Chapter 4
Multipoint Analysis of Mendelian Loci

Each Mendelian locus occupies a specific point on a chromosome. A linkage analysis requires two or more Mendelian loci and thus involves two or more points. When a linkage analysis involves two Mendelian loci, as we have seen in Chap. 2 for estimating the recombination fraction between two loci, the analysis is called two-point analysis. When more than two Mendelian loci are analyzed simultaneously, the method is called multipoint analysis (Jiang and Zeng 1997). Multipoint analysis can extract more information from the data if markers are not fully informative, e.g., missing genotypes, dominance alleles, and so on.

When there is no interference between the crossovers of two consecutive chromosome segments, the joint distribution of genotypes of marker loci is Markovian. We can imagine that the entire chromosome behaves like a Markov chain, in which the genotype of one locus depends only on the genotype of the "previous" locus. A Markov chain has a direction, but a chromosome has no meaningful direction. Its direction is defined in an arbitrary fashion. Therefore, we can use either a forward Markov chain or a backward Markov chain to define a chromosome, and the result will be identical, regardless of which direction has been taken.

A Markov chain is used to derive the joint distribution of all marker genotypes. The joint distribution is eventually used to construct a likelihood function for estimating multiple recombination fractions. Given the recombination fractions, one can derive the conditional distribution of the genotype of a locus bracketed by two marker loci given the genotypes of the markers. The conditional distribution is fundamentally important in genetic mapping for complex traits, a topic to be discussed in a later chapter.

4.1 Joint Distribution of Multiple-Locus Genotype

When three loci are considered jointly, the method is called three-point analysis. Theory developed for three-point analysis applies to arbitrary number of loci.

S. Xu, *Principles of Statistical Genomics*, DOI 10.1007/978-0-387-70807-2_4,

4.1.1 BC Design

Let ABC be three ordered loci on the same chromosome with pairwise recombination fractions denoted by r_{AB}, r_{BC}, and r_{AC}. We can imagine that these loci form a Markov chain as either A $\longrightarrow$ B $\longrightarrow$ C or A $\longleftarrow$ B $\longleftarrow$ C. The direction is arbitrary. Each locus represents a discrete variable with two or more distinct values (states). For an individual from a BC population, each locus takes one of two possible genotypes and thus two states. Let A_1A_1 and A_1A_2 be the two possible genotypes for locus A, B_1B_1 and B_1B_2 be the two possible genotypes for locus B, and C_1C_1 and C_1C_2 be the two possible genotypes for locus C. For convenience, each state is assigned a numerical value. For example, $A = 1$ or $A = 2$ indicates that an individual takes genotype A_1A_1 or A_1A_2. Let us take A $\longrightarrow$ B $\longrightarrow$ C as the Markov chain; the joint distribution of the three-locus genotype is

$$\Pr(A, B, C) = \Pr(A) \Pr(B|A) \Pr(C|B), \tag{4.1}$$

where $\Pr(A = 1) = \Pr(A = 2) = \frac{1}{2}$ assuming that there is no segregation distortion. The conditional probabilities, $\Pr(B|A)$ and $\Pr(C|B)$, are called the transition probabilities between loci A and B and between loci B and C, respectively. The transition probabilities depend on the genotypes of the two loci and the recombination fractions between the two loci. These transition probabilities can be found from the following 2×2 transition matrix:

$$T_{AB} = \begin{bmatrix} \Pr(B = 1|A = 1) & \Pr(B = 2|A = 1) \\ \Pr(B = 1|A = 2) & \Pr(B = 2|A = 2) \end{bmatrix}. \tag{4.2}$$

Because $\Pr(B = 1|A = 1) = \Pr(B = 2|A = 2) = 1 - r_{AB}$ represents the probability of no recombination between the two loci and $\Pr(B = 2|A = 1) = \Pr(B = 1|A = 2) = r_{AB}$ represents the probability of recombination between the two loci, the exact form of the transition matrix between loci A and B is

$$T_{AB} = \begin{bmatrix} T_{AB}(1,1) & T_{AB}(1,2) \\ T_{AB}(2,1) & T_{AB}(2,2) \end{bmatrix} = \begin{bmatrix} 1 - r_{AB} & r_{AB} \\ r_{AB} & 1 - r_{AB} \end{bmatrix}, \tag{4.3}$$

where $T_{AB}(k,l)$ $\forall k,l = 1,2$ denotes the kth row and the lth column of matrix T_{AB}. It is now obvious that $T_{AB}(k,l) = \Pr(B = l|A = k)$. Note that we have used a special notation "$\forall k,l = 1,2$" to indicate that k and l each takes a value from 1 to 2. Verbally, "$\forall k,l = 1,2$" means "for all $k = 1,2$ and $l = 1,2$". When using this kind of notation, we should particularly pay attention to the positions of k and l in $T_{AB}(k,l) = \Pr(B = l|A = k)$. It is a conditional probability that $B = l$ given $A = k$. Replacing the conditional probabilities by the elements of the transition matrix, we rewrite the joint probability of the three-locus genotype as

$$\Pr(A, B, C) = \frac{1}{2} T_{AB}(A, B) T_{BC}(B, C). \tag{4.4}$$

For example, the probability that $A = 1$, $B = 2$, and $C = 2$ is

$$\Pr(A = 1, B = 2, C = 2) = \frac{1}{2} T_{AB}(1,2) T_{BC}(2,2) = \frac{1}{2} r_{AB}(1 - r_{BC}).$$

