

1

Basic Biological Concepts in Genetics

1.1: Learning Objectives

By the end of this chapter, the reader should be able to: describe the structure of DNA and chromosomes; distinguish genotype from phenotype; define allele, locus, gene, and genetic marker; describe the modes of single-gene inheritance; and relate these concepts to Nigerian examples such as the sickle cell gene and the ABO blood group system.

1.2: Introduction

Statistical genetics is the branch of statistics concerned with the analysis of genetic data — that is, with using probability theory and statistical inference to study how traits and diseases are transmitted from parents to offspring, how genes are distributed within and between populations, and how genes are associated with disease. In Nigeria, where inherited disorders such as sickle cell disease (SCD), glucose-6-phosphate dehydrogenase (G6PD) deficiency, and albinism carry a heavy public-health burden, statistical genetics provides the quantitative tools needed to map, estimate, and predict the transmission of these conditions.

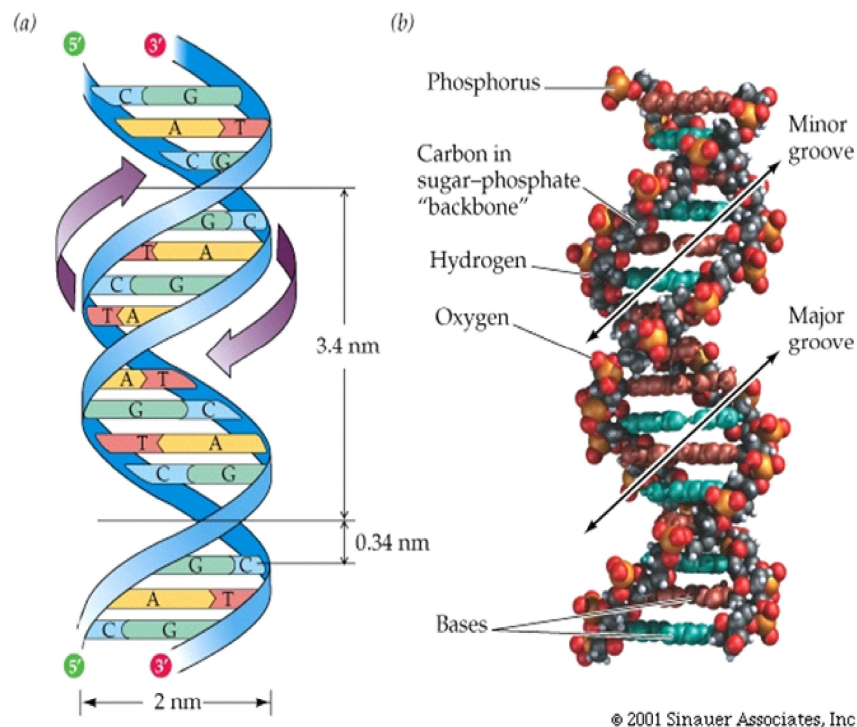
The major areas of interest in statistical genetics include:

- **Gene mapping:** locating genes on chromosomes, particularly disease genes, so that their DNA sequence and

protein products can eventually be studied and, where possible, targeted for treatment.

- **Population genetics:** describing how allele and genotype frequencies behave in populations under mating, migration, selection, and drift.
- **Quantitative genetics:** partitioning the variation in continuously measured traits (height, blood pressure, birth weight) into genetic and environmental components.
- **Association and linkage analysis:** testing whether genetic markers are associated with, or co-inherited with, disease.

1.3: The Cell, the Chromosome and DNA



The DNA double helix, showing the sugar-phosphate backbone and complementary base pairing.

Every human cell (apart from mature red blood cells) contains a nucleus in which the genetic material is stored. The DNA (deoxyribonucleic acid) molecule is packaged into thread-like structures called **chromosomes**. Each chromosome is a very long molecule of DNA wound tightly around proteins called histones, which give it structural support and allow it to be tightly compacted inside the nucleus.

Each chromosome has a constriction point called the **centromere**, dividing it into a short arm (the “p” arm, from the French *petit*) and a long arm (the “q” arm). Humans normally have 23 pairs of chromosomes — 46 in total. Twenty-two of these pairs are called **autosomes**, and they look alike in males and females. The 23rd pair are the **sex chromosomes**: females carry two X chromosomes (46,XX) while males carry one X and one Y (46,XY).

