The Robustness of Recombination Frequency Estimates in Intercrosses With Dominant Markers

Torbjörn Säll and Nils-Otto Nilsson

Department of Genetics, Lund University, S-223 62 Lund, Sweden

Manuscript received September 23, 1993

Accepted for publication February 8, 1994

ABSTRACT

The robustness of the maximum likelihood estimates of recombination frequencies has been investigated in double intercrosses with complete dominance at both loci. The robustness was investigated with respect to bias in the recombination frequency estimates due to: (1) limited sample sizes, (2) heterogeneity in recombination frequencies between sexes or among meioses and (3) factors that distort the segregation—misclassification or differential viability. In the coupling phase, the recombination frequency estimates are quite robust with respect to most of the investigated factors. Potentially, the most serious cause of a bias is misclassifications, which tend to increase the recombination frequency estimates. In the repulsion phase, misclassifications are particularly serious, leading to extreme discrepancies between true and observed values. In addition, limited sample size and sex differences in recombination can also bias recombination frequency estimates in repulsion. These effects may pose serious problem in genetic mapping with random amplified polymorphic DNA (RAPD) markers.

THE estimation of recombination frequencies is a crucial step in genetic mapping. The mathematical treatment of linkage was developed during the first half of this century, and a considerable body of theory concerning recombination frequency estimates for different types of data has been developed (MATHER 1938; ALLARD 1956; BAILEY 1961). During this early phase of genetic mapping and recombination studies, the markers used were normally morphological, and offspring sample sizes were usually large. Many hundreds to thousands of offspring were usually studied. The major problem was to construct parental stocks that could yield information on more than a few loci in a specific cross. This problem has been overcome lately by the use of DNA markers in genetic analyses and mapping. The major benefits of these new markers are their general availability and their ability to reveal suitable levels of genetic variation without the need to construct parental stocks. On the other hand, mapping projects using many DNA markers are usually limited to the analysis of only a few hundred offspring.

One important DNA marker type is the restriction fragment length polymorphism, or RFLP (BOTSTEIN *et al.* 1980). RFLP markers have been used successfully in genetic mapping for several reasons. First, the detection of RFLPs is a well established technique which has proved to be experimentally robust. In addition, RFLP markers are usually codominant and, thus, can be useful in most types of crosses, particularly in intercrosses where both homologs in the individuals of the segregating offspring are informative (TANKSLEY *et al.* 1989). The usefulness of RFLP markers in mapping has led also to a renewed interest in the estimation of recom-

bination frequencies (ZHAO *et al.* 1990; RITTER *et al.* 1990; OTT 1991).

Recently, a new class of DNA markers has come into use—the random amplified polymorphic DNA markers, or RAPDs (WILLIAMS *et al.* 1990). The benefit of the RAPD technique is that it is based on PCR amplifications and, therefore, requires only a small amount of DNA. In addition, the assay is very fast and can be automated (RAFALSKI *et al.* 1991). One drawback with RAPD markers, however, is that they are experimentally less robust and, hence, RAPD data are susceptible to errors (WEEDEN *et al.* 1992). The fact that RAPDs are usually dominant also influences their usefulness in mapping projects. In diploids, the ideal situation would be to analyze backcross progeny, recombinant inbred lines or doubled haploid offspring, where dominant alleles are as useful as codominant alleles (WILLIAMS *et al.* 1991; REITER *et al.* 1992). In intercrosses, however, dominant markers are less informative because, under dominance, one cannot always deduce the offspring genotypes. Moreover, a dihybrid cross with complete dominance on both loci can be made with the dominant alleles either in coupling or repulsion. Crosses in the coupling phase are considerably more informative than crosses in the repulsion phase, because the variance of the estimate is smaller in coupling phase, especially for small recombination frequencies (ALLARD 1956; WILLIAMS *et al.* 1993). Because of the large variance in the repulsion phase, RAPD mapping based on intercrosses is usually accomplished by mapping the dominant markers from either parent separately, to create two different maps (TINGEY *et al.* 1992, GRATAPAGLIA *et al.* 1992). By using

TABLE 1

Expected proportions and observed numbers of the four phenotype classes in double intercrossoes with complete dominance

	<i>AB</i>	<i>Ab</i>	<i>aB</i>	<i>ab</i>	Total
Expected proportion	$\frac{2+\theta}{4}$	$\frac{1-\theta}{4}$	$\frac{1-\theta}{4}$	$\frac{\theta}{4}$	1
Observed number	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>n</i>

For a cross in the coupling phase $\theta = (1-r)^2$ and for a cross in the repulsion phase $\theta = r^2$, where *r* is the frequency of recombination between the two loci.

a scaffold of codominant markers, the two maps can then be aligned into a final map.