This joint probability can be written in matrix notation. Let us use a 2×2 diagonal matrix D_A to denote the genotype of locus A. This matrix is defined as

$$D_A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \texttt{ for } A = 1 \texttt{ and } D_A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \texttt{ for } A = 2.$$

Diagonal matrices D_B and D_C are defined similarly for loci B and C, respectively. The original data are in the form of genotype indicator variables, A, B, and C, but the new form of the data is represented by the diagonal matrices. Let us define a 2×1 unity vector by $J = [1 \quad 1]'$. The joint distribution given in (4.4) is rewritten in matrix notation as

$$\Pr(A, B, C) = \frac{1}{2} J' D_A T_{AB} D_B T_{BC} D_C J. \tag{4.5}$$

One can verify that

$$\begin{aligned} &\Pr(A = 1, B = 2, C = 2) \\ &\quad = \frac{1}{2} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 - r_{AB} & r_{AB} \\ r_{AB} & 1 - r_{AB} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 - r_{BC} & r_{BC} \\ r_{BC} & 1 - r_{BC} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &\quad = \frac{1}{2} r_{AB}(1 - r_{BC}). \end{aligned}$$

4.1.2 F_2 Design

Taking into consideration the order of the two alleles carried by an F_2 individual, we have four possible genotypes: A_1A_1, A_1A_2, A_2A_1, and A_2A_2. The first and the last genotypes are homozygotes, while the second and third genotypes are heterozygotes. The two forms of heterozygote represent two different origins of the alleles. They are indistinguishable from each other. Therefore, we adopt a special notation, (A_1A_2), to denote the unordered heterozygote. The alleles and genotypes for the other loci are expressed using similar notation. Let $A = k, \forall k = 1, \ldots, 4$ be an indicator variable to indicate the four genotypes of locus A. Variables B and C are similarly defined for loci B and C, respectively. The joint probability of the three-locus genotype is $\Pr(A, B, C) = \Pr(A) \Pr(B|A) \Pr(C|B)$ where $\Pr(A{=}k){=}\frac{1}{4}, \forall k = 1, \ldots, 4$. $\Pr(B|A)$ and $\Pr(C|B)$ are the transition probabilities

from locus A to locus B and from locus B to locus C, respectively. The transition probabilities from locus A to locus B can be found from the following 4×4 transition matrix:

$$T_{AB} = \begin{bmatrix} (1-r_{AB})^2 & (1-r_{AB})r_{AB} & r_{AB}(1-r_{AB}) & r_{AB}^2 \\ (1-r_{AB})r_{AB} & (1-r_{AB})^2 & r_{AB}^2 & r_{AB}(1-r_{AB}) \\ r_{AB}(1-r_{AB}) & r_{AB}^2 & (1-r_{AB})^2 & 1-r_{AB})r_{AB} \\ r_{AB}^2 & r_{AB}(1-r_{AB}) & 1-r_{AB})r_{AB} & (1-r_{AB})^2 \end{bmatrix}. \tag{4.6}$$

The transition matrix from locus B to locus C is denoted by T_{BC}, which is equivalent to matrix (4.6) except that the subscript $_{AB}$ is replaced by subscript $_{BC}$.

Note that this transition matrix is obtained by the Kronecker square (denoted by a superscript $^{[2]}$) of a 2×2 transition matrix,

$$H_{AB} = \begin{bmatrix} 1-r_{AB} & r_{AB} \\ r_{AB} & 1-r_{AB} \end{bmatrix}, \tag{4.7}$$

that is,

$$T_{AB} = \begin{bmatrix} 1-r_{AB} & r_{AB} \\ r_{AB} & 1-r_{AB} \end{bmatrix}^{[2]} = \begin{bmatrix} 1-r_{AB} & r_{AB} \\ r_{AB} & 1-r_{AB} \end{bmatrix} \otimes \begin{bmatrix} 1-r_{AB} & r_{AB} \\ r_{AB} & 1-r_{AB} \end{bmatrix}.$$

The 4×4 transition matrix (4.6) may be called the zygotic transition matrix, and the 2×2 transition matrix (4.7) may be called the gametic transition matrix. That the zygotic transition matrix is the Kronecker square of the gametic transition matrix is very intuitive because a zygote is the product of two gametes. Let $T_{AB}(k, l)$ be the kth row and the lth column of the 4×4 transition matrix T_{AB}, $\forall k, l = 1, \ldots, 4$. The joint probability of the three-locus genotype is expressed as

$$\Pr(A, B, C) = \frac{1}{4} T_{AB}(A, B) T_{BC}(B, C). \tag{4.8}$$

For example, the joint three-locus genotype $A_1A_1B_1B_2C_2C_1$ is numerically coded as $A = 1$, $B = 2$, and $C = 3$, whose probability is

$$\Pr(A = 1, B = 2, C = 3) = \frac{1}{4} T_{AB}(1, 2) T_{BC}(2, 3) = \frac{1}{4}(1 - r_{AB}) r_{AB} r_{BC}^2.$$

