S. J. Knapp · J. L. Holloway · W. C. Bridges
B.-H. Liu

Mapping dominant markers using F_2 matings

Received: 22 March 1994 / Accepted: 11 November 1994

Abstract The development of efficient methods for amplifying random DNA sequences by the polymerase chain reaction has created the basis for mapping virtually unlimited numbers of mixed-phase dominant DNA markers in one population. Although dominant markers can be efficiently mapped using many different kinds of matings, recombination frequencies and locus orders are often mis-estimated from repulsion F_2 matings. The major problem with these matings, apart from excessive sampling errors of recombination frequency (θ) estimates, is the bias of the maximum-likelihood estimator (MLE) of θ (θ_{ML}). $\hat{\theta}_{ML} = 0$ when the observed frequency of double-recessive phenotypes is 0 and the observed frequency of double-dominant phenotypes is less than 2/3 – the bias for those samples is $-\theta$. We used simulation to estimate the mean bias of θ_{ML}. Mean bias is a function of n and θ and decreases as n increases. Valid maps of dominant markers can be built by using sub-sets of markers linked in coupling, thereby creating male and feamle coupling maps, as long as the maps are fairly dense (about 5 cM) – the sampling errors of θ increase as θ increases for coupling linkages and are equal to those for backcross matings when $\theta = 0$. The use of F_2 matings for mapping dominant markers is not necessarily proscribed because they yield twice as many useful markers as a backcross population, albeit in two maps, for the same number of DNA extractions and PCR assays; however, dominant markers can be more effeciently exploited by using doubled-haploid, recombinant-inbred, or other inbred populations.

Key words Bias · Recombination frequency · Genetic maps · DNA markers

Introduction

The development of efficient methods for assaying random DNA polymorphisms using the polymerase chain reaction has led to the widespread use of dominant DNA markers for genetic mapping (Tingey et al. 1992). Random amplified polymorphic DNAs (RAPDs), sequence characterized amplified regions (SCARs), and amplified fragment length polymorphisms (AFLPs) (Welsh and McClelland 1990; Williams et al. 1990: Rafalski and Tingey 1993; Zabeau 1993) are the major examples.

Mixed coupling- and repulsion-linkages arise within F_2 populations assayed for randomly selected dominant DNA markers because the distribution of dominant and null alleles between lines is random. A given oligonucleotide sequence usually amplifies DNA at some loci and fails to amplify DNA at other loci, thus giving rise to dominant and null alleles at different loci (Tingey et al. 1992).

Restriction fragment length polymorphism (RFLP) maps are often built using phase-known F_2 progeny. The use of F_2 progeny for building these maps poses few problems because most RFLP markers are co-dominant. Phase-known F_2 progeny can be rapidly developed in many species, and recombination frequencies (θ) between co-dominant markers are efficiently estimated from phase-known F_2 matings (Fisher 1926; Mather 1936; Ott 1991). These matings are, however, one of the least efficient for estimating recombination frequencies between dominant markers – repulsion F_2 matings are the least efficient, while coupling F_2 matings are nearly as efficient as backcross matings for tightly linked loci (Mather 1936).

Communicated by J. W. Snape

S. J. Knapp (☒) · J. L. Holloway
Department of Crop and Soil Science, Oregon State University, Corvallis, OR 97331, USA

W. C. Bridges
Department of Experimental Statistics, Clemson University, Clemson, SC29634, USA

B.-H. Liu
Department of Forestry, North Carolina State University, Raleigh, NC 27695, USA

Although the MLEs of many genetic parameters are biased (Weir 1990), these biases are usually not serious; the bias is often significantly less than the variance of the estimator. Huether and Murphy (1980) found the bias of a gene frequency estimator to be significant and proposed estimators other than the MLE. Clerget-Darpoux (1982) and Ott (1991) described "genetic biases" affecting the estimation of recombination frequencies. The biases of the MLEs of θ for different matings, however, are not known. In this paper we describe the bias of the maximum-likelihood estimator (MLE) of θ for repulsion F_2 matings.

The maximum-likelihood estimator of θ

The log of the likelihood of θ given $n\hat{f}$ for a repulsion F_2 mating is

$$\ln L(\theta|n\hat{f}) = n\hat{f}_1 \ln\left[\frac{1}{4}(2+\theta^2)\right] + n\hat{f}_2 \ln\left[\frac{1}{4}(1-\theta^2)\right]$$
$$+ n\hat{f}_3 \ln\left[\frac{1}{4}(1-\theta^2)\right] + n\hat{f}_4 \ln\left[\frac{1}{4}\theta^2\right],$$

where $\ln\ ((n\hat{f}_1 + n\hat{f}_2 + n\hat{f}_3 + n\hat{f}_4)!/(n\hat{f}_1!n\hat{f}_2!n\hat{f}_3!n\hat{f}_4!))$ is dropped from the likelihood, n is the sample size, f_1 and $\hat{f}_1$ are the expected and observed frequencies of the A_B_ phenotypes, f_2 and $\hat{f}_2$ are the expected and observed frequencies of the A_bb phenotypes, f_3 and $\hat{f}_3$ are the expected and observed frequencies of the *aaB_* phenotypes, and f_4 and $\hat{f}_4$ are the expected and observed frequencies of the *aabb* phenotype (Ott 1991) (Table 1). The MLE of θ is

$$\frac{\partial}{\partial\theta}\ln L = n\hat{f}_1\frac{\partial \ln f_1}{\partial\theta} + n\hat{f}_2\frac{\partial \ln f_2}{\partial\theta} + n\hat{f}_3\frac{\partial \ln f_3}{\partial\theta} + n\hat{f}_4\frac{\partial \ln f_4}{\partial\theta}$$
$$= n\hat{f}_1\left[\frac{2\theta}{2+\theta^2}\right] + n(\hat{f}_2 + \hat{f}_3)\left[\frac{-2\theta}{1-\theta^2}\right]$$
$$+ n\hat{f}_4\left[\frac{2}{\theta}\right] = 0 \tag{1}$$