DNA structure: DNA is a double helix made of two complementary strands. Each strand is a chain of nucleotides, each nucleotide consisting of a phosphate group, a deoxyribose sugar, and one of four nitrogenous bases: adenine (A), thymine (T), guanine (G), or cytosine (C). The bases pair specifically across the two strands — A pairs with T, and G pairs with C — a rule known as **complementary base pairing**. This specificity is what allows DNA to be faithfully copied (replicated) each time a cell divides, and it underlies the amplification and sequencing technologies used in modern genetic laboratories, including those now operating in Nigerian teaching hospitals such as the University College Hospital (UCH), Ibadan.

1.4: Genes, Alleles, and Genetic Markers

A **gene** is a segment of DNA that carries the instructions for building a protein or a functional RNA molecule, or that regulates the activity of other genes. In humans, a gene may range from a few hundred to more than two million DNA bases in length. Every

individual carries two copies of almost every gene — one inherited from the mother and one from the father.

An **allele** is one of the alternative forms that a gene can take at a given position (locus) on a chromosome. A locus is said to be **polymorphic** if two or more alleles are found in the population at appreciable frequency (conventionally, the rarer allele has frequency at least 1%); otherwise the locus is **monomorphic**.

A **genetic marker** is a polymorphic DNA segment of known chromosomal location, used as a reference point (a “landmark”) in genetic mapping. Classical markers include blood groups and enzyme variants; modern markers include microsatellites and single nucleotide polymorphisms (SNPs).

A taxonomy of genetic variants: Modern genotyping and sequencing technology (Chapter 20) reveals several distinct classes of DNA variation, each with different statistical properties and uses:

- **Single nucleotide polymorphisms (SNPs):** a change at a single DNA base pair (e.g., an A where most chromosomes carry a G). SNPs are by far the most common form of human genetic variation (occurring, on average, roughly once every 300 base pairs genome-wide) and are the workhorse marker of modern association studies (Chapters 11, 12, 18) because they are abundant, stable, and cheap to genotype on high-density arrays.
- **Insertions and deletions (indels):** the gain or loss of a short stretch of DNA sequence, ranging from a single base pair to a few hundred base pairs.
- **Copy number variants (CNVs):** larger duplications or deletions of DNA segments (often thousands to millions of base pairs), which can alter the number of functional copies of one or more genes.
- **Microsatellites (short tandem repeats, STRs):** runs of a short repeated motif (e.g., “CA” repeated 8 to 20 times),

which are highly polymorphic (many alleles per locus, distinguished by repeat number) and were the marker of choice for linkage analysis (Chapter 17) before SNP genotyping became affordable.

- **Structural variants:** larger-scale rearrangements including inversions and translocations, which can disrupt gene function or alter recombination patterns (Chapter 8) at the breakpoint.

Not every genetic variant affects a trait or disease: a variant may be **functional** (directly altering a protein's sequence or a gene's regulation) or merely a **marker in linkage disequilibrium** (Chapter 15) with a nearby, unobserved functional variant — a distinction central to every association study in this book.

Genotype refers to the (unordered) pair of alleles an individual carries at a locus — for example, at the sickle-cell (HBB) locus, an individual may be genotype AA (normal), AS (carrier, “sickle-cell trait”), or SS (sickle-cell disease). **Phenotype** refers to the observable characteristic that results, in part, from the genotype — together with environmental influences. An individual who is genotype AA and SS at the HBB locus is homozygous (identical alleles); one who is AS is heterozygous (different alleles).

Worked Example 1.1: (ABO blood group, Nigerian setting)

The ABO blood group locus has three alleles: I^A , I^B , and i . I^A and I^B are co-dominant with respect to each other, while both are dominant to i . Genotype $I^A I^A$ or $I^A i$ gives blood group A; $I^B I^B$ or $I^B i$ gives blood group B; $I^A I^B$ gives group AB; and ii gives group O. In a study of blood donors at a Lagos teaching hospital, suppose 620 donors out of 1,000 typed as group O, 210 as group A, 140 as group B, and 30 as group AB. These are phenotype (blood group) frequencies; Chapter 3 shows how the

underlying **allele** frequencies of I^A , I^B , and i can be estimated from such data under Hardy–Weinberg assumptions.

Worked Example 1.2: (Distinguishing marker classes)

A Nigerian genomics laboratory characterising a patient sample report: (i) a single base change at position chr11:5,227,002 (HBB); (ii) a 3.5 kb duplication spanning part of the *CYP2D6* gene, relevant to drug metabolism; and (iii) a CA-repeat marker on chromosome 13 with alleles of 8, 10, and 12 repeats observed across a family. Classify each: (i) is a SNP; (ii) is a copy number variant; (iii) is a microsatellite (STR). Only the first is potentially the functional (causal) variant itself in this example; the third would typically serve only as a linkage marker (Chapter 17), not a direct cause of any trait.