In practice, people will never observe a three-locus genotype like $A_1A_1B_1B_2C_2C_1$ because the two forms of the heterozygote are not distinguishable. The joint three-locus genotype $A_1A_1(B_1B_2)(C_1C_2)$ is actually what we can observe. The numerical code for the first locus is $A = 1$, but the codes for loci B and C are ambiguous. For example, locus B can be coded as either $B = 2$ or $B = 3$ with an equal

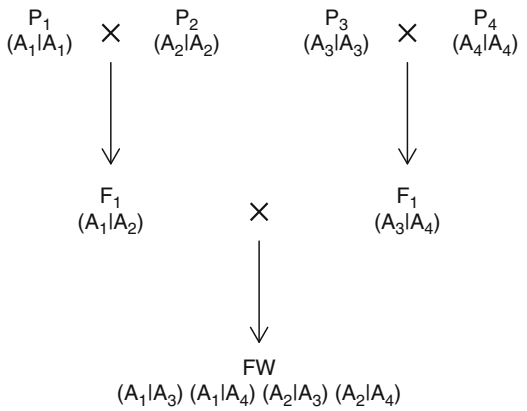

Fig. 4.1 Four-way (FW) cross mating design

probability. This ambiguous situation is denoted by $B = (2,3)$. Similar notation applies to locus C as $C = (2,3)$. The joint distribution for $A_1A_1(B_1B_2)(C_1C_2)$ is

$$\begin{aligned}\Pr[A = 1, B = (2,3), C = (2,3)] &= \frac{1}{4}\sum_{k=2}^{3}\left[T_{AB}(1,k)\sum_{l=2}^{3}T_{BC}(k,l)\right] \\ &= \frac{1}{2}r_{AB}(1-r_{AB})[r_{BC}^2 + (1-r_{BC})^2].\end{aligned}$$

Again, the joint distribution of the three-locus genotype (4.8) can be expressed in matrix notation. We now use a 4×4 diagonal matrix to denote the genotype of a locus. For locus A, this diagonal matrix is defined as

$$D_A = \begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix},\ D_A = \begin{bmatrix}0&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix}\ \texttt{and}\ D_A = \begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1\end{bmatrix},$$

for $A = 1$, $A = (2,3)$ and $A = 4$, respectively. Verbally, matrix D_A is a diagonal matrix with unity values for the diagonal elements corresponding to the positions pointed by the value of A. Having defined these diagonal matrices for all loci, we can rewrite the joint distribution of the three-locus genotype as

$$\Pr(A, B, C) = \frac{1}{4}J' D_A T_{AB} D_B T_{BC} D_C J, \tag{4.9}$$

where J is now a 4×1 vector of unity, rather than a 2×1 vector as in the BC design (Fig. 2.1).

4.1.3 Four-Way Cross Design

A four-way cross design involves two different crosses and four different inbred parents. Let $F_1^{(12)}$ be the hybrid progeny derived from the cross of P_1 and P_2 and $F_1^{(34)}$ be the progeny derived from the cross of P_3 and P_4. The cross between $F_1^{(12)}$ and $F_1^{(34)}$ is called the four-way cross. Such a design is called the four-way cross design (FW) as illustrated in Fig. 4.1. Let $A_k A_k B_k B_k C_k C_k$ be the three-locus genotype for the k parent, $\forall k = 1, \ldots, 4$. The three-locus genotypes for $F_1^{(12)}$ and $F_1^{(34)}$ are $A_1A_2B_1B_2C_1C_2$ and $A_3A_4B_3B_4C_3C_4$, respectively. Consider a single locus, say locus A. An FW progeny can take one of the four genotype: A_1A_3, A_1A_4, A_2A_3, and A_3A_4. Let $A = 1, \ldots, 4$ denote the numerical code for each of the four genotypes. The joint three-locus genotype is still expressed by (4.9) with the same transition matrices as defined earlier in the F_2 design. The diagonal matrices, D_A, D_B, and D_C, are defined similarly to those in the F_2 design except that the second and third genotypes are distinguishable. The numerical code of $A = k$ is translated into a D_A matrix whose elements are all zero except that the kth row and the kth column are unity. For example, the joint probability that $A = 3$, $B = 1$, and $C = 4$ is

$$\begin{aligned}\Pr(A = 3, B = 1, C = 4) &= \frac{1}{4} J' D_A T_{AB} D_B T_{BC} D_C J \\ &= \frac{1}{4} T_{AB}(3,1) T_{BC}(1,4) \\ &= \frac{1}{4} r_{AB}(1 - r_{AB}) r_{BC}^2 .\end{aligned}$$

4.2 Incomplete Genotype Information

4.2.1 Partially Informative Genotype

The FW cross design described earlier represents a situation where all the four genotypes in the progeny are distinguishable. In reality, it is often that not all genotypes are distinguishable. This may happen when two or more of the grandparents carry the same allele at the locus of interest. The consequence is that the F_1 hybrid initiated by the first level of the cross may be homozygous or the two F_1 parents may have the same genotype. Assume that $F_1^{(34)}$ has a genotype of A_3A_3, which is homozygous. This may be caused by a cross between two parents, both of which are fixed at A_3 allele. Regardless of the reason that causes the homozygosity of the F_1 hybrid, let us focus on the genotypes of the two F_1 parents and consider the four possible genotypes of the FW progeny. Assume that $F_1^{(12)}$ and $F_1^{(34)}$ have genotypes of A_1A_2 and A_3A_3, respectively.

