Direct fast estimator for recombination frequency in the F_2 cross

Mathias Lorieux
DIADE, University of Montpellier, CIRAD, IRD, Montpellier, France
E-mail: mathias.lorieux@ird.Fr
ORCID: 0000-0001-9864-3933
April 2021

Abstract

A new unbiased maximum-likelihood estimator is proposed for the recombination frequency in the F2 cross. The estimator is faster to calculate than with the EM algorithm, yet as efficient according to simulation studies.

Keywords: Recombination fraction, F_2 cross, EM algorithm, efficient estimator

Problem

Whole-genome sequencing allows constructing genetic maps with millions of loci. Determining the order of the loci on linkage groups requires the computation of the recombination fractions for all pairs of loci. Computing time can be extremely long, especially in the F_2 cross where an iterative procedure is usually used. A faster yet efficient estimator is thus needed.

Development

Consider two loci for which we want to know their recombination frequency *r*. In the F_2 cross, the double heterozygotes class is issued from either two recombined (AB|BA or BA|AB) or two non-recombined (AA|BB) gametes, if A and B are the parental alleles. Those two configurations are generally indistinguishable. Thus, we cannot estimate *r* directly since we need to solve the 3rd degree equation to get the efficient, unbiased maximum-likelihood estimator for *r*

$$2(c+g)+(b+d+f+h)-2\hat{r}n+2\frac{\hat{r}^2}{(1-\hat{r})^2+\hat{r}^2}e=0 \tag{1}$$

Where *c* and *g* are the counts for the double recombinants AB|AB or BA|BA, *b*, *d*, *f* and *h* are the counts for the single recombinants AA|AB, AB|AA, AB|BB and BB|AB, and *e* is the count for the double heterozygotes class. Note that the non-recombinant classes AA|AA and BB|BB, *a* and *i*, are not used since they are not informative for the estimation of *r*.

The common way to solve (1) is to use an iterative algorithm to estimate the fraction of the AB|BA or BA|AB in the double heterozygote class. The most used in this case are the Newton-Raphson (Edwards 1972) and Expectation-maximization (EM) algorithms (Dempster *et al*, 1977). For instance, an EM algorithm to estimate the fraction of double recombinants in the *e* class was proposed by Mangin (1991)

$$\hat{t}=\frac{\hat{r}^2}{(1-\hat{r})^2+\hat{r}^2} \tag{2.1}$$

$$\hat{r} = \frac{2(c+g)+(b+d+f+h)+2\hat{t}e}{2n} \tag{2.2}$$

Its asymptotic variance is

$$V_{\hat{r}} = \frac{r(1-r)(1-2r+2r^2)}{2n(1-3r+3r^2)} \tag{2.3}$$

The problem with EM is it may require several iterations to reach convergence, depending on the starting parameter value for r, making it slow to calculate. Benito and Gallego (2014) proposed a simplified, faster estimator for r (here noted as $\hat{p}$). Following our notations, this estimator is

$$\hat{p} = \frac{1}{n}\left[2(c+g)+\frac{1}{2}(b+d+f+h)\right] \tag{3.1}$$

With asymptotic variance

$$V_{\hat{p}} = \frac{p(1+p)}{2n} \tag{3.2}$$

However, this estimator has greater variance than the EM estimator, especially when r is close to 0.5, which is problematic for testing linkage.

Now, note that if a value close enough to the true r is chosen to initiate the EM process, then convergence is reached in one step, allowing to reduce it to a single estimator formula for r. One can use the unbiased Benito's estimator, $\hat{p}$, as the starting value in a one-step EM process. For this, we replace $\hat{r}$ in the t EM formula (2.1) by $\hat{r}$, then plug in t in (2.2)

$$\hat{r} = \frac{1}{2n}\left[2(c+g)+(b+d+f+h)+2e\frac{\hat{p}^{\,2}}{(1-\hat{p})^2+\hat{p}^2}\right] \tag{4}$$

Then we plug (3.1) in (4)

$$\hat{r} = \frac{1}{2n}\left[2(c+g)+(b+d+f+h)+2e\frac{(2(c+g)/n+(b+d+f+h)/2n)^2}{\left(1-(2(c+g)/n+(b+d+f+h)/2n)\right)^2+(2(c+g)/n+(b+d+f+h)/2n)^2}\right] \tag{5}$$

Let R = c + g and H = b + d + f + h. Then

$$\hat{r} = \frac{1}{2n}\left[2R+H+2e\frac{(2R/n+H/2n)^2}{\left(1-(2R/n+H/2n)\right)^2+(2R/n+H/2n)^2}\right] \tag{6}$$

