Theor Appl Genet (1995) 90:81–89

M. Lorieux · X. Perrier · B. Goffinet
C. Lanaud · D. González de León

Maximum-likelihood models for mapping genetic markers showing segregation distortion. 2. F_2 populations

Received: 11 August 1993 / Accepted: 2 April 1994

Abstract In F_2 populations, gametic and zygotic selection may affect the analysis of linkage in different ways. Therefore, specific likelihood equations have to be developed for each case, including dominant and codominant markers. The asymptotic bias of the "classical" estimates are derived for each case, in order to compare them with the standard errors of the suggested estimates. We discuss the utility and the efficiency of a previous model developed for dominant markers. We show that dominant markers provide very poor information in the case of segregation distortion and, therefore, should be used with circumspection. On the other hand, the estimation of recombination fractions between codominant markers is less affected by selection than is that for dominant markers. We also discuss the analysis of linkage between dominant and codominant markers.

Key words Genetic mapping · Maximum-likelihood · Molecular markers · Gametic selection · Zygotic selection

Introduction

It has been shown that the estimation of recombination fractions may be biased by deviations of single-locus segregation ratios from expected frequencies (Bailey 1949; Allard and Alder 1960; Heun and Gregorius 1987). A typical source of deviation is the upsets in the formation or function of gametes or zygotes, due to the selection of one or more selected genes on the chromosomes. Other possible sources, such as partial manifestation or structural rearrangements like translocations, are not considered in this paper.

Bailey (1949) treated the case of the analysis of linkage between two dominant loci under zygotic selection, and Heun and Gregorius (1987) the case of one dominant locus under gametic or zygotic selection. In F_2 populations, these two types of selection do not similarly affect the estimation of the recombination fraction. The case of codominant loci has not yet been considered. We present here maximum-likelihood methods for the estimation of the recombination fractions between dominant or codominant markers showing segregation distortion. The utility and efficiency of the Heun and Gregorius model are discussed for dominant markers. We also discuss the analysis of linkage between a dominant and a codominant marker. The asymptotic bias of the "classical" estimates are derived, in order to compare them with the standard errors of the suggested estimates.

It will be assumed for simplicity that the markers show segregation distortion, because they are both exactly located on genes affected by gametic or zygotic selection. For gametic selection, we will always assume that only male gametes are affected. This assumption seems to be realistic, because pollen grains are more often affected by differential viability, or by differential capacity to fertilization, than are ovules.

Dominant markers

Consider a coupling mating of the type $AB/ab \times AB/ab$, involving two dominant markers, A and B. Four phenotypic classes are obtained. Table 1 shows the expected and observed frequencies of the four classes, assuming (1) gametic selection on A (2) independent

Communicated by G. Wenzel

M. Lorieux[1] (✉) · C. Lanaud
CIRAD-BIOTROP, B.P. 5035, 34032 Montpellier Cedex 1, France

X. Perrier
CIRAD-FLHOR, B.P. 5035, 34032 Montpellier Cedex 1, France

B. Goffinet
Station de Biométrie et d'Intelligence Artificielle, INRA, B.P. 27, 31326 Castanet-Tolosan Cedex, France

D. González de León
CIMMYT, Lisboa 27, Colonia Juárez, Apdo. Postal 6-641, 06600 México, D.F., México

Present address:
[1]LRGAPT-ORSTOM, BP 5045, 34032 Montpellier Cedex 1, France

Table 1 Expected and observed frequencies for an F_2 in coupling, involving two dominant markers, A and B. For matings in repulsion, r is replaced by $1-r$

Phenotypes	AB	Ab	aB	ab
Expected frequencies (gametic selection on A)	$n\frac{1+r(r-1-u_g)+2u_g}{2u_g+2}$	$n\frac{u_g r+r-r^2}{2u_g+2}$	$n\frac{2r-r^2}{2u_g+2}$	$n\frac{(1-r)^2}{2u_g+2}$
Expected frequencies (gametic selection on A and B)	$n[1+r(r+u_g+v_g-2)+(1-r)(2u_g v_g)]/D$	$n\frac{u_g r+r-r^2}{D}$	$n\frac{v_g r+r-r^2}{D}$	$n\frac{(1-r)^2}{D}$
Expected frequencies (zygotic selection)	$n\frac{uv(3-2r+r^2)}{D}$	$n\frac{u(2r-r^2)}{D}$	$n\frac{v(2r-r^2)}{D}$	$n\frac{(1-r)^2}{D}$
Observed frequencies	a	b	c	d

$D=2(u_g v_g+1)(1-r)+2(u_g+v_g)r$ (gametic selection)
$D=uv(3-2r+r^2)+(u+v)(2r-r^2)+(1-r)^2$ (zygotic selection)
u_g = viability of A gametes relative to a gametes; v_g = viability of B gametes relative to b gametes; u = viability of A – zygotes relative to aa zygotes; v = viability of B – zygotes relative to bb zygotes

gametic selection on A and B and (3) independent zygotic selection on A and B. By independent selection, we mean that the two markers are affected by two types of selection from different sources, i.e., that two selected genes are involved. Now, consider that for gametic selection, the viability of A gametes relative to a is u_g, and the viability B gametes relative to b is v_g. For zygotic selection, the viability of A phenotypes relative to a is u, and the viability of B phenotypes relative to b is v. As each phenotype or haplotype can be favored by selection, the values of u, u_g, v or v_g fall between 0 and $+\infty$. The case $u=v=1$ or $u_g=v_g=1$ is that of no selection, i.e., Mendelian segregation.

Zygotic selection

It has been shown that when only one marker is affected by zygotic selection, the classical estimate, $\hat{r}$, of the recombination fraction is consistent (Bailey 1949). This estimate can be expressed as

$$\hat{r}=1-\sqrt{\frac{d+3a}{2n}+\sqrt{1+\frac{d-3a}{n}+\frac{(d+3a)^2}{4n^2}}-1}.$$

However, the asymptotic variance of $\hat{r}$ depends on the intensity of selection (Fig. 1).