The four possible genotypes of the progeny are A_1A_3, A_1A_3, A_2A_3, and A_2A_3. The first and the second genotypes are not distinguishable, although the A_3 allele carried by the two genotypes has different origins. This situation applies to the third and fourth genotypes. Considering the allelic origins, we have four ordered genotypes, but we only observe two distinguishable genotypes. This phenomenon is called incomplete information for the genotype. Such a genotype is called partially informative genotype. If we observe a genotype A_1A_3, the numerical code for the genotype is $A = (1, 2)$. In matrix notation, it is represented by

$$D_A = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix}.$$

If an observed genotype is A_2A_3, the numerical code becomes $A = (3, 4)$, represented by

$$D_A = \begin{bmatrix} 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix}.$$

If both parents are homozygous and fixed to the same allele, say A_1, then all the four genotypes of the progeny have the same observed form, A_1A_1. The numerical code for the genotype is $A = (1, 2, 3, 4)$, a situation called no information. Such a locus is called uninformative locus and usually excluded from the analysis. The diagonal matrix representing the genotype is simply a 4×4 identity matrix.

The following is an example showing how to calculate the three-locus joint genotype using the FW cross approach with partial information. Let $A_1A_3B_2B_3C_1C_1$ and $A_4A_4B_2B_3C_1C_2$ be the three-locus genotypes for two parents. The linkage phases of markers in the parents are assumed to be known so that the order of the two alleles within a locus is meaningful. In fact, the phase-known genotypes of the parents are better denoted by $\frac{A_1B_2C_1}{A_3B_3C_1}$ and $\frac{A_4B_2C_1}{A_4B_3C_2}$, respectively, for the two parents. Assume that a progeny has a genotype of $A_3A_4B_2B_2C_1C_1$. We want to calculate the probability of observing such a progeny given the genotypes of the parents. First, we examine each single-locus genotype to see which one of the four possible genotypes this individual belongs to. For locus A, the parental genotypes are A_1A_2 and A_4A_4. The four possible genotypes of a progeny are A_1A_4, A_1A_4, A_3A_4, and A_3A_4, respectively. The single-locus genotype of the progeny is A_3A_4, matching the third and fourth genotypes, and thus $A = (3, 4)$. For locus B, the parental genotypes are B_2B_3 and B_2B_3. The four possible genotypes of a progeny are B_2B_2, B_2B_3, B_3B_2, and B_3B_3, respectively. The single-locus genotype B_2B_2 for the progeny matches the first genotype and thus $B = 1$. For locus C, the parental genotypes are C_1C_1 and C_1C_2. The four possible genotypes of a progeny are C_1C_1, C_1C_2, C_1C_1, and C_1C_2, respectively. The single-locus genotype of the progeny C_1C_1 matches the first and the third genotypes and thus $C = (1, 3)$. In summary, the numerical codes

for the three loci are $A = (3, 4)$, $B = 1$, and $C = (1, 3)$, respectively. We now convert the three single-locus genotypes into their corresponding diagonal matrices,

$$D_A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \; D_B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } D_C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Substituting these matrices into (4.9), we have

$$\begin{aligned} \Pr[A = (3, 4), B = 1, C = (1, 3)] &= \frac{1}{4} J' D_A T_{AB} D_B T_{BC} D_C J \\ &= \frac{1}{4} [T_{AB}(3, 1) + T_{AB}(4, 1)][T_{BC}(1, 1) + T_{BC}(1, 3)] \\ &= \frac{1}{4} r_{AB}(1 - r_{BC}) \end{aligned}$$

