La-Net

BIO 201

GENETICS I

- Understand genes, inheritance and variation
- Explore patterns of inheritance and genetic analysis
- Build a strong foundation for modern genetics

Downoad on the App Store

GET IT ON Google Play

Course video is available on our app

lanetonline.com

Course code: BIO 201

Course title: Genetics I

Course credit load: 2 Units

Series 1: Heritable and Non-Heritable Characteristics

- **Part 1.1: Fundamentals of Heredity**
  - Sub Part 1.1.1: Definition and scope of genetics
  - Sub Part 1.1.2: Historical background of genetics
  - Sub Part 1.1.3: Chromosomes and the basis of heredity
- **Part 1.2: Heritable Characteristics**
  - Sub Part 1.2.1: Definition and nature of heritable characteristics
  - Sub Part 1.2.2: Mendelian inheritance: monohybrid crosses
  - Sub Part 1.2.3: Mendelian inheritance: dihybrid crosses
- **Part 1.3: Non-Heritable Characteristics**
  - Sub Part 1.3.1: Definition and nature of non-heritable characteristics
  - Sub Part 1.3.2: Environmental influences on phenotype
  - Sub Part 1.3.3: Distinguishing heritable from non-heritable traits
- **Part 1.4: Extensions of Mendelian Genetics**
  - Sub Part 1.4.1: Incomplete dominance and codominance
  - Sub Part 1.4.2: Multiple alleles and the ABO blood group system
  - Sub Part 1.4.3: Pleiotropy and epistasis

Introduction

Genetics is the branch of biology concerned with heredity and variation, the mechanisms by which traits are transmitted from parents to offspring, and the causes of differences and similarities among related individuals. At the heart of genetics lies the distinction between characteristics that are inherited through the genetic material and those that arise through the influence of the environment on gene expression. This distinction is fundamental to understanding both normal biological variation and the genetic basis of disease.

This Series establishes the conceptual and historical foundations of genetics, introduces chromosomes as the physical basis of heredity, distinguishes heritable from non-heritable characteristics with supporting evidence, presents the classical Mendelian model of inheritance through monohybrid and dihybrid crosses, examines the role of environment in shaping phenotypes, and extends the Mendelian framework to cover incomplete dominance, codominance, multiple alleles, pleiotropy, and epistasis. These topics provide the essential framework for the quantitative and population genetics topics that follow in later Series.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 1.1** define genetics and describe its scope, outline the key historical milestones in the development of genetics, and explain the chromosomal basis of heredity (Linked to Part 1.1),
- **SLO 1.2** define heritable characteristics and apply Mendel's laws to predict offspring ratios in monohybrid and dihybrid crosses (Linked to Part 1.2),
- **SLO 1.3** define non-heritable characteristics, describe the environmental influences that modify phenotype, and explain how heritable and non-heritable traits are distinguished experimentally (Linked to Part 1.3), and
- **SLO 1.4** explain incomplete dominance, codominance, and multiple alleles with examples, and describe how pleiotropy and epistasis extend the Mendelian model (Linked to Part 1.4).

1.1 Fundamentals of Heredity

1.1.1 Definition and scope of genetics

Genetics is the scientific study of genes, heredity, and genetic variation in living organisms. The word genetics was coined by William Bateson in 1905 from the Greek ‘genesis’ (origin). A gene is the fundamental unit of heredity: a segment of DNA that encodes information used to produce a functional product (usually a protein or structural RNA) that contributes to the phenotype of the organism. Heredity is the process by which traits are passed from parents to offspring through the transmission of genes. Variation refers to the differences in traits observed among individuals of the same species.

The scope of genetics is broad and interdisciplinary: classical (Mendelian) genetics studies the patterns of trait transmission from parent to offspring using crosses and statistical analysis; molecular genetics examines the structure, function, and regulation of genes and their products at the molecular level (DNA, RNA, protein); cytogenetics studies the structure and behaviour of chromosomes during cell division and their relationship to heredity; population genetics studies allele frequencies and genetic variation in populations over time; genomics is the large-scale study of the complete genetic content of organisms; epigenetics studies heritable changes in gene expression that do not involve changes in the DNA sequence itself. Genetics underlies all of biology: it explains evolution (through heritable variation and natural selection), development (through differential gene expression), disease (through mutation and genetic disorders), and crop and livestock improvement (through selective breeding and genetic engineering).

Fill in the blank: Genetics is the scientific study of genes, _____, and genetic variation; the word was coined by William Bateson in 1905; a gene is the fundamental unit of heredity and encodes information to produce a functional product that contributes to the phenotype.

1.1.2 Historical background of genetics

The history of genetics spans more than 150 years and includes several landmark discoveries that transformed biology. Gregor Mendel (1822 to 1884): an Augustinian monk working in Brünn (now Brno, Czech Republic), Mendel conducted carefully controlled cross-pollination experiments on garden peas (Pisum sativum) between 1856 and 1863, analysing the inheritance of seven pairs of contrasting characters; his two laws (the Law of Segregation and the Law of

Independent Assortment) were published in 1866 but were largely ignored until 1900. The Rediscovery of Mendel (1900): three botanists working independently, Hugo de Vries (Netherlands), Carl Correns (Germany), and Erich von Tschermak (Austria), each arrived at results similar to Mendel's and searched the earlier literature to find Mendel's pre-existing work; this rediscovery in 1900 launched the modern science of genetics. Walter Sutton and Theodor Boveri (1902 to 1903): independently proposed the chromosomal theory of inheritance, noting that the behaviour of chromosomes during meiosis parallels the segregation and independent assortment of Mendelian factors; this identified chromosomes as the physical carriers of Mendel's hereditary units.

Thomas Hunt Morgan and colleagues (1908 to 1920s): working with Drosophila melanogaster (fruit fly), Morgan demonstrated that genes are located on chromosomes and discovered sex-linkage, gene linkage, and genetic recombination; his group produced the first genetic maps; Morgan received the Nobel Prize in Physiology or Medicine in 1933. Frederick Griffith (1928): demonstrated the transforming principle in bacteria; heat-killed virulent Streptococcus pneumoniae could transform non-virulent bacteria into virulent forms, suggesting a heritable transforming substance. Oswald Avery, Colin MacLeod, and Maclyn McCarty (1944): identified DNA (not protein) as Griffith's transforming principle, providing the first evidence that DNA is the genetic material. Alfred Hershey and Martha Chase (1952): confirmed DNA as the genetic material using bacteriophage T2 and radioactive tracers. James Watson and Francis Crick (1953): proposed the double-helix structure of DNA based on X-ray crystallography data from Rosalind Franklin and Maurice Wilkins; this opened the molecular era of genetics. Nirenberg, Khorana, and colleagues (1961 to 1966): deciphered the genetic code. The Human Genome Project (1990 to 2003): sequenced the complete human genome (approximately 3.2 billion base pairs; approximately 20,000 to 25,000 protein-coding genes).

Fill in the blank: Gregor Mendel published his laws of inheritance in _____; his work was ignored until 1900 when it was independently rediscovered by Hugo de Vries, Carl Correns, and Erich von Tschermak; Watson and Crick proposed the double-helix structure of DNA in 1953.

1.1.3 Chromosomes and the basis of heredity

The chromosomal theory of heredity, proposed by Sutton and Boveri in 1902 to 1903, states that genes are located on chromosomes and that the behaviour of chromosomes during meiosis explains Mendel's laws. Evidence supporting the chromosomal theory: (1) Parallel behaviour: chromosomes segregate during meiosis (Meiosis I) in exactly the way that Mendel's hereditary factors segregate; homologous chromosomes from different pairs assort independently during Meiosis I, exactly as Mendel's law of independent assortment predicts. (2) Sex determination: the presence or absence of specific sex chromosomes (X and Y in humans and Drosophila) correlates precisely with sex-linked traits. (3) Linkage and recombination: genes on the same chromosome are linked (tend to be inherited together) unless separated by crossing over (recombination); the frequency of crossing over between two loci is proportional to the physical distance between them on the chromosome, enabling the construction of genetic maps.

Chromosome structure: each eukaryotic chromosome is a single, linear, double-stranded DNA molecule (complexed with histone proteins to form chromatin; further coiled into the visible chromosome structure during cell division). Humans have 46 chromosomes: 22 pairs of autosomes (identical in both sexes) and one pair of sex chromosomes (XX in females; XY in males). The karyotype is the complete set of chromosomes of an organism or cell, typically displayed in a standardised arrangement of homologous pairs ordered by size. Changes in chromosome number (aneuploidy: trisomy, monosomy) or structure (deletion, inversion, translocation, duplication) have profound effects on phenotype. Mitosis: cell division in somatic (body) cells that produces two genetically identical daughter cells (each with the same number of chromosomes as the parent cell; 2n); maintains the chromosome number. Meiosis: the specialised cell division that produces haploid gametes (1n) with half the chromosome number of the parent cell; consists of two successive divisions (Meiosis I: reductive division; separates homologous chromosomes; Meiosis II: equational division; separates sister chromatids); crossing over between homologous chromosomes during Prophase I of Meiosis I creates genetic recombination.

Fill in the blank: The chromosomal theory of heredity proposed by Sutton and Boveri states that genes are located on _____; the behaviour of chromosomes during meiosis parallels the segregation and independent assortment of Mendel's hereditary factors; humans have 46 chromosomes arranged in 23 pairs.

Pilot questions 1.1

1. Define genetics and state three sub-disciplines within it.
2. State four key milestones in the history of genetics between 1866 and 1953.
3. Explain the chromosomal theory of heredity and state two lines of evidence supporting it.
4. Distinguish between mitosis and meiosis in terms of the number of divisions, ploidy of daughter cells, and biological function.
5. Explain what a karyotype is and describe what information it can reveal about an individual.

1.2 Heritable Characteristics

1.2.1 Definition and nature of heritable characteristics

A heritable (or hereditary) characteristic is a trait that is encoded in the genetic material (DNA) and transmitted from parents to offspring through reproduction. Heritable traits are determined by the alleles present in the organism's genotype. Key terminology: gene (a specific DNA sequence that encodes a functional product and occupies a specific location (locus) on a chromosome); allele (one of two or more alternative forms of a gene that can occupy the same locus on homologous chromosomes); locus (the specific position on a chromosome where a gene is located; plural: loci); genotype (the complete genetic constitution of an organism; specifically the combination of alleles at one or more loci under consideration); phenotype (the observable characteristics of an organism resulting from the interaction of its genotype with the environment); homozygous (having two identical alleles at a locus: AA or aa); heterozygous (having two different alleles at a locus: Aa); dominant (the allele that is expressed in the phenotype of a heterozygote; conventionally written with a capital letter); recessive (the allele that is expressed only when homozygous; conventionally written in lower case); wild-type (the normal or most common allele or phenotype in a natural population).

Evidence that a characteristic is heritable: (1) It is transmitted from parents to offspring across generations in a predictable pattern. (2) The pattern of transmission matches the predictions of Mendelian genetics (or its extensions). (3) Identical (monozygotic) twins reared apart show greater trait similarity than non-identical (dizygotic) twins reared together, suggesting genetic

rather than environmental causation. (4) Selective breeding programmes successfully alter the trait in a population over generations, demonstrating a heritable basis. (5) Molecular genetics identifies specific DNA sequence variants (mutations, polymorphisms) associated with the trait. Examples of heritable characteristics: blood group (determined by ABO and Rh loci); sickle-cell disease (caused by a single nucleotide substitution in the beta-globin gene); eye colour (determined primarily by the OCA2 and HERC2 genes); height (substantially heritable but also influenced by nutrition); skin colour (polygenic; determined by multiple genes controlling melanin production).

Fill in the blank: A heritable characteristic is encoded in the _____ and transmitted from parents to offspring; the genotype is the complete genetic constitution of an organism; the phenotype is the observable characteristic resulting from the interaction of genotype with the environment.

1.2.2 Mendelian inheritance: monohybrid crosses

Gregor Mendel's experiments with pea plants established the fundamental rules of heredity. Mendel chose the garden pea (Pisum sativum) for his experiments because of its short generation time, ease of controlled pollination, many contrasting characters with clear-cut alternatives, and availability of pure-breeding (true-breeding) lines. He studied seven pairs of contrasting characters: seed shape (round vs wrinkled), seed colour (yellow vs green), pod shape (inflated vs constricted), pod colour (green vs yellow), flower colour (purple vs white), flower position (axial vs terminal), and stem height (tall vs dwarf). A monohybrid cross involves two parents that differ in a single pair of contrasting characters. Procedure for a monohybrid cross: (1) Parental generation (P): cross two true-breeding parents with contrasting phenotypes (e.g. round seeds AA x wrinkled seeds aa). (2) First filial generation (F1): all offspring are heterozygous (Aa) and show the dominant phenotype (round seeds); the recessive character has disappeared. (3) F1 self-fertilisation (F1 x F1): F2 offspring are in the ratio 1AA : 2Aa : 1aa (genotypic ratio) giving a phenotypic ratio of 3 round : 1 wrinkled.

Mendel's Law of Segregation (First Law): the two alleles of a gene separate (segregate) during gamete formation so that each gamete receives only one allele; the two alleles are reunited at fertilisation. Test cross: crossing an individual of unknown genotype (showing the dominant

phenotype) with a homozygous recessive individual (aa). If the unknown is AA: all offspring show the dominant phenotype. If the unknown is Aa: offspring are 1/2 dominant phenotype (Aa) : 1/2 recessive phenotype (aa). The Punnett square is a diagrammatic method for predicting the genotypic and phenotypic ratios of offspring from a cross. For the F1 x F1 cross (Aa x Aa): Punnett square gives 1/4 AA : 2/4 Aa : 1/4 aa; phenotypic ratio = 3 dominant : 1 recessive (3:1). Probability rules applied to genetics: product rule (the probability of two independent events occurring together = the product of their individual probabilities; used for combining the probabilities of alleles from each parent); sum rule (the probability that one of two mutually exclusive events occurs = the sum of their individual probabilities; used for combining probabilities of genotypes that give the same phenotype).

Fill in the blank: Mendel's Law of _____ states that the two alleles of a gene separate during gamete formation so that each gamete receives only one allele; the F2 ratio from a monohybrid cross (Aa x Aa) gives a phenotypic ratio of 3 dominant : 1 recessive (3:1).

1.2.3 Mendelian inheritance: dihybrid crosses

A dihybrid cross involves two parents that differ in two pairs of contrasting characters. Mendel's Law of Independent Assortment (Second Law): the alleles of two (or more) different genes assort independently of one another during gamete formation (i.e. the allele of gene A that enters a gamete is independent of which allele of gene B enters that same gamete). This law holds only for genes on different (non-homologous) chromosomes or genes far apart on the same chromosome (unlinked genes). Procedure: P: true-breeding round yellow seeds (RRYY) x true-breeding wrinkled green seeds (rryy). F1: all RrYy (round yellow; both dominant characters expressed). F1 x F1 dihybrid cross: RrYy x RrYy; F1 gametes: RY, Ry, rY, ry each at frequency 1/4; F2 Punnett square (4 x 4 = 16 combinations): phenotypic ratio of 9 round yellow : 3 round green : 3 wrinkled yellow : 1 wrinkled green (9:3:3:1). The 9:3:3:1 ratio is the expected F2 phenotypic ratio for a dihybrid cross involving two independently assorting genes with complete dominance. It can be derived by multiplying the two independent 3:1 ratios: (3 round : 1 wrinkled) x (3 yellow : 1 green) = 9:3:3:1.

Dihybrid test cross: RrYy x rryy; gametes from RrYy: RY, Ry, rY, ry; all gametes from rryy are ry; offspring ratio: 1 RrYy : 1 Rryy : 1 rrYy : 1 rryy = 1 round yellow : 1 round green : 1

wrinkled yellow : 1 wrinkled green (1:1:1:1); confirms independent assortment. Forked-line (branch diagram) method: an alternative to the Punnett square for dihybrid and higher-order crosses; uses the multiplication rule of probability to combine the results of each gene pair analysed separately; for each character: Aa x Aa gives 3/4 A_ : 1/4 aa; then multiply: 3/4 round x 3/4 yellow = 9/16 round yellow; 3/4 round x 1/4 green = 3/16 round green; 1/4 wrinkled x 3/4 yellow = 3/16 wrinkled yellow; 1/4 wrinkled x 1/4 green = 1/16 wrinkled green; total = 16/16.

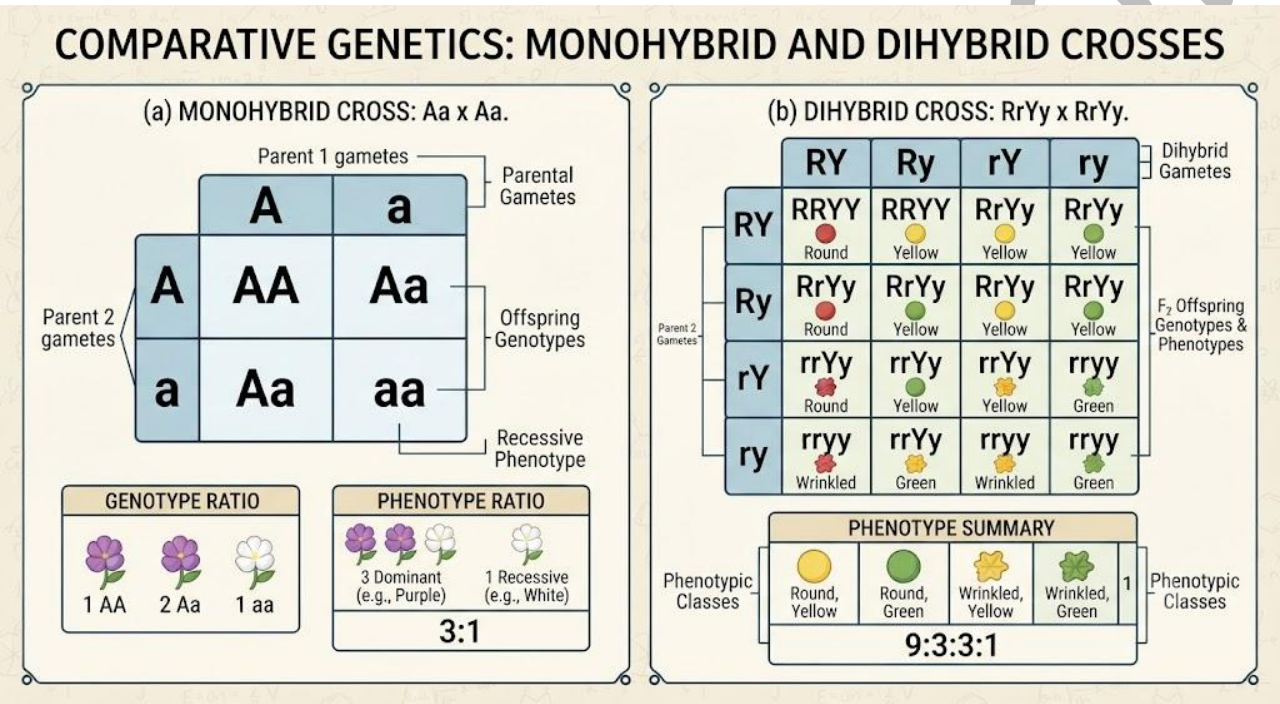

Figure 1.1: Punnett squares for (a) the monohybrid F2 cross (Aa x Aa) showing the 3:1 phenotypic ratio, and (b) the dihybrid F2 cross (RrYy x RrYy) showing the 9:3:3:1 phenotypic ratio

Pilot questions 1.2

1. Define the terms gene, allele, locus, genotype, and phenotype.
2. In peas, round seed shape (R) is dominant over wrinkled (r). Write the F2 genotypic and phenotypic ratios expected from crossing two F1 heterozygotes (Rr x Rr).
3. Explain Mendel's Law of Segregation.
4. A tall plant of unknown genotype (TT or Tt) is test-crossed with a dwarf plant (tt). All 40 offspring are tall. What is the most likely genotype of the tall parent? Explain.

5. In peas, round (R) is dominant over wrinkled (r) and yellow (Y) is dominant over green (y). What phenotypic ratio is expected in the F2 offspring of a dihybrid cross (RrYy x RrYy)?

1.3 Non-Heritable Characteristics

1.3.1 Definition and nature of non-heritable characteristics

A non-heritable (or acquired) characteristic is a trait that arises through the organism's interaction with its environment during its lifetime and is not encoded in the genetic material; it cannot be transmitted to offspring through the germ cells. Non-heritable characteristics arise through environmental influences on the phenotype without any change in the DNA sequence. The distinction between heritable and non-heritable traits is captured by the genotype-phenotype relationship: phenotype = genotype + environment (+ genotype x environment interaction). Because the environment can alter the expression of a gene without changing the DNA sequence (through epigenetic mechanisms), the boundary between heritable and non-heritable is not always sharp; however, in classical genetics, the following are considered non-heritable: (1) Acquired skills and learned behaviours (e.g. speaking a language, playing an instrument, riding a bicycle); these are encoded in neural connections formed by experience, not in the DNA. (2) Physically or chemically induced body modifications (e.g. a farmer's callused hands, a bodybuilder's enlarged muscles, scars from injury); these reflect tissue remodelling in response to use or damage, not heritable genetic changes. (3) Most environmentally induced phenotypic changes in individuals (e.g. skin tanning in response to UV exposure; stunted growth due to malnutrition; weight gain due to diet).

The Weismann barrier (proposed by August Weismann in 1892): there is a strict separation between the somatic (body) cells and the germ cells (cells that produce gametes); changes occurring in somatic cells during an individual's lifetime cannot be transmitted to the germ cells and therefore cannot be inherited by offspring; this principle refutes Lamarck's theory of inheritance of acquired characteristics (the idea that traits acquired during an organism's life can be passed to offspring). Epigenetic modifications (DNA methylation, histone modification)

represent a partial exception: some environmentally induced epigenetic marks in somatic cells can affect germ cells and be transmitted for a limited number of generations (transgenerational epigenetic inheritance); however, these marks are generally erased and reset in each generation, and the underlying DNA sequence is unchanged.

Fill in the blank: A non-heritable characteristic arises through the organism's interaction with its _____ during its lifetime and cannot be transmitted to offspring; the Weismann barrier separates somatic cells from germ cells, ensuring that somatic changes are not inherited.

1.3.2 Environmental influences on phenotype

Even for traits that are substantially heritable, the environment can profoundly influence the phenotype. This is expressed through the concept of the norm of reaction (reaction norm): the range of phenotypes that a given genotype can produce across different environmental conditions; the norm of reaction shows that the same genotype does not always produce the same phenotype. Examples of environmental influence on heritable traits: (1) Height in humans: height is substantially heritable (heritability $h^2 \approx 0.80$ in well-nourished populations); however, average human height has increased dramatically over the past 150 years in developed countries (secular trend) due to improved nutrition, healthcare, and living conditions, not genetic change; malnutrition during childhood causes permanent height reduction even in individuals with 'tall' genotypes. (2) Phenylketonuria (PKU): PKU is caused by a recessive mutation in the phenylalanine hydroxylase (PAH) gene; without treatment, affected individuals (genotype: pp) develop intellectual disability due to toxic accumulation of phenylalanine; however, when the diet is restricted in phenylalanine from early infancy, normal cognitive development proceeds; the genotype is heritable but the phenotype is profoundly modified by the environmental treatment.

(3) Himalayan rabbit and temperature-sensitive alleles: the Himalayan allele of the tyrosinase gene in rabbits produces the enzyme only at low temperatures; at body temperature (above about 33°C) the enzyme is inactive; at cooler extremities (ears, nose, paws, tail), pigment is produced (black); if a patch of white fur on the body is shaved and the skin chilled, black fur grows in that patch; the same genotype produces different pigmentation depending on local temperature. (4) Coat colour in Siamese cats: controlled by a similar temperature-sensitive allele; darker

colouration at the cooler extremities. (5) Hydrangea flower colour: the same genotype produces blue flowers in acidic soil (high aluminium availability) and pink flowers in alkaline soil; the soil pH modifies the expression of the colour gene. (6) Plant phenotypic plasticity: a single genotype of the aquatic plant Ranunculus aquatilis produces broad leaves above water and finely divided leaves underwater; a classic demonstration of phenotypic plasticity controlled by environmental signalling.

Fill in the blank: The norm of _____ is the range of phenotypes that a given genotype can produce across different environmental conditions; it demonstrates that the same genotype does not always produce the same phenotype; PKU is an example where a heritable genetic defect produces a normal phenotype when the environment (diet) is controlled.