Theorem 1.1: (The Law of Equal Segregation, formal statement)

Let an individual be heterozygous Aa at an autosomal locus. Then, at meiosis, the two alleles A and a are transmitted to a gamete with equal probability: $P(\text{gamete carries } A) = P(\text{gamete carries } a) = 1/2$, and this probability is independent across the gametes produced (and, in particular, independent of the sex or birth order of any resulting offspring).

Proof sketch. At meiosis I, homologous chromosomes (one of maternal origin carrying A , one of paternal origin carrying a , in a heterozygote) align and then separate to opposite poles of the dividing cell, a process called **disjunction**. Because the two homologues are physically indistinguishable to the cellular machinery that pulls them apart (the spindle apparatus attaches to each via functionally identical centromere/kinetochore structures, with no mechanism discriminating “maternal” from “paternal” origin), each homologue is equally likely to be pulled to either pole. Since a given gamete receives the chromosome at whichever pole it derives from, and each pole is equally likely to carry A or

a, it follows that $P(\text{gamete carries } A) = P(\text{gamete carries } a) = 1/2$. Independence across successive meioses follows because each meiotic division is a physically separate event, with the outcome of disjunction in one cell having no causal influence on disjunction in another. “

This theorem is the formal, cytological justification for the probabilistic calculations used throughout Chapters 1–9 of this book — every Punnett square, every Hardy–Weinberg genotype proportion, and every recombination-fraction calculation rests on this single foundational fact about how chromosomes physically separate at meiosis.

Worked Example 1.3: (Applying the Law of Equal Segregation: a Yoruba family planning consultation)

A genetic counsellor at a Nigerian teaching hospital advises a Yoruba couple in which the husband is heterozygous AS at the sickle-cell locus and the wife is also AS. The couple asks: “Is our first child more likely to be unaffected because we already worry about the condition?” By Theorem 1.1, each parent transmits A or S to any given child with probability $1/2$, independently of anything the parents believe, hope, or have experienced with previous pregnancies (a formal restatement of the fact that meiosis has no “memory” — sometimes informally called the **gambler’s fallacy** trap in genetic counselling). The correct probabilities for this and *every subsequent* child, independently, are $P(AA) = 1/4$, $P(AS) = 1/2$, $P(SS) = 1/4$, regardless of the genotypes of any older siblings.

1.5: Relationships Between Alleles

At a diploid locus, the two alleles an individual carries may relate to one another in several ways:

- **Dominant/recessive.** A **dominant** allele produces its phenotypic effect whether it is present in one copy

(heterozygote) or two (homozygote). A **recessive** allele produces its effect only when present in two copies. Sickle-cell disease is recessive: only SS individuals are clinically affected, while AS carriers are usually healthy (though they carry “sickle-cell trait,” which confers partial protection against severe *Plasmodium falciparum* malaria — a classic example of **heterozygote advantage**, well documented in Nigerian populations).

- **Co-dominant.** Both alleles are expressed simultaneously and are separately detectable in the heterozygote, as in the AB blood group.
- **Additive.** The heterozygote’s phenotype is intermediate between the two homozygotes — common for many quantitative traits (Chapter 8).

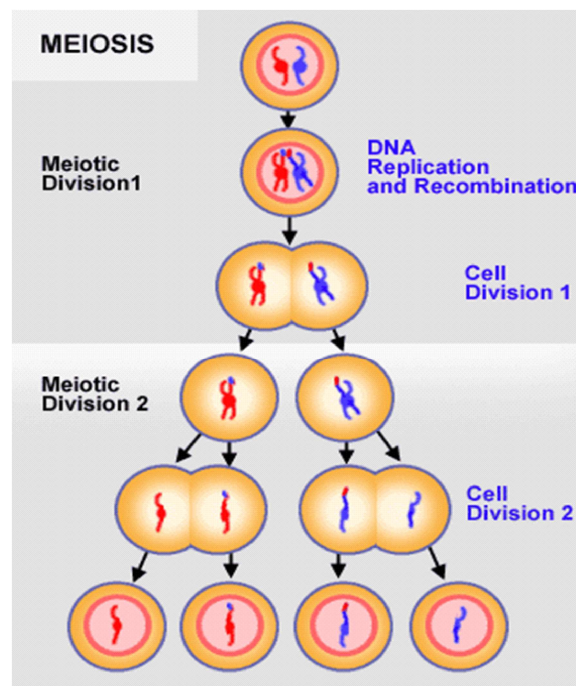
1.6: Autosomal versus Sex-Linked (X-linked) Inheritance



A normal human karyotype (22 autosome pairs plus the X and Y sex chromosomes).