(Mather 1956; Ott 1991) (Table 1). The four solutions of (1) are

$$\pm 2^{-1/2}\{2\hat{f}_1 - \hat{f}_2 - \hat{f}_3 - 1 \pm [8\hat{f}_4 + (1 - 2\hat{f}_1 + \hat{f}_2 + \hat{f}_3)^2]^{1/2}\}^{1/2},$$

of which

$$\hat{\theta}_{ML} = 2^{-1/2}\{2\hat{f}_1 - \hat{f}_2 - \hat{f}_3 - 1 + [8\hat{f}_4 + (1 - 2\hat{f}_1 + \hat{f}_2 + \hat{f}_3)^2]^{1/2}\}^{1/2} \tag{2}$$

is the MLE of θ; two of the solutions are imaginary numbers while another solution is outside the parameter space (Table 1). When no double-recessive phenotypes are observed ($\hat{f}_4 = 0$), (2) simplifies to

$$\hat{\theta}_{ML} = 2^{-1/2}\{(3\hat{f}_1 - 2) + [(2 - 3\hat{f}_1)^2]^{1/2}\}^{1/2}; \tag{3}$$

hence, when $\hat{f}_4 = 0$ and $\hat{f}_1 \leq 2/3$, $\hat{\theta}_{ML} = 0$, and when $\hat{f}_4 = 0$ and $\hat{f}_1 > 2/3$, $\hat{\theta}_{ML} = (3\hat{f}_1 - 2)^{1/2}$. This means that the recombination frequency estimates from samples with no double-recessive phenotypes and fewer than two-thirds double-dominant phenotypes are equal to 0.0 regardless of the number of recombinants; recombination is under estimated by $-\theta$ from these samples. This is a major source of the bias of (2); however, bias is not limited to these samples. Recombination frequencies estimated from other samples, e.g., $\hat{f}_4 = 0.01$ and $\hat{f}_2 = 0.02$, are biased up or down for different θ as shown below.

Phenotype probabilities

Double-recessive phenotypes are rare in repulsion F_2 populations for many n and θ. The probability of observing $n\hat{f}_4$ double-recessive phenotypes in an F_2 population of size n is

$$\binom{n}{n\hat{f}_4}(f_4)^{n\hat{f}_4}(1 - f_4)^{n - n\hat{f}_4},$$

Table 1 Expected (f_i) and observed ($\hat{f}_i$) phenotype frequencies and expected likelihood odds ($E[Z_i(\theta)]$ for dominant markers linked in coupling (AB/ab) or repuslion (Ab/aB) in F_2 populations where A and B are dominant alleles, a and b are recessive alleles, and θ is the recombination frequency between A and B

Phenotype	$\hat{f}_i$	Repulsion		Coupling	
		f_i	$E[Z_i(\theta)]$	f_i	$E[Z_i(\theta)]$
A_B_	$\hat{f}_1$	$f_i = \frac{1}{4}(2+\theta^2)$	$\mathrm{Log}_{10}\left[\frac{4(2+\theta^2)}{9}\right]$	$f_1 = \frac{1}{4}(3 - 2\theta + \theta^2)$	$\mathrm{Log}_{10}\left[\frac{4(3-2\theta+\theta^2)}{9}\right]$
A_bb + aaB_	$\hat{f}_2 + \hat{f}_3$	$f_2 + f_3 = \frac{1}{2}(1-\theta^2)$	$\mathrm{Log}_{10}\left[\frac{4(1-\theta^2)}{3}\right]$	$f_2 + f_3 = \frac{1}{2}(2\theta - \theta^2)$	$\mathrm{Log}_{10}\left[\frac{4(2\theta-\theta^2)}{3}\right]$
aabb	$\hat{f}_4$	$f_4 = \frac{1}{4}\theta^2$	$\mathrm{Log}_{10}[4\theta^2]$	$f_4 = \frac{1}{4}(1-\theta)^2$	$\mathrm{Log}_{10}[4(1-\theta)^2]$

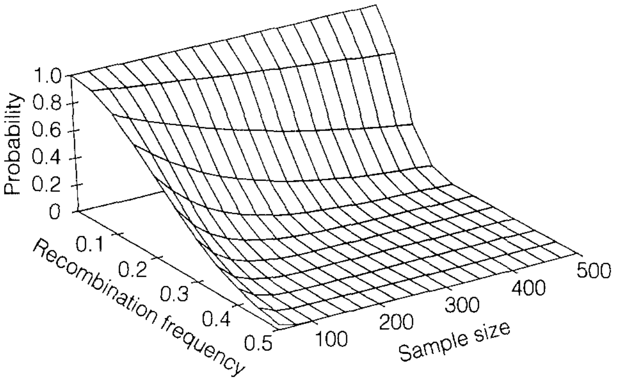


Fig. 1 Probability of failing to observe double-recessive phenotypes (aabb) in F_2 populations of size n when two loci A and B are linked in repulsion for different recombination frequencies where a and b are recessive alleles

where f_4 and $\hat{f}_4$ are the expected and observed frequencies of double-recessive phenotypes (Table 1). The probability of failing to observe double-recessive phenotypes ($n\hat{f}_4 = 0$) in a sample of n F_2 progeny is

$$Pr[n\hat{f}_4 = 0|n] = \binom{n}{0}(f_4)^0(1-f_4)^n = (1-f_4)^n; \tag{4}$$

thus, for coupling and repulsion we have $(\frac{3}{4}+\frac{1}{2}\theta-\frac{1}{4}\theta^2)^n$ and $(1-\frac{1}{4}\theta^2)^n$, respectively (Table 1). These probabilities estimate the frequency of F_2 samples lacking double-recessive phenotypes. The probability of failing to observe double-recessive phenotypes increases as n and θ decrease for repulsion matings – the probability ranges from 0.00157 to 1.0, for example, for $n = 100$ (Fig. 1). The probability of failing to observe double-recessive phenotypes decreases as sample size increases (Fig. 1) for either phase. The probability of failing to observe double-recessive phenotypes for coupling matings is negligible for most n and θ.