1.3.3 Distinguishing heritable from non-heritable traits

Distinguishing whether a characteristic is heritable or non-heritable requires a combination of experimental approaches: (1) Twin studies: compare trait similarity between monozygotic (MZ; identical) twins and dizygotic (DZ; fraternal) twins; MZ twins share virtually all their DNA (100 percent); DZ twins share on average 50 percent of their DNA; if a trait is heritable, MZ twins should be more similar (higher concordance) than DZ twins; concordance rate is the percentage of twin pairs in which both members show the trait; if MZ concordance is higher than DZ concordance, the trait is at least partly heritable; if MZ concordance is 100 percent, the trait is fully determined by genetics (rare); if MZ concordance equals DZ concordance, the trait is entirely non-heritable (environmental). (2) Adoption studies: compare adopted children with their biological parents (shared genes, different environment) and with their adoptive parents (different genes, shared environment); greater similarity with biological parents suggests heritability.

(3) Selective breeding: if artificial selection on a trait successfully alters its frequency or magnitude in a population over generations, the trait is heritable; non-heritable traits do not respond to selection. (4) Heritability analysis: heritability (h^2) is a statistical measure of the proportion of phenotypic variance in a population that is due to genetic variance: $h^2 = V^G/V_p$, where V^G = genetic variance and V_p = total phenotypic variance; h^2 ranges from 0 (no genetic component; entirely environmental) to 1.0 (entirely genetic); h^2 is a property of a population in a

specific environment, not of a trait in isolation. (5) Molecular genetic approaches: identifying specific DNA variants (single nucleotide polymorphisms, SNPs; copy number variants, CNVs; mutations) that are consistently associated with a trait confirms a genetic component; genome-wide association studies (GWAS) scan the entire genome for SNPs associated with complex traits. (6) Experiments eliminating environmental variation: growing plants of different genotypes in identical controlled environments; if phenotypic differences persist, they reflect genotype; if phenotypic differences disappear, they were environmental.

Fill in the blank: Heritability (h^2) = V^G/V_p is the proportion of phenotypic variance due to _____ variance; $h^2 = 0$ means the trait is entirely environmental; $h^2 = 1.0$ means the trait is entirely genetic; h^2 is a property of a population in a specific environment, not of a trait in isolation.

Pilot questions 1.3

1. Define non-heritable characteristics and give three examples from everyday life.
2. Explain the Weismann barrier and state how it relates to the inheritance of acquired characteristics.
3. Describe the norm of reaction with one biological example.
4. Explain how twin studies are used to distinguish heritable from non-heritable characteristics.
5. Define heritability (h^2) and state what values of $h^2 = 0$ and $h^2 = 1.0$ indicate about a trait.

1.4 Extensions of Mendelian Genetics

1.4.1 Incomplete dominance and codominance

In classical Mendelian genetics, one allele (dominant) completely masks the expression of the other (recessive) in the heterozygote. However, many genes do not follow this simple dominant-recessive pattern. Incomplete dominance (partial dominance): neither allele is fully dominant; the heterozygote shows an intermediate phenotype between the two homozygotes. Classic example: flower colour in the four o'clock plant (Mirabilis jalapa) and snapdragon (Antirrhinum majus): RR = red flowers; R'R' = white flowers; RR' (heterozygote) = pink flowers (intermediate). F2 from RR' x RR': genotypic ratio = 1 RR : 2 RR' : 1 R'R'; phenotypic ratio = 1 red : 2 pink : 1

white (1:2:1); the phenotypic ratio equals the genotypic ratio (unlike complete dominance where the phenotypic ratio is 3:1). Explanation: in incomplete dominance, a single copy of the R allele produces only enough functional protein (e.g. pigment enzyme) for an intermediate phenotype; two copies of R are needed for the full red phenotype.

Codominance: both alleles are fully expressed simultaneously in the heterozygote; neither allele is dominant over the other; the heterozygote shows the phenotype of both homozygotes simultaneously, not an intermediate. Classic example: MN blood group in humans: genotype MM = M antigen on red blood cell surface; genotype NN = N antigen; genotype MN = both M and N antigens simultaneously (not an intermediate). Another example: roan coat colour in cattle: RR = red; R'R' = white; RR' = roan (a mixture of red and white hairs; when the individual hairs of a roan animal are examined, each hair is either entirely red or entirely white; the hairs are not pink; this distinguishes roan from incomplete dominance where the colour is intermediate). Distinction between incomplete dominance and codominance: in incomplete dominance, the heterozygote shows an intermediate blend of the two parental phenotypes; in codominance, the heterozygote shows both parental phenotypes simultaneously and distinctly. Both patterns produce a 1:2:1 phenotypic ratio in F2 crosses, but for different underlying reasons.

Fill in the blank: In incomplete _____, the heterozygote shows an intermediate phenotype between the two homozygotes (e.g. pink flowers in Mirabilis jalapa); in codominance, both alleles are fully expressed simultaneously in the heterozygote (e.g. MN blood type); both produce a 1:2:1 phenotypic ratio in F2 crosses.

1.4.2 Multiple alleles and the ABO blood group system

Multiple alleles: a gene is said to have multiple alleles when more than two alternative forms of that gene (alleles) exist in a population; however, any diploid individual can carry only two alleles (one on each homologous chromosome). The ABO blood group system: the best-known example of multiple alleles in humans, first described by Karl Landsteiner in 1900. The ABO locus on chromosome 9 has three principal alleles: I^A (also written I^A): encodes an enzyme (glycosyl transferase A) that adds a terminal N-acetylgalactosamine sugar to the H antigen on the surface of red blood cells, producing the A antigen; I^B: encodes a different glycosyl transferase

(B) that adds a terminal galactose sugar to the H antigen, producing the B antigen; i (also written i^0): a non-functional allele that does not modify the H antigen; blood group O (both red blood cells and serum lack A or B antigens; only H antigen present). Dominance relationships: I^A and I^B are both dominant over i (recessive); I^A and I^B are codominant with each other.

Genotype-phenotype table for ABO blood groups: I^AI^A or I^Ai = blood group A; I^BI^B or I^Bi = blood group B; I^AI^B = blood group AB; ii = blood group O. Blood group clinical significance: antibodies (agglutinins) are present in the plasma against the ABO antigens not present on the individual's own red blood cells (by convention: anti-A in group B; anti-B in group A; both anti-A and anti-B in group O; neither in group AB). Transfusion compatibility: group O (universal donor for red blood cells); group AB (universal recipient). ABO blood group inheritance in genetics problems: a couple both with blood group A could be I^AI^A x I^AI^A (all children group A), I^AI^A x I^Ai (all children A), or I^Ai x I^Ai (children: 1/4 group O possible); blood group typing is used in forensic genetics, paternity testing, and population studies. Other blood group systems: Rh (RhD antigen; Rh+ or Rh-; clinically important in haemolytic disease of the newborn); MN; Kell; Duffy; Kidd.

Fill in the blank: The ABO blood group system is controlled by three alleles at the same locus: I^A, I^B, and i; I^A and I^B are _____ to each other and both dominant over i; genotype I^AI^B gives blood group AB (both A and B antigens are expressed simultaneously); genotype ii gives blood group O.

1.4.3 Pleiotropy and epistasis

Pleiotropy: a single gene influences two or more apparently unrelated phenotypic traits. In pleiotropy, a mutation in one gene produces multiple phenotypic effects because the gene product (protein) is involved in multiple developmental or physiological pathways. Examples: (1) Sickle-cell disease: caused by a single nucleotide change in the beta-globin gene (HBB); the single mutation affects the shape of haemoglobin S under low oxygen tension, causing red blood cell sickling and a cascade of pleiotropic effects: haemolytic anaemia, vaso-occlusive crises, splenomegaly, increased susceptibility to infection, stroke risk, kidney disease, and retinal damage. (2) Marfan syndrome: caused by mutations in the FBN1 gene (fibrillin-1); single gene mutation produces phenotypic effects in multiple systems: skeletal (tall stature, long limbs,

arachnodactyly), cardiovascular (aortic aneurysm, mitral valve prolapse), and ocular (lens dislocation). (3) Phenylketonuria (PKU): PAH gene mutation causes hyperphenylalanemia that affects multiple systems: intellectual disability, seizures, abnormal EEG, eczema, unusual urine odour, and hypopigmentation of hair and skin.

Epistasis: the interaction between two or more genes such that the phenotypic expression of one gene (the hypostatic gene) is masked by or modified by the alleles of another gene (the epistatic gene) at a different locus. Epistasis differs from dominance (which involves allelic interactions at the same locus); epistasis involves interactions between alleles at different loci. Types of epistasis and their modified F2 ratios (from a dihybrid cross AaBb x AaBb): recessive epistasis (gene B is hypostatic; bb masks the expression of gene A; F2 ratio: 9 A_B_ : 3 A_bb : 3 aaB_ : 1 aabb becomes 9:3:4 when the aaB_ and aabb classes both show the same phenotype); dominant epistasis (at least one dominant allele of B epistatic over gene A; F2 ratio modified to 12:3:1); duplicate dominant epistasis (dominant allele at either locus gives same phenotype; 15:1); duplicate recessive epistasis (recessive at either locus masks phenotype; 9:7). Classic example of recessive epistasis: coat colour in Labrador retrievers (gene B: B_ = black; bb = brown/chocolate; gene E: E_ = colour expressed; ee = yellow regardless of B genotype; if ee, dog is yellow whether BB, Bb, or bb); F2 ratio from BbEe x BbEe: 9 B_E_ (black) : 3 bbE_ (chocolate) : 3 B_ee (yellow) : 1 bbee (yellow) = 9 black : 3 chocolate : 4 yellow.

Genetic Interaction	Cross	Expected F_2 Phenotypic Ratio	Example Organism and Trait
Complete Dominance	Monohybrid (Aa × Aa)	3:1	Pea plant – tall vs. dwarf height
Incomplete Dominance	Monohybrid (Rr × Rr)	1:2:1	Four-o'clock plant (*Mirabilis jalapa*) – red, pink, white flowers
Codominance	Monohybrid (IAIB × IAIB)	1:2:1	Human – ABO blood group (A, AB, B)
Recessive Epistasis	Dihybrid (AaBb × AaBb)	9:3:4	Mouse coat colour – black, brown, albino
Dominant Epistasis	Dihybrid (AaBb × AaBb)	12:3:1	Squash fruit colour – white, yellow, green

Duplicate Dominant Epistasis	Dihybrid (AaBb × AaBb)	15:1	Shepherd's purse (*Capsella bursa-pastoris*) – triangular vs. ovoid seed shape
Duplicate Recessive Epistasis	Dihybrid (AaBb × AaBb)	9:7	Sweet pea (*Lathyrus odoratus*) – purple vs. white flower colour

Box 1.1: Summary of Mendelian and extended Mendelian phenotypic ratios

Pilot questions 1.4

1. Distinguish between incomplete dominance and codominance with one example of each.
2. In snapdragon flowers, red (RR) x white (R'R') gives F1 pink (RR'). What are the F2 genotypic and phenotypic ratios from RR' x RR'?
3. State the three alleles at the ABO locus and write the genotype for each of the four blood groups.
4. Define pleiotropy and give one human example.
5. Define epistasis and explain how recessive epistasis modifies the standard dihybrid F2 ratio of 9:3:3:1.

Series Summary

This Series introduced **genetics as the study of genes, heredity, and genetic variation**, tracing key milestones from **Mendel's experiments (1866)**, the term *genetics* by **Bateson (1905)**, the rediscovery of Mendel's work (1900), the **chromosomal theory** by **Sutton and Boveri (1902–1903)**, and the discovery of the **DNA structure** by **Watson and Crick (1953)**. It explained that heritable traits are encoded in **DNA** and transmitted from parents to offspring, distinguishing **genotype** (genetic makeup) from **phenotype** (observable traits influenced by genotype × environment). Mendel's **Law of Segregation** states that paired alleles separate during gamete formation, producing a **3:1 F_2 monohybrid ratio**, while the **Law of Independent Assortment** states that alleles of different genes assort independently, giving a **9:3:3:1 F_2 dihybrid ratio**. Test crosses produce **1:1** (Aa × aa) and **1:1:1:1** (AaBb × aabb) offspring ratios. The Series also distinguished **heritable** from **non-heritable** traits, explained the **Weismann barrier**, introduced the **norm of reaction**, and described methods for studying inheritance, including **twin studies, adoption studies, selective breeding, heritability analysis ($h^2 = V^G/V_p$), molecular genetics, and controlled-environment experiments**.

The Series also covered **non-Mendelian inheritance patterns**. **Incomplete dominance** produces an intermediate heterozygous phenotype with a **1:2:1 ratio**, while **codominance** results in full expression of both alleles, also producing a **1:2:1 ratio**. **Multiple alleles** were illustrated by the **ABO blood group system**, where **I^A and I^B are codominant**, both dominant over **i**, producing four blood groups from six genotypes, with **group O** as the universal red-cell donor and **group AB** as the universal recipient. The Series further explained **pleiotropy**, where a single gene influences multiple traits (e.g., sickle-cell disease, Marfan syndrome, PKU), and **epistasis**, where one gene masks or modifies another, covering major types and modified phenotypic ratios: **recessive epistasis (9:3:4), dominant epistasis (12:3:1), duplicate dominant epistasis (15:1), and duplicate recessive epistasis (9:7)**, with the **Labrador coat-colour** example illustrating recessive epistasis.

Glossary of Terms

- **Allele:** one of two or more alternative forms of a gene that can occupy the same locus on homologous chromosomes; a diploid organism carries two alleles at each locus.
- **Codominance:** a pattern of inheritance in which both alleles in a heterozygote are fully expressed simultaneously; the heterozygote shows the phenotype of both homozygotes, not an intermediate; example: MN blood group (MN genotype shows both M and N antigens).
- **Dihybrid cross:** a cross between two individuals that differ in two pairs of contrasting characters; F2 offspring of a dihybrid cross show a 9:3:3:1 phenotypic ratio under complete dominance and independent assortment.
- **Dominance:** the relationship between two alleles in which one allele (dominant) is expressed in the phenotype of a heterozygote, masking the expression of the other allele (recessive).
- **Epistasis:** the interaction between alleles at two or more different loci such that the expression of one gene (hypostatic) is masked or modified by the alleles of another gene (epistatic) at a different locus; modifies the standard 9:3:3:1 dihybrid ratio.
- **Gene:** a specific DNA sequence at a defined locus on a chromosome that encodes a functional product (protein or RNA) contributing to the phenotype of the organism.
- **Genotype:** the complete genetic constitution of an organism; specifically the combination of alleles at the locus or loci under consideration.
- **Heritability (h^2):** the proportion of total phenotypic variance in a population that is attributable to genetic variance; $h^2 = V^G/V_p$; ranges from 0 (entirely environmental) to 1.0 (entirely genetic); a population-specific statistic.
- **Incomplete dominance:** a pattern of inheritance in which neither allele is fully dominant; the heterozygote shows an intermediate phenotype between the two homozygotes; produces a 1:2:1 phenotypic ratio in F2 crosses; example: flower colour in Mirabilis jalapa (snapdragon).
- **Locus:** the specific, fixed position on a chromosome where a gene (or a specific DNA sequence) is located; plural: loci.
- **Mendel's Law of Independent Assortment:** the alleles of two (or more) different genes on different chromosomes assort independently of one another during gamete formation; produces a 9:3:3:1 F2 ratio in dihybrid crosses.
- **Mendel's Law of Segregation:** the two alleles of a gene separate during gamete formation so that each gamete receives only one allele; the two alleles are reunited at fertilisation.

- **Non-heritable characteristic:** a trait that arises through the organism's interaction with its environment during its lifetime; not encoded in the genetic material; cannot be transmitted to offspring through the germ cells.
- **Norm of reaction:** the range of phenotypes that a given genotype can produce when exposed to different environmental conditions; demonstrates that phenotype is not determined by genotype alone.
- **Phenotype:** the observable characteristics of an organism resulting from the interaction of its genotype with the environment.
- **Pleiotropy:** the phenomenon whereby a single gene influences two or more apparently unrelated phenotypic traits; example: the HBB mutation in sickle-cell disease causes multiple phenotypic effects in different organ systems.
- **Test cross:** a cross between an individual showing the dominant phenotype but of unknown genotype (AA or Aa) and a homozygous recessive individual (aa); used to determine the genotype of the dominant phenotype individual; offspring ratio of 1:1 indicates the unknown was Aa; all dominant offspring indicates it was AA.
- **Weismann barrier:** the physiological separation between somatic (body) cells and germ cells (gamete-producing cells); ensures that changes occurring in somatic cells during an organism's lifetime are not transmitted to offspring; refutes Lamarckian inheritance of acquired characteristics.

Concept Check Questions [Series 1]

Objective Questions

1. Genetics is the scientific study of genes, _____, and genetic variation; the word was coined by William Bateson in 1905. (Linked to SLO 1.1)
2. Mendel's Law of _____ states that the two alleles of a gene separate during gamete formation so that each gamete receives only one allele; the F2 monohybrid ratio is 3 dominant : 1 recessive. (Linked to SLO 1.2)
3. A non-heritable characteristic arises through the organism's interaction with its _____ during its lifetime and cannot be transmitted to offspring through the germ cells. (Linked to SLO 1.3)
4. In incomplete dominance, the heterozygote (RR') shows an _____ phenotype between the two homozygotes; in codominance, both alleles are fully expressed simultaneously in the heterozygote. (Linked to SLO 1.4)
5. The ABO blood group system has three alleles at the same locus: I^A, I^B, and i; genotype I^AI^B gives blood group _____ because both alleles are codominant and both A and B antigens are expressed simultaneously. (Linked to SLO 1.4)

Short-Answer Questions

6. Define genetics and state three sub-disciplines within it. (Linked to SLO 1.1)
7. State Mendel's two laws and explain the F2 ratios they predict. (Linked to SLO 1.2)
8. Define non-heritable characteristics and give two examples with explanations. (Linked to SLO 1.3)
9. Explain how twin studies are used to distinguish heritable from non-heritable traits. (Linked to SLO 1.3)
10. Define pleiotropy and epistasis and give one example of each. (Linked to SLO 1.4)

Long-Answer Questions

11. Describe the historical development of genetics from Mendel to the double helix. Explain the chromosomal basis of heredity. Define heritable characteristics, apply Mendel's laws to monohybrid and dihybrid crosses, and define non-heritable characteristics with examples. (Linked to SLOs 1.1, 1.2, 1.3)
12. Explain how heritable and non-heritable traits are distinguished using twin studies, heritability analysis, and other methods. Describe incomplete dominance, codominance, multiple alleles

(ABO system), pleiotropy, and epistasis with examples and modified F2 ratios. (Linked to SLOs 1.3, 1.4)

Answers to Concept Check Questions [Series 1]

Objective Questions

1. Heredity.
2. Segregation.
3. Environment.
4. Intermediate.
5. AB.

Short-Answer Questions

6. Genetics is the scientific study of genes, heredity, and genetic variation. Three sub-disciplines: classical genetics (patterns of inheritance using crosses); molecular genetics (structure and function of genes at the molecular level); population genetics (allele frequencies and genetic variation in populations over time).
7. Law of Segregation: the two alleles of a gene separate during gamete formation; each gamete receives one allele; F2 monohybrid ratio = 3 dominant : 1 recessive (3:1). Law of Independent Assortment: alleles of genes on different chromosomes assort independently during gamete formation; F2 dihybrid ratio = 9:3:3:1.
8. Non-heritable characteristics arise from environmental interactions during an organism's lifetime and cannot be transmitted to offspring. Example 1: a farmer's callused hands (tissue remodelling in response to mechanical stress; not encoded in DNA). Example 2: skin tanning in response to UV radiation (increased melanin production in existing skin cells; not a heritable genetic change).
9. Monozygotic (MZ; identical) twins share 100 percent of their DNA; dizygotic (DZ; fraternal) twins share approximately 50 percent. If a trait is heritable, MZ twins show higher concordance than DZ twins. If MZ concordance = DZ concordance, the trait is entirely environmental. If MZ concordance > DZ concordance, the trait has a heritable genetic component.
10. Pleiotropy: one gene affects multiple phenotypic traits; example: the HBB gene mutation in sickle-cell disease causes haemolytic anaemia, vaso-occlusive crises, splenomegaly, and other effects. Epistasis: alleles at one locus mask the expression of alleles at a different locus; example:

recessive epistasis in Labrador retriever coat colour (ee genotype masks B locus expression, giving yellow coat; 9:3:4 ratio).

Long-Answer Questions

11. History: Mendel (1866, pea cross laws); rediscovery by de Vries, Correns, von Tschermak (1900); Sutton and Boveri chromosomal theory (1902 to 1903); Morgan and Drosophila genetics (1908 to 1920s); Avery et al. DNA as genetic material (1944); Watson and Crick double helix (1953). Chromosomal basis: genes on chromosomes; chromosome behaviour in meiosis parallels Mendel's laws (segregation and independent assortment); evidence from sex-linked traits and genetic maps; 46 chromosomes in humans (22 autosome pairs + XX or XY). Heritable characteristics: encoded in DNA; transmitted across generations; examples: blood group, sickle-cell disease, eye colour. Monohybrid cross: Aa x Aa → F2 = 3:1 (phenotypic); Law of Segregation. Dihybrid cross: RrYy x RrYy → F2 = 9:3:3:1; Law of Independent Assortment; forked-line method (3/4 x 3/4 = 9/16 etc.). Test cross: dominant phenotype x aa → 1:1 (if Aa) or all dominant (if AA). Non-heritable: arise from environment; not transmitted; examples: tanned skin, callused hands, acquired muscle, learned behaviours; Weismann barrier separates soma from germline.
12. Distinguishing heritable from non-heritable: twin studies (MZ vs DZ concordance; higher MZ concordance = heritable); adoption studies; selective breeding response; heritability $h^2 = V^G/V_p$ (0 = entirely environmental; 1.0 = entirely genetic); GWAS and molecular genetics; controlled environment experiments. Incomplete dominance: heterozygote intermediate; F2 ratio 1:2:1; example: Mirabilis jalapa pink flowers (RR'). Codominance: both alleles fully expressed; F2 ratio 1:2:1; example: MN blood group (MN genotype shows both antigens). Multiple alleles (ABO): I^A, I^B, i; I^A and I^B codominant; both dominant over i; genotypes: I^AI^A/I^Ai = A; I^BI^B/I^Bi = B; I^AI^B = AB; ii = O. Pleiotropy: one gene affects multiple traits; sickle-cell disease (HBB mutation); Marfan syndrome (FBN1). Epistasis: gene at one locus masks another; recessive epistasis (9:3:4); dominant epistasis (12:3:1); Labrador coat colour (B and E loci; 9 black:3 chocolate:4 yellow).

Answers to Pilot Questions [Series 1]

Part 1.1

1. Genetics is the scientific study of genes, heredity, and genetic variation. Three sub-disciplines: classical genetics; molecular genetics; population genetics.

2. Mendel published his laws (1866). Rediscovery of Mendel (1900). Chromosomal theory by Sutton and Boveri (1902 to 1903). Watson and Crick proposed the double-helix structure of DNA (1953).
3. The chromosomal theory states that genes are located on chromosomes and that chromosome behaviour during meiosis explains Mendel's laws. Two lines of evidence: parallel behaviour (chromosomes segregate during meiosis exactly as Mendel's alleles segregate; independent assortment of different chromosome pairs parallels independent assortment of Mendel's factors); sex-linked traits (sex-linked traits are inherited with the sex chromosomes, confirming that genes are physically on chromosomes).
4. Mitosis: one division; daughter cells diploid (2n); function is growth and somatic cell replacement. Meiosis: two successive divisions; daughter cells haploid (1n); function is gamete production; involves crossing over in Prophase I.
5. A karyotype is the complete set of chromosomes of an individual arranged in standardised pairs by size and morphology. It reveals: chromosome number (e.g. trisomy 21 in Down syndrome); sex chromosome constitution (XX or XY); structural abnormalities (deletions, translocations, inversions).

Part 1.2

1. Gene: a specific DNA sequence that encodes a functional product and occupies a defined locus on a chromosome. Allele: one of two or more alternative forms of a gene at the same locus. Locus: the specific position on a chromosome where a gene is located. Genotype: the combination of alleles an organism carries at a locus or loci under study. Phenotype: the observable characteristics resulting from genotype interacting with the environment.
2. Cross: Rr x Rr. Genotypic ratio: 1 RR : 2 Rr : 1 rr. Phenotypic ratio: 3 round : 1 wrinkled (3:1).
3. Mendel's Law of Segregation: the two alleles of a gene separate (segregate) during gamete formation so that each gamete receives only one allele; the two alleles are reunited at fertilisation when two gametes fuse.
4. The most likely genotype is TT (homozygous dominant). If the plant were Tt, a test cross (Tt x tt) would produce approximately half tall (Tt) and half dwarf (tt) offspring; obtaining 40 tall offspring from 40 progeny is highly unlikely for a Tt parent but expected for a TT parent.
5. F2 phenotypic ratio: 9 round yellow : 3 round green : 3 wrinkled yellow : 1 wrinkled green (9:3:3:1).