The fastest way to compute $\hat{r}$ may depend on the programming language. However, multiplication is usually faster than division in computers, so one might first calculate $R, H, D = 1/2n$ and $E = 4DR + DH$, then apply the formula:

$$\hat{r} = D\left[2R+H+2e\frac{E^2}{(1-E)^2+E^2}\right] \tag{7}$$

Results

This new estimator is equivalent to the EM estimator after convergence and is therefore also unbiased and has the same variance. Moreover, it can be more precise than the EM estimator especially when few iterations are allowed. An estimator's efficiency is obtained by calculating its mean-squared error (MSE), which is the sum of its variance and the squared difference between its estimated and expected values (*i.e.*, its bias). The smaller the MSE, the greater the efficiency. Note that if an estimator is unbiased, the asymptotic MSE is simply the variance, however in a finite number of observations it is preferable to keep the observed bias in the MSE formula. Figure 1 shows plots of the MSEs and standard deviations of three estimators for *r*: EM, Benito and Gallego and formula (7) in simulated F_2 populations. We see that the new estimator is increasingly more efficient than Benito and Gallego's $\hat{p}$ as r increases. In the test carried in this study, this formula was about 30-50 times faster than the EM algorithm estimator to calculate. In Supplemental data 1, an Excel workbook is provided to replicate the simulation studies with any population size, number of repetitions and value for *r*.

Conclusion

The new proposed estimator has the desirable properties of efficiency and is considerably faster to compute than the EM algorithm. It is more efficient than Benito and Gallego's estimator. Also, in case of segregation distortion that would favor the double heterozygotes class, Benito and Gallego's estimator will rely on a smaller number of observations than expected and would have a larger standard deviation. The new estimator should thus be used in F_2 crosses when the number of loci is very large.

Conflict of interest statement: The author declared he has no conflict of interest.

References

Benito, C., & Gallego, A. (2004). A simple method for the estimation of recombination frequencies and genetic distances. *Cellular & Molecular Biology Letters*, *9*(4A), 617–634.

Dempster A.P., N.M. Laird et D.B. Rubin. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society* 39: 1-38.

Edwards. 1972. *Likelihood*. The John Hopkins University Press, Baltimore, 275 pp.

Mangin B. 1991. Construction de cartes génétiques : quelques méthodes. *In: Méribel 91*, Méribel, France, p. 1-4.

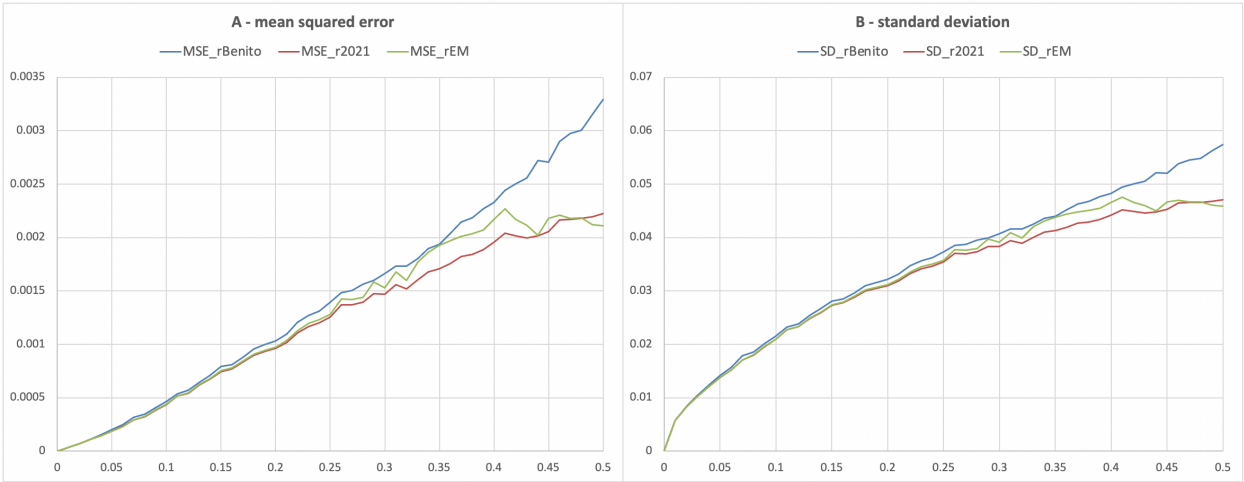

Figure 1. Mean-squared errors and standard deviations of three estimators for r, obtained from 1000,000 simulated F_2 populations of size $n = 100$. Horizontal axis: expected recombination frequency. **A:** Mean-squared errors for maximum-likelihood/EM (MSE_rEM) r, Benito and Gallego's $\hat{p}$ (MSE_rBenito) and formula (6) (MSE_r2021). **B:** standard deviations. Convergence of r_EM was stopped at 0.00001 and the initiation parameter was set to 0.25.