If it is known that the two markers are under zygotic selection, the product formula method must be used, as it leads to a consistent and fully-efficient estimate of r (Fisher and Balmukand 1928; Bailey 1949). An estimate is consistent, or asymptotically unbiased, if it converges to the "true" value of the parameter as the population size increases. It is efficient if no other estimate has a smaller variance. In coupling phase, the product formula method leads to the estimate

$$\hat{r}=1-\sqrt{\frac{\sqrt{bc^2+3abcd}-ad-bc}{bc-ad}}. \qquad (1)$$

Let us put $\theta=(1-r)^2$ for the coupling phase. The asymptotic variance of $\hat{r}$ is given by

$$V_{\hat{r}}=V_{\hat{\theta}}\times\left(\frac{dr}{d\theta}\right)^2=\frac{V_{\hat{\theta}}}{(2r-2)^2}$$

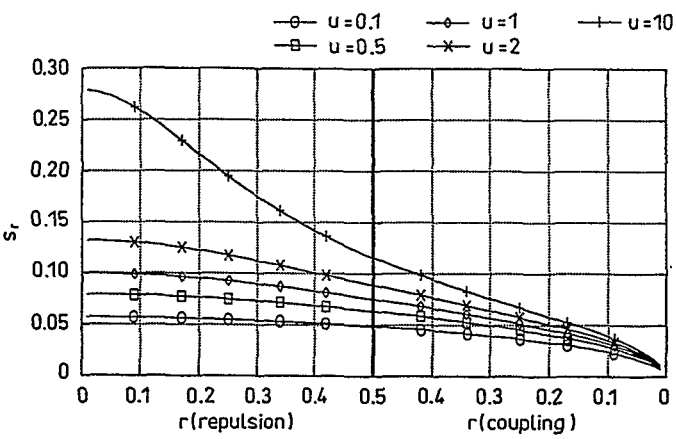
where

$$V_{\hat{\theta}}=\frac{\theta(1-\theta)(2+\theta)}{4nuv(1+2\theta)^2}[uv(2+\theta)+(u+v)(1-\theta)+\theta]\times[\theta(1-\theta)+(u+v)\theta(2+\theta)+uv(2+\theta)(1-\theta)].$$

In repulsion phase, we have $\theta=r^2$, and $V_{\hat{r}}$ is equal to $V_{\hat{\theta}}/4r^2$. Figure 2 gives the values of the standard error of $\hat{r}$, $s_{\hat{r}}$, against $\hat{r}$, for coupling and repulsion phases.

On the other hand, if only one marker is selected, the estimate (1) is consistent, but not fully efficient. Nevertheless, the loss of efficiency in this case is negligible (Bailey 1949).

Fig. 1 Asymptotic standard error of the classical estimate of the recombination fraction, r, between two dominant markers, one being affected by a zygotic selection with intensity u (F_2 population of 100 individuals)



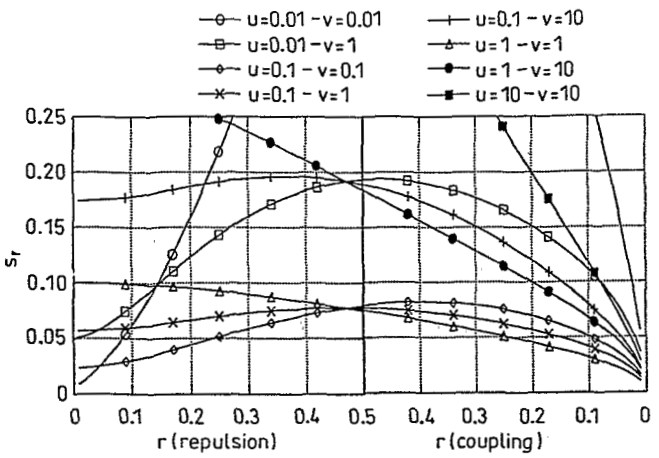


Fig. 2 Asymptotic standard error of the product formula estimate of the recombination fraction, r, between two dominant markers being affected by zygotic selections with intensities u and v (F_2 population of 100 individuals)

Selection of unknown type

We will see below that gametic selection does not affect the estimation of recombination fractions in the same way as zygotic selection. For dominant markers, no statistical method allows one to precisely distinguish gametic from zygotic selection, because allelic frequencies cannot be well estimated if zygotic selection occurs. Heun and Gregorius (1987) developed a model for estimating linkage between two dominant loci, A and B, when the selection type is unknown. The authors assume that only one locus, A, is under selection of an unknown type. Under this condition, only two classes (aB and ab) may be used to estimate the recombination fraction, leading to the estimate:

$$\hat{r}_{H\&G} = \sqrt{\frac{d}{c+d}}$$

where c and d are the observed frequencies of the two classes aB and ab. Suppose that the viability of A phenotypes relative to a is u if zygotic selection occurs, and that the viability of A gametes relative to a is u_g if gametic selection occurs. We have the relationship

$$\hat{u} = \frac{\hat{u}_g^2 + 2\hat{u}_g}{3} \quad \text{or} \quad \hat{u}_g = \sqrt{3\hat{u}+1} - 1$$

and, deriving Fisher's information matrix (Fisher 1937), we find that the asymptotic standard error of $\hat{r}_{H\&G}$ is

$$S_{\hat{r}_{H\&G}} = \sqrt{\frac{(u_g+1)(4r^3 - r^4 - 5r^2 + 2r)}{2n(1-r)^2}} \tag{3}$$

$$= \sqrt{\frac{(\sqrt{3u+1})(4r^3 - r^4 - 5r^2 + 2r)}{2n(1-r)^2}}. \tag{3bis}$$

Figure 3 shows the value of $s_{\hat{r}_{H\&G}}$ as a function of r and u_g. It can be compared with the asymptotic bias of the classical maximum likelihood estimate

$$B_{\hat{r}} = 1 - \sqrt{\frac{d+3a}{2n} + \sqrt{1 + \frac{d-3a}{n} + \frac{(d+3a)^2}{4n^2}}} - 1 - r \tag{4}$$

where the observations are replaced by their expectations (coupling phase). Figure 4 shows the values of $B_{\hat{r}}$ against r for several intensities of gametic selection. The comparison between Figs. 3 and 4 clearly indicates that, for strong gametic selection in favor of the dominant allele ($u \gg 1$), $s_{\hat{r}_{H\&G}}$ is much larger than $B_{\hat{r}}$. This means that this estimate must be used with circumspection. A further comparison with the asymptotic bias of the estimate (1) obtained by the product formula would give rise to the same conclusions (Fig. 5). An additional limitation of the Heun and Gregorius model is that it assumes that we know which locus is selected. In a practical situation, the determination of this locus is not

