

# A note on the estimation of recombination fractions in an intercross

## The Problem

Certain mutations in a mouse gene are believed to be responsible for a recessive phenotypic trait. Let  $A$  and  $a$  be the wildtype and mutant alleles for the phenotypic trait. Let  $B$  and  $b$  be the wildtype and mutant alleles for the candidate gene. Crossing a wildtype mouse ( $AB/AB$ ) with a mutant mouse ( $ab/ab$ ) produced progeny mice (F1) which were all heterozygous for both loci and phenotypically normal ( $AB/ab$ ). An intercross with F1 mice gave twenty seven offspring (F2), which fell into six categories as given in Table 1.

	BB	Bb	bb
$A^*$	8	12	0
$a^*$	0	0	7

Table 1: Observed frequencies from F2 intercross.

Here  $A^*$  and  $a^*$  are the wildtype and mutant phenotypes. Hence  $A^*$  mice have either the  $AA$  or  $Aa$  genotypes, which are indistinguishable. The genotypic composition at the  $B$  locus can be distinguished. The problem is to estimate the maternal and paternal recombination fractions between the two loci. A recent paper by Blixt, Mahlapuu, Aitola, Pelto-Huikko, Enerbäck and Carlsson [1] has data with the same structure. The candidate gene approach is routinely used to find genes in mouse genetics; see Silver [2]. A similar problem was considered in Fisher and Balmukand [3].

## Genotypic and phenotypic frequencies

Let  $r_m$  and  $r_p$  be the maternal and paternal recombination fractions respectively. Write  $\bar{r}_m = 1 - r_m$ ,  $\bar{r}_p = 1 - r_p$ . The genotype of an egg from an F1 female mouse is  $AB$ ,  $Ab$ ,  $aB$  or  $ab$  with probability

$$\bar{r}_m/2, r_m/2, r_m/2, \bar{r}_m/2,$$

respectively. The genotype of a sperm from an F1 male mouse has a similar distribution, with  $p$  replacing the subscript  $m$ . Assuming independence, the genotypic frequencies of F2 mice are obtained by multiplying probabilities. These are presented in the Punnett square. The top row shows the paternal gametes and the first column shows the maternal gametes. Thus, about  $\bar{r}_m\bar{r}_p/4$  of the F2 mice have genotype  $AB/ab$  and get the chromosome with  $AB$  from the mother and that with  $ab$  from the father.  $\bar{r}_m\bar{r}_p/4$  of the F2 mice have the same genotype but the parental origins of the chromosomes are switched. These two classes are not distinguishable.

The relative frequencies of the observable phenotypic classes,  $p_1, p_2, \dots, p_6$ , numbered across the rows of Table 2, can be derived from the Punnett square by appropriate aggregating.

Haploid genotype	AB	Ab	aB	ab
AB	$\bar{r}_m \bar{r}_p / 4$	$\bar{r}_m r_p / 4$	$\bar{r}_m r_p / 4$	$\bar{r}_m \bar{r}_p / 4$
Ab	$r_m \bar{r}_p / 4$	$r_m r_p / 4$	$r_m r_p / 4$	$r_m \bar{r}_p / 4$
aB	$r_m \bar{r}_p / 4$	$r_m r_p / 4$	$r_m r_p / 4$	$r_m \bar{r}_p / 4$
ab	$\bar{r}_m \bar{r}_p / 4$	$\bar{r}_m r_p / 4$	$\bar{r}_m r_p / 4$	$\bar{r}_m \bar{r}_p / 4$

Table 2: Expected genotypic frequencies.

