Direct fast estimator for recombination frequency in the F_2 cross

Mathias Lorieux
DIADE, University of Montpellier, Cirad, IRD, Montpellier, France
ORCID: 0000-0001-9864-3933
E-mail: mathias.lorieux@ird.fr

April 2021

Abstract

In this short note, a new unbiased maximum-likelihood estimator is proposed for the recombination frequency in the F_2 cross. The estimator is much faster to calculate than its EM algorithm equivalent, yet as efficient. Simulation studies are carried to illustrate the gain over another simple estimate proposed by Benito & Gallego (2004).

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

1 Introduction

Whole-genome sequencing allows constructing genetic maps with millions of loci. Unless the order of loci is known *a priori*, determining the order of the loci on linkage groups requires computing the recombination fractions for all pairs of loci, in the order of trillions. Computing time can thus be extremely long, especially in the F_2 cross where an iterative procedure is usually used by most of the genetic mapping software. A faster yet efficient estimator is proposed, which can be used as replacement for the iterative methods.

2 Methods

Consider two loci for which we want to know their recombination frequency r. In the F_2 cross, it is not possible to count the recombinant individuals directly, since the double heterozygotes class is issued from either two recombinant ($AB|BA$ or $BA|AB$) or two non-recombinant ($AA|BB$) gametes, if A and B are the parental alleles. Those two configurations are generally indistinguishable (see Appendix 1 for a detailed explanation). Thus, we cannot estimate r directly since we need to solve the 3rd degree equation to obtain the efficient, unbiased maximum-likelihood estimator for r

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

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 recombinants, $AB|BA$ or $BA|AB$, in the double heterozygote class. The most used in this case are the Newton-Raphson (see for instance 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)

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

$$\hat{r} = \frac{2(c+g) + (b+d+f+h) + 2\hat{t}e}{2n} \tag{3}$$

After convergence, we obtain the $\hat{r}_{EM}$ estimator, which asymptotic variance is

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

Iterative algorithms can require many iterations to reach convergence, depending on the starting parameter value for r, making them slow to calculate. Benito and Gallego (2014) proposed a simplified, faster estimator for r (here noted as $\hat{p}$) that ignores the double heterozygotes class. Following our notations, this estimator is

$$\hat{r}_{Benito} = \hat{p} = \frac{1}{n}\left(2(c+g) + \frac{1}{2}(b+d+f+h)\right) \tag{4}$$

With asymptotic variance

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

This estimator is unbiased. However, it has greater variance than the $\hat{r}_{EM}$ estimator, the ratio $V\hat{p}/V\hat{r}$ being

$$\frac{(1+r)(1-3r+3r^2)}{(1-r)(1-2r+2r^2)}$$

This ratio increases with r, which is problematic for testing linkage ($H_0 : r = 0.5$).

Now, note that if a value close enough to the true r is chosen to initiate the EM process, then convergence is reached in very few steps. Even the first step gives a better estimation than the initialization value. We can thus try to reduce the EM process to a direct estimator for r, that is, without having to iterate. 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 $\hat{t}$ EM formula (2) by $\hat{p}$

$$\hat{r}_{2021} = \frac{1}{2n}\left(2(c+g) + (b+d+f\ \ +h) + 2e\frac{\hat{p}^2}{(1-\hat{p})^2 + \hat{p}^2}\right) \tag{5}$$

then we plug in (4) in (5)

$$\begin{aligned}\hat{r}_{2021} = \frac{1}{2n}&\left(2(c+g) + (b+d+f+h)\right.\\ &\left.+2e\frac{(2(c+g)/n + (b+d+f+h)/2n)^2}{(1-(2(c+g)/n + (b+d+f+h)/2n))^2 + (2(c+g)/n + (b+d+f+h)/2n)^2}\right)\end{aligned}$$

Let $R = c+g$ and $H = b+d+f+h$ in (4). Then

$$\hat{r}_{2021} = \frac{1}{2n}\left(2R + H + 2e\frac{(2R/n + H/2n)^2}{(1-(2R/n + H/2n))^2 + (2R/n + H/2n)^2}\right) \tag{6}$$

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 want to pre-calculate R, H, $D = 1/2n$, $E = 4DR + DH$ and $F = E^2$ then apply the formula

$$\hat{r}_{2021} = D\left(2R + H + 2e\frac{F}{1-2E+2F}\right) \tag{7}$$

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. Monte-Carlo simulation studies were carried to verify the properties of $\hat{r}_{2021}$ compared with $\hat{r}_{16g}$ (defined in Appendix 1), $\hat{r}_{EM}$, $\hat{r}_{Benito}$. For this, for each value of r (0-0.5, step 0.01) 1,000,000 F_2 populations were simulated and the four estimators calculated. Convergence of $\hat{r}_{EM}$ was stopped using 0.00001 as stop criterion to ensure comparable precision with the two other estimators, and $\hat{p}$ as initialization parameter.

