Chapter 2
Recombination Fraction

Recombination fraction (also called recombination frequency) between two loci is defined as the ratio of the number of recombined gametes to the total number of gametes produced. Recombination fraction, denoted by r throughout the book, however, has a domain of $0 \leq r \leq 0.5$, with $r = 0$ indicating perfect linkage and $r = 0.5$ meaning complete independence of the two loci. In most situations, gametes are not directly observable. Therefore, special mating designs are required to infer the number of recombined gametes. When a designed mating experiment cannot be carried out, data collected from pedigrees can be used for estimating recombination fractions. However, inferring the number of recombined gametes in pedigrees is much more complicated than that in designed mating experiments. This book only deals with designed mating experiments.

2.1 Mating Designs

Two mating designs are commonly used in linkage study, the backcross (BC) design and the $\mathtt{F}_2$ design. Both designs require two inbred lines, which differ in both the phenotypic values of traits (if marker-trait association study is to be performed) and allele frequencies of marker loci used for constructing the linkage map. We will use two marker loci as an example to show the mating designs and methods for estimating recombination fraction. The BC design is demonstrated in Fig. 2.1. Let A and B be the two loci under investigation. Let A_1 and A_2 be the two alleles at locus A and B_1 and B_2 be the two alleles at locus B. Let $\mathtt{P}_1$ and $\mathtt{P}_2$ be the two parents that initiate the line cross. Since both parents are inbred, we can describe the two-locus genotype for $\mathtt{P}_1$ and $\mathtt{P}_2$ by $\frac{A_1B_1}{A_1B_1}$ and $\frac{A_2B_2}{A_2B_2}$, respectively. The hybrid progeny of cross between $\mathtt{P}_1$ and $\mathtt{P}_2$ is denoted by $\mathtt{F}_1$ whose genotype is $\frac{A_1B_1}{A_2B_2}$. The horizontal line in the $\mathtt{F}_1$ genotype separates the two parental gametes, i.e., A_1B_1 is the gamete from $\mathtt{P}_1$ and A_2B_2 is the gamete from $\mathtt{P}_2$. The $\mathtt{F}_1$ hybrid crosses back to one of the parents

S. Xu, *Principles of Statistical Genomics*, DOI 10.1007/978-0-387-70807-2_2,

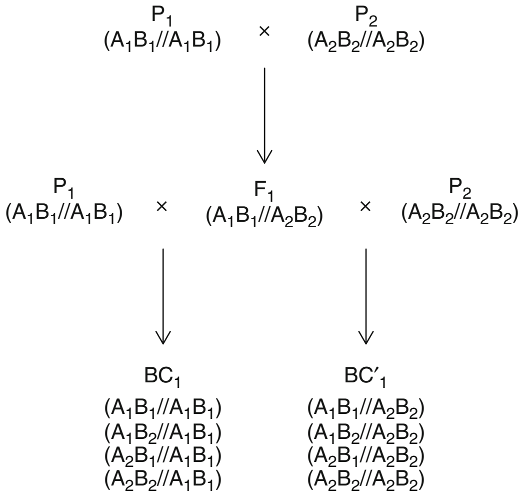

Fig. 2.1 The backcross (BC) mating design. The BC progeny generated by $\mathtt{F}_1 \times \mathtt{P}_1$ is called $\mathtt{BC}_1$, whereas the BC population generated by $\mathtt{F}_1 \times \mathtt{P}_2$ is called $\mathtt{BC}'_1$

Table 2.1 Count data of two-locus genotypes collected from a $\mathtt{BC}_1$ family

Genotype	Count	Frequency	Type
$\frac{A_1B_1}{A_1B_1}$	n_{11}	$\frac{1}{2}(1-r)$	Parental
$\frac{A_1B_2}{A_1B_1}$	n_{12}	$\frac{1}{2}r$	Recombinant
$\frac{A_2B_1}{A_1B_1}$	n_{21}	$\frac{1}{2}r$	Recombinant
$\frac{A_2B_2}{A_1B_1}$	n_{22}	$\frac{1}{2}(1-r)$	Parental

to generate multiple BC progeny, which will be used for linkage study. The BC population is a segregating population. Linkage analysis can only be conducted in such a segregating population. A segregating population is defined as a population that contains individuals with different genotypes. The two parental populations and the $\mathtt{F}_1$ hybrid population are not segregating populations because individuals within each of the three populations are genetically identical. The BC progeny generated by $\mathtt{F}_1 \times \mathtt{P}_1$ is called $\mathtt{BC}_1$, whereas the BC population generated by $\mathtt{F}_1 \times \mathtt{P}_2$ is called $\mathtt{BC}'_1$.