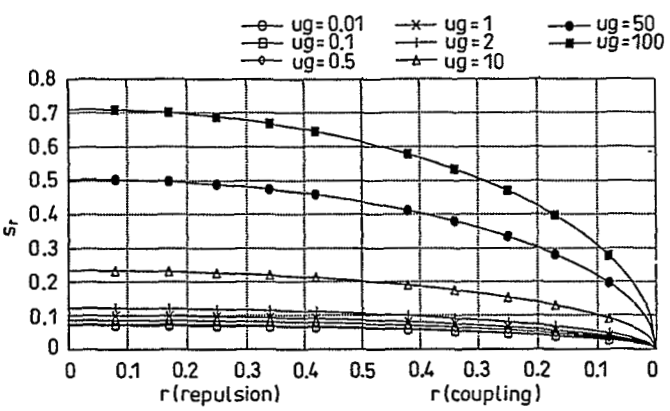
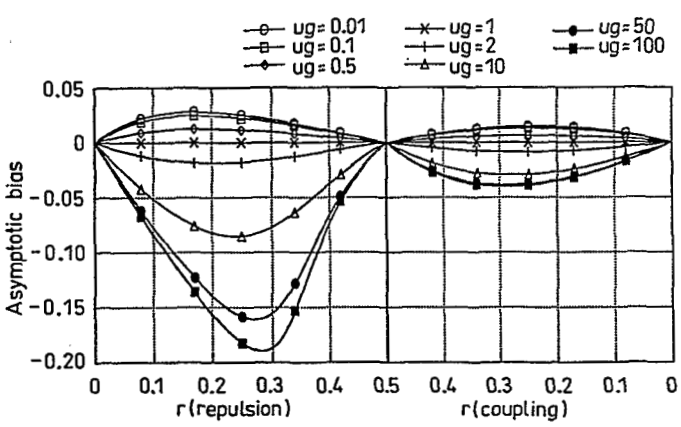


Fig. 3 Asymptotic standard error of Heun and Gregorius estimate of the recombination fraction, r, between two dominant markers, one being affected by a gametic selection with intensity u_g (F_2 population of 100 individuals)

Fig. 4 Asymptotic bias of the classical estimate of the recombination fraction, r, between two dominant markers, one being affected by a gametic selection with intensity u_g (F_2 population)



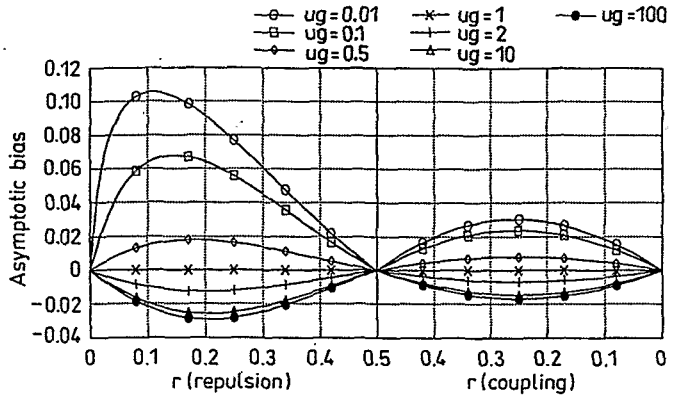


Fig. 5 Asymptotic bias of the product formula estimate of the recombination fraction, r, between two dominant markers, one being affected by a gametic selection with intensity u_g (F_2 population)

straightforward. Moreover, this model does not take into account the possibility that the two markers may be affected by independent selection. When the selection type is unknown, we therefore suggest the use of estimate (1), given by the method of the product formula, since it is consistent in the case of zygotic selection, and only slightly biased in the case of gametic selection.

Gametic selection

For certain species, such as rice, it has been established that gametic selection is almost the unique form of selection (Lin et al. 1992). It is thus possible to use the full information available in the four phenotypic classes. If only A is under selection, then the likelihood equation to be solved is (see Table 1, gametic selection on A)

$$\frac{\partial L}{\partial r} = a\frac{2r-1-u_g}{1+r(r-1-u_g)+2u_g} + b\frac{u_g+1-2r}{u_g r+r-r^2} + c\frac{2-2r}{2r-r^2} + d\frac{2}{r-1} = 0 \qquad (5)$$

where u_g is estimated by

$$\hat{u}_g = \frac{n-2(c+d)}{2(c+d)}.$$

The asymptotic variance of $\hat{r}$ cannot be easily derived, because the covariance between r and u_g is not null; its calculation requires the inversion of Fisher's expected information matrix. However, it can be shown that a good approximation is obtained by simply inverting the expected information for r, because the covariance between the two parameters is always close to zero. We get the following approximation, illustrated in Fig. 6,

$$V_{\hat{r}} = r(2/n)(2-r)(r-u_g-1)(1+u_g) \times (1-r+r^2+2u_g-ru_g)/[17r-20r^2+8r^3+u_g \times (44r-36r^2+8r^3-20)+u_g^2(25r-8r^2+18) + u_g^3(2r-4)-6]. \qquad (6)$$