The probability of observing more than two-thirds double-dominant phenotypes in an F_2 population of size n is

$$\sum_{i=k}^{n}\binom{n}{k}(f_1)^k(1-f_1)^k = \sum_{i=k}^{n}\binom{n}{k}\left[\frac{1}{4}(2+\theta^2)\right]^k \times\left[1-\frac{1}{4}(2+\theta^2)\right]^{n-k},$$

where f_1 is the expected frequency of double-dominant phenotypes (Table 1), $k \geq n2/3$, and k is the smallest integer greater than $n2/3$. The probability of observing $\hat{f}_1 > 2/3$ ranges from 0.0033 to 0.0329 for $n = 50$, and from 0.0002 to 0.0109 for $n = 100$, for θ from 0.00 to 0.50, respectively; thus, samples satisfying $\hat{f}_1 > 2/3$ are much rarer than samples satisfying $\hat{f}_1 < 2/3$, and the number of $\hat{f}_4 = 0$ samples for which $\hat{\theta}_{ML} = 0$ is close to (4).

Bias of the MILE of θ

We used simulation to estimate the bias of (2). Repulsion F_2 samples were simulated for factorial combinations of $n = 50$, 100, 150, and 200, and $\theta = 0.00$ to 0.50, by 0.05. Twenty-five-thousand samples were simulated for each n and θ. A sample was simulated by drawing n random uniform numbers (0, 1) and counting the number of observations within intervals defined by the expected genotype frequencies for two loci with known θ (Table 1). A_B_, A_bb + aaB_, and aabb intervals for repulsion F_2 matings are 0 to f_1, f_1 to $f_1+f_2+f_3$, and $f_1+f_2+f_3$ to 1.0, respectively (Table 1). Mean bias was estimated by $\theta - \bar{\theta}_{ML}$, where θ is the true recombination frequency, $\bar{\theta}_{ML} = (\Sigma_{j=1}^{k}\hat{\theta}_{ML_j})/n_s$, $\hat{\theta}_{ML_j}$ is the maximum-likelihood estimate of θ from the jth simulated sample, and $n_s = 25\,000$ is the number of simulated F_2 samples for each n and θ.

We found $\theta - \bar{\theta}_{ML} \neq 0$ for most n and θ (Fig. 2); thus, (2) is a biased estimator of θ (Mood et al. 1974; Ott 1991). The mean bias of (2) is shown for different n in Fig. 2, while the sampling distribution of $\hat{\theta}_{ML}$ is shown for $n = 100$ in Fig. 3. $\bar{\theta}_{ML}$ was consistently less than θ; (2) is a downwardly biased estimator. The mean bias ranged from 0.00 to 0.08 for $n = 50$ and from 0.00 to 0.04 for $n = 200$ (Fig. 2). Mean bias decreases as sample size increases (Fig. 2).

The sampling distribution of $\hat{\theta}_{ML}$ is shown for $n = 100$ (Fig. 3) to illustrate some of the key features of (2). The $\hat{\theta}_{ML}$ distribution has a planar ridge parallel to the θ-axis at $\hat{\theta}_{ML} = 0$. This ridge is comprised of estimates from $\hat{f}_4 = 0$ samples (Fig. 3). The frequency of $\hat{\theta}_{ML} = 0$ estimates along this ridge (Fig. 1) is close to (4) (the probability of failing to observe double-recessive phenotypes).

There are three other prominent features of the sampling distribution of $\hat{\theta}_{ML}$ (Fig. 3). First, there is a major ridge of estimates with coordinates along the peak of a ridge defined by $\theta = \hat{\theta}_{ML}$ (Fig. 3). This ridge is comprised of unbiased estimates with a normal dispersion of $\hat{\theta}_{ML}$ along the $\hat{\theta}_{ML}$-axis (Fig. 3). An unbiased estimator of θ should only yield estimates along the $\theta = \hat{\theta}_{ML}$ axis. Second, there are no estimates between $\hat{\theta}_{ML} = 0.0$ and $\hat{\theta}_{ML} \cong 0.17$. Third, a ridge can be seen along a line parallel to the θ-axis between $\theta = 0.0$ and $\theta \cong 0.20$ and between $\hat{\theta}_{ML} \cong 0.17$ and $\hat{\theta}_{ML} \cong 0.22$. This ridge is comprised of estimates from $\hat{f}_4 = 0.01$ samples. A second less perceptible ridge can be seen along a line parallel to the θ-axis between $\theta = 0.0$ and $\theta \cong 0.20$ and between $\hat{\theta}_{ML} \cong 0.24$ and $\hat{\theta}_{ML} \cong 0.27$. This ridge is comprised of estimates from $\hat{f}_4 = 0.02$ samples. The $\hat{\theta}_{ML} = 0.0$ planar ridge, lack of estimates between $\hat{\theta}_{ML} = 0.0$ and $\hat{\theta}_{ML} \cong 0.17$, and major ridges parallel to the θ-axis (for some n), illustrate the bias of (2). Some of these features are a function of periodicity – the observed phenotype frequencies are discontinuous for small sample sizes which leads to the discontinuities of $\hat{\theta}_{ML}$ (Fig. 3).