Statistical errors, quantified by the variance, are important to consider in linkage studies, as described above. There are, however, other causes of errors in recombination estimates. In the present paper, we investigate in detail the properties of recombination estimators under complete dominance in intercrossoes. Our main objective is to determine which type of errors can seriously disturb the estimate and which are less important. First, we investigate the maximum likelihood estimate of the recombination frequency *per se*, to assess the size of the bias in expectation with limited sample sizes. Then we investigate the effect of departures from the assumptions underlying the recombination frequency estimators. The assumptions have been divided into those concerning the homogeneity of recombination frequencies and those concerning factors that may distort the segregation ratio.

MODEL

In the following we will consider two linked loci, *A* and *B*. The recombination frequency between the two loci is *r*. At each locus two alleles occur where one allele is completely dominant. The dominant alleles are symbolized by capital letters, and the recessives are symbolized by small letters (*A/a* and *B/b* for the two loci, respectively). In a RAPD situation a band at a particular position corresponds to the dominant allele, whereas the absence of a band corresponds to the recessive allele. As pointed out above, we consider only intercrossoes, *i.e.*, when the *F*₂ generation is produced by selfing heterozygous *F*₁ individuals or by crossing different *F*₁ individuals with each other. The arrangement of alleles in the heterozygous *F*₁ generation can be either in the coupling phase, *AB/ab*, or in the repulsion phase, *Ab/aB*, depending on the allelic composition of the parents. Below, we assume that phase is known and, therefore, we do not place an upper limit on *r* at 0.5, which would be necessary if phase were unknown.

The *F*₂ generation segregates into four phenotypes, *AB*, *Ab*, *aB* and *ab*. Given that Mendelian segregation occurs at each of the two loci, the expected proportions of the phenotypes are shown in Table 1, using the no-

tations of BAILEY (1961). Here $\theta = (1-r)^2$ if the *F*₁ individuals that gave rise to the *F*₂ generation were in the coupling phase, and $\theta = r^2$ if they were in the repulsion phase. The use of the parameter θ thus makes the representation in Table 1 independent of the linkage phase. The maximum likelihood estimate, $\hat{\theta}$, of θ is

$$\hat{\theta} = \frac{a - 2b - 2c - d}{2n} + \sqrt{\left(\frac{a - 2b - 2c - d}{2n}\right)^2 + \frac{2d}{n}} \quad (1)$$

for both coupling and repulsion. The recombination frequency estimate, $\hat{r}$, of the true recombination frequency, *r*, is consequently calculated as

$$\hat{r} = 1 - \sqrt{\hat{\theta}} \quad (2)$$

in the coupling phase, and

$$\hat{r} = \sqrt{\hat{\theta}} \quad (3)$$

in the repulsion phase (BAILEY 1961). In the following we will consider the properties of these two recombination frequency estimators under different assumptions.

RESULTS

Properties of the estimators with limited sample sizes

Given a fixed offspring size, the numbers of observed individuals in the four classes in the *F*₂ generation are described by a multinomial distribution with the respective probabilities given in Table 1. The relative proportion of a specific class is simply the number of observed individuals in the class divided by the total number of offspring. Such an estimate is unbiased irrespective of the number of observed offspring. Estimators that involve nonlinear functions of the observations, however, are often biased (OTT 1991). The estimators (1), (2) and (3) are typical cases of such functions. Because they are maximum likelihood estimators they are asymptotically unbiased, *i.e.*, as the sample size grows, the bias of the estimators becomes smaller and smaller (BAILEY 1961), but for small sample sizes they are biased.

The expectation of any function *g*(*x*) of a discrete random variable *x* (or vector of random variables) is given by

$$E(g(x)) = \sum_x g(x)p(x), \quad (4)$$

where *p*(*x*) is the probability function of *x*. The expectation of $\hat{\theta}$ will then be

$$E(\hat{\theta}) = \sum_{a,b,c} \hat{\theta} \frac{n!}{a! b! c! d!} \left(\frac{2+\theta}{4}\right)^a \left(\frac{1-\theta}{4}\right)^{b+c} \left(\frac{\theta}{4}\right)^d, \quad (5)$$

FIGURE 1.—The bias, ϵ , in the recombination frequency estimate as a function of sample size, n , for different recombination frequencies, r . The bias is calculated as $\epsilon = E(\hat{r}) - r$ with crosses in the coupling phase.

where the sample size, $n = a + b + c + d$ and $\hat{\theta}$, is given by Equation 1.