We know that the following equalities hold:

$$p_1 + p_2 + p_3 = P(A^*) = 3/4,$$

$$p_4 + p_5 + p_6 = P(a^*) = 1/4,$$

$$p_1 + p_4 = P(BB) = 1/4,$$

$$p_2 + p_5 = P(Bb) = 1/2,$$

$$p_3 + p_6 = P(bb) = 1/4.$$

To get the six probabilities in terms of the recombination fractions, we just need to calculate two of them:

$$p_4 = P(a^*BB) = \frac{1}{4} r_m r_p.$$

$$p_6 = P(a^*bb) = \frac{1}{4} \bar{r}_m \bar{r}_p.$$

The rest are easily obtained:

$$p_1 = 1/4 - p_4 = \frac{1}{4}(1 - r_m r_p),$$

$$p_3 = 1/4 - p_6 = \frac{1}{4}(1 - \bar{r}_m \bar{r}_p),$$

$$p_2 = \frac{3}{4} - p_1 - p_3 = \frac{1}{4}(1 + r_m r_p + \bar{r}_m \bar{r}_p),$$

$$p_5 = 1/2 - p_2 = \frac{1}{4}(r_m \bar{r}_p + \bar{r}_m r_p).$$

It turns out that these probabilities depend on the recombination fractions only through the two products  $\phi = r_m r_p$  and  $\theta = \bar{r}_m \bar{r}_p$ . We have the following phenotypic frequency table.

Note that if the two loci segregate independently, i.e.,  $r_m = r_p = 1/2$ , then  $\phi = \theta = 1/4$ . At the other extreme, if  $r_m = r_p = 0$ , then  $\phi = 0$  and  $\theta = 1$ .

	BB	Bb	bb
A*	$(1 - \phi)/4$	$(1 + \theta + \phi)/4$	$(1 - \theta)/4$
a*	$\phi/4$	$(1 - \theta - \phi)/4$	$\theta/4$

Table 3: Expected phenotypic frequencies.

### Estimation

A standard method to estimate the recombination fractions from the observed frequencies is via the maximum likelihood principle. Let  $n_i$  be the observed frequencies in the classes  $i = 1, \dots, 6$  labeled as the  $p_i$ 's, and denote the sum by  $n$ . The probability of observing these numbers is proportional to

$$L(r_m, r_p) = p_1^{n_1} \cdots p_6^{n_6}.$$

This expression is known as the likelihood function and we seek a point  $(\hat{r}_m, \hat{r}_p)$  in the square  $[0, 1/2] \times [0, 1/2]$  so that  $L(\hat{r}_m, \hat{r}_p)$  is larger than at any other point in the same set. This is equivalent to maximizing the logarithm of the likelihood, which is often simpler to manipulate. The loglikelihood is

$$\begin{aligned} LL(r_m, r_p) = & n_1 \log(1 - r_m r_p) + n_2 \log(1 - \bar{r}_m \bar{r}_p) + n_3 \log(1 + r_m r_p + \bar{r}_m \bar{r}_p) \\ & + n_4 \log(r_m r_p) + n_5 \log(\bar{r}_m \bar{r}_p) + n_6 \log(r_m \bar{r}_p + \bar{r}_m r_p) - n \log 4. \end{aligned}$$

Operationally, one may directly maximize the loglikelihood function  $LL(r_m, r_p)$  with calculus or use an iterative procedure such as the EM algorithm. Alternatively, one can view the loglikelihood as a function of the transformed parameter  $(\phi, \theta)$ , in which case, one maximizes the expression

$$\begin{aligned} LL(\phi, \theta) = & n_1 \log(1 - \phi) + n_2 \log(1 - \theta) + n_3 \log(1 + \phi + \theta) + \\ & n_4 \log \phi + n_5 \log \theta + n_6 \log(1 - \phi + \theta) - n \log 4, \end{aligned}$$

over the subset defined by

$$0 \leq \phi \leq 1/4, \quad 1/2 - \phi \leq \theta \leq (1 - \sqrt{\phi})^2.$$

Details about the EM algorithm and the constraints are discussed in the appendix.

