_3, P_4 and P_6.

Cost effectiveness

Let y be the cost of phenotyping each line; t the ratio of the cost of phenotyping to genotyping m markers, where $t = y/xm$; and s the proportion of excess RILs (in comparison to QSA-RILs)

required to be phenotyped and genotyped to achieve the same precision of QTL location estimate as in QSA-RILs.

The total cost of genotyping and phenotyping in RILs is the sum of phenotyping and genotyping costs of n and ns RILs, i.e., $(xmn + ny) + (xmns + nsy)$. Expressing the phenotyping costs in terms of genotyping costs, the total cost is $xmn(1 + t)(1 + s)$.

The total cost of genotyping and phenotyping in QSA-RILs is the sum of genotyping costs of all markers for all lines in the first generation, genotyping costs of proportion of markers of all lines in successive generations, and phenotyping cost of n lines, *i.e.*, $xmnp(1 + k) + ny$. Total cost in terms of genotyping costs is $xmn(p(1 + k) + t)$.

Although QSA-RILs are expensive to generate, their development can be cost-effective, depending on the ratio of the cost of phenotyping to genotyping (see Appendix B, Fig. 3.5) when the purpose of population development is to increase precision of QTL location estimates. For instance, 300 QSA-RILs would afford the same precision as 500 RILs (Fig. 3.2). In this scenario QSA-RIL development would be cost-effective when the ratio of cost of phenotyping to genotyping is at least five. The cost of genotyping can also be reduced by means of a low-marker-density panel (of, say, 15 or 20 cM) during QSA-RIL development. If ten regions are selected for recombination enrichment, three markers per region, making a total of thirty genotypes per plant, would be sufficient for QSA-RIL development.

Figures and Tables

Figure 3.1 Average recombination in a 30-cM interval around a simulated QTL

The population size is 300 and marker density is 5 cM. QTLs 1 to 5 explain 20, 16, 12, 8, and 4% of phenotypic variance. The number of populations simulated is 500. The distance between a simulated QTL and the nearest marker, *d*, is cM; family size is 10.

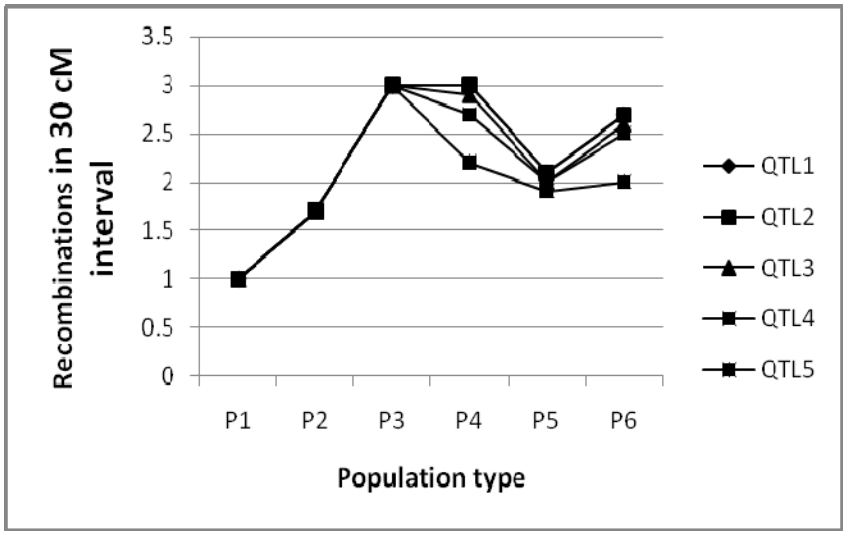


Figure 3.2 Accuracy of position estimates of a simulated QTL for RILs of several sizes and QSA-RILs of several family sizes

RIL_*p*_*n* denotes RILs with family size *p* and population size *n*. The number of simulations is 500 and the marker density is 5 cM. When $p > 1$, QSA-RILs are selected for recombination around the QTL peaks discovered in an F_2 population. The distance of the simulated QTL from a marker is shown in panels a) with $d = 0$, b) $d = 10$, c) $d = 25$ mM. The position estimate is presented in milliMorgans.

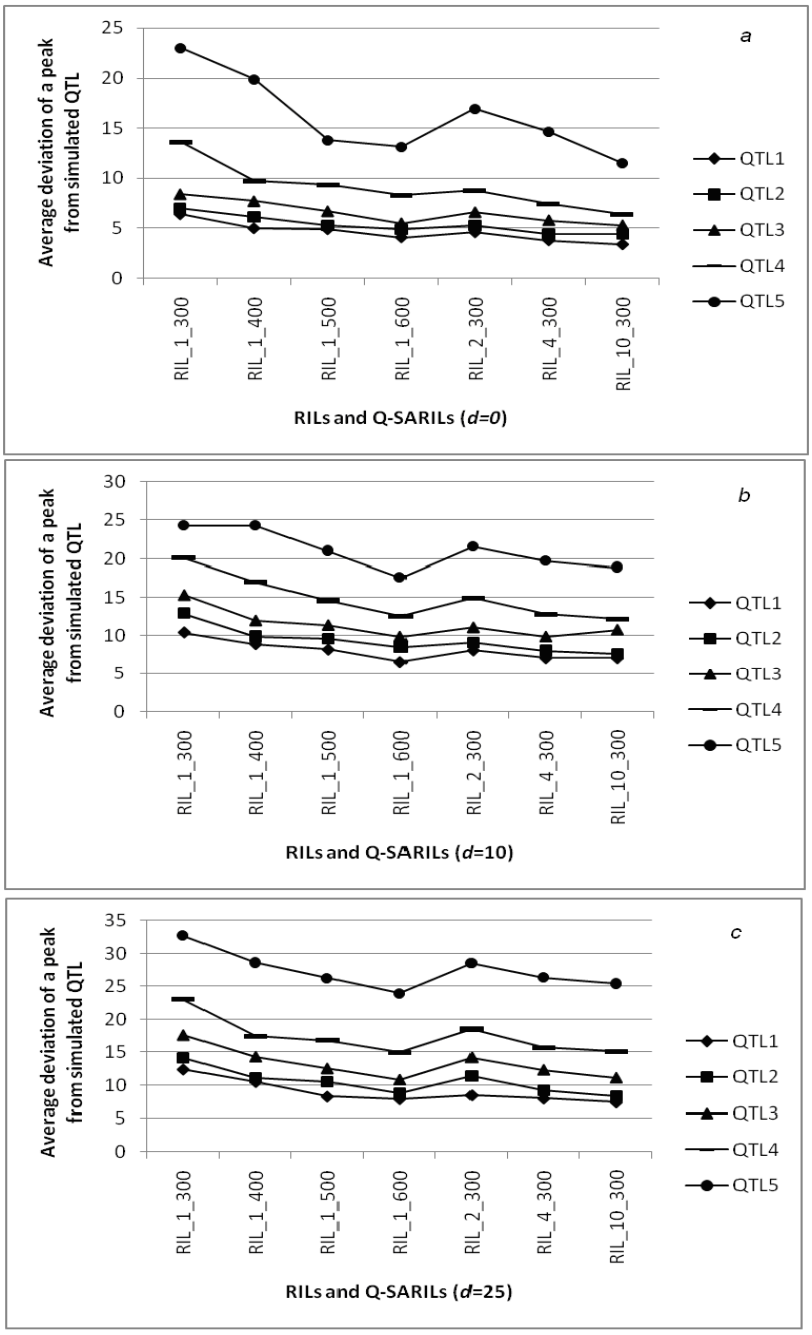


Figure 3.3 Accuracy of position estimates of simulated QTL for QSA-RILs developed with six different selection criteria

The population size is 300 and the marker density is 5 cM. The number of simulations is 500. The position estimate is presented in milliMorgans**.** The distance between the nearest marker and the simulated QTL is d = 0, 10, and 25 mM as shown in panels *a*, *b* and *c*. Selection criteria P1–6 are described in the text.