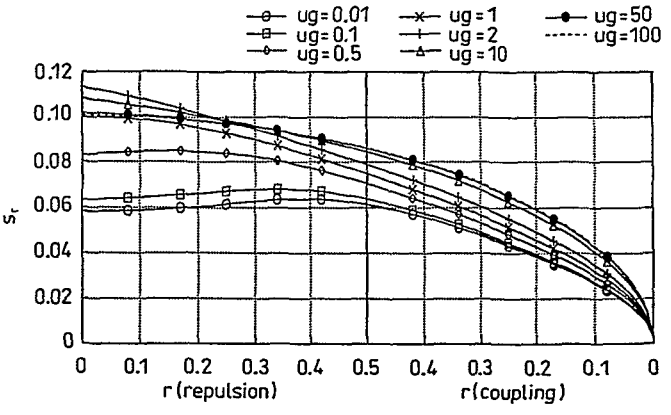
The relative efficiency of $\hat{r}_{H\&G}$, compared to this estimate, is obtained by dividing $V_{\hat{r}}$ by $V_{\hat{r}_{H\&G}}$ (Fig. 7). This figure shows that the Heun and Gregorius estimate is drastically inefficient for strong selection in favor of the dominant allele. This confirms the conclusions of the comparison of Figs. 3 and 4, i.e., that the Heun and Gregorius estimate has to be used carefully.

As for Heun and Gregorius model, the utilization of (5) requires one to determine which marker is affected by selection. This problem is circumvented by using a more general model which assumes that A and B are under gametic selection. Then, a system of three maximum-likelihood equations have to be solved iteratively, and the calculation of the asymptotic variance of $\hat{r}$ requires the inversion of the expected information matrix (see Appendix).

Codominant markers

Codominant markers are more informative than dominant markers, since heterozygotes are distinguished from homozygotes. We show here that, in case of segregation distortion, the advantage of codominant markers is enhanced. Consider a coupling mating of the type: $AB/ab \times AB/ab$, involving two codominant markers, A and B. Nine phenotypic classes are obtained, with the observed and expected frequencies summarized in Table

Fig. 6 Asymptotic standard error of the fully efficient estimate of the recombination fraction, r, between two dominant markers, one being affected by a gametic selection with intensity u_g (F_2 population of 100 individuals)



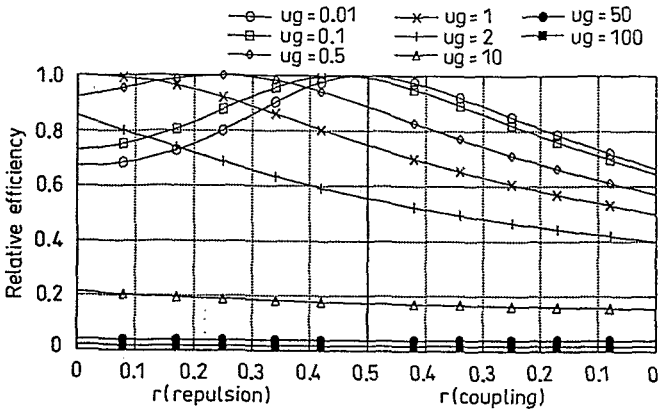


Fig. 7 Relative efficiency of the Heun and Gregorius estimate, compared to the fully efficient estimate (5)

2. It can be inferred from this table that the expected frequencies are not the same if gametic or if zygotic selection occurs. Note that the case where only one marker is selected is obtained in Table 2 by setting u or v to one. In coupling phase, when both markers show Mendelian segregation, the MLE of the recombination fraction is given by solving

$$\frac{\partial L}{\partial r}=(a+i)\frac{2}{r-1}+(b+d+f+h)\frac{1-2r}{r(1-r)}+(c+g)\frac{2}{r}+e\frac{4r-2}{1+2r^2-2r}=0. \quad (7)$$

The variance of the estimate obtained by solving (7) is

$$V_{\hat{r}}=\frac{2(1-3r+3r^2)}{r(1-r)(1-2r+2r^2)} \text{ (Allard 1956).}$$

Since (7) has no analytical solution, it has to be solved iteratively, using a Newton-Raphson, or an EM, algorithm (Edwards 1972; Mangin 1991). It has been shown that these two algorithms lead to maximum-likelihood estimates (MLEs) with different speeds of convergence (Dempster et al. 1977; Wu 1983). The asymptotic bias of the MLE of r obtained by (7), in case of gametic or zygotic selection, is obtained by subtracting the true value of r from its estimation. We derived this bias in case of zygotic selection, using a Newton-Raphson algorithm (Fig. 8a). In both cases of selection, it can be shown that this MLE of r, which ignores selection, is consistent if only one marker is under gametic or zygotic selection. The demonstration is given by deriving the log-likelihood for gametic and zygotic selection, using the expected frequencies of Table 2 and setting one of the two selection parameters equal to 1. For both selection cases, the derivative is identical to (7). This consistency is an additional advantage of codominant over dominant markers, since for dominant markers, $\hat{r}$ is biased in the case of gametic selection on one of the two markers. Moreover, it can be shown by deriving Fisher's information matrix, that, for codominant markers, the variance of $\hat{r}$ is unchanged for gametic selection, i.e. it does not depend on the intensity of the distortion. Note that the two markers will generally show segregation distortion, even when only one marker is under selection. This effect will be proportional to the intensity of linkage between the markers. In practice, this situation is not distinguishable from the situation where gametic or zygotic selection occurs on both markers. Therefore, when both markers show segregation distortion, it is preferable to always assume that they are located on two selected genes. Under this assumption, systems of maximum-likelihood equations have to be solved iteratively in both cases (see Appendix, systems A.6 and A.7). The derivation of the asymptotic variances of the estimates of r requires the inversion of the expected information matrix, which is (3, 3) for gametic selection or (5,5) for zygotic selection. Analytical expressions for the two cases were derived using Mathematica (Wolfram 1988). Figure 8b shows the values of the asymptotic standard error of $\hat{r}$ estimated by the resolution of system (A.6), which has to be used in the case of zygotic selection ($n = 100$ individuals). Comparing these values with the