Because a multinomial distribution has a finite number of outcomes at a fixed sample size, it is possible to calculate numerically the expected values for this estimator. The expectation for $\hat{r}$ is furthermore achieved by substituting $\hat{\theta}$ in the sum with the respective functions of $\hat{\theta}$ given in (2) and (3). We calculated the expected value of $\hat{r}$ in coupling and repulsion for sample sizes up to 50. For larger sample sizes the expected value was calculated through simulation. We then calculated the bias of the estimates, ϵ , as $\epsilon = E(\hat{r}) - r$.

The results are shown in Figures 1 and 2; observe that the bias axes have different scales in the two figures. The bias shows a marked difference between the coupling and the repulsion phase. For the coupling phase, the bias is very small for sample sizes above 20. The repulsion phase exhibits a complex pattern. At low values of n the bias is positive; it then becomes negative, reaches a minimum and as n increases it approaches 0. For small and moderate sample sizes, the bias in the repulsion phase is very strong. For example, at a recombination frequency of 0.1, the expected value of $\hat{r}$ is 0.0475 at a sample size of 100. Furthermore, simulations show that at a sample size of 800 the expected value is 0.0891, and at 1200 the expectation is 0.0939. A surprising feature of the curves in Figures 1 and 2 is that the bias fluctuates at small sample sizes rather than decreasing monotonically with the sample size. Similar fluctuations were observed by BOLLING and MURPHY (1979), who investigated the bias in backcrosses with unknown phase.

We also compared the bias of the recombination estimate with its mean square error (MSE) by calculating the ratio $\epsilon^2/(\epsilon^2 + V(\hat{r}))$, where $\epsilon^2 + V(\hat{r})$ is the MSE of $\hat{r}$, and $V(\hat{r})$ is the variance of $\hat{r}$ given by $E(\hat{r}^2) - E(\hat{r})^2$. In coupling phase, the bias is negligible relative to the variance of the estimates. In repulsion, the bias constitutes a considerable proportion of the MSE, particularly at small recombination frequencies. For example, at a

FIGURE 2.—The bias, ϵ , in the recombination frequency estimate as a function of sample size, n , for different recombination frequencies, r . The bias is calculated as $\epsilon = E(\hat{r}) - r$ with crosses in the repulsion phase.

sample size of $n = 100$, the ratio is 0.379 for $r = 0.05$ and 0.439 for $r = 0.01$.

To compensate for the bias we tried the jackknife method, a general method to reduce bias in estimates (see, *e.g.*, WEIR 1990). However, the jackknife method did not function satisfactory for this purpose. For low recombination frequencies of 1 or 2%, the jackknife estimates have an even higher bias than the original estimate—also for sample sizes above 100. We believe that the failure of the jackknife method is attributable to the strong fluctuations of the bias as seen in Figure 2. These fluctuations can be explained by the shifting signs of the different terms in the Taylor's series expansion of $E(\hat{r})$. Because the expected value, $E(\hat{r})$, has a periodicity of three, we also tried a different jackknife method with $E_{\text{jack}}(\hat{r}) = \frac{1}{3}[nE_n(\hat{r}) - (n-3)E_{n-3}(\hat{r})]$. This method turned out to be only slightly better than the traditional jackknife, $E_{\text{jack}}(\hat{r}) = nE_n(\hat{r}) - (n-1)E_{n-1}(\hat{r})$. However, for small recombination frequencies and limited sample sizes, this method also appears unsatisfactory.

The bias described above is an inherent property of the recombination estimator that will occur even if the assumptions underlying the estimates in Table 1 are met. This bias occurs with limited samples sizes, but as n grows large the expected value of the estimate will approach r , *i.e.*, the asymptotic expectation of the estimate, $AsE(\hat{r}) = r$. If any of the underlying assumptions is violated, additional bias will influence the results. This means that the asymptotic expectation of the estimate will not approach r , but will approach some other value.

Heterogeneity of recombination frequencies

Heterogeneity of recombination frequencies is one possible violation of the basic assumptions that can have an effect on the asymptotic expectation of the estimates. Two cases of such heterogeneity are investigated. The first involves differences between the sexes. The second involves differences in recombination frequency among different meioses in the F_1 generation.

