

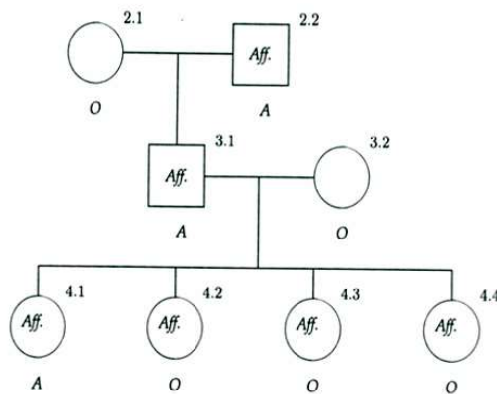
6

Pedigree Analysis: Structure, Notation, and Relatedness

6.1: Learning Objectives

By the end of this chapter, the reader should be able to draw and interpret a pedigree using standard symbols, recognise the inheritance pattern (autosomal dominant, autosomal recessive, X-linked) suggested by a pedigree, compute coefficients of relationship and Wright's coefficient of inbreeding (F) by path-counting, construct a kinship matrix for a small pedigree, and use the R package **kinship2** and Python to draw and analyse pedigrees computationally.

6.2: The Pedigree: Structure and Notation



A three-generation pedigree with standard symbols, showing genotypes and phenotypes segregating across generations.

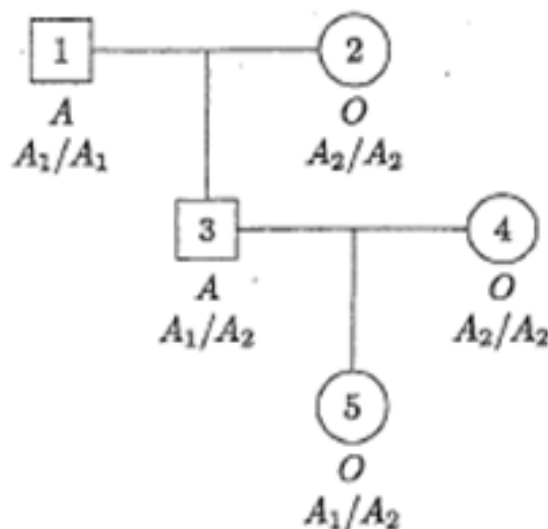
A **pedigree** is a diagram summarising the pattern of a trait's occurrence across the members of a family, over multiple generations. By convention: squares represent males, circles represent females, and a diamond is used when sex is not recorded; a filled (shaded) symbol denotes an affected individual, and an unfilled symbol an unaffected individual; a horizontal line connecting two symbols represents a mating; a vertical line descending from a mating line connects to the couple's offspring, drawn left to right in birth order; generations are numbered with Roman numerals (I, II, III, ...) from the top (earliest) generation downward, and individuals within a generation are numbered left to right with Arabic numerals (I-1, I-2, II-1, II-2, ...). A double horizontal mating line denotes a consanguineous union. A small circle placed above a symbol, or a slash through it, denotes a deceased individual, and an arrow marks the **proband** (propositus) — the affected individual through whom the family first came to clinical or research attention.

Reading inheritance patterns from a pedigree. Several visual clues distinguish the classical Mendelian modes of inheritance:

- **Autosomal recessive** (e.g., sickle-cell disease, most inborn errors of metabolism): affected individuals are typically born to unaffected (carrier) parents; the trait can skip generations; males and females are affected in roughly equal proportion; consanguinity in the parents raises the recurrence risk (Section 6.4).
- **Autosomal dominant** (e.g., Huntington disease, achondroplasia): affected individuals typically have an affected parent; the trait does not skip generations (barring reduced penetrance or a new mutation); males and females are affected equally; unaffected individuals do not transmit the trait.

- **X-linked recessive** (e.g., haemophilia, red-green colour blindness, G6PD deficiency): affected individuals are predominantly male; an affected male's daughters are obligate carriers, but his sons are unaffected (no father-to-son transmission); the trait typically passes from an affected grandfather through carrier daughters to affected grandsons.
- **X-linked dominant**: affected males transmit the trait to all daughters and no sons; affected heterozygous females transmit to half their children of either sex.