We now use $\mathtt{BC}_1$ progeny as an example to demonstrate the BC analysis. The gametes generated by the $\mathtt{P}_1$ parent are all of the same type A_1B_1. However, the $\mathtt{F}_1$ hybrid can generate four different gametes and thus four distinguished genotypes. Let r be the recombination fraction between loci A and B. Let n_{ij} be the number of gametes of type A_iB_j or the number of genotype of $\frac{A_iB_j}{A_1B_1}$ kind for $i, j = 1, 2$. The four genotypes and their frequencies are given in Table 2.1. This table provides the data for the maximum likelihood estimation of recombination fraction. The maximum likelihood method will be described later.

The $\mathtt{F}_2$ mating design requires mating of the hybrid with itself, called selfing and denoted by the symbol $\otimes$ (see Fig. 2.2 for the $\mathtt{F}_2$ design). When selfing is impossible, e.g., in animals and self-incompatible plants, intercross between

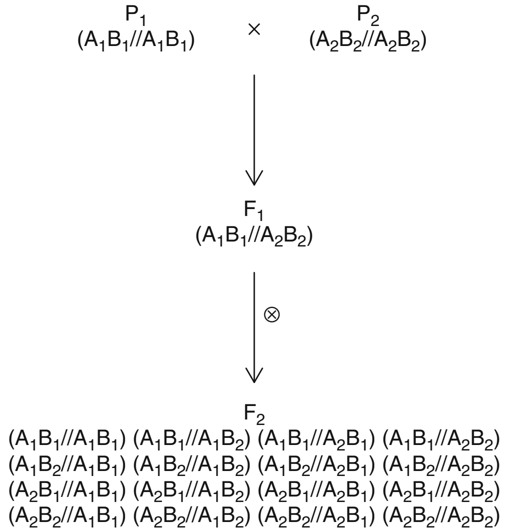
P1
(A1B1//A1B1)
×
P2
(A2B2//A2B2)
F1
(A1B1//A2B2)
⊗
F2
(A1B1//A1B1) (A1B1//A1B2) (A1B1//A2B1) (A1B1//A2B2)
(A1B2//A1B1) (A1B2//A1B2) (A1B2//A2B1) (A1B2//A2B2)
(A2B1//A1B1) (A2B1//A1B2) (A2B1//A2B1) (A2B1//A2B2)
(A2B2//A1B1) (A2B2//A1B2) (A2B2//A2B1) (A2B2//A2B2)

Fig. 2.2 The $\mathtt{F}_2$ mating design

Table 2.2 The 16 possible genotypes and their observed counts in an $\mathtt{F}_2$ family

	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_1	$\frac{A_1B_1}{A_1B_1}, n_{11}$	$\frac{A_1B_1}{A_1B_2}, n_{12}$	$\frac{A_1B_1}{A_2B_1}, n_{13}$	$\frac{A_1B_1}{A_2B_2}, n_{14}$
A_1B_2	$\frac{A_1B_2}{A_1B_1}, n_{21}$	$\frac{A_1B_2}{A_1B_2}, n_{22}$	$\frac{A_1B_2}{A_2B_1}, n_{23}$	$\frac{A_1B_2}{A_2B_2}, n_{24}$
A_2B_1	$\frac{A_2B_1}{A_1B_1}, n_{31}$	$\frac{A_2B_1}{A_1B_2}, n_{32}$	$\frac{A_2B_1}{A_2B_1}, n_{33}$	$\frac{A_2B_1}{A_2B_2}, n_{34}$
A_2B_2	$\frac{A_2B_2}{A_1B_1}, n_{41}$	$\frac{A_2B_2}{A_1B_2}, n_{42}$	$\frac{A_2B_2}{A_2B_1}, n_{43}$	$\frac{A_2B_2}{A_2B_2}, n_{44}$

different $\mathtt{F}_1$ individuals initiated from the same cross is required. The progeny of selfing $\mathtt{F}_1$ or intercross between two $\mathtt{F}_1$ hybrids is called an $\mathtt{F}_2$ progeny. An $\mathtt{F}_2$ family consists of multiple $\mathtt{F}_2$ progeny. The $\mathtt{F}_2$ family represents another segregating population for linkage analysis. Recall that an $\mathtt{F}_1$ hybrid can generate four possible gametes for loci A and B jointly. Therefore, selfing of $\mathtt{F}_1$ can generate 16 possible genotypes, as illustrated in Table 2.2. Let n_{ij} be the number of individuals combining the ith gamete from one parent and the jth gamete from the other parent, for $i, j = 1, \ldots, 4$. The frequencies of all the 16 possible genotypes are listed in Table 2.3. This table is the basis from which the maximum likelihood estimation of recombination fraction will be derived.