4.2.2 *BC and $\mathbf{F_2}$ Are Special Cases of FW*

The four-way cross design is a general design where the BC and F_2 designs are special cases of the general design with partial information. For example, the two parents of the BC_1 design have genotypes of A_1A_2 and A_1A_1, respectively. If we treat a BC progeny as a special FW progeny, the four possible genotypes are A_1A_1, A_1A_1, A_2A_1, and A_2A_1, only two distinguishable observed types. If a progeny has a genotype A_1A_1, the numerical code of the genotype in terms of an FW cross is $A = (1, 2)$. If a progeny has a genotype of A_2A_1, its numerical codes become $A = (3, 4)$. The two parents of a BC_1' design have genotypes of A_1A_2 and A_2A_2, respectively. In terms of an FW cross, the four possible genotypes are A_1A_2, A_1A_2, A_2A_2, and A_2A_2. Again, there are only two distinguishable genotypes. The two parents of an F_2 design have genotypes of A_1A_2 and A_1A_2, respectively. If we treat an F_2 progeny as a special FW progeny, the four possible genotypes are A_1A_1, A_1A_2, A_2A_1, and A_2A_2, only three distinguishable genotypes. The numerical codes for the two types of homozygote are $A = 1$ and $A = 4$, respectively, whereas the numerical code for the heterozygote is $A = (2, 3)$. In summary, when the general FW design is applied to a BC design, only two of the four possible genotypes are distinguishable, and the numerical codes are $A = (1, 2)$ for one observed genotype and $A = (3, 4)$ for the other observed genotype. When the general FW design is applied to the F_2 design, the two forms of heterozygote are not distinguishable. When coding the genotype, we use $A = (2, 3)$ to represent the heterozygote and $A = 1$ and $A = 4$ to represent the two types of homozygote, respectively. The transition matrices remain the same as those used in an FW cross design.

We have learned the BC design in Sec. 4.1.1 using the 2×2 transition matrix. When using the FW design for the BC problem, we have combined the first and second genotypes to form the first observable genotype and combined the third and fourth genotypes to form the second observable genotype for the BC design. It can be shown that the joint probability calculated by the Markov chain with two states (using the 2×2 transition matrix) and that calculated by the Markov chain with four states (the 4×4 transition matrix) are identical.

The $\mathtt{F}_2$ design we learned earlier can be handled by combining the second and third genotypes into the observed heterozygote. The 4×4 transition matrix is converted into a 3×3 transition matrix,

$$T_{AB} = \begin{bmatrix} (1-r_{AB})^2 & 2(1-r_{AB})r_{AB} & r_{AB}^2 \\ (1-r_{AB})r_{AB} & r_{AB}^2 + (1-r_{AB})^2 & (1-r_{AB})r_{AB} \\ r_{AB}^2 & 2(1-r_{AB})r_{AB} & (1-r_{AB})^2 \end{bmatrix}.$$

The joint probability of multiple-locus genotype for an $\mathtt{F}_2$ individual can be calculated using a Markov chain with the 3×3 transition matrix. The numerical code for a genotype must be redefined in the following way. The three defined genotypes, A_1A_1, A_1A_2, and A_2A_2, are numerically coded by $A = 1$, $A = 2$, and $A = 3$, respectively.

In matrix notation, the three genotypes are denoted by

$$D_A = \begin{bmatrix} 1\,0\,0 \\ 0\,0\,0 \\ 0\,0\,0 \end{bmatrix}, \; D_A = \begin{bmatrix} 0\,0\,0 \\ 0\,1\,0 \\ 0\,0\,0 \end{bmatrix} \text{ and } D_A = \begin{bmatrix} 0\,0\,0 \\ 0\,0\,0 \\ 0\,0\,1 \end{bmatrix},$$

respectively.

The general FW design using a Markov chain with four states is computationally more intensive when applied to BC and $\mathtt{F}_2$ designs compared to the specialized BC (with 2×2 transition matrix) and $\mathtt{F}_2$ (with 3×3 transition matrix) algorithm. However, the difference in computing times is probably unnoticeable given the current computing power. In addition, the 3×3 transition matrix is not symmetrical, a factor that may easily cause a programming error. Therefore, the general FW design is recommended for all line crossing experiments.

4.2.3 Dominance and Missing Markers

A dominance marker is a type of marker whose heterozygous genotype cannot be distinguished from one of the two homozygous genotypes. Therefore, dominance markers cannot be used in a BC design. However, partial information can be extracted from dominance markers in an F_2 design. Consider locus A with four possible genotypes in an F_2 population under a biallelic system, alleles A_1 vs A_2. The four ordered genotypes are A_1A_1, A_1A_2, A_2A_1, and A_2A_2. Dominance can be found in two directions. If A_1 is dominant over A_2, we cannot distinguish the three

genotypes, A_1A_1, A_1A_2, and A_2A_1. If A_2 is dominant over A_1, however, we cannot distinguish the three genotypes, A_1A_2, A_2A_1, and A_2A_2. Therefore, we can only observe two possible genotypes for a particular locus. The two possible genotypes are represented by A_1A_* and A_2A_2 if A_1 dominates over A_2, or A_2A_* and A_1A_1 if A_2 dominates over A_1. Allele A_* is a wild card and can be either A_1 or A_2. When A_1 dominates over A_2, we use $A = (1, 2, 3)$ to code genotype A_1A_* and $A = 4$ to code genotype A_2A_2. If A_2 dominates over A_1, we use $A = 1$ to code genotype A_1A_1 and $A = (2, 3, 4)$ to code genotype A_2A_*. The numerical code for each locus is then converted into an appropriate diagonal matrix, D_A, D_B, or D_C, for calculating the joint probability of a joint three-locus genotype.