Part 1.3

1. Non-heritable characteristics arise from environmental influences during an organism's lifetime and cannot be transmitted to offspring. Three examples: a farmer's callused hands (mechanical friction causes epidermal thickening; not encoded in DNA); skin tanning (UV radiation increases melanin production; not a heritable change); scar tissue (repair of injured skin; not transmitted to offspring).
2. The Weismann barrier is the physiological separation between somatic (body) cells and germ cells (gamete-producing cells). It ensures that changes occurring in somatic cells during an organism's lifetime are not transmitted to offspring. It refutes Lamarck's hypothesis of the inheritance of acquired characteristics.
3. The norm of reaction is the range of phenotypes a given genotype can produce across different environmental conditions. Example: the Himalayan rabbit allele of the tyrosinase gene produces pigment only at cool temperatures; the same genotype produces black fur at the cool extremities and white fur at the warmer body surface; shaving fur on the body and cooling that area causes black fur to grow there, demonstrating that phenotype depends on both genotype and environment.
4. Monozygotic (MZ) twins share 100 percent of DNA; dizygotic (DZ) twins share approximately 50 percent. Concordance rate = percentage of twin pairs where both members show the trait. If MZ concordance significantly exceeds DZ concordance, the trait has a heritable component. If MZ concordance equals DZ concordance, the trait is entirely environmental.
5. Heritability $h^2 = V^G/V_p$ = genetic variance / total phenotypic variance. $h^2 = 0$: the trait is entirely determined by the environment; no genetic component in this population. $h^2 = 1.0$: all phenotypic variation in the population is due to genetic differences; the environment does not contribute to variation.

Part 1.4

1. Incomplete dominance: heterozygote shows an intermediate phenotype; example: Mirabilis jalapa (four o'clock) flower colour (RR = red; R'R' = white; RR' = pink). Codominance: both alleles are fully expressed simultaneously in the heterozygote; example: MN blood group (genotype MN produces both M and N antigens on red blood cell surface).
2. F2 cross: RR' x RR'. Genotypic ratio: 1 RR : 2 RR' : 1 R'R'. Phenotypic ratio: 1 red : 2 pink : 1 white (1:2:1).
3. Three alleles: I^A, I^B, i. Genotypes: blood group A = I^AI^A or I^Ai; blood group B = I^BI^B or I^Bi; blood group AB = I^AI^B; blood group O = ii.

4. Pleiotropy: a single gene influences two or more apparently unrelated phenotypic traits. Human example: sickle-cell disease (HBB gene mutation causes haemolytic anaemia, vaso-occlusive crises, splenomegaly, stroke risk, and retinal damage from a single amino acid substitution in beta-globin).
5. Epistasis: alleles at one locus (epistatic gene) mask the phenotypic expression of alleles at a different locus (hypostatic gene). Recessive epistasis: homozygous recessive at the epistatic locus (bb) masks the expression of the hypostatic gene (A locus); the standard 9 A_B_ : 3 A_bb : 3 aaB_ : 1 aabb becomes 9:3:4 (the 3 aaB_ and 1 aabb classes both show the same masked phenotype).

References and Attributions [Series 1]

Textual References (Books and External Sources)

- Brooker, R. J. (2015). Genetics: Analysis and principles (5th ed.). McGraw-Hill, New York.
- Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., & Doebley, J. (2015). Introduction to genetic analysis (11th ed.). W. H. Freeman, New York.
- Lewin, B. (2008). Genes IX. Jones and Bartlett Publishers, Sudbury.
- Snustad, D. P., & Simmons, M. J. (2012). Principles of genetics (6th ed.). John Wiley & Sons, Hoboken.

Figures, Plates, Tables, and Boxes

- Figure 1.1: Punnett squares for monohybrid (Aa x Aa) and dihybrid (RrYy x RrYy) F2 crosses. Attribution: AI generated. Suggested OER search string: “monohybrid dihybrid Punnett square F2 ratio genetics OER”.
- Box 1.1: Summary of Mendelian and extended Mendelian phenotypic ratios. Attribution: AI generated. Suggested OER search string: “Mendelian ratios extensions incomplete dominance codominance epistasis table OER”.

Course code: BIO 201

Course title: Genetics I

Course credit load: 2 Units

Series 2: Probability in Genetic Events

- **Part 2.1: Foundations of Probability**
  - Sub Part 2.1.1: Definition and basic rules of probability
  - Sub Part 2.1.2: The product rule and the sum rule in genetics
  - Sub Part 2.1.3: Conditional probability and Bayes' theorem
- **Part 2.2: Probability Applied to Mendelian Genetics**
  - Sub Part 2.2.1: Predicting offspring genotype and phenotype ratios
  - Sub Part 2.2.2: Probability in multiple-gene crosses
  - Sub Part 2.2.3: The binomial probability distribution in genetics
- **Part 2.3: Sex Linkage and Probability**
  - Sub Part 2.3.1: Sex determination and sex chromosomes
  - Sub Part 2.3.2: X-linked inheritance and probability
  - Sub Part 2.3.3: Y-linked and sex-influenced inheritance
- **Part 2.4: Linkage, Recombination, and Probability**
  - Sub Part 2.4.1: Linked genes and reduced recombination
  - Sub Part 2.4.2: Calculating recombination frequency
  - Sub Part 2.4.3: Genetic mapping and map distances

Introduction

Probability is the mathematical language of genetics. Because meiosis is a random process (which homologous chromosome goes to which pole is determined by chance) and fertilisation is a random union of gametes, the transmission of alleles from parents to offspring follows the laws of probability. Understanding probability is therefore essential for predicting the outcomes of genetic crosses, interpreting pedigrees, diagnosing genetic disorders, and analysing population genetic data.

This Series builds systematically on the Mendelian foundations of Series 1 by introducing the mathematical framework of probability and applying it to genetic events. It covers the basic rules of probability (product rule, sum rule, conditional probability, and Bayes' theorem); the application of probability to predict offspring ratios in single-gene and multi-gene crosses; the binomial probability formula for calculating the probability of specific outcomes among multiple offspring; sex-linked inheritance and the different probabilities that apply to X-linked traits in males and females; and the relationship between linkage, recombination frequency, and the probability of inheriting two alleles together.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 2.1** define probability, state and apply the product rule and sum rule to genetic problems, and apply conditional probability and Bayes' theorem to carrier probability calculations (Linked to Part 2.1),
- **SLO 2.2** use probability rules to predict offspring genotype and phenotype ratios in single-gene and multi-gene crosses, and apply the binomial formula to calculate the probability of specific offspring combinations (Linked to Part 2.2),
- **SLO 2.3** explain sex determination in humans, predict the outcomes of X-linked crosses using probability, and distinguish X-linked from Y-linked and sex-influenced inheritance (Linked to Part 2.3), and

- **SLO 2.4** explain how linked genes affect expected offspring ratios, calculate recombination frequency from testcross data, and convert recombination frequencies to map distances (Linked to Part 2.4).

2.1 Foundations of Probability

2.1.1 Definition and basic rules of probability

Probability is a numerical measure of the likelihood that a specific event will occur, expressed on a scale from 0 (the event is impossible) to 1 (the event is certain). In genetics, probability applies to the random processes of meiosis (allele segregation) and fertilisation (gamete union). The probability of an event A is defined as: P(A) = number of favourable outcomes / total number of equally likely outcomes. For a fair coin: P(heads) = 1/2; for the roll of a fair die: P(6) = 1/6. Basic probability rules: (1) Complement rule: P(not A) = 1 - P(A); the probability that an event does NOT occur = 1 minus the probability that it does occur. Example: if the probability that a child inherits the recessive allele from both parents is 1/4, then the probability it does NOT inherit both recessive alleles = 1 - 1/4 = 3/4. (2) Addition rule for mutually exclusive events (sum rule; see Part 2.1.2): if events A and B cannot both occur at the same time, P(A or B) = P(A) + P(B). (3) Multiplication rule for independent events (product rule; see Part 2.1.2): if events A and B are independent (the occurrence of one does not affect the probability of the other), P(A and B) = P(A) × P(B).

Genetic context for probability: the segregation of alleles at a single locus during meiosis is like flipping a coin: a heterozygous parent (Aa) produces A-bearing gametes and a-bearing gametes each with probability 1/2; this is because the two homologous chromosomes (one carrying A, one carrying a) are separated randomly during Meiosis I, so each has a 50 percent chance of going to a given pole. Fertilisation is also random: any sperm has an equal probability of fertilising any egg. Because meiosis and fertilisation are independent events (the allele segregated in one parent's meiosis does not affect the allele segregated in the other parent's meiosis), the product rule applies to combining probabilities from each parent. This is the mathematical basis for Punnett square calculations: the probability of a specific genotype in the offspring equals the product of the probabilities that each parent contributes the relevant allele.

Fill in the blank: Probability ranges from _____ (impossible) to 1 (certain); the probability of an event is defined as the number of favourable outcomes divided by the total number of equally likely outcomes; the complement rule states that P(not A) = 1 - P(A).

2.1.2 The product rule and the sum rule in genetics

The product rule (multiplication rule for independent events): the probability that two or more independent events will all occur together equals the product (multiplication) of their individual probabilities. P(A and B) = P(A) × P(B). In genetics: the alleles a heterozygous mother contributes and the alleles a heterozygous father contributes are independent events (one parent's meiosis does not affect the other's). Therefore, to find the probability that a specific genotype appears in an offspring of two Aa parents: P(AA offspring) = P(mother contributes A) × P(father contributes A) = 1/2 × 1/2 = 1/4. P(Aa offspring) = P(mother A, father a) + P(mother a, father A) = (1/2 × 1/2) + (1/2 × 1/2) = 1/4 + 1/4 = 1/2. P(aa offspring) = 1/2 × 1/2 = 1/4. These three probabilities sum to 1 (1/4 + 1/2 + 1/4 = 1), confirming they are exhaustive. The product rule is particularly powerful for multi-gene crosses: for independent genes, the probability of a combined genotype equals the product of the separate gene probabilities.

The sum rule (addition rule for mutually exclusive events): the probability that one or another of two mutually exclusive outcomes will occur equals the sum of their individual probabilities. P(A or B) = P(A) + P(B) (when A and B cannot both occur simultaneously). In genetics: if an offspring can be either Aa OR aA (two different ways of being heterozygous), these are mutually exclusive outcomes (a given offspring cannot simultaneously be both a 'mother A, father a' offspring and a 'mother a, father A' offspring), so we sum them: P(heterozygous) = P(Aa) + P(aA) = 1/4 + 1/4 = 1/2. The product rule and sum rule work together: use the product rule to calculate the probability of each specific combination; use the sum rule to add up the probabilities of all combinations that give the same phenotype or genotype. Example: what is the probability of a dominant phenotype from Aa x Aa? P(A_) = P(AA) + P(Aa) = 1/4 + 1/2 = 3/4.

Fill in the blank: The product rule states P(A and B) = P(A) × P(B) for _____ events; the sum rule states P(A or B) = P(A) + P(B) for mutually exclusive events; both rules together enable calculation of any genotype or phenotype probability without a Punnett square.

2.1.3 Conditional probability and Bayes' theorem

Conditional probability is the probability of an event A given that another event B has already occurred: P(A|B) = P(A and B) / P(B). In genetics, conditional probability arises when we know something about the phenotype or family history of an individual and want to revise our estimate of their genotype probability. Classic example: a man with normal phenotype has an affected brother (aa) and unaffected parents; the parents must both be Aa (carriers); the man's prior probability of being AA or Aa is 1/3 AA : 2/3 Aa (from the 1:2 ratio of unaffected offspring from Aa x Aa, because 1/4 AA + 1/2 Aa = 3/4 unaffected, of which 1/3 are AA and 2/3 are Aa). Bayes' theorem provides a systematic method for revising prior probabilities in the light of new information (conditional evidence). The formula: P(A|B) = [P(B|A) × P(A)] / P(B), where P(A) is the prior probability of hypothesis A (before considering the new evidence); P(B|A) is the conditional probability of the evidence given hypothesis A is true; P(B) is the total probability of the evidence across all hypotheses.

Bayesian probability in pedigree analysis: Bayesian analysis is used in genetic counselling to calculate the posterior probability (revised probability after considering all available information) that a phenotypically normal individual is a carrier of a recessive allele. Example: a woman's mother is a known carrier of an autosomal recessive condition. Prior probability the woman is a carrier = 1/2 (she either inherited the recessive allele or she did not). Now suppose she has three unaffected children with an unaffected, non-carrier partner. If she is a carrier, each child has a 1/2 chance of being unaffected; probability of three unaffected children given she is a carrier = $(1/2)^3$ = 1/8. If she is not a carrier, all children will be unaffected; probability = 1. Using Bayes' theorem: posterior P(carrier) = (1/2 × 1/8) / [(1/2 × 1/8) + (1/2 × 1)] = (1/16) / (1/16 + 1/2) = (1/16) / (9/16) = 1/9 ≈ 11 percent. The posterior probability of being a carrier has dropped from 50 percent to 11 percent because her three unaffected children provide evidence against carrier status.

Fill in the blank: Bayes' theorem P(A|B) = [P(B|A) × P(A)] / P(B) is used in genetic counselling to calculate the _____ probability that an individual is a carrier after considering both prior family history information and new conditional evidence such as unaffected children.

Pilot questions 2.1

6. Define probability and state the three basic rules: complement rule, product rule, and sum rule.
7. Two carriers (Aa) of an autosomal recessive disorder have a child. Using the product rule, calculate the probability that the child is (a) homozygous dominant (AA), (b) a carrier (Aa), and (c) affected (aa).
8. Explain why the product rule and sum rule are both needed to calculate the probability of the dominant phenotype in a monohybrid cross.
9. Define conditional probability and explain why it is useful in pedigree analysis.
10. A phenotypically normal woman has a 1/2 prior probability of being a carrier for an autosomal recessive disorder. She has two unaffected children (partner is non-carrier). Use Bayes' theorem to calculate her posterior probability of being a carrier.

2.2 Probability Applied to Mendelian Genetics

2.2.1 Predicting offspring genotype and phenotype ratios

Every Punnett square calculation can be replaced by a probability calculation using the product and sum rules, which is faster for complex crosses. For a monohybrid cross Aa x Aa: the gametes from each parent are A (p = 1/2) and a (p = 1/2). Offspring genotype probabilities: P(AA) = 1/2 × 1/2 = 1/4; P(Aa) = (1/2 × 1/2) + (1/2 × 1/2) = 1/2 (two paths to heterozygosity); P(aa) = 1/2 × 1/2 = 1/4. Offspring phenotype probabilities (with A dominant): P(dominant phenotype, A_) = P(AA) + P(Aa) = 1/4 + 1/2 = 3/4; P(recessive phenotype, aa) = 1/4. For a cross between two different genotypes: example Aa x AA. Gametes from Aa: A (1/2) or a (1/2); gametes from AA: A (all gametes = 1). P(AA offspring) = 1/2 × 1 = 1/2; P(Aa offspring) = 1/2 × 1 = 1/2; P(aa offspring) = 0; phenotypic ratio: all dominant (1/2 AA + 1/2 Aa = 1).

For a cross Aa x aa (test cross): gametes from Aa: A (1/2), a (1/2); gametes from aa: all a. P(Aa offspring) = 1/2 × 1 = 1/2; P(aa offspring) = 1/2 × 1 = 1/2; phenotypic ratio 1:1 (1/2 dominant : 1/2 recessive). For incomplete dominance (e.g. RR' x RR' in snapdragons): gametes R (1/2) and R' (1/2) from each parent; P(RR) = 1/4 (red); P(RR') = 1/2 (pink); P(R'R') = 1/4 (white); phenotypic ratio 1:2:1. For the ABO blood group: if both parents are I^AI^B (blood group AB): gametes I^A (1/2) or I^B (1/2) from each parent; P(I^AI^A) = 1/4 (type A); P(I^AI^B) = 1/2 (type AB);

$P(I^BI^B) = 1/4$ (type B); no offspring can be type O (ii). If one parent is I^Ai and the other is I^Bi: gametes from I^Ai are I^A (1/2) or i (1/2); gametes from I^Bi are I^B (1/2) or i (1/2); offspring: $P(I^AI^B) = 1/4$ (AB); $P(I^Ai) = 1/4$ (A); $P(I^Bi) = 1/4$ (B); $P(ii) = 1/4$ (O).

Fill in the blank: For the cross Aa x Aa, P(AA) = _____ and P(Aa) = 1/2 (using the product and sum rules); P(dominant phenotype) = P(AA) + P(Aa) = 3/4; this probability approach replaces the Punnett square and gives the same result.

2.2.2 Probability in multiple-gene crosses

For genes on different chromosomes (independently assorting genes), the probability of a specific multi-gene genotype is the product of the probabilities for each individual gene, calculated independently. This is the mathematical expression of Mendel's Law of Independent Assortment. Example: dihybrid cross AaBb x AaBb. Probability of offspring genotype A_B_ (dominant phenotype for both genes): P(A_) = 3/4 (from Aa x Aa); P(B_) = 3/4 (from Bb x Bb); P(A_B_) = 3/4 × 3/4 = 9/16. P(A_bb) = 3/4 × 1/4 = 3/16. P(aaB_) = 1/4 × 3/4 = 3/16. P(aabb) = 1/4 × 1/4 = 1/16. These sum to 9/16 + 3/16 + 3/16 + 1/16 = 16/16 = 1, confirming the 9:3:3:1 ratio. The forked-line method is the graphical application of this multiplication approach.

Trihybrid cross AaBbCc x AaBbCc: P(A_) = 3/4; P(B_) = 3/4; P(C_) = 3/4; P(dominant for all three) = 3/4 × 3/4 × 3/4 = 27/64; P(dominant for A and B, recessive for C) = 3/4 × 3/4 × 1/4 = 9/64; etc. The full trihybrid ratio is 27:9:9:9:3:3:3:1 (8 phenotypic classes summing to 64). The general rule: for n independently assorting heterozygous gene pairs, the number of different gamete types = 2^n; the number of genotypic classes in F2 = 3^n; the number of phenotypic classes in F2 (with complete dominance) = 2^n; the F2 fraction with all dominant phenotypes = $(3/4)^n$. Worked example: what is the probability that an offspring of AaBbCcDd x AaBbCcDd will show the dominant phenotype for all four traits? $P = (3/4)^4 = 81/256 \approx 31.6$ percent.

Fill in the blank: For independently assorting genes, the probability of any multi-gene genotype equals the _____ of the individual gene probabilities; for a trihybrid cross AaBbCc x AaBbCc, the fraction of offspring showing dominant phenotype for all three genes is $(3/4)^3 = 27/64$.

2.2.3 The binomial probability distribution in genetics

The binomial probability formula calculates the probability of obtaining exactly k occurrences of a specific outcome in n independent trials, where each trial has only two possible outcomes (success with probability p; failure with probability q = 1 - p). Formula: P(k successes in n trials) = [n! / (k! × (n-k)!)] × p^k × q^{n-k}, where n! = n factorial (n × (n-1) × ... × 2 × 1); the term n!/[k!(n-k)!] is the binomial coefficient, also written as C(n,k) or $\binom{n}{k}$, which counts the number of different ways k successes can be arranged among n trials. Application to genetics: if a couple are both carriers of an autosomal recessive allele (Aa x Aa), the probability that any single child is affected = p = 1/4; the probability of being unaffected = q = 3/4. What is the probability that exactly 2 of their 5 children will be affected? n = 5; k = 2; p = 1/4; q = 3/4. P = [5!/(2! × 3!)] × $(1/4)^2$ × $(3/4)^3$ = 10 × 1/16 × 27/64 = 10 × 27/1,024 = 270/1,024 ≈ 0.264 (approximately 26.4 percent).

Additional examples: P(exactly 3 of 5 children affected, p = 1/4) = [5!/(3! × 2!)] × $(1/4)^3$ × $(3/4)^2$ = 10 × 1/64 × 9/16 = 90/1,024 ≈ 8.8 percent. P(at least 1 of 5 affected) = 1 - P(0 affected) = 1 - $(3/4)^5$ = 1 - 243/1,024 = 781/1,024 ≈ 76.3 percent (using the complement rule). The binomial formula is also used to calculate the probability of specific sequences of sexes in offspring: if each child has p = 1/2 probability of being female: P(exactly 3 girls in a family of 5 children) = [5!/(3! × 2!)] × $(1/2)^3$ × $(1/2)^2$ = 10 × $(1/2)^5$ = 10/32 = 5/16 ≈ 31.3 percent. Important: the binomial formula assumes each trial is independent (each birth is independent of previous births) and p is constant (the probability of the outcome is the same for each trial).

Fill in the blank: The binomial formula P(k) = [n!/(k!(n-k)!)] × p^k × q^{n-k} calculates the probability of exactly _____ successes in n independent trials; for genetics problems, p is the probability of the genotype/phenotype of interest, and the binomial coefficient counts the different birth orders that give k successes.

Pilot questions 2.2

1. Using the product and sum rules (without a Punnett square), calculate the probability of each genotype and phenotype in the cross Aa x Aa.
2. In a dihybrid cross AaBb x AaBb (A and B on different chromosomes), calculate the probability of each of the four phenotypic classes.

3. Both parents are Aa (carriers). What is the probability that exactly 1 of their 4 children will be affected (aa)?
4. What is the probability that exactly 3 of 4 children from two Aa parents are unaffected?
5. For genes A, B, and C on different chromosomes, both parents are AaBbCc. What fraction of offspring will be dominant for all three traits? What fraction will be recessive for all three?

2.3 Sex Linkage and Probability

2.3.1 Sex determination and sex chromosomes

Sex determination in humans and many other organisms depends on a specialised pair of sex chromosomes (allosomes) that differ between the sexes, in contrast to the autosomes (the 22 other chromosome pairs in humans), which are the same in both sexes. In the XX/XY system (humans, Drosophila, most mammals): females have two X chromosomes (XX; homogametic sex; produce only X-bearing eggs); males have one X and one Y chromosome (XY; heterogametic sex; produce X-bearing and Y-bearing sperm in approximately equal numbers). Sex of offspring is determined by which sperm fertilises the egg: X-bearing sperm + X egg = XX (female); Y-bearing sperm + X egg = XY (male); therefore the probability of each sex = 1/2. The Y chromosome in humans is much smaller than the X chromosome and carries relatively few genes; key Y-specific genes include SRY (sex-determining region Y), which triggers testis development and the male developmental pathway; most Y-linked genes are expressed only in males. Other sex determination systems: ZW/ZZ (birds, some fish, some reptiles: females are ZW, males are ZZ); XX/XO (grasshoppers: females XX; males X0, having only one sex chromosome); haplodiploidy (Hymenoptera: bees, ants, wasps; females are diploid; males develop from unfertilised haploid eggs).

The X chromosome carries many genes that are unrelated to sex determination; these genes are said to be X-linked (or sex-linked). In males, X-linked genes are hemizygous (only one copy present; on the single X chromosome); males therefore express the phenotype of whatever allele is present on their single X chromosome, whether dominant or recessive. In females, X-linked genes behave more like autosomal genes (two copies present; dominance relationships apply). This hemizygosity in males is why X-linked recessive conditions (red-green colour blindness,

haemophilia A and B, Duchenne muscular dystrophy) are far more common in males than in females: a male needs only one copy of the recessive allele (on his single X) to be affected; a female needs two copies (one on each X). X-inactivation (Lyon hypothesis): in somatic cells of female mammals, one of the two X chromosomes in each cell is randomly inactivated (condensed into a Barr body; transcriptionally silenced) early in embryonic development; as a result, females are functional mosaics (some cells express the paternal X allele; others express the maternal X allele); this explains the patchy (tortoiseshell/calico) coat colour pattern in heterozygous female cats.

Fill in the blank: In the XX/XY sex determination system, females are the _____ sex (produce only X-bearing eggs); the sex of offspring is determined by which sperm (X-bearing or Y-bearing) fertilises the egg; X-linked genes in males are hemizygous, explaining why X-linked recessive conditions are much more common in males than females.

2.3.2 X-linked inheritance and probability

X-linked inheritance follows special probability rules because males have only one X chromosome. Notation: X^A = dominant allele on X; X^a = recessive allele on X; X^AY = unaffected male; X^aY = affected male; X^AX^A = homozygous dominant female (unaffected); X^AX^a = heterozygous female carrier (unaffected); X^aX^a = homozygous recessive female (affected). Classic X-linked recessive example: red-green colour blindness in humans. Cross 1: carrier female (X^AX^a) x normal-vision male (X^AY). Gametes from female: X^A (1/2) or X^a (1/2). Gametes from male: X^A (1/2) or Y (1/2). Offspring: X^AX^A (1/4, normal female); X^AX^a (1/4, carrier female); X^AY (1/4, normal male); X^aY (1/4, colour-blind male). Probability of colour blindness among all offspring = 1/4. Probability of colour blindness among sons only = 1/2. Probability among daughters = 0 (daughters cannot be affected in this cross).