2.2 Maximum Likelihood Estimation of Recombination Fraction

In a BC design, the four types of gametes are distinguishable. Therefore, the recombination fraction can be directly calculated by taking the ratio of the number of recombinants to the total number of gametes. We use $\mathtt{BC}_1$ as an example to demonstrate the method. The count data are given in Table 2.1. Let $n_p = n_{11} + n_{22}$ be the number of individuals carrying the parental gametes and $n_r = n_{12} + n_{21}$ be the number of recombinants. The estimated recombination fraction between loci A and B is simply

$$\hat{r} = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} = \frac{n_r}{n_r + n_p}. \tag{2.1}$$

We use a hat above r to indicate estimation of r. The true value of recombination fraction is not known, but if the sample size is infinitely large, the estimated r will approach to the true value, meaning that the estimation is unbiased.

We now prove that $\hat{r}$ is the maximum likelihood estimate (MLE) of r. We introduce the ML method because it provides a significance test on the hypothesis that $r = 0.5$. To construct the likelihood function, we need a probability model, a sample of data and a parameter. The probability model is the binomial distribution, the data are the counts of the two possible genotypes, and the parameter is r. The binomial probability of the data given the parameter is

$$\Pr(n_r, n_p | r) = \frac{n!}{n_r! n_p!} \left(\frac{1}{2}\right)^{n_r + n_p} r^{n_r} (1 - r)^{n_p}, \tag{2.2}$$

where $n = n_r + n_p$ is the sample size. The value of r for $0 \le r \le 0.5$ that maximizes the probability is the MLE of r. Two issues need to be emphasized here for any maximum likelihood analysis, including this one. First, the probability involves a factor that does not depend on the parameter,

$$\texttt{const} = \frac{n!}{n_r! n_p!} \left(\frac{1}{2}\right)^{n}. \tag{2.3}$$

It is a constant with respect to the parameter r. This constant is irrelevant to the ML analysis and thus should be ignored. Secondly, the r value that maximizes a monotonic function of the probability also maximizes this probability. For computational convenience, we can maximize the logarithm of the probability. Therefore, it is the log likelihood function that is maximized in the ML analysis. The log likelihood function is defined as

$$L(r) = n_r \ln r + n_p \ln(1 - r). \tag{2.4}$$

To find the MLE of r, we need to find the derivative of $L(r)$ with respect to r,

$$\frac{\mathrm{d}}{\mathrm{d}r} L(r) = \frac{n_r}{r} - \frac{n_p}{1 - r}. \tag{2.5}$$

Letting $\frac{\mathrm{d}}{\mathrm{d}r}L(r) = 0$ and solving for r, we have

$$\hat{r} = \frac{n_r}{n_r + n_p}. \tag{2.6}$$

which is identical to that given in (2.1).

2.3 Standard Error and Significance Test

A parameter is a fixed but unknown quantity. The estimate of the parameter, however, is a variable because it varies from one sample to another. As the sample size increases, the estimate will approach to the true value of the parameter, provided that the estimate is unbiased. The deviation of the estimate from the true parameter can be measured by the standard error of the estimate. In this section, we will learn a method to calculate the standard error of $\hat{r}$. To calculate the standard error, we need the second derivative of the log likelihood function with respect to r and obtain a quantity called information, from which the variance of the estimated r can be approximated. Let us call the first derivative of $L(r)$ with respect to r the score function, denoted by $S(r)$,

$$S(r) = \frac{\mathrm{d}}{\mathrm{d}r}L(r) = \frac{n_r}{r} - \frac{n_p}{1-r}. \tag{2.7}$$

The second derivative of $L(r)$ with respect to r is called the Hessian matrix, denoted by $H(r)$,

$$H(r) = \frac{\mathrm{d}}{\mathrm{d}r}S(r) = \frac{\mathrm{d}^2}{\mathrm{d}r^2}L(r) = \frac{n_p}{(1-r)^2} - \frac{n_r}{r^2}. \tag{2.8}$$

