CHAPTER 2. GENE IDENTITY BY DESCENT

2.1 Kinship and inbreeding coefficients

A *gene*, as opposed to an allele or a locus, is the *DNA segment* that is copied from parents to offspring. Underlying the patterns of phenotypes observed on related individuals are the *genotypes*, but underlying the genotypes are the patterns of gene identity by descent. Phenotypes of relatives are similar because they have similar genotypes and may share a common environment. Genotypes are similar because relatives share genes that are identical by descent (*ibd*) — identical copies of a gene segregating from a common ancestor within the defined pedigree. Disregarding mutation (which for modern microsattelite markers one probably shouldn't), genes that are *ibd* must be of the same allelic type, while genes that are not *ibd* are of independent allelic types.

Gene identity by descent is defined only within the context of a defined pedigree. A pedigree specifies the two parents of every non-founder individual. A founder has neither parent specified, and by definition the genes in founders are not *ibd*. It will often be convenient if a pedigree is ordered in such a way that every individual is preceded in the listing by his parents; this is clearly always possible.

Mendel's first law states that:

> a diploid individual receives at any given locus a copy of a randomly chosen one of the two genes in his father and (independently) a copy of a randomly chosen one of the two genes in his mother, and will pass on a copy of a randomly and independently chosen one of these two genes to each of his offspring.

This simple law leads to complex patterns of gene identity on an extended pedigree, due to the huge number of alternative events; 2^k for k segregations. The segregating genes determine the patterns of gene identity by descent on the pedigree, and hence the patterns of similarity among relatives.

We start with coefficients of *inbreeding* and *kinship*, since these provide an introduction to the ideas of gene identity by descent, to alternative computational approaches, and to Monte Carlo estimation of expectations. Kinship and inbreeding are best thought of as relationships between gametes rather than between individuals. The coefficient of kinship between two individuals B and C, $\psi(B, C)$, is the probability that homologous genes on gametes segregating from B and from C are *ibd*, while the inbreeding coefficient of an individual B, f_B is the probability that homologous genes on the two gametes uniting to form individual B are *ibd*. Hence

$$f_B = \psi(M_B, F_B)$$

where M_B and F_B are the parents of B. An individual is inbred if his parents are related. He is *autozygous* at a given locus if, at that locus, his two genes are *ibd*; his inbreeding coefficient is the prior probability of this event, based only on the pedigree.

2.2 Methods of computation

2.2.1 Path-counting

There are (at least) three methods for computing kinship coefficients. The early approach of *path-counting* (Wright 1922) simply enumerates all the possibilities (in an efficient way). Each path from the individual, B, to common ancestor, A, of his parents, descending via a disjoint set of individuals to B again contributes a term $2^{-(n_M+n_F+1)}(1 + f_A)$ to the inbreeding coefficient f_B, where n_M and n_F are the number of segregations in the maternal and paternal lines of the path. For example, for the