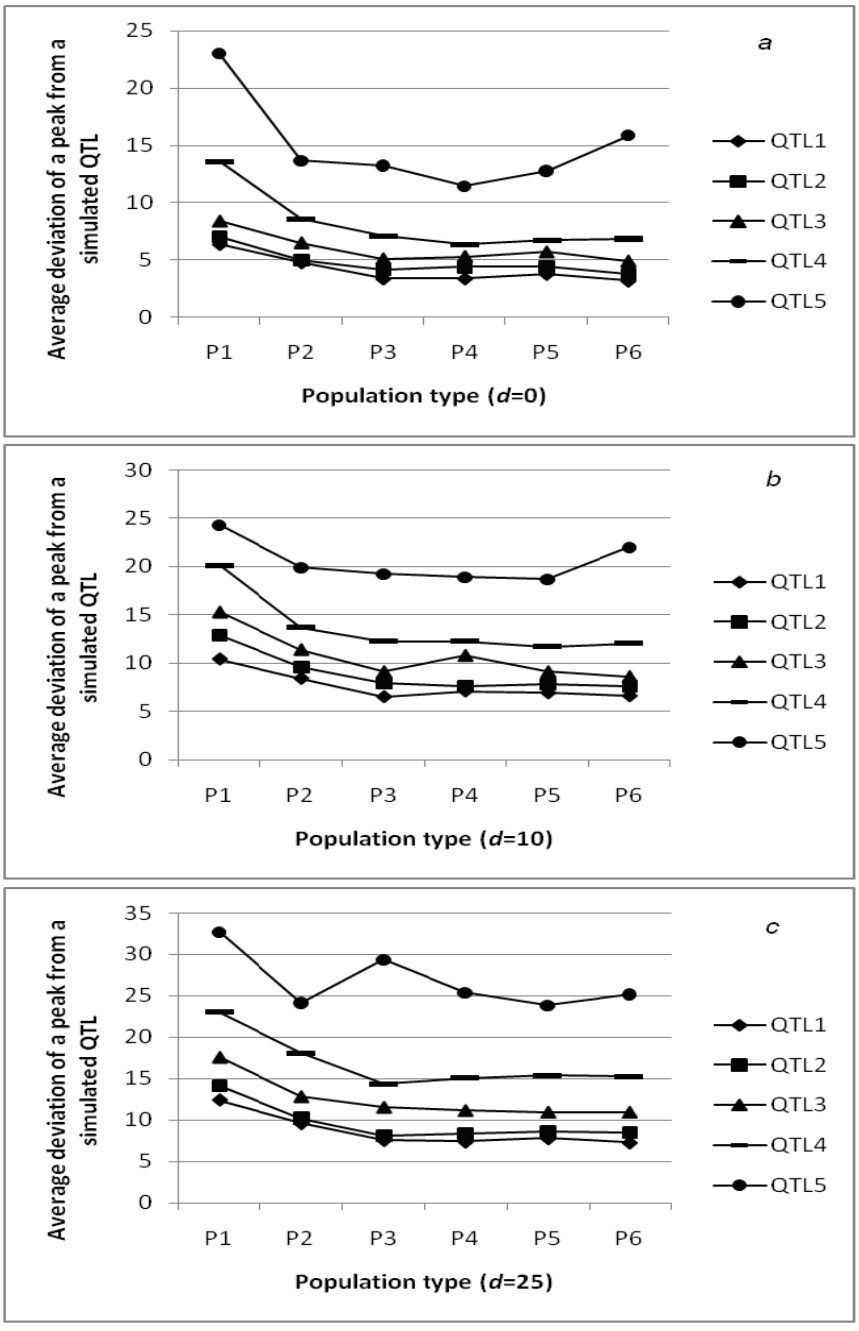


Figure 3.4 Number of simulated QTLs detected (true positives) for RILs and QSA-RILs

RIL_*p*_*n* denotes RILs with a family size *p* and population size *n*. The number of simulations is 500. The marker density is 5 cM.. When $p > 1$, QSA-RILs are selected for recombination around the QTL peaks discovered in an F_2 population. The distance of the simulated QTL from the nearest marker is $d = 0$, 10, and 25 mM in panels *a*, *b* and *c*.

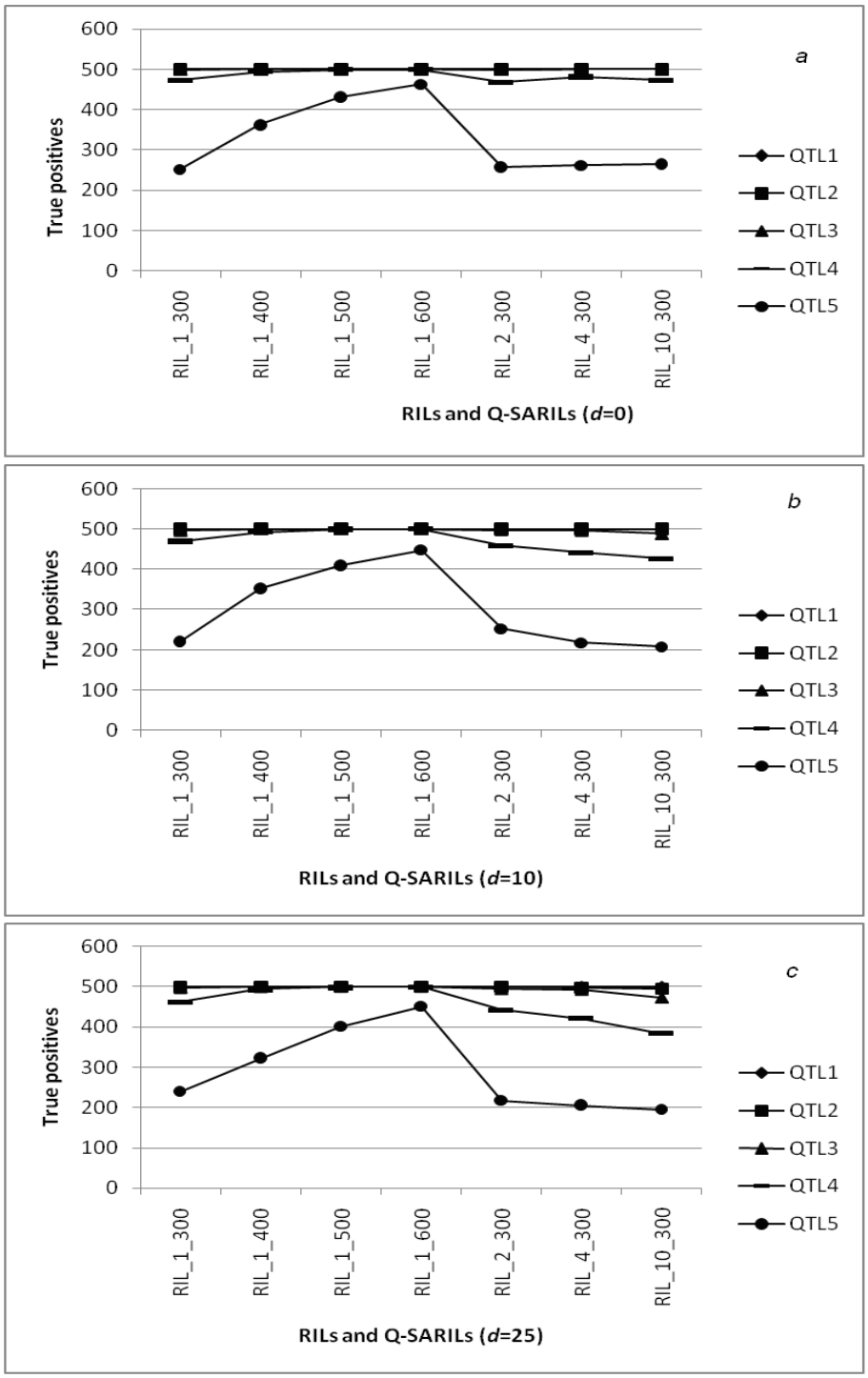
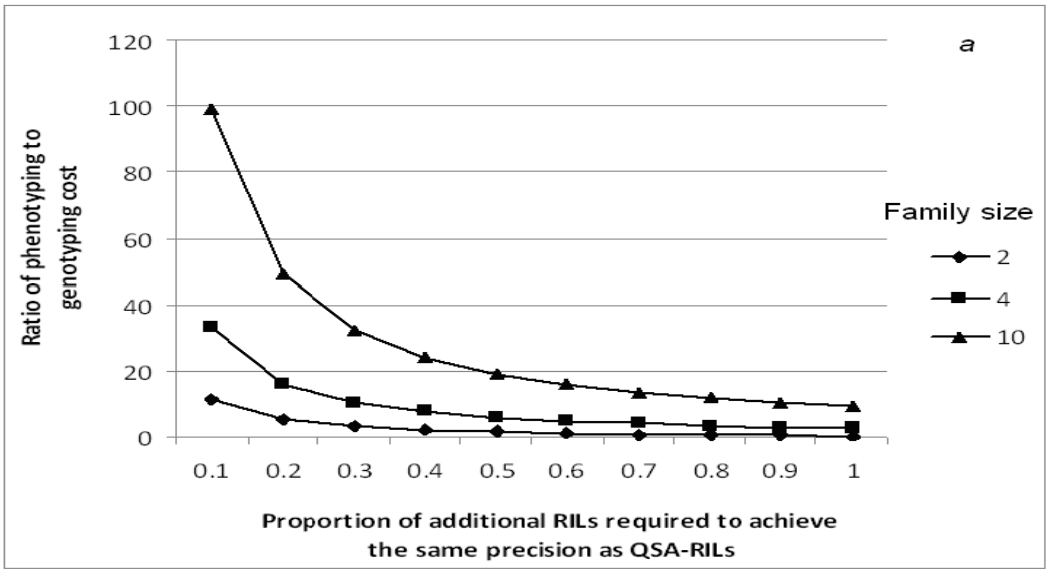