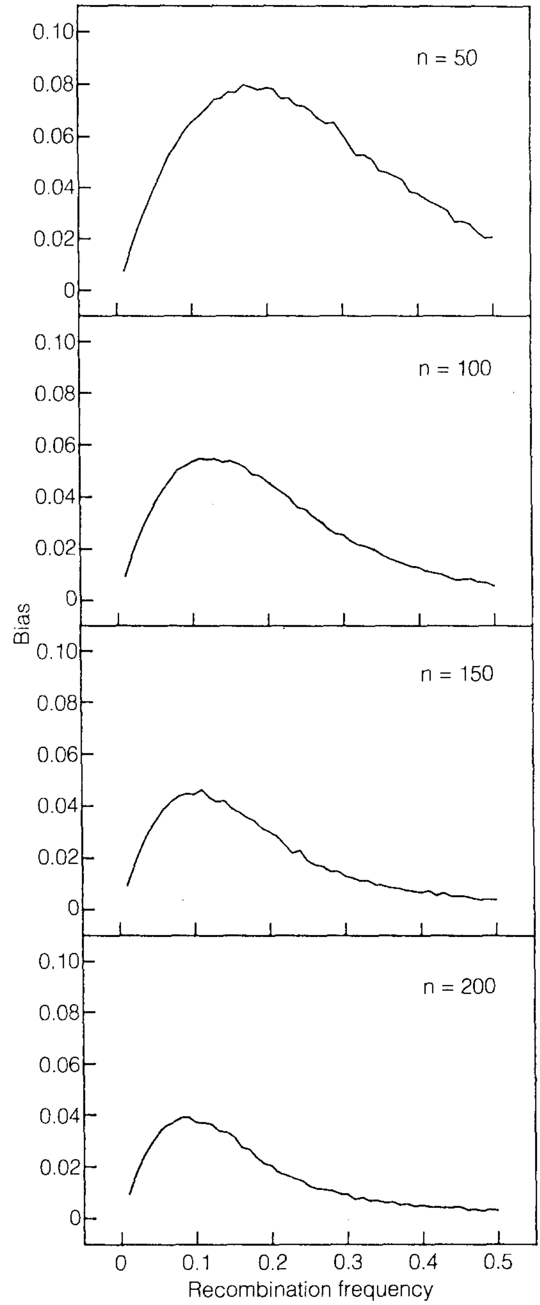


Fig. 2 Bias of $\hat{\theta}_{ML}$ estimated from simulated repulsion F_2 samples for recombination frequencies from 0.00 to 0.50 and sample sizes of 50, 100, 150, and 200 for two dominant markers linked in repulsion

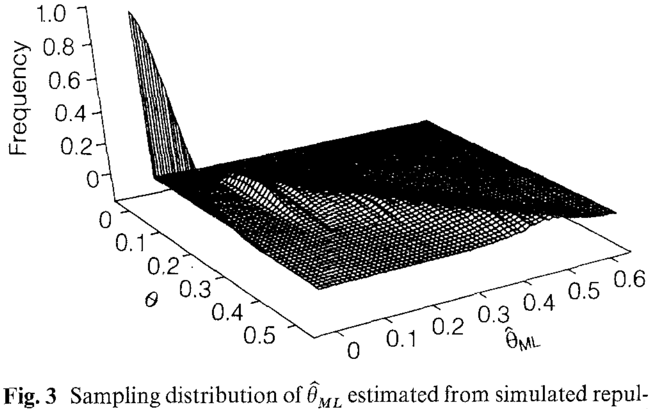


Fig. 3 Sampling distribution of $\hat{\theta}_{ML}$ estimated from simulated repulsion F_2 samples for recombination frequencies from 0.00 to 0.50 and $n = 100$ for two dominant markers linked in repulsion

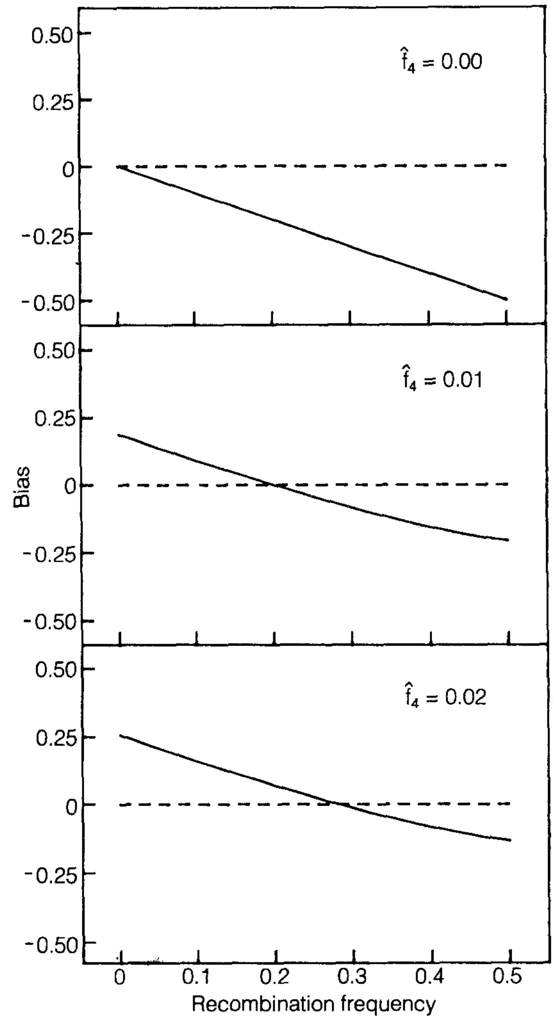


Fig. 4 Sample-wise biases of $\hat{\theta}_{ML}$ for observed double-recessive phenotype frequencies ($\hat{f}_4$) of 0.00, 0.01, and 0.02

