Theor Appl Genet (2011) 123:897–906
DOI 10.1007/s00122-011-1635-3

ORIGINAL PAPER

Linkage analysis in unconventional mating designs in line crosses

James C. Nelson

Received: 3 January 2011 / Accepted: 3 June 2011 / Published online: 18 June 2011

Abstract Linkage estimation and genetic map construction with genotyped DNA markers in plants preferentially employ a few maximally informative early-generation or recombinant-inbred mating designs. Fitting their recombination models to unconventional designs adapted to cultivar development (series of backcrossing, selfing, haploid-doubling, random-intercrossing, and sib-mating steps) distorts single- and multipoint linkage estimates even with dense marker coverage. Two methods are provided for correct linkage estimation in unconventional designs: fitting a correct multigeneration model, or correcting the estimates produced by fitting a one-generation model with any conventional software. These methods also support calculation of multilocus genotype frequencies and QTL-genotype distributions and are available in software.

Introduction

Constructing a multilocus genetic map in plant species, using DNA marker phenotype data, is now routine. Favored research mating designs are those maximizing informative meioses or minimizing development time. Among these are F_2, BC_1, and F_1-doubled-haploid designs, as well as recombinant inbred lines or RILs modeled as having been selfed or sib-mated to full homozygosity (Haldane and Waddington 1931; Martin and Hospital 2006), and advanced intercross lines or AILs (Darvasi and Soller 1995; Falque 2005; Liu et al. 1996; Rockman and Kruglyak 2008; Teuscher and Broman 2007; Winkler et al. 2003). In cultivar development the need to combine genetic discovery with breeding progress may lead to the use of any of a large class of "unconventional" mating designs (henceforth UMDs), defined here as arbitrary series of backcrossing, selfing, haploid-doubling, random-intercrossing, and/or sib-mating steps. Linkage in UMDs is wrongly estimated under standard genetic recombination models, but correct models have not been elaborated in any published work known to the author.

What errors might be expected from a mismatched recombination model? An estimate r of the recombination probability θ between a pair of loci is based on the ratio of the estimate g of the number γ of meiotic crossovers between those loci to the estimate m of the number μ of recombinationally informative meioses in the mating design. These are meioses in which recombination could potentially be observed, and require heterozygosity at both loci. All meioses are recombinationally informative in F_1 backcross or selfing designs, but not in UMDs. Underestimating μ or γ will bias recombination estimates in opposite directions, spuriously and nonlinearly expanding or contracting single- and multipoint distances. Locus *order*, in contrast, should be insensitive to the choice of model. Ordering can be based solely on g and carried out via seriation, minimum-spanning-tree (Wu et al. 2008) or other algorithms (for an overview see Wu et al. (2007), p. 72), since m, however, inaccurately estimated, will have a nondecreasing relationship with μ. Our focus here is on linkage and not on map ordering.

Correct linkage and QTL analysis in UMDs as in all mating designs bear directly on decisions governing marker-assisted selection. One strategy employing UMDs for

Communicated by M. Sillanpaa.

J. C. Nelson (✉)
Department of Plant Pathology, 4024 Throckmorton Plant
Sciences Center, Kansas State University,
Manhattan, KS 66506, USA
e-mail: jcn@ksu.edu

accelerated cultivar development is “advanced backcrossing” (Tanksley and Nelson 1996), inspired by an earlier scheme of Wehrhahn and Allard (1965). Here genetic analysis is begun after one or more backcross and selfing steps, commonly in a wild × cultivated cross where novel alleles are sought. The later authors proposed limited selection, apt to cause only local distortion in linkage estimation near genes governing the selected trait. In many reported advanced-backcrossing studies (Blair et al. 2006; Huang et al. 2004; Li et al. 2005; Ramchiary et al. 2007; Septiningsih et al. 2003a; Septiningsih et al. 2003b; Stevens et al. 2007; Thomson et al. 2003), linkage analyses have been based on an early-generation model available in conventional linkage and/or QTL-analysis software. Some researchers avoid these analyses and conduct only single-marker QTL tests (Gyenis et al. 2007; Naz et al. 2008), sacrificing some genetic map information.

Random-intercrossing steps, which prolong the recombinational mixing of alleles without altering allele or genotype frequencies, might be found useful in a mapping/breeding experiment if an analytic linkage treatment were available. Sib mating might be required in dioecious or nonhermaphroditic species. Genetic analysis of UMDs has not dealt with such steps, though designs featuring only intercrossing followed by selfing to homozygosity have attracted study (Darvasi and Soller 1995; Falque 2005; Liu et al. 1996; Rockman and Kruglyak 2008).

A suitable method for constructing QTL-genotype distributions would generalize interval mapping to any mating design. Interval QTL mapping requires these distributions conditional on flanking-marker genotypes. To compute them with incompletely informative markers, Jiang and Zeng (1997) presented a Markov-chain algorithm employing matrices of genotype transition probabilities between adjacent loci, giving no general method for the construction of these matrices. Potential error in QTL position and effect estimation arising from use of incorrect conditional QTL probabilities in advanced-generation selfing designs was demonstrated by Kao and Zeng (2009). These authors used numerical transition matrices of dimension 36 to generate these three-locus *diplotype* (multilocus genotype on both chromosomes, in contrast to *haplotype*) probabilities numerically. This approach was similar to that of Fisch et al. (1996), whose purpose was the construction of QTL genotype probability distributions in generation F_y based on flanking-marker genotype data from some preceding generation F_x.

Linkage analysis as well as marker-based selection in UMDs requires prediction of multilocus genotype probabilities in advanced generations. A method based on recurrence equations for computing such probabilities in some designs when r values are given has been described (Hospital et al. 1996). But like the numerical matrix methods described above, such methods are not adapted to linkage estimation, since they would require, for each of many proposed values of r, extensive matrix multiplication for recreating haplotype probabilities starting at the F_1 generation. Purely algebraic probability expressions computed only once and thereafter evaluated as often as desired for the unknown variable would have obvious advantages.

This study was undertaken to develop, for unconventional mating designs following a diploid line cross, correct and practical methods for linkage estimation, and to assess the consequences of mismatch between genetic model and mating design.