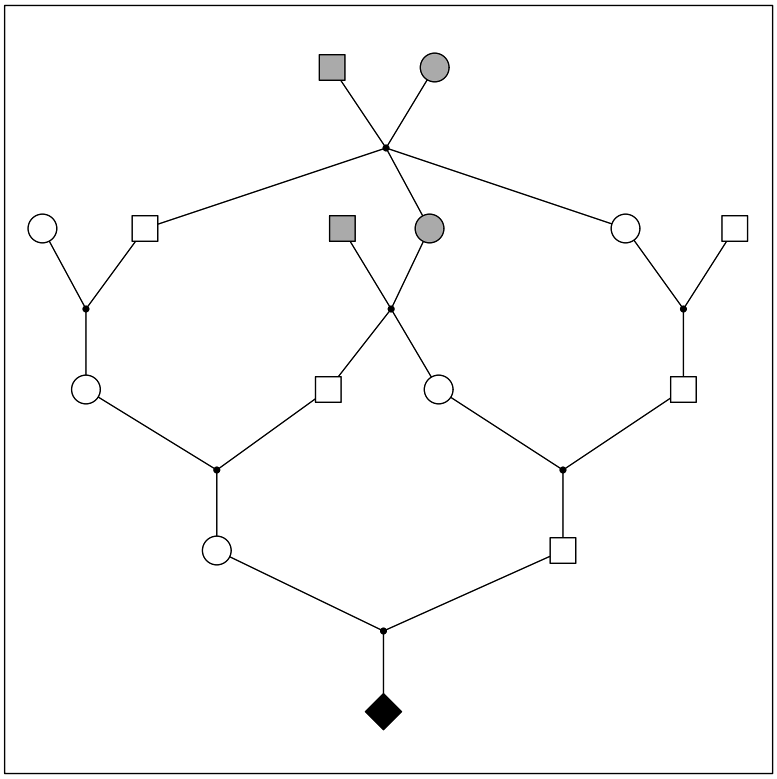

Figure 1: An example pedigree

offspring of a first cousin marriage, there are 2 paths, one via each of the two grandparents shared by his parents, each having $n_M = n_F = 2$, providing an inbreeding coefficient of $2 \times 2^{-5} = 1/16$. In the complex pedigree of figure 1, the common ancestors of the parents of the final individual are shaded grey. The individual is the offspring of a first cousin marriage, but so also is each of his parents. Here there are two paths via his great-grandparents, each having $n_M = n_F = 2$ as for the simple cousin marrriage, and 3 paths via each of his parents' 2 shared great-grandparents, each with $n_M = n_F = 3$, providing a total inbreeding coefficient of $2 \times 2^{-5} + 2 \times 3 \times 2^{-7} = 7/64$.

2.2.2 Recursive approach

There are equations for kinship which follow from the segregation indicators (1.2).

Provided B is not an ancestor of C, we may condition on the segregation S from B, where $\Pr(S=0) = \Pr(S=1) = \frac{1}{2}$. If $S = 0$, the segregating gene is B's maternal gene; that is, a gene from the mother of B. If $S = 1$, the gene is B's paternal gene. Thus we obtain immediately

$$\psi(B,C) = \psi(M_B,C)P(S=0) + \psi(F_B,C)P(S=1) = (\psi(M_B,C) + \psi(F_B,C))/2 \qquad (1)$$

where M_B and F_B are the mother and the father of B. Also, from the definition, we have symmetry: $\psi(B,C) = \psi(C,B)$. Thus the only additional equation needed is for the case $B = C$. In this case, we must consider two independent segregations from B, S_1 and S_2:

$$\Pr(S_1 = S_2) = \Pr(S_1 \neq S_2) = \frac{1}{2}$$

giving

$$\psi(B,B) = P(S_1 = S_2) + \psi(M_B,F_B)P(S_1 \neq S_2) = (1 + \psi(M_B,F_B))/2$$

Together with the boundary conditions

$$\begin{array}{rll} \psi(B,B) & = \frac{1}{2} & \text{for any founder } B, \\ \text{and } \psi(B,C) & = 0 & \text{if } B \text{ is a founder not an ancestor of } C, \end{array}$$

these equations determine the function $\psi()$ on the pedigree.

A recursive algorithm based on these equations is very easily implemented, and works well even on large and complex pedigrees. However, it is not necessarily computationally efficient; the same expansion may be repeated many times. In principle, this can be avoided, by saving key $\psi(C, D)$, but the simplicity of the method is then lost.

2.2.3 Sequential computation

Equations need not determine the computational procedure. Another way to implement these equations is via a top-down sequential brute-force matrix method, computing kinship coefficients between ancestors arriving finally at the descendant individuals of interest. This is computationally trivial, but expensive on store. All computation is a trade-off between time and store.

2.2.4 Monte Carlo estimation of inbreeding coefficients

The earliest Monte Carlo estimation on pedigrees was to estimate inbreeding coeffcients. Edwards (1967) dropped genes down pedigrees to estimate inbreeding coeficients. Wright and McPhee (1925) traced random paths up pedigrees for the same purpose.

2.3 Multi-gamete kinship coefficients

An important extension to these equations was made by Karigl (1981), who considered the probability of simultaneous identity by descent, $\psi(B_1, ..., B_k)$, of genes segregating from a set of (not necessarily distinct) individuals $B_1, B_2, ..., B_k$. Again if B_1 is not an ancestor of any of $B_2, ..., B_k$, conditioning on the segregation from B_1 gives

$$\psi(B_1, B_2, ..., B_k) \;=\; \frac{1}{2}\Big(\psi(M_{B_1}, B_2, ..., B_k) \;+\; \psi(F_{B_1}, B_2, ..., B_k)\Big) \qquad (2)$$