Table 2 Expected and observed frequencies for an F_2 in coupling, involving two codominant markers, A and B. For matings in repulsion, r is replaced by $1-r$

Phenotypes	Expected frequencies (gametic selection)	Expected frequencies (zygotic selection)	Observed frequencies
AABB	$nu_g v_g(1-r)^2/D$	$nu_1 v_1(1-r)^2/D$	a
AABb	$n(u_g+u_g v_g)(r-r^2)/D$	$n2u_1 v_2(r-r^2)/D$	b
AAbb	$nu_g r^2/D$	$nu_1 r^2/D$	c
AaBB	$n(v_g+u_g v_g)(r-r^2)/D$	$n2u_2 v_1(r-r^2)/D$	d
AaBb	$n[(1-r)^2(1+u_g v_g)+r^2(u_g+v_g)]/D$	$n2u_2 v_2(1-2r+2r^2)/D$	e
Aabb	$n(1+u_g)(r-r^2)/D$	$n2u_2(r-r^2)/D$	f
aaBB	$nv_g r^2/D$	$nv_1 r^2/D$	g
aaBb	$n(1+v_g)(r-r^2)/D$	$n2v_2(r-r^2)/D$	h
aabb	$n(1-r)^2/D$	$n(1-r)^2/D$	i

$D=2(u_g v_g+1)(1-r)+2(u_g+v_g)r$ (gametic selection)
$D=(1-r)^2(u_1 v_1+1)+r(1-r)[2u_2(1+v_1)+2v_2(1+u_1)]+r^2(u_1+v_1)+(1+2r^2-2r)(2u_2 v_2)$ (zygotic selection)
u_g = viability of A gametes relative to a gametes; v_g = viability of B gametes relative to b gametes; u_1 = viability of AA zygotes relative to aa zygotes; u_2 = viability of Aa zygotes relative to aa zygotes; v_1 = viability of BB zygotes relative to bb zygotes; v_2 = viability of Bb zygotes relative to bb zygotes

bias of the classical estimate (Fig. 8 a), the advantage of using system (A.6) instead of the classical estimate is clearly seen. For a smaller population size, however, simulations should be done in order to study the bias. It is noticeable that when no selection occurs, the standard errors of the estimates given by (A.6) or (A.7) are equal to that of the estimate given by (7). This means that no information is lost when using these systems, even when the assumption of two selected genes is false.

It follows that the selection type has to be determined in order to choose the appropriate system to be solved. With codominant markers, it is possible to test for the selection type by using successive χ^2 tests (Pham et al. 1990). Consider a codominant marker, A, with two alleles A and a. The individuals in the F_2 population are AA, Aa or aa, with expected Mendelian proportions, 1:2:1. This expectation is tested by

$$\chi_2^2 = \frac{4n_{AA}^2 + 2n_{Aa}^2 + 4n_{aa}^2}{n} - n. \tag{8}$$

If it is found to be significant, the segregation is distorted, and the type of selection has to be determined. Let p be the allelic frequency of A, and q the frequency of a. The MLEs of p and q are

$$\hat{p} = n_{AA} + n_{Aa}/2$$
$$\hat{q} = n_{aa} + n_{Aa}/2. \tag{9}$$

The hypothesis $p = q$, is tested by

$$\chi_1^2 = \frac{(2n\hat{p} - n)^2 + (2n\hat{q} - n)^2}{n}. \tag{10}$$

If gametic selection occurred, this test will be significant. If zygotic selection occurred, it may or may not be significant, depending on which genotype is selected. One then has to test the hypothesis that the associations of gametes occurred at random, i.e., the phenotypes are distributed according to $p^2{:}2pq{:}q^2$. The test is given by

$$\chi_1^2 = \frac{(n_{AA} - n\hat{p})^2}{n\hat{p}^2} + \frac{(n_{Aa} - 2n\hat{p}\hat{q})^2}{2n\hat{p}\hat{q}} + \frac{(n_{aa} - n\hat{q})^2}{n\hat{q}^2}. \tag{11}$$

If only gametic selection occurred, this test will not be significant. On the other hand, if zygotic selection occurred it will be significant. The following table summarizes the interpretation of (10) and (11):

Test (10)	Test (11)	Selection type
Non significant	Significant	Zygotic
Significant	Significant	Zygotic
Significant	Non significant	Gametic

It should be noted that it is not possible to use a similar method for dominant markers, since the estimation of p and q uses only the phenotypes aa. Thus, the selection would always appear to be of gametic type, and, consequently, a test equivalent to (11) would always be equal to zero.

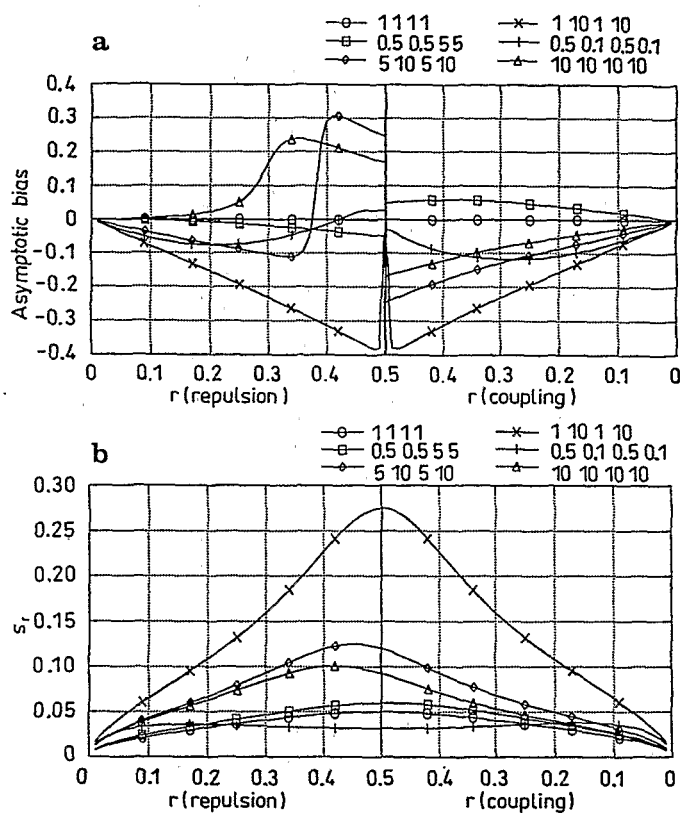


Fig. 8a, b **a** Asymptotic bias of the classical MLE (using Newton-Raphson iterations) of the recombination fraction, r, between two markers affected by a zygotic selection. **b** Asymptotic standard error of the MLE (equation A.6; see Appendix) in the same conditions, for an F_2 population of 100 individuals. The numbers in the legends denote the values of u_1, u_2, v_1 and v_2, respectively