Sample-wise bias of the MLE of θ

The mean bias ($\theta - \bar{\theta}_{ML}$) must be distinguished from the bias in one sample or the sample-wise bias. Sample-wise biases are important for understanding how bias affects the estimation of recombination frequencies and locus orders from a specific sample or population. The recombination frequency estimate for a pair of dominant markers from one sample can be unbiased or biased up or down (Figs. 2, 3). The sample-wise bias of (2) is equal to $0.0 - \theta = -\theta$ when $\hat{f}_4 = 0$ and $\hat{f}_1 < 2/3$ (Fig. 4). The range of estimates from $\hat{f}_4 = 0.01$ and $\hat{f}_4 = 0.02$ samples is narrow for many n and decreases as n decreases (Fig. 3). Substituting $\hat{f}_4 = 0.01$ and $\hat{f}_1 = 1.00 - \hat{f}_2 - \hat{f}_3 - 0.01$, and using θ to define $\hat{f}_2 = \hat{f}_3$, we found that $\hat{\theta}_{ML}$ ranged from 0.189 to 0.294 for θ from 0.00 to 0.50, respectively, and substituting $\hat{f}_4 = 0.02$ and $\hat{f}_1 = 1.00 - \hat{f}_2 - \hat{f}_3 - 0.02$, and using θ to define $\hat{f}_2 = \hat{f}_3$, we found that $\hat{\theta}_{ML}$ ranged from 0.257 to 0.366 for θ from 0.00 to 0.50, respectively; thus, θ is under- or over-estimated with these samples (Fig. 4).

Whether or not an estimate is biased can be determined by checking the frequencies of A_B_ and aabb phenotypes; however, because the bias of (2) is a function of θ, and θ is unknown, the bias of an estimate cannot be estimated from an experiment. For any sample and pair of loci, the bias of the recombination frequency estimate for a pair of loci can be significantly greater than the mean bias (Figs. 2–4). Suppose, for example, a repulsion F_2 sample is observed with $\hat{f}_1 < 2/3$ and $\hat{f}_4 = 0$, where the true recombination frequency between two loci is $\theta = 0.30$, then $\hat{\theta}_{ML} = 0, \theta$ is underestimated by 0.30, and the sample-wise bias is -0.30. The mean bias for $\theta = 0.30$ is negligible for most sample sizes (Fig. 2), but the sample-wise bias is substantial (-0.30) for those samples which satisfy $\hat{f}_4 = 0$ and $\hat{f}_1 < 2/3$. The frequency of biased estimates and mean bias decrease as n increases (Fig. 2), whereas the sample-wise bias is equal to $-\theta$ within every sample satisfying $\hat{f}_4 = 0$ and $\hat{f}_1 < 2/3$ for every n (Fig. 4).

Grouping mixed-phase dominant marker loci

Besides yielding misleading estimates of recombination, locus orders are not efficiently estimated from mixed-phase F_2 matings; however, the estimation of locus groups seldom poses a problem because a locus can be assigned to a group on the basis of one of $(k^2 - k)/2$ tests. The hypothesis of no linkage ($\theta = 0.50$) is tested against the composite hypothesis of linkage ($0.00 \leq \theta < 0.50$) using likelihood statistics (Ott 1991). The expected likelihood odds (ELOD) for a mating type can be used to estimate the sample size needed to reject the null hypothesis with a stated Type-I error probability (LOD threshold). The minimum sample size needed to reject the null hypothesis of no linkage is found by dividing the LOD threshold by the ELOD for different θ (Ott 1991). This function is maximum for some θ, and the maximum determines the sample size needed to reject a false null hypothesis.

The ELOD is estimated by

$$E[Z(\theta)] = \sum_{i=1}^{k} f_i Z_i(\theta),$$

where i indexes phenotype classes, f_i is the expected frequency of ith phenotype, k is the number of phenotypes, $Z_i(\theta) = \mathrm{Log}_{10}[(f_i(\theta))/(f_i(0.50))]$ is the ELOD for the ith phenotype, $f_i(\theta)$ is the expected frequency for the alternate hypothesis, and $f_i(0.50)$ is the expected frequency for the null hypothesis (Ott 1991). The ELOD for a double-recessive phenotype in a repulsion F_2 sample, for example, is

$$\mathrm{Log}_{10}\left[\frac{\theta^2/4}{0.5^2/4}\right] = \mathrm{Log}_{10}[4\theta^2]$$

(Table 1).

The ELOD for F_2 matings for dominant and co-dominant markers are shown in the Appendix. The minimum number of observations ($n_{\min}$) needed to reject the hypothesis of no linkage is found by dividing the LOD threshold by the ELOD for different θ (Ott 1991); thus, using a LOD threshold of 3.0, as is often done (Ott 1991), the number of repulsion F_2 progeny needed for rejecting the hypothesis of no linkage between two dominant markers is

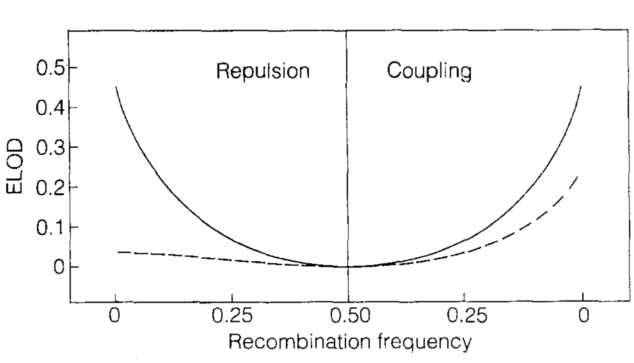
$$n_{\min} = \frac{12}{\theta^2 \mathrm{Log}_{10}[4\theta^2] + 2(1-\theta^2)\mathrm{Log}_{10}\left[\frac{4(1-\theta^2)}{3}\right] + (2+\theta^2)\mathrm{Log}_{10}\left[\frac{4(2+\theta^2)}{9}\right]}$$

(See Appendix).

The ELOD for dominant markers are phase-dependent, while those for co-dominant markers are phase-independent (Fig. 5). The ELOD for coupling-phase are greater than those for repulsion-phase matings (for dominant markers alone), and those for co-dominant markers are significantly greater than those for repulsion-phase dominant markers (Fig. 5). The ELOD for repulsion-phase dominant markers range from 0.037 to 0.0 for θ from 0.0 to 0.5, while the ELOD for repulsion-phase co-dominant markers range from 0.45 to 0.0 for θ from 0.0 to 0.5 (Fig. 5). The ELOD for repulsion-phase dominant markers is fairly flat throughout the θ range; thus, the power of this mating for detecting linkage is not much greater for two tightly linked loci than it is for two loci widely separated loci (Fig. 5). Using a LOD threshold of 3.0, $n_{\min} = 81$ for two completely linked dominant marker loci ($\theta = 0.0$), while $n_{\min} = 285$ for two dominant marker loci spaced 0.30 apart ($\theta = 0.30$).