The symmetry of the definition provides that we may collect the arguments for some B_1 who is not an ancestor of any of the others to the first t arguments of ψ. Then, considering the t independent segregations from B_1, either the segregating gene is the same in every case, being a random gene from B_1, or both the maternal and the paternal genes of B_1 are among the t genes. Since

$$\Pr(S_1 = S_2 = ... = S_t) \;=\; 2^{-t+1},$$

we obtain

$$\begin{aligned}\psi(B_1, ..., B_1, B_2, ..., B_k) &= 2^{-t+1}\Big(\psi(B_1, B_2, ..., B_k) \;+ \\ &\qquad (2^{t-1} - 1)\;\psi(M_{B_1}, F_{B_1}, B_2, ..., B_k)\Big) \\ &= 2^{-t}\Big(\psi(B_1, B_2, ..., B_k) \;+ \\ &\qquad (2^{t-1} - 1)\;\psi(M_{B_1}, F_{B_1}, B_2, ..., B_k)\Big) \qquad (3)\end{aligned}$$

Together with symmetry and boundary conditions, these equations determine the multiple kinship coefficients on any pedigree, although practical implementation can be problematic on a large multi-generation pedigree if $k \geq 7$.

2.4 Patterns of gene IBD in pairs of individuals

Karigl (1981) was interested primarily in the case $k = 4$, and in the determination of the probabilities of patterns of IBD, J, among the four genes of two individuals, at a single genetic locus.

Between inbred individuals there are 15 states of gene identity at a single autosomal locus (Cotterman 1974). These correspond simply to the number of partitions of the four genes into classes of genes that are *ibd*. Ordering the individuals, and the two genes within each, states can be labelled by labelling the first gene 1, and labelling each successive gene with the same label as any previously labelled gene to which it is *ibd*, and with the next available integer if it is not *ibd* to any previously labelled gene. However, there are only 9 genotypically relevant classes of states, since with regard to genotypes the maternal and paternal origins of genes are irrelevant, so the identities of the two genes within each individual can be interchanged.

For two non-inbred diploid individuals, there are three possible *gene identity states* at a single autosomal locus. That is, the individuals can have both genes *ibd*, or one, or neither. These events have probabilities $(k_2,\; k_1,\; k_0)$ say, $(k_2 + k_1 + k_0 \;=\; 1)$, determined by the pedigree. Individuals are related if $k_0 < 1$. Each relationship may thus be represented by a point in an equilateral triangle of unit height, the vertices corresponding to unrelated pairs ($k_0 = 1$), parent-offspring ($k_1 = 1$), and the identity (monozygous twin) relationship ($k_2 = 1$). The kinship coefficient is the probability that homologous genes segregating from two individuals are identical by decent and thus $\psi \;=\; (2\;k_2 \;+\; k_1)/4$.

While each relationship determines a point $\mathbf{k}$, the converse is not true, not only may several relationships give the same point $\mathbf{k}$, but some points are not (even in the limit) attainable by any relationship. In fact, it can be shown that $k_1^2 \geq 4k_0k_2$ (Thompson 1986). This result follows from the fact that, for non-inbred individuals

$$\begin{aligned}\psi &= (1/4)(\psi_{mm} + \psi_{ff} + \psi_{mf} + \psi fm) \\ \text{and}\;\; k_2 &= (\psi_{mm}\psi_{ff} \;+\; \psi mf\psi_{fm})\end{aligned}$$

where the subscripted kinship coefficients are those between a parent (mother (m) or father (f)) of one individual, and a parent of the other. Then the arithmetic-geometric mean inequality gives

$$\begin{aligned}
4k_2 &\leq (\psi_{mm} + \psi_{ff})^2 + (\psi_{mf} + \psi fm)^2 \\
&\leq (\psi_{mm} + \psi_{ff} + \psi_{mf} + \psi fm)^2 \\
&= (4\psi)^2 \quad = \quad (k_1 \ + \ 2k_2)^2 \\
&= k_1^2 \ + \ 4k_2(k_1 + k_2) \quad \text{or} \\
4k_2k_0 &= 4k_2(1 - (k_1 + k_2)) \quad \leq \quad k_1^2
\end{aligned}$$

Relationships such as sibs and double-cousins of any given degree fall on the boundary parabola. Note that it is possible for the mother and father of each individual to be related to both the mother and the father of the other, without either individual being inbred. That is, all four of the cross-parental kinship coefficients in the above equation may be non-zero. The simplest example is that of quadruple-half-first-cousins, for which $\psi_{mm} = \psi_{ff} = \psi_{mf} = \psi fm = (1/2)$. This relationship gives $k_2 = 1/32,\ k_1 = 7/16,\ k_0 = 17/32$ and $\psi = 1/8$. The point in the triangle lies midway between that for half-sibs and for double-first-cousins, which also each have $\psi = 1/8$.

More details of the material of this section, and references to earlier work, can be found in Chapter 2 of (Thompson 1986).