Dominant and codominant markers

The analysis of linkage is affected by either gametic selection on either one of two dominant markers. It can be shown that if one of the two markers is codominant, then the classical estimate of r stays consistent when gametic or zygotic selection occurs, provided that one marker only is under selection. The value of the variance of $\hat{r}$ depends on the type of selection and of the segregation mode of the marker selected. As for codominant markers, a system of maximum-likelihood equations has to be solved iteratively for each case when it is not known if one or both markers are under selection (see systems A.8 and A.9 in the Appendix, and Table 3). However, assumptions have to be made about the selection type occurring on the dominant marker, since it cannot be inferred from the distribution of phenotypes (Lorieux 1993).

Table 3 Expected and observed frequencies for an F_2 in coupling, involving one dominant marker, *A* and one codominant marker, *B*. For matings in repulsion, *r* is replaced by $1-r$

Phenotypes	Expected frequencies (gametic selection)	Expected frequencies (zygotic selection)	Observed frequencies
A-BB	$n[rv_g(1-r-u_g)+u_gv_g]/D$	$nuv_1(1-r^2)/D$	*a*
A-Bb	$n[r(u_g+rv_g)+(1-r)^2+(1-r)(u_gv_g)]/D$	$n2uv_2(1-r+r^2)/D$	*b*
A-bb	$n[r(1+u_g)-r^2]/D$	$nu(2r-r^2)/D$	*c*
aaBB	$n(v_gr^2)/D$	nv_1r^2/D	*d*
aaBb	$n[r(1-r)(1+v_g)]/D$	$n2v_2(r-r^2)/D$	*e*
aabb	$n(1-r)^2/D$	$n(1-r)^2/D$	*f*

$D=2(u_gv_g+1)(1-r)+2(u_g+v_g)r$ (gametic selection)
$D=r^2(1-2v_2+v_1-u+2uv_2-uv_1)+2r(u-uv_2+v_2-1)+uv_1+2uv_2+1$ (zygotic selection)
u = viability of *A* gametes relative to *a* gametes; v_g = viability of *B* gametes relative to *b* gametes; u_1 = viability of *A* zygotes relative to *aa* zygotes; v_1 = viability of *BB* zygotes relative to *bb* zygotes; v_2 = viability of *Bb* zygotes relative to *bb* zygotes

Discussion

Specific maximum-likelihood estimates for each case of selection were derived for the analysis of the recombination fraction between two genetic markers, for F_2 populations.

It is noticeable that the variance of the recombination estimate is lower for certain selection patterns than for Mendelian segregation of the markers (see Figs. 6 and 8b). This phenomenon is simply explained by the fact that, for these patterns, the most informative phenotypes (e.g. the *ab* phenotype for a coupling mating involving two dominant markers, whose genotype is known with certainty) are favored.

Dominant markers, such as RAPDs, provide poor information, especially in repulsion phase (Allard 1956). This problem increases when zygotic selection occurs, implying that the use of such markers should be done carefully. However, when codominant markers such as RFLPs are associated with dominant markers, some properties of the codominant loci are retained. First, gametic selection on one of the two loci does not affect the estimation of the recombination fraction. Secondly, matings are as informative in repulsion as in coupling. We therefore suggest the use of as many codominant markers as possible for a species where segregation distortion is known to be frequent. If only dominant markers are available, then the product formula (equation 11) should be used, since it leads to a consistent estimate in case of zygotic selection, and to a less-biased estimate in case of gametic selection. However, if it is known that only gametic selection occurs, then the system (A.2) should be solved.

Gametic and zygotic selection does not modify the estimation of *r* in the same way. Thus, when the type of selection is not known, general models could be used, which take into account the two possibilities of selection. For dominant markers, such a model which uses the full information (i.e. the four phenotypic classes) cannot work, and the Heun and Gregorius estimate, which is theoretically correct, has a large variance. For codominant markers, a general model is not necessary, since it is possible to determine what type of selection occurred at a locus by using two successive χ^2-tests: the first tests the equality of the allelic frequencies, and the second tests for independent assortment of the alleles (Pham et al. 1990). These tests cannot apply to dominant markers because the estimates of allelic frequencies are biased in case of zygotic selection.

It should be noted that in the case of two markers affected by two independent selections, we have only presented the situation where the two selections are of the same type. For the situation where one marker is under gametic selection and one marker under zygotic selection, formulas and curves are available from the corresponding author.

As an alternative to selected genes, structural rearrangements such as translocations may affect the viability of gametes (Fauré et al. 1993). The models described above do not apply in this case, and probably the answer is not a statistical one.

Appendix

We present here the maximum-likelihood treatment of segregation data with segregation distortion, in the case of two markers under selection. The case of gametic selection on each of two dominant markers is detailed. Systems to be solved in cases of codominant and dominant/codominant markers are also presented, for gametic and zygotic selection.