Worked Example 6.1 (Sickle-cell pedigree, autosomal recessive pattern)



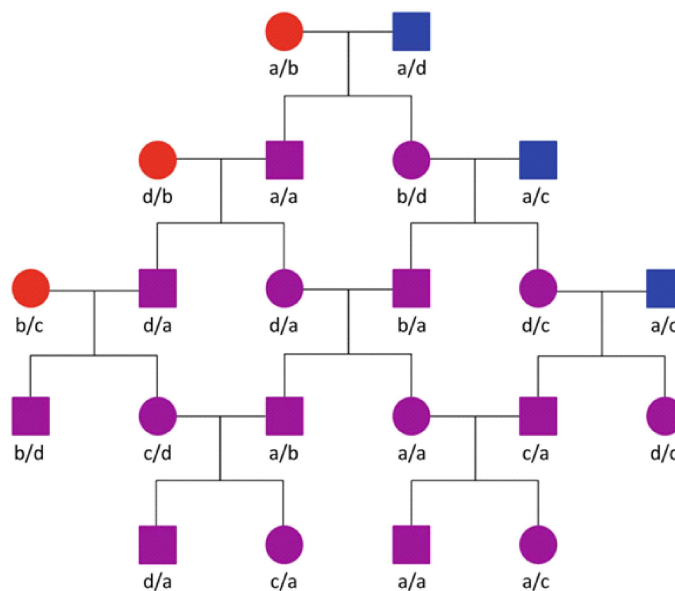
A worked pedigree with ABO and AK1 phenotypes, illustrating genotype inference from a family tree (Section 6.2).

Consider a three-generation Nigerian family pedigree in which the grandparents (generation I) are both AS (unaffected carriers); they have three children (generation II), one of whom, II-2, is affected (SS); II-2 marries an unrelated AS spouse, and they have

two children (generation III), one AA and one AS. The trait “skips” the appearance of affection in generation I despite both grandparents carrying the S allele, and both grandparents (obligate carriers) are phenotypically normal — a signature of autosomal recessive inheritance. Segregation analysis (Chapter 7) uses exactly such data, pooled across many families, to check whether the pattern of affected/unaffected offspring is statistically consistent with the hypothesised recessive mode.

Worked Example 6.2 (A four-generation pedigree with full genotypes)

Real pedigree analysis often records the full genotype (not just the phenotype) at a locus of interest for every family member, allowing exact tracking of allele transmission generation by generation.



A four-generation pedigree with genotypes fully labelled at every individual, allowing exact allele-transmission tracking across generations — the level of detail typically recorded in a modern genotyped family study.

In such a pedigree, each individual's two-letter genotype (e.g., a/b, d/c) can be read directly, and the rules of Mendelian transmission (Chapter 2) can be applied to verify, for any child, that each of their two alleles is consistent with having been inherited from one parent's genotype — a manual consistency check that is the conceptual ancestor of the automated **Mendelian error checking** step in modern genotyping quality-control pipelines (Chapter 19), which flags any parent–offspring genotype trio that is *not* consistent with simple Mendelian transmission as a likely genotyping error (or, occasionally, a sample mix-up or non-paternity) before any further statistical analysis is attempted.

6.3: The Coefficient of Relationship and Identity by Descent

The **coefficient of relationship** (r), also called the **kinship coefficient** when expressed as $\phi = r/2$, between two relatives measures the expected proportion of alleles they share **identical by descent** (IBD) — that is, inherited from a common ancestor via a traceable path of transmission, as distinct from alleles that merely happen to be identical in state by chance. For each path connecting two relatives through a common ancestor, the contribution to r is $(\frac{1}{2})^L$, where L is the number of meioses (generation-links, i.e. parent-to-child steps) along that path; if there are multiple independent paths (e.g., two common ancestors, as with full siblings sharing both parents), the contributions are summed:

$$r = \sum_{paths} \left(\frac{1}{2}\right)^L (1 + F_A)$$

where F_A is the inbreeding coefficient of the common ancestor A (taken as 0 unless A is itself known to be inbred).

Standard tabulated values. Parent–offspring: $r = \frac{1}{2}$ (one meiosis). Full siblings: $r = \frac{1}{2}$ (two paths, each $(\frac{1}{2})^2 = \frac{1}{4}$, through the shared mother and shared father, summing to $\frac{1}{2}$). Half-siblings: $r = \frac{1}{4}$ (one shared parent, path of two meioses). Grandparent–grandchild: $r = \frac{1}{4}$. Avuncular (uncle/aunt–nephew/niece): $r = \frac{1}{4}$. First cousins: $r = \frac{1}{8}$ (two paths through the two shared grandparents, each of four meioses, each contributing $(\frac{1}{2})^4 = 1/16$, summing to $1/8$). Double first cousins (both pairs of grandparents shared, as when two brothers marry two sisters): $r = \frac{1}{4}$.