Genes on the 22 autosomes are inherited without regard to the sex of the parent or offspring, and both sexes carry two alleles. Genes on the X chromosome behave differently: females (XX) carry two alleles at an X-linked locus, but males (XY) carry only one, because the Y chromosome carries almost no equivalent genes. A male is therefore said to be **hemizygous** at X-linked loci. This has an important statistical consequence, developed fully in Chapter 4: recessive X-linked conditions are far more common in males than in females. G6PD deficiency — common in Nigeria, especially in malaria-endemic areas, and important because it can trigger severe haemolysis in response to certain drugs (e.g., primaquine) and foods (e.g., fava beans) — is inherited in exactly this X-linked recessive manner.

1.7: The Inheritance of Chromosomes: Meiosis



The two divisions of meiosis, producing four genetically distinct haploid gametes from one diploid cell.

Gametes (sperm and egg cells) are produced by a special form of cell division called **meiosis**, in which the chromosome number is halved, so that each gamete carries only one member of each chromosome pair. When gametes from mother and father fuse at fertilisation, the full complement of 46 chromosomes (23 pairs) is restored in the offspring. Because which member of each pair goes into a given gamete is determined at random (independent assortment, Chapter 2), and because chromosomes may exchange segments through **crossing-over** (recombination) during meiosis, each gamete carries a unique combination of parental genetic material. This randomness is the biological basis for the probability laws that govern the transmission of genes — the subject of the rest of this book.

Theorem 1.2: (Independence of Transmission at Unlinked Loci)

Let an individual be heterozygous at two autosomal loci located on different chromosome pairs. Then the allele transmitted to a gamete at the first locus is statistically independent of the allele transmitted at the second locus.

Proof. By Theorem 1.1, transmission at each locus is governed by which of the two homologous chromosomes of the relevant pair is included in the gamete, and this is determined at metaphase I by the independent, random orientation of that chromosome pair on the meiotic spindle. Because the two loci lie on different (non-homologous) chromosome pairs, the orientation of one pair at metaphase I is physically unconnected to the orientation of the other pair — each pair aligns independently. Hence the two transmission events are outcomes of independent random mechanisms, and the joint probability of any combination of alleles at the two loci is the product of their individual transmission probabilities, establishing independence. This is the cytological

basis of Mendel's Law of Independent Assortment, formalised in probabilistic terms.

1.8: Chapter Summary

- The gene is the fundamental unit of heredity, physically a segment of DNA; alleles are alternative forms of a gene at a locus.
- Chromosomes are the vehicles of inheritance: 22 autosome pairs plus one pair of sex chromosomes (XX in females, XY in males) in humans.
- Genotype and phenotype are distinct; genotypes combine as homozygous or heterozygous, and alleles may act as dominant, recessive, co-dominant, or additive.
- X-linked inheritance differs fundamentally from autosomal inheritance because males are hemizygous.
- Meiosis and recombination introduce the randomness that statistical genetics formalises mathematically.

1.9: Computational Corner: Simulating Mendelian Transmission

R.

Simulating Mendelian transmission at the ABO locus for a Nigerian family

```
set.seed(762)
```

```
mother <- c("A", "O") # phenotype A (genotype AO)
```

```
father <- c("B", "O") # phenotype B (genotype BO)
```

```
# Correct genotype -> phenotype map: check the homozygous-O case FIRST,
```

```
# since a naive "subset of {A,O}" test would otherwise misclassify OO as A
```

```
genotype_to_phenotype <- function(g) {  
  s <- unique(g)
```

```

if(setequal(s, "O")) "O"
elseif(setequal(s, c("A", "B")) "AB"
elseif("A"%in%s) "A"
elseif("B"%in%s) "B"
else"O"}

# self-check against all four possible offspring genotypes before
# trusting the simulation
stopifnot(genotype_to_phenotype(c("O", "O")) == "O")
stopifnot(genotype_to_phenotype(c("A", "O")) == "A")
stopifnot(genotype_to_phenotype(c("O", "B")) == "B")
stopifnot(genotype_to_phenotype(c("A", "B")) == "AB")

# Simulate a large number of offspring (meiosis = one allele drawn
# from each parent)
n_offspring <- 2000
offspring <- replicate(n_offspring, c(sample(mother, 1),
sample(father, 1)))
offspring <- t(offspring)
phenotypes <- apply(offspring, 1, genotype_to_phenotype)

sim_table <- table(phenotypes)
sim_props <- prop.table(sim_table)
print(sim_table)
print(round(sim_props, 3))    # should each be close to 0.25

# Bar chart comparing simulated to theoretical (Mendelian)
# proportions
theoretical <- c(A=0.25, AB=0.25, B=0.25, O=0.25)
barplot(rbind(sim_props[names(theoretical)], theoretical),
beside=TRUE,
col = c("steelblue", "darkorange"), names.arg
=names(theoretical),
legend.text=c("Simulated", "Theoretical"), ylab="Proportion
of offspring",