Figure 3.5 The ratio of phenotyping to genotyping costs for several family sizes and proportions of additional RILs required to achieve the same precision of QTL location estimate as in QSA-RILs

The proportion of additional RILs is with respect to number of QSA-RILs phenotyped required to achieve the same precision of QTL location estimate as QSA-RILs at the same cost. The proportion of markers genotyped in each generation after the F_2 is 10%. Panel *b* shows the same on the log scale.



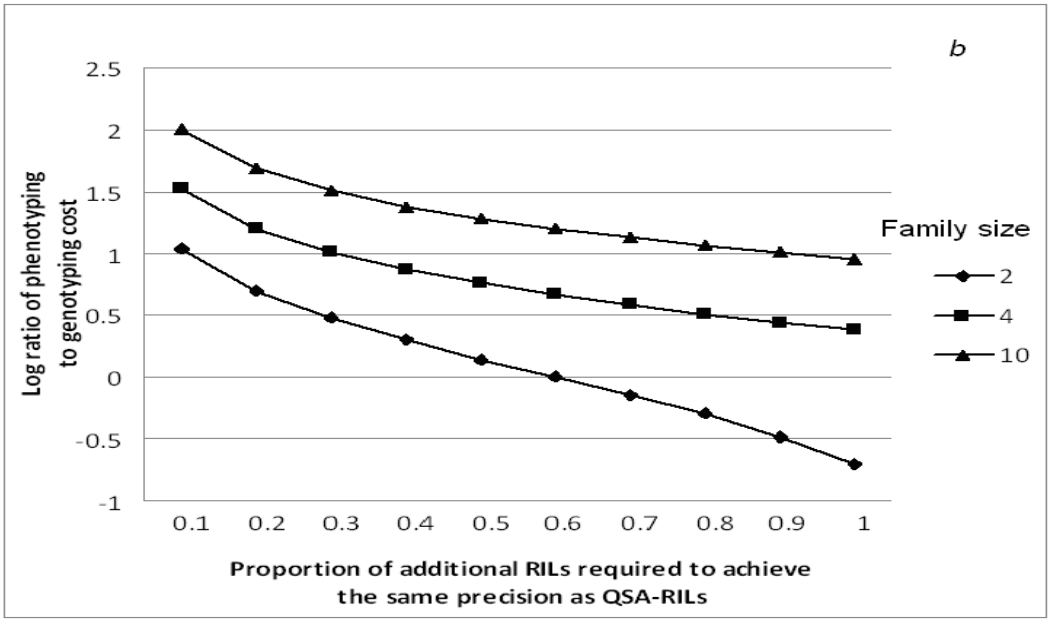


Table 3.1 QTL positions and effect signs used in all simulations

Chromosome	QTL position, cM	Effect sign
1	20.0	+
2	35.0	+
3	40.0	+
4	55.0	-
5	60.0	+

Table 3.2. Recombination between adjacent markers *A* and *B* for different two-locus genotypes

c is the recombination fraction between loci *A* and *B*.

Two-locus genotype	# Recombinations
AABb	1
aaBB or *AAbb*	2
AABb or *aaBb*	1
AaBB or *Aabb*	1
AaBb	$2c^2/(c^2+(1-c)^2)$

CHAPTER 4 - Derivations for Estimating Main and Interaction Effects in the BayesA (Xu) Method

Introduction

Epistasis, defined as the phenotypic effect of interactions among two or more loci, explains a considerable portion of phenotypic variation in quantitative traits (Carlborg and Haley, 2004). Most QTL mapping studies ignore the epistatic portion of the phenotypic variation because of the difficulty involved in the estimation of QTL interaction effects. The number of such effects to be estimated is often many times larger than the sample size.

In order to identify a set of markers and marker pairs that best explain the variation of a trait, we wish to evaluate several models, *i.e.,* combinations of markers and marker pairs. With increasing markers, the number of models in the model space increases exponentially. Because exploring the entire model space is infeasible, only a portion of the model space is evaluated. We select this portion by choosing a subset of markers based on some (often subjective) model-selection criteria. Several model-selection approaches with varying criteria for inclusion and exclusion of model variables have been proposed (Carlborg et al. 2000 and Yi et al. 2003, 2005, 2007).

In this chapter I extend the model-selection-free approach proposed by Xu (2003), which has been successfully used to identify multiple markers with main effects. I call this the BayesA (Xu) approach as the approach is similar to the BayesA method of Meuwissen et al. (2001) with some modifications by Xu (2003) made for linkage-based QTL mapping in experimental populations derived from inbred crosses. The advantage of BayesA (Xu) is twofold: firstly, the method is model-selection-free since all the marker effects across the genome are evaluated simultaneously, and secondly, the Bayesian framework allows estimation of a large number of effects, even more than the sample number. Although the implementation of BayesA (Xu) for estimating main effects was described by Xu (2003), the equations for the posterior mean and variance, used in estimation of the effects, were not elaborated. Here I derive the equations used

for the estimation of main effects for doubled-haploid and F_2 populations. I then extend these equations to estimate interaction effects in doubled-haploid populations.

BayesA (Xu) method

Main-effects model for a doubled-haploid (DH) population

Let y_i be the phenotype of an individual i and x_{ij} the genotype of an individual i for locus j. Let there be p loci. Two genotypes are possible at each locus in doubled-haploid populations. The genotypes are coded as +1 and -1, so that the additive effect can be calculated as the difference in genetic effect due to genotype *AA* and that due to genotype *aa*. The model for a DH individual can be written as follows.

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + e_i \tag{4.1}$$

where e_i follows $N(0, \sigma^2_0)$ and b_j is the additive effect due to marker j. Marker effects are considered a random variable in Bayesian analysis. The prior distributions for the effects are $p(b_0) \sim 1/\sigma^2_0$, $p(b_j) \sim \mathrm{N}(0, \sigma^2_j)$, $p(\sigma^2_j) \sim 1/\sigma^2_j$. So the parameters to be estimated are $[b,v] = b_0, b_1, b_2, \ldots. b_p\ \sigma^2_0, \sigma^2_1, \sigma^2_2, \ldots, \sigma^2_p$

The likelihood of the phenotype data given the DH model parameters

$$p(y \mid b,v) = \prod_{i=1}^{n} p(y_i \mid b,v)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\frac{1}{2\sigma_0^2}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2}$$

$$p(y \mid b,v) \sim \sigma_0^{-n/2} e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2} \tag{4.2}$$

The posterior distribution of the DH model parameters given the phenotypic data

$$p(b,v \mid y) \sim p(y \mid b,v)p(b,v)$$

$$\sim \sigma_0^{-n/2} e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2} \frac{1}{\sigma_0^2} \prod_j^p \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{1}{2\sigma^2} b_j^2} \prod_j^p \frac{1}{\sigma_j^2}$$

$$\sim \sigma_0^{-n/2} e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2} \frac{1}{\sigma_0^2} \prod_j^p \frac{1}{\sigma_j} e^{-\frac{1}{2\sigma^2} b_j^2} \prod_j^p \frac{1}{\sigma_j^2} \tag{4.3}$$

Estimation of b_0

Ignoring terms other than b_0 we have the kernel of the distribution of b_0

$$P(b_0 \mid b - b_0, v, y) \sim c e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2}$$

where c is a constant and $b - b_0$ includes all parameters of b except b_0.