Computing times for $\hat{r}_{2021}$ and $\hat{r}_{EM}$ were compared by the same simulation strategy, this time varying the initialization parameter (0-0.5, step 0.01). The only difference is that 500,000 populations were used.

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.

3 Results

The new estimator $\hat{r}_{2021}$ is equivalent to $\hat{r}_{EM}$ after one iteration as it is unbiased and has approximately the same asymptotic variance.

Figure 1 shows plots of the MSEs and standard deviations of four estimators for r: $\hat{r}_{16g}$, $\hat{r}_{EM}$, $\hat{r}_{Benito}$ or $\hat{p}$ and formula (7) $\hat{r}_{2021}$ in simulated F_2 populations. We see that the relative efficiency of the new estimator over Benito and Gallego's $\hat{p}$ increases continuously with r.

Moreover, $\hat{r}_{2021}$ can be more precise than $\hat{r}_{EM}$ (Figure 1C), especially when an inadequate starting value is chosen and few iterations are allowed to speed-up the computation.

Regarding computing time, in the simulation study Formula (7) was 3 to 17.5 times faster than the EM algorithm estimator to compute in 95% of the cases (Figure 2). The relative computation time depended on both the values of r and the initialization value r_{Start}. Computing time of course also depends on the stop criterion set for the convergence.

4 Discussion

The new proposed estimator, $\hat{r}_{2021}$, has the desirable properties of efficiency — it is unbiased and has minimal variance — and is considerably faster to compute than the EM algorithm. As $\hat{r}_{EM}$, it is more efficient than Benito and Gallego's estimator, especially for r values greater than 0.3.

Also, it should be noted that in case of segregation distortion that, for instance, would favor the double heterozygotes class e, Benito and Gallego's estimator would rely on a smaller number of observations than expected and would have a larger standard deviation. This is not the case for $\hat{r}_{EM}$ and $\hat{r}_{2021}$.

Using the new estimator $\hat{r}_{2021}$ in lieu of r_{EM} in F_2 crosses when the number of loci is very large could help researcher computing genetic maps in less time, especially in "orphan" species that do not benefit from a reference sequence. $\hat{r}_{2021}$ can also replace r_{Benito} in most cases.

5 References

Benito, C., and Gallego, A. (2004). A simple method for the estimation of recombination frequencies and genetic distances. Cellular and 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.