Methods

Assumptions

We assume no extensive selection affecting marker-allele frequencies, no crossover interference, and independence of meiotic behavior from generation (invariance of θ to genetic environment), or similar consequences of these phenomena for all mating designs. We assume the application of the same mating steps to all progeny of a cross between two parents homozygous at all genetic loci.

Need of a recombination probability model for linkage estimation

Linkage estimation involves fitting a recombination model to a real data set and maximizing the likelihood of the observed segregation of marker phenotypes by an iterative process. One such is the expectation–maximization (EM) algorithm described in Lander and Green (1987), which iteratively alternates between estimating the number of crossovers and estimating θ at a pair of loci, given observed progeny marker genotypes and a two-locus segregation-probability-distribution model expressed as functions of θ. A second familiar method (Liu 1997; Wu et al. 2007) maximizes the data likelihood given the segregation model, without need of a crossover model, by finding a zero of the derivative of the log likelihood. Both of these direct methods require a model for the distribution of two-locus diplotypes, precisely what is lacking for UMDs.

It is well known that numerical genotype probabilities may be generated from a vector of F_1 probabilities by multiplication with real-valued genotype–genotype transition matrices. Using conventional *AB* notation for expressing two-locus genotypes, we may define backcrossing matrix $\boldsymbol{\Psi}_B$ (with the recurrent parent defined as homozygous for the lower-case allele at every locus), selfing matrix $\boldsymbol{\Psi}_S$, and haploid-doubling matrix $\boldsymbol{\Psi}_D$ in Tables 1, 2, 3. Defining F_1 genotype distribution vector

$\pi_{\mathrm{F}_1} = [0\,0\,0\,0\,1\,0\,0\,0\,0\,0]^{\mathrm{T}}$ (the superscript denoting the transpose), we can produce the distribution resulting from any desired sequence of B, S, or D steps. For example, for two backcrosses followed by three selfs,

$$\pi_{\mathbf{bbsss}} = \Psi_{\mathrm{S}}\Psi_{\mathrm{S}}\Psi_{\mathrm{S}}\Psi_{\mathrm{B}}\Psi_{\mathrm{B}}\pi_{\mathrm{F}_1}. \qquad (1)$$

Hitherto, this Markovian calculation has been carried out with scalar probability matrices after substitution of θ with a numerical value. But generality and speed are much increased by purely algebraic multiplication, yielding as product a vector of polynomials in θ that can then be evaluated by substitution of any desired value for the variable.

Polynomial matrices for genetic algebra

Polynomial matrices (here denoted by Greek letters) allow matrix operations on polynomial probability expressions in θ. They obey the same rules of matrix multiplication and addition as do scalar matrices (Roman letters), except in forming cell products. Each cell contains a vector of real values representing polynomial coefficients; thus $(1 - \theta)^2$ is described as [1 –2 1]. This vector follows the rules of polynomial multiplication, where the zth term of the product of $\mathbf{m}$ and $\mathbf{n}$ is calculated as $\mathbf{p}_z = \sum_{x \le N_m, y \le N_n, y=z-x} \mathbf{m}_x \mathbf{n}_y$, with the Ns denoting the numbers of terms in the vectors to be multiplied, and the length of the product vector equal to $N_m + N_n - 1$. This is only grade-school arithmetic in base θ instead of 10 and without the carry operation.

Each element of a polynomial matrix such as $\boldsymbol{\pi}$ can be numerically evaluated by multiplication with scalar vector $\boldsymbol{\varphi}_r$ containing powers of r (in the example, $\boldsymbol{\varphi}_r = [r^0\ r^1\ r^2]^{\mathrm{T}}$). This operation is expressed as $\boldsymbol{\pi} \circ^{v} \boldsymbol{\Phi}_r$, where $\boldsymbol{\Phi}_r$ is a matrix of the same dimension as $\boldsymbol{\pi}$ each of whose elements is $\boldsymbol{\varphi}_r$, $^\circ$ denotes elementwise multiplication at the matrix level, and ^{v} denotes vector, rather than polynomial, multiplication at the cell level.

Table 1 Backcross transition operator matrix $\boldsymbol{\Psi}_{\mathrm{B}}$

	aabb	*Aabb*	*AAbb*	*aaBb*	*AaBb*	*aABb*	*AABb*	*aaBB*	*AaBB*	*AABB*
aabb	1.0	0.5	0	0.5	$0.5\ (1 - \theta)$	**0.5θ**	0	0	0	0
Aabb	0	0.5	1.0	0	**0.5θ**	$0.5\ (1 - \theta)$	0.5	0	0	0
AAbb	0	0	0	0	0	0	0	0	0	0
aaBb	0	0	0	0.5	**0.5θ**	$0.5\ (1 - \theta)$	0	1.0	0.5	0
AaBb	0	0	0	0	$0.5\ (1 - \theta)$	**0.5θ**	0.5	0	0.5	1.0
aABb	0	0	0	0	0	0	0	0	0	0
AABb	0	0	0	0	0	0	0	0	0	0
aaBB	0	0	0	0	0	0	0	0	0	0
AaBB	0	0	0	0	0	0	0	0	0	0
AABB	0	0	0	0	0	0	0	0	0	0

Each entry describes the probability of occurrence of the row-labeled two-locus genotype in progeny of a backcross of a parent carrying the column-labeled genotype to a parent carrying genotype *aabb*. θ is the true meiotic crossover probability between the two loci. Boldface terms denote transitions involving one crossover