Dominant markers–gametic selection

From Table 1 (gametic selection on *A* and *B*), we can write the log-likelihood

$$L=a\log[1+r(r-2-2u_gv_g+u_g+v_g)+2u_gv_g]+b\log[u_gr+r-r^2]+c\log[v_gr+r-r^2]+d\log(1-r)^2-n\log[(u_gv_g)(1-r)+(u_g+v_g)r]. \quad (A.1)$$

The system of maximum-likelihood equations to be solved in order to estimate the parameters is obtained by partially deriving (A.1)

$$\begin{cases}
\dfrac{\partial L}{\partial r} = a\dfrac{2r-2-2u_gv_g+u_g+v_g}{1+r(r-2-2u_gv_g+u_g+v_g)+2u_gv_g} \\
\quad +b\dfrac{1-2r+u_g}{u_gr+r-r^2}+c\dfrac{2-2r+v_g}{v_gr+r-r^2}+2d\dfrac{1}{r-1} \\
\quad -n\dfrac{u_g+v_g-u_gv_g-1}{(u_gv_g+1)(1-r)+(u_g+v_g)r}=0 \\
\dfrac{\partial L}{\partial u_g} = a\dfrac{r+2v_g-2rv_g}{1+r(r-2-2u_gv_g+u_g+v_g)+2u_gv_g} \\
\quad +b\dfrac{r}{u_gr+r-r^2}-n\dfrac{v_g(1-r)+r}{(u_gv_g+1)(1-r)+(u_g+v_g)r}=0 \\
\dfrac{\partial L}{\partial v_g} = a\dfrac{r+2u_g-2ru_g}{1+r(r-2-2u_gv_g+u_g+v_g)+2u_gv_g} \\
\quad +c\dfrac{r}{v_gr+r-r^2}-n\dfrac{u_g(1-r)+r}{(u_gv_g+1)(1-u)+(u_g+v_g)u}=0.
\end{cases} \tag{A.2}$$

An iterative method, such as Newton-Raphson's algorithm, may be used (see Edwards 1972) to solve this system. Derivation of the asymptotic variance of $\hat{r}$ requires the inversion of Fisher's expected information matrix, i

$$i = \begin{array}{c|ccc} & r & u_g & v_g \\ \hline r & i_{r,r} & i_{r,u} & i_{r,v} \\ u_g & & i_{u,u} & i_{u,v} \\ v_g & & & i_{v,v} \end{array} \tag{A.3}$$

where $i_{\theta,\phi}$ is the expected information for parameters θ and ϕ, given by the formula

$$i_{\theta,\phi} = \sum_{j=1}^{t}\left[\frac{1}{m_j}\left(\frac{\partial m_j}{\partial \theta}\right)\left(\frac{\partial m_j}{\partial \phi}\right)\right] \tag{A.4}$$

where t is the number of phenotypic classes, and m_j is the expected frequency of class j (obvious g subscripts were removed for clarity). Then, inversion of i leads to

$$V_{\hat{r}} = \frac{i_{u,u}i_{v,v}-i_{u,v}^2}{n(2i_{r,u}i_{r,v}i_{u,v}-i_{r,r}i_{u,v}^2-i_{u,u}i_{r,v}^2-i_{v,v}i_{r,u}^2+i_{r,r}i_{u,u}i_{v,v})} \tag{A.5}$$

which is a function of r, u_g and v_g.

Codominant markers–zygotic selection

The same method as for dominant markers can be applied for codominant markers. From Table 2, we obtain the system

$$\begin{cases}
\dfrac{\partial L}{\partial r} = (a+i)\dfrac{2}{r-1}+(b+d+f+h)\dfrac{1-2r}{r(1-r)}+(c+g)\dfrac{2}{r}+e\dfrac{4r-2}{1+2r^2-2r} \\
\quad -n\dfrac{(2r-2)(u_1v_1+1)+(1-2r)[2u_2(v_1+1)+2v_2(u_1+1)]+2r(u_1+v_1)+(4r-2)(2u_2v_2)}{D}=0 \\
\dfrac{\partial L}{\partial u_1} = \dfrac{a+b+c}{u_1}-\dfrac{n[v_1(1-2r+r^2)+2v_2(r-r^2)+r^2]}{D}=0 \\
\dfrac{\partial L}{\partial u_2} = \dfrac{d+e+f}{u_2}-\dfrac{2n[v_2(1-2r+r^2)+v_1(r-r^2)+r-r^2]}{D}=0 \\
\dfrac{\partial L}{\partial v_1} = \dfrac{a+d+g}{v_1}-\dfrac{n[u_1(1-2r+r^2)+2u_2(r-r^2)+r^2]}{D}=0 \\
\dfrac{\partial L}{\partial v_2} = \dfrac{b+e+h}{v_2}-\dfrac{2n[u_2(1-2r+r^2)+u_1(r-r^2)+r-r^2]}{D}=0
\end{cases} \tag{A.6}$$

where $D=(1-r)^2(u_1v_1+1)+r(1-r)[2u_2(v_1+1)+2v_2(u_1+1)]+r^2(u_1+v_1)+(1+2r^2-2r)(2u_2v_2)$.

The values of the standard error of $\hat{r}$ are shown in Fig. 8 b. Their calculations require the inversion of a (5, 5) information matrix. The standard error for a population size n can be obtained by multiplying the value of interest in Fig. 8 b by $\sqrt{100/n}$.

Codominant markers–gametic selection (Table 2)

$$\begin{cases}
\dfrac{\partial L}{\partial r} = (a+i)\dfrac{2}{r-1}+(b+d+f+h)\dfrac{1-2r}{r(1-r)}+(c+g)\dfrac{2}{r} \\
\quad +e\dfrac{2(r-1+u_gv_gr-u_gv_g+u_gr+v_gr)}{(1+u_gv_g)(1-r)^2-(u_g+v_g)r^2}-n\dfrac{u_g+v_g-u_gv_g-1}{D}=0 \\
\dfrac{\partial L}{\partial u_g} = \dfrac{a+c}{u_g}+\dfrac{dv_g}{v_g+u_gv_g}+\dfrac{b(1+v_g)}{u_g+u_gv_g}+\dfrac{f}{1+u_g} \\
\quad +e\dfrac{v_g(1-r)^2+r^2}{(1+u_gv_g)(1-r)^2+(u_g+v_g)r^2}-n\dfrac{v_g(1-r)+r}{D}=0 \\
\dfrac{\partial L}{\partial v_g} = \dfrac{a+g}{v_g}+\dfrac{d(1+u_g)}{v_g+u_gv_g}+\dfrac{bu_g}{u_g+u_gv_g}+\dfrac{h}{1+v_g} \\
\quad +e\dfrac{u_g(1-r)^2+r^2}{(1+u_gv_g)(1-r)^2+(u_g+v_g)r^2}-n\dfrac{u_g(1-r)+r}{D}=0
\end{cases} \tag{A.7}$$