Another way to use the ELOD is to determine the maximum θ for grouping loci for some n. Using $n = 100$ F_2 progeny and a LOD threshold of 3.0, linkage can be detected between repulsion-phase dominant markers spaced $\theta = 0.10$ or less apart, while linkage can be detected between coupling-phase dominant markers

Fig. 5 Expected likelihood odds (ELOD) for repulsion and coupling F_2 matings for dominant (*dashed*) and co-dominant (*solid*) markers



spaced $\theta = 0.25$ or less apart, or between co-dominant markers spaced $\theta = 0.30$ or less apart (Fig. 5).

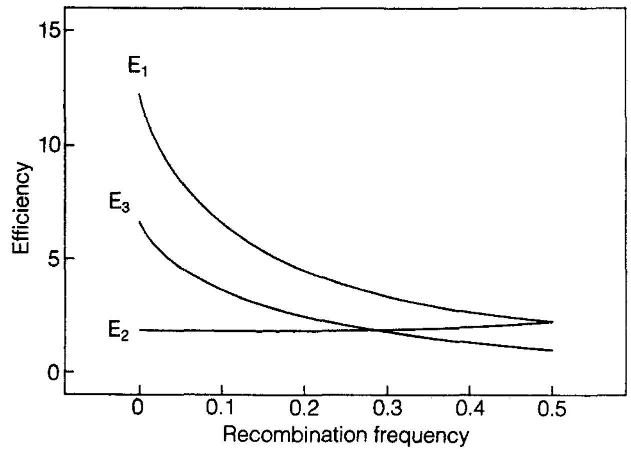
The efficiency (E) of two mating types for grouping loci is estimated by

$$E = \frac{n_{\min_1}}{n_{\min_2}} = \frac{(LOD_{\min})/(ELOD_1)}{(LOD_{\min})/(ELOD_2)} = \frac{ELOD_2}{ELOD_1}$$

(Ott 1991). Efficiencies are shown for F_2 matings for co-dominant markers (either phase) relative to repulsion-phase dominant markers (E_1) and coupling-phase dominant markers (E_2) and coupling-phase relative to repulsion-phase dominant markers (E_3) (Fig. 6). For tightly linked loci ($\theta \cong 0.00$), for example, sample sizes must be 12.0 times greater for repulsion-phase dominant markers than for either-phase co-dominant markers or 6.0 times greater for repulsion-phase than for coupling-phase dominant markers (Fig. 6). Co-dominant markers require roughly one-half as many progeny as coupling-phase dominant markers (E_2) and anywhere from one-half to one-twelfth as many progeny as repulsion-phase dominant markers (E_1) for almost any θ (Fig. 6).

Despite these efficiency differences, repulsion-phase dominant markers can often be grouped when using samples of 100 or fewer progeny because a locus is added to a group when the hypothesis of no linkage is rejected between this locus and any other locus in the group. The sample size needed to do this is determined by the most efficient mating. When coupling-phase and repulsion-phase dominant markers are dispersed throughout a linkage group, the ELOD for the former determines the minimum sample size needed for rejecting a false null hypothesis–the hypothesis of no linkage might not be rejected for repulsion-phase pairs of dominant markers assigned to a group even when their recombination frequencies are 0.0! The distribution of coupling-phase dominant markers or co-dominant markers throughout a linkage group affects the grouping of mixed-phase dominant markers.

Fig. 6 Releative efficiencies ($E_1 = (ELOD_{F_2-C})/(ELOD_{F_2-D_R})$, $E_2 = (ELOD_{F_2-C})/(ELOD_{F_2-D_C})$, and $E_3 = (ELOD_{F_2-D_C})/(ELOD_{F_2-D_R})$ of F_2 mating types for rejecting the hypothesis of no linkage



Ordering mixed-phase dominant marker loci

Ordering poses much more of a problem than grouping mixed-phase dominant markers. This is shown by using F_2 populations segregating for three linked dominant markers. With linked dominant markers A, B, and C and the locus order ABC, the mating ABC/abc is pure coupling, whereas the matings AbC/aBc, ABc/abC, and aBC/Abc are mixed coupling and repulsion. These mixed-phase matings pose a problem because two of the three linkage phases are repulsion and their recombination frequency estimates are frequently biased.

Suppose three dominant markers (A, B and C) are assayed using F_2 populations from AbC/aBc, ABc/abC, and aBC/Abc F_1 hybrids, the true recombination frequencies are $\theta_{AC} = 9.5$, $\theta_{AB} = 5.0$, and $\theta_{BC} = 5.0$, the true locus order is ABC, and no double recessive phenotypes are observed between loci linked in repulsion. The estimated recombination frequency matrices for the AbC/aBc, ABc/abC, and aBC/Abc F_2 matings are exemplified by

$$\begin{bmatrix} \cdot & \hat{\theta}_{AB} & \hat{\theta}_{AC} \\ \hat{\theta}_{AB} & \cdot & \hat{\theta}_{BC} \\ \hat{\theta}_{AC} & \hat{\theta}_{BC} & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & 0.0 & 0.095 \\ 0.0 & \cdot & 0.0 \\ 0.095 & 0.0 & \cdot \end{bmatrix},$$

$$\begin{bmatrix} \cdot & 0.05 & 0.0 \\ 0.05 & \cdot & 0.0 \\ 0.0 & 0.0 & \cdot \end{bmatrix},$$

and

$$\begin{bmatrix} \cdot & 0.0 & 0..0 \\ 0.0 & \cdot & 0.05 \\ 0.0 & 0.05 & \cdot \end{bmatrix},$$

respectively. Recombination frequency estimates for the repulsion-phase pairs of loci are biased by $-\theta$, whereas recombination frequency estimates for the coupling-phase pairs of loci are unbiased; there are two biased estimates and one unbiased estimate for each of these matings. Valid maps cannot be built from these F_2 samples. The locus orders estimated from the AbC/aBc, ABc/abC and aBC/Abc F_2 samples are ABC, ACB and BAC, respectively. The coupling pair of loci is always selected as the outer pair because the estimated recombination frequency is greatest for this pair when the sample lacks double-recessive phenotypes – the likelihood is maximum and map length is minimum for the order defined by placing the coupling pair as the outer pair of loci (Ott 1991).