Table 2 Selfing transition operator $\boldsymbol{\Psi}_{\mathrm{S}}$

	aabb	*Aabb*	*AAbb*	*aaBb*	*AaBb*	*aABb*	*AABb*	*aaBB*	*AaBB*	*AABB*
aabb	1.0	0.25	0	0.25	$0.25 - 0.5\theta + 0.25\theta^2$	**$\underline{0.25\theta^2}$**	0	0	0	0
Aabb	0	0.5	0	0	**$0.5\theta - 0.5\theta^2$**	**$0.5\theta - 0.5\theta^2$**	0	0	0	0
AAbb	0	0.25	1.0	0	**$\underline{0.25\theta^2}$**	$0.25 - 0.5\theta + 0.25\theta^2$	0.25	0	0	0
aaBb	0	0	0	0.5	**$0.5\theta - 0.5\theta^2$**	**$0.5\theta - 0.5\theta^2$**	0	0	0	0
AaBb	0	0	0	0	$0.5 - \theta + 0.5\theta^2$	**$\underline{0.5\theta^2}$**	0	0	0	0
aABb	0	0	0	0	**$\underline{0.5\theta^2}$**	$0.5 - \theta + 0.5\theta^2$	0	0	0	0
AABb	0	0	0	0	**$0.5\theta - 0.5\theta^2$**	**$0.5\theta - 0.5\theta^2$**	0.5	0	0	0
aaBB	0	0	0	0.25	**$\underline{0.25\theta^2}$**	$0.25 - 0.5\theta + 0.25\theta^2$	0	1.0	0.25	0
AaBB	0	0	0	0	**$0.5\theta - 0.5\theta^2$**	**$0.5\theta - 0.5\theta^2$**	0	0	0.5	0
AABB	0	0	0	0	$0.25 - 0.5\theta + 0.25\theta^2$	**$\underline{0.25\theta^2}$**	0.25	0	0.25	1.0

Each entry describes the probability of occurrence of the row-labeled two-locus genotype in selfed progeny of a parent carrying the column-labeled genotype. Boldface terms denote transitions involving one or (if underlined) two crossovers

Table 3 Doubled-haploid transition operator $\mathbf{\Psi}_{\mathrm{D}}$

	aabb	*Aabb*	*AAbb*	*aaBb*	*AaBb*	*aABb*	*AABb*	*aaBB*	*AaBB*	*AABB*
aabb	1.0	0.5	0	0.5	$0.5\ (1-\theta)$	$\mathbf{0.5\theta}$	0	0	0	0
Aabb	0	0	0	0	0	0	0	0	0	0
AAbb	0	0.5	1.0	0	$\mathbf{0.5\theta}$	$0.5\ (1-\theta)$	0.5	0	0	0
aaBb	0	0	0	0	0	0	0	0	0	0
AaBb	0	0	0	0	0	0	0	0	0	0
aABb	0	0	0	0	0	0	0	0	0	0
AABb	0	0	0	0	0	0	0	0	0	0
aaBB	0	0	0	0.5	$\mathbf{0.5\theta}$	$0.5\ (1-\theta)$	0	1.0	0.5	0
AaBB	0	0	0	0	0	0	0	0	0	0
AABB	0	0	0	0	$0.5\ (1-\theta)$	$\mathbf{0.5\theta}$	0.5	0	0.5	1.0

Each entry describes the probability of occurrence of the row-labeled two-locus genotype in doubled-haploid progeny of a parent carrying the column-labeled genotype. Boldface terms denote transitions involving one crossover

Calculation for random intermating

Computation of genotype distributions following a random-intermating step employs transition matrix $\mathbf{\Delta}_{\mathrm{I}}$ (Table 4), which has been given a different symbol because, unlike $\mathbf{\Psi}$, it describes the transition from diplotype to haplotype (gametic) probabilities. Indexing generations with t, the new genotype vector is formed as

$$\pi_{t+1} = pool\left(\Delta_{\mathrm{I}}\pi_t(\Delta_{\mathrm{I}}\pi_t)^{\mathrm{T}}\right)$$

where matrix-rearrangement operation *pool* forms the sums of the six symmetrically paired off-diagonal entries of the 4×4 product, representing reciprocal crosses, and composes them, with the four on the main diagonal, into the ten rows of $\boldsymbol{\pi}_{t+1}$.

Calculation for sib mating

For computing the generation transitions for full-sib mating (which we will encode as X), we note first that each column $\mathbf{c}_g$ in the $\mathbf{\Psi}_e$ diplotype–diplotype transition matrix (where e denotes B, S, or D as above or I as shown below) describes the conditional diplotype probabilities within a full-sib family generated by the e mating operation applied to diplotype g in the previous generation (an X operation applied to the F_1 is simply replaced with the identical S). Our purpose is to generate the corresponding gth column in $\mathbf{\Psi}_{\mathrm{X}(\boldsymbol{e})}$ by intermating the sibs in all possible and reciprocal combinations and accumulating the progeny diplotypes. For any non-sib-mating step followed by any number of sib matings only one transition, composing the entire series of steps, will be applied to update the $\boldsymbol{\pi}$ state vector.

We compute $\mathbf{c}_g\mathbf{c}_g^{\mathbf{T}}$, a 10×10 matrix of polynomial probabilities that for most g will be mostly zero. Consider the 1×1 product of the ith and jth elements $\mathbf{c}_{gi}\mathbf{c}_{gj}$. The haplotype probabilities for this sib mating are given in the ith and jth columns of $\mathbf{\Psi}_{\mathrm{I}}$ (Table 4). Their outer product gives the progeny diplotype probabilities and *pool* rearranges these into a 10×1 vector $\mathbf{g}_{ij(e)(g)}$. The weighted sum

$$\sum_{i=1\ldots10, j=1\ldots10} c_i c_j \left(pool\left(\Delta_{\mathrm{I}[,i]}\Delta_{\mathrm{I}[,j]}\right)^{\mathrm{T}}\right)$$

of all of these $\mathbf{g}$ vectors gives the gth column of the one-generation sib-mating transition from a generation produced by mating operation e. The transition matrix for multiple sib-mating steps in succession is computed by recurrent application, to each column, of the two steps described: generating all nonzero pairwise probabilities for sibling diplotypes and using them as weights for summing the corresponding vectors of progeny probabilities. For sib mating following random intermating we construct