Let $y_i - \sum_j^p x_{ij} b_j$ be denoted as k_i. Then

$$P(b_0 \mid b - b_0, v, y) \sim c e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(b_0 - k_i)^2}$$

$$\sim c e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(b_0^2 + k_i^2 - 2b_0 k_i)}$$

$$\sim c e^{-\frac{1}{2\sigma_0^2}(n b_0^2 + \sum_i^n k_i^2 - 2b_0 \sum_i^n k_i)}$$

$$P(b_0 \mid b - b_0, v, y) \sim c e^{-\frac{1}{2(\sigma_0^2/n)}(b_0^2 + \frac{1}{n}\sum_i^n k_i^2 - 2b_0 \frac{1}{n}\sum_i^n k_i)}$$

The mean of the posterior distribution is

$$\frac{1}{n}\sum_i^n k_i = \frac{1}{n}\sum_1^n (y_i - \sum_j^p x_{ij} b_j) \tag{4.4}$$

The variance of the distribution is

$$\frac{\sigma_0^2}{n} \tag{4.5}$$

Estimation of b_t

$$p(b_t \mid b - b_t, v, y) \sim c e^{-\frac{1}{2\sigma_0^2}\sum_i^n (y_i - b_0 - \sum_{j \neq t}^{p} x_{ij} b_j - x_{it} b_t)^2} e^{-\frac{1}{2\sigma_t^2} b_t^2}$$

Let $y_i - b_0 - \sum_{j \neq t}^{p} x_{ij} b_j$ be denoted as k_i. Then

$$p(b_t \mid b - b_t, v, y) \sim c e^{-\frac{1}{2\sigma_0^2}\sum_i^n (k_i - x_{it} b_t)^2 - \frac{1}{2\sigma_t^2} b_t^2}$$

$$\sim ce^{-\frac{1}{2\sigma_0^2}\sum_i^n (k_i^2 + x_{it}^2\ b_t^2 - 2x_{it}b_t\ k_i) - \frac{1}{2\sigma_t^2}b_t^2}$$

$$\sim ce^{-\frac{1}{2\sigma_0^2}\sum_i^n k_i^2 + \sum_i^n\ x_{it}^2\ b_t^2 - 2b_t\sum_i^n\ x_{it}k_i) - \frac{1}{2\sigma_t^2}b_t^2}$$

$$\sim\ ce^{-\frac{1}{2\sigma_0^2\ /\ \sum_i^n x_{it}^2}(b_t^2 + \frac{1}{\sum_i^n x_{it}^2}\sum_i^n k_i^2 - 2b_t\sum_i^n\ x_{it}k_i) - \frac{1}{2\sigma_t^2}b_t^2}$$

$$\sim ce^{-\frac{1}{2}(\frac{\sum_i^n x_{it}^2}{\sigma_0^2} + \frac{1}{\sigma_t^2})b_t^2 - \frac{\sum_i^n x_{it}^2}{2\sigma_0^2}\frac{1}{\sum_i^n x_{it}^2}\sum_i^n k_i^2 - \frac{2b_t}{\left(\frac{\sum_i^n x_{it}^2}{\sigma_0^2} + \frac{1}{\sigma_t^2}\right)^{-1}}\left(\frac{\left(\frac{\sum_i^n x_{it}^2}{\sigma_0^2} + \frac{1}{\sigma_t^2}\right)^{-1}}{\sigma_0^2}\right)\sum_i^n\ x_{it}k_i}$$

The mean of this posterior distribution is

$$b_t = \left(\frac{\sum_i^n x_{it}^2}{\sigma_0^2} + \frac{1}{\sigma_t^2}\right)^{-1}\sum_i^n x_{it}k_i$$

$$= \left(\frac{\sum_i^n x_{it}^2}{\sigma_0^2} + \frac{1}{\sigma_t^2}\right)^{-1}\sum_i^n x_{it}(y_i - b_0 - \sum_{j\neq t}^p x_{ij}b_j) \tag{4.6}$$

The variance of the distribution is

$$\sigma_0^2\left(\sum_i^n x_{it}^2 + \frac{\sigma_0^2}{\sigma_t^2}\right)^{-1} \tag{4.7}$$

Derivation of σ_0^2 posterior

$$p(v_0 \mid b, v - v_0, y) \sim c\sigma_0^{2(-n/2)}e^{-\frac{1}{2\sigma_0^2}\sum_i^n (y_i - b_0 - \sum_j^p x_{ij}b_j)^2\frac{1}{\sigma_0^2}}$$

Let $(y_i - b_0 - \sum_{j=1}^p b_j x_{ij})^2$ be k_i

$$p(v_0 \mid b, v - v_0, y) \sim c\sigma_0^{2(-\frac{n}{2}-1)}e^{-\frac{1}{2\sigma_0^2}\sum_i^n k_i}$$

$$\sim c\frac{1}{(\sum_i^n k_i)^{2(-\frac{n}{2}-1)}}(\sum_i^n k_i\sigma)^{2(-\frac{n}{2}-1)}e^{-\frac{1}{2\sigma_0^2}\sum_i^n k_i}$$

$$\sim \frac{1}{\chi_n^2}(\sum_i^n k_i)$$

$$\sim \frac{1}{\chi_n^2}\sum_i^n (y_i - b_0 - \sum_{j=1}^p b_j x_{ij})^2$$

$$\sim \frac{1}{\chi_n^2}\sum_i^n (y_i - b_0 - \sum_{j=1}^p b_j x_{ij})^2 \tag{4.8}$$

Derivation of σ_t^2 posterior

$$p(v_t \mid b, v - v_t, y) \sim c\sigma_t^{-1} e^{-\frac{1}{2\sigma_t^2} b_t^2}$$

$$\sim \frac{1}{\chi_1^2} b_t^2 \tag{4.9}$$

Main-effects model for F_2 population

The model for an F_2 design can be written as follows:

$$y_i = b_0 + \sum_{j=1}^p b_j x_{ij} + \sum_j^p d_j w_{ij} + e_i \tag{4.10}$$

x_{ij} are coded as -√2, 0, √2 for genotypes *AA*, *Aa*, *aa* and w_{ij} are coded as -1 1 and -1 for genotypes *AA*, *Aa*, *aa*.

The parameters to be estimated are

$$\Theta = [b_0, b, d, v_b, v_d, v_0] = b_0, b_1, b_2 \ldots b_p, d_1, d_2, \ldots d_p, \sigma_0^2, \sigma_{b1}^2, \ldots \sigma_{bp}^2, \sigma_{d1}^2, \ldots \sigma_{dp}^2$$

The priors for the parameters are

$$p(b_0) = 1, p(b_j) \sim N(0, \sigma_{bj}^0), p(d_j) \sim N(0, \sigma_{dj}^2), p(\sigma_{bj}^2) \sim \frac{1}{\sigma_{bj}^2}, p(\sigma_{dj}^2) \sim \frac{1}{\sigma_{dj}^2}$$

The likelihood of the phenotype data given the F_2 model parameters

$$p(y \mid \Theta) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(y_i - b_0 - \sum_{j=1}^p b_j x_{ij} - \sum_j^p d_j w_{ij})^2}$$

$$p(y \mid \Theta) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(y_i - b_0 - \sum_{j=1}^p b_j x_{ij} - \sum_j^p d_j w_{ij})^2}$$

$$p(y \mid \Theta) \sim (\sigma_0^2)^{-n/2} e^{-\frac{1}{2\sigma_0^2}\sum_i (y_i - b_0 - \sum_{j=1}^p b_j x_{ij} - \sum_j^p d_j w_{ij})^2}$$

$$p(y \mid \Theta) \sim (\sigma_0^2)^{-n/2} e^{-\frac{1}{2\sigma_0^2}\sum_i (y_i - b_0 - \sum_{j=1}^p b_j x_{ij} - \sum_j^p d_j w_{ij})^2}$$

The posterior distribution of F_2 model parameters given the phenotypic data

$$p(\Theta \mid y) \sim p(y \mid \Theta) p(\Theta)$$

$$\sim (\sigma_0^2)^{-n/2} e^{-\frac{1}{2\sigma_0^2}\sum_i (y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} - \sum_j^p d_j w_{ij})^2} \prod_i^n \frac{1}{\sqrt{2\pi\sigma_{bj}^2}} e^{-\frac{1}{2\sigma_{bj}^2} b_j^2} \frac{1}{\sqrt{2\pi\sigma_{dj}^2}} e^{-\frac{1}{2\sigma_{dj}^2} d_j^2} \frac{1}{\sigma_{bj}^2 \sigma_{dj}^2}$$

Note: For convenience the derivations for the parameters will be done on the log scale.