If the genotype for a locus, say locus A, is missing, the numerical code for the locus is $A = (1, 2, 3, 4)$, and the corresponding diagonal matrix D_A is simply a 4×4 identity matrix. Missing marker genotypes are treated the same way as genotypes of uninformative loci.

4.3 Conditional Probability of a Missing Marker Genotype

An important application of the three-point analysis to genetic mapping is to calculate the probability of genotype of a locus conditional on genotypes of flanking markers. Note that flanking markers are the two nearby markers of a locus, one in each side. Consider three loci, ABC, where A and C are two markers with known genotypes and B is a locus whose genotype is not observable. The conditional probability of genotype of locus B is

$$\Pr(B|A, C) = \frac{\Pr(A, B, C)}{\Pr(A, C)}. \tag{4.10}$$

The joint probability of the three-locus genotype in the numerator can be rewritten as

$$\Pr(A, B, C) = \Pr(B) \Pr(A, C|B) = \Pr(B) \Pr(A|B) \Pr(C|B).$$

We are able to write $\Pr(A, C|B) = \Pr(A|B) \Pr(C|B)$ because conditional on the genotype of B, the genotypes of A and C are independent due to the Markovian property of Mendelian loci. The joint probability of the two-locus genotype in the denominator of (4.10) is expressed as

$$\Pr(A, C) = \sum_{B=1}^{4} \Pr(A, B, C) = \sum_{B=1}^{4} \Pr(B) \Pr(A|B) \Pr(C|B).$$

Eventually, the conditional probability is expressed as

$$\Pr(B|A, C) = \frac{\Pr(B) \Pr(A|B) \Pr(C|B)}{\sum_{B=1}^{4} \Pr(B) \Pr(A|B) \Pr(C|B)} \tag{4.11}$$

We realize that $\Pr(A|B)$ and $\Pr(C|B)$ are the transition probabilities and $\Pr(B = k) = \frac{1}{4}$, $\forall k = 1, \ldots, 4$, is the marginal probability. The conditional probability expressed this way (4.11) is an expression of Bayes' theorem.

We now use matrix notation to express the conditional probability. Assume that we want to calculate $\Pr(B = k|A, C)$, $\forall k = 1, \ldots, 4$, where the genotypes of loci A and C are known and represented by matrices D_A and D_C. Since marker B is treated as a missing marker, its genotype is represented by $D_B = I_{4\times 4}$, an identity matrix. The matrix version of the numerator of (4.11) is

$$\Pr(B = k)\Pr(A|B = k)|\Pr(C|B = k) = \frac{1}{4}J'D_A T_{AB} D_{(k)} T_{BC} D_C J, \quad (4.12)$$

where $D_{(k)}$ is a diagonal matrix with all elements equal to zero except the element at the kth row and the kth column, which is unity. The matrix expression of the denominator of (4.11) is

$$\sum_{B=1}^{4} \Pr(B = k)\Pr(A|B = k)|\Pr(C|B = k) = \frac{1}{4}J'D_A T_{AB} D_B T_{BC} D_C J. \quad (4.13)$$

Therefore, the matrix expression of the conditional probability is

$$\Pr(B = k|A, C) = \frac{J'D_A T_{AB} D_{(k)} T_{BC} D_C J}{J'D_A T_{AB} D_B T_{BC} D_C J}. \quad (4.14)$$

We now use an F_2 progeny as an example to show how to calculate the conditional probabilities of a locus given genotypes of the flanking markers. Let A_1A_1 and (C_1C_2) be the genotypes of loci A and C, respectively. Recall that (C_1C_2) means that locus C is heterozygous, which has two forms, C_1C_2 and C_2C_1. We want to calculate the conditional probability that locus B is B_1B_1. The numerical codes for the genotypes of A and C are $A = 1$ and $C = (2, 3)$, respectively, which are translated into matrices of $D_A = D_{(1)}$ and $D_C = D_{(2)} + D_{(3)}$, respectively. Let $D_B = I_{4\times 4}$ because locus B is a missing marker. The numerator and the denominator of the conditional probability are

$$\begin{aligned} J'D_A T_{AB} D_{(1)} T_{BC} D_C J &= T_{AB}(1,1)(T_{BC}(1,2) + T_{BC}(1,3)) \\ &= 2(1 - r_{AB})^2 r_{BC}(1 - r_{BC}) \end{aligned}$$

and

$$\begin{aligned} J'D_A T_{AB} D_B T_{BC} D_C J &= \sum_{k=1}^{4} T_{AB}(1,k)(T_{BC}(k,2) + T_{BC}(k,3)) \\ &= 2r_{AC}(1 - r_{AC}) \end{aligned}$$

respectively. Therefore, the conditional probability is

$$\Pr[B = 1|A = 1, C = (2,3)] = \frac{2(1-r_{AB})^2 r_{BC}(1-r_{BC})}{2r_{AC}(1-r_{AC})}$$
$$= \frac{(1-r_{AB})^2 r_{BC}(1-r_{BC})}{r_{AC}(1-r_{AC})}$$

where

$$r_{AC} = r_{AB}(1-r_{BC}) + r_{BC}(1-r_{AB})$$

For the same genotypes of marker A and C, what is the conditional probability that marker B is heterozygous? This probability is represented by

$$\Pr[B = (2,3)|A = 1, C = (2,3)] = \frac{J' D_A T_{AB}(D_{(2)} + D_{(3)}) T_{BC} D_C J}{J' D_A T_{AB} D_B T_{BC} D_C J}$$