where $D=(u_gv_g+1)(1-r)+(u_g+v_g)r$

One dominant and one codominant marker–zygotic selection
(Table 3)

$$\begin{cases}
\dfrac{\partial L}{\partial r} = a\dfrac{2r}{r^2-1}+b\dfrac{2r-1}{1+r^2-r}+c\dfrac{2-2r}{2r-r^2}+d\dfrac{2r}{r^2}+e\dfrac{1-2r}{r-r^2}+f\dfrac{2}{r-1} \\
\quad -n\dfrac{2r(1-2v_2+v_1-u+2uv_2-uv_1)+2(u-uv_2+v_2-1)}{D}=0 \\
\dfrac{\partial L}{\partial u} = \dfrac{a+b+c}{u}-n\dfrac{r^2(2v_2-v_1-1)+2r(1-v_2)+v_1+2v_2}{D}=0 \\
\dfrac{\partial L}{\partial v_1} = \dfrac{a+d}{v_1}-n\dfrac{r^2(1-u)+u}{D}=0 \\
\dfrac{\partial L}{\partial v_2} = \dfrac{b+e}{v_2}-n\dfrac{r^2(2u-2)+2r(1-u)+2u}{D}=0
\end{cases} \tag{A.8}$$

where $D = r^2(1-2v_2+v_1-u+2uv_2-uv_1)+2r(u-uv_2+v_2-1)+uv_1+2uv_2+1$.

One dominant and one codominant marker–gametic selection (Table 3)

$$
\begin{cases}
\dfrac{\partial L}{\partial r} = a\dfrac{v_g(1-2r)-u_gv_g}{r(v_g-u_gv_g)-r^2v_g+u_gv_g} \\
\quad + b\dfrac{2r(v_g+1)-(u_g-u_gv_g-2)}{1+u_gv_g+r^2(v_g+1)-r(u_g-u_gv_g-2)} \\
\quad + c\dfrac{u_g+1-2r}{r(u_g+1)-r^2} + d\dfrac{2rv_g}{r^2v_g} + e\dfrac{(1-2r)(v_g+1)}{(r-r^2)(v_g+1)} \\
\quad + f\dfrac{2}{r-1} - n\dfrac{u_g+v_g-u_gv_g-1}{(u_g+v_g)r+(u_gv_g+1)(1-r)} = 0 \\
\dfrac{\partial L}{\partial u_g} = a\dfrac{v_g(1-r)}{+u_gv_g+r(v_g-rv_g-u_gv_g)} \\
\quad + b\dfrac{(v_g-rv_g+r)}{r^2(v_g+1)+r(u_g-2-u_gv_g)+u_gv_g+1} \\
\quad + c\dfrac{r}{r(u_g+1-r)} - n\dfrac{v_g(1-r)+r}{(u_gv_g+1)(1-r)+(u_g+v_g)r} = 0 \qquad \text{(A.9)} \\
\dfrac{\partial L}{\partial v_g} = a\dfrac{(1-r)(r+u_g)}{u_gv_g+r(v_g-rv_g-u_gv_g)} \\
\quad + b\dfrac{(u_g-ru_g+r^2)}{r^2(v_g+1)+r(u_g-2-u_gv_g)+u_gv_g+1} \\
\quad + d\dfrac{r^2}{v_gr^2} + e\dfrac{(r-r^2)}{(v_g+1)(r-r^2)} - n\dfrac{u_g(1-r)+r}{(u_gv_g+1)(1-r)+(u_g+v_g)r} = 0.
\end{cases}
$$

References

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

Allard AW, Alder HL (1960) The effect of incomplete penetrance on the estimation of recombination values. Heredity 15:263–282

Bailey NTJ (1949) The estimation of linkage with differential viability, II and III. Heredity 3:220–228

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc 39:1–38

Edwards AWF (1972) Likelihood. The John Hopkins University Press, Baltimore

Fauré S, Noyer JL, Horry JP, Bakry F, Lanaud C, González de León D (1993) A molecular marker-based linkage map of diploid bananas (*Musa acuminata*). Theor Appl Genet 87:517–526

Fisher RA (1937) The design of experiments. Oliver and Boyd, Edinburgh London

Fisher RA, Balmukand B (1928) The estimation of linkage from the offspring of selfed heterozygotes. J Genet 20:79–92

Heun M, Gregorius HR (1987) A theoretical model for estimating linkage in F_2 populations with distorted single gene segregation. Biomet J 29:397–406

Lin SY, Ikehashi H, Yanagihara S, Kawashima A (1992) Segregation distortion via male gametes in hybrids between Indica and Japonica or wide-compatibility varieties in rice (*Oryza sativa*). Theor Appl Genet 84:812–818

Lorieux (1993) Cartographie des marqueurs moléculaires et distorsions de ségrégation: modèles mathématiques. Thèse de Doctorat en Sciences, Université Montpellier II, France, 135 pp.

Mangin B (1991) Construction de cartes génétiques: quelques méthodes. In: Méribel 91, Méribel, France, 1–4

Pham JL, Glaszmann JC, Sano R, Barbier P, Ghesquière A, Second G (1990) Isozyme markers in rice: genetic analysis and linkage relationships. Genome 33:348–359

Wolfram S (1988) Mathematica, a system for doing mathematics by computer. Addison-Wesley Publishing Company, Inc., Redwood City, California

Wu CJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11:95–103