Cross 2: carrier female (X^AX^a) x colour-blind male (X^aY). Offspring: X^AX^a (1/4, carrier female); X^aX^a (1/4, affected female); X^AY (1/4, normal male); X^aY (1/4, colour-blind male). P(affected) = 1/2; P(affected female) = 1/4; P(affected male) = 1/4; P(affected among daughters) = 1/2; P(affected among sons) = 1/2. Cross 3 (X-linked dominant): X-linked dominant conditions (e.g. X-linked hypophosphataemia; Rett syndrome; incontinentia pigmenti) are more common in females; an affected male (X^AY) with an unaffected female (X^aX^a): offspring are X^AX^a (1/2,

affected female) and X^aY (1/2, unaffected male); notably, an X-linked dominant condition passes from an affected father to ALL daughters but NO sons (fathers give their X to daughters and their Y to sons); an affected mother passes the condition to 1/2 daughters and 1/2 sons. Criss-cross inheritance: a hallmark of X-linked recessive conditions; affected males receive the X-linked recessive allele from their carrier mothers (not from their fathers, who give only Y to sons); the allele 'criss-crosses' from grandfather to carrier daughter to affected grandson.

Fill in the blank: In X-linked recessive inheritance, a carrier female (X^AX^a) crossed with a normal male (X^AY) produces colour-blind _____ with probability 1/2 among sons, but no affected daughters; fathers cannot transmit an X-linked trait to their sons because sons receive the Y chromosome (not X) from their father.

2.3.3 Y-linked and sex-influenced inheritance

Y-linked (holandric) inheritance: genes located only on the Y chromosome (in the non-recombining region of the Y) are transmitted exclusively from father to all sons; Y-linked traits skip females entirely (daughters never inherit the Y). True Y-linked traits are rare because the human Y chromosome has relatively few genes. Proposed examples include hairy ear pinnae (though the evidence is disputed) and some forms of Y-chromosome-linked male infertility (e.g. AZF region deletions on the Y chromosome causing azoospermia). All sons of an affected male will be affected; no daughters will be affected; the condition appears exclusively in males and the pattern is perfectly paternal (father-to-all-sons). Distinguishing Y-linked from X-linked recessive: in X-linked recessive inheritance, affected males cannot transmit the trait to sons (sons get Y from father); in Y-linked inheritance, affected males transmit the trait to ALL sons.

Sex-influenced inheritance: genes located on autosomes (not sex chromosomes) whose expression is influenced by the sex of the individual, typically due to differences in sex hormone levels between males and females. The dominance relationships of the alleles differ between the two sexes. Classic example: male-pattern baldness (androgenetic alopecia) in humans, controlled by the autosomal gene at the AR locus modified by androgen receptor sensitivity; the baldness allele (B) acts as a dominant in males (BB or Bb = bald; bb = non-bald) but as a recessive in females (only BB = bald; Bb = non-bald; bb = non-bald); therefore, a heterozygous male (Bb) is bald but a heterozygous female (Bb) is not bald. Second example: horn length in Dorset sheep;

the horned allele (H^A) is dominant in males but recessive in females. Sex-limited inheritance: a more extreme form in which a trait is expressed only in one sex even if the genotype is present in both sexes; e.g. milk production (expressed only in females); egg-laying capacity; sperm production; antler development. Sex-limited traits are autosomal but their phenotypic expression is limited to one sex by anatomical or physiological differences.

Fill in the blank: Y-linked (_____) inheritance is transmitted from father to ALL sons and to no daughters; sex-influenced inheritance involves autosomal genes whose expression differs between males and females because of differing hormone environments, as in male-pattern baldness where the baldness allele is dominant in males but recessive in females.

Pilot questions 2.3

1. Explain the XX/XY sex determination system and state the probability that any given child is female.
2. Explain why males are hemizygous for X-linked genes and why X-linked recessive conditions are more common in males than females.
3. A carrier woman (X^AX^a) for colour blindness marries a normal-vision man (X^AY). List all offspring genotypes and their probabilities; state the probability of having a colour-blind son.
4. Distinguish Y-linked inheritance from X-linked recessive inheritance in terms of the pattern of transmission.
5. Define sex-influenced inheritance and give one example with the dominance relationships stated for each sex.

2.4 Linkage, Recombination, and Probability

2.4.1 Linked genes and reduced recombination

Mendel's Law of Independent Assortment holds only for genes located on different (non-homologous) chromosomes or genes that are very far apart on the same chromosome. Genes that are located on the same chromosome are said to be linked and tend to be inherited together as a unit (they are part of the same linkage group). When genes are fully linked (no recombination

between them), a dihybrid testcross (AaBb x aabb) would give only two offspring classes (AB and ab parental types; each at frequency 1/2) instead of the expected four classes at 1:1:1:1 ratio predicted by independent assortment. In reality, some crossing over (physical exchange of segments between homologous chromosomes) occurs during Prophase I of Meiosis I, producing recombinant (non-parental) chromosomes; the frequency of recombination depends on the physical distance between the two genes: genes far apart on the same chromosome recombine at higher frequency (approaching 50 percent for genes very far apart, which behaves like independent assortment); genes close together recombine at lower frequency (less than 50 percent). The phenomenon of genes being physically close on the same chromosome and therefore showing less than 50 percent recombination is called genetic linkage.

Parental (non-recombinant) gametes are those that carry the same combination of alleles as the original parental chromosomes; recombinant gametes carry new combinations of alleles produced by crossing over. In a cross AB/ab x ab/ab (coupling configuration; both dominant alleles on the same chromosome): parental gametes are AB and ab; recombinant gametes are Ab and aB. If the recombination frequency between A and B is 20 percent (0.20), then: P(AB gamete from double heterozygote) = (1-0.20)/2 = 0.40; P(ab gamete) = 0.40; P(Ab gamete) = 0.20/2 = 0.10; P(aB gamete) = 0.10; total = 1.0. Testcross offspring frequencies (AB/ab x ab/ab): P(AB/ab) = 0.40 (parental, dominant for both); P(ab/ab) = 0.40 (parental, recessive for both); P(Ab/ab) = 0.10 (recombinant); P(aB/ab) = 0.10 (recombinant). Note that parental classes (0.40 each) are more frequent than recombinant classes (0.10 each) because the genes are linked.

Fill in the blank: Genes on the same chromosome are _____ and tend to be inherited together; the frequency of separation (recombination) depends on the physical distance between them; linked genes produce more parental-type offspring than recombinant-type offspring in a testcross.

2.4.2 Calculating recombination frequency

Recombination frequency (RF; also called map frequency or crossover frequency) is calculated from the results of a testcross between a double heterozygote (AaBb) and a double homozygous recessive (aabb): RF = (number of recombinant offspring) / (total number of offspring) × 100 percent. Worked example: a testcross of a fruit fly dihybrid (AB/ab) with ab/ab yields the

following offspring: 406 AB/ab (parental); 404 ab/ab (parental); 48 Ab/ab (recombinant); 42 aB/ab (recombinant); total = 900. RF = (48 + 42) / 900 × 100 = 90/900 × 100 = 10 percent. The two genes are 10 map units (centimorgans; cM) apart. RF ranges from 0 percent (completely linked; no recombination occurs between them) to 50 percent (unlinked; on different chromosomes or very far apart on the same chromosome; recombination is so frequent that assortment appears random). An RF of 50 percent is indistinguishable from independent assortment; genes showing RF = 50 percent may be unlinked or very distantly linked.

Three-point testcross (mapping gene order and distances): a three-point testcross involves a triple heterozygote (AaBbCc) crossed with aabbcc. The offspring fall into eight classes (three parental types and their recombinants; counting class pairs). The most common classes are the parental (non-recombinant) classes. The least common classes are the double recombinant classes (require two simultaneous crossover events; less likely than single crossovers). Steps to map gene order and distances from a three-point testcross: (1) Identify the parental classes (most frequent). (2) Identify the double recombinant classes (least frequent). (3) By comparing double recombinants with parentals, identify which gene is in the middle (the middle gene is the one whose allele combination is different in the double recombinant compared to the parental). (4) Calculate RF values between adjacent gene pairs by adding the classes that require crossing over between each pair. (5) Assemble the genetic map with distances in cM.

Fill in the blank: Recombination frequency (RF) = (number of _____ offspring) / total offspring × 100 percent; RF ranges from 0 percent (completely linked) to 50 percent (unlinked or very far apart); in a three-point testcross, the gene order is determined by identifying which allele is inverted in the double-recombinant compared with the parental class.

2.4.3 Genetic mapping and map distances

A genetic map (linkage map) is a diagram showing the relative positions of genes on a chromosome and the distances between them, expressed in centimorgans (cM) or map units (mu). One centimorgan = one percent recombination frequency; named after Thomas Hunt Morgan. Genetic maps are constructed from recombination frequency data; the greater the RF between two genes, the further apart they are on the genetic map. Additive map distances: the map distance between two genes is approximately additive (the distance A-C ≈ A-B + B-C when

B is between A and C) for short distances; for longer distances, the observed RF underestimates the true map distance because of double crossovers (two crossovers between the same pair of genes produce parental-type gametes and go undetected), reducing the apparent recombination frequency below the actual number of crossover events. The Haldane mapping function: a mathematical correction for multiple crossovers that converts observed RF values into true map distances: map distance (M) = -1/2 × ln(1 - 2RF). The Kosambi mapping function provides an alternative correction that accounts for crossover interference (the presence of one crossover inhibiting a nearby crossover).

Interference and coincidence: in a three-point testcross, the observed frequency of double crossovers is often less than the product of the two single crossover frequencies (because the occurrence of one crossover inhibits another nearby crossover). Coefficient of coincidence (C) = (observed double crossover frequency) / (expected double crossover frequency); where expected double crossover frequency = RF(A-B) × RF(B-C). Interference (I) = 1 - C; if I = 0 (C = 1): crossovers are independent (no interference); if I > 0 (C < 1): positive interference (one crossover inhibits nearby crossovers; common in most eukaryotes); if I < 0 (C > 1): negative interference (one crossover stimulates nearby crossovers; rare). Worked example: if RF(A-B) = 0.15; RF(B-C) = 0.20; expected double crossover = 0.15 × 0.20 = 0.03 (3 percent); if observed double crossover = 0.018 (1.8 percent): C = 0.018/0.03 = 0.60; I = 1 - 0.60 = 0.40 (40 percent positive interference). Comparison with physical maps: genetic map distances (in cM) may not correspond uniformly to physical distances (in base pairs or Mb); recombination hot spots and cold spots cause uneven correspondence between genetic and physical distances.

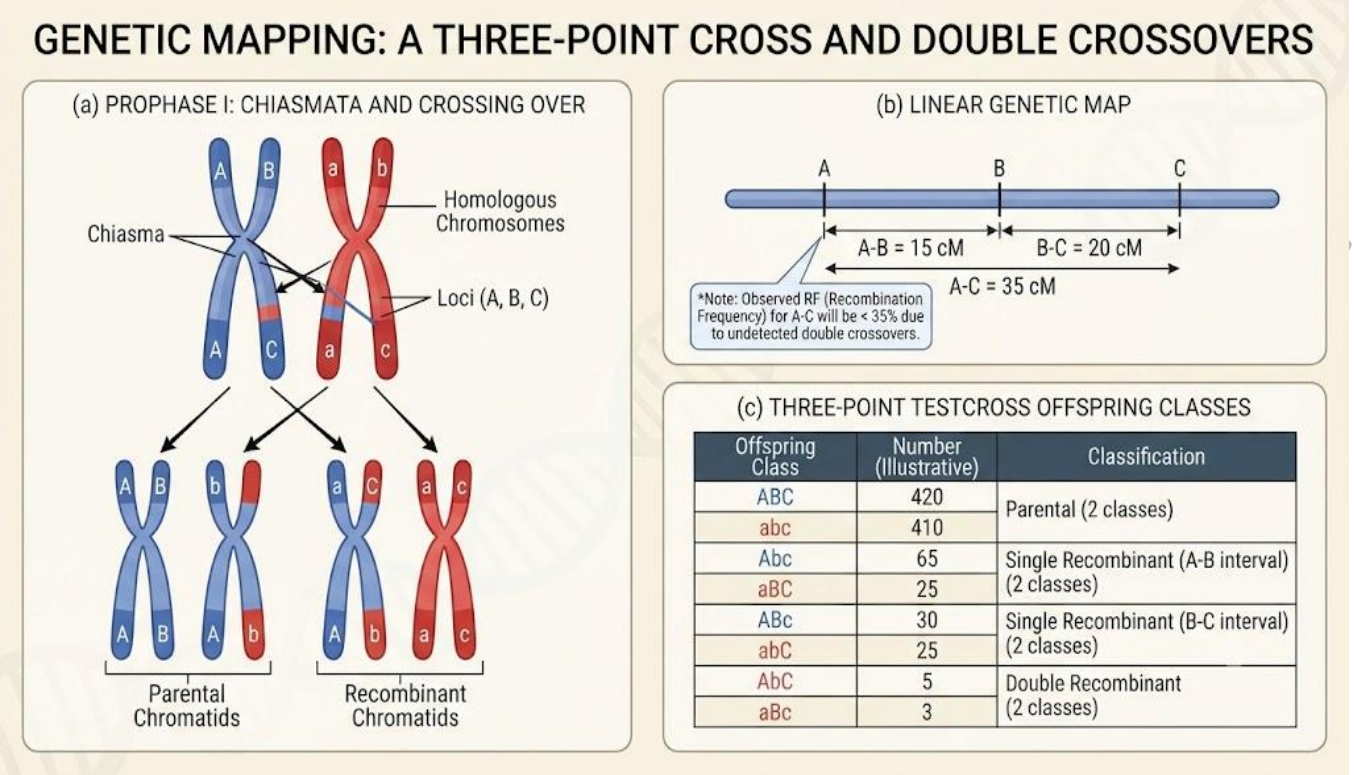

Figure 2.1: (a) Crossover events between linked loci during Meiosis I; (b) the resulting genetic map with distances in centimorgans; (c) offspring classes from a three-point testcross

Pilot questions 2.4

1. Explain why linked genes do not follow Mendel's Law of Independent Assortment.
2. In a testcross of AB/ab x ab/ab, the offspring are: 430 AB, 420 ab, 48 Ab, 52 aB. Calculate the recombination frequency and state the map distance in centimorgans.
3. Define parental and recombinant gametes and explain which class is more frequent in a testcross when genes are linked.
4. Explain what is meant by a coefficient of coincidence of 0.60 and an interference value of 0.40.
5. Distinguish between a genetic map and a physical map of a chromosome.

Series Summary

This Series introduced the fundamental principles of **genetics**, explaining how hereditary information is transmitted from one generation to the next and how genetic variation arises. It covered the foundations of inheritance, including genotype, phenotype, Mendel's Laws of Segregation and Independent Assortment, monohybrid and dihybrid crosses, and the distinction between heritable and non-heritable traits. It also explored more complex inheritance patterns, including incomplete dominance, codominance, multiple alleles, pleiotropy, and epistasis, while demonstrating how probability principles are applied to predict genetic outcomes.

Building on these foundations, the Series examined sex determination systems, sex-linked and sex-influenced inheritance, genetic linkage, recombination, and chromosome mapping. Together, these topics provide the basis for understanding how genes are inherited, expressed, and organised on chromosomes. By completing this Series, you should now be able to explain the principles of Mendelian and non-Mendelian inheritance, apply probability to genetic problems, interpret patterns of sex-linked inheritance and linkage, and use these concepts to achieve all four Series Learning Outcomes (SLOs).

Glossary of Terms

- **Bayes' theorem:** a formula for revising probability estimates in the light of new conditional evidence: $P(A|B) = [P(B|A) \times P(A)] / P(B)$; used in genetic counselling to calculate posterior carrier probabilities.
- **Binomial probability formula:** $P(k) = [n!/(k!(n-k)!)] \times p^k \times q^{n-k}$; calculates the probability of exactly k occurrences of an outcome with probability p in n independent trials.
- **Centimorgans (cM):** the unit of genetic map distance; 1 cM = 1 percent recombination frequency; named after Thomas Hunt Morgan.
- **Coefficient of coincidence (C):** the ratio of observed double crossover frequency to expected double crossover frequency; C = observed DCO/expected DCO; used to quantify crossover interference.
- **Conditional probability:** the probability of event A given that event B has already occurred: P(A|B) = P(A and B)/P(B); used in pedigree analysis to revise genotype probabilities in light of phenotypic information.

- **Criss-cross inheritance:** the pattern of X-linked recessive inheritance in which the trait appears predominantly in males; affected males received the X-linked recessive allele from their carrier mothers; the allele 'criss-crosses' from grandfather (affected) to carrier daughter to affected grandson.
- **Genetic map (linkage map):** a diagram showing the relative positions of genes on a chromosome and the genetic distances between them in centimorgans; constructed from recombination frequency data.
- **Hemizygous:** having only one copy of a gene (rather than the usual two); applies to X-linked genes in males (who have only one X chromosome) and to genes in any region where the homologous chromosome is absent.
- **Interference (I):** I = 1 - C; measures the degree to which one crossover inhibits (positive interference) or stimulates (negative interference) another nearby crossover; most eukaryotes show positive interference.
- **Linkage:** the tendency of genes on the same chromosome to be inherited together because they are physically connected; linked genes show less than 50 percent recombination in testcrosses.
- **Parental (non-recombinant) gametes:** gametes that carry the same combination of alleles as the original parental chromosomes; produced at higher frequency than recombinant gametes when genes are linked.
- **Product rule:** P(A and B) = P(A) × P(B) for independent events; used to calculate the probability that two independent genetic events both occur simultaneously.
- **Recombinant gametes:** gametes that carry new combinations of alleles produced by crossing over between linked loci during Meiosis I; less frequent than parental gametes when genes are linked.
- **Recombination frequency (RF):** the proportion of offspring in a testcross that are recombinant: RF = recombinants/total × 100 percent; ranges from 0 percent (completely linked) to 50 percent (unlinked); equals map distance in centimorgans for short distances.
- **Sex-influenced inheritance:** autosomal inheritance in which the expression of a gene differs between males and females because of differences in sex hormone environment; example: male-pattern baldness (dominant in males; recessive in females).
- **Sum rule:** P(A or B) = P(A) + P(B) for mutually exclusive events; used to combine the probabilities of different genotypic combinations that give the same phenotypic outcome.

- **X-inactivation (Lyon hypothesis):** the random inactivation of one X chromosome in each somatic cell of female mammals early in embryonic development; the inactive X is condensed into a Barr body; females with X-linked heterozygosity are functional mosaics.
- **X-linked inheritance:** inheritance of genes located on the X chromosome; X-linked recessive conditions are more common in males (hemizygous) than in females (who need two copies of the recessive allele to be affected).
- **Y-linked (holandric) inheritance:** inheritance of genes located only on the Y chromosome; transmitted from affected father to ALL sons and to no daughters; Y-linked traits skip females entirely.

Concept Check Questions [Series 2]

Objective Questions

1. The product rule states P(A and B) = P(A) × P(B) for _____ events; the probability that offspring of Aa x Aa parents are homozygous recessive (aa) = 1/2 × 1/2 = 1/4. (Linked to SLO 2.1)
2. The binomial formula $P(k) = [n!/k!(n-k)!] \times p^k \times q^{n-k}$ calculates the probability of exactly _____ successes in n independent trials with probability p per trial. (Linked to SLO 2.2)
3. In the XX/XY sex determination system, the sex of offspring is determined by which _____ (X-bearing or Y-bearing) fertilises the egg; males are hemizygous for X-linked genes. (Linked to SLO 2.3)
4. Genes on the same chromosome are _____ and tend to be inherited together; they show less than 50 percent recombination in testcrosses. (Linked to SLO 2.4)
5. Recombination frequency (RF) = (number of recombinant offspring / total offspring) × 100 percent; one percent RF is equivalent to one _____ (cM) of map distance. (Linked to SLO 2.4)

Short-Answer Questions

1. State the product rule and the sum rule and explain how they are used together to calculate genotype frequencies. (Linked to SLO 2.1)
2. Both parents are Aa. Using probability rules, calculate the probability of each genotype and each phenotype among their offspring. (Linked to SLO 2.2)
3. Apply the binomial formula to calculate the probability that exactly 2 of 5 children from two Aa parents will be affected (aa). (Linked to SLO 2.2)
4. A carrier female (X^AX^a) marries a normal male (X^AY). List all offspring genotypes and their probabilities and state the probability of having an affected son. (Linked to SLO 2.3)
5. In a testcross of a dihybrid (AB/ab × ab/ab), 440 parental AB offspring, 452 parental ab offspring, 56 recombinant Ab offspring, and 52 recombinant aB offspring were obtained. Calculate RF and the map distance. (Linked to SLO 2.4)

Long-Answer Questions

1. Explain the product rule, sum rule, conditional probability, and Bayes' theorem with genetic examples. Describe the application of probability to monohybrid and dihybrid crosses and explain the binomial formula with a worked example. (Linked to SLOs 2.1, 2.2)
2. Explain sex determination and X-linked inheritance with probability calculations for carrier female crosses. Distinguish Y-linked and sex-influenced inheritance. Explain genetic linkage,

calculate recombination frequency from testcross data, describe three-point mapping, and explain interference and genetic maps. (Linked to SLOs 2.3, 2.4)

Answers to Concept Check Questions [Series 2]

Objective Questions

1. Independent.
2. k.
3. Sperm.
4. Linked.
5. Centimorgan.

Short-Answer Questions

1. Product rule: P(A and B) = P(A) × P(B); used to calculate the probability that both independent events occur (e.g. P(AA) = P(mother contributes A) × P(father contributes A) = 1/2 × 1/2 = 1/4). Sum rule: P(A or B) = P(A) + P(B) for mutually exclusive events; used to add up the probabilities of combinations that give the same phenotype (e.g. P(dominant phenotype) = P(AA) + P(Aa) = 1/4 + 1/2 = 3/4).
2. P(AA) = 1/2 × 1/2 = 1/4; P(Aa) = (1/2 × 1/2) + (1/2 × 1/2) = 1/2; P(aa) = 1/4. Phenotype: P(dominant, A_) = P(AA) + P(Aa) = 3/4; P(recessive, aa) = 1/4.
3. n = 5; k = 2; p = 1/4 (affected); q = 3/4 (unaffected). P(2 affected of 5) = [5!/(2! × 3!)] × $(1/4)^2 \times (3/4)^3$ = 10 × 1/16 × 27/64 = 270/1,024 ≈ 26.4 percent.
4. Offspring: X^AX^A (1/4, normal female); X^AX^a (1/4, carrier female); X^AY (1/4, normal male); X^aY (1/4, affected male). P(affected son) = 1/4 among all offspring = 1/2 among sons.
5. Total = 440 + 452 + 56 + 52 = 1,000. Recombinants = 56 + 52 = 108. RF = 108/1,000 × 100 = 10.8 percent. Map distance = 10.8 cM.

Long-Answer Questions

1. Product rule: P(A and B) = P(A) × P(B); independent events; P(AA from Aa x Aa) = 1/4. Sum rule: P(A or B) = P(A) + P(B); mutually exclusive events; P(A_) = 3/4. Conditional probability: P(A|B) = P(A and B)/P(B); used in pedigree analysis (e.g. an unaffected offspring from Aa x Aa has 2/3 probability of being Aa, not 1/2, because being unaffected is the condition). Bayes' theorem: posterior P(carrier) = [P(evidence|carrier) × P(carrier)] / P(evidence); example: prior P(carrier) = 1/2; three unaffected children (non-carrier partner) reduces posterior P(carrier) to 1/9 ≈ 11 percent. Monohybrid: Aa x Aa; P(AA) = 1/4; P(Aa) = 1/2; P(aa) = 1/4; P(dominant) = 3/4; P(recessive) = 1/4. Test cross Aa x aa: P(Aa) = 1/2; P(aa) = 1/2 (1:1 ratio). Dihybrid: AaBb x AaBb; P(A_B_) = 9/16; P(A_bb) = 3/16; P(aaB_) = 3/16; P(aabb) = 1/16. Trihybrid: P(dominant all three) = $(3/4)^3$ = 27/64. Binomial: P(k) = [n!/k!(n-

k)!] × p^k × q^{n-k}; example: P(2 of 5 affected, p=1/4) = 10 × (1/16) × (27/64) ≈ 26.4 percent; P(at least 1 affected of 5) = 1 - $(3/4)^5$ ≈ 76.3 percent.