In either of the two types of heterogeneity, a complete description would present estimates of all the separate recombination frequencies. In most cases, however, the geneticist is unaware of the heterogeneity and will use the data to estimate only one recombination frequency. The expectation of the estimate can be compared with the "true" recombination frequencies in different ways. We have chosen to compare the arithmetic mean, $\bar{r}$, of the two underlying recombination frequencies, with the expectation of the estimate. Thus, in this section bias is defined as $\epsilon = AsE(\hat{r}) - \bar{r}$. The arithmetic mean is not only the most natural function of the recombination values to use, it can also be shown that, given random mating, the rate of breakdown of linkage disequilibria, in the case of varying recombination frequencies, will be equal to a situation where all recombination frequencies take the value of the arithmetic mean.

Different recombination frequencies in male and female: To model a situation where recombination frequencies in the two sexes are different, we assume these to be r and kr , respectively. The k value can vary between 0 and 1, where $k = 0$ means that one of the sexes has no recombination at all, and $k = 1$ means no sex difference. It is unimportant which sex has the higher recombination frequency. The value of θ , as it appears in the model in Table 1, will then be $\theta = (1 - r)(1 - kr)$ in the coupling phase and $\theta = kr^2$ in the repulsion phase. Thus, Equation 1 is still the maximum likelihood estimator of θ . If (3) and (4) are used to calculate the estimates of the recombination frequencies, however, these will have the asymptotic expectations

$$AsE(\hat{r}) = 1 - \sqrt{(1 - r)(1 - kr)} \quad (6)$$

in the coupling phase, and

$$AsE(\hat{r}) = r\sqrt{k} \quad (7)$$

in the repulsion phase, respectively. These expectations are compared to the arithmetic mean of the recombination frequencies, $\bar{r} = (r + kr)/2$, to give the bias, $\epsilon = AsE(\hat{r}) - \bar{r}$.

For the coupling phase, ϵ is relatively small for all values of r and kr . The most extreme difference that can occur is when $r = 0.5$ and $kr = 0$. In this case $\epsilon = 0.043$, with $\bar{r} = 0.25$ and $AsE(\hat{r}) = 0.293$, *i.e.*, a difference of 17% of the mean value. For smaller values of r , the difference is smaller in both absolute and relative terms. For example, if $r = 0.1$ and $kr = 0$, then $\bar{r} = 0.05$ and $\epsilon = 0.0013$ or 3% of the mean (Figure 3). Given a fixed value of $\bar{r}$, the difference decreases for smaller differences between the sexes ($k > 0$).

For the repulsion phase, however, the effect is more pronounced, as seen in Figure 3. The bias, ϵ , grows to a large negative value when k is small. In particular, if $k = 0$ the estimate of the recombination frequency will be 0 irrespective of the recombination frequency in the sex with recombination. The case of no recombination

FIGURE 3.—The asymptotic bias, ϵ , as a function of the degree of sex difference in recombination frequencies, $1 - k$. The bias is calculated as $\epsilon = AsE(\hat{r}) - \bar{r}$, where $\bar{r}$ is the arithmetic mean of the sexes. Solid curves; repulsion phase. Dotted curve; coupling phase.

in one sex is, of course, unusual, but the example illustrates an undesirable property of the estimator.

Different recombination frequencies among meioses: Now consider a case in which the F_1 individuals that are intercrossed are heterogeneous in such a way that a proportion p of all meioses has the recombination frequency r , whereas $1 - p$ has kr . In this case we also assume $0 \leq k \leq 1$ so that r is the highest occurring recombination frequency. The arithmetic mean will then be $\bar{r} = pr + (1 - p)kr$. Furthermore, it is assumed that the gametes that produce a zygote always come from the same class of meiosis. It can be shown that $\theta = p\theta_1 + (1 - p)\theta_2$, where $\theta_1 = (1 - r)^2$ and $\theta_2 = (1 - kr)^2$ in the coupling phase, and $\theta_1 = r^2$ and $\theta_2 = (kr)^2$ in the repulsion phase. Thus, Equation 1 in this case is also the maximum likelihood estimate of θ . When the recombination frequency is estimated through (3) and (4), the asymptotic expectations are

$$AsE(\hat{r}) = 1 - \sqrt{p(1 - r)^2 + (1 - p)(1 - kr)^2} \quad (8)$$

in the coupling phase and

$$AsE(\hat{r}) = r\sqrt{p + (1 - p)k^2} \quad (9)$$

in the repulsion phase, respectively. In this case not only k , but also p , influences the bias. The strongest effect is again reached when $k = 0$, *i.e.*, when one of the recombination frequencies is zero. Given that $k = 0$ and a specific value of r , the maximum difference in the coupling phase occurs at $p = (4 - r)/4(2 - r)$. For $r = 0.5$, with a maximum at $p = 0.583$, gives $\bar{r} = 0.292$ and $AsE(\hat{r}) = 0.25$, *i.e.*, an underestimate of 14%. Smaller values of r and larger values of k also give smaller bias at the value of p that maximizes the bias.