6 Figures

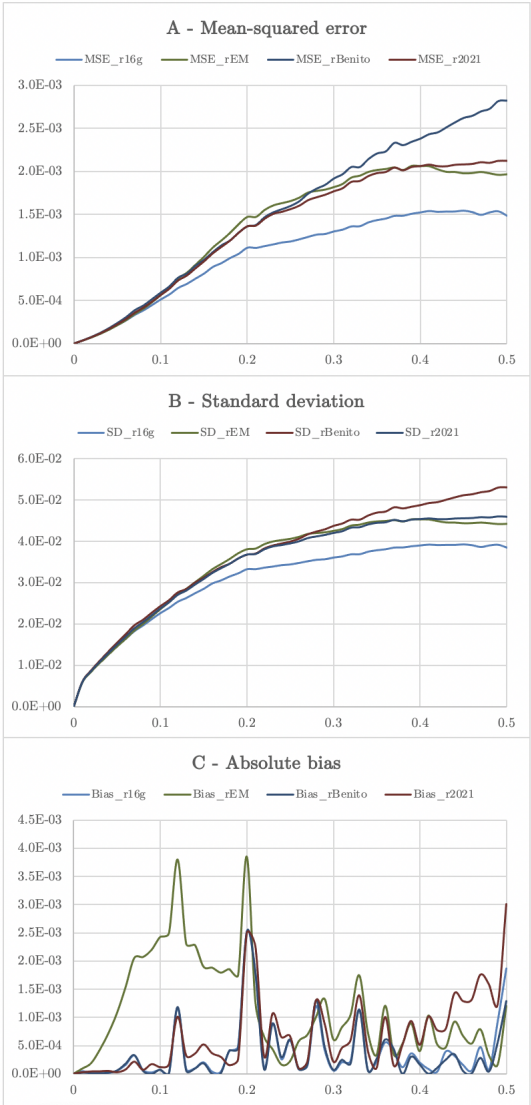

Figure 1: Mean-squared errors and standard deviations of three estimators for r, obtained by Monte-Carlo simulation of 1,000,000 F_2 populations of size $n = 100$. Horizontal axis: expected recombination frequency r. A: Mean-squared errors for maximum-likelihood 16 genotypes (MSE_r16g, defined in Appendix 1), maximum-likelihood/EM (MSE_rEM) r, Benito and Gallego's $\hat{p}$ (MSE_rBenito) and Formula (7) (MSE_r2021). B: standard deviations. C: Absolute bias. Convergence of $\hat{r}_{EM}$ was stopped at 0.00001 and $\hat{p}$ was used as initialization parameter.

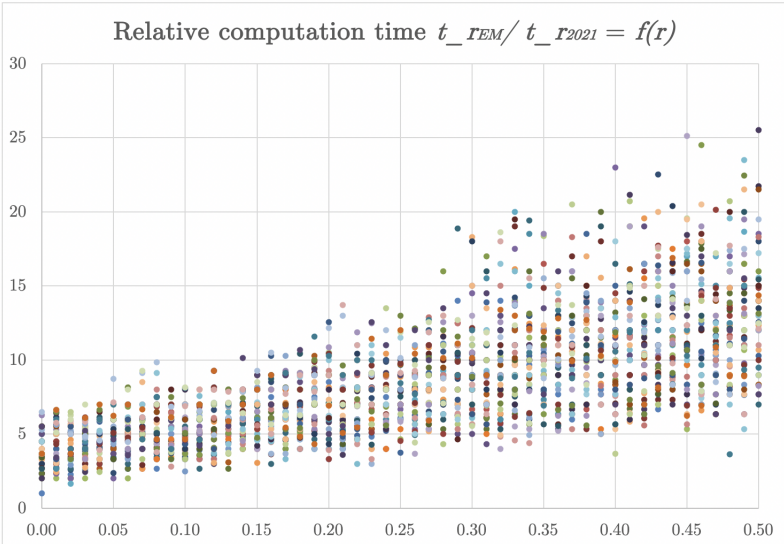

Figure 2: Relative computation time t_{rEM}/t_{r2021}, obtained by simulation of 500,000 F_2 populations of size $n = 100$. Horizontal axis: expected recombination frequency r. Convergence of $\hat{r}_{EM}$ was stopped at 0.00001. Each dot color represents a different value for the initialization parameter, which varied between 0 and 0.5.

7 Appendix 1

Details of the estimation of recombination fraction in the F_2 cross (two codominant loci, two alleles)

We consider the case of two loci segregating in an F_2 progeny. If A and B are the parental alleles, 16 genotypes are produced in the F_2 (Table A1).

Table A1: The true genotypes (female/male): 16 genotypes

	Male gamete	*non-recombinant*	*non-recombinant*	*recombinant*	*recombinant*
	Female gamete	**AA**	**BB**	**AB**	**BA**
non-recombinant	**AA**	AA|AA	AA|BB	AA|AB	AA|BA
non-recombinant	**BB**	BB|AA	BB|BB	BB|AB	BB|BA
recombinant	**AB**	AB|AA	AB|BB	AB|AB	AB|BA
recombinant	**BA**	BA|AA	BA|BB	BA|AB	BA|BA

In pure lines crosses, there are only two possible alleles, so only 9 genotypes can be observed (Table A2).

Table A2: What we see: 9 genotypes

	Male gamete	*non-recombinant*	*non-recombinant*	*recombinant*	*recombinant*
	Female gamete	**AA**	**BB**	**AB**	**BA**
non-recombinant	**AA**	A|A (1)	H|H (5)	A|H (2)	H|A (4)
non-recombinant	**BB**	H|H (5)	B|B (9)	H|B (6)	B|H (8)
recombinant	**AB**	A|H (2)	H|B (6)	A|B (3)	H|H (5)
recombinant	**BA**	H|A (4)	B|H (8)	H|H (5)	B|A (7)

In the double heterozygotes class $H|H$, the double recombinants cannot be distinguished from the non-recombinants (Table A3).