2. Sex determination: XX/XY (females homogametic; males heterogametic; P(female) = 1/2); ZW/ZZ (birds); XX/X0 (grasshoppers); haplodiploidy (Hymenoptera). X-linked inheritance: males hemizygous (X^AY or X^aY); X-linked recessive more common in males; carrier female (X^AX^a) x normal male (X^AY): 1/4 normal female; 1/4 carrier female; 1/4 normal male; 1/4 colour-blind male; P(affected son) = 1/2 among sons; X-linked dominant: affected father → all daughters affected, no sons affected (criss-cross); X-inactivation: random in females; functional mosaic (Barr body). Y-linked: father to all sons; no daughters; holandric. Sex-influenced: autosomal; different dominance in males vs females; male-pattern baldness (dominant in males; recessive in females). Linkage: genes on same chromosome; parental classes > recombinant classes in testcross; RF = recombinants/total × 100 percent; 1 percent RF = 1 cM; example RF = 108/1,000 = 10.8 cM. Three-point testcross: parental classes most frequent; double recombinants least frequent; middle gene identified by inversion in double recombinant vs parental; map distances summed; interference I = 1 - C (positive I = inhibition of nearby crossovers); Haldane function corrects for multiple crossovers.

Answers to Pilot Questions [Series 2]

Part 2.1

1. Probability is a numerical measure of the likelihood that a specific event will occur, ranging from 0 (impossible) to 1 (certain). Complement rule: P(not A) = 1 - P(A). Product rule: P(A and B) = P(A) × P(B) for independent events. Sum rule: P(A or B) = P(A) + P(B) for mutually exclusive events.
2. P(AA) = 1/2 × 1/2 = 1/4. P(Aa) = (1/2 × 1/2) + (1/2 × 1/2) = 1/2. P(aa) = 1/2 × 1/2 = 1/4.
3. The dominant phenotype includes two mutually exclusive genotypes (AA and Aa). Use the product rule to find P(AA) = 1/4 and P(Aa) = 1/2 separately; then use the sum rule to combine them: P(dominant) = 1/4 + 1/2 = 3/4.
4. Conditional probability P(A|B) = P(A and B)/P(B) is the probability of event A given that event B has already occurred. In pedigree analysis, knowing a person is phenotypically normal conditions on them being either AA or Aa (not aa), which changes the probability ratio from 1:2:1 to 1:2 (AA:Aa). This is used to calculate the probability that an unaffected individual is a carrier.
5. Prior P(carrier) = 1/2; prior P(non-carrier) = 1/2. P(2 unaffected children | carrier, non-carrier partner) = $(1/2)^2$ = 1/4. P(2 unaffected children | non-carrier) = 1. Posterior P(carrier) = (1/2 × 1/4) / [(1/2 × 1/4) + (1/2 × 1)] = (1/8) / (1/8 + 1/2) = (1/8) / (5/8) = 1/5 = 20 percent.

Part 2.2

1. P(AA) = 1/4; P(Aa) = 1/2; P(aa) = 1/4. Phenotypes (A dominant): P(dominant, A_) = 3/4; P(recessive, aa) = 1/4.
2. P(A_) = 3/4; P(aa) = 1/4; P(B_) = 3/4; P(bb) = 1/4 (independent genes). P(A_B_) = 3/4 × 3/4 = 9/16. P(A_bb) = 3/4 × 1/4 = 3/16. P(aaB_) = 1/4 × 3/4 = 3/16. P(aabb) = 1/4 × 1/4 = 1/16.
3. n = 4; k = 1; p = 1/4; q = 3/4. P(exactly 1 affected of 4) = [4!/(1! × 3!)] × $(1/4)^1$ × $(3/4)^3$ = 4 × 1/4 × 27/64 = 108/256 = 27/64 ≈ 42.2 percent.
4. n = 4; k = 3 unaffected; p(unaffected) = 3/4; q(affected) = 1/4. P(exactly 3 unaffected of 4) = [4!/(3! × 1!)] × $(3/4)^3$ × $(1/4)^1$ = 4 × 27/64 × 1/4 = 108/256 = 27/64 ≈ 42.2 percent. (Same calculation as the previous question by symmetry.)
5. P(dominant for all three) = $(3/4)^3$ = 27/64 ≈ 42.2 percent. P(recessive for all three, aabbcc) = $(1/4)^3$ = 1/64 ≈ 1.6 percent.

Part 2.3

1. Females are XX (homogametic; produce only X eggs); males are XY (heterogametic; produce X-bearing and Y-bearing sperm in equal numbers). X-bearing sperm + egg = XX (female); Y-bearing sperm + egg = XY (male). P(female) = 1/2; P(male) = 1/2.
2. Males have only one X chromosome; they are hemizygous for X-linked genes and must express whatever allele is on their single X, whether dominant or recessive. A female needs two copies of the recessive allele (one on each X) to express an X-linked recessive condition; a male needs only one copy, so X-linked recessive conditions are much more common in males.
3. Gametes from X^AX^a: X^A (1/2) or X^a (1/2). Gametes from X^AY: X^A (1/2) or Y (1/2). Offspring: X^AX^A (1/4, normal female); X^AX^a (1/4, carrier female); X^AY (1/4, normal male); X^aY (1/4, colour-blind male). P(colour-blind son) = 1/4 among all offspring; = 1/2 among sons only.
4. Y-linked: transmitted from affected father to ALL sons; no daughters ever affected; trait appears exclusively in males; example: AZF deletions causing azoospermia. X-linked recessive: affected males received the recessive allele from their carrier mothers; affected males cannot transmit the trait to sons (sons receive Y from father, not X); trait can appear in females if homozygous recessive.
5. Sex-influenced inheritance involves autosomal genes whose expression differs between males and females because of different sex hormone environments. Example: male-pattern baldness; the baldness allele (B) acts as dominant in males (Bb = bald) but as recessive in females (Bb = non-bald); only BB females are bald.

Part 2.4

1. Linked genes are on the same chromosome and are physically connected; they do not assort independently because they tend to travel together during meiosis. Independent assortment requires that different gene pairs separate independently during Meiosis I, which only applies to genes on different chromosomes or to genes so far apart on the same chromosome that recombination between them occurs at 50 percent frequency.
2. Recombinants = 48 + 52 = 100. Total = 430 + 420 + 48 + 52 = 950. RF = 100/950 × 100 = 10.5 percent. Map distance = 10.5 cM.
3. Parental gametes carry the same allele combinations as the original parents (e.g. AB and ab from a parent of genotype AB/ab); they are produced without crossover and are the more frequent class when genes are linked. Recombinant gametes (Ab and aB) are produced by crossing over between the two loci and are the less frequent class when genes are linked; their frequency equals the recombination frequency.

4. Coefficient of coincidence C = 0.60 means the observed frequency of double crossovers is 60 percent of the frequency expected if the two crossover events were independent. Interference I = 1 - 0.60 = 0.40 means 40 percent positive interference: the occurrence of one crossover reduces the probability of a second crossover nearby by 40 percent, which is the typical situation in eukaryotes.
5. A genetic map shows the relative positions of genes on a chromosome and the distances between them in centimorgans (cM) based on recombination frequency data; distances are additive for short intervals but underestimate true distances for long intervals due to multiple crossovers. A physical map shows the actual positions of genes in base pairs (bp) or megabases (Mb) of DNA sequence determined by molecular techniques; the correspondence between genetic and physical distances is uneven due to recombination hot spots and cold spots.

References and Attributions [Series 2]

Textual References (Books and External Sources)

- Brooker, R. J. (2015). Genetics: Analysis and principles (5th ed.). McGraw-Hill, New York.
- Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., & Doebley, J. (2015). Introduction to genetic analysis (11th ed.). W. H. Freeman, New York.
- Snustad, D. P., & Simmons, M. J. (2012). Principles of genetics (6th ed.). John Wiley & Sons, Hoboken.
- Strachan, T., & Read, A. (2011). Human molecular genetics (4th ed.). Garland Science, New York.

Figures, Plates, Tables, and Boxes

- Figure 2.1: Genetic linkage map and three-point testcross diagram. Attribution: AI generated. Suggested OER search string: “genetic linkage map three-point testcross recombination frequency centimorgans OER”.

Course code: BIO 201

Course title: Genetics I

Course credit load: 2 Units

Series 3: Tests of Goodness of Fit

- **Part 3.1: Introduction to Statistical Hypothesis Testing in Genetics**
  - Sub Part 3.1.1: Why statistics are needed in genetics
  - Sub Part 3.1.2: Null and alternative hypotheses
  - Sub Part 3.1.3: Type I and Type II errors and significance levels
- **Part 3.2: The Chi-Square Test of Goodness of Fit**
  - Sub Part 3.2.1: Rationale and formula for chi-square
  - Sub Part 3.2.2: Degrees of freedom and chi-square distribution
  - Sub Part 3.2.3: Worked examples in monohybrid and dihybrid crosses
- **Part 3.3: Chi-Square Test of Independence**
  - Sub Part 3.3.1: Rationale and differences from the goodness-of-fit test
  - Sub Part 3.3.2: Setting up a contingency table
  - Sub Part 3.3.3: Worked examples in genetics
- **Part 3.4: Other Goodness-of-Fit Approaches in Genetics**
  - Sub Part 3.4.1: The G-test (likelihood ratio test)
  - Sub Part 3.4.2: Poisson distribution and rare genetic events
  - Sub Part 3.4.3: Power, sample size, and limitations of goodness-of-fit tests

Introduction

When Gregor Mendel reported his pea crossing results, he did not obtain perfect 3:1 or 9:3:3:1 ratios. He obtained results that were close to those ratios but deviated from them because meiosis and fertilisation are random processes and any small sample of offspring will show random sampling variation. The central question in every genetic experiment is therefore: are the observed numbers close enough to the expected numbers to be explained by chance, or is the deviation so large that it suggests the hypothesis (e.g. the 3:1 ratio of Mendelian segregation) is wrong?

This Series introduces the statistical methods used to answer this question, focusing on the chi-square (χ^2) goodness-of-fit test as the primary tool for evaluating whether observed genetic data are consistent with a specific expected ratio. The Series covers the logic of statistical hypothesis testing (null hypothesis, significance level, Type I and II errors); the chi-square formula, degrees of freedom, and critical values; worked examples with monohybrid and dihybrid cross data; the chi-square test of independence for analysing contingency tables; and additional approaches including the G-test, the Poisson distribution for rare genetic events, and the practical limitations of goodness-of-fit testing in genetics.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 3.1** explain why statistical testing is necessary in genetics, formulate null and alternative hypotheses for genetic data, and distinguish Type I from Type II errors (Linked to Part 3.1),
- **SLO 3.2** apply the chi-square goodness-of-fit formula to genetic data, determine degrees of freedom, interpret the chi-square value using a critical values table, and draw a conclusion about whether observed data fit a hypothesised ratio (Linked to Part 3.2),
- **SLO 3.3** explain the rationale for the chi-square test of independence, construct a contingency table, calculate expected values, compute chi-square, and interpret the result in a genetic context (Linked to Part 3.3), and

- **SLO 3.4** describe the G-test as an alternative to chi-square, apply the Poisson distribution to rare genetic events, and explain the effect of sample size on the power of goodness-of-fit tests (Linked to Part 3.4).

3.1 Introduction to Statistical Hypothesis Testing in Genetics

3.1.1 Why statistics are needed in genetics

Genetic experiments produce counts of individuals in discrete phenotypic or genotypic classes. These counts are subject to sampling variation: because meiosis and fertilisation are random, the actual offspring obtained in any experiment deviate from the theoretical expected ratio simply by chance. For example, if a 3:1 ratio is expected, a small experiment of 4 offspring might give 4:0 or 2:2 simply by chance, even if the 3:1 ratio is the true underlying ratio. As sample size increases, the observed ratio approaches the true underlying ratio (the Law of Large Numbers), but there will always be some random deviation. Statistics provides the tools to decide whether the deviation between observed and expected counts is within the range of normal random variation (and the hypothesis should be accepted) or whether the deviation is so large that it is unlikely to be due to chance alone (and the hypothesis should be rejected). Without statistical testing, a geneticist cannot distinguish a real biological signal from random noise.

The fundamental question answered by goodness-of-fit tests: do the observed data fit (are consistent with) a specific theoretical model (e.g. a Mendelian 3:1 ratio, a 9:3:3:1 ratio, or a 1:2:1 ratio)? The answer is probabilistic, not certain: statistical tests yield a probability (p-value) that expresses the likelihood of obtaining the observed deviation from the expected ratio by chance alone, assuming the theoretical model is correct. A high p-value (e.g. 0.8) indicates that the observed data are very consistent with the theoretical model; a low p-value (e.g. 0.02) indicates that the observed data are unlikely to arise by chance if the theoretical model is correct. The most widely used goodness-of-fit test in genetics is the chi-square (χ^2) test, introduced by Karl Pearson in 1900. It is particularly suitable for genetic data because it works with counts of individuals in discrete categories.

Fill in the blank: Statistical goodness-of-fit tests in genetics answer the question: do the _____ data fit (are consistent with) a specific theoretical ratio such as the Mendelian 3:1? The test yields a p-value expressing the probability of obtaining the observed deviation by chance alone if the theoretical model is correct.

3.1.2 Null and alternative hypotheses

Statistical hypothesis testing always involves two competing hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_1 or H_a). The null hypothesis (H_0): the hypothesis of no difference, no effect, or no association; in genetics, H_0 typically states that the observed data are consistent with a specific theoretical model (e.g. 'the data fit a 3:1 ratio'; 'the two genes assort independently'; 'the observed frequencies conform to Hardy-Weinberg equilibrium'). The null hypothesis is the hypothesis that is directly tested by the statistical test. The alternative hypothesis (H_1): the hypothesis that is accepted if H_0 is rejected; it states that the observed data are NOT consistent with the theoretical model (e.g. 'the data do not fit a 3:1 ratio'; 'the two genes are linked'; 'the population is not in Hardy-Weinberg equilibrium'). The alternative hypothesis is not directly tested; it is accepted by default when H_0 is rejected.

Directional (one-tailed) vs non-directional (two-tailed) tests: in most genetic goodness-of-fit tests, a two-tailed test is used because we are asking whether the data deviate from the expected in either direction (too many or too few of any class). One-tailed tests are used when the alternative hypothesis specifies the direction of deviation. In genetics, the chi-square goodness-of-fit test is inherently one-tailed (only large positive values of χ^2 indicate poor fit; the test statistic cannot be negative). The p-value in hypothesis testing: the p-value is the probability of obtaining a test statistic as extreme as or more extreme than the observed value, assuming H_0 is true. A small p-value means the observed data would be very unlikely if H_0 were true; this provides evidence against H_0. The p-value does NOT give the probability that H_0 is true or false; it only tells us how surprising the data are under H_0.

Fill in the blank: The _____ hypothesis (H_0) in a genetic chi-square test states that the observed data are consistent with the theoretical ratio; it is rejected when the p-value falls below the chosen significance level (α); the alternative hypothesis is then accepted by default.

3.1.3 Type I and Type II errors and significance levels

Because statistical decisions are probabilistic, two types of error can occur: Type I error (false positive; α error): rejecting H_0 when it is actually true; concluding that the data do not fit the expected ratio when in fact they do (the deviation was just due to chance). The probability of committing a Type I error is set by the significance level (α): if $\alpha = 0.05$, there is a 5 percent probability of rejecting H_0 when it is actually true; this means that approximately 1 in 20 experiments that truly follow the expected ratio will incorrectly be classified as deviating from it. Type II error (false negative; β error): failing to reject H_0 when it is actually false; concluding that the data fit the expected ratio when in fact they do not (the experiment lacked sufficient power to detect the true deviation). The probability of committing a Type II error is β; statistical power = 1 - β is the probability of correctly rejecting a false H_0.

Significance level (α): the threshold probability below which H_0 is rejected. In genetics (and most biological sciences), $\alpha = 0.05$ is the conventional threshold: if the p-value from the chi-square test is less than 0.05, H_0 is rejected; the result is said to be statistically significant. If $p \geq 0.05$, H_0 is not rejected (but it is not proved to be true; there may simply be insufficient evidence to reject it). In some contexts (e.g. medical genetics or genome-wide association studies), more stringent significance levels ($\alpha = 0.01$ or even $\alpha = 5 \times 10^{-8}$ for GWAS) are used to reduce the probability of false positives. The relationship between α and β: for a fixed sample size, reducing α (making it harder to reject H_0) increases β (increases the probability of missing a true effect); the only way to reduce both simultaneously is to increase sample size. Power analysis: calculating the minimum sample size needed to detect a true deviation of a given magnitude with a specified power (1 - β) and significance level (α).

Fill in the blank: A Type _____ error (false positive) occurs when H_0 is rejected even though it is true; its probability equals the significance level α (conventionally 0.05 in genetics); a Type II error (false negative) occurs when H_0 is not rejected even though it is false; its probability is β.

Pilot questions 3.1

11. Explain why observed genetic ratios never exactly match the expected theoretical ratio even when Mendel's laws are operating correctly.

12. Define the null hypothesis and the alternative hypothesis in the context of a chi-square test of a 3:1 genetic ratio.
13. What is a p-value? What does a p-value of 0.03 indicate in a genetic chi-square test (using $\alpha = 0.05$)?
14. Distinguish between Type I error and Type II error in statistical hypothesis testing.
15. Why is $\alpha = 0.05$ the conventional significance threshold in genetics and what does it mean in practice?

3.2 The Chi-Square Test of Goodness of Fit

3.2.1 Rationale and formula for chi-square

The chi-square (χ^2) goodness-of-fit test measures how well observed frequencies in different categories match the frequencies expected under a specific theoretical model. It converts the differences between observed (O) and expected (E) counts in each category into a single test statistic that can be compared to a theoretical distribution (the chi-square distribution) to obtain a p-value. The chi-square formula: $\chi^2 = \Sigma [(O - E)^2 / E]$, where the sum ($\Sigma$) is taken over all categories; O = observed count in a category; E = expected count in that category. The expected count for each category = total number of offspring × the expected proportion for that category (from the theoretical ratio). Properties of the χ^2 statistic: it is always non-negative (because squared values are used); a value of 0 means perfect agreement between observed and expected; larger values indicate greater discrepancy; the magnitude of χ^2 that is considered significant depends on the degrees of freedom.

Calculating expected values: for a monohybrid cross of two heterozygotes (Aa x Aa) producing n total offspring, the expected numbers are: E(dominant) = 3n/4; E(recessive) = n/4. Example: 100 total offspring expected as 75 dominant and 25 recessive. If observed: 72 dominant and 28 recessive, then: $\chi^2 = [(72-75)^2/75] + [(28-25)^2/25] = [9/75] + [9/25] = 0.12 + 0.36 = 0.48$. This small χ^2 value suggests the data fit the 3:1 ratio well (the deviation of 3 in each class is minor). Requirement: the chi-square test is only valid when expected values are sufficiently large; the conventional rule is that all expected values should be at least 5 (some texts say at least 1, with no more than 20 percent below 5); if expected values are too small, categories should be

combined or an exact test (e.g. Fisher's exact test) should be used instead. The χ^2 test is a large-sample approximation; it performs poorly with very small samples.

Fill in the blank: The chi-square formula $\chi^2 = \Sigma [(O - E)^2 / E]$ sums the squared _____ between observed and expected counts divided by the expected count over all categories; a larger χ^2 value indicates greater discrepancy between observed and expected data.

3.2.2 Degrees of freedom and chi-square distribution

Degrees of freedom (df) are the number of categories minus the number of independent constraints on the data. For the goodness-of-fit test: df = number of phenotypic classes - 1. The reason for subtracting 1 is that once all but the last category count is fixed, the last is determined by the total (the counts must sum to the total n, removing one degree of freedom). Examples: monohybrid cross (2 phenotypic classes): df = 2 - 1 = 1; dihybrid cross (4 phenotypic classes): df = 4 - 1 = 3; trihybrid cross (8 phenotypic classes): df = 8 - 1 = 7. The chi-square distribution is a family of probability distributions, each defined by its degrees of freedom; it is right-skewed for low df and becomes more symmetric and bell-shaped as df increases. Critical values of χ^2 for different df and significance levels are given in chi-square tables; the critical value χ^{2T} is the value above which only $\alpha \times 100$ percent of samples fall by chance alone when H_0 is true.

How to use the chi-square table: find the row corresponding to the degrees of freedom of the test; find the column corresponding to the desired significance level ($\alpha = 0.05$ is the standard); the number at the intersection is the critical value. Decision rule: if the calculated χ^2 exceeds the critical value, reject H_0 ($p < \alpha$; statistically significant). If the calculated χ^2 is less than or equal to the critical value, do not reject H_0 ($p \geq \alpha$; not statistically significant). Critical values at $\alpha = 0.05$: df = 1: $\chi^2 = 3.841$; df = 2: $\chi^2 = 5.991$; df = 3: $\chi^2 = 7.815$; df = 4: $\chi^2 = 9.488$; df = 5: $\chi^2 = 11.070$. Critical values at $\alpha = 0.01$: df = 1: $\chi^2 = 6.635$; df = 2: $\chi^2 = 9.210$; df = 3: $\chi^2 = 11.345$. A very small χ^2 value (close to 0) with $p > 0.99$ may also be suspicious: it suggests that the observed data fit the expected ratio almost too well, which may indicate data fabrication or selection (Mendel's own data showed suspiciously good fit, noted by Ronald Fisher in 1936).

Fill in the blank: For the chi-square goodness-of-fit test, degrees of freedom = number of phenotypic _____ - 1; for a dihybrid cross with four phenotypic classes, df = 3 and the critical value at $\alpha = 0.05$ is 7.815; if the calculated χ^2 exceeds this critical value, H_0 is rejected.

3.2.3 Worked examples in monohybrid and dihybrid crosses

Worked example 1 (monohybrid cross): Mendel crossed two heterozygous tall pea plants (Tt x Tt). He obtained 787 tall and 277 short offspring (total = 1,064). The expected ratio is 3:1. H_0: the data fit a 3:1 ratio. H_1: the data do not fit a 3:1 ratio. Expected: E(tall) = 1,064 × 3/4 = 798; E(short) = 1,064 × 1/4 = 266. $\chi^2 = [(787 - 798)^2 / 798] + [(277 - 266)^2 / 266] = [121/798] + [121/266] = 0.152 + 0.455 = 0.607$. df = 2 - 1 = 1. Critical value at $\alpha = 0.05$, df = 1: 3.841. Since $0.607 < 3.841$, do NOT reject H_0. Conclusion: the data are consistent with a 3:1 ratio ($p > 0.05$); the deviation from the expected ratio is within the range expected by chance. The p-value can be estimated from chi-square tables as approximately 0.44 (well above 0.05).

Worked example 2 (dihybrid cross): Mendel crossed RrYy x RrYy pea plants. He obtained 315 round yellow, 108 round green, 101 wrinkled yellow, and 32 wrinkled green offspring (total = 556). Expected ratio: 9:3:3:1. H_0: the data fit a 9:3:3:1 ratio. Expected values: E(round yellow) = 556 × 9/16 = 312.75; E(round green) = 556 × 3/16 = 104.25; E(wrinkled yellow) = 556 × 3/16 = 104.25; E(wrinkled green) = 556 × 1/16 = 34.75. $\chi^2 = [(315-312.75)^2/312.75] + [(108-104.25)^2/104.25] + [(101-104.25)^2/104.25] + [(32-34.75)^2/34.75] = [5.0625/312.75] + [14.0625/104.25] + [10.5625/104.25] + [7.5625/34.75] = 0.016 + 0.135 + 0.101 + 0.218 = 0.470$. df = 4 - 1 = 3. Critical value at $\alpha = 0.05$, df = 3: 7.815. Since $0.470 < 7.815$, do NOT reject H_0. Conclusion: Mendel's dihybrid data are consistent with a 9:3:3:1 ratio ($p \approx 0.92$).

Phenotype	Observed (O)	Expected (E)	O − E	(O − E)²	(O − E)² / E
Tall	787	798	−11	121	0.152
Short	277	266	+11	121	0.455
Total	1,064	—	—	—	**$\chi^2 = 0.607$**
Phenotype	Observed (O)	Expected (E)	O − E	(O − E)²	**(O − E)² / E**
Round Yellow	315	312.75	+2.25	5.06	**0.016**
Round Green	108	104.25	+3.75	14.06	**0.135**
Wrinkled Yellow	101	104.25	−3.25	10.56	**0.101**

Wrinkled Green	32	34.75	−2.75	7.56	**0.218**
Total	556	—	—	—	**$\chi^2 = 0.470$**

Box 3.1: Chi-square goodness-of-fit calculations for Mendel's monohybrid (3:1) and dihybrid (9:3:3:1) data

Pilot questions 3.2

1. State the chi-square formula and explain what each term represents.
2. In a cross of Aa x Aa, 180 offspring are obtained: 140 dominant and 40 recessive. Perform a chi-square goodness-of-fit test to determine whether these data fit a 3:1 ratio (use $\alpha = 0.05$).
3. Explain how degrees of freedom are determined for a chi-square goodness-of-fit test and state the df for a test of a 9:3:3:1 ratio.
4. A chi-square goodness-of-fit test of a 3:1 ratio gives $\chi^2 = 5.2$ with df = 1. What is the conclusion? (Critical value at $\alpha = 0.05$, df = 1 is 3.841.)
5. Explain what it means if a chi-square value is suspiciously close to zero (e.g. $p > 0.99$) in a genetic experiment.

