DOI: 10.1534/genetics.107.078956

Note

Precision of Recombination Frequency Estimates After Random Intermating With Finite Population Sizes

Matthias Frisch and Albrecht E. Melchinger[1]

Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

Manuscript received July 17, 2007
Accepted for publication October 17, 2007

ABSTRACT

Random intermating of F_2 populations has been suggested for obtaining precise estimates of recombination frequencies between tightly linked loci. In a simulation study, sampling effects due to small population sizes in the intermating generations were found to abolish the advantages of random intermating that were reported in previous theoretical studies considering an infinite population size. We propose a mating scheme for intermating with planned crosses that yields more precise estimates than those under random intermating.

MARKER applications such as marker-assisted backcrossing, marker-assisted selection, and map-based cloning require linkage maps with precise estimates of the recombination frequency r between tightly linked loci. The amount of information per individual

$$i_{\mathrm{p}} = \frac{1}{n_{\mathrm{m}}\sigma_{\mathrm{r}}^2} \qquad (1)$$

(Mather 1936; Allard 1956), where n_{m} is the size of the mapping population and σ_{r}^2 the expected variance of the recombination frequency estimate, is a statistic to compare alternative types of mapping populations with respect to the precision of recombination frequency estimates. To obtain a high mapping precision for tightly linked loci, t times intermated F_2 mapping populations ($F_2^{(t)}$ populations) were suggested (Darvasi and Soller 1995) and developed in Arabidopsis (Liu *et al.* 1996) and maize (Lee *et al.* 2002). Liu *et al.* (1996) derived i_{p} for $F_2^{(t)}$ populations and found that i_{p} for their $F_2^{(4)}$ population was greater than that for an F_2 population if $r < 0.131$.

In their derivations, Liu *et al.* (1996) assumed random mating and infinite population sizes n_{i} in the intermating generations. However, Falke *et al.* (2006) hypothesized that for finite n_{i} sampling effects might overrule the increase in precision of estimates due to intermating. Martin and Hospital (2006) investigated estimation of recombination frequencies in recombinant inbred lines and found that maximum-likelihood estimates of r are biased if the relationship $R = g(r)$ between r and the frequency R of recombinant gametes in the mapping population is nonlinear. The bias is determined by the size n_{m} of the mapping population. For intermated populations, g is nonlinear and, hence, maximum-likelihood estimates of r from intermated populations are biased. Knowledge about the relative extent of (a) the reduction in i_{p} due to finite sizes of intermating populations and (b) the bias of recombination frequency estimates due to finite sizes of mapping populations is important to assess the actual advantage of intermated populations over F_2 base populations for linkage mapping. However, no results are available.

Our objectives were to (1) investigate with computer simulations the extent of the bias of maximum-likelihood estimates of r depending on the finite size n_{m} of the mapping population assuming random mating with population size $n_{\mathrm{i}} = \infty$ in the intermating generations, (2) investigate with computer simulations the effect of finite population sizes n_{i} in the intermating generations on the amount of information per individual i_{p} in the mapping population, and (3) propose a mating scheme for intermating with planned crosses that results in the same i_{p} values as random intermating with infinite population size.

Bias: For intermated $F_2^{(t)}$ mapping populations, the relation g between the recombination frequency r and the frequency of recombinant gametes R is

$$R = g(r) = \frac{1}{2}[1 - (1 - r)^t(1 - 2r)] \qquad (2)$$

(Darvasi and Soller 1995). Because g is nonlinear, the maximum-likelihood estimator $r^\star$ (*cf.* Bailey 1961) of

[1] *Corresponding author:* Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. E-mail: melchinger@uni-hohenheim.de

r is biased. Martin and Hospital (2006) employed a Taylor series expansion to derive a bias correction for arbitrary nonlinear g. Equation 18 of their derivations needs a correction. For $g = f^{-1}$ it should read

$$f'' = -\frac{g''}{[g']^3}. \tag{3}$$

With this modification, the general form of the bias correction according to Martin and Hospital (2006) is

$$c = -\frac{g''(r^\star)}{[g'(r^\star)]^3}\frac{g(r^\star)[1 - g(r^\star)]}{2n_\mathrm{m}}, \tag{4}$$

where g' and g'' are the first and second derivatives of g with respect to r. The bias-corrected estimator is then

$$\hat{r} = r^\star - c. \tag{5}$$

For $\mathrm{F}_2^{(t)}$ mapping populations, it can be calculated by using

$$g'(r) = \frac{(1-r)^t t(1-2\,r)}{2(1-r)} + (1-r)^t \tag{6}$$

and

$$g''(r) = -\frac{(1-r)^t t^2(1-2\,r)}{2(1-r)^2} + \frac{(1-r)^t t(1-2\,r)}{2(1-r)^2} - \frac{2(1-r)^t t}{1-r}. \tag{7}$$