Estimation of b_0

Let k_i denote $y_i - \sum_j^p x_{ij} b_j - \sum_j^p d_j w_{ij}$. Ignoring terms other than b_0 ,

$$\log(p(b_0 \mid y, \Theta - b_0) \sim \frac{1}{\sigma_0^2} \sum_1^n (b_0 - k_i)^2$$

$$\sim \frac{1}{\sigma_0^2} (n b_0^2 - 2 b_0 \sum_i k_i + \sum_i k_i^2)$$

The mean of the posterior distribution is

$$\frac{1}{n} \sum_i (y_i - \sum_j^p x_{ij} b_j - \sum_j^p w_{ij} d_j) \tag{4.11}$$

The variance of the distribution is

$$\sigma_0^2 / n \tag{4.12}$$

Estimation of b_t

Let k_i denote $y_i - b_0 - \sum_{j \neq t}^{p} x_{ij} b_j - \sum_j^p d_j w_{ij}$. Ignoring terms other than b_t,

$$\log(p(b_t \mid y, \Theta - b_t) \sim \frac{1}{\sigma_0^2} \sum_i^n (x_{it} b_t - k_i)^2 + \frac{b_t^2}{\sigma_{bt}^2}$$

$$\sim \frac{1}{\sigma_0^2} \sum_i^n (x_{it}^2 b_t^2 - 2 x_{it} b_t k_i + k_i^2) + \frac{b_t^2}{\sigma_{bt}^2}$$

$$\sim \frac{1}{\sigma_0^2} ((\sum_i^n x_{it}^2 + \frac{\sigma_0^2}{\sigma_{bt}^2}) b_t^2 - 2 b_t \sum_i x_{it} k_i + \sum_i k_i^2)$$

The mean of the posterior distribution is

$$\left(\sum_i x_{it}^2 + \frac{\sigma_0^2}{\sigma_{bt}^2} \right)^{-1} \sum_i x_{it} \left(y_i - b_0 - \sum_{j \neq t}^{p} x_{ij} b_j - \sum_j^p d_j w_{ij} \right) \tag{4.13}$$

The variance of the distribution is

$$\sigma_0^2\left(\sum_i x_{it}^2+\frac{\sigma_0^2}{\sigma_{bt}^2}\right)^{-1} \tag{4.14}$$

Estimation of d_t

Let k_i denote $y_i-b_0-\sum_j^p x_{ij}b_j-\sum_{j\neq t}^p d_j w_{ij}$. Ignoring terms other than d_t,

$$\log(p(d_t\mid y,\Theta-d_t)\sim\frac{1}{\sigma_0^2}\sum_i^n(w_{it}d_t-k_i)^2+\frac{d_t^2}{\sigma_{dt}^2}$$

$$\sim\frac{1}{\sigma_0^2}\sum_i^n(w_{it}^2d_t^2-2w_{it}d_tk_i+k_i^2)+\frac{d_t^2}{\sigma_{dt}^2}$$

$$\sim\frac{1}{\sigma_0^2}((\sum_i^n w_{it}^2+\frac{\sigma_0^2}{\sigma_{dt}^2})d_t^2-2d_t\sum_i w_{it}k_i+\sum_i k_i^2)$$

The mean of the posterior distribution is

$$\left(\sum_i^n w_{it}^2+\frac{\sigma_0^2}{\sigma_{dt}^2}\right)^{-1}\sum_i^n w_{it}\left(y_i-b_0-\sum_j^p x_{ij}b_j-\sum_{j\neq t}^p d_j w_{ij}\right) \tag{4.15}$$

The variance of the distribution is

$$\sigma_0^2\left(\sum_i^n w_{it}^2+\frac{\sigma_0^2}{\sigma_{dt}^2}\right)^{-1} \tag{4.16}$$

Derivation of σ_0^2 posterior

Ignoring the terms other than σ_0^2, we have

$$p(v_0\mid y,\Theta-v_0)\sim\sigma_0^{2(-n/2)}e^{-\frac{1}{2\sigma_0^2}\sum_i^n\left(y_i-b_0-\sum_j^p b_j x_{ij}-\sum_j^p d_j w_{ij}\right)^2}\frac{1}{\sigma_0^2}$$

Let $\sum_i^n\left(y_i-b_0-\sum_j^p b_j x_{ij}-\sum_j^p d_j w_{ij}\right)^2$ be k_i then

$$p(v_0\mid y,\Theta-v_0)\sim\sigma_0^{2(-\frac{n}{2}-1)}e^{-\frac{1}{2\sigma_0^2}k_i}$$

$$\sim\frac{1}{(k_i)^{-\frac{n}{2}-1}}(k_i\sigma_0)^{2(-\frac{n}{2}-1)}e^{-\frac{1}{2\sigma_0^2}k_i}$$

$$\sim (k_i)\frac{1}{\chi_n^2}$$

$$\sim \frac{1}{\chi_n^2}\sum_i^n\left(y_i - b_0 - \sum_j^p b_j x_{ij} - \sum_j^p d_j w_{ij}\right)^2 \tag{4.17}$$

Derivation of σ_{bt}^2 posterior

Ignoring the terms other than σ_{bt}^2, we have

$$p(v_{bt} \mid y, \Theta - v_{bt}) \sim \frac{1}{\sigma_t} e^{-\frac{1}{2\sigma_{bt}^2} b_t^2} \frac{1}{\sigma_{bt}^2}$$

$$\sim \frac{1}{\chi_1^2} b_t^2 \tag{4.18}$$

Derivation of σ_{dt}^2 posterior

Ignoring the terms other than σ_{dt}^2, we have

$$p(v_{dt} \mid y, \Theta - v_{dt}) \sim \frac{1}{\sigma_t} e^{-\frac{1}{2\sigma_{dt}^2} d_t^2} \frac{1}{\sigma_{dt}^2}$$

$$\sim \frac{1}{\chi_1^2} b_t^2 \tag{4.19}$$

Interaction-effects model for a doubled-haploid design

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + \sum_{j}^{j=pg} g_{ij}^{aa} w_{ij}^{aa} + e_i \tag{4.20}$$

The parameters to be estimated are

$$\Theta = [b_0, b, g, v_b, v_{g^{aa}}, v_0] = b_0, b_1, b_2 .... b_p, g_1^{aa}, g_2^{aa}, .., g_{p_{ii}}^{aa}, \sigma_0^2, \sigma_{b1}^2, ... \sigma_{g_1^{aa}}^2, \sigma_{g_2^{aa}}^2, ... \sigma_{g_{pii}^{aa}}^2$$

The priors for the parameters are

$$p(b_0) = 1, p(b_j) \sim N(0, \sigma_{bj}^0), p(g_j^{aa}) \sim N(0, \sigma_{g_t^{aa}}^2), p(\sigma_{bj}^2) \sim \frac{1}{\sigma_{bj}^2}, p(\sigma_{g_t^{aa}}^2) \sim \frac{1}{\sigma_{g_t^{aa}}^2}$$

Likelihood equation for the phenotype data given DH interaction model parameters

$$p(y \mid \Theta) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} - \sum_j^{pg} g_j^{aa} w_{ij})^2}$$

$$p(y \mid \Theta) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} - \sum_j^{pg} g_j^{aa} w_{ij})^2}$$

$$\sim (\sigma_0^2)^{-n/2} e^{-\frac{1}{2\sigma_0^2}\sum_i (y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} - \sum_j^{pg} g_j^{aa} w_{ij})^2}$$

The posterior distribution of the interaction model parameters given the phenotype data

$$p(\Theta \mid y) \sim p(y \mid \Theta) p(\Theta)$$

$$\sim (\sigma_0^2)^{-n/2} e^{-\frac{1}{2\sigma_0^2}\sum_i (y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} - \sum_j^{pg} g_j^{aa} w_{ij})^2} \prod_i^n \frac{1}{\sqrt{2\pi\sigma_{bj}^2}} e^{-\frac{1}{2\sigma_{bj}^2} b_j^2} \frac{1}{\sqrt{2\pi\sigma_{g_j^{aa}}^2}} e^{-\frac{1}{2\sigma_{g_j^{aa}}^2} g_j^{aa2}} \frac{1}{\sigma_{bj}^2 \sigma_{g_j^{aa}}^2}$$

Note: For convenience the derivations for the parameters will be done on the log scale.