The effect is much stronger in the repulsion phase. Given $k = 0$, the maximum difference occurs at $p = 0.25$, irrespective of r . The $AsE(\hat{r})$ will then be twice $\bar{r}$, *i.e.*, there will be a 100% overestimate irrespective of the value of r . The largest possible bias is reached for $r = 0.5$,

FIGURE 4.—Asymptotic bias, ϵ , as a function of probability of misclassification, λ . The bias is calculated as $\epsilon = AsE(\hat{r}) - r$. Solid curves: repulsion phase. Dotted curves: coupling phase.

which gives $\bar{r} = 0.125$ and $AsE(\hat{r}) = 0.25$. For $k > 0$ the effect on the bias is always smaller.

The fact that both cases above can be expressed in terms of θ means that it is impossible to detect the presence of these factors from the segregation ratio.

Factors distorting segregation

The model presented in Table 1 assumes a 3:1 segregation at both loci. There are, however, several factors that may cause a deviation from that expectation. We will focus on two specific factors that may cause a distorted segregation, namely misclassification and differential viability. Either of these factors may be due purely to experimental conditions or have genuine biological causes. In contrast to the case of heterogeneity of recombination frequencies, there exists a single correct value of recombination both for misclassifications and for differential viability. Thus, the bias reported below is the difference between the true value and the asymptotic expectation of the estimate, *i.e.*, $\epsilon = AsE(\hat{r}) - r$.

Misclassification: It is obvious that misclassifications may disturb the results in linkage studies. Lately, this problem has been given a formal treatment (SHIELDS *et al.* 1991; LINCOLN and LANDER 1992). Both papers point out that misclassifications systematically lead to a positive bias in estimates of recombination frequencies and, consequently, in map distances. Neither of the papers investigates the case of complete dominance at two loci.

To illustrate the effect of varying probability of misclassification, the bias, $\epsilon = AsE(\hat{r}) - r$, is shown in Figure 4 for different probabilities of misclassification, λ , varying from 0 to 0.02, and for different true recombination frequencies. The figure illustrates an extremely strong effect of misclassifications on the recombination frequency estimates in the repulsion phase. At low recombination frequencies, the estimates are completely unrelated to the actual value even for low probabilities of misclassification. For example, if $r = 0.01$ and $\lambda = 0.01$

TABLE 2

The bias in recombination estimates for different combinations of misclassification frequencies at a true recombination frequency of 0.10

Frequencies of misclassifications (%)				Bias, $\epsilon = AsE(\hat{r}) - r$	
$\lambda_{A \rightarrow a}$	$\lambda_{a \rightarrow A}$	$\lambda_{B \rightarrow b}$	$\lambda_{b \rightarrow B}$	In coupling	In repulsion
1	1	0	0	0.0086	0.0387
1	1	1	1	0.0170	0.0674
1	0	0	0	0.0066	0.0381
0	1	0	0	0.0019	0.0002
1	0	1	0	0.0113	0.0663
0	1	0	1	0.0038	0.0005
1	0	0	1	0.0086	0.0384
1	1	1	0	0.0151	0.0668
1	1	0	1	0.0105	0.0390

the bias is approximately 0.14. In the coupling phase the effect is not so marked, but misclassifications can be serious, particularly for low recombination frequencies and for the coupling phase. Figure 4 reveals the undesirable property that the bias is larger for smaller recombination frequencies. Compared to backcrosses, the distortion is larger in intercrosses with full dominance.

Moreover, in the case of complete dominance at a locus, there are actually two types of misclassification that may occur. First, individuals with a dominant allele may be classified as recessives and, second, recessives may be classified as the dominant phenotype. Thus, for two loci a general model of misclassification includes four parameters, in our notation $\lambda_{A \rightarrow a}$, $\lambda_{B \rightarrow b}$, $\lambda_{a \rightarrow A}$ and $\lambda_{b \rightarrow B}$. In Table 2 the effect of the four parameters is shown for a case in which there is a recombination frequency of 0.1 between the loci, and each probability of misclassification may take the values 0 or 1%. It is clearly seen that misclassifications of dominant into recessives have a much stronger effect than misclassifications in the reverse direction. It is also obvious that for almost all combinations, the effect is stronger in repulsion than in coupling. Moreover, the effects on the bias are nearly additive.