Although $H(r)$ is a single variable, we still call it a matrix because in subsequent chapters we will deal with multiple dimension of parameters, in which case $H(r)$ is a matrix. From $H(r)$, we can find the information of r, which is

$$I(r) = -\mathtt{E}[H(r)] = \frac{\mathtt{E}(n_r)}{r^2} - \frac{\mathtt{E}(n_p)}{(1-r)^2}. \tag{2.9}$$

The symbol E represents expectation of the data given the parameter value. Here, the data are referred to n_r and n_p, not n, which is the sample size (a constant). Suppose that we know the true parameter r, what is the expected number of recombinants if we sample n individuals? This expected number is $\mathtt{E}(n_r) = r\,n$. The expected number of the parental types is $\mathtt{E}(n_p) = (1-r)n$. Therefore, the information is

$$I(r) = -\mathtt{E}[H(r)] = \frac{rn}{r^2} - \frac{(1-r)n}{(1-r)^2} = \frac{n}{r(1-r)}. \tag{2.10}$$

The variance of the estimated r takes the inverse of the information, with the true parameter replaced by $\hat{r}$,

$$\mathtt{var}(\hat{r}) \approx I^{-1}(\hat{r}) = \frac{\hat{r}(1-\hat{r})}{n}. \tag{2.11}$$

Therefore, the standard error of $\hat{r}$ is

$$\mathtt{se}(\hat{r}) = \sqrt{\mathtt{var}(\hat{r})} = \sqrt{\frac{\hat{r}(1-\hat{r})}{n}}. \tag{2.12}$$

The standard error is inversely proportional to the square root of the sample size and thus approaches zero as n becomes infinitely large.

When we report the estimated recombination fraction, we also need to report the estimation error in a form like $\hat{r} \pm \mathtt{se}(\hat{r})$. In addition to the sample size, the estimation error is also a function of the recombination fraction, with the maximum error occurring at $r = \frac{1}{2}$, i.e., when the two loci are unlinked. To achieve the same precision of estimation, it requires a larger sample to estimate a recombination fraction between two loosely linked loci than between two closely linked loci.

Because of the sampling error, even two unlinked loci may look like being linked as the estimated r may be superficially smaller than 0.5. How small an $\hat{r}$ is sufficiently small so that we can claim that the two loci are linked in the same chromosome? This requires a significance test.

The null hypothesis for such a test is denoted by $H_0 : r = \frac{1}{2}$. Verbally, H_0 is stated that the two loci are not linked. The alternative hypothesis is $H_A : r < 1/2$, i.e., the two loci are linked on the same chromosome. When the sample size is sufficiently large, we can always use the z-test to decide which hypothesis should be accepted. Here, we will use the usual likelihood ratio test statistic to declare the statistical significance of $\hat{r}$. Let $L(r)|_{r=\hat{r}} = L(\hat{r})$ be the log likelihood function evaluated at the MLE of r using (2.4). Let $L(r)|_{r=\frac{1}{2}} = L(1/2)$ be the log likelihood function evaluated under the null hypothesis. The likelihood ratio test statistic is defined as

$$\lambda = -2[L(1/2) - L(\hat{r})]. \tag{2.13}$$

where

$$L(\hat{r}) = n_r \ln \hat{r} + n_p \ln(1-\hat{r}). \tag{2.14}$$

and

$$L(1/2) = -n \ln 2 = -0.6931n. \tag{2.15}$$

If the null hypothesis is true, λ will approximately follow a chi-square distribution with one degree of freedom. Therefore, if $\lambda > \chi^2_{1,1-\alpha}$, we will claim that the two loci are linked, where $\chi^2_{1,1-\alpha}$ is the $(1-\alpha) \times 100$ percentile of the central χ^2_1 distribution and α is the type I error determined by the investigator. In human linkage studies, people often use LOD (log of odds) score instead. The relationship between LOD and λ is

$$\texttt{LOD} = \frac{\lambda}{2\ln(10)} \approx 0.2171\lambda. \tag{2.16}$$

Conventionally, $\texttt{LOD} > 3$ is used as a criterion to declare a significant linkage. This converts to a likelihood ratio criterion of $\lambda > 3 \times \ln(100) = 13.81551$. The LOD criterion has an intuitive interpretation. An LOD of k means that the alternative model (linkage) is 10^k times more likely than the null model.