4.4 Joint Estimation of Recombination Fractions

The three-locus genotype distribution can be used to estimate r_{AB} and r_{BC} jointly. Again, let ABC be the three ordered loci under consideration. Assume that we have collected n progeny from a line cross family. The family can be a BC, an F_2, or an FW, but all represented by the generalized FW family so that the 4×4 transition matrix between consecutive markers applies to all designs. Let A^i be the numerical code for the genotype of individual i at locus A, $\forall i = 1, \ldots, n$, where A^i can take a subset of $\{1, 2, 3, 4\}$, depending on the actual genotype of individual i. The three-locus genotype is denoted by $A^i B^i C^i$. The corresponding diagonal matrices for the individual locus genotypes are denoted by D_A^i, D_B^i, and D_C^i, respectively. The joint three-locus genotype for individual i is

$$\Pr(A^i B^i C^i) \propto J' D_A^i T_{AB} D_B^i T_{BC} D_C^i J. \tag{4.15}$$

The equal sign is replaced by the sign of "proportional to" because the expression in the right-hand side of the equation differs from that in the left-hand side by a constant factor ($\frac{1}{4}$). The log likelihood function of the recombination fractions established from all the n individuals is

$$L(r_{AB}, r_{BC}) = \sum_{i=1}^{n} \ln(J' D_A^i T_{AB} D_B^i T_{BC} D_C^i J). \tag{4.16}$$

Explicit solutions for the ML estimates of the recombination fractions are possible if there are no missing genotypes of the markers. In this case, the above log likelihood function can be rewritten as

$$L(r_{AB}, r_{BC}) = \sum_{i=1}^{n} \ln T_{AB}(A_i, B_i) + \sum_{i=1}^{n} \ln T_{BC}(B_i, C_i). \tag{4.17}$$

The first term is simply a function of r_{AB}, and the second term is a function of r_{BC}, which are denoted by $L(r_{AB})$ and $L(r_{BC})$, respectively. Therefore, the log likelihood function for the three-point analysis is simply the sum of the two pairwise log likelihood functions,

$$L(r_{AB}, r_{BC}) = L(r_{AB}) + L(r_{BC}). \tag{4.18}$$

As a consequence, the three-point analysis provides identical results for the estimated recombination fractions as the pairwise analysis. Therefore, when markers are all fully informative, there is no reason to invoke the three-point analysis. The three-point analysis, however, can extract additional information from the data if partially informative markers are present or there are missing marker genotypes. One reason for the increased efficiency of the three-point analysis is the incorporation of the marker order. For the pairwise analysis of three markers, one would have to estimate r_{AC} also from the same data. However, the three-point analysis treats the estimated r_{AC} as a function of the other two recombination fractions, i.e., $\hat{r}_{AC} = \hat{r}_{AB} + \hat{r}_{BC} - 2\hat{r}_{AB}\hat{r}_{BC}$. Therefore, information about the order of the three markers has been incorporated implicitly in the three-point analysis.

In general, there is no explicit solution for the joint estimate of the two recombination fractions, unless all markers are fully informative and there are no missing marker genotypes. A general numerical algorithm, e.g., the simplex method of Nelder and Mead (1965), can be adopted here to search for the MLE of the parameters. For problems with two clearly bounded parameters, such as this one with $(0 < r_{AB}, r_{BC} < 0.5)$, we may even use the simple grid search algorithm, which guarantees that the global optimal solutions for the parameters are obtained.

4.5 Multipoint Analysis for m Markers

We have just learned the three-point analysis ($m = 3$) as a special case of the general multipoint analysis. We now extend the methods to situations where $m > 3$. Let us use $j = 1, \ldots, m$ to index the locus. We now have $m-1$ consecutive recombination fractions and thus $m - 1$ transition matrices. The recombination fraction between loci j and $j + 1$ is denoted by $r_{j(j+1)}$, and the corresponding transition matrix is denoted by $T_{j(j+1)}$. Let D_j be the diagonal matrix for the genotype of locus j. We now use $G_j = k, \forall k = 1, \ldots, 4$, to denote the numerical code for the genotype of the jth locus. Recall that there are four possible genotypes for the generalized four-way cross design. Again, D_j is a matrix version of the numerical code for the genotype of locus j with $D_j = D_{(k)}$ for $G_j = k$. For an ambiguous genotype like $G_j = (2, 3)$ or $G_j = (1, 2, 3, 4)$, the corresponding diagonal matrix is denoted by $D_j = D_{(2)} + D_{(3)}$ or $D_j = D_{(1)} + D_{(2)} + D_{(3)} + D_{(4)} = I_{4\times4}$, respectively.