### Discussion of special case

For the present problem, it is simpler to work with the new parameters  $\phi$  and  $\theta$ . The loglikelihood for  $n_1 = 8$ ,  $n_2 = 12$ ,  $n_3 = 0$ ,  $n_4 = 0$ ,  $n_5 = 0$ ,  $n_6 = 7$ , is

$$LL(\phi, \theta) = -27 \log 4 + 8 \log(1 - \phi) + 12 \log(1 + \phi + \theta) + 7 \log \theta.$$

To get the global maximum, we first let  $\phi$  be some fixed number in  $[0, 1/4]$ . Notice that on this strip of the parameter space, the loglikelihood is monotone increasing in  $\theta$ , so it is maximum at a  $\theta(\phi)$  given by

$$\theta(\phi) = (1 - \sqrt{\phi})^2.$$

Substituting this condition back into the loglikelihood expression, we now maximize

$$-42 \log 2 + 8 \log(1 - \phi) + 12 \log(1 + \phi - \sqrt{\phi}) + 14 \log(1 - \sqrt{\phi}),$$

which is monotone decreasing in  $\phi$  on  $[0, 1/4]$ . So the loglikelihood is maximized at

$$\hat{\phi} = 0, \hat{\theta} = 1,$$

or

$$\hat{r}_m = \hat{r}_p = 0.$$

The conclusion is that the gene and the phenotype locus are extremely close.

Geneticists are familiar with the LOD score, which is defined as the base 10 logarithm of the ratio of the probability of observed progeny to the probability of observed progeny if there were no linkage, or equivalently

$$\text{LOD} = \log_{10} \left\{ \frac{L(r_m, r_p)}{L(1/2, 1/2)} \right\}.$$

Since the true recombination fractions are unknown, we have to plug in our estimates  $\hat{r}_m = 0, \hat{r}_p = 0$ . The approximate probability of observing the progeny is proportional to

$$L(0, 0) = \left(\frac{1}{4}\right)^8 \left(\frac{1}{2}\right)^{12} \left(\frac{1}{4}\right)^7.$$

If the gene and the phenotype locus were unlinked, the probability is proportional to

$$L(1/2, 1/2) = \left(\frac{3}{16}\right)^8 \left(\frac{3}{8}\right)^{12} \left(\frac{1}{16}\right)^7.$$

So  $\text{LOD} \sim 6.7$ , indicating an odds ratio of more than one million. In genetics, an odds ratio of one thousand ( $\text{LOD} = 3$ ) is considered strong evidence for linkage.

Of course, this does not prove that the recombination fractions are zero. One may want to find an interval, such that any recombination fraction in this interval is consistent with our observations. For simplicity, let us assume both the maternal and paternal recombination fractions are equal to  $r$ . The likelihood based on the data is

$$L(r) = \{(1 - r^2)/4\}^8 \{(1 - r + r^2)/2\}^{12} \{(1 - r)^2/4\}^7.$$

Postulating that a priori  $r$  is equally likely to be any number in  $[0, 1/2]$ , the likelihood can be viewed as an updated version of the distribution of  $r$  in light of the observed data, with appropriate scaling. Then a reasonable 95% confidence interval for  $r$  is  $[0, r_{0.95}]$ , where  $r_{0.95}$  is the 95th percentile of the distribution. It is found by solving the equation

$$\int_0^{r_{0.95}} L(r) dr = 0.95 \int_0^{1/2} L(r) dr.$$

Using numerical integration,  $r_{0.95}$  is about 0.11, so that 11 cM is an approximate upper 95% confidence limit.

## Appendix

### *The constraints on $\phi$ and $\theta$*

Recall that  $\phi = r_m r_p$  and  $\theta = (1 - r_m)(1 - r_p)$  with both the recombination fractions constrained in the interval  $[0, 1/2]$ . To derive the constraints on  $\phi$  and  $\theta$ , first notice that  $\phi$  ranges from 0 to  $1/4$ . Fix  $\phi$  in  $[0, 1/4]$ , and we see that

$$\theta = 1 + \phi - (r_m + r_p), \quad r_m r_p = \phi.$$

It is readily checked that provided the product of the recombination fractions is fixed at  $\phi$ , the sum is minimum when they are equal ( $r_m = r_p = \sqrt{\phi}$ ), and is maximum when they are most unequal ( $r_m = 1/2, r_p = 2\phi$  or  $r_p = 1/2, r_m = 2\phi$ ). This gives the range of  $\theta$ :

$$\begin{aligned} \theta &\leq 1 + \phi - 2\sqrt{\phi} = (1 - \sqrt{\phi})^2, \\ \theta &\geq 1 + \phi - (1/2 + 2\phi) = 1/2 - \phi. \end{aligned}$$