We conducted simulations with Plabsoft (Maurer *et al.* 2008) to investigate the extent of the bias of $r^\star$ and $\hat{r}$ in $\mathrm{F}_2^{(1)}$ and $\mathrm{F}_2^{(5)}$ mapping populations of size $n_\mathrm{m} = 50$, 100, 500, 100, and 5000, employing large population sizes $n_\mathrm{i} = 25{,}000$ in the intermating generations. For each n_m we simulated 50,000 mapping populations in which $r^\star$ and $\hat{r}$ were estimated for locus pairs with map distances $r = 0.01, 0.02, \ldots, 0.39, 0.40$. From the 50,000 simulated mapping experiments, the bias of $r^\star$ and $\hat{r}$ was estimated as $\bar{\hat{r}} - r$ and $\overline{r^\star} - r$.

For large population sizes ($n_\mathrm{m} \geq 500$) and small recombination frequencies ($r < 0.1$), the bias of $r^\star$ was $<10^{-4}$ in the $\mathrm{F}_2^{(1)}$ and $<3 \times 10^{-4}$ in the $\mathrm{F}_2^{(5)}$ mapping populations (Figure 1). However, for small populations ($n_\mathrm{m} = 50$, 100) and $r = 0.1$ the bias amounted to 10^{-3} and 4×10^{-3} in the $\mathrm{F}_2^{(1)}$ and $\mathrm{F}_2^{(5)}$ mapping populations, respectively. Its absolute value was reduced efficiently by the bias correction. For example, for $n_\mathrm{m} = 50$ and $r = 0.05$ the bias of $r^\star$ in the $\mathrm{F}_2^{(1)}$ was 3.6×10^{-4} and that of $\hat{r}$ was -1.2×10^{-4}. In the $\mathrm{F}_2^{(5)}$ mapping population the bias of $r^\star$ was 10^{-3} and that of $\hat{r}$ was -7×10^{-5}. For $n_\mathrm{m} = 50$ and recombination frequencies >0.1, the bias of $r^\star$ was considerable, reaching its maximum value of ~0.04 in the interval $0.2 < r < 0.3$. For recombination frequencies $r > 0.25$ the bias correction resulted in a serious overcorrection (Figure 1).

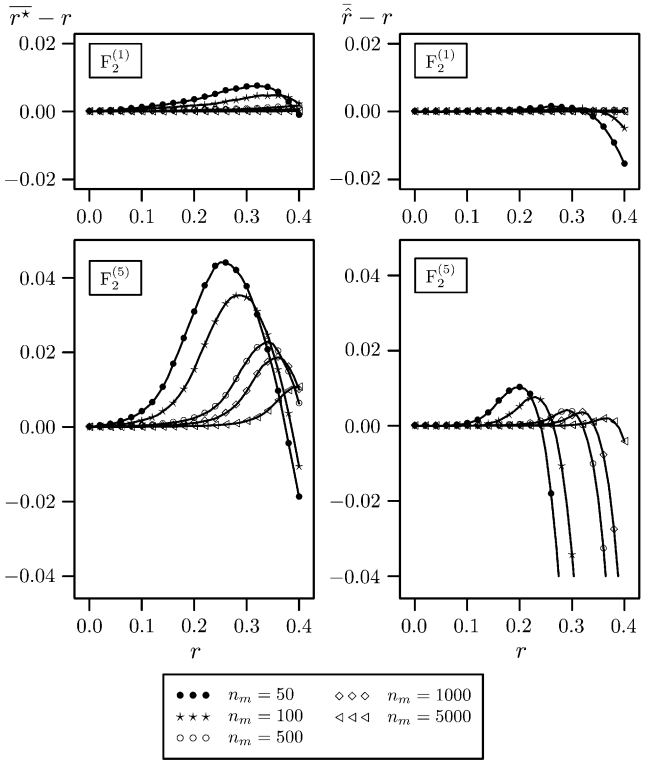

Figure 1.—Estimates of the bias $\overline{r^\star} - r$ of the maximum-likelihood estimator $r^\star$ (left) and the bias $\bar{\hat{r}} - r$ of the bias-corrected estimator $\hat{r}$ (right) for $\mathrm{F}_2^{(1)}$ (top) and $\mathrm{F}_2^{(5)}$ (bottom) mapping populations depending on the recombination frequency r. In the intermating generations, populations sizes $n_\mathrm{i} = 25{,}000$ were used. Sizes of the mapping populations were $n_\mathrm{m} = 50$, 100, 500, 1000, and 5000.