Differential viability: A differential viability may be associated with a specific allele, a genotype or a phenotype. In the case of morphological markers, there are primarily two models that have been investigated (BAILEY 1961). These correspond to model 1 and 2 in Table 3. In model 1, homozygotes for the recessive allele at one of the loci have a lower viability. Under this model the expectation of the estimate is not biased in either phase. Only the variance of the estimate is influenced (BAILEY 1961). In model 2, the homozygous recessives at either locus have a lowered viability with an independent (multiplicative) effect on the double recessive. In this case the recombination estimate is influenced. For example, if $r = 0.1$ and the viabilities are $s = t = 0.8$ on either of the recessives, the expectation of the estimate is 0.0977 in the coupling phase and 0.1032 in the repulsion phase.

TABLE 3

The expected proportions of phenotype classes in four investigated models of differential viability

Model	AB	Ab	aB	ab	Total
1	$\frac{2 + \theta}{D}$	$\frac{1 - \theta}{D}$	$\frac{s(1 - \theta)}{D}$	$\frac{s\theta}{D}$	$1; D = 3 + s$
2	$\frac{2 + \theta}{D}$	$\frac{t(1 - \theta)}{D}$	$\frac{s(1 - \theta)}{D}$	$\frac{ts\theta}{D}$	$1; D = 2 + t + s + \theta(1 - t)(1 - s)$
3	$\frac{s(2 + \theta)}{D}$	$\frac{1 - \theta}{D}$	$\frac{1 - \theta}{D}$	$\frac{s\theta}{D}$	$1; D = 2(1 + s - \theta(1 - s))$
4	$\frac{2 + \theta}{D}$	$\frac{s(1 - \theta)}{D}$	$\frac{s(1 - \theta)}{D}$	$\frac{\theta}{D}$	$1; D = 2(1 + s + \theta(1 - s))$

The parameters s and t are the relative viabilities of the phenotype. The expectations are for a double intercross with complete dominance.

The relative difference is approximately constant for different recombination frequencies. In either of the two models there is a symmetrical relation to the dominant phenotypes; for example, a lowered fitness on only one of the dominant markers does not influence the estimate.

Besides these two models, a large number of different cases may occur. For instance, when selectively neutral DNA markers are used in recombination studies, and these markers are tightly linked to alleles that are under selection, any combination of phenotype viability is possible. Instead of discussing all such cases, we illustrate the potential effects through two models which give the strongest effects on the estimates. The first case is model 3 in Table 3. If the true recombination frequency is 0.1 and the differential viability $s = 0.8$, then the expectation of the estimates is 0.122 and 0.0819 in coupling and repulsion, respectively. A similar effect is achieved under model 4, where the expectation of the estimates is 0.0813 and 0.128 for the two phases, given the same values of r and s . This means that for a viability of 0.8 on any combination of the four phenotypes, the maximum bias is 19% in the coupling phase and 28% in the repulsion phase. It is not surprising that models 3 and 4 give the strongest effects because the viabilities in both cases are connected with the classes showing either a recombinant or a parental phenotype.

For stronger differences in viability the bias, of course, gets stronger, but there is also an increasing probability of observing the skewed segregations and adjusting for them. In the case of model 2, for example, an asymptotically unbiased estimator of the recombination frequency is then given by the product formula or moments estimator of θ (BAILEY 1961; WEIR 1990). This method, however, is problematic at limited sample sizes because the estimator is not defined for several outcomes (*i.e.*, because the method depends on a product ratio where the denominator can be zero).

DISCUSSION

The present paper investigates the properties of the maximum likelihood estimates of recombination frequencies in double intercrosses with complete dominance. To evaluate the different causes of bias in the

estimates, two fundamental aspects must be considered. One is the likelihood that a disturbance will occur, and the other is the potential effect if the disturbance occurs. The problems considered here vary with respect both to how often they might occur and to the size of their effect.