3.3 Chi-Square Test of Independence

3.3.1 Rationale and differences from the goodness-of-fit test

The chi-square test of independence (also called the chi-square test of association) tests whether two categorical variables are statistically independent of each other, or whether they are associated. In genetics, the test of independence is used to ask questions such as: Are two genes independent (do they assort independently)? Is the distribution of a trait independent of sex? Is a genetic marker independent of a disease phenotype? The key difference from the goodness-of-fit test: the goodness-of-fit test compares observed frequencies to expected frequencies derived from a pre-specified theoretical model (e.g. the 3:1 ratio comes from Mendel's laws); the test of independence does NOT use a pre-specified ratio; instead, the expected frequencies are calculated from the observed marginal totals of a two-way contingency table, under the assumption that the two variables are independent. The null hypothesis for the test of

independence: H_0: the two variables are independent (not associated); deviations from independence are due to chance. Alternative hypothesis: H_1: the two variables are associated (not independent).

When is the test of independence used in genetics? (1) Testing whether two genes assort independently: observe offspring in a dihybrid testcross and arrange the counts in a 2 x 2 or larger contingency table (rows: alleles at locus A; columns: alleles at locus B); if independent assortment holds, the genotype at one locus is independent of the genotype at the other locus. (2) Testing whether a trait is associated with sex: rows = male/female; columns = trait present/absent; test whether the proportion with the trait is the same in both sexes. (3) Testing for association in genetic epidemiology: rows = disease status (affected/unaffected); columns = genetic variant (allele 1/allele 2); test whether the variant frequency differs between cases and controls (a genetic association study). This is the foundation of genome-wide association studies (GWAS).

Fill in the blank: The chi-square test of _____ tests whether two categorical variables are statistically associated; the expected values are calculated from the observed marginal totals of a contingency table (not from a pre-specified theoretical ratio); H_0 states that the two variables are independent.

3.3.2 Setting up a contingency table

A contingency table (cross-tabulation) arranges the observed counts in a two-dimensional grid, with the categories of one variable along the rows and the categories of the other variable along the columns. For a 2 x 2 contingency table: two row categories (e.g. genotype at locus A: A_ and aa) and two column categories (e.g. genotype at locus B: B_ and bb); four cells contain the observed counts; row totals and column totals (marginals) are computed; the grand total n is placed in the lower right corner. Generic 2 x 2 table: rows are categories of variable 1 (R1; R2); columns are categories of variable 2 (C1; C2); cells are a (R1, C1); b (R1, C2); c (R2, C1); d (R2, C2); row totals R1 = a+b; R2 = c+d; column totals C1 = a+c; C2 = b+d; grand total n = a+b+c+d. Calculating expected values under independence: E(cell) = (row total × column total) / grand total. E(a) = R1 × C1 / n; E(b) = R1 × C2 / n; E(c) = R2 × C1 / n; E(d) = R2 × C2 / n.

These expected values represent what would be observed if the two variables were perfectly independent.

The formula for chi-square in a contingency table: $\chi^2 = \Sigma [(O - E)^2 / E]$, summed over all cells. Degrees of freedom for a contingency table of r rows and c columns: df = (r - 1)(c - 1). For a 2 x 2 table: df = (2-1)(2-1) = 1. For a 2 x 3 table: df = (2-1)(3-1) = 2. For a 3 x 3 table: df = (2)(2) = 4. Yates' continuity correction for 2 x 2 tables: the chi-square approximation is less accurate for 2 x 2 tables with small expected values; Yates' correction applies a continuity correction: $\chi^2 = \Sigma [(|O - E| - 0.5)^2 / E]$; this correction reduces the χ^2 value and makes the test more conservative; it is recommended when any expected value is between 5 and 10; for expected values below 5, Fisher's exact test is preferred. For large samples and all expected values well above 10, Yates' correction is not necessary.

Fill in the blank: In a contingency table, expected values are calculated as E(cell) = (_____ total × column total) / grand total; degrees of freedom = (r - 1)(c - 1); for a 2 x 2 contingency table, df = 1 and the critical value at $\alpha = 0.05$ is 3.841.

3.3.3 Worked examples in genetics

Worked example 1 (testing independent assortment): a testcross of a doubly heterozygous plant (AaBb × aabb) gives offspring. If the genes are independently assorting, we expect equal numbers of the four phenotypic classes. Observed: 112 A_B_; 86 A_bb; 94 aaB_; 108 aabb; total n = 400. Under independent assortment, expected ratio is 1:1:1:1, so E = 400/4 = 100 for each class. $\chi^2 = [(112-100)^2/100] + [(86-100)^2/100] + [(94-100)^2/100] + [(108-100)^2/100] = [144/100] + [196/100] + [36/100] + [64/100] = 1.44 + 1.96 + 0.36 + 0.64 = 4.40$. df = 4 - 1 = 3. Critical value at $\alpha = 0.05$, df = 3: 7.815. Since 4.40 < 7.815, do NOT reject H_0. Conclusion: the data are consistent with independent assortment ($p \approx 0.22$).

Worked example 2 (2 x 2 test of independence): researchers test whether a genetic variant (G or A allele at a candidate locus) is associated with a disease. They genotype 200 affected patients and 200 healthy controls (total 400 individuals; 800 alleles). Observed counts (allele counts): Disease cases: G allele = 140; A allele = 60; total = 200. Controls: G allele = 100; A allele = 100; total = 200. Grand total: G = 240; A = 160; total = 400. Expected values: E(G allele, cases) =

200 × 240/400 = 120; E(A allele, cases) = 200 × 160/400 = 80; E(G allele, controls) = 200 × 240/400 = 120; E(A allele, controls) = 200 × 160/400 = 80. χ^2 = [(140-120)²/120] + [(60-80)²/80] + [(100-120)²/120] + [(100-80)²/80] = [400/120] + [400/80] + [400/120] + [400/80] = 3.33 + 5.00 + 3.33 + 5.00 = 16.67. df = (2-1)(2-1) = 1. Critical value at $\alpha = 0.05$, df = 1: 3.841. Since 16.67 > 3.841, REJECT H_0. Conclusion: the G allele is significantly associated with the disease ($p < 0.001$); the two variables are not independent.

Pilot questions 3.3

1. Distinguish between the chi-square goodness-of-fit test and the chi-square test of independence in terms of the null hypothesis and how expected values are calculated.
2. Set up a 2 x 2 contingency table for a study comparing allele frequencies in 100 affected and 100 unaffected individuals (G allele: 70 in affected, 50 in unaffected; A allele: 30 in affected, 50 in unaffected). Calculate the expected values for each cell.
3. Calculate χ^2 for the contingency table in question 2 and state the df and the conclusion (critical value at $\alpha = 0.05$, df = 1 is 3.841).
4. Explain when Yates' continuity correction should be applied to a 2 x 2 chi-square test.
5. In a dihybrid testcross, 200 offspring are obtained: 60 A_B_, 40 A_bb, 55 aaB_, 45 aabb. Test whether the genes assort independently (use chi-square goodness-of-fit for a 1:1:1:1 ratio, $\alpha = 0.05$).

3.4 Other Goodness-of-Fit Approaches in Genetics

3.4.1 The G-test (likelihood ratio test)

The G-test (also called the likelihood ratio chi-square test or log-likelihood ratio test) is an alternative to the Pearson chi-square test for testing goodness of fit and independence. The G-test formula: $G = 2 \Sigma [O \times \ln(O/E)]$, where the sum is over all categories; O = observed count; E = expected count; ln = natural logarithm. The G statistic follows a chi-square distribution with the same degrees of freedom as the equivalent Pearson χ^2 test (df = number of categories - 1 for goodness of fit; df = (r-1)(c-1) for independence tests). Decision rule: same as for chi-square; reject H_0 if G exceeds the chi-square critical value for the appropriate df and α. Advantages of the G-test over the Pearson chi-square test: the G-test is based on the likelihood ratio framework,

which has more favourable theoretical properties; G values are additive across independent tests (making it useful for combining results from multiple experiments); the G-test is increasingly recommended in modern statistical textbooks for large samples.

Worked example: from the monohybrid example of Part 3.2 (787 tall, 277 short; expected 798, 266): $G = 2 \times [787 \times \ln(787/798) + 277 \times \ln(277/266)] = 2 \times [787 \times \ln(0.9862) + 277 \times \ln(1.0414)] = 2 \times [787 \times (-0.01391) + 277 \times (0.04058)] = 2 \times [(-10.949) + (11.241)] = 2 \times 0.292 = 0.584$. df = 1; critical value at $\alpha = 0.05$: 3.841. Since $0.584 < 3.841$, do NOT reject H_0. This is very close to the Pearson $\chi^2 = 0.607$ from the same data, illustrating that both tests give very similar results with large samples. With small samples, the G-test may give slightly different results; both tests have the same known limitations with small expected counts. For large samples ($n > 100$ and all expected values > 5), both tests are equivalent and either can be used.

Fill in the blank: The G-test formula $G = 2 \Sigma [O \times \ln(O/E)]$ uses the _____ logarithm; like chi-square, it follows the chi-square distribution with the same degrees of freedom; G values from independent tests are additive, which is an advantage over the Pearson chi-square statistic.

3.4.2 Poisson distribution and rare genetic events

The Poisson distribution describes the probability of a given number of rare, independent events occurring in a fixed interval of time, space, or another unit, when the average rate of occurrence is known. It is appropriate when: the events are rare (low probability); the events occur independently of each other; the average rate of occurrence (λ; lambda) is constant. The Poisson probability formula: $P(X = k) = (e^{-\lambda} \times \lambda^k) / k!$, where k = number of events (0, 1, 2, ...); λ = mean number of events per interval; e = Euler's number (approximately 2.718). For the Poisson distribution: mean = λ; variance = λ; standard deviation = $\sqrt{\lambda}$. Applications in genetics: (1) Mutation rate estimation: rare spontaneous mutations at a locus per cell per generation follow a Poisson distribution; if the average mutation rate is λ mutations per locus per gamete per generation, $P(0 \text{ mutations}) = e^{-\lambda}$; $P(1 \text{ mutation}) = \lambda e^{-\lambda}$; etc. (2) Radiation genetics: the number of chromosome breaks induced per cell by ionising radiation follows a Poisson distribution at low radiation doses. (3) Rare disease allele counts in families. (4) Number of crossovers per chromosome per meiosis.

Worked example: the average rate of a specific spontaneous mutation is 2×10^{-5} per gene per gamete (i.e. $\lambda = 2 \times 10^{-5}$ per gamete). What is the probability of a gamete carrying exactly one new mutation at this locus? $P(X = 1) = (e^{-2x10-5} \times (2 \times 10^{-5})^1) / 1! \approx e^{—n521} \times 2 \times 10^{-5} \approx 1 \times 2 \times 10^{-5} = 2 \times 10^{-5}$. P(0 mutations) = $e^{—n521} \approx 0.99998$. Goodness-of-fit test using Poisson: if the observed counts of rare events in multiple experiments are available, a chi-square or G-test can test whether the observed distribution fits a Poisson distribution with a given λ (testing whether events are truly random and independent). Departure from a Poisson distribution (overdispersion: variance > mean; or underdispersion: variance < mean) can indicate biological structure in the data (e.g. mutation clustering, non-random distribution of crossovers).

Fill in the blank: The Poisson distribution describes the probability of rare independent events with mean rate λ; the formula $P(X = k) = (e^{-}\lambda \times \lambda^k) / k!$ applies; the mean and _____ of a Poisson distribution are both equal to λ; departure from Poisson expectations (overdispersion) may indicate non-random clustering of events.

3.4.3 Power, sample size, and limitations of goodness-of-fit tests

Statistical power is the probability that a test correctly rejects H_0 when H_0 is false (i.e. the probability of detecting a true difference when one exists): Power = 1 - β. Power depends on: (1) Sample size (n): larger samples give more power; more offspring means the observed ratio is closer to the true underlying ratio, making it easier to detect deviations from the hypothesised ratio; (2) Effect size: larger deviations from the expected ratio are easier to detect; a true 2.5:1 ratio is easier to distinguish from a 3:1 hypothesis than a 2.9:1 ratio; (3) Significance level (α): more stringent α (e.g. 0.01 instead of 0.05) reduces power (increases Type II error rate) because the threshold for rejection is higher; (4) Number of categories (df): tests with more categories (higher df) have less power per unit of sample size for detecting a given deviation. Power analysis in genetics: before conducting a genetic cross experiment, a researcher can calculate the minimum sample size needed to detect a specific alternative ratio with a desired power (e.g. 80 percent or 90 percent) at a given significance level. Standard formula (for goodness-of-fit with 2 categories, df = 1): $n \approx (\chi^2\alpha + \chi^2\beta)^2 / w^2$, where w is the effect size (Cohen's $w = \sqrt{(\chi^2/n)}$ from the expected non-central chi-square).

Limitations of chi-square goodness-of-fit tests in genetics: (1) Minimum expected count requirement: all expected values should be ≥ 5 (or at most 20 percent below 5); violations require combining categories or using Fisher's exact test (for 2 x 2 tables) or mid-P test; this is particularly problematic in crosses with many phenotypic classes and small total offspring. (2) Independence of observations: chi-square requires that each observed individual is independent; if families or sibling groups are used as observational units without accounting for relatedness, the test is invalid. (3) Chi-square tests on derived ratios: computing chi-square on ratios or percentages rather than raw counts is incorrect; the test requires raw counts. (4) Multiple testing: performing many chi-square tests on the same dataset inflates the probability of a Type I error; Bonferroni correction or other multiple testing corrections are required. (5) Directionality: chi-square is non-directional; it detects any deviation from the expected ratio but cannot tell which specific class is in excess or deficit; this requires inspection of the contributions of individual categories to the total χ^2. (6) Mendel's data paradox: Fisher (1936) calculated that Mendel's combined data across all his experiments fit the expected ratios suspiciously well (χ^2 values too small; p-values too high), suggesting possible data selection or unconscious bias in recording.

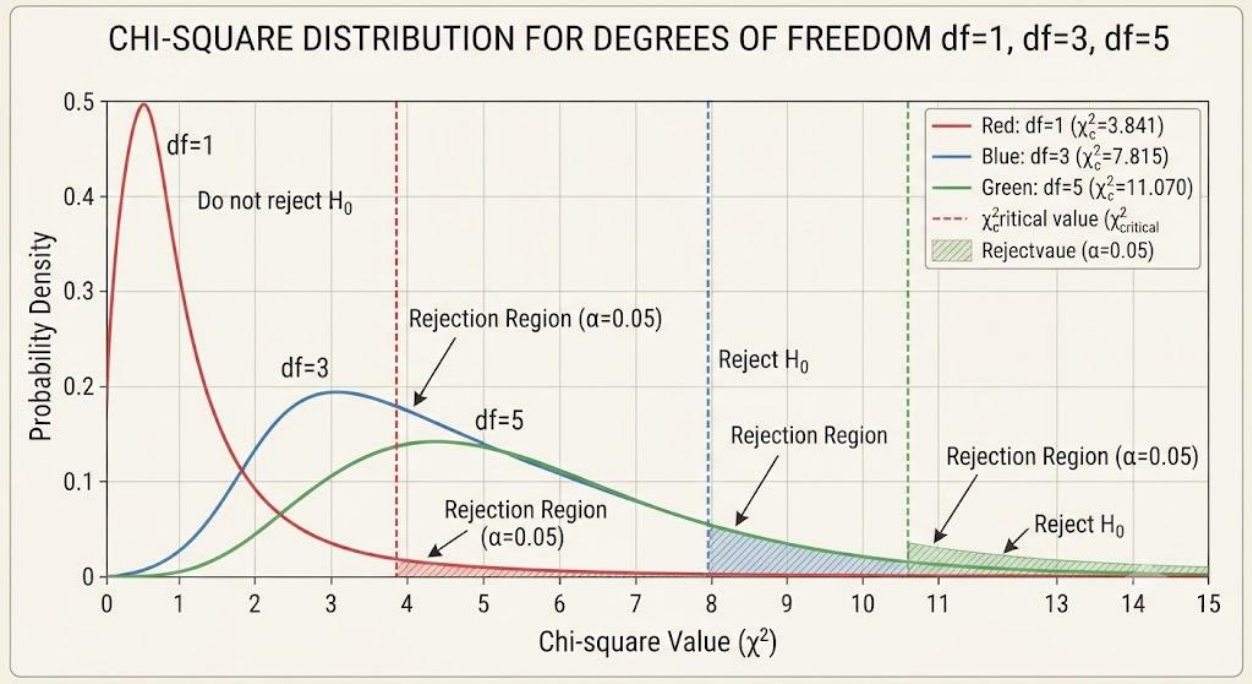

Figure 3.1: Chi-square probability distributions for df = 1, df = 3, and df = 5, showing critical values at α = 0.05 and the rejection regions

Pilot questions 3.4

1. State the G-test formula and explain one advantage it has over the Pearson chi-square test.
2. Calculate the G statistic for observed counts of 140 tall and 60 short from a cross where a 3:1 ratio is expected (total = 200; expected 150 tall and 50 short). Use $\alpha = 0.05$.
3. Define the Poisson distribution and state two genetic situations where it applies.
4. State four limitations of the chi-square goodness-of-fit test in genetic experiments.
5. Explain the effect of sample size on the power of a chi-square goodness-of-fit test in genetics.

Series Summary

This Series introduced the use of **statistical methods in genetics**, explaining why observed genetic results often differ from theoretical expectations because of random variation during meiosis and fertilisation. It covered the principles of hypothesis testing, including the **null and alternative hypotheses**, **p-values**, **Type I and Type II errors**, statistical power, and the **chi-square goodness-of-fit test** for comparing observed and expected genetic ratios. The Series also demonstrated the calculation and interpretation of chi-square values, degrees of freedom, critical values, and the application of these concepts using Mendel's classic monohybrid and dihybrid experiments.

Building on these foundations, the Series explored the **chi-square test of independence**, expected frequencies in contingency tables, Yates' continuity correction, Fisher's exact test, the **G-test**, and the **Poisson distribution** for analysing rare genetic events. It also discussed the assumptions and limitations of these statistical methods, including minimum expected frequencies, independent observations, multiple testing corrections, and statistical power. By completing this Series, you should now be able to apply appropriate statistical tests to genetic data, interpret their results correctly, evaluate the fit of observed inheritance patterns to theoretical expectations, and achieve all four Series Learning Outcomes (SLOs).

Glossary of Terms

- **Alternative hypothesis (H_1):** the hypothesis that is accepted if the null hypothesis is rejected; in genetics, it states that the observed data do not fit the theoretical model (e.g. the ratio is not 3:1).
- **Chi-square (χ^2) goodness-of-fit test:** a statistical test that measures how well observed frequencies match expected frequencies from a theoretical model; $\chi^2 = \Sigma\ [(O-E)^2/E]$; used to evaluate whether observed genetic ratios are consistent with Mendelian predictions.
- **Chi-square test of independence:** a statistical test of whether two categorical variables are associated or independent; expected values are calculated from the marginal totals of a contingency table; df = (r-1)(c-1).
- **Contingency table:** a two-dimensional table of observed counts for two categorical variables; used in the chi-square test of independence; expected values are calculated as (row total × column total) / grand total.

- **Degrees of freedom (df):** the number of independent data values free to vary after constraints are imposed; for chi-square goodness-of-fit: df = number of categories - 1; for contingency table: df = (r-1)(c-1).
- **Fisher's exact test:** an exact test of independence for 2 x 2 contingency tables with small expected values (below 5); computes the exact probability of the observed or more extreme table configurations.
- **G-test (likelihood ratio test):** an alternative to the Pearson chi-square test; $G = 2 \Sigma [O \times \ln(O/E)]$; follows the chi-square distribution with the same df; G values are additive across independent tests.
- **Null hypothesis (H_0):** the hypothesis directly tested by a statistical test; in genetics, it typically states that the observed data are consistent with the theoretical model (e.g. the data fit a 3:1 ratio).
- **p-value:** the probability of obtaining a test statistic as extreme as or more extreme than the observed value, assuming H_0 is true; a small p-value (below α) indicates evidence against H_0.
- **Poisson distribution:** a probability distribution describing the number of rare, independent events in a fixed interval; $P(X=k) = (e^{-\lambda} \times \lambda^k)/k!$; mean = variance = λ; applies to mutation rates, radiation-induced breaks, and rare genetic events.
- **Power (statistical):** the probability of correctly rejecting a false null hypothesis; Power = 1 - β; increases with larger sample size, larger effect size, and less stringent significance level.
- **Significance level (α):** the threshold probability below which H_0 is rejected; conventionally α = 0.05 in genetics; equals the probability of a Type I error.
- **Sampling variation:** random deviation of observed counts from expected counts in any finite sample; the reason why genetic experiments never produce exactly the theoretical ratio; decreases as sample size increases.
- **Type I error (α error):** rejecting H_0 when it is actually true (false positive); its probability equals the significance level α; in genetics: concluding that data do not fit the expected ratio when they actually do.
- **Type II error (β error):** failing to reject H_0 when it is actually false (false negative); its probability equals β; in genetics: concluding that data fit the expected ratio when they actually do not.
- **Yates' continuity correction:** a correction applied to 2 x 2 chi-square tests with small expected values: $\chi^2 = \Sigma [(|O-E|-0.5)^2/E]$; reduces the chi-square value and makes the test more conservative.

Concept Check Questions [Series 3]

Objective Questions

1. Statistical goodness-of-fit tests in genetics yield a _____ expressing the probability that the observed deviation from the expected ratio arose by chance alone, assuming the null hypothesis is true. (Linked to SLO 3.1)
2. The chi-square formula is $\chi^2 = \Sigma [(O - E)^2 / E]$; a value of zero indicates _____ agreement between observed and expected; larger values indicate greater discrepancy. (Linked to SLO 3.2)
3. For a chi-square goodness-of-fit test of a dihybrid 9:3:3:1 ratio (four phenotypic classes), the degrees of freedom = 4 - 1 = _____; the critical value at $\alpha = 0.05$ is 7.815. (Linked to SLO 3.2)
4. In a chi-square test of independence, expected values are calculated as (_____ total × column total) / grand total; this differs from the goodness-of-fit test where expected values come from a pre-specified theoretical ratio. (Linked to SLO 3.3)
5. The G-test formula is $G = 2 \Sigma [O \times \ln(O/E)]$ and follows the chi-square distribution; an advantage over the Pearson chi-square is that G values from independent tests are _____ . (Linked to SLO 3.4)

Short-Answer Questions

1. Define the null hypothesis and the alternative hypothesis for a chi-square test of Mendelian 3:1 ratio, and explain what the p-value represents. (Linked to SLO 3.1)
2. State the chi-square formula and show, step by step, how to calculate χ^2 for observed counts of 85 dominant and 35 recessive offspring from a cross expected to give a 3:1 ratio (total n = 120). (Linked to SLO 3.2)
3. Distinguish between Type I and Type II errors and state the conventional significance level used in genetics. (Linked to SLO 3.1)
4. Explain the difference between the chi-square goodness-of-fit test and the chi-square test of independence in terms of the null hypothesis and calculation of expected values. (Linked to SLO 3.3)
5. State two limitations of the chi-square goodness-of-fit test that a genetics student should be aware of. (Linked to SLO 3.4)

Long-Answer Questions

1. Explain why statistical testing is necessary in genetics. Describe the logic of hypothesis testing (null hypothesis, p-value, Type I and II errors, significance level). Explain the chi-

square goodness-of-fit test (formula, degrees of freedom, critical values) and perform the test on a monohybrid dataset. (Linked to SLOs 3.1, 3.2)

2. Describe the chi-square test of independence, explain how expected values are calculated from a contingency table, and perform the test on a genetic dataset. Describe the G-test, the Poisson distribution in genetics, and the limitations of goodness-of-fit tests including power and sample size. (Linked to SLOs 3.3, 3.4)

Answers to Concept Check Questions [Series 3]

Objective Questions

1. p-value.
2. Perfect.
3. 3.
4. Row.
5. Additive.