Table 4 Intercrossing transition operator $\mathbf{\Delta}_{\mathrm{I}}$

	aabb	*Aabb*	*AAbb*	*aaBb*	*AaBb*	*aABb*	*AABb*	*aaBB*	*AaBB*	*AABB*
ab	1	0.5	0	0.5	$0.5-0.5\theta$	$\mathbf{0.5\theta}$	0	0	0	0
Ab	0	0.5	1	0	$\mathbf{0.5\theta}$	$0.5-0.5\theta$	0.5	0	0	0
aB	0	0	0	0.5	$\mathbf{0.5\theta}$	$0.5-0.5\theta$	0	1	0.5	0
AB	0	0	0	0	$0.5-0.5\theta$	$\mathbf{0.5\theta}$	0.5	0	0.5	1

Each entry describes the probability of occurrence of the row-labeled two-locus gametic genotype arising from a parent carrying the column-labeled genotype. Boldface terms denote transitions involving one crossover

Table 5 The joint distribution $\boldsymbol{\Gamma}_{\mathbf{S}}$ of two-locus genotype and crossover number in the F_2 generation of a line cross

Genotype	Crossovers		
	0	1	2
aabb	$0.25 - 0.5\theta + 0.25\theta^2$	0	0
Aabb	0	$0.5 - 0.5\theta$	0
AAbb	0	0	$0.25\theta^2$
aaBb	0	$0.5 - 0.5\theta$	0
AaBb	$0.5 - \theta + 0.5\theta^2$	0	0
aABb	0	0	$0.5\theta^2$
AABb	0	$0.5 - 0.5\theta$	0
aaBB	0	0	$0.25\theta^2$
AaBB	0	$0.5 - 0.5\theta$	0
AABB	$0.25 - 0.5\theta + 0.25\underline{\theta^2}$	0	0

diplotype–diplotype transition matrix $\boldsymbol{\Psi}_{\mathbf{I}}$ whose *g*th column is computed as $pool(\boldsymbol{\Delta}_{\mathrm{I}[,g]}(\boldsymbol{\Delta}_{\mathrm{I}}\boldsymbol{\pi}_t)^{\mathrm{T}})$. This construction was skipped in the random-intermating section above because families did not need to be kept distinct.

Calculating linkage in UMDs by the correction method

To obtain an expression allowing the correction of linkage estimates made for UMDs under conventional models, we must determine the relationship between linkage estimates under the same two-locus-segregation model applied to different observed genotype segregations. To do this, we can decompose the $\boldsymbol{\pi}_{F_2}$ genotype distribution according to the numbers of crossovers represented by each of the ten possible two-locus genotypes, to form the joint diplotype and crossover-number distribution $\boldsymbol{\Gamma}_{\mathrm{S}}$ shown in Table 5. Addition of the crossover numbers scaled by their probabilities will yield γ. In matrix terms, $\gamma = \mathbf{1}\boldsymbol{\Gamma}_{\mathrm{S}}\mathbf{c}$, where $\mathbf{1}_{1\times 10}$ is a summing vector and $\mathbf{c}$ is crossover vector $[0\ 1\ 2]^{\mathrm{T}}$. Since $\gamma = 2\theta$ and $\mu = 2$, the linkage is θ under the F_2 model when the true matches the expected genotypic segregation.

Now suppose we rescale the segregation to some design *e* while preserving the F_2 genotype-conditional crossover probabilities in $\boldsymbol{\Gamma}_{\mathrm{S}}$. We can do this by constructing$\boldsymbol{\pi}_e$, forming the elementwise product $\boldsymbol{\delta} = \boldsymbol{\pi}_{\mathrm{S}}^{\mathrm{rec}} \circ \boldsymbol{\pi}_e$ where $\boldsymbol{\pi}_{\mathrm{S}}^{\mathrm{rec}}$ contains the reciprocals of the elements of $\boldsymbol{\pi}_{\mathrm{s}}$, and recomputing the expected crossover number as $\gamma_{e(\mathrm{S})} = \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{S}}\mathbf{c}$ and the $r_{e(\mathrm{S})}$ as $\gamma_{e(\mathrm{S})}/2$. This expression in θ predicts the *r* that will be estimated under the F_2 recombination model when the observed segregation is that of mating design *e*. Thus, for design **bs**, $r_{\mathrm{bs(S)}} = \theta(2 - \theta)/2$. All such $r_{e(\mathrm{S})} = f(\theta_e)$, or $r_{e(\mathrm{B})} = f(\theta_e)$ for designs ending in a backcross, are seen to be polynomials with alternating signs on the coefficients, as illustrated in Fig. 1 for selected UMDs. Though in the example $\theta_{\mathrm{bs(S)}}$ may be calculated as $f^{-1}\left(r_{\mathrm{bs(S)}}\right) = 1 - \sqrt{1 - 2r_{\mathrm{bs(S)}}}$, numerically solving for θ given *r* is easier than algebraic function inversion for converting single- to multigeneration linkage estimates.

Calculating linkage in UMDs by the direct method

Either of the direct-estimation approaches reviewed at the top of the Methods section may be used in UMDs. Only the second, for which the computation is simpler to describe because no crossover model is needed, is outlined here. The likelihood equation is $L = \prod_k^{N_p} p_k^{N_k}$ where N_p denotes the number of observed marker-phenotype classes, p_k the corresponding probability term (or sum of terms, in case of dominant markers) in $\boldsymbol{\pi}_e$, and N_k the observed frequency of the class in the data. Then $\log L = \sum_{k=1}^{N_p} N_k p_k$ and, expressing each p_k as polynomial function f_k (θ), $d(\log L)/d\theta = \sum_{k=1}^{N_p} N_k(df_k(\theta)/f_k(\theta)d\theta)$, where differentiation is also done algebraically. This expression is set to 0 and solved for θ. The only novelty here is the production of the necessary $\boldsymbol{\pi}_e$ and derivatives by purely algebraic computation.

Multipoint linkage estimation in UMDs

The likelihood of a multipoint map is maximized over a chain of ordered loci via the hidden-Markov model of Lander and Green (1987). Its implementation in the software CarthaGene (de Givry et al. 2005) computes, in the E step, *g* for each individual in each interval in turn, maximizing the likelihood over genotype and crossover number. The M step updates the *r* for each interval. The modification necessary for UMDs is maximization over crossover numbers higher than the conventional 1 or 2. This requires for mating design *e* a joint distribution $\boldsymbol{\Gamma}_e$ of two-locus genotype and crossover number, whose construction is omitted here for brevity.