First, the bias due to a limited sample size differs strongly in effect between the two types of linkage phases. In the coupling phase, the effect is already negligible at sample sizes above 20. In repulsion, however, the bias may be very serious even at sample sizes up to several hundred, especially for low recombination frequencies. It is important to point out that this effect is always present, independent of external conditions. It would be valuable, therefore, to find an effective method to reduce the bias. The fact that the jackknife method does not work satisfactory calls for alternative methods. One possibility is to calculate the bias for a number of combinations of sample sizes and observed recombination frequencies to provide a relationship which, in a specific case, can be used to correct the recombination estimate through interpolation. This method has been suggested by P. STAM (personal communication) for the JOINMAP mapping program (STAM 1993). However, this seems not to be a trivial solution to the problem because, for example, in cases when the estimate equals zero, an *a priori* distribution must be assumed to be able to correct for the bias (P. STAM, personal communication).

Second, heterogeneities in recombination frequencies also show a clear difference in effect between the coupling and repulsion phase. In the coupling phase, the maximum possible bias is low for both sex differences and differences among meioses. Thus, there is no reason for concern that heterogeneities in recombination frequencies can cause serious distortions in the coupling phase. In repulsion, however, the bias is again more pronounced. In particular, sex differences can give an incorrect picture of the recombination pattern in a species. This result may have empirical relevance because large sex differences in recombination frequencies are known in many organisms (BURT *et al.* 1991). In the most extreme case, this difference in recombination

between the sexes is complete with one sex having all the occurring recombination, such as in *Drosophila*.

The observation that sex differences have a stronger effect on the recombination estimates than differences between meioses contradicts the effects of the two types of variation on the coefficient of coincidence in triple crosses. Sex differences can only cause minor disturbances, whereas differences among meioses may have a considerable effect (SÄLL and BENGTSSON 1989). Such intra-individual heterogeneity in recombination has been observed in several species (JONES 1987). It is important to note that the described types of heterogeneity are undetectable with complete dominance because they cause no discrepancy with the basic model, in contrast to intercrosses with codominance. To detect such differences, it is necessary to test for them experimentally, for example, through reciprocal test crosses.

Third, factors that distort segregation also influence the recombination estimates. The initial effect of misclassification is particularly strong for the repulsion phase, but the coupling phase may also be seriously affected, particularly at low recombination frequencies. Misclassifications include everything from incomplete penetrance to misscoring and typing errors. In mapping experiments using DNA markers, the risk of misclassification is especially pronounced because large data matrices are produced and subsequently interpreted, scored and digitized in several steps. Furthermore, for RAPD data where reproducible results require great experimental care, misclassifications must be considered even more seriously. WEEDEN *et al.* (1992) concluded that error rates might be as high as 10% in segregating pea and apple populations, but if extra care is taken this can be kept below 4%. However, as seen above, an error rate of 4% would cause a tremendous bias in recombination estimates.

Misclassifications cause a deviation from the 3:1 segregation and, in principle should be possible to detect through a skewed segregation. However, even with a low probability of misclassification, large biases in recombination estimates are expected, leaving the segregation ratio only weakly distorted. For differential viability, however, strong effects on the recombination bias require large differences in viability, which in turn should be possible to detect as a deviation from a 3:1 ratio. Given this situation, it should be possible to adjust for the disturbance by choosing a different estimator of θ . However, because differential viability appears less serious than misclassification the problem of differential viability should not be overemphasized.

In conclusion: The most striking result is the pronounced difference between recombination estimates in coupling and repulsion. In the coupling phase, the estimates are quite robust. The major concern is to use a sample size which is large enough to compensate for the inherent statistical variation of the estimate. The repulsion phase, however, is very unstable. The situation

is particularly complicated because the different disturbing factors may act in different directions. The original bias is negative, and so is the effect of variation between the sexes. However, differences between meioses and misclassifications tend to increase the value.

An obvious piece of advice is, of course, to avoid situations which lead to crosses in repulsion. Nevertheless, there might very well occur situations in breeding or other applications where it is necessary to estimate recombination fractions in the repulsion phase. In such cases, it is of utmost importance to consider the results above. In particular, it is important to minimize the risk of misclassification.

We thank Drs. C. HALLDÉN, P. PETERSSON and P. STAM, and Prof. J. LANKE for valuable suggestions on the manuscript. The study was supported by a grant from the Erik Philip-Sørensen Foundation to T.S.