6.4: Wright’s Coefficient of Inbreeding (F)

The **inbreeding coefficient**, F , of an individual is the probability that the two alleles the individual carries at a randomly chosen autosomal locus are identical by descent — i.e., that both alleles trace back to a single copy carried by a common ancestor of the individual’s *parents*. Sewall Wright’s path-counting formula states that F of an individual equals the coefficient of relationship of that individual’s two parents:

$$F(\text{offspring}) = r(\text{mother}, \text{father})/2 = \sum_{\text{paths}} \left(\frac{1}{2}\right)^L (1 + F_A)$$

where, for each path connecting the two parents through a shared ancestor A , L counts the number of individuals in the path linking the two parents through A (both parents included, offspring excluded), and F_A is A ’s own inbreeding coefficient.

Theorem 6.1 (Wright’s Path-Counting Formula)

Let X be an individual whose parents are M and F , connected by one or more paths through common ancestors. Then the inbreeding coefficient of X is

$$F_X = \sum_{\text{paths}} \left(\frac{1}{2}\right)^L (1 + F_A)$$

where the sum is over all distinct paths connecting M and F through a common ancestor A , L is the number of links (meioses) in the path, and F_A is the inbreeding coefficient of A .

Proof sketch. By definition, F_X is the probability that the two alleles X carries at a locus are identical by descent (IBD). This can only happen if both of X 's alleles trace back to a single copy carried by some common ancestor A of M and F . Consider one specific path of length L connecting M to F through A . The probability that a specific allele of A is transmitted, unchanged in identity, down one particular branch of this path to M (or to F) is $(1/2)$ per meiosis, by Theorem 1.1 (each meiosis transmits either of the two parental alleles with probability $1/2$, independently); the probability that the *same* allele of A reaches both M and F via this path, and is then transmitted from each of them to X , is therefore $(1/2)^L$, where L counts the total number of meioses along the full path from M through A to F . The factor $(1+F_A)$ corrects for the case where A is itself inbred: even when the *specific* path traced does not itself constitute the whole story, A 's own two alleles may already be IBD with probability F_A , in which case any single allele “chosen” at A is, in effect, doubly represented; a careful accounting (given in full in Wright, 1922) shows this contributes the multiplicative correction $(1+F_A)$. Since paths through different common ancestors (or genuinely distinct paths through the same ancestor) correspond to mutually exclusive ways for X 's two alleles to be IBD, their probabilities are added, giving the summation over all paths.

Worked Example 6.3 (Wright's formula applied to a Yoruba royal lineage marriage pattern)

Historically, some Yoruba ruling houses practised marriages designed to keep certain lineage ties close, occasionally including unions between children of two brothers' children (i.e., ordinary first cousins) as well as, in some documented

cases, more complex unions where the couple shared common ancestors through *two* independent routes (for example, if the couple's grandparents were themselves already first cousins to one another, rather than wholly unrelated). Suppose a couple are ordinary first cousins (two paths of length 4 each through the two shared grandparents, $F_A = 0$ for both, contributing $2 \times (1/2)^4 = 1/8$, matching Section 6.3), but it is additionally known that the shared paternal grandfather was himself the offspring of a first-cousin marriage one generation earlier, giving him an inbreeding coefficient of $F_A = 1/16$ (Worked Example 6.4) rather than 0. Applying Theorem 6.1, the path through this particular grandparent contributes $(1/2)^4 \times (1 + 1/16) = (1/16)(17/16) = 17/256$ instead of the simplified $1/16$, while the path through the (assumed non-inbred) grandmother still contributes $1/16 = 16/256$; the couple's coefficient of relationship becomes $r = 17/256 + 16/256 = 33/256 \approx 0.129$ (compared with exactly $1/8 = 0.125$ when both grandparents are assumed non-inbred) — a small but non-zero increase, illustrating how Theorem 6.1's correction factor $(1+F_A)$ captures the compounding effect of ancestral inbreeding across generations, a phenomenon of genuine relevance in lineages with a sustained historical pattern of close-kin marriage.