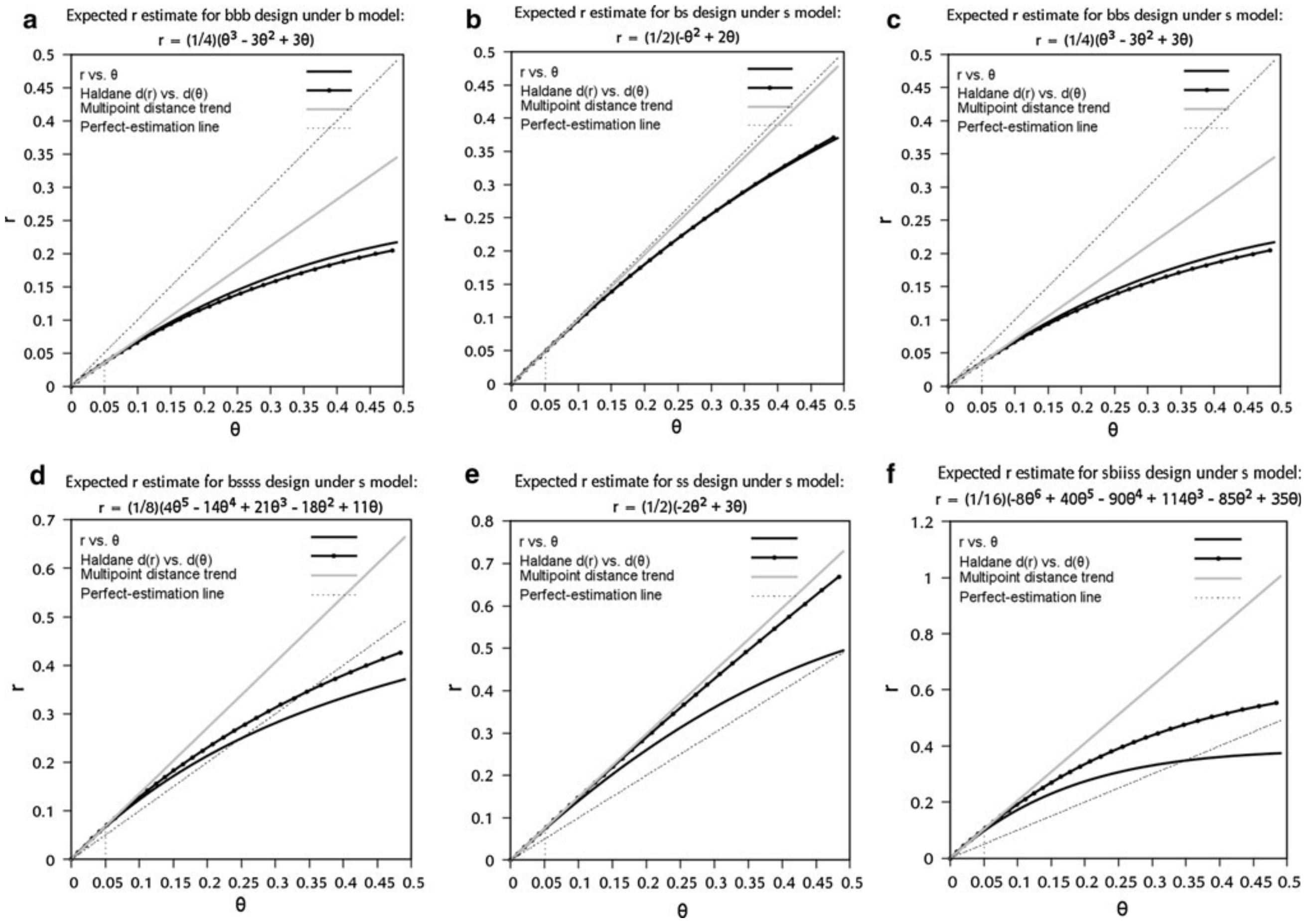

Fig. 1 Theoretical fidelity of linkage estimates made with a one-generation model in multigeneration mating designs. Horizontal axis: true crossover probability θ or Haldane distance depending on the plotted curve. *Black curve* shows the r estimate of θ (computed from the equation at top of each panel) and *black patterned curve* shows the Haldane distance corresponding to r plotted against that corresponding to θ under a BC$_1$ model (a: **bbb** design) or an F$_2$ model (b: **bs**; c: **bbs**, d: **bssss**, e: **ss**, f: **sbiiss**), where **b**, **s**, and **i** denote **b**ackcross, **s**elfing, or random **i**ntercrossing steps in a mating series. A **bb**, not shown**,** is equivalent to a **bs** design with respect to average estimation error (see text). *Dotted diagonal line* corresponds to correct estimation. *Dotted vertical segment* indicates the r estimated at the average distance between adjacent loci in simulated maps. *Solid gray diagonal* indicates the expected extension, under multipoint mapping, of the corresponding distance distortion to all intervals, whether or not between adjacent loci

Experimental methods

Software implementation

Algorithms described above were implemented in Perl (available from the author or http://coding.plantpath.ksu.edu/exotic_linkage/Exotic_linkage.html) for demonstrating and applying the correction method and displaying the correction expression and fidelity profile for any desired design. They were also implemented in CarthaGene 1.2.1 for direct single- and multipoint linkage analysis of UMDs, while QGene 4.0 (Joehanes and Nelson 2008) was adapted for population simulation and QTL mapping in UMDs.

Simulation of populations under UMDs

For assessing the effects of linkage model on multipoint map fidelity, five UMDs were simulated, of which four, **bbs**, **bs**, **bssss**, and **ss** (**S** designs) ended in a selfing and one, **bb** (**B** design) in a backcrossing step. For each design two levels (0, 20) of missing-marker percent and for **S** designs two levels (0, 20) each of maternal and paternal dominant marker percent were imposed. QGene was used to simulate 50 replicate datasets for 100 progeny for each parameter set. Each data set of a mating design was simulated from the same set of linkage relationships (“map”) among 20 markers randomly placed along a 120-cM chromosome.