LITERATURE CITED

- ALLARD, R. W., 1956 Formulas and tables to facilitate the calculation of recombination values in heredity. *Hilgardia* **24**: 235–278.
- BAILEY, N. T. J., 1961 *Introduction to the Mathematical Theory of Genetic Linkage*. Clarendon Press, Oxford.
- BOLLING, D. R., and E. A. MURPHY, 1979 Finite sample properties of maximum likelihood estimates of the recombination fraction in double backcross matings in man. *Am. J. Med. Genet.* **3**: 81–95.
- BOTSTEIN, D., R. L. WHITE, M. SKOLNICK and R. W. DAVIS, 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**: 314–331.
- BURT, A., G. BELL, and P. H. HARVEY, 1991 Sex differences in recombination. *J. Evol. Biol.* **4**: 259–277.
- GRATTAPAGLIA, D., J. CHAPARRO, P. WILCOX, S. MCCORD, D. WERNER *et al.*, 1992 Mapping in woody plants with RAPD markers: application to breeding in forestry and horticulture, pp. 37–40 in *Proceedings of the Symposium on Applications of RAPD Technology to Plant Breeding*, Minneapolis.
- JONES, G. H., 1987 Chiasmata, pp. 213–244 in *Meiosis*, edited by P. B. MOENS. Academic Press, London.
- LINCOLN, S. E., and E. S. LANDER, 1992 Systematic detection of errors in genetic linkage data. *Genomics* **14**: 604–610.
- MATHER, K., 1938 *The Measurement of Linkage in Heredity*. Methuen, London.
- OTT, J., 1991 *Analysis of Human Linkage Data*. Johns Hopkins Press, Baltimore.
- RAFALSKI, J. A., S. V. TINGEY and J. G. K. WILLIAMS, 1991 RAPD markers—a new technology for genetic mapping and plant breeding. *Agric. Biotechnol. News* **3**: 645–648.
- REITER, R. S., J. G. K. WILLIAMS, K. A. FELDMANN, J. A. RAFALSKI, S. V. TINGEY *et al.*, 1992 Global and local genome mapping in *Arabidopsis thaliana* by using recombinant inbred lines and random amplified polymorphic DNAs. *Proc. Natl. Acad. Sci. USA* **89**: 1477–1481.
- RITTER, E., C. GERHARDT and F. SALAMINI, 1990 Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. *Genetics* **125**: 645–654.
- SHEILDS, D. C., A. COLLINS, K. H. BUETOW and N. E. MORTON, 1991 Error filtration, interference, and the human linkage map. *Proc. Natl. Acad. Sci. USA* **88**: 6501–6505.
- SÄLL, T., and B. O. BENGTSSON, 1989 Apparent negative interference due to variation in recombination frequencies. *Genetics* **122**: 935–942.
- STAM, P., 1993 Construction of integrated genetic linkage maps by means of a new computer package: JOINMAP. *Plant J.* **3**: 739–744.
- TANKSLEY, S., N. D. YOUNG, A. H. PATERSON and M. W. BONIERBALE, 1989 RFLP mapping in plant breeding: new tools for an old science. *Biotechnology* **7**: 257–264.

- TINGEY, S. V., J. A. RAFALSKI and J. G. K. WILLIAMS, 1992 Genetic analysis with RAPD markers, pp. 3–8 in *Proceedings of the Symposium on Applications of RAPD Technology to Plant Breeding*, Minneapolis.
- WEEDEN, N. F., G. M. TIMMERMAN, M. HEMMAT, B. E. KNEEN and M. A. LODHI, 1992 Inheritance and reliability of RAPD markers, pp. 12–17 in *Proc. Conf. Appl. RAPD Tech. Plant Breed.*
- WEIR, B. S., 1990 *Genetic Data Analysis*. Sinauer, Sunderland, MA.
- WILLIAMS, J. G. K., A. R. KUBELIK, K. J. LIVAK, J. A. RAFALSKI and S. V. TINGEY, 1990 DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* **18**: 6531–6535.
- WILLIAMS, J. G. K., A. R. KUBELIK, J. A. RAFALSKI and S. V. TINGEY, 1991 Genetic analysis with RAPD markers, pp. 431–439 in *More Gene Manipulations in Fungi*, edited by J. W. BENNETT and L. L. LASURE. Academic Press, San Diego.
- WILLIAMS, J. G. K., M. K. HANAFEY, J. A. RAFALSKI and S. V. TINGEY, 1993 Genetic analysis using random amplified polymorphic DNA markers. *Methods Enzymol.* **218**: 704–740.
- ZHAO, L. P., E. THOMPSON and R. PRENTICE, 1990 Joint estimation of recombination fractions and interference coefficients in multilocus linkage analysis. *Am. J. Hum. Genet.* **47**: 255–265.

Communicating editor: B. S. WEIR