The goal of using intermated mapping populations is to increase the precision of recombination frequency estimates for tightly linked loci. Therefore, the properties of an estimator must be favorable for small values of r. For these, biasedness is not a serious problem of the maximum-likelihood estimator $r^\star$. Nevertheless, the bias correction of Martin and Hospital (2006) with the modification presented in Equation 3 provides a means to reduce the bias to negligibly small values.

Amount of information per individual: The precision of alternative types of mapping populations can be compared by expressing their i_p value as a proportion of the i_p value of an F_2 individual (Mather 1936):

$$i_\mathrm{r} = \frac{i_\mathrm{p}}{i_\mathrm{p}(\mathrm{F}_2)}. \tag{8}$$

For F_2 individuals Mather (1936) derived

$$i_\mathrm{p}(\mathrm{F}_2) = \frac{2(1 - 3r + 3r^2)}{r(1-r)(1-2r+2r^2)}, \tag{9}$$

and for $\mathrm{F}_2^{(t)}$ individuals Liu *et al.* (1996) derived

$$i_\mathrm{p} = \frac{(1-r)^{2t-2}[2(1-r) + t(1-2r)]^2[1 + 3(1-2r)^2(1-r)^{2t}]}{1 - (1-2r)^4(1-r)^{4t}}. \tag{10}$$

The derivations of Liu *et al.* (1996) assume infinite population sizes $n_{\mathrm{i}} = \infty$ in the intermating generations and, therefore, do not take into account an increase in the variance σ_{r}^2 due to sampling effects caused by finite population sizes n_{i}.

Our investigations focus on the effect of finite population sizes n_{i} in the intermating generations and a finite population size n_{m} of an $F_2^{(t)}$ mapping population on the amount of information per individual i_{p}. The effect of finite n_{i} is accounted for by carrying out simulations with finite population sizes. The effect of finite n_{m} is accounted for by using a modified definition of the information content,

$$i_{\mathrm{p}} = \frac{1}{n_{\mathrm{m}}\mathrm{MSE}_{\mathrm{r}}}, \qquad (11)$$

in which the variance σ_{r}^2 is replaced by the mean squared error $\mathrm{MSE}_r = \sigma_{\mathrm{r}}^2 + \mathrm{bias}^2$ and, hence, the effect of the bias is considered.

We investigated the effect of finite population sizes $n_{\mathrm{i}} = 100, 200, 500$ in the intermating generations on the amount of information i_{p} of individuals in the $F_2^{(1)} - F_2^{(4)}$ mapping populations of size $n_{\mathrm{m}} = 100$. For each type of mapping population and each n_{i}, we simulated 50,000 mapping populations in which $r^{\star}$ was estimated for locus pairs with map distances $r = 0.01, 0.02, \ldots, 0.29, 0.3$. From the 50,000 simulated mapping experiments, $\mathrm{MSE}_{\mathrm{r}}$ was estimated as $\overline{(r^{\star} - r)^2}$, from which i_{p} and i_{r} (Equations 11 and 8) were determined.

For $n_{\mathrm{i}} = 100$, the i_{r} values were <1 for all types of mapping populations, irrespective of the recombination frequencies r (Figure 2). For $n_{\mathrm{i}} = 200$ and 500, the i_{r} values were >1 if the recombination frequencies were $> \approx 0.05$ and ≈ 0.1, respectively. Even with $n_{\mathrm{i}} = 500$, the i_{r} values were considerably smaller than the i_{r} values for infinite population sizes $n_{\mathrm{i}} = \infty$ calculated with Equation 10 (Liu *et al.* 1996).

We conclude that the population sizes n_{i} of the intermating generations are the crucial factor for obtaining precise estimates of small r from $F_2^{(t)}$ populations. A substantial gain in precision compared to estimation of recombination frequencies from the F_2 base populations is achieved only if $n_{\mathrm{i}} \geq 500$ are employed.

Mating scheme with independent recombinations: From the assumption of infinite population sizes in the intermating generations it follows that the individuals of a mapping population do not have common ancestors in the F_2 or intermating generations. Therefore, the recombination events in different individuals of the mapping population are stochastically independent. This stochastic independence is the key property of the model with infinite population sizes in the intermating generations, for which Liu *et al.* (1996) derived the information content per individual (Equation 10). For finite population sizes and random intermating, the above property of stochastic independence does not hold, because two individuals of the mapping populations can have a common ancestor with a probability larger than zero. This increases the standard error of the recombination frequency estimate and, hence, decreases the information content i_{p} per individual. A mapping population consisting of n_{m} individuals that have no common ancestors in F_2 or later generations, *i.e.*, with stochastic independence of recombination events in different individuals, can be generated with the following planned crossing scheme and finite population sizes n_{i}.