2.4 Fisher's Scoring Algorithm for Estimating *r*

The $\texttt{F}_2$ mating design is demonstrated in Fig. 2.2. The ML analysis described for the $\texttt{BC}_1$ mating design is straightforward. The MLE of r has an explicit form. In fact, there is no need to invoke the ML analysis for the BC design other than to demonstrate the basic principle of the ML analysis. To estimate r using an $\texttt{F}_2$ design, the likelihood function is constructed using the same probability model (multinomial distribution), but finding the MLE of r is complicated. Therefore, we will resort to some special maximization algorithms. The algorithm we will learn is the Fisher's scoring algorithm (Fisher 1946).

Let us look at the genotype table (Table 2.2) and the table of genotype counts and frequencies (Table 2.3) for the $\texttt{F}_2$ design. If we were able to observe all the 16 possible genotypes, the same ML analysis used in the BC design would apply here to the $\texttt{F}_2$ design. Unfortunately, some of the genotypes listed in Table 2.2 are not distinguishable from others. For example, genotypes $\frac{A_1B_1}{A_1B_2}$ and $\frac{A_1B_2}{A_1B_1}$ are not distinguishable. These two genotypes appear to be the same because they both carry an A_1B_1 gamete and an A_1B_2 gamete. However, the origins of the two gametes are different for the two genotypes. Furthermore, the four genotypes in the minor diagonal of Table 2.2 actually represent four different linkage phases of the same observed genotype (double heterozygote). If we consider the origins of the alleles, there are four possible genotypes for each locus. However, the two configurations of the heterozygote are not distinguishable. Therefore, there are only three observable genotypes for each locus, making a total of nine observable joint

Table 2.3 The counts (in parentheses) and frequencies of the 16 possible genotypes in an $\texttt{F}_2$ family

	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_1	$(n_{11})\ \frac{1}{4}(1-r)^2$	$(n_{12})\ \frac{1}{4}r(1-r)$	$(n_{13})\ \frac{1}{4}r(1-r)$	$(n_{14})\ \frac{1}{4}(1-r)^2$
A_1B_2	$(n_{21})\ \frac{1}{4}r(1-r)$	$(n_{22})\ \frac{1}{4}r^2$	$(n_{23})\ \frac{1}{4}r^2$	$(n_{24})\ \frac{1}{4}r(1-r)$
A_2B_1	$(n_{31})\ \frac{1}{4}r(1-r)$	$(n_{32})\ \frac{1}{4}r^2$	$(n_{33})\ \frac{1}{4}r^2$	$(n_{34})\ \frac{1}{4}r(1-r)$
A_2B_2	$(n_{41})\ \frac{1}{4}(1-r)^2$	$(n_{42})\ \frac{1}{4}r(1-r)$	$(n_{43})\ \frac{1}{4}r(1-r)$	$(n_{44})\ \frac{1}{4}(1-r)^2$

Table 2.4 The nine observed genotypes and their counts in an $\mathtt{F}_2$ population

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	$A_1A_1B_1B_1$ (m_{11})	$A_1A_1B_1B_2$ (m_{12})	$A_1A_1B_2B_2$ (m_{13})
A_1A_2	$A_1A_2B_1B_1$ (m_{21})	$A_1A_2B_1B_2$ (m_{22})	$A_1A_2B_2B_2$ (m_{23})
A_2A_2	$A_2A_2B_1B_1$ (m_{31})	$A_2A_2B_1B_2$ (m_{32})	$A_2A_2B_2B_2$ (m_{33})

two-locus genotypes, as shown in Table 2.4. Let m_{ij} be the counts of the joint genotype combining the ith genotype of locus A and the jth genotype of locus B, for $i, j = 1, \ldots, 3$. These counts are the data from which a likelihood function can be constructed.

Before we construct the likelihood function, we need to find the probability for each of the nine observed genotypes. These probabilities are listed in Table 2.5. The count data in the second column and the frequencies in the third column of Table 2.5 are what we need to construct the log likelihood function, which is

$$\begin{aligned}L(r) &= \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}\ln(q_{ij})\\ &= [2(m_{11}+m_{33}) + m_{12} + m_{21} + m_{23} + m_{32}]\ln(1-r)\\ &\quad + [2(m_{13}+m_{31}) + m_{12} + m_{21} + m_{23} + m_{32}]\ln(r)\\ &\quad + m_{22}\ln[r^2 + (1-r)^2].\end{aligned} \tag{2.17}$$