Figure 1 shows the parameter space in a barycentric representation. For any point in the triangle, the first coordinate is the length of the perpendicular dropped to the “right-hand” side of the triangle. Similarly, the second and third coordinate are the lengths of perpendiculars to the “left-hand” and bottom sides. The three lengths sum to 1.

Notice that the upper bound and lower bound coincide when  $\phi = 1/4$ , so the condition  $\phi \leq 1/4$  is unnecessary. Also, the two points  $(r_m, r_p)$  and  $(r_p, r_m)$  are mapped to the same point in the  $(\phi, \theta)$ -space. This means that we can estimate the recombination fractions but will not be able to tell which is which.

The equation  $\theta = (1 - \sqrt{\phi})^2$  implies  $r_m = r_p = r$  and is the same as  $4\phi\theta = (1 - \phi - \theta)^2$ , which expresses the Hardy-Weinberg equilibrium for a di-allelic locus if  $\phi$  and  $\theta$  are respectively the frequencies of the two homozygotes. Then  $r$  is one of the allelic frequencies.

### *The EM algorithm*

This is an iterative procedure to find a local maximum of the loglikelihood  $LL(r_m, r_p)$ . In general, it is impossible to tell if the maximum found is a global maximum.

Let  $m_1, \dots, m_{16}$  be the frequencies of the sixteen cells in the Punnett table. Then we

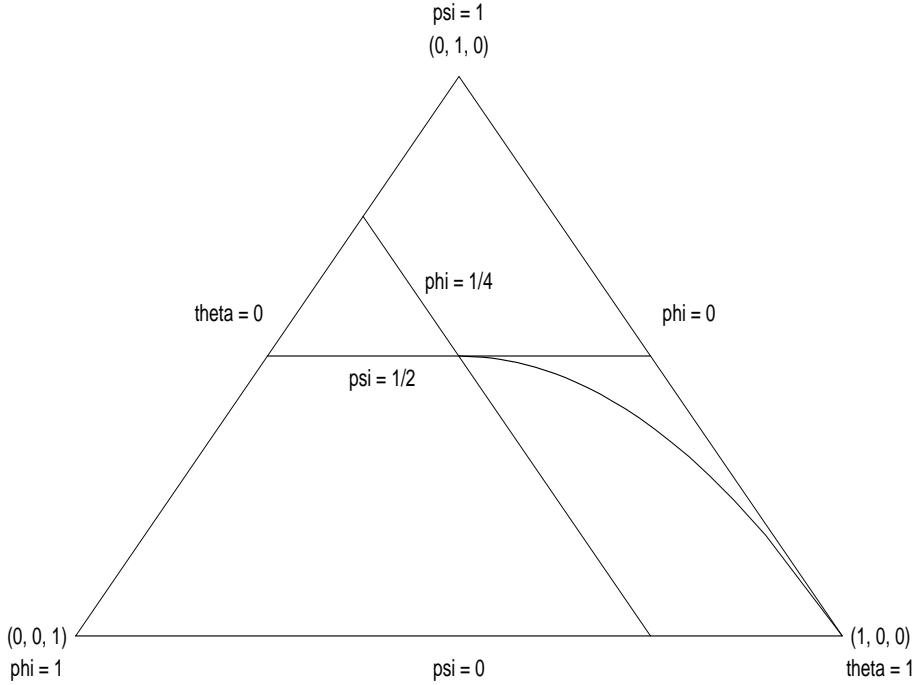


Figure 1: *Baricentric representation.* The curve is  $\theta = (1 - \sqrt{\phi})^2$ .  $\phi = \phi$ ;  $\theta = \theta$ ;  $\psi = 1 - \phi - \theta$ .