Estimation of b_0.

Let k_i denote, $y_i - \sum_j^p x_{ij} b_j - \sum_j^{pg} g_j^{aa} w_{ij}$. Ignoring terms other than b_0,

$$\log(p(b_0 \mid y, \Theta - b_0) \sim \frac{1}{\sigma_0^2} \sum_1^n (b_0 - k_i)^2$$

$$\sim \frac{1}{\sigma_0^2}(n b_0^2 - 2 b_0 \sum_i k_i + \sum_i k_i^2)$$

The mean of the posterior distribution is

$$\frac{1}{n}\sum_i (y_i - \sum_j^p x_{ij} b_j - \sum_j^{pg} w_{ij} g_j^{aa}) \tag{4.21}$$

The variance of the distribution is

$$\sigma_0^2 / n \tag{4.22}$$

Estimation of b_t

Let k_i denote, $y_i - b_0 - \sum_{j \neq t}^p x_{ij} b_j - \sum_j^{pg} g_j^{aa} w_{ij}$. Ignoring terms other than b_t,

$$\log(p(b_t \mid y, \Theta - b_t) \sim \frac{1}{\sigma_0^2} \sum_i^n (x_{it} b_t - k_i)^2 + \frac{b_t^2}{\sigma_{bt}^2}$$

$$\sim \frac{1}{\sigma_0^2}\sum_i^n (x_{it}^2 b_t^2 - 2x_{it}b_t k_i + k_i^2) + \frac{b_t^2}{\sigma_{bt}^2}$$

$$\sim \frac{1}{\sigma_0^2}((\sum_i^n x_{it}^2 + \frac{\sigma_0^2}{\sigma_{bt}^2})b_t^2 - 2b_t\sum_i x_{it}k_i + \sum_i k_i^2)$$

The mean of the posterior distribution is

$$\left(\sum_i x_{it}^2 + \frac{\sigma_0^2}{\sigma_{bt}^2}\right)^{-1}\sum_i x_{it}\left(y_i - b_0 - \sum_{j\neq t}^{p} x_{ij}b_j - \sum_j^{pg} g_j^{aa} w_{ij}\right) \tag{4.23}$$

The variance of the distribution is

$$\sigma_0^2\left(\sum_i x_{it}^2 + \frac{\sigma_0^2}{\sigma_{bt}^2}\right)^{-1} \tag{4.24}$$

Derivation of g_t^{aa} posterior

Let k_i denote $y_i - b_0 - \sum_j^p x_{ij}b_j - \sum_{j\neq t}^{pg} g_j^{aa} w_{ij}$. Ignoring terms other than g_t^{aa},

$$\log(p(g_t^{aa} \mid y, \Theta - g_t^{aa}) \sim \frac{1}{\sigma_0^2}\sum_i^n (w_{it}g_t^{aa} - k_i)^2 + \frac{g_t^{aa2}}{\sigma_{g_t^{aa}}^2}$$

$$\sim \frac{1}{\sigma_0^2}\sum_i^n (w_{it}^2 g_t^{aa2} - 2w_{it}g_t^{aa}k_i + k_i^2) + \frac{g_t^{aa}2}{\sigma_{g_t^{aa}}^2}$$

$$\sim \frac{1}{\sigma_0^2}((\sum_i^n w_{it}^2 + \frac{\sigma_0^2}{\sigma_{g_t^{aa}}^2})g_t^{aa2} - 2g_t^{aa}\sum_i w_{it}k_i + \sum_i k_i^2)$$

The mean of the posterior distribution is

$$\left(\sum_i w_{it}^2 + \frac{\sigma_0^2}{\sigma_{g_t^{aa}}^2}\right)^{-1}\sum_i w_{it}\left(y_i - b_0 - \sum_j^p x_{ij}b_j - \sum_{j\neq t}^{pg} g_j^{aa} w_{ij}\right) \tag{4.25}$$

The variance of the posterior distribution is

$$\sigma_0^2\left(\sum_i w_{it}^2 + \frac{\sigma_0^2}{\sigma_{g_t^{aa}}^2}\right)^{-1} \tag{4.26}$$

Derivation of σ_0^2 posterior

Ignoring the terms other than σ_0^2 we have

$$p(v_0 \mid y, \Theta - v_0) \sim \sigma_0^{2(-n/2)} e^{-\frac{1}{2\sigma_0^2}\sum_i^n \left(y_i - b_0 - \sum_j^p b_j x_{ij} - \sum_j^{pg} g_j^{aa} w_{ij} \right)^2} \frac{1}{\sigma_0^2}$$

Let $\sum_i^n \left(y_i - b_0 - \sum_j^p b_j x_{ij} - \sum_j^{pg} g_j^{aa} w_{ij} \right)^2$ be k_i, then

$$p(v_0 \mid y, \Theta - v_0) \sim \sigma_0^{2(-\frac{n}{2}-1)} e^{-\frac{1}{2\sigma_0^2} k_i}$$

$$\sim \frac{1}{(k_i)^{-\frac{n}{2}-1}} (k_i \sigma_0)^{2(-\frac{n}{2}-1)} e^{-\frac{1}{2\sigma_0^2} k_i}$$

$$\sim (k_i) \frac{1}{\chi_n^2}$$

$$\sim \frac{1}{\chi_n^2} \sum_i^n \left(y_i - b_0 - \sum_j^p b_j x_{ij} - \sum_j^{pg} g_j^{aa} w_{ij} \right)^2 \tag{4.27}$$

Derivation of σ_{bt}^2 posterior

Ignoring the terms other than σ_{bt}^2 we have

$$p(v_{bt} \mid y, \Theta - v_{bt}) \sim \frac{1}{\sigma_t} e^{-\frac{1}{2\sigma_{bt}^2} b_t^2} \frac{1}{\sigma_{bt}^2}$$

$$\sim \frac{1}{\chi_1^2} b_t^2 \tag{4.28}$$

Derivation of $\sigma_{g_t^{aa}}^2$ posterior

Ignoring the terms other than $\sigma_{g_t^{aa}}^2$ we have

$$p(v_{g_t^{aa}} \mid y, \Theta - v_{g_t^{aa}}) \sim \frac{1}{\sigma_t} e^{-\frac{1}{2\sigma_{g_t^{aa}}^2} g_t^{aa2}} \frac{1}{\sigma_{g_t^{aa}2}^2}$$

$$\sim \frac{1}{\chi_1^2} g_j^{aa2} \tag{4.29}$$

MCMC implementation of the BayesA (Xu) method

The implementation steps, described by Xu (2003), are presented here with some modifications to suit the interaction model.

Step 1: Initialize

The parameters are denoted as $\Theta = b_0, b_1, b_2 .... b_p, g_1^{aa}, g_2^{aa}, .., g_{p_g}^{aa}, \sigma_0^2, \sigma_{b1}^2, ... \sigma_{g_1^{aa}}^2, \sigma_{g_2^{aa}}^2, ... \sigma_{g_{pg}^{aa}}^2$

The effect parameters $b_0, b_1, b_2 .... b_p, g_1^{aa}, g_2^{aa}, .., g_{pg}^{aa}$ are initialized to zeros and the variance parameters, $\sigma_0^2, \sigma_{b1}^2, ..., \sigma_{g_1^{aa}}^2, \sigma_{g_2^{aa}}^2, ... \sigma_{g_{pg}^{aa}}^2$ are initialized to a positive number (here 0.5)

Step 2: Update effects

Mean: b_0^{r+1} is sampled from a normal distribution with mean and variance as specified in equations 4.21 and 4.22. b_0^{r+1} will replace b_0^r in all the equations, where *r* indexes the MCMC iterations.

Additive effects: b_t^{r+1} is sampled from a normal distribution with mean and variance specified in equations 4.23 and 4.24. b_t^{r+1} will replace b_t^r during subsequent calculations.