Prediction of linkage estimates under UMDs

Without use of simulation, the correction method was employed to predict, for the simulated UMDs (except that the **bb** was substituted with a **bbb** design because misestimation and correction are identical for **bb** and **bs**), the r that would be estimated for a conventional model mismatched to the mating design, for all values of θ between 0 and 0.5. The **sbiiss** design was included to show the application of the calculation method to an intercrossing design not simulated in the mapping experiment.

Multipoint map construction

Two recombination models were fitted: the correct multigenerational model, and an F_2 or BC_1 model for **S** and **B** designs, respectively. Multipoint orders and distances were calculated with CarthaGene in script-driven mode. The ordering method was *mfmapd*, a seriation method based on pairwise genetic distances, followed by *flips*, a series of successive permutations of 5-adjacent-marker groups aimed at finding orders of increased likelihood and applied until improvement ceased. This fast method usually produces a map identical to those from more elaborate methods.

Measures of map fidelity

The fidelity of ordering and distance estimation with respect to the known true order and pairwise marker distances was assessed with order and distance statistics. The order statistics were the sum of absolute values of differences between the assigned and the correct rank of a marker (sum absolute rank difference, SARD) and Kendall's tau coefficient (Van Os et al. 2005). Both are nonparametric order statistics, the first taking its minimum value of 0 under perfect correspondence and the second ranging between 1 for perfect and –1 for reverse correspondence. For either, the statistic was calculated for both forward and reverse orders and the value closest to that representing perfect correspondence was taken. The distance statistic was the estimated distance between each pair of markers, averaged for each distinct value of the true distance within each parameter set.

Results

Predicted linkage distortion in UMDs under an incorrect model

Plotting the predicted r and its Haldane distance transformation against θ and its distance for six UMDs (Fig. 1) shows consistent misestimation of r when a one-generation recombination model is applied for linkage estimation. In these plots the diagonal broken line corresponds to perfect correspondence of the estimate r with the true value θ. Misestimation is seen to vary nonlinearly with θ and sometimes (Fig. 1d, f) to pass from over- to underestimation in the same design. These plots show that in UMDs, typically in only a narrow range of θ are linkage and distance estimates made under the wrong model expected to be correct, and that they may err by twofold or more.

Observed multipoint map order under an incorrect model

Because SARD and tau statistics for the multipoint maps showed similar trends, only SARD is presented. Designs with fewer generations and fewer backcrosses showed more order fidelity (Fig. 2); the **bbs** design produced the most distortion even under the correct model. Model correctness did not affect fidelity, irrespective of data incompleteness.

Observed interlocus map distance under an incorrect model

Application of an incorrect model resulted in over- or underestimation of interlocus distances (Fig. 3) comparable with the theoretical predictions of Fig. 1. The correct model approximated the true distances in all designs, with some divergence in the **bbs** and **bssss**. Incomplete data owing to dominant and/or missing marker genotypes increased the dispersion of distance estimates but not their average values (result not shown).

Discussion

Relative fidelity of map ordering among different UMDs and models

Linkage-estimation error showed, as predicted, negligible effect on ordering. Differences among designs in the frequency and magnitude of ordering errors were associated more with the information content of the design (half of mapping information being lost at any backcross step) than with its correct modeling. In practice, ordering error may arise from noisy or missing marker data, the concern of Wu et al. (2008). Their algorithm is subject to the same linkage-estimation errors in UMDs as are all others lacking a UMD model.

Single- and multipoint distance misestimation

The one-generation model overestimated distances in UMDs dominated by selfing, evidently owing mainly to underestimation of μ, while underestimating distances in

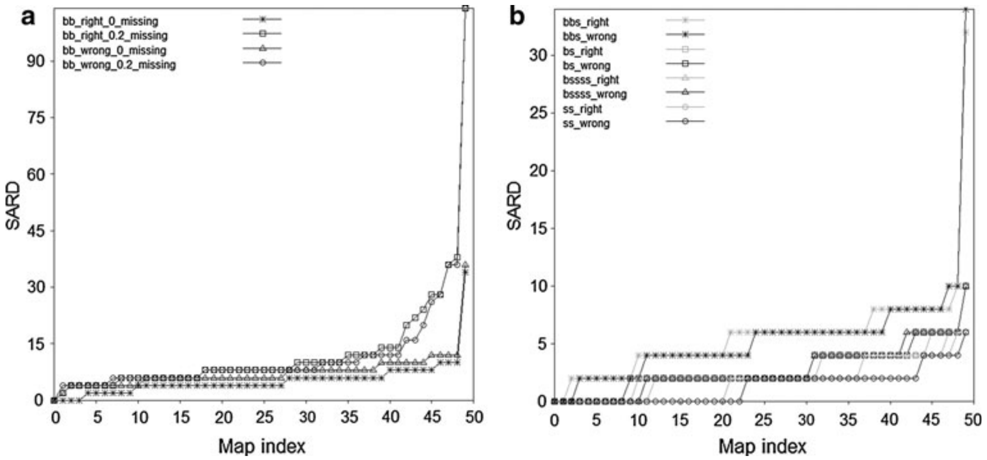

Fig. 2 Observed effects of mating design, correctness of genetic model, and completeness of simulated data on multipoint genetic map order in two sets of mating designs. SARD: sum of absolute rank differences; low values correspond to fewer order inversions and hence higher mapping fidelity. Values for 50 maps for each parameter combination are plotted in ascending order. **a** Effects of model (wrong: **b** model, right: **bb** model) and proportion (0, 0.2) of missing data on SARD order statistic in **bb** (BC_2) mating design; **b** Effects of design and genetic model (wrong: **s** model; right, model matching the respective design) in the presence of completely informative marker data (no dominance or missing markers) on SARD in **bs**, **bbs**, **bssss**, and **ss** selfing designs