Worked Example 6.4 (First-cousin marriage, Northern Nigeria)

A couple are first cousins ($r = 1/8$ between them, Section 6.3); their shared grandparents are assumed non-inbred ($F_A = 0$). Since $F(\text{child}) = r(\text{parents})/2$:

$$F(\text{child of first cousins}) = (1/8)/2 = 1/16$$

This is the classical figure quoted in medical genetics: a child of first cousins has a $1/16$ (6.25%) chance, at any given autosomal

locus, of carrying two alleles identical by descent — raising the risk that any rare recessive disease allele carried by the shared grandparents will appear in homozygous (affected) form in the child, relative to the population baseline risk of q^2 under random mating with the same allele frequency q . This underlies the well-documented increase in birth defects and recessive disorders among the offspring of consanguineous marriages reported in genetic-counselling studies from Northern Nigeria and elsewhere in the sub-region where cousin marriage remains culturally common.

Worked Example 6.5 (Full-sibling union)

For a (hypothetical, and in reality prohibited) full-sibling mating, $r(\text{parents}) = 1/2$, so $F(\text{child}) = (1/2)/2 = 1/4$ — the same magnitude of risk as an $AS \times AS$ Mendelian cross, and four times the risk from a first-cousin union, illustrating why closer unions carry sharply higher genetic risk.

6.5: The Kinship Matrix

For a pedigree of n individuals, the **kinship matrix**, Φ , is the $n \times n$ matrix whose (i,j) entry, ϕ_{ij} , is the kinship coefficient between individuals i and j (with $\phi_{ii} = (1 + F_i)/2$ on the diagonal). The kinship matrix is the fundamental building block of modern pedigree-based statistical genetics: it defines the expected additive-genetic covariance between relatives used in variance-components linkage analysis (Chapter 9) and mixed-model GWAS in family-based samples (Chapter 18), and is computed recursively:

$$\phi_{ij} = \frac{1}{2}(\phi_{fj} + \phi_{mj})$$

for i a non-founder with parents f and m , and j not a descendant of i ,

with founders (individuals with no parents recorded in the pedigree) assigned $\phi_{ii} = 1/2$ (assuming non-inbred founders) and $\phi_{ij} = 0$ for unrelated founders.

Theorem 6.2 (Offspring Inbreeding Coefficient Equals Parental Kinship)

Let individuals X and Y have kinship coefficient $\phi(X,Y)$. If X and Y mate and produce an offspring Z , then Wright's inbreeding coefficient of Z equals $F(Z) = \phi(X,Y)$.

Proof. By definition, the inbreeding coefficient $F(Z)$ is the probability that the two alleles Z carries at a randomly chosen autosomal locus are identical by descent (IBD). Z receives one allele from X and one allele from Y . By definition, the kinship coefficient $\phi(X,Y)$ is precisely the probability that an allele drawn at random from X and an allele drawn at random from Y (at the same locus) are IBD. Since Z 's two alleles are exactly one randomly transmitted allele from each of X and Y , the probability that Z 's two alleles are IBD is identical to the probability that a random allele from X and a random allele from Y are IBD — that is, $F(Z) = \phi(X,Y)$ by direct correspondence of the two probabilistic definitions. This identity is what allows the kinship matrix (Section 6.5) to be built up recursively from founder individuals through the pedigree.

6.6: Chapter Summary

- A pedigree summarises trait transmission across family generations using standard symbols and numbering; the visual pattern of affection (which relatives are affected, and the sex distribution) suggests the likely mode of inheritance.
- The coefficient of relationship, r , quantifies expected allele sharing (IBD) between relatives via path-

counting; standard values range from $\frac{1}{2}$ for parent–offspring and full-sibling pairs down to $\frac{1}{8}$ for first cousins.

- The inbreeding coefficient, F , of an individual equals half the coefficient of relationship of that individual’s parents, and rises sharply with parental relatedness ($F = 1/16$ for children of first cousins; $F = 1/4$ for children of full siblings).

The kinship matrix, Φ , collects pairwise kinship coefficients for an entire pedigree and underlies variance-components methods used throughout later chapters.

6.7 Computational Corner: Building and Analysing Pedigrees in R and Python

R. The **kinship2** package is the standard tool for drawing pedigrees and computing kinship matrices in R.

```
# install.packages("kinship2")
library(kinship2)

# A small three-generation Nigerian pedigree (sickle-cell family,
# Worked Example 6.1)
id    <-c(1,2,3,4,5,6,7)
dadid <-c(0,0,1,1,0,3,3)
momid <-c(0,0,2,2,0,4,4)
sex    <-c(1,2,1,2,2,1,2)    # 1 = male, 2 = female
status <-c(0,0,0,1,0,0,0)    # 1 = affected (SS), 0 = unaffected

ped <-pedigree(id, dadid, momid, sex, affected = status)
plot(ped)                        # draws the pedigree with standard
                                # symbols

Phi <-kinship(ped)              # full kinship matrix
round(Phi, 4)
```

```

F_child <- 2*Phi["6", "7"] - 1 # not meaningful here (siblings);
for F of an
# individual, read the diagonal: 2*Phi[i,i]-1
diag(Phi)