The derivative of $L(r)$ with respect to r is

$$\begin{aligned}S(r) &= \frac{\mathrm{d}}{\mathrm{d}r}L(r)\\ &= -\frac{2(m_{11}+m_{33})}{1-r} + \frac{(m_{12}+m_{21}+m_{23}+m_{32})(1-2r)}{r(1-r)}\\ &\quad -\frac{2m_{22}(1-2r)}{1-2r+2r^2} + \frac{2(m_{13}+m_{31})}{r}.\end{aligned} \tag{2.18}$$

The MLE of r is obtained by setting $S(r) = 0$ and solving for r. Unfortunately, there is no explicit solution for r. Therefore, an iterative algorithm is resorted to solve for r. Before introducing the Fisher's scoring algorithm (Fisher 1946), we first try the Newton method, which also requires the second derivative of $L(r)$ with respect to r,

Table 2.5 Frequencies of the nine observed genotypes in an $\mathtt{F}_2$ population

Genotype		Count		Probability	
$A_1A_1B_1B_1$	$= \frac{A_1B_1}{A_1B_1}$	$m_{11} =$	n_{11}	$q_{11} =$	$\frac{1}{4}(1-r)^2$
$A_1A_1B_1B_2$	$= \frac{A_1B_1}{A_1B_2}, \frac{A_1B_2}{A_1B_1}$	$m_{12} =$	$n_{12} + n_{21}$	$q_{12} =$	$\frac{1}{2}r(1-r)$
$A_1A_1B_2B_2$	$= \frac{A_1B_2}{A_1B_2}$	$m_{13} =$	n_{22}	$q_{13} =$	$\frac{1}{4}r^2$
$A_1A_2B_1B_1$	$= \frac{A_1B_1}{A_2B_1}, \frac{A_2B_1}{A_1B_1}$	$m_{21} =$	$n_{13} + n_{31}$	$q_{21} =$	$\frac{1}{2}r(1-r)$
$A_1A_2B_1B_2$	$= \frac{A_1B_1}{A_2B_2}, \frac{A_1B_2}{A_2B_1}$	$m_{22} =$	$n_{14} + n_{23} +$	$q_{22} =$	$\frac{1}{2}[r^2 + (1-r)^2]$
	$\frac{A_2B_1}{A_1B_2}, \frac{A_2B_2}{A_1B_1}$		$n_{32} + n_{41}$		
$A_1A_2B_2B_2$	$= \frac{A_1B_2}{A_2B_2}, \frac{A_2B_2}{A_1B_2}$	$m_{23} =$	$n_{24} + n_{42}$	$q_{23} =$	$\frac{1}{2}r(1-r)$
$A_2A_2B_1B_1$	$= \frac{A_2B_1}{A_2B_1}$	$m_{31} =$	n_{33}	$q_{31} =$	$\frac{1}{4}r^2$
$A_2A_2B_1B_2$	$= \frac{A_2B_1}{A_2B_2}, \frac{A_2B_2}{A_2B_1}$	$m_{32} =$	$n_{34} + n_{43}$	$q_{32} =$	$\frac{1}{2}r(1-r)$
$A_2A_2B_2B_2$	$= \frac{A_2B_2}{A_2B_2}$	$m_{33} =$	n_{44}	$q_{33} =$	$\frac{1}{4}(1-r)^2$

$$\begin{aligned} H(r) &= \frac{\mathrm{d}}{\mathrm{d}r}S(r) = \frac{\mathrm{d}^2}{\mathrm{d}r^2}L(r) \\ &= -\frac{2(m_{11}+m_{33})}{(1-r)^2} - \frac{(m_{12}+m_{21}+m_{23}+m_{32})(1-2r+2r^2)}{r^2(1-r)^2} \\ &\quad + \frac{8m_{22}r(1-r)}{(1-2r+2r^2)^2} - \frac{2(m_{13}+m_{31})}{r^2}. \end{aligned} \tag{2.19}$$

The Newton method starts with an initial value of r, denoted by $r^{(t)}$ for $t = 0$, and update the value by

$$r^{(t+1)} = r^{(t)} - \frac{S(r^{(t)})}{H(r^{(t)})}. \tag{2.20}$$

The iteration process stops if

$$|r^{(t+1)} - r^{(t)}| \le \epsilon, \tag{2.21}$$

where ϵ is a small positive number, say 10^{-8}.