2.5 Multigamete gene IBD specification and computation

Among larger sets of individuals, the number of possible states of *gibd* increases rapidly (Thompson 1974). For the 12 genes of 6 individuals, there are more than 4 million gene identity states (partitions of 12 ordered objects). However, there are only just over 198,000 genotypically distinct classes of states. Although this is not a small number, with modern computers and an efficient indexing of state classes it is not impossible to consider all the possible state classes given data on 6 individuals.

2.6 Observations on related individuals

Phenotypic similarities among relatives result from the genes they share IBD. Among an ordered set of genes, a partition of the set may be used to specify which subsets of the genes are IBD. Among a set of observed individuals, we denote this partition of their genes by $\mathbf{J}$, and refer to it as the *pattern* of gene identity by descent among the individuals. The segregation indicators $\mathbf{S} = \{S_i; i = 1, ..., m\}$ of equation (**??**) determine the pattern, $\mathbf{J}$, of genes IBD in any currently observed set of individuals; $\mathbf{J} = \mathbf{J}(\mathbf{S})$. The probability of any data (i.e. observed characteristics of the individuals) depends on $\mathbf{S}$ only through $\mathbf{J}(\mathbf{S})$, and we may write

$$\Pr(data) \quad = \quad \sum_{\mathbf{J}} \Pr(data \mid \mathbf{J}(\mathbf{S}))\ \Pr(\mathbf{J}) \quad = \quad \sum_{\mathbf{S}} \Pr(data \mid \mathbf{J}(\mathbf{S}))\ \Pr(\mathbf{S}) \tag{4}$$

In partitioning the likelihood in this way, the "genetic model" is separated from the effects of genealogical and genetic structure. The probability of a set of meiosis indicators $\mathbf{S}$ at a single locus is trivial; the components are independent, each 0 or 1 with probability 1/2. The probability of a given pattern $\mathbf{J}(\mathbf{S})$ depends on the genealogical relationship among the observed individuals: in principle it may be computed by the methods of sections 2.3 and 2.4. Given the gene identity pattern, $\mathbf{J}(\mathbf{S})$, the probability of data depends on the different types of genes, their frequencies, and how they affect observable phenotypes. Thus, the passage of genes in pedigrees provides the connection between observable genetic characteristics and the pedigree structure, whether we are estimating relationships from genetic data, estimating the genetic basis of traits knowing the pedigree, or inferring the ancestry and descent of particular genes, knowing both the genetic model and the data.

The finally we must consider the probability $\Pr(data \mid \mathbf{J}(\mathbf{S}))$, for a specified pattern of gibd among the observed individuals. The probability any distinct gene, j, is of allelic type $\alpha = a(j)$ is the population frequency of the allele α, and distinct genes j have independent allelic types. Thus, $\Pr(data \mid \mathbf{J}(\mathbf{S}))$ is the sum over all possible assignments $\mathcal{A}$ of allelic types to genes of the product of allele frequencies q_α of assigned alleles α:

$$\Pr(data \mid \mathbf{J}(\mathbf{S})) = \sum_{\mathcal{A}} \prod_{j} q_{a(j)}$$

This equation was given by Thompson (1974) who gave an example of ABO blood types on three individuals. The special case of two individuals (9 states $\mathbf{J}$) is discussed in Chapter 2 of Thompson (1986). In general, efficient determination of assignments $\mathcal{A}$ compatible with data is straightforward for genotypic data (e.g. DNA marker phenotypes) (see, for example, Thompson and Heath (1998)), and can be generalized to more complex phenotypes (Heath – ref ??).

Literature Cited

Fisher 1922
Fisher 1936
Mendel 1866

Cotterman CW (1974) A Calculus for Statistico-Genetics. Ph.D. Thesis 1940. Ohio State University. In PA Ballonoff, ed., *Genetics and Social Structure*. Academic Press, New York

Edwards AWF (1967) Automatic construction of genealogies from phenotypic information (AUTOKIN). Bulletin of the European Society of Human Genetics 1:42–43

Karigl G (1981) A recursive algorithm for the calculation of gene identity coefficients. Annals of Human Genetics 45:299–305

Thompson EA (1974) Gene identities and multiple relationships. Biometrics 30:667–680

— (1986) Pedigree Analysis in Human Genetics. Johns Hopkins University Press, Baltimore

Thompson EA, Heath SC (1998) Estimation of conditional multilocus gene identity among relatives. IMS Lecture Note Series In Press

Wright S (1922) Coefficients of inbreeding and relationship. American Naturalist 56:330–338

Wright S, McPhee HC (1925) An approximate method of calculating coefficients of inbreeding and relationship from livestock pedigrees. Journal of Agricultural Research 31:377–383