```

Python. Python has no single dominant pedigree package as mature as *kinship2*; a pedigree can be represented as a directed graph with **networkx**, and the kinship coefficient computed recursively.

```

import networkx as nx

```

```

# parents[child] = (father, mother); None = founder
parents = {1: (None, None), 2: (None, None), 3: (1, 2), 4: (1, 2),
5: (None, None), 6: (3, 5), 7: (3, 5)}

```

```

phi_cache = {}

```

```

def kinship(i, j):
    if i > j:
        i, j = j, i
    if (i, j) in phi_cache:
        return phi_cache[(i, j)]
    if i == j:
        f, m = parents[i]
        val = 0.5 if f is None else 0.5 * (1 + kinship(f, m))
    else:
        f, m = parents[j]
        if f is None:
            # j is a founder, unrelated to i unless i==j
            val = 0.0
        else:
            val = 0.5 * (kinship(i, f) + kinship(i, m))
        phi_cache[(i, j)] = val
    return val

```

```
for pair in [(1, 2), (3, 4), (6, 7), (1, 6)]:
print(pair, "->", round(kinship(*pair), 4))
```

Both implementations return $\phi = 0.25$ for the full-sibling pair (individuals 6 and 7), corresponding to $r = 2\phi = 0.5$, exactly matching the path-counting result of Section 6.3, and confirm that a computational pedigree tool is simply an automated, error-checked version of the by-hand path-counting rules developed above — indispensable once pedigrees grow beyond a handful of individuals, as in real Nigerian extended-family sickle-cell or consanguinity studies.

Mini-Project

Mini-Project 6: Building and Auditing a Multi-Generation Pedigree

Construct a three-generation pedigree with at least one instance of consanguinity and compute its key relatedness statistics.

- Step 1.** Draw (by hand or using the R/Python tools of Section 6.7) a three-generation pedigree including a first-cousin mating.
- Step 2.** Compute the coefficient of relationship between the first cousins using path-counting.
- Step 3.** Compute Wright's inbreeding coefficient F for their offspring.
- Step 4.** Construct the kinship matrix for all individuals in the pedigree.
- Step 5.** Write a short paragraph explaining how the inbreeding coefficient you calculated would affect the offspring's expected homozygosity at a marker locus.

Deliverable: *A pedigree diagram, relatedness/inbreeding calculations, the kinship matrix, and a short written interpretation.*

6.8: Practice Questions

- Q1.** Two individuals are double first cousins (sharing both sets of grandparents, i.e. two independent lines of cousin relationship, as occurs when two brothers marry two sisters). Using the path-counting rule, find their coefficient of relationship, r .
- Q2.** A genetic counsellor in Kano is evaluating a couple who are uncle and niece. Using the standard avuncular relationship $r = 1/4$, compute the inbreeding coefficient, F , of a child of this union, and compare it with the first-cousin result of Worked Example 6.4.
- Q3.** Draw (in words, describing the pedigree symbols and layout) a three-generation pedigree consistent with X-linked recessive inheritance of G6PD deficiency, in which an affected grandfather (I-1) has two daughters (II-1, II-2), both unaffected carriers, and II-1 has an affected son (III-1).
- Q4.** For the pedigree in Section 6.2 (Worked Example 6.1), use the R or Python code of Section 6.7 to compute the kinship coefficient between the two generation-III children (individuals 6 and 7) and confirm it equals the full-sibling value obtained by hand.
- Q5.** Explain, in one or two sentences, why the kinship coefficient between an individual and a founder to whom they are unrelated is exactly zero, while the kinship coefficient of a non-inbred founder with themselves is $1/2$ rather than 1.
- Q6.** In the pedigree of Worked Example 6.2, explain what it would mean, biologically, if a child's genotype contained an allele that matched *neither* of the corresponding parent's two

alleles at that locus, and name the modern quality-control procedure (Chapter 19) built on exactly this logic.

- Q7.** Two individuals share one grandparent only (i.e., they are half-first-cousins, each descended from a different child of that one shared grandparent, with no other shared ancestors). Using the path-counting rule of Section 6.3, find their coefficient of relationship, r .