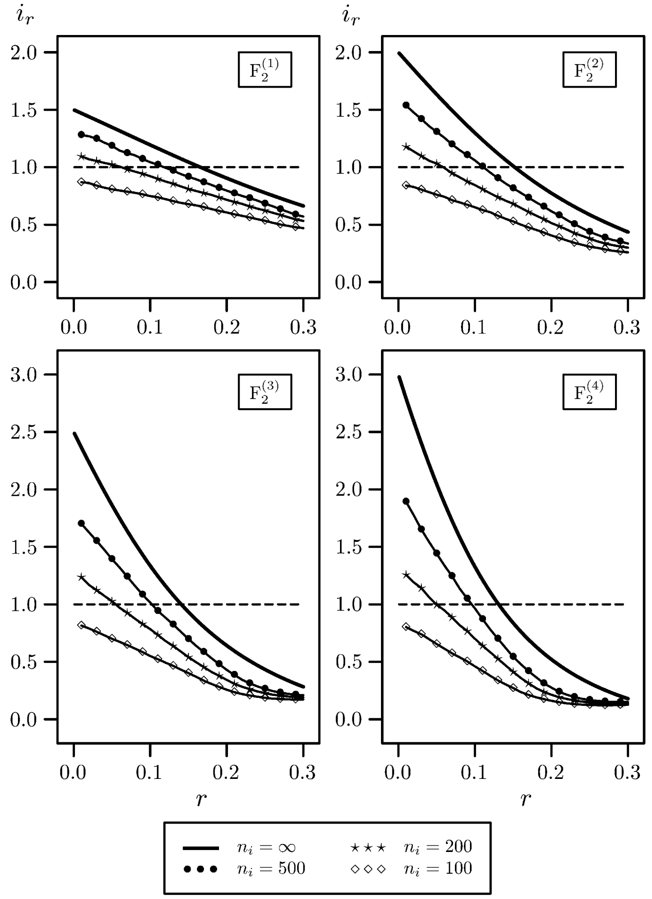

Figure 2.—Relative amount of information per individual i_{r} in $F_2^{(1)} - F_2^{(4)}$ mapping populations of size $n_{\mathrm{m}} = 100$ depending on the recombination frequency r. Simulation results are given for population sizes $n_{\mathrm{i}} = 500$, 200, and 100 in the intermating generations. For $n_{\mathrm{i}} = \infty$ the theoretical values (Equation 10) are presented.

For generating an $F_2^{(t)}$ mapping population of size n_{m}, an F_2 population of size $2^t n_{\mathrm{m}}$ is generated. Then, $2^{t-1} n_{\mathrm{m}}$ pairs of F_2 plants are crossed and from each cross one single $F_2^{(1)}$ plant is generated, resulting in an $F_2^{(1)}$ population of size $2^{t-1} n_{\mathrm{m}}$. The procedure is repeated in each subsequent generation, by producing exactly one progeny from the cross of two individuals of the parental population. Continuing the procedure for t generations results in a $F_2^{(t)}$ mapping population of size n_{m}.

Mapping populations generated with this "mating scheme with independent recombinations" have the same properties as mapping populations derived from large random-mating populations. In such populations, the amount of information i_p per individual is the same as in Equation 10. Hence, the mating scheme guarantees the maximum possible information content in the mapping population but reduces the efforts of employing large intermating populations.

We thank Frank M. Gumpert for checking the derivations, Jasmina Muminović for her editorial work, and the anonymous reviewers for helpful comments and suggestions.

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 *Mathematical Theory of Genetical Linkage*. Oxford University Press, Oxford.

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

Falke, K. C., A. E. Melchinger, C. Flachenecker, B. Kusterer and M. Frisch, 2006 Comparison of linkage maps from F2 and three times intermated generations in two populations of European flint maize (*Zea mays* L.). Theor. Appl. Genet. **113:** 857–866.

Lee, M., N. Sharopova, W. D. Beavis, D. Grant, M. Katt *et al.*, 2002 Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant Mol. Biol. **48:** 453–461.

Liu, S.-C., S.-P. Kowalski, T.-H. Lan, K. A. Feldmann and A. H. Paterson, 1996 Genome wide high resolution mapping by recurrent intermating using *Arabidopsis thaliana* as a model. Genetics **142:** 247–258.

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

Mather, K., 1936 Types of linkage data and their values. Ann. Eugen. **7:** 251–264.

Maurer, H. P., A. E. Melchinger and M. Frisch, 2008 Population genetic simulation and data analysis with Plabsoft. Euphytica (in press).

Communicating editor: R. W. Doerge