Table A3: Observable vs. true genotypes

#	***Observed genotype***	***True genotype (DR)***	***True genotype (SR)***	***True genotype (NR)***	
1	A|A			AA|AA	*distinguishable*
2	A|H		AB|AA, AA|AB		*distinguishable*
3	A|B	AB|AB			*distinguishable*
4	H|A		AA|BA, BA|AA		*distinguishable*
5	H|H	AB|BA, BA|AB		AA|BB, BB|AA	*cannot distinguish DR and NR*
6	H|B		BB|AB, AB|BB		*distinguishable*
7	B|A	BA|BA			*distinguishable*
8	B|H		BB|BA, BA|BB		*distinguishable*
9	B|B			BB|BB	*distinguishable*

DR: double recombinant; SR: single recombinant; NR: non-recombinant

Now let the recombination frequency between the two loci be r. Thus, the probabilities of the 16 possible genotypes are given in Table A4, they are derived from the gamete probabilities (Table A5).

If we could distinguish all the classes (for instance, if the male and female parents had distinguish-

Table 4: Genotypes probabilities or expected frequencies of the 16 true genotypes

#	True genotype	Type	p(NR)	p(DR)	p(SR)	Sum	Name class
1	AA|AA	NR|NR = NR	$(1 - r)^2/4$			$(1 - r)^2/4$	A
2	AA|BB	NR|NR = NR	$(1 - r)^2/4$			$(1 - r)^2/4$	B
3	BB|AA	NR|NR = NR	$(1 - r)^2/4$			$(1 - r)^2/4$	C
4	BB|BB	NR|NR = NR	$(1 - r)^2/4$			$(1 - r)^2/4$	D
5	AB|AB	R|R = DR		$r^2/4$		$r^2/4$	E
6	AB|BA	R|R = DR		$r^2/4$		$r^2/4$	F
7	BA|AB	R|R = DR		$r^2/4$		$r^2/4$	G
8	BA|BA	R|R = DR		$r^2/4$		$r^2/4$	H
9	AA|AB	NR|R = SR			$r(1 - r)/4$	$r(1 - r)/4$	I
10	AA|BA	NR|R = SR			$r(1 - r)/4$	$r(1 - r)/4$	J
11	BB|AB	NR|R = SR			$r(1 - r)/4$	$r(1 - r)/4$	K
12	BB|BA	NR|R = SR			$r(1 - r)/4$	$r(1 - r)/4$	L
13	AB|AA	R|NR = SR			$r(1 - r)/4$	$r(1 - r)/4$	M
14	AB|BB	R|NR = SR			$r(1 - r)/4$	$r(1 - r)/4$	N
15	BA|AA	R|NR = SR			$r(1 - r)/4$	$r(1 - r)/4$	O
16	BA|BB	R|NR = SR			$r(1 - r)/4$	$r(1 - r)/4$	P
		Sum	$(1 - r)^2$	r^2	$2r(1 - r)$	1	n

Table A5: Gametes probabilities (male or female)

	Gamete genotype	Probability	Explanation
non-recombinant	AA	$(1 - r)/2$	p (A) = 0.5 and p (non recombinant) = $1 - r$
non-recombinant	BB	$(1 - r)/2$	p (B) = 0.5 and p (non recombinant) = $1 - r$
recombinant	AB	$r/2$	p (A) = 0.5 and p (recombinant) = r
recombinant	BA	$r/2$	p (B) = 0.5 and p (recombinant) = r

able alleles), then the estimator for r would be:

$$\hat{r}_{16g} = \frac{E + F + G + H + (I + J + K + L + M + N + O + P)/2}{n}$$

which is the estimator of the average of male and female recombination frequencies, with asymptotic variance

$$V_{\hat{r}} = \frac{r(1 - r)}{2n}$$

The expected frequency of SR-equivalents is 2r. This is because the F_2 progeny results from two meioses, while in the backcross only one meiosis is observed. Note that we can estimate female and male recombination separately

$$r_{female} = \frac{E + F + G + H + M + N + O + P}{n}$$

$$r_{male} = \frac{E + F + G + H + I + J + K + L}{n}$$

Now, consider the probability table for the 9 observable genotypes (Table A6).