Short-Answer Questions

1. H_0: the observed data are consistent with a 3:1 ratio (any deviation is due to chance). H_1: the observed data do not fit a 3:1 ratio. The p-value is the probability of obtaining the observed chi-square value or larger by chance alone, assuming H_0 is true; a p-value below $\alpha = 0.05$ leads to rejection of H_0.
2. Formula: $\chi^2 = \Sigma [(O-E)^2/E]$. n = 120; E(dominant) = 120 × 3/4 = 90; E(recessive) = 120 × 1/4 = 30. $\chi^2 = [(85-90)^2/90] + [(35-30)^2/30] = [25/90] + [25/30] = 0.278 + 0.833 = 1.111$. df = 1; critical value at $\alpha = 0.05$: 3.841. Since $1.111 < 3.841$, do not reject H_0; the data are consistent with a 3:1 ratio.
3. Type I error: rejecting H_0 when it is true (false positive); probability = α. Type II error: failing to reject H_0 when it is false (false negative); probability = β. The conventional significance level in genetics is $\alpha = 0.05$.
4. Goodness-of-fit test: H_0 states that data fit a pre-specified theoretical ratio (e.g. 3:1); expected values are calculated from that ratio (E = n × expected proportion). Test of independence: H_0 states that two categorical variables are independent; expected values are calculated from the observed marginal totals of a contingency table (E = row total × column total / grand total).

5. All expected values must be at least 5; violations require combining categories or using Fisher's exact test. The chi-square test requires independent observations; data from related individuals (siblings, families) violate this assumption without appropriate correction.

Long-Answer Questions

1. Need for statistics: meiosis and fertilisation are random; any finite sample shows sampling variation (observed ratio deviates from theoretical by chance); statistics determines whether the deviation is within the range of chance variation. Hypothesis testing: H_0 (data fit the theoretical model); H_1 (data do not fit); p-value (probability of observed or more extreme deviation assuming H_0); Type I error ($\alpha = 0.05$; rejecting true H_0); Type II error (β; failing to reject false H_0); Power = 1 - β; significance level $\alpha = 0.05$ in genetics. Chi-square formula: $\chi^2 = \Sigma [(O-E)^2/E]$; df = categories - 1; critical values at $\alpha = 0.05$ (df=1: 3.841; df=3: 7.815); reject H_0 if χ^2 > critical value. Monohybrid example: Mendel's data 787 tall, 277 short (n=1,064); E(tall) = 798; E(short) = 266; $\chi^2 = 0.607$; df = 1; 0.607 < 3.841; do not reject H_0; data consistent with 3:1 ratio ($p \approx 0.44$).
2. Chi-square test of independence: H_0 (two variables are independent); expected values from marginal totals (E = row total × column total / n); $\chi^2 = \Sigma [(O-E)^2/E]$; df = (r-1)(c-1); Yates' correction for 2 x 2 with small E; Fisher's exact test when E < 5. Dihybrid testcross example: 112:86:94:108 (n=400); E=100 each; $\chi^2 = 4.40$; df=3; < 7.815; do not reject H_0 (independent assortment). Disease association example: G allele 140 cases/100 controls; $\chi^2 = 16.67$; df=1; > 3.841; reject H_0 (G allele associated with disease). G-test: $G = 2\Sigma[O \times \ln(O/E)]$; same df as Pearson χ^2; G values additive across independent tests; preferred in modern statistics. Poisson distribution: $P(X=k) = (e^{-\lambda} \times \lambda^k)/k!$; mean = variance = λ; applies to mutation rates, radiation breaks, rare crossover events. Limitations: $E \geq 5$ required; independent observations; raw counts not ratios; multiple testing correction (Bonferroni); non-directional; Mendel's data paradox. Power: 1 - β; increases with n, effect size; decreases with stringent α; power analysis determines minimum n.

Answers to Pilot Questions [Series 3]

Part 3.1

1. Meiosis and fertilisation are random processes; any finite sample of offspring will deviate from the true underlying ratio simply due to random sampling variation; as sample size increases, the observed ratio approaches the theoretical ratio (Law of Large Numbers), but some deviation always remains.

2. H_0: the observed data are consistent with a 3:1 ratio (any deviation is due to random chance). H_1: the observed data do not fit a 3:1 ratio; the deviation is too large to be explained by chance alone.
3. A p-value is the probability of obtaining a chi-square value as large as or larger than the observed value by chance, assuming H_0 is true. A p-value of 0.03 means there is only a 3 percent probability of obtaining the observed deviation (or larger) if the 3:1 ratio is correct. Since $0.03 < \alpha = 0.05$, H_0 is rejected; the data do not fit the 3:1 ratio.
4. Type I error (false positive): rejecting H_0 when it is actually true; probability equals $\alpha = 0.05$ (one in 20 correct H_0s will be wrongly rejected). Type II error (false negative): failing to reject H_0 when it is actually false; probability equals β; decreases with larger sample size.
5. $\alpha = 0.05$ means there is a 5 percent probability of rejecting H_0 when it is true (one false positive in 20 tests on average); it is the conventional balance between being too lenient (high Type I error rate with larger α) and too stringent (low power with smaller α); it is widely accepted in biological research as a reasonable threshold.

Part 3.2

1. $\chi^2 = \Sigma [(O-E)^2/E]$; O = observed count in a category; E = expected count = total × expected proportion; the sum is over all phenotypic classes; larger χ^2 indicates greater discrepancy between observed and expected.
2. n = 180; E(dominant) = 180 × 3/4 = 135; E(recessive) = 180 × 1/4 = 45. $\chi^2 = [(140-135)^2/135] + [(40-45)^2/45] = [25/135] + [25/45] = 0.185 + 0.556 = 0.741$. df = 1; critical value at $\alpha = 0.05$: 3.841. Since $0.741 < 3.841$, do not reject H_0; the data are consistent with a 3:1 ratio.
3. df = number of phenotypic classes - 1. For a 9:3:3:1 ratio: four phenotypic classes; df = 4 - 1 = 3. One degree of freedom is lost because the counts must sum to the known total n, leaving three free to vary.
4. $\chi^2 = 5.2$; df = 1; critical value at $\alpha = 0.05$: 3.841. Since $5.2 > 3.841$, REJECT H_0. Conclusion: the data do not fit the 3:1 ratio ($p < 0.05$; the deviation is statistically significant and is unlikely to be due to chance alone).
5. A very small χ^2 ($p > 0.99$) means the observed counts match the expected counts almost perfectly, more closely than would normally occur by chance in a random process. This is suspicious because random sampling should produce some deviation. It may indicate data fabrication, selective data reporting, or unconscious bias in recording ambiguous phenotypes (as Ronald Fisher suggested regarding Mendel's combined data in 1936).

Part 3.3

1. Goodness-of-fit test: H_0 states that data fit a pre-specified theoretical ratio (e.g. 3:1); expected values = n × expected proportion from the theoretical model. Test of independence: H_0 states that two categorical variables are independent; expected values = (row total × column total) / grand total from the observed marginal totals; no pre-specified ratio is used.
2. Contingency table: affected (G=70, A=30, total=100); unaffected (G=50, A=50, total=100); column totals (G=120, A=80); grand total=200. E(G affected) = 100 × 120/200 = 60; E(A affected) = 100 × 80/200 = 40; E(G unaffected) = 100 × 120/200 = 60; E(A unaffected) = 100 × 80/200 = 40.
3. $\chi^2 = [(70-60)^2/60] + [(30-40)^2/40] + [(50-60)^2/60] + [(50-40)^2/40] = [100/60] + [100/40] + [100/60] + [100/40] = 1.667 + 2.500 + 1.667 + 2.500 = 8.333$. df = (2-1)(2-1) = 1; critical value at $\alpha = 0.05$: 3.841. Since 8.333 > 3.841, REJECT H_0; the G allele is significantly associated with the disease.
4. Yates' continuity correction [$\chi^2 = \Sigma [(|O-E|-0.5)^2/E]$] should be applied to 2 x 2 contingency tables when any expected value falls between 5 and 10; it makes the chi-square approximation more accurate by correcting for the discreteness of count data. When expected values fall below 5, Fisher's exact test is preferred over any chi-square approximation.
5. n = 200; E = 200/4 = 50 for each class. $\chi^2 = [(60-50)^2/50] + [(40-50)^2/50] + [(55-50)^2/50] + [(45-50)^2/50] = [100/50] + [100/50] + [25/50] + [25/50] = 2.00 + 2.00 + 0.50 + 0.50 = 5.00$. df = 4 - 1 = 3; critical value at $\alpha = 0.05$: 7.815. Since 5.00 < 7.815, do not reject H_0; the data are consistent with independent assortment (1:1:1:1 ratio).

Part 3.4

1. G-test formula: $G = 2 \Sigma [O \times \ln(O/E)]$; sum over all categories; ln = natural logarithm. One advantage over Pearson chi-square: G values from independent tests are additive (the G values from two separate tests on independent data can be summed and the combined G value tested against the chi-square distribution with the combined degrees of freedom), making it easier to combine results across experiments.
2. n = 200; E(tall) = 200 × 3/4 = 150; E(short) = 200 × 1/4 = 50. $G = 2 \times [140 \times \ln(140/150) + 60 \times \ln(60/50)] = 2 \times [140 \times \ln(0.9333) + 60 \times \ln(1.200)] = 2 \times [140 \times (-0.06899) + 60 \times (0.18232)] = 2 \times [(-9.659) + (10.939)] = 2 \times 1.280 = 2.560$. df = 1; critical value at $\alpha = 0.05$: 3.841. Since 2.560 < 3.841, do not reject H_0; the data are consistent with a 3:1 ratio.
3. Poisson distribution: describes the probability of k rare, independent events in a fixed interval when the mean rate λ is known; $P(X=k) = (e^{-\lambda} \times \lambda^k)/k!$; mean = variance = λ. Two genetic applications: spontaneous mutation rate estimation (probability of k new mutations at a locus

per gamete when mean rate λ is known); number of radiation-induced chromosome breaks per cell at low doses follows a Poisson distribution.

4. Four limitations: all expected values must be ≥ 5 (otherwise chi-square approximation is inaccurate); observations must be independent (data from related individuals violate this); the test must use raw counts (not ratios or percentages); the chi-square test is non-directional and does not identify which specific class is in excess or deficit.
5. Power = 1 - β = the probability of correctly detecting a true deviation from the expected ratio. Larger sample size increases power because more offspring means the observed ratio is closer to the true underlying ratio, making even small deviations detectable; with too few offspring, a real 2.5:1 ratio may not be distinguishable from a 3:1 hypothesis and H_0 will not be rejected even though it is false (Type II error). Power analysis before an experiment determines the minimum n needed to detect a given effect size with a desired power (e.g. 80 percent) at α = 0.05.

References and Attributions [Series 3]

Textual References (Books and External Sources)

- Brooker, R. J. (2015). Genetics: Analysis and principles (5th ed.). McGraw-Hill, New York.
- Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., & Doebley, J. (2015). Introduction to genetic analysis (11th ed.). W. H. Freeman, New York.
- Sokal, R. R., & Rohlf, F. J. (1995). Biometry (3rd ed.). W. H. Freeman, New York.
- Fisher, R. A. (1936). Has Mendel's work been rediscovered? Annals of Science, 1(2), 115–137.

Figures, Plates, Tables, and Boxes

- Box 3.1: Chi-square goodness-of-fit calculations for Mendel's monohybrid and dihybrid data. Attribution: AI generated.
- Figure 3.1: Chi-square probability distributions for df = 1, 3, and 5. Attribution: AI generated. Suggested OER search string: "chi-square distribution degrees of freedom critical values genetics OER".

Course code: BIO 201

Course title: Genetics I

Course credit load: 2 Units

Series 4: Polygenic and Quantitative Inheritance

- **Part 4.1: Introduction to Quantitative Genetics**
  - Sub Part 4.1.1: Distinction between qualitative and quantitative traits
  - Sub Part 4.1.2: The polygenic model of inheritance
  - Sub Part 4.1.3: Continuous variation and the normal distribution
- **Part 4.2: Components of Phenotypic Variance**
  - Sub Part 4.2.1: Partitioning phenotypic variance into genetic and environmental components
  - Sub Part 4.2.2: Types of genetic variance: additive, dominance, and epistatic
  - Sub Part 4.2.3: Genotype-by-environment interaction
- **Part 4.3: Heritability**
  - Sub Part 4.3.1: Broad-sense and narrow-sense heritability
  - Sub Part 4.3.2: Methods of estimating heritability
  - Sub Part 4.3.3: Interpretation and misinterpretation of heritability
- **Part 4.4: Selection and Response to Selection**
  - Sub Part 4.4.1: Selection differential and response to selection
  - Sub Part 4.4.2: The breeder's equation and its applications

- Sub Part 4.4.3: QTL mapping and molecular approaches to quantitative traits

Introduction

Most traits of agricultural and medical importance do not show the simple, discrete phenotypic classes characteristic of Mendelian inheritance. Instead, they vary continuously across a range of values: crop yield, body height, milk production, grain protein content, blood pressure, intelligence, and susceptibility to complex diseases are all examples of traits that cannot be assigned to a few discrete categories. These traits are called quantitative traits (or complex traits) and their inheritance is studied by quantitative genetics, a branch of genetics that combines principles from Mendelian genetics, statistics, and evolutionary biology.

This Series introduces the polygenic model of inheritance, which explains how the combined action of many genes at different loci produces continuous variation; the partitioning of phenotypic variation into genetic and environmental components; the concept of heritability and its estimation using various experimental methods; genotype-by-environment interaction; and the principles of artificial selection, the breeder's equation, and quantitative trait locus (QTL) mapping. These topics are foundational for plant and animal breeding, human complex disease genetics, and evolutionary genetics.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 4.1** distinguish between qualitative and quantitative traits, explain the polygenic model of inheritance with examples, and describe continuous variation and the normal distribution (Linked to Part 4.1),
- **SLO 4.2** partition total phenotypic variance into genetic and environmental components, distinguish between additive, dominance, and epistatic genetic variance, and explain genotype-by-environment interaction (Linked to Part 4.2),

- **SLO 4.3** define broad-sense and narrow-sense heritability, describe methods of estimating heritability, and correctly interpret heritability values (Linked to Part 4.3), and
- **SLO 4.4** apply the breeder's equation to predict response to selection, explain the factors that affect the rate of response to selection, and describe QTL mapping as a molecular approach to understanding quantitative traits (Linked to Part 4.4).

4.1 Introduction to Quantitative Genetics

4.1.1 Distinction between qualitative and quantitative traits

Traits can be classified into two broad categories based on the nature of their variation. Qualitative (discontinuous) traits: show a small number of discrete, non-overlapping phenotypic classes; individuals can be assigned unambiguously to one class or another; the classes are typically determined by alleles at a single locus (or a small number of loci) with simple dominance relationships; examples include: ABO blood groups (four discrete classes: A, B, AB, O); pea seed colour as studied by Mendel (yellow vs green); sickle-cell disease (normal vs affected); coat colour in Labrador retrievers (black, chocolate, yellow); Huntington's disease (affected vs unaffected). Qualitative traits do not require statistical description because all individuals fall into one of a few well-defined categories. Quantitative (continuous) traits: show a continuous range of phenotypic values; the variation is graded and there are no clear-cut boundaries between phenotypic classes; individuals cannot be assigned to discrete classes; examples include: human height (ranges continuously from approximately 130 cm to 220 cm in adults); body weight; crop grain yield; milk production; intelligence quotient (IQ); blood pressure; number of eggs laid per year in poultry.

Reasons why quantitative traits show continuous variation: (1) Polygenic determination: the trait is influenced by alleles at many loci (polygenes), each contributing a small, roughly additive amount to the phenotype; the combined effect of many loci with small individual effects produces a continuous distribution of phenotypes. (2) Environmental influence: the environment modifies the expression of each genotype, adding additional variability to the phenotypic distribution; this environmental component of variation smooths out the small between-genotype differences and produces a continuous distribution. Continuous traits that are also substantially

influenced by the environment tend to show even more overlap between genotypic classes than continuous traits with lower environmental variance. The key practical consequence of quantitative traits: they cannot be analysed using Punnett squares or simple Mendelian ratios; they require statistical analysis of populations (means, variances, correlations) rather than analysis of individual genotypes; the tools of quantitative genetics (variance partitioning, heritability analysis, selection response equations) are required.

Fill in the blank: Quantitative traits show continuous variation because they are _____ (controlled by alleles at many loci, each contributing a small additive effect) and are also influenced by the environment; they cannot be analysed by Mendelian ratios and require statistical methods applied to populations.

4.1.2 The polygenic model of inheritance

The polygenic model of inheritance proposes that quantitative traits are determined by the combined additive action of alleles at many loci (polygenes), each contributing equally and in the same direction to the phenotype. The model was first formally proposed by Swedish geneticist Herman Nilsson-Ehle in 1908, based on his experiments crossing wheat varieties with different grain colours. Key assumptions of the basic polygenic model: (1) The trait is controlled by n loci, each with two alleles (capital letter = contributing allele; lowercase = non-contributing allele). (2) Each contributing allele adds an equal amount to the phenotype. (3) The loci contribute additively (no dominance; no epistasis at the basic model level). (4) The loci are unlinked (assort independently). Nilsson-Ehle's wheat grain colour experiment: red-grained (AABB) x white-grained (aabb) parents; F1: AaBb (intermediate red; two contributing alleles); F2 from AaBb x AaBb: distribution of contributing alleles from 0 to 4; number of phenotypic classes = 2n + 1 = 2(2) + 1 = 5 (for 2 loci); F2 ratio: AABB (1/16, darkest red); AABb or AaBB (4/16, dark red); AAbb or AaBb or aaBB (6/16, intermediate red); Aabb or aaBb (4/16, light red); aabb (1/16, white); this 1:4:6:4:1 ratio corresponds to the binomial distribution of contributing alleles.

Generalising the polygenic model to n loci: number of phenotypic classes in F2 = 2n + 1; proportion of F2 progeny in the extreme parental classes = $(1/4)^n$ each; as n increases, the number of phenotypic classes increases, the proportion in extreme classes decreases, and the

distribution increasingly resembles a normal (Gaussian) distribution; for large n (above about 5 to 6 loci), the discrete classes become so numerous and their frequency differences so small that the phenotypic distribution appears continuous. This is why quantitative traits, controlled by large numbers of loci, show continuous variation in populations. Effect of environment: even if only a few loci controlled a trait, environmental variation would cause the discrete phenotypic classes to overlap, producing what appears to be a continuous distribution; the combination of polygenic determination and environmental modification is what produces the smooth, bell-shaped distributions observed for most quantitative traits in natural and agricultural populations.

Fill in the blank: Nilsson-Ehle's two-locus model for wheat grain colour predicts _____ phenotypic classes in the F2 ($2n + 1 = 5$ for $n = 2$ loci) in a 1:4:6:4:1 ratio; as the number of polygene loci increases, the distribution increasingly approximates the normal curve and the trait appears continuous.

4.1.3 Continuous variation and the normal distribution

A normal (Gaussian) distribution is a symmetric, bell-shaped probability distribution completely described by two parameters: the mean (μ; the central tendency, the arithmetic average of all values in the population) and the variance (σ^2; the average squared deviation from the mean, which measures the spread of the distribution). The standard deviation ($\sigma = \sqrt{\sigma^2}$) is the square root of the variance and is expressed in the same units as the trait. Properties of the normal distribution relevant to quantitative genetics: (1) Symmetry: the distribution is symmetric about the mean; the mean = median = mode. (2) 68-95-99.7 rule: approximately 68 percent of observations fall within one standard deviation of the mean ($\mu \pm \sigma$); approximately 95 percent fall within two standard deviations ($\mu \pm 2\sigma$); approximately 99.7 percent fall within three standard deviations ($\mu \pm 3\sigma$). (3) Total area under the curve = 1 (all probabilities sum to 1).

Why quantitative traits in populations often follow a normal distribution: the Central Limit Theorem (CLT) states that the sum (or average) of a large number of independent random variables tends toward a normal distribution, regardless of the individual distributions; because quantitative traits are affected by many independently assorting genes and many independent environmental factors, the sum of their effects tends toward normality by the CLT. Statistical description of quantitative traits in populations: phenotypic mean ($\dot{x}$; sample mean = $\Sigma x_i/n$);

phenotypic variance ($V_p = s^2 = \Sigma(x_i - \dot{x})^2 / (n-1)$); phenotypic standard deviation ($s = \sqrt{V_p}$); coefficient of variation ($CV = s/\dot{x} \times 100$ percent; allows comparison of variability across traits with different means). The normal distribution is the basis for many statistical tests used in quantitative genetics (t-test, ANOVA, regression) and for the quantitative genetic models of selection response covered in Part 4.4.

Fill in the blank: The normal distribution is characterised by two parameters: the _____ (μ; central tendency) and the variance (σ^2; spread); approximately 95 percent of observations fall within two standard deviations of the mean; the Central Limit Theorem explains why quantitative traits controlled by many loci tend toward normality.

Pilot questions 4.1

1. Distinguish between qualitative and quantitative traits and give two examples of each.
2. Explain the polygenic model of inheritance and describe Nilsson-Ehle's wheat grain colour experiment.
3. For a trait controlled by three independently assorting polygene loci (n = 3), how many phenotypic classes are expected in the F2? What proportion of F2 offspring show the parental extreme phenotypes?
4. Define the mean and variance of a quantitative trait and explain the 68-95-99.7 rule for the normal distribution.
5. Explain why quantitative traits show continuous rather than discontinuous variation even if controlled by a finite number of polygene loci.

4.2 Components of Phenotypic Variance

4.2.1 Partitioning phenotypic variance into genetic and environmental components

The total phenotypic variance (V_p) observed in a population for a quantitative trait can be partitioned into components attributable to genetic and environmental sources. This partitioning is the foundation of quantitative genetics and heritability analysis. The fundamental equation: $V_p = V^G + V^E + V^{GE}$ (or $V_p = V^G + V^E$ when genotype-by-environment interaction is negligible or not estimated separately), where V_p = total phenotypic variance; V^G = genetic variance (due to differences in genotype among individuals); V^E = environmental variance (due to differences in

the environments experienced by individuals); V^{GE} = genotype-by-environment interaction variance (due to different genotypes responding differently to different environments). Genetic variance (V^G): the component of phenotypic variance that is due to differences in genotype among individuals in the population; V^G is further partitioned into additive genetic variance (V_a), dominance genetic variance (V^D), and epistatic (interaction) genetic variance (V^I): $V^G = V_a + V^D + V^I$. Environmental variance (V^E): the component of phenotypic variance due to non-genetic factors (nutrition, temperature, light, disease, measurement error, random developmental noise); V^E includes both 'macro-environmental' effects (between-environment differences that are systematic and predictable) and 'micro-environmental' effects (random developmental and measurement variation experienced even by genetically identical individuals).

Estimating V^E: if a pure-breeding line or a population of genetically identical clones is grown in a range of environments, all phenotypic variance observed within that line is attributable to environmental effects (since all individuals have the same genotype); V^E estimated in this way can be subtracted from V_p measured in a genetically variable population to estimate V^G. Covariance between genotype and environment: in natural populations, genotype and environment may be positively correlated (e.g. larger, healthier genotypes may be placed in more favourable environments by parents in some agricultural systems); this genotype-environment covariance (Cov^{GE}) is sometimes added to the equation: $V_p = V^G + V^E + V^{GE} + 2Cov^{GE}$; in designed experiments with random assignment of genotypes to environments, $Cov^{GE} = 0$. Importance of variance partitioning: it allows breeders to assess how much of the phenotypic variation in a crop or livestock population is under genetic control (and therefore can be altered by selection) versus environmental (and cannot be changed by genetic selection).

Fill in the blank: Total phenotypic variance $V_p = V^G + V^E$ (+ V^{GE} when interaction is estimated); V^G is the component due to _____ differences among individuals; V^E is due to non-genetic environmental factors; partitioning these components allows breeders to determine how much phenotypic variation can be altered by genetic selection.

4.2.2 Types of genetic variance: additive, dominance, and epistatic

Genetic variance (V^G) can be further partitioned into three components based on the type of gene action responsible: (1) Additive genetic variance (V_a): the component of genetic variance due to

the average effects of alleles at individual loci, considered independently; it reflects the degree to which parents transmit their phenotypes to their offspring; it is the component of genetic variance that responds to selection (because selection acts on phenotypes and the additive effects are what parents reliably pass to offspring); high V_a means that offspring resemble their parents closely for the trait; narrow-sense heritability (h^2) is defined as V_a/V_p (see Part 4.3). Additive gene action: each additional contributing allele at a locus adds the same increment to the phenotype (no dominance; the heterozygote is exactly intermediate between the two homozygotes). Example: if AA = 10 cm, Aa = 8 cm, aa = 6 cm (no dominance; linear scale), gene A acts additively. (2) Dominance genetic variance (V^D): the component due to the interaction between alleles at the same locus (within-locus interaction; dominance, recessiveness, overdominance); it arises when the heterozygote phenotype deviates from the midpoint between the two homozygotes. Example: if AA = 10 cm, Aa = 9.5 cm, aa = 6 cm (Aa closer to AA than to aa), partial dominance of A is present; the deviation of the heterozygote from the midpoint contributes to V^D. V^D is not transmitted from parents to offspring (offspring do not inherit genotypic combinations, only alleles) and therefore does not contribute to narrow-sense heritability.