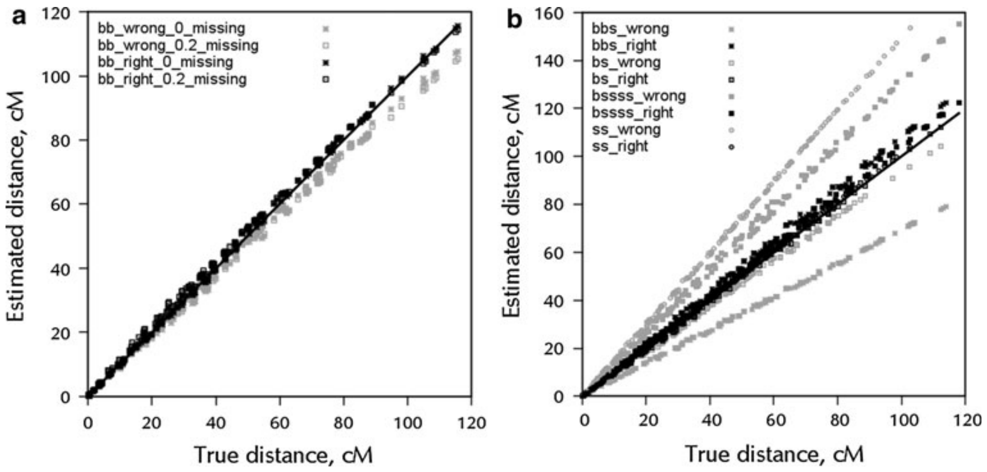

Fig. 3 Effect of mating design, correctness of genetic model, and completeness of data on multipoint genetic map average interlocus distances from simulated data in five mating designs. **a** Effects of model (wrong: **b** model, right: **bb** model) and proportion (0, 0.2) of missing data on estimated interlocus distance in **bb** (BC_2) mating design; **b** effect of model (wrong: **s** model, right: model matching the respective design) on estimated interlocus distance in **bs**, **bbs**, **bssss**, and **ss** designs. In all plots, the estimated interlocus distance, averaged for each distinct value of the true distance, is plotted against true distance for all pairs of loci, whether or not adjacent. *Diagonal line* shows the equality expected under perfect estimation. Effects and labels are as described for Fig. 2

those with more backcrossing steps owing to underestimation of γ. In longer **S** mating series the correct model still slightly overestimated distances. Two-point estimates cannot account for multiple crossovers in nonadjacent marker intervals, accumulating over meioses.

Why do map distances show a linear relationship for multipoint linkage (Fig. 2) but a nonlinear one for single linkage (Fig. 1)? The first of these figures describes *all* pairwise distances in the calculated maps. Because the multipoint algorithm maximizes linkage likelihood only for *adjacent* loci, average distance misestimation will reflect average adjacent-interval size and will accumulate linearly across multiple loci. In the simulated maps the average interval size was around 5 cM, and accordingly in Fig. 1 a line from the origin through the distance corresponding to r when the distance corresponding to θ is 0.05 resembles the trend lines in Fig. 2. But for individual two-point linkage estimates, distortion under the wrong genetic model is nonlinear, as is obvious from the polynomial correction expression.

It is close and not distant genetic linkage that is of practical interest, and Fig. 1 shows that increased marker density need not improve recombination estimates made under an inappropriate model. In most of the example designs the slope of the fidelity curve at 0 is still far from 1. Applications that depend on accurate linkage estimates, even over short intervals, should take this result into account.

Choosing estimation algorithm and software for a mating design

The linkage algorithm presented applies to conventional designs as well as UMDs. However, in some cases fast shortcuts are available. RIL designs should be handled with the Haldane and Waddington (1931) correction, improved by Martin and Hospital (2006). AILs may be treated as RILs followed by distance correction (Falque 2005). CarthaGene, besides handling all UMDs by the direct method, provides a "boosted" algorithm for fast mapping in RIL and BC_1 designs. But any linkage or mapping software offering BC_1 and F_2 recombination models may be used for linkage estimation or genetic map construction in any UMD according to a simple rule: if the design ends with a selfing or intercrossing step, use the F_2 model, and if in a backcross or haploid-doubling step, the BC_1 model. Each adjacent-marker interval in the output must then be corrected, using the resources described above or a spreadsheet-based numerical solver provided with the correction expression that may be obtained from them.

Extensions of the linkage-estimation algorithm

To accommodate other mating systems

The matrix machinery for calculating genotype probabilities could be used with advantage to simplify and accelerate the numerical computation described by other authors (Fisch et al. 1996; Kao and Zeng 2009). It could also be extended to designs more complex than line crosses in diploids, possibly accommodating more allele combinations, higher ploidy levels, and intermating among families descended from different crosses.

To compute hidden-genotype distributions

The transition matrices needed for computing QTL and other missing-genotype distributions by the algorithm of Jiang and Zeng (1997) can be obtained from $\boldsymbol{\pi}$ by conditioning on the genotype of either of the two component loci; for example, probability $\Pr(G_R = \mathrm{BB} \mid G_L = aa) = \Pr(G_{LR} = aaBB)/\Pr(G_L = aa)$.

To compute multilocus genotype distributions

The expected probability of occurrence of any linearly ordered set of L loci for which recombination probabilities $r_1, r_2, \ldots, r_{L-1}$ may be specified is readily calculated. Fast one-time construction of $\boldsymbol{\pi}$ by means of Eq. 1 is followed by substitution of the r_is to yield the unconditional genotype probability vector for each interval. Next, each two-locus probability entry in every vector but the leftmost is conditioned on the left-hand genotype term. Thus, the entries for *aabb*, *aaBb*, and *aaBB* are divided by their sum, as are the entries of the form *Aaxx* and *AAxx*. Now the probability of any given multilocus genotype may be calculated as the products of these terms, chained by the shared genotype at each locus. For example,

$$\Pr(AabbCcDD) = \Pr(Aabb)\Pr(bbCc|bb)\Pr(CcDD|Cc),$$

where the conditioning expression $|gg$ is taken to mean "given genotype gg at the left-hand locus". The calculation is easily extended to multiple chromosomes under the assumptions of independent assortment and, where applicable, additive gene action.