```

```
main =paste("Simulated vs. theoretical ABO phenotypes (n =",
n_offspring, ")")
```

Python.

```
import random
from collections import Counter
import matplotlib.pyplot as plt
```

```
random.seed(762)
```

```
mother = ["A", "O"] # phenotype A (genotype AO)
```

```
father = ["B", "O"] # phenotype B (genotype BO)
```

```
def genotype_to_phenotype(g):
```

```
    """Map an unordered ABO genotype (two allele letters) to its
    phenotype.
```

```
    The homozygous-O case is checked FIRST: a naive 'is {O} a
    subset of {A,O}?'

```

```
    test would otherwise wrongly classify OO as phenotype A."""
```

```
    s = set(g)
```

```
    if s == {"O"}:
```

```
        return "O"
```

```
    if s == {"A", "B"}:
```

```
        return "AB"
```

```
    if "A" in s:
```

```
        return "A"
```

```
    if "B" in s:
```

```
        return "B"
```

```
    return "O"
```

```
# self-check against all four possible offspring genotypes before
trusting the simulation
```

```
assert genotype_to_phenotype(("O", "O")) == "O"
```

```
assert genotype_to_phenotype(("A", "O")) == "A"
```

```
assert genotype_to_phenotype(("O", "B")) == "B"
```

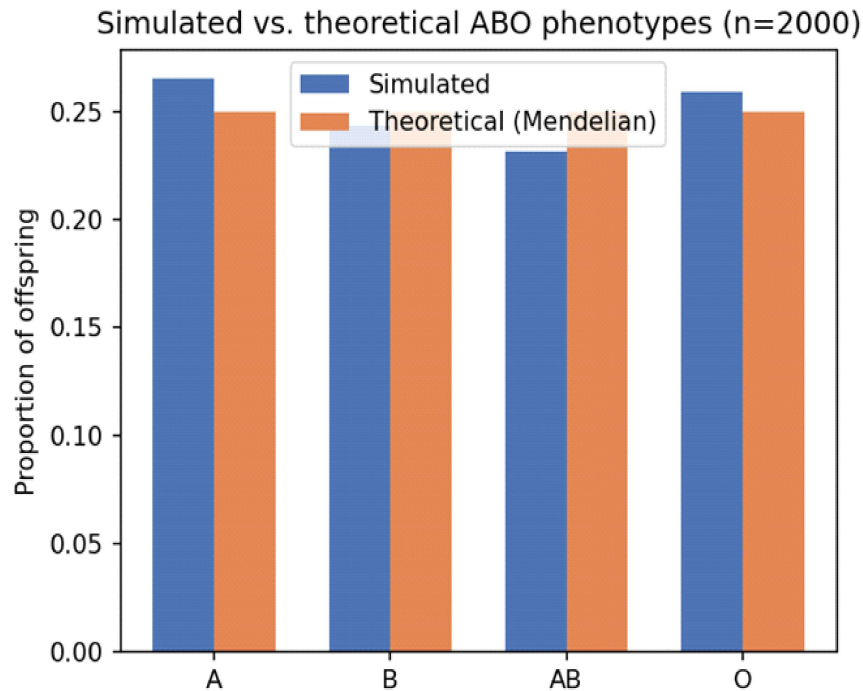
```
assert genotype_to_phenotype(("A", "B")) == "AB"
```

```

n_offspring = 2000
offspring = [(random.choice(mother), random.choice(father))
for _ in range(n_offspring)]
phenotypes = [genotype_to_phenotype(g) for g in offspring]
counts = Counter(phenotypes)
props = {k: v / n_offspring for k, v in counts.items()}
print("Simulated counts:", dict(counts))
print("Simulated proportions:", {k: round(v, 3) for k, v in
props.items()})

labels = ["A", "B", "AB", "O"]
theoretical = {l: 0.25 for l in labels}
x = range(len(labels))
width = 0.35
fig, ax = plt.subplots(figsize=(5, 4))
ax.bar([i - width/2 for i in x], [props.get(l, 0) for l in labels],
width,
    label="Simulated", color="#4C72B0")
ax.bar([i + width/2 for i in x], [theoretical[l] for l in labels],
width,
    label="Theoretical (Mendelian)", color="#DD8452")
ax.set_xticks(list(x)); ax.set_xticklabels(labels)
ax.set_ylabel("Proportion of offspring")
ax.set_title(f"Simulated vs. theoretical ABO phenotypes
(n={n_offspring})")
ax.legend()
plt.tight_layout()
plt.savefig("ch01_mendelian_simulation.png", dpi=150)