The locus orders estimated from the ABc/abC and aBC/Abc matings are wrong, whereas the locus order estimated from the AbC/aBc mating matches the true order. There is, of course, no way to know this in practice. The three-locus estimate of θ_{AC} for the

AbC/aBc mating is

$$\hat{\theta}_{AC}=\hat{\theta}_{AB}+\hat{\theta}_{BC}-2\hat{\theta}_{AB}\hat{\theta}_{BC}=0.0+0.0-(2\cdot 0.0\cdot 0.0)=0.0$$

based on null interference. And the three-locus estimate of θ_{AB} for the ABc/abC mating is

$$\hat{\theta}_{AB}=\hat{\theta}_{AC}+\hat{\theta}_{CB}-2\hat{\theta}_{AC}\hat{\theta}_{CB}=0.0+0.0-(2\cdot 0.0\cdot 0.0)=0.0,$$

whereas the three-locus estimate of θ_{BC} for the aBC/Abc mating is

$$\hat{\theta}_{BC}=\hat{\theta}_{AB}+\hat{\theta}_{AC}-2\hat{\theta}_{AB}\hat{\theta}_{AC}=0.0+0.0-(2\cdot 0.0\cdot 0.0)=0.0.$$

These biased estimates imply complete linkage between A, B, and C, which, of course, is wrong. Maps cannot be estimated for any of these mixed-phase matings.

These outcomes apply to samples lacking double-recessive phenotypes, but misleading locus order estimates are not restricted to these samples. Locus orders are mis-estimated when one or two recombinants are observed between one or more locus pairs linked in repulsion. Suppose the AbC/aBc F_2 mating is used, one double-recessive phenotype is observed between A and B, and no double-recessive phenotypes are observed between B and C. The recombination frequency matrix for this sample is exemplified by

$$\begin{bmatrix} \cdot & \hat{\theta}_{AB} & \hat{\theta}_{AC} \\ \hat{\theta}_{AB} & \cdot & \hat{\theta}_{BC} \\ \hat{\theta}_{AC} & \hat{\theta}_{BC} & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & 0.19 & 0.095 \\ 0.19 & \cdot & 0.0 \\ 0.095 & 0.0 & \cdot \end{bmatrix}.$$

The estimated locus order for this example is ACB (the recombination frequency between A and B is greater than between A and C or B and C), and the estimate of θ_{AB} is much greater from this sample than from a sample lacking double-recessive homozygotes (0.19 as opposed to 0.0). The recombination frequency between A and B is overestimated by 0.14 (Fig. 4), the estimated locus order is wrong, and the two-locus and three-locus estimates of θ_{AB} disagree (the two-locus estimate is $\hat{\theta}_{AB}=0.19$, whereas the three-locus estimate is $\hat{\theta}_{AB}=0.095$). Many other examples can be concocted to show how locus ordering is affected by the bias of (2). The basic features of the problem are nonetheless clear from the examples shown.

Discussion

The bias of (2) is a significant source of error in the estimation of θ from finite samples (Figs. 2–4). The "inefficiency" of repulsion F_2 matings for estimating θ has been noted by many authors (Mather 1936, 1951; Allard 1956; Tingey et al. 1992; Rafalski and Tingey 1993) who ascribed the problem to the variance of (2); however, the problem is greater than can be ascribed to the variance alone. The mean squared error (bias^2 + variance), rather than the variance, defines the estimation errors of (2).

The locus order which minimizes map length (yields the minimum distance map) is usually the most probable locus order (Olson and Boehnke 1990; Ott 1991); however, this only holds for mating types which yield unbiased estimates of θ. This rule breaks down with matings which yield recombination frequency matrices comprised of biased and unbiased estimates. Observed phenotype numbers can be inspected to determine which θ estimates are biased, but this does not ameliorate the problem or lead to superior recombination-frequency or locus-order estimates.

Widely used methods for defining unbiased or less-biased estimators (Heuther and Murphy 1980; Efron 1982) fail when applied to the problem of estimating θ because the bias of (2) is a function of the unknown parameter θ. Jackknifing and bootstrapping, for example, can often be used to estimate and correct for bias (Efron 1982). Delete-one jackknifing, however, yields n estimates of θ equal to 0.0 when the jackknife samples are drawn from an original sample of n repulsion F_2 progeny lacking double-recessive phenotypes. The outcome is no different for bootstrapping.

Most samples sizes are insufficient for mapping mixed-phase dominant markers. The mean squared error of (2) decreases as sample size increases; however, increasing sample size is not cost efficient and seldom ameliorates the bias problem. When assaying 500 F_2 progeny, for example, at least 73.1% and 28.6% of the recombination-frequency estimates are biased for loci spaced 0.05 and 0.10 recombination-units apart (Fig. 1); so even though the number of biased estimates decreases as sample size increases, this number can still be substantial for closely linked loci.