We now discuss the joint distribution for the m locus genotype, the conditional distribution of a missing marker genotype given the observed genotypes of $m-1$ markers, and the log likelihood function for jointly estimating $m-1$ recombination fractions using m markers. The joint distribution of the m locus genotype is denoted by

$$\Pr(G_1, G_2, \ldots, G_m) = \frac{1}{4} J' D_1 T_{12} D_2 \ldots T_{(j-1)j} D_j T_{j(j+1)} \ldots D_{m-1} T_{(m-1)m} D_m J. \tag{4.19}$$

Assume that the genotype of the jth marker is missing. The conditioning probability of $G_j = k$ given the genotypes of all the $m-1$ markers is

$$\begin{aligned} &\Pr(G_j = k | G_1, \ldots, G_m) \\ &\quad = \frac{J' D_1 T_{12} D_2 \ldots T_{(j-1)j} D_{(k)} T_{j(j+1)} \ldots D_{m-1} T_{(m-1)m} D_m J}{J' D_1 T_{12} D_2 \ldots T_{(j-1)j} D_j T_{j(j+1)} \ldots D_{m-1} T_{(m-1)m} D_m J}. \end{aligned} \tag{4.20}$$

Recall that $D_j = I_{4\times 4}$ because j is the missing marker. The probability that $G_j = (2,3)$ is simply obtained by substituting $D_{(k)}$ in the numerator of the above equation by $D_{(2)} + D_{(3)}$. Let D_j^i be the matrix representation of G_j for individual i for $i = 1, \ldots, n$. The log likelihood function for estimating $\theta = \{r_{12}, r_{23}, \ldots, r_{(m-1)m}\}$ is

$$L(\theta) = \sum_{i=1}^{n} \ln J' D_1^i T_{12} D_2^i \ldots T_{(j-1)j} D_j^i T_{j(j+1)} \ldots D_{m-1}^i T_{(m-1)m} D_m^i J. \tag{4.21}$$

One property of the multipoint analysis is that

$$\Pr(G_j = k | G_1, \ldots, G_m) = \Pr(G_j = k | G_{j-1}, G_{j+1}), \tag{4.22}$$

if markers $j-1$ and $j+1$ are fully informative. Verbally, this property is stated as "the genotype of a marker only depends on the genotypes of the flanking markers." This can be proved by the following argument. If loci $j-1$ and $j+1$ are fully informative, the numerator of (4.20) can be rewritten as

$$H_l \times (J' D_{j-1} T_{(j-1)j} D_{(k)} T_{j(j+1)} D_{j+1} J) \times H_r \tag{4.23}$$

and the denominator of (4.20) can be rewritten as

$$H_l \times (J' D_{j-1} T_{(j-1)j} D_j T_{j(j+1)} D_{j+1} J) \times H_r, \tag{4.24}$$

where

$$H_l = J' D_1 T_{12} D_2 \ldots D_{j-1} J$$

and

$$H_r = J' D_{j+1} \ldots D_{m-1} T_{(m-1)m} D_m J.$$

Note that H_l and H_r are scalars and they appear in both the numerator and the denominator. Therefore, they are canceled out in the conditional probability, leaving

$$\begin{aligned}\Pr(G_j = k|G_1, \ldots, G_m) &= \frac{H_l \times (J' D_{j-1} T_{(j-1)j} D_{(k)} T_{j(j+1)} D_{j+1} J) \times H_r}{H_l \times (J' D_{j-1} T_{(j-1)j} D_j T_{j(j+1)} D_{j+1} J) \times H_r} \\ &= \frac{J' D_{j-1} T_{(j-1)j} D_{(k)} T_{j(j+1)} D_{j+1} J}{J' D_{j-1} T_{(j-1)j} D_j T_{j(j+1)} D_{j+1} J} \\ &= \Pr(G_j = k|G_{j-1}, G_{j+1}), \end{aligned} \tag{4.25}$$

which is the conditional probability we have learned in the three-point analysis.

4.6 Map Construction with Unknown Recombination Fractions

The multipoint analysis described so far has only been used when the order of the markers is known, in which only $m - 1$ recombination fractions are estimated. Recombination fractions between nonconsecutive markers are irrelevant and thus are not estimated. The recombination fraction between any two nonconsecutive markers can be obtained using Haldane map function if such information is required. Taking the map ABCD for example, the multipoint analysis only provides estimates for r_{AB}, r_{BC}, and r_{CD}. One can obtain the remaining three recombination fractions by $r_{AC} = r_{AB} + r_{BC} - 2r_{AB}r_{BC}$, $r_{AD} = r_{AC} + r_{CD} - 2r_{AC}r_{CD}$, and $r_{BD} = r_{BC} + r_{CD} - 2r_{BC}r_{CD}$. Alternatively, one may convert the recombination fractions into additive distances and join the additive distances to make an additive map, from which all pairwise recombination fractions can be calculated using the Haldane map function.

For those understudied species, marker orders may be unknown. The multipoint method provides a mechanism to order markers and estimate recombination fractions simultaneously. The marker order and the estimated recombination fractions under that order should be the joint ML estimates if such an order and the estimated recombination fractions under that order generate the maximum likelihood value compared to all other orders.