```



Simulated ABO phenotype proportions from 2,000 offspring of an $A \times B$ ($AO \times BO$) cross, closely tracking the theoretical 1:1:1:1 Mendelian ratio.

A worked debugging note: An earlier, more careless version of the `genotype_to_phenotype` function tested membership with `set(g) <= {"A", "O"}` (“is this genotype’s allele set a subset of {A, O}?”) to detect phenotype A. This test is a classic off-by-logic bug: the homozygous genotype OO has allele set {O}, and {O} is a subset of {A, O} — so every OO offspring was silently misclassified as phenotype A, inflating the simulated frequency of A and making phenotype O appear far too rarely (or never). The fix, shown above, checks the homozygous-O case *first*, before testing for A or B, so that OO is correctly resolved to phenotype O. This is a useful general lesson in defensive coding: when

classification rules are tested in sequence, always check the most specific or most easily “accidentally matched” condition first, and add a small self-check (as the *assert/stopifnot* lines above do) against every possible input before trusting a simulation’s output.

This simulation makes concrete the biological process described in Sections 1.4–1.7: each offspring genotype is formed by drawing exactly one allele from each parent (a simulated meiosis), and repeating the draw many times recovers, on average, the Mendelian proportions expected from the parental genotypes — the same logic that underlies every simulation-based method used later in the book (Chapters 6, 16 and 20).

Mini-Project

Mini-Project 1: Simulating and Verifying Mendelian Transmission

Extend the Section 1.9 simulation to verify, empirically, that the Law of Equal Segregation (Theorem 1.1) holds for a marker of your choice.

Step 1. Choose a hypothetical autosomal locus with two alleles and simulate 10,000 gametes from a heterozygous parent.

Step 2. Tabulate the observed proportion of each allele transmitted and compare it to the theoretical 1:1 ratio.

Step 3. Repeat the simulation for a heterozygous X-linked locus in a male versus a female parent, and note the structural difference.

Step 4. Perform a chi-square goodness-of-fit test (Section 2.5, previewed here) on your simulated transmission counts.

Step 5. Write two to three sentences relating your simulation result back to the proof of Theorem 1.1.

Deliverable: *Simulation code (R or Python), a summary table of observed vs. expected transmission ratios, and a short-written interpretation.*

1.10: Practice Questions

Q1. Distinguish between genotype and phenotype, using the sickle-cell locus (HBB) as your example.

Q2. A newborn screening programme in Enugu State finds that a baby is genotype SS at the HBB locus. State (a) the baby's phenotype and (b) whether the baby is homozygous or heterozygous at this locus.

Q3. Explain, in one sentence each, why (a) a male cannot be a “carrier” of an X-linked recessive disease in the same sense that a female can, and (b) G6PD deficiency is more common among Nigerian males than among Nigerian females.

Q4. A woman is genotype $I^A i$ at the ABO locus and her husband is genotype $I^B i$. List all possible genotypes, and the corresponding phenotypes, of their children.

Q5. Define “genetic marker” and give one example of a marker used in classical (pre-DNA-sequencing) human genetics in Nigeria.

Q6. A microsatellite marker on chromosome 15 has alleles of repeat length 6, 9, and 11. Explain why this marker, on its own, would be a poor choice for directly identifying the causal mutation for a disease, but could still be extremely useful for linkage mapping (Chapter 17).

Q7. Using the R or Python code of Section 1.9, change the parental genotypes to a cross between an AB individual and an OO individual, and predict (before running the code) what phenotype proportions you expect to see; then run the simulation to check your prediction.