have

$$\begin{aligned}
 n &= \sum_{i=1}^{16} m_i, \\
 n_1 &= m_1 + m_3 + m_9, \\
 n_2 &= m_2 + m_4 + m_5 + m_7 + m_{10} + m_{13}, \\
 n_3 &= m_6 + m_8 + m_{14}, \\
 n_4 &= m_{11}, \\
 n_5 &= m_{12} + m_{15}, \\
 n_6 &= m_{16}.
 \end{aligned}$$

The  $n_i$ 's are known; the  $m_i$ 's are not observable. To get started, one has to choose some

values for the recombination fractions, say,  $r_{m0}, r_{p0}$ . The following two steps are repeated until one is satisfied that the successive estimates have settled down.

**Imputing.** Calculate the complete data by setting

$$\hat{m}_i = E_{r_{m0}, r_{p0}}(m_i | n_1, n_2, n_3, n_4, n_5, n_6).$$

For example,

$$\begin{aligned} \hat{m}_1 &= E_{r_{m0}, r_{p0}}(m_1 | n_1) \\ &= (P(m_1) + P(m_3) + P(m_9)) / P(n_1) \times n_1 \\ &= (\bar{r}_{m0}\bar{r}_{p0} + \bar{r}_{m0}r_{p0} + r_{m0}\bar{r}_{p0}) / (1 - r_{m0}r_{p0}) \times n_1, \\ \hat{m}_{16} &= E_{r_{m0}, r_{p0}}(m_{16} | n_6) \\ &= P(m_{16}) / P(n_6) \times n_6 \\ &= n_6. \end{aligned}$$

**Estimation.** Obtain the maximum likelihood estimates of  $r_m$  and  $r_p$  based on the imputed  $\hat{n}$ 's:

$$\begin{aligned} r_{m1} &= (\hat{m}_5 + \hat{m}_6 + \hat{m}_7 + \hat{m}_8 + \hat{m}_9 + \hat{m}_{10} + \hat{m}_{11} + \hat{m}_{12}) / n, \\ r_{p1} &= (\hat{p}_2 + \hat{p}_3 + \hat{p}_6 + \hat{p}_7 + \hat{p}_{10} + \hat{p}_{11} + \hat{p}_{14} + \hat{p}_{15}) / n. \end{aligned}$$

If  $(r_{m1}, r_{p1})$  is close to  $(r_{m0}, r_{p0})$ , stop. If not, set  $(r_{m0}, r_{p0}) = (r_{m1}, r_{p1})$  and repeat the imputing step.

For our problem, let us choose  $r_{m0} = 0, r_{p0} = 0$ . This implies no recombination. So we have

$$\begin{aligned} \hat{m}_1 &= E_{0,0}(m_1 | n_1) = n_1 = 8, \\ \hat{m}_4 &= E_{0,0}(m_4 | n_2) = n_2 / 2 = 6, \\ \hat{m}_{13} &= E_{0,0}(m_{13} | n_2) = n_2 / 2 = 6, \\ \hat{m}_{16} &= E_{0,0}(m_{16} | n_6) = n_6 = 7, \\ \hat{m}_i &= 0, \text{ all other } i. \end{aligned}$$

It is easily seen that the  $(r_{m1}, r_{p1}) = (0, 0)$  and we have reached stability in one step. This concludes the EM algorithm. A separate argument would be needed to show this is a global maximum.

## References

- [1] Å. Blixt, M Mahlapuu, M. Aitola, M. Peltto-Huikko, S. Enerbäck and P. Carlsson (2000). A forkhead gene, *FoxE3*, is essential for lens epithelial proliferation and closure of the lens vesicle. *Genes and Development* 14:245-254.
- [2] L. M. Silver (1995). *Mouse Genetics*. (Oxford University Press, New York).
- [3] R. A. Fisher and B. Balmukand (1928). The estimation of linkage from the offspring of selfed heterozygotes. *Journal of Genetics* 20:79-92.