The derivation of the Newton method is very simple. It uses the Taylor series expansion to approximate the score function. Let $r^{(0)}$ be the initial value of r. The score function $S(r)$ can be approximated in the neighborhood of $r^{(0)}$ by

$$\begin{aligned} S(r) &= S(r^{(0)}) + \frac{\mathrm{d}}{\mathrm{d}r}S(r^{(0)})(r-r^{(0)}) + \frac{1}{2!}\frac{\mathrm{d}^2}{\mathrm{d}r^2}S(r^{(0)})(r-r^{(0)})^2 + \cdots \\ &\approx S(r^{(0)}) + \frac{\mathrm{d}}{\mathrm{d}r}S(r^{(0)})(r-r^{(0)}). \end{aligned} \tag{2.22}$$

The approximation is due to ignorance of the higher order terms of the Taylor series. Recall that $H(r^{(0)}) = \frac{\mathrm{d}}{\mathrm{d}r}S(r^{(0)})$ and thus

$$S(r) \approx S(r^{(0)}) + H(r^{(0)})(r - r^{(0)}). \tag{2.23}$$

Letting $S(r) = 0$ and solving for r, we get

$$r = r^{(0)} - \frac{S(r^{(0)})}{H(r^{(0)})}. \tag{2.24}$$

We have moved from $r^{(0)}$ to r, one step closer to the true solution. Let $r = r^{(t+1)}$ and $r^{(0)} = r^{(t)}$. The Newton's equation of iteration (2.20) is obtained by substituting r and $r^{(0)}$ into (2.24).

The Newton method does not behave well when r is close to zero or 0.5 for the reason that $H^{-1}(r)$ can be easily overflowed. The Fisher's scoring method is a modified version of the Newton method for avoiding the overflow problem. As such, the method behaves well in all range of the parameter in the legal domain $0 \le r \le \frac{1}{2}$. In the Fisher's scoring method, the second derivative involved in the iteration is simply replaced by the so-called expectation of the second derivative. The iteration equation becomes

$$r^{(t+1)} = r^{(t)} - \frac{S(r^{(t)})}{\mathtt{E}[H(r^{(t)})]}, \tag{2.25}$$

where

$$\mathtt{E}[H(r^{(t)})] = -\frac{2n[1 - 3r^{(t)} + 3(r^{(t)})^2]}{r^{(t)}(1 - r^{(t)})[1 - 2r^{(t)} + 2(r^{(t)})^2]}. \tag{2.26}$$

Let $I(r^{(t)}) = -\mathtt{E}[H(r^{(t)})]$ be the Fisher's information. The iteration process can be rewritten as

$$r^{(t+1)} = r^{(t)} + I^{-1}(r^{(t)})S(r^{(t)}). \tag{2.27}$$

Assume that the iteration converges at the $t + 1$ iteration. The MLE of r is $\hat{r} = r^{(t+1)}$. The method provides an automatic way to calculate the variance of the estimate,

$$\mathtt{var}(\hat{r}) \approx I^{-1}(\hat{r}) = \frac{\hat{r}(1 - \hat{r})(1 - 2\hat{r} + 2\hat{r}^2)}{2n(1 - 3\hat{r} + 3\hat{r}^2)}, \tag{2.28}$$

where $n = \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}$ is the sample size. Note that when $\hat{r} \to 0$, $\hat{r}^2$ becomes negligible and $1 - 2\hat{r} \approx 1–3\hat{r}$, leading to $1–2\hat{r} + 2\hat{r}^2 \approx 1–3\hat{r} + 3\hat{r}^2$. Therefore,

$$\mathtt{var}(\hat{r}) \approx \frac{\hat{r}(1 - \hat{r})}{2n}. \tag{2.29}$$

Comparing this variance with the one in the BC design shown in (2.11), we can see that the variance has been reduced by half. Therefore, using the $\mathtt{F}_2$ design is more efficient than the BC design.