Additive by additive interaction effect: $g_t^{aa(r+1)}$ is sampled from a normal distribution with mean and variance specified in equations 4.25 and 4.26.

Step 3: Update variances

Error variance: $\sigma_0^{2(r+1)}$ is sampled from a distribution specified in equation 4.27.

Additive effect variances: $\sigma_{b1}^{2(r+1)}$ is sampled from a distribution specified in equation 4.28.

Additive x additive effect variance: $\sigma_{ii_1^{aa}}^{2(r+1)}$ is sampled from a distribution specified in equation 4.29.

Step 4: Repeat steps 2 and 3 until the MCMC chain converges to a stationary distribution

Code optimization

BayesA (Xu) implementation is computationally intensive because of the total number of iterations (often in excess of 30,000) involved in Gibbs sampling and the requirement for estimation of all additive and interaction effects during every iteration (cycle). One of the time-consuming steps during estimation of additive interaction effects in each cycle, $g_t^{aa(r+1)}$ is the following operation:

$$\sum_i w_{it}\left(y_i - b_0 - \sum_j^p x_{ij} b_j - \sum_{j \neq t}^{pg} g_j^{aa} w_{ij} \right).$$

The equation can be elaborated as

$$= \sum_i w_{it} y_i - \sum_i w_{it} b_0 - \sum_i w_{it} \sum_j^p x_{ij} b_j - \sum_i w_{it} \sum_{j \neq t}^{pg} g_j^{aa} w_{ij}$$

$$= \sum_i w_{it} y_i - \sum_i w_{it} b_0 - \sum_i w_{it} \sum_j^p x_{ij} b_j - \sum_i w_{it} \sum_j^{pg} g_j^{aa} w_{ij} + \sum_i w_{it}^2 g_t^{aa}$$

The number of operations required for this computation is $n + n + np + np(p\text{-}1)/2 + n$ and is of order np^2. For instance, if $n = 200$ and $p = 300$, the total number of operations computed in each cycle for estimating each interaction exceeds one million. Substantial savings can be achieved by storing in variables the values of the terms $\sum_i w_{it} y_i$, $\sum_i w_{it}$, $\sum_j^p x_{ij} b_j$, $\sum_j^{pg} g_j^{aa} w_{ij}$, and $\sum_i w_{it}^2$. Of these only the term $\sum_j^{pg} g_j^{aa} w_{ij}$ is updated. This is done by addition of $(g_t^{aa(r+1)} - g_t^{aa(r)})\, w_{it}$ to $\sum_j^{pg} g_j^{aa} w_{ij}$. The number of operations per cycle per interaction effect can thus be reduced to order n from np^2.

References

Balint-Kurti, P.J., J.C. Zwonitzer, R.J. Wisser, M.L. Carson, M.A. Oropeza-Rosas, J.B. Holland, and S.J. Szalma. 2007. Precise mapping of quantitative trait loci for resistance to southern leaf blight, caused by *Cochliobolus heterostrophus* race O, and flowering time using advanced intercross maize lines. Genetics 176:645–657.

Becanovic, K., M. Jagodic, J.R. Sheng, I. Dahlman, F. Aboul-Enein, E. Wallstrom, P. Olofsson, R. Holmdahl, H. Lassmann and T. Olsson. 2006. Advanced intercross line mapping of *Eae5* reveals *ncf-1* and *CLDN4* as candidate genes for experimental autoimmune encephalomyelitis. J. Immunol. 176:6055-6064

Boddhireddy, P., J.L. Jannink , J.C. Nelson JC. 2009. Selective advance for accelerated development of recombinant inbred QTL-mapping populations. Crop Sci. 49: 1284-1294.

Carlborg, O. and C.S. Haley. 2004. Epistasis: Too often neglected in complex trait studies? Nat. Rev. Genet. 5:618-625.

Carlborg, O., L. Andersson and B. Kinghorn. 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. Genetics 155:2003-2010.

Carson, M.L.1998. Aggressiveness and perennation of isolates of *Cochliobolus heterostophus* from North Carolina. Plant Dis. 82:1043-1047.

Carson, M.L., C.W. Stuber and M.L. Senior. 2004. Identification and mapping of quantitative trait loci conditioning resistance to southern leaf blight of maize caused by *Cochliobolus heterostrophus* race O. Phytopathology 94:862-867.

Churchill, G.A. and R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963-971.

Darvasi, A. 1998. Experimental strategies for the genetic dissection of complex traits in animal models. Nat. Genet. 18:19-24.

Darvasi, A. and M. Soller. 1995. Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141:1199-1207.

Darvasi, A. and M. Soller. 1997. A simple method to calculate resolving power and confidence interval of QTL map location. Behav. Genet. 27:125-132.

Darvasi, A., and M. Soller. 1995. Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141:1199–1207.

Dekkers, J.C., and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3:22–32.

Falconer, D.S. and T.F.C. Mackay. 1996. Introduction to Quantitative Genetics. Prentice Hall, Harlow, UK.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinburgh 52:399-433

Haldane, J.B.S. 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. Genetics 8:299-309.

Haley, C.S. and S.A. Knott. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315-324.

Haseman, J.K. and R.C. Elston. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. 2:3-19.

Hayes, B. and M.E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33:209-229.
Hubner, N., C.A. Wallace, H. Zimdahl, E. Petretto, H. Schulz, F. Maciver, M. Mueller, O. Hummel, J. Monti, V. Zidek, A. Musilova, V. Kren, H. Causton, L. Game, G. Born, S. Schmidt, A. Muller, S.A. Cook, T.W. Kurtz, J. Whittaker, M. Pravenec, and T.J. Aitman. 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat. Genet. 37:243–253.
Jannink, J.L. 2005. Selective phenotyping to accurately map quantitative trait loci. Crop Sci. 45: 901-908.
Jansen, R.C. 1993. Interval mapping of multiple quantitative trait loci. Genetics 135:205-211.
Jansen, R.C. and P. Stam. 1994. High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447-1455.
Jiang, C. and Z.B. Zeng. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140:1111-1127.
Jin, C., H. Lan, A.D. Attie, G.A. Churchill, D. Bulutuglo and B.S. Yandell. 2004. Selective phenotyping for increased efficiency in genetic mapping studies. Genetics 168:2285-2293.
Joehanes, R., and J.C. Nelson. 2008. QGene 4.0, an extensible Java QTL-analysis platform. Bioinformatics 24:2788–2789.
Karlin, S. 1984. Theoretical aspects of genetic map functions in recombination processes. In Human Population Genetics: The Pittsburgh symposium edited by A Chakravarti, 209-228. Van Nostrand Reinhold, New York.
Kearsey, M.J., and A.G. Farquhar. 1998. QTL analysis in plants: Where are we now? Heredity 80:137–142.
Kosambi, D. D. 1944. The estimation of map distances from recombination values. Ann. Eugen. 12: 172-175.
Lander, E.S. and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199.
Lee, M., N. Sharopova, W.D. Beavis, D. Grant, M. Katt, D. Blair, and A. Hallauer. 2002. Expanding the genetic map of maize with the intermated *B73 × Mo17* (IBM) population. Plant Mol. Biol. 48:453–461.
Liu, S.C., S.P. Kowalski, T.H. Lan, K.A. Feldmann, and A.H. Paterson. 1996. Genome-wide high-resolution mapping by recurrent intermating using *Arabidopsis thaliana* as a model. Genetics 142:247–258.
Luo, Z.W. and M.J. Kearsey. 1992. Interval mapping of quantitative trait loci in an F_2 population. Heredity 69:236-242.
Lynch, M., and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. Sinauer, Sunderland, MA.
Mangin, B., and B. Goffinet. 1997. Comparison of several confidence intervals for QTL location. Heredity 78:345–353.
Mangin, B., B. Goffinet and A. Rebai. 1994. Constructing confidence intervals for QTL location. Genetics 138:1301-1308.
Martin, O.C. and F. Hospital. 2006. Two- and three-locus tests for linkage analysis using recombinant inbred lines. Genetics 173:451-459.
Meuwissen, T.H., B.J. Hayes and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Nadeau, J.H. and W.N. Frankel. 2000. The roads from phenotypic variation to gene discovery: Mutagenesis versus QTLs. Nat. Genet. 25:381-384.
Otto, S.P. and C.D. Jones. 2000. Detecting the undetected: Estimating the total number of loci underlying a quantitative trait. Genetics 156:2093-2107.
Rao, D.C., N.E. Morton, J. Lindsten, M. Hulten and S. Yee. 1977. A mapping function for man. Hum. Hered. 27:99-104.
Sharopova, N., M.D. McMullen, L. Schultz, S. Schroeder, H. Sanchez-Villeda, J. Gardiner, D. Bergstrom, K. Houchins, S. Melia-Hancock, T. Musket, N. Duru, M. Polacco, K. Edwards, T. Ruff, J.C. Register, C. Brouwer, R. Thompson, R. Velasco, E. Chin, M. Lee, W. Woodman-Clikeman, M.J. Long, E. Liscum, K. Cone, G. Davis and E.H. Coe Jr. 2002. Development and mapping of SSR markers for maize. Plant Mol. Biol. 48:463-481.
Shrimpton, A.E. and A. Robertson. 1988. The isolation of polygenic factors controlling bristle score in *drosophila melanogaster*. II. distribution of third chromosome bristle effects within chromosome sections. Genetics 118:445-459.
Spielman, R.S., R.E. McGinnis and W.J. Ewens. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52:506-516.
Sturtevant, A.H. 1913. A third group of linked genes in *drosophila ampelophila*. Science 37:990-992.
Van Ooijen, J. W. 1992. Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet. 84:803-811.
Visscher, P.M. and M.E. Goddard. 2004. Prediction of the confidence interval of quantitative trait loci location. Behav. Genet. 34:477-482.
Visscher, P.M., R. Thompson and C.S. Haley. 1996. Confidence intervals in QTL mapping by bootstrapping. Genetics 143:1013-1020.
Wang, M., W.J. Lemon, G. Liu, Y. Wang, F.A. Iraqi, A.M. Malkinson, and M. You. 2003. Fine mapping and identification of candidate pulmonary adenoma susceptibility 1 genes using advanced intercross lines. Cancer Res. 63:3317–3324.
Weller, J.I., Y. Kashi and M. Soller. 1990. Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. J. Dairy Sci. 73:2525-2537
Winkler, C.R., N.M. Jensen, M.P. Cooper, W.S. Dean, and S. Oscar. 2003. On the determination of recombination rates in intermated recombinant inbred populations. Genetics 164:741–745.
Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. Genetics 163:789-801.
Xu, Z., F. Zou and T.J. Vision. 2005. Improving quantitative trait loci mapping resolution in experimental crosses by the use of genotypically selected samples. Genetics 170:401-408.
Yi, N., S. Xu and D.B. Allison. 2003. Bayesian model choice and search strategies for mapping interacting quantitative trait loci. Genetics 165:867-883.
Yi, N., D. Shriner, S. Banerjee, T. Mehta, D. Pomp and B.S. Yandell. 2007. An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. Genetics 176:1865-1877.
Yi, N., B.S. Yandell, G.A. Churchill, D.B. Allison, E.J. Eisen and D. Pomp. 2005. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics 170:1333-1344.

