Chapter 3

Genetic Linkage

3.1 Chromosomes and linkage: data at linked loci

Where our data, are dependent on the alleles at more than one locus on a chromosome pair, it is no longer sufficient to consider the inheritance of genes at each locus separately. For ease of notation, we shall assume that the L loci of interest are ordered $1, \ldots, L$ along the chromosome. The components of the meiosis indicator vector for meiosis i, $S_{i*} = (S_{i,1}, \ldots, S_{iL})$, are dependent. We introduce also the notation $S_{*l} = (S_{1,l}, \ldots, S_{m,l})$ for the set of all meiosis indicators on the pedigree, at locus l. In considering the probability of data on related individuals in a pedigree:

$$L \;=\; \Pr(\text{data}) \;=\; \sum_{\mathbf{S}} \Pr(\text{data} \mid \mathbf{S}) \Pr(\mathbf{S}) \tag{3.1}$$

Often (although not always), our data will be specific to a given locus. Let Y_l denote the data pertaining to locus l, so the full data pertaining to this chromosomal region is $\mathbf{Y} = (Y_1, \ldots, Y_L)$, and

$$\Pr(\mathbf{Y} \mid \mathbf{S}) \;=\; \prod_l \Pr(Y_l \mid \mathbf{J}(S_{*l}))$$

where $\mathbf{J}(S_{*l})$ is the pattern of gene identity by descent among observed individuals, at locus l, which is determined by S_{*l}. Since meioses i are independent, equation (3.1) becomes

$$L \;=\; \Pr(\mathbf{Y}) \;=\; \sum_{\mathbf{S}} \left(\prod_l \Pr(Y_l \mid \mathbf{J}(S_{*l})) \right) \left(\prod_i \Pr(S_{i*}) \right) \tag{3.2}$$

To proceed further, we need a model for the vector S_{i*}. Such models may derive from our model for the process of meiosis (Chapter 4) or may be based on computationally convenient assumptions (section 3.4). However, in either case, it is the binary segregation indicators (1.2) which provide a means to trace the descent and ancestry of genes, at multiple linked loci. They determine patterns of gene-identity-by-descent (gibd), which in turn determine patterns of phenotypic similarity among relatives.

Somewhere define multi-locus genotype.

3.2 Two linked loci: recombination

Recall the segregation indicators of section 1.2 (equation (1.2)):

$$\begin{aligned} S_{il} &= 0 \quad \text{if copied gene at segregation } i \text{ locus } l \text{ is parent's maternal gene} \\ &= 1 \quad \text{if copied gene at segregation } i \text{ locus } l \text{ is parent's paternal gene.} \end{aligned}$$

Here $i = 1, ..., m$ indexes the segregations of the pedigree, and $l = 1, ..., L$ indexes the genetic loci. The marginal distribution of each S_{il} is as before: $\Pr(S_{il} = 0) = \Pr(S_{il} = 1) = \frac{1}{2}$. For different segregations i, the S_{il} are independent.

We say that, in a given meiosis, *recombination* has occurred between two loci j and l (or within the chromosomal interval bounded by loci j and l), if the genes segregating to the gamete at these two loci are from different paretal chromosomes. That is, they derive from different grandparents. For two loci, we do not need a full model for the S_{i*}; the pairwise distribution of $(S_{i,j}, S_{i,l})$ is determined by the *recombination frequency*, which is a measure of the dependence in inheritance between the two loci. For two given loci (l and j) the recombination frequency θ between them is

$$\theta \ = \ \Pr(S_{i,l} \neq S_{i,j}) \quad \text{for each } i, \quad 0 \leq \theta \leq \frac{1}{2}$$

Here we assume that θ does not vary with i. In practice, recombination frequencies vary among segregations, a major factor in this variation being the sex of the parent. Computationally, such variation can be easily accommodated. For loci that are close together on a chromosome, θ is close to 0. For independently segregating loci, $\theta = \frac{1}{2}$.

To introduce notation, for a chromosome with L ordered loci, $1, \ldots, L$, let the intervals between successive loci be $I_1, \ldots, I_{L-1}$. Let $R_j = 1$ if a gamete is recombinant on interval I_j, and $R_j = 0$ otherwise ($j = 1, \ldots, L-1$). Each vector $(R_1, \ldots, R_{L-1})$ determines two equiprobable vectors $S_{i*} = (S_{i,1}, \ldots, S_{i,L})$. A model for S_{i*} is equivalent to a model for $(R_1, \ldots, R_{L-1})$. The joint distributions of S_{i*} over more than two loci will be considered in section 3.4 in one special case, and more generally models for $(R_1, \ldots, R_{L-1})$ will be considered in Chapter 4.

3.3 Two-locus kinship and gene identity

The recursive equations for multiple kinship coefficients of section 2.3 (equations (2.2) and (2.3)) extend to multiple loci, conditioning on the segregation indicators in a given segregation, over the loci in question. Consider, for example, the case of $\psi_2(L_1(B^{(1)}, C), L_2(B^{(1)}, D))$. This expression denotes the two locus kinship probability, that, in a single gamete segregating from B, the gene at locus L_1 is IBD to that on a gamete segregating from individual C, while the gene at locus L_2 is IBD to that on a gamete segregating from individual D. The identical superscript "(1)" on the individual B indicates that we are considering here a single segregation i from B, rather than two separate segregations to different offspring. Now if B is not an ancestor of C or D, we may condition on the four events $(S_{i,1},\ S_{i,2}) = (0,0),\ (0,1),\ (1,1),\ (1,0)$ with probabilities $\frac{1}{2}(1-\theta),\ \frac{1}{2}\theta,\ \frac{1}{2}(1-\theta),\ \frac{1}{2}\theta$ respectively, where θ is the recombination frequency between locus 1 and locus 2. Thus we obtain

$$\begin{aligned} \psi_2(L_1(B^{(1)},\ C), L_2(B^{(1)},\ D)) \ &= \ \frac{1}{2}(1-\theta)\psi_2(L_1(M_B^{(1)}, C), L_2(M_B^{(1)}, D)) \ + \\ \tfrac{1}{2}\theta \quad & \psi_2(L_1(M_B, C), L_2(F_B, D)) + \frac{1}{2}(1-\theta)\psi_2(L_1(F_B^{(1)}, C), L_2(F_B^{(1)}, D)) \\ & + \ \frac{1}{2}\theta\psi_2(L_1(F_B, C), L_2(M_B, D)) \end{aligned}$$