We see that the male and female recombination frequencies cannot be estimated separately since the M, N, O and P classes are indistinguishable from the classes I, J, K and L. We can however estimate an average r in simply counting the observable recombinants:

$$\hat{p} = \frac{2(c + g) + (b + d + f + h)/2}{n}$$

Table A6: Genotypes probabilities or expected frequencies of the 9 observable genotypes

#	*Observed genotype*	*Name observed class*	*Corresponding true class*	*True genotype (DR)*	*True genotype (SR)*	*True genotype (NR)*	*p(observed genotype)*	Try an r value: 0.1 *Expected frequency = f(r)*	Try a n value: 100 *Expected count*
1	A|A	a	A			$(1 - r)^2/4$	$(1 - r)^2/4$	0.2025	20.3
2	A|H	b	I + M		$r(1 - r)/4 + r(1 - r)/4$		$r(1 - r)/2$	0.0450	4.5
3	A|B	c	E	$r^2/4$			$r^2/4$	0.0025	0.3
4	H|A	d	J + O		$r(1 - r)/4 + r(1 - r)/4$		$r(1 - r)/2$	0.0450	4.5
5	H|H	e	B + C + F + G	$r^2/4 + r^2/4$		$(1 - r)^2/4 + (1 - r)^2/4$	$r^2/2 + (1 - r)^2/2$	0.4100	41.0
6	H|B	f	K + N		$r(1 - r)/4 + r(1 - r)/4$		$r(1 - r)/2$	0.0450	4.5
7	B|A	g	H	$r^2/4$			$r^2/4$	0.0025	0.3
8	B|H	h	L + P		$r(1 - r)/4 + r(1 - r)/4$		$r(1 - r)/2$	0.0450	4.5
9	B|B	i	D			$(1 - r)^2/4$	$(1 - r)^2/4$	0.2025	20.3
	Sum	n	n	r^2	$2r(1 - r)$	$(1 - r)^2$	1	1	100
	alternative formula	n	n	r^2	$2r - 2r^2$	$1 - 2r + r^2$	1	1	100

$\hat{p}$ is the Benito & Gallego's (2014) estimator, which is unbiased as shown in Table A7. Its variance is

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

Table A7: estimating r using the distinguishable classes only

#	*Observed genotype*	*Name observed class*	Try an r value: 0.5 *Expected frequency = f(r)*	Try a n value: 300 *Expected count*
1	A|A	a	0.0625	18.8
2	A|H	b	0.1250	37.5
3	A|B	c	0.0625	18.8
4	H|A	d	0.1250	37.5
5	H|H	e	0.2500	75.0
6	H|B	f	0.1250	37.5
7	B|A	g	0.0625	18.8
8	B|H	h	0.1250	37.5
9	B|B	i	0.0625	18.8
		n	1	300

Result: $p = 0.5$

However, if ignoring the $H|H$ class does not affect the informative sample size when r is small (< 0.3), this results in the loss of 12.5% of informative recombinants when $r = 0.5$.

This affects the variance of the estimator for r (Table A8).

The way to take the $H|H$ class into account is to solve the equation

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

This can be attained using an iterative process like the Newton-Raphson or the EM algorithms. For instance, Mangin (1991) proposed the following EM:

Table A8: How do p(DR) and p(NR) within the H/H class vary with r?

r	p(H|H)	p(DR)	p(NR)	% (DR) in H|H	% (NR) in H|H	*Var(r) (full information is used) = Var1*	*Var(r) (distinguishable genotypes only) = Var2*	*Var2/Var1*
0	0.500	0.000	0.500	0.000	1.000	0.000	0.000	1.000
0.05	0.453	0.001	0.451	0.003	0.997	0.025	0.026	1.047
0.1	0.410	0.005	0.405	0.012	0.988	0.051	0.055	1.088
0.15	0.373	0.011	0.361	0.030	0.970	0.077	0.086	1.121
0.2	0.340	0.020	0.320	0.059	0.941	0.105	0.120	1.147
0.25	0.313	0.031	0.281	0.100	0.900	0.134	0.156	1.167
0.3	0.290	0.045	0.245	0.155	0.845	0.165	0.195	1.185
0.35	0.273	0.061	0.211	0.225	0.775	0.195	0.236	1.210
0.4	0.260	0.080	0.180	0.308	0.692	0.223	0.280	1.256
0.45	0.253	0.101	0.151	0.401	0.599	0.243	0.326	1.344
0.5	0.250	0.125	0.125	0.500	0.500	0.250	0.375	1.500

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

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

This provides, after convergence, a maximum-likelihood, unbiased and efficient estimator for r, with variance

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