F_2 matings are not necessarily proscribed for mapping dominant markers. They can be exploited by building separate 'pure-coupling maps'. Dominant markers can be split into two groups and mapped, one with dominant alleles from the male and one with dominant alleles from the female, thereby creating two pure-coupling F_2 maps. Building separate coupling-phase F_2 maps should yield twice as many markers as a backcross population for the same number of DNA extractions and PCR assays. The number of useful markers in a backcross population is equal to the number of recessive alleles in the backcross parent, and this number is equal to the number of markers in either the male or female pure-coupling F_2 map. The number of useful markers is equal to the number of markers in the male coupling F_2 map when the male of the F_1 is used as the backcross parent, whereas the number of useful markers is equal to the number of markers in the female coupling F_2 map when the female of the F_1 is used as the backross parent. Thus, one F_2 population should yield twice as many useful dominant markers as one backcross population for the same number of DNA extractions and PCR assays when a randomly selected set of dominant DNA markers is used.

Two coupling F_2 maps can be built at half the cost of two backcross maps. Coupling F_2 matings are only as efficient as backcross matings for closely linked loci (Mather 1936; Ott 1991). The validity of a coupling map is affected by mean map density – sampling errors for coupling F_2 maps are equal to those for backcross maps when $\theta = 0$, but increase as θ increases, and are significantly greater than those for backcross progeny for many n and θ (Mather 1936). The pitfalls of coupling F_2 matings can be minimized by using sufficiently stringent θ and LOD thresholds to estimate groups. This might mean estimating more groups than there are chromosomes.

RAPDs (Welsh and McClelland 1990; Williams et al. 1990) and AFLPs (Zabeau 1993) create the basis for quickly and efficiently assaying hundreds of dominant DNA markers. There are many ways to exploit the power of RAPDs and AFLPs which are not affected by dominance, e.g., marker-assisted backcrossing breeding. These markers are most efficiently exploited for mapping by using doubled-haploid, recombinant-inbred, or other inbred line populations (Tingey et al. 1992; Rafalski and Tingey 1993). Linkage statistics for dominant and co-dominant markers for either linkage phase are estimated with equal efficiency in these populations, and they exploit the maximum number of markers (they yield the same number of useful markers as F_2 populations).

Appendix

Expected likelihood odds (ELOD) for F_2 matings

The ELOD for repulsion or coupling F_2 matings for co-dominant markers is

$$\begin{aligned} ELOD_{F_2-C} &= \tfrac{1}{2}(1-\theta)^2 \mathrm{Log}_{10}[4(1-\theta)^2] \\ &\quad + 2(1-\theta)\theta \mathrm{Log}_{10}[4(1-\theta)\theta] + \tfrac{1}{2}\theta^2 \mathrm{Log}_{10}[4\theta^2] \\ &\quad + \tfrac{1}{2}[(1-\theta)^2 + \theta^2] \mathrm{Log}_{10}[2(1-\theta)^2 + 2\theta^2] \end{aligned}$$

as shown by Ott (1991). The ELOD for repulsion and coupling F_2 matings for dominant markers are

$$\begin{aligned} ELOD_{F_2-D_R} &= \tfrac{1}{4}\theta^2 \mathrm{Log}_{10}[4\theta^2] + \tfrac{1}{2}(1-\theta^2] \mathrm{Log}_{10}\left[\frac{4(1-\theta^2)}{3}\right] \\ &\quad + \tfrac{1}{4}(2+\theta^2) \mathrm{Log}_{10}\left[\frac{4(2+\theta^2)}{9}\right] \end{aligned}$$

and

$$\begin{aligned} ELOD_{F_2-D_C} &= \tfrac{1}{4}(3-2\theta+\theta^2) \mathrm{Log}_{10}\left[\frac{4(3-2\theta+\theta^2)}{9}\right] \\ &\quad + \tfrac{1}{2}(2\theta-\theta^2) \mathrm{Log}_{10}\left[\frac{4(2\theta-\theta^2)}{3}\right] \\ &\quad + \tfrac{1}{4}(1-\theta)^2 \mathrm{Log}_{10}[4(1-\theta)^2], \end{aligned}$$

respectively (Table 1).

Acknowledgements This research was supported by a grant from the USDA (#90-43213-5190) to S.J.K. Oregon Agricultural Experiment Station Technical Paper No. 9833.

References

Allard RW (1956) Formulas and tables to facilitate the calculation of recombination values in heredity. Hilgardia 24:235–278

Clerget-Darpoux F (1982) Bias of the estimated recombination fraction and lod score due to an association between a disease gene and a marker gene. Ann Hum Genet 46:363–372

Efron B (1982) The jackknife, the bootstrap, and other resampling plans. Soc Indus Appl Math, Philadelphia

Fisher RA (1925) Theory of statistical estimation. Proc Camb Philos Soc 22:700–725

Huether CA, Murphy EA (1980) Reduction of bias in estimating the frequency of recessive genes. Am J Hum Genet 32:212–222

Mather K (1936) Types of linkage and their value. Ann Eugen 7:251–264

Mather K (1951) The measurement of linkage in heredity. Methuen, London

Mood AM, Graybill FA, Boes DC (1974) Introduction to the theory of statistics. McGraw-Hill, New York

Olson JM, Boehnke M (1990) Monte Carlo comparison of preliminary methods for ordering multiple genetic loci. Am J Hum Genet 47:470–482

Ott J (1991) Analysis of human genetic linkage. Johns Hopkins University Press, Baltimore, Maryland

Rafalski JA, Tingey SV (1993) Genetic diagnostics in plant breeding: RAPDs, microsattelites, and machines. Trends Genet 9:275–280

Tingey SV, Rafalski JA, Williams JGK (1992) Genetic analysis with RAPD markers. In: Proc Pl Breed Symp, CSSA-ASHS, Applications of RAPD technology to plant breeding. CSSA, Madison, Wisconsin, pp 1–8

Weir BS (1990) Genetic data analysis. Sinauer Assoc, Sunderland, Massachusetts

Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18:7213–7218

Williams JGK, Kublelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535

Zabeau M (1993) Selective restriction fragment length amplification: a general method for DNA fingerprinting. Europ Patent Applic No 92402629.7