2.5 EM Algorithm for Estimating *r*

The EM algorithm was developed by Dempster et al. (1977) for handling missing data problems. The algorithm repeatedly executes an E-step and an M-step for iterations. The E-step stands for expectation and the M-step for maximization. The problem of estimating recombination fraction in $\mathtt{F}_2$ can be formulated as a missing value problem and thus solved by the EM algorithm. The derivation of the EM algorithm is quite involved and will be introduced later when we deal with a simpler problem. We now only give the final equation of the EM iteration. Recall that the $\mathtt{F}_1$ hybrid can produce four possible gametes, two of them are of parental type (A_1B_1 and A_2B_2) and the other two are recombinants (A_1B_2 and A_2B_1). Therefore, an $\mathtt{F}_2$ progeny can be classified into one of three categories in terms of the number of recombinant gametes contained: 0, 1, or 2. From Table 2.3, we can see that each of the following observed genotypes carries one recombinant gamete: $A_1A_1B_1B_2$, $A_1A_2B_1B_1$, $A_1A_2B_2B_2$, and $A_2A_2B_1B_2$, and each of the following observed genotypes carries two recombinant gametes: $A_1A_1B_2B_2$ and $A_2A_2B_1B_1$. Let $n_1 = m_{12} + m_{21} + m_{23} + m_{32}$ be the number of individuals of category 1 and $n_2 = m_{13} + m_{31}$ be the number of individuals of category 2. The double heterozygote $A_1A_2B_1B_2$ is an ambiguous genotype because it may carry 0 recombinant gamete, $(\frac{A_1B_1}{A_2B_2}, \frac{A_2B_2}{A_1B_1})$, or two recombinant gametes, $(\frac{A_1B_2}{A_2B_1}, \frac{A_2B_1}{A_1B_2})$. The number of double heterozygote individuals that carry two recombinant gametes is $n_{23} + n_{32}$. Unfortunately, this number is not observable. If it were, we would be able to take the ratio of the number of recombinant gametes to the total number of gametes in the $\mathtt{F}_2$ progeny ($2n$) to get the estimated recombination fraction right away,

$$\hat{r} = \frac{1}{2n}[2(n_{23} + n_{32} + n_2) + n_1] \tag{2.30}$$

The EM algorithm takes advantage of this simple expression by substituting the missing values ($n_{23} + n_{32}$) by its expectation. The expectation, however, requires knowledge of the parameter, which is what we want to estimate. Therefore, iterations are required. To calculate the expectation, we need the current value of r, denoted by $r^{(t)}$, and the number of double heterozygote individuals (m_{22}). Recall that the overall proportion of the double heterozygote is $\frac{1}{2}[r^2 + (1-r)^2]$, where $\frac{1}{2}r^2$ represents the proportion of individuals carrying two recombinant gametes and $\frac{1}{2}(1-r)^2$ represents the proportion of individuals carrying no recombinant gametes. The conditional expectation of $n_{23} + n_{32}$ is

$$\mathtt{E}(n_{23} + n_{32}) = \frac{(r^{(t)})^2}{(r^{(t)})^2 + (1 - r^{(t)})^2} m_{22} = w^{(t)} m_{22}. \tag{2.31}$$

The iterative equation may be written as

$$r^{(t+1)} = \frac{1}{2n}\{2[\mathtt{E}(n_{23} + n_{32}) + n_2] + (n_1)\}. \tag{2.32}$$

The final equation of the EM iteration becomes

$$r^{(t+1)} = \frac{1}{2n}\left[2(w^{(t)}m_{22} + m_{13} + m_{31}) + (m_{12} + m_{21} + m_{23} + m_{32})\right]. \quad (2.33)$$

Calculating $\mathtt{E}(n_{23} + n_{32})$ using (2.31) represents the E-step, and updating $r^{(t+1)}$ using (2.32) represents the M-step of the EM algorithm. The final result of the EM algorithm is so simple, yet it behaves extremely well with regard to the small number of iterations required for convergence and the insensitiveness to the initial value of r. A drawback of the EM algorithm is the difficulty in calculating the standard error of the estimate. Since the solution is identical to the Fisher's scoring method, the variance (square of the standard error) of the estimate given in (2.28) can be used as the variance of the EM estimate.

To test the hypothesis of no linkage, $r = \frac{1}{2}$, we will use the same likelihood ratio test statistic, as described in the BC design. The log likelihood value under the null model, however, needs to be evaluated in a slightly different way, that is, $L(\frac{1}{2}) = \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}\ln(q_{ij})$, where q_{ij} is a function of $r = \frac{1}{2}$ (see Table 2.5). The log likelihood value under the alternative model is evaluated at $r = \hat{r}$, using $L(\hat{r}) = \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}\ln(\hat{q}_{ij})$, where $\hat{q}_{ij}$ is a function of $r = \hat{r}$ (see Table 2.5).

Springer

http://www.springer.com/978-0-387-70806-5

Principles of Statistical Genomics
Xu, S.
2013, XV, 428 p. 46 illus., 12 illus. in color., Hardcover
ISBN: 978-0-387-70806-5