Conclusions and perspectives

Elegantly derived algebraic expressions given in works such as Teuscher and Broman (2007) for haplotype probabilities in specialized intermated RIL populations appear fully adequate for linkage and QTL computation in the mating designs they address, all involving mating to fixation for the production of immortal research populations. What the use of symbolic transition matrices offers, by expressing each design as the sequential application of individual steps and abstracting out the recombination parameter, is the automatic determination of haplotype probability expressions for a far greater assortment of mating designs not necessarily ending in fixation. This approach may reduce the requirement for manual derivation of estimators for novel designs. It also directly generates the expression $r_e = f(\theta)$, whose derivative, as explained by the above authors, quantifies the map "expansion" associated with analysis of UMD e as a single-generation design.

Inference over increasing numbers of mating steps, particularly those including backcrossing, can reduce information by increasing the variance of linkage estimates; all linkage-estimation methods are asymptotic at best. But while the methods developed here are intended for linkage and QTL analysis aimed at marker-assisted selection in individual crosses, unconventional mating designs also provide linkage evidence that may be of value for meta-analyses of multiple crosses. The increasing application of high-throughput marker genotyping and genomic selection may reduce, though not eliminate, the importance of biparental linkage and interval-QTL-mapping experiments.

Acknowledgments Hermann Buerstmayr posed a motivating question and provided sample data. Thomas Schiex and colleagues compiled contributed code into CarthaGene. Prof. Eberhard Weber pointed out errors and an alternative calculation for crossover expectation. The editor is thanked for bringing additional references to my attention. This is contribution 11-117-J from the Kansas Agricultural Experiment Station.

References

Blair MW, Iriarte G, Beebe S (2006) QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean x wild common bean (*Phaseolus vulgaris* L.) cross. Theor Appl Genet 112:1149–1163

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

de Givry S, Bouchez M, Chabrier P, Milan D, Schiex T (2005) CarthaGene: multipopulation integrated genetic and radiation hybrid mapping. Bioinformatics 21:1703–1704

Falque M (2005) IRILmap: linkage map distance correction for intermated recombinant inbred lines/advanced recombinant inbred strains. Bioinformatics 21:3441–3442

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

Gyenis L, Yun SJ, Smith KP, Steffenson BJ, Bossolini E, Sanguineti MC, Muehlbauer GJ (2007) Genetic architecture of quantitative trait loci associated with morphological and agronomic trait differences in a wild by cultivated barley cross. Genome 50: 714–723

Haldane JBS, Waddington CH (1931) Inbreeding and linkage. Genetics 16:357–374

Hospital F, Dillmann C, Melchinger AE (1996) A general algorithm to compute multilocus genotype frequencies under various mating systems. Comput Appl Biosci 12:455–462

Huang XQ, Kempf H, Ganal MW, Röder MS (2004) Advanced backcross QTL analysis in progenies derived from a cross between a German elite winter wheat variety and a synthetic wheat (*Triticum aestivum* L.). Theor Appl Genet 109:933–943

Jiang C, Zeng Z-B (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica 101:47–58

Joehanes R, Nelson JC (2008) QGene 4.0, an extensible Java QTL-analysis platform. Bioinformatics 24:2788–2789

Kao CH, Zeng MH (2009) A study on the mapping of quantitative trait loci in advanced populations derived from two inbred lines. Genet Res 91:85–99

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Li J, Huang XQ, Heinrichs F, Ganal MW, Röder MS (2005) Analysis of QTLs for yield, yield components, and malting quality in a BC3-DH population of spring barley. Theor Appl Genet 110:356–363

Liu BH (1997) Statistical genomics: linkage mapping and QTL analysis. CRC Press, New York

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

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

Naz A, Kunert A, Lind V, Pillen K, Léon J (2008) AB-QTL analysis in winter wheat: II. Genetic analysis of seedling and field resistance against leaf rust in a wheat advanced backcross population. Theor Appl Genet 116:1095–1104

Ramchiary N, Bisht NC, Gupta V, Mukhopadhyay A, Arumugam N, Sodhi YS, Pental D, Pradhan AK (2007) QTL analysis reveals context-dependent loci for seed glucosinolate trait in the oilseed *Brassica juncea*: importance of recurrent selection backcross scheme for the identification of 'true' QTL. Theor Appl Genet 116:77–85

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

Septiningsih EM, Prasetiyono J, Lubis E, Tai TH, Tjubaryat T, Moeljopawiro S, McCouch SR (2003a) Identification of quantitative trait loci for yield and yield components in an advanced backcross population derived from the *Oryza sativa* variety IR64 and the wild relative *O. rufipogon*. Theor Appl Genet 107:1419–1432

Septiningsih EM, Trijatmiko KR, Moeljopawiro S, McCouch SR (2003b) Identification of quantitative trait loci for grain quality in an advanced backcross population derived from the *Oryza sativa* variety IR64 and the wild relative *O. rufipogon*. Theor Appl Genet 107:1433–1441

Stevens R, Buret M, Duffe P, Garchery C, Baldet P, Rothan C, Causse M (2007) Candidate genes and quantitative trait loci affecting fruit ascorbic acid content in three tomato populations. Plant Physiol 143:1943–1953

Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor Appl Genet 92:191–203

Teuscher F, Broman KW (2007) Haplotype probabilities for multiple-strain recombinant inbred lines. Genetics 175:1267–1274

Thomson MJ, Tai TH, McClung AM, Lai XH, Hinga ME, Lobos KB, Xu Y, Martinez CP, McCouch SR (2003) Mapping quantitative trait loci for yield, yield components and morphological traits in an advanced backcross population between *Oryza rufipogon* and the *Oryza sativa* cultivar Jefferson. Theor Appl Genet 107: 479–493

Van Os H, Stam P, Visser RGF, Van Eck H (2005) RECORD: a novel method for ordering loci on a genetic linkage map. Theor Appl Genet 112:30–40

Wehrhahn C, Allard RW (1965) The detection and measurement of the effects of individual genes involved in the inheritance of a quantitative character in wheat. Genetics 51:109–119

Winkler CR, Jensen NM, Cooper M, Podlich DW, Smith OS (2003) On the determination of recombination rates in intermated recombinant inbred populations. Genetics 164:741–745

Wu R, Ma C-X, Casella G (2007) Statistical genetics of quantitative traits: linkage maps and QTL. Springer, New York

Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genet 4:e1000212