Yu, X., K. Bauer, P. Wernhoff, D. Koczan, S. Moller, H.J. Thiesen, and S.M. Ibrahim. 2006. Fine mapping of collagen-induced arthritis quantitative trait loci in an advanced intercross line. J. Immunol. 177:7042–7049.
Zeng, Z.B. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA 90:10972–10976.
Zeng, Z.B. 1994. Precision mapping of quantitative trait loci. Genetics 136:1457-1468.

Appendix A - Calculation of recombination due to selection

The approach proposed by Jannink (2005) for calculating recombination due to selection can be extended to the case of SA-RILs. For this derivation the following assumptions are made. The distance between flanking markers is close to zero so that the number of intervals, *n*, is large. The amount of genome that contributes toward new recombination falls by half on average at every generation. Because of the assumption of higher marker density than that used in our experiments, the formula derived here gives an estimate of the maximum recombination gains obtainable. For this derivation we follow the event of meiosis in one parental meiocyte. For application to selfing, where two meioses occur, *n* should be replaced by 2*n*. In *n* intervals of *x* cM each, the event of a crossover in a given interval at a given generation can be considered a Bernoulli random variable with probability *c*, the recombination fraction. The relationship between *c* and *x* under Haldane's mapping function is $c = \frac{1}{2}(1 - e^{-2x})$. The sum of such random variables approximates a Poisson distribution with mean $E(R)$ and standard deviation $S(R)$. The expected number of recombinations $E(R)$ is then *nc*, with standard deviation $S(R) = \sqrt{nc}$. The recombination in conventionally developed RILs at generation *i*, $E_i(R)$, can be calculated as $E_1(R) = nc$; then

$$E_2(R) = E_1(R) + \frac{n}{2}c \qquad [A1]$$

and

$$E_k(R) = nc\left\{1 + \frac{1}{2} + \frac{1}{4} \ldots \frac{1}{2^{k-1}}\right\} \approx 2nc \text{ as } k \text{ becomes large} \qquad [A2]$$

Without selection, the expected recombination after eight or more generations is twice that at F_2, in agreement with Liu et al. (1996). Similarly one can estimate the recombination under selection by applying the concepts of response to selection (Falconer and Mackay, 1996). Here the phenotype would be the number of recombinations, which follows a Poisson distribution. However, with large (>8 M) maps the distribution approaches normality (Fig. 6). A truncated Poisson distribution is thus used to compute the intensity of selection, *i*, which is the mean deviation of individuals exceeding the truncation point in units of standard deviation from

the population mean. The truncation point is derived from the proportion of individuals selected (Falconer and Mackay 1996). The expected recombinations due to selection are

$$E_1(R) = nc + \sqrt{nc}i \quad [A3]$$

$$E_2(R) = E_1(R) + \frac{n}{2}c + \sqrt{\frac{n}{2}c}i \quad [A4]$$

$$E_k(R) = E_{k-1}(R) + \frac{n}{2^{k-1}}c + \sqrt{\frac{n}{2^{k-1}}c}i \quad [A5]$$

$$E_k(R) = nc\left\{1 + \frac{1}{2} + \frac{1}{4} \ldots \frac{1}{2^{k-1}}\right\} + \sqrt{nc}i\left\{1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}^2} + \ldots \frac{1}{\sqrt{2}^{k-1}}\right\} \quad [A6]$$

$$E_k(R) = 2nc + \sqrt{nc}t(2 + \sqrt{2}) \quad [A7]$$

The relative increase of recombinations in SA-RILs at any generation can thus be expressed as

$$\frac{i\left\{1 + \frac{1}{\sqrt{2}} + \ldots . \frac{1}{\sqrt{2}^{k-1}}\right\}}{\sqrt{nc}\left\{1 + \frac{1}{2} + \frac{1}{4} \ldots . \frac{1}{2^{k-1}}\right\}} \quad [A8]$$

The advantage of selection is then $i/\sqrt{nc}$, a quantity that, as expected, increases with decreasing interval size or map length and increases with family size. Jannink (2005) and Xu et al. (2005) also found that the efficiency of selective phenotyping is inversely proportional to the square root of map length.

Permission to Reprint

Permission Granted
Frances Katz
Director of Publications
ASA-SSSA-CSSA

From: Prashanth Boddhireddy [mailto:reddy@ksu.edu]
Sent: Tuesday, August 11, 2009 11:35 AM
To: Frances Katz
Subject: Request to Reprint Content

Name	Prashanth Boddhireddy
Institution	Kansas State University
Department	Plant Pathology
Line 1	4022, Throckmorton, Kansas State University
City, State/Prov	Manhattan, KS
Postal Code	66506
Country	USA
Phone, Fax	2697167005, 7855325692
EMail	reddy@ksu.edu

I am the author of the work that is to be used.

Original Work

Original Format	Journal
Title of Work	Selective Advance for Accelerated Development of Recombinant Inbred QTL Mapping Populations
Author Name	P. Boddhireddy
Year	2009
Volume	49
Article Title	Selective Advance for Accelerated Development of Recombinant Inbred QTL Mapping Populations
Page Number(s)	1284-1294

Text I would like to reproduce all figures and tables

Reprinted Material

The work will be reproduced/used in the following format(s): Dissertation

Title	Development of highly recombinant inbred QTL mapping populations
Publisher	Kansas State University
Publisher's Location	Manhattan, KS
Quantity	1
Publication Date	August 14, 2009
Publisher Type	Non-Profit Publisher