The full set of equations for determining two-locus gene identity probabilities between genes segregating from up to four individuals are given by (Thompson 1988). These equations can be used to determine two-locus IBD state probabilities, even on a large and complex pedigree.

At two linked loci, there are also many more possible gene identity patterns (Denniston 1975). Some relationships which have identical *gibd* probabilities at a single locus can, in principle, be distinguished by data at linked loci. The simplest example is for the three unilineal ($k_2 = 0$) pairwise relationships of grandmother-granddaughter (G), half-sisters (H), and aunt-niece (N). Each of these relationships has $\mathbf{k} = (\frac{1}{2}, \frac{1}{2}, 0)$, and hence they are indistinguishable on the basis of data at independently segregating loci. For such relationships, gene identity at two linked loci is defined by

$$k_{11}(\theta) \quad = \quad P(share\ 1\ gene\ ibd\ at\ each\ of\ 2\ loci\ recombination\ \theta)$$

For the three relationships above we have

$$\begin{aligned}
G &: k_{11}(\theta) = \frac{1}{2}(1-\theta) \\
H &: k_{11}(\theta) = \frac{1}{2}(\theta^2 + (1-\theta)^2) = \frac{1}{2}R \\
N &: k_{11}(\theta) = \frac{1}{2}((1-\theta)R + \theta/2)
\end{aligned}$$

Thus the relationships are identifiable of the basis of data at two linked loci ($0 < \theta < \frac{1}{2}$), but not on the basis of data at unlinked loci. All the three relationships have $k_{11}(0) = \frac{1}{2}$ and $k_{11}(\frac{1}{2}) = \frac{1}{4}$.

Some relationships which provide identical two-locus kinship coefficients have different three-locus kinship coefficients (Thompson 1988). Thus, there are relationships that are non-identifiable on the basis of data at pairs of loci (whatever the values of the recombination frequencies between them), but that are identifiable on the basis of data at trios of loci. One may conjecture that there are relationships non-identifiable on the basis of data at L loci, but identifiable on the basis of data at $L+1$ loci.

3.4 Conditional independence model for the segregation indicators

Recall (section 3.2), the recombination indicators. In a given meiosis i,

$$R_j = 1 \text{ if } S_{i,j} \neq S_{i,j+1}, \text{ and} \tag{3.3}$$
$$R_j = 0 \text{ if } S_{i,j} = S_{i,j+1}, \quad j = 1, \ldots, L-1. \tag{3.4}$$

The simplest models for meiosis assume *no intrference*: this implies that the R_j are independent. Under this model, the dependence structure of the S_{il} takes a simple form, with a first-order Markov property over loci l, and with segregations i being independent. The probability of any given indicator S_{il} conditional on all the others, $\mathbf{S}_{-\{il\}}$, depends only on the indicators for the same segregation and the two neighboring loci:

$$\begin{aligned}
\Pr(S_{i,l} = s \mid \mathbf{S}_{-\{i,l\}}) &= \Pr(S_{i,l} = s \mid S_{i,l+1}, S_{i,l-1}) \\
&= \theta_{l-1}^{|s-S_{i,l-1}|}(1-\theta_{l-1})^{1-|s-S_{i,l-1}|}\theta_l^{|s-S_{i,l+1}|}(1-\theta_l)^{1-|s-S_{i,l+1}|}
\end{aligned} \tag{3.5}$$

for $s = 0, 1$, where $\theta_j = \Pr(R_j = 1) = \Pr(S_{i,j} \neq S_{i,j+1})$ is the recombination frequency between locus j and locus $j+1$. (For notational convenience we assume the loci to be numbered sequentially along the chromosome.) Note that equation (3.5), is just counting the recombination/non-recombination events implied in intervals $l-1$ and l, implied by the three indicators $(S_{i,l-1}, S_{i,l} = s, S_{i,l+1})$.

3.5 Multi-locus kinship and gene identity

Under the assumptions of conditional independence, computation and Monte Carlo are, in principle, straightforward. The recursive equations for two-locus kinship generalize to the multilocus case, although becoming progressively more horrendous. The probability of a *recombination pattern* in the intervals between marker loci is simple.

The process of $\{S_{i,l}\}$ is independent over segregations i, and in the absence of interference is Markov over a sequence of loci l along a chromosome. However, if the the resulting patterns of gene identity by descent among observed individuals that determine probabilities of observed data (equation (3.2)). Although the component $S_{i,l}$ are Markov over loci l, note that this is not necessarily so for the resulting patterns of gibd $\mathbf{J}(S_{*l})$ among observed individuals. Different values of S_{*l} may give rise to the same gibd pattern. Along the chromosome, gibd process is an agglomeration of the S_{*l} process. Grouping states of a Markov chain, does not, in general, produce a Markov chain.

As an example, consider again the pedigree of figure 2.1 (section 2.2). Summarize results. See (Thompson 1994) for more details.