(3) Epistatic (interaction) genetic variance (V^I; also written V^I or V^I): the component due to interactions between alleles at different loci (between-locus interactions; epistasis); includes digenic interactions (between two loci), trigenic interactions, and higher-order interactions; V^I is generally estimated as $V^G - V_a - V^D$ (the residual after subtracting additive and dominance components); V^I is complex, not easily estimated, and generally smaller than V_a for most agricultural traits. The relative magnitudes of V_a, V^D, and V^I have important implications for breeding strategy: if V_a is large (additive gene action predominates), individual selection and mass selection programmes are effective; if V^D is large (dominance is important), selection for specific combining ability, hybrid breeding programmes, and heterosis (hybrid vigour) become more valuable strategies; if V^I is large, epistatic effects are important and selection may not respond predictably.

Fill in the blank: Additive genetic variance (V_a) is the component of genetic variance that responds to _____ because it reflects the effects of alleles that parents reliably transmit to

offspring; narrow-sense heritability $h^2 = V_a/V_p$; dominance variance (V^D) does not contribute to h^2 because heterozygous genotypes are not transmitted intact from parent to offspring.

4.2.3 Genotype-by-environment interaction

Genotype-by-environment interaction (G×E interaction) occurs when different genotypes respond differently to changes in environment; in other words, the rank order of genotypes for a trait changes (or the magnitude of phenotypic differences between genotypes changes) across different environments. G×E interaction is distinct from genotype-environment correlation (where certain genotypes consistently occur in certain environments): G×E describes differential sensitivity of genotypes to environmental change. Types of G×E interaction: (1) Scale interaction (non-crossover interaction): the rank order of genotypes remains the same across environments, but the magnitude of differences changes; genotype A always outperforms genotype B, but the gap is larger in one environment than another; this type is less problematic for plant breeding because the best genotype in one environment is still the best in another. (2) Crossover interaction (rank-change interaction): the rank order of genotypes changes across environments; genotype A outperforms B in environment 1 but B outperforms A in environment 2; this type is the most problematic for breeding because it means no single genotype is 'best' across all environments.

Detecting G×E interaction: G×E is detected and measured using multi-environment trials (METs) in which a set of genotypes (varieties, lines, hybrids) is evaluated in multiple locations and/or years; analysis of variance (ANOVA) partitions variation into genotype (G), environment (E), and G×E interaction effects; a significant G×E interaction term indicates that genotypes respond differently to environments. Norm of reaction (reaction norm): the graph showing how a genotype's phenotype changes across different environments; if all genotypes show parallel norms of reaction (slopes are equal; lines do not cross), there is no G×E interaction; if the norms of reaction cross (genotype A has a steeper slope than genotype B; their ranks swap across environments), crossover G×E interaction is present. Agricultural importance of G×E interaction: G×E makes it necessary to evaluate new crop varieties across multiple locations and years before release; breeders must decide whether to develop broadly adapted varieties (good across many environments; genotypes with small G×E) or specifically adapted varieties (best in

particular environments; narrow adaptation). Stability analysis (Finlay-Wilkinson regression; Eberhart-Russell stability) quantifies the stability of a genotype across environments.

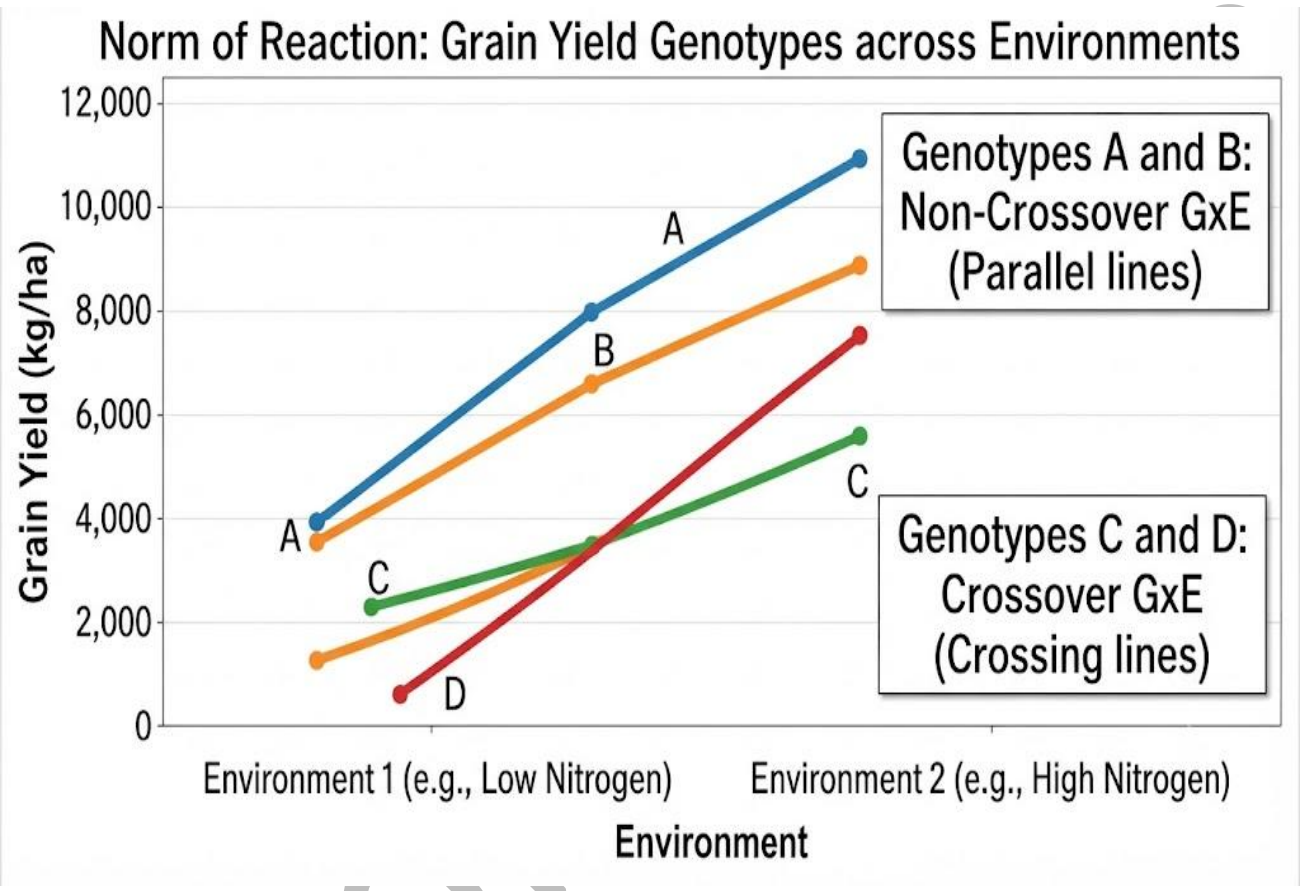

Figure 4.1: Norms of reaction for four genotypes across two environments, illustrating non-crossover (genotypes A and B) and crossover (genotypes C and D) genotype-by-environment interaction

Pilot questions 4.2

1. State the equation $V_p = V^G + V^E$ and explain what each term represents.
2. Distinguish between additive genetic variance (V_a) and dominance genetic variance (V^D) in terms of the type of gene action involved and the contribution of each to narrow-sense heritability.
3. Explain why V^D does not contribute to the response to selection.
4. Define G×E interaction and distinguish between scale (non-crossover) and crossover interactions.

5. Explain why crossover G×E interaction is problematic for plant breeders selecting new crop varieties.

4.3 Heritability

4.3.1 Broad-sense and narrow-sense heritability

Heritability is a statistical measure of the degree to which phenotypic variation in a population is attributable to genetic variation. It is one of the most important concepts in quantitative genetics. There are two forms: Broad-sense heritability (H^2): the proportion of total phenotypic variance attributable to total genetic variance: $H^2 = V^G / V_p = (V_a + V^D + V^I) / (V^G + V^E)$. H^2 measures the total extent to which phenotypic differences are genetically determined; it ranges from 0 (all phenotypic variation is environmental) to 1 (all phenotypic variation is genetic). Narrow-sense heritability (h^2): the proportion of total phenotypic variance attributable to additive genetic variance only: $h^2 = V_a / V_p$. h^2 is more important for plant and animal breeding because it is the additive component of genetic variance that is transmitted from parents to offspring and that determines the response to selection; h^2 is always $\leq H^2$; for traits where dominance and epistasis are small relative to additive variance, $h^2 \approx H^2$.

Illustrative values of heritability: high h^2 (above 0.60): bristle number in Drosophila ($h^2 \approx 0.50$ to 0.60); body weight in cattle ($h^2 \approx 0.45$ to 0.65); plant height in maize ($h^2 \approx 0.70$ to 0.90); human height ($h^2 \approx 0.80$ in well-nourished populations); egg weight in poultry ($h^2 \approx 0.55$ to 0.75). Moderate h^2 (0.30 to 0.60): milk production in cattle ($h^2 \approx 0.25$ to 0.40); litter size in pigs ($h^2 \approx$ 0.10 to 0.15); wool weight in sheep ($h^2 \approx 0.40$ to 0.60). Low h^2 (below 0.30): reproductive fitness traits (fertility, litter size, clutch size) generally have low heritability because strong natural selection has depleted additive genetic variance; disease resistance often has moderate to low heritability. General pattern: fitness-related traits tend to have lower h^2 (because selection has already reduced V_a for fitness); morphological and physiological traits tend to have higher h^2. Important caveats: heritability is a population-specific statistic (it depends on the genetic composition and environmental range of the specific population studied); the same trait can have very different h^2 values in different populations or in different environments.

Fill in the blank: Narrow-sense heritability $h^2 = V_a / V_p$ is more useful for plant and animal breeding than broad-sense heritability $H^2 = V^G / V_p$ because it measures the _____ genetic variance that is transmitted from parents to offspring and that determines response to selection; h^2 is always equal to or less than H^2.

4.3.2 Methods of estimating heritability

Several experimental approaches are used to estimate heritability. (1) Parent-offspring regression: regress the mean phenotype of offspring on the mean phenotype of their parents (mid-parent value); the regression slope (b) of offspring on mid-parent equals h^2 under standard quantitative genetic assumptions; alternatively, regress offspring on one parent: $b = 0.5\ h^2$ (because offspring share only half their genes with one parent). This method is widely used for estimating h^2 in livestock and human populations where parent-offspring data are available. Limitations: parent and offspring may share environments (causing overestimation of h^2 if environment is positively correlated across generations); assortative mating (non-random mating by phenotype) can bias estimates. (2) Twin studies: compare the phenotypic similarity of monozygotic (MZ; identical) twins with that of dizygotic (DZ; fraternal) twins; MZ twins share all their DNA; DZ twins share on average 50 percent; $h^2 \approx 2(r^{MR} - r^{DR})$, where r^{MR} is the MZ intraclass correlation coefficient and r^{DR} is the DZ intraclass correlation coefficient; $H^2 \approx r^{MR}$ (the MZ correlation approximates broad-sense heritability under the assumption that MZ twins share environments as similarly as DZ twins). Widely used in human genetics; subject to the equal environments assumption (MZ and DZ twin pairs experience equally similar environments, which may not always hold).

(3) Full-sib and half-sib analysis: uses family structure in designed experiments; half-sibs (sharing one parent) share on average 1/4 of their genes; the among-half-sib family variance component (σ^{2A}) = $V_a/4$; $h^2 = 4\sigma^{2A} / V_p$; this approach is widely used in livestock breeding (groups of half-sib offspring from the same sire) and in designed plant breeding experiments. Full-sibs share on average 1/2 of their genes; the full-sib variance component includes both $V_a/2$ and $V^D/4$. (4) Response to selection (realised heritability): after a generation of artificial selection, $h^2 = R / S$, where R = response to selection (change in mean between selected offspring and parental population mean) and S = selection differential (mean of selected parents minus

mean of parental population); this is called realised heritability and is estimated after the fact from actual selection experiments. (5) Molecular marker-based methods (GCTA; GREML): genome-wide association study data and whole-genome sequencing can be used to estimate the genetic relatedness between all pairs of unrelated individuals in a population; the covariance in phenotype between related individuals is compared to their genetic relatedness estimated from SNP markers; this 'SNP heritability' estimates the proportion of variance explained by common SNPs.

Fill in the blank: In the parent-offspring regression method, the slope b of offspring on _____ value equals h^2; in the twin study method, $h^2 \approx 2(r^{MR} - r^{DR})$; in the half-sib method, $h^2 = 4\sigma^{2A} / V_p$; and in selection experiments, realised heritability = R / S.

4.3.3 Interpretation and misinterpretation of heritability

Heritability is one of the most frequently misunderstood concepts in genetics. Common misinterpretations and the correct understanding: (1) 'Heritability tells us how much of a trait is caused by genes': INCORRECT. Heritability tells us what fraction of the phenotypic VARIATION in a POPULATION is due to genetic variation; a trait can be 100 percent genetically determined in terms of the mechanism (e.g. having two eyes is determined by genes) yet have a heritability of 0 if all individuals have the same genotype and all variation is environmental. (2) 'A high heritability means the trait cannot be changed by the environment': INCORRECT. PKU (phenylketonuria) is a highly heritable metabolic disorder whose most severe phenotypic consequence (intellectual disability) is completely preventable by dietary management; human height is highly heritable ($h^2 \approx 0.80$) yet has increased dramatically with improved nutrition over the past 150 years. (3) 'Heritability of a trait in one population tells us about genetic differences between populations': INCORRECT. Even if a trait has high h^2 within each of two groups, the difference in mean trait value between the groups could be entirely environmental.

Classic illustration of this misinterpretation (Richard Lewontin's plant analogy, 1972): imagine two batches of genetically diverse seeds (representing populations A and B); grow batch A in rich soil (all in the same rich soil) and batch B in poor soil (all in the same poor soil); within each batch, all variation in plant height is genetic ($h^2 = 1.0$ in both batches, since environments are

uniform); yet the difference in mean height between batch A (taller, rich soil) and batch B (shorter, poor soil) is entirely environmental (due to soil quality, not genetic differences between the batches); heritability within groups tells us nothing about the cause of differences between groups. Correct interpretation of heritability: (1) h^2 is a property of a specific population in a specific environment at a specific time; it can change if the population's genetic composition changes (e.g. after selection) or if the range of environments changes. (2) h^2 predicts the resemblance between parents and offspring for the trait; parents with high phenotypic values will on average have offspring with values above the population mean, by a factor h^2 times the parental deviation from the mean. (3) h^2 predicts the rate of response to artificial selection (see Part 4.4).

Fill in the blank: Heritability estimates the proportion of phenotypic _____ in a population attributable to genetic variation; it does NOT indicate that the trait cannot be changed by the environment, and it does NOT explain differences between populations because between-group differences may be entirely environmental even when within-group heritability is high.

Pilot questions 4.3

1. Define broad-sense heritability (H^2) and narrow-sense heritability (h^2) and state the formula for each.
2. Explain the parent-offspring regression method of estimating heritability and state what the regression slope estimates.
3. In a twin study, the MZ intraclass correlation for a trait is 0.80 and the DZ intraclass correlation is 0.50. Estimate h^2.
4. Explain why $h^2 = 0.80$ for human height does NOT mean that height cannot be changed by nutrition.
5. Explain Lewontin's plant analogy and what it demonstrates about the interpretation of heritability.

4.4 Selection and Response to Selection

4.4.1 Selection differential and response to selection

Artificial selection (selective breeding) is the deliberate choice of individuals with superior phenotypes as parents of the next generation, with the goal of shifting the population mean for a trait in a desired direction. The effectiveness of selection depends on both the heritability of the trait and the intensity of selection applied. Key terms in selection theory: Selection differential (S): the difference between the mean phenotype of the selected parents and the mean phenotype of the population before selection: $S = \dot{x}_s - \dot{x}$, where $\dot{x}_s$ = mean of selected parents and $\dot{x}$ = mean of the whole population; S measures how superior the selected parents are relative to the population average; a large S means the selected parents are far above average. Response to selection (R): the difference between the mean phenotype of the offspring of selected parents and the mean phenotype of the parental population before selection: $R = \dot{x}_o^{ff} - \dot{x}$, where $\dot{x}_o^{ff}$ = mean of offspring of selected parents; R is what is actually achieved (the change in the population mean from one generation to the next). Realised heritability: estimated from selection experiments as $h^2 = R/S$ (see Part 4.3).

Factors affecting S: (1) Truncation selection threshold: when a fixed proportion of the population (the top p percent) is selected, $S = i \times \sigma_p$, where i is the standardised selection differential (intensity of selection; the value of i depends on the proportion p selected; i increases as p decreases, i.e. more intense selection; for p = 0.10, $i \approx 1.755$; for p = 0.20, $i \approx 1.400$; for p = 0.50, $i \approx 0.798$; for p = 0.01, $i \approx 2.665$) and σ_p is the phenotypic standard deviation. (2) Phenotypic variance: a population with more phenotypic variance (larger σ_p) will have a larger S for a given proportion selected. Selection intensity (i): the standardised form of S (measured in units of σ_p): $i = S / \sigma_p$. More intense selection (smaller proportion selected, larger i) increases S and therefore tends to produce a larger response, but leaves fewer parents and reduces effective population size (risks inbreeding).

Fill in the blank: The selection differential $S = \dot{x}_s - \dot{x}$ is the difference between the mean of selected _____ and the mean of the whole parental population; it measures how superior the selected parents are relative to the population; response to selection $R = h^2 \times S$ (the breeder's equation).

4.4.2 The breeder's equation and its applications

The breeder's equation is the fundamental equation relating response to selection, heritability, and selection differential: $R = h^2 \times S$, where R = response to selection (change in population mean per generation); h^2 = narrow-sense heritability; S = selection differential. An equivalent form using selection intensity: $R = h^2 \times i \times \sigma_p$, where i = selection intensity ($i = S/\sigma_p$). Or, using the genetic standard deviation: $R = h \times \sigma_a$ (since $h^2 = V_a/V_p$, $h = \sigma_a/\sigma_p$, and $R = h \times \sigma_p \times (S/\sigma_p) = h \times \sigma_a \times i/1$). The breeder's equation has several important implications: (1) If $h^2 = 0$, no selection response is possible regardless of S (all variance is environmental; selection has nothing to act on genetically). (2) If $h^2 = 1$, R = S exactly (the offspring mean exactly equals the selected parents' mean; all phenotypic variation is additive genetic). (3) Larger S (more intense selection) gives larger R for a given h^2. (4) Traits with higher h^2 respond faster to selection. Worked example: a dairy cattle population has mean milk yield = 6,000 L/lactation; phenotypic standard deviation σ_p = 800 L; breeders select the top 20 percent of cows as dams (selection intensity $i \approx 1.400$); h^2 = 0.30. $S = i \times \sigma_p = 1.400 \times 800 = 1{,}120$ L. $R = h^2 \times S = 0.30 \times 1{,}120 = 336$ L per generation. New population mean after one generation of selection = 6,000 + 336 = 6,336 L.

The response to selection per unit time (R/year): in plant and animal breeding, generation interval (L; the average age of parents when offspring are born) affects how rapidly genetic gain is achieved per year: genetic gain per year = $R / L = (i \times h \times \sigma_a) / L$. Strategies to maximise genetic gain per year: increase selection intensity (i) by selecting a smaller proportion as parents; increase h^2 (by reducing V^E through better environmental control); increase σ_a (by introducing new genetic variation); decrease L (by using shorter generation intervals, e.g. early selection, embryo transfer, genomic selection in livestock). Limits to selection response: as selection proceeds, (1) V_a decreases as favourable alleles increase in frequency toward fixation (the population becomes more homozygous for the selected alleles, reducing genetic variance); (2) inbreeding depression may reduce performance; (3) correlated responses in other traits may be undesirable (if two traits are genetically correlated, selection for one changes the other). Cumulative response to selection across multiple generations: if h^2 and S remain constant (not always realistic), cumulative R after t generations = $t \times R_1$ (linear response initially).

Fill in the blank: The breeder's equation $R = h^2 \times S$ predicts that response to selection equals _____ heritability multiplied by the selection differential; genetic gain per year = $(i \times h \times \sigma_a) / L$, where L is the generation interval; longer generation intervals reduce the annual rate of genetic improvement.

4.4.3 QTL mapping and molecular approaches to quantitative traits

Quantitative trait locus (QTL) mapping is a statistical technique that identifies regions of the genome (QTL) containing genes that contribute to variation in quantitative traits. The underlying principle: if a QTL (a chromosomal region containing one or more genes affecting the trait) is close to a molecular genetic marker (a polymorphic DNA sequence of known chromosomal location), then individuals with different marker genotypes will on average have different phenotypic values for the trait; the association between marker genotype and trait value reveals the presence and approximate chromosomal location of the QTL. QTL mapping procedure: (1) Develop a segregating population: cross two parents that differ substantially in the quantitative trait (P1 × P2) to produce an F1; self or intercross F1 to produce F2 (or backcross, recombinant inbred lines (RILs), doubled haploids); this segregating population contains individuals varying in genotype at many loci, including QTL. (2) Genotype all individuals at a dense set of molecular markers distributed throughout the genome (microsatellites in earlier studies; SNPs in current studies). (3) Measure the phenotype of each individual for the quantitative trait of interest. (4) Statistically test for association between each marker genotype (or marker interval) and the trait phenotype.

QTL mapping methods: (1) Single-marker analysis: for each marker, compare the mean phenotypic values of the marker genotypic classes (e.g. MM vs Mm vs mm) using ANOVA or t-test; a significant difference indicates that the marker is associated with a QTL; this is the simplest method but lacks precision and statistical power. (2) Interval mapping (Lander and Botstein 1989): uses pairs of flanking markers to estimate the position and effect of a QTL within marker intervals; produces a LOD score (logarithm of the odds ratio; LOD = $\log_{10}$ [probability that data fit a model with a QTL at the test position / probability that data fit a model without a QTL]); QTL are declared at positions where LOD exceeds a threshold (typically LOD ≥ 3.0 for inbred line crosses, corresponding to $p \approx 0.001$ after correction for multiple testing). (3)

Composite interval mapping (CIM): adds background marker cofactors to the interval mapping model to control for the effects of QTL elsewhere in the genome; improves resolution and reduces false positives. (4) Genome-wide association study (GWAS): rather than using a designed cross, tests for association between hundreds of thousands to millions of SNP markers and a trait in a large sample of unrelated individuals from a natural or breeding population; applicable to humans, livestock, and crops; identifies candidate regions but resolution is limited by linkage disequilibrium patterns. QTL information in breeding: once QTL are identified and validated, the associated molecular markers can be used in marker-assisted selection (MAS): breeders select individuals that carry the favourable QTL allele at the marker, without needing to phenotype all offspring; MAS is particularly valuable for traits that are expensive or difficult to measure (e.g. disease resistance after challenge; carcass traits that can only be measured after slaughter; drought tolerance).

Concept	Formula	Interpretation
Phenotypic Variance	$V_p = V_G + V_E + V_{GE}$	Total observed variation in a trait is due to genetic, environmental, and gene–environment interaction effects.
Genetic Variance Components	$V_G = V_A + V_D + V_I$	Genetic variance is partitioned into additive (VA), dominance (VD), and epistatic (VI) components.
Broad-Sense Heritability	$H^2 = V_G / V_p$	Proportion of total variance due to genetic factors; ranges from 0 to 1.
Narrow-Sense Heritability	$h^2 = V_A / V_p$	Proportion of variance due to additive genetic effects; ranges from 0 to 1; key for predicting response to selection.
Breeder's Equation	$R = h^2 \times S$	Response to selection (R) depends on narrow-sense heritability and selection differential.
Selection Differential	$S = \bar{x}_s - \bar{x}$ or $S = i \times \sigma_p$	Difference between mean of selected parents and population mean; measures selection intensity.
Genetic Gain per Year	$\Delta G = i \times h \times \sigma_A / L$	Annual genetic improvement depends on selection intensity (i), accuracy (h), additive SD (σ_A), and generation interval (L).
Realised Heritability	$h^2 = R / S$	Empirical estimate of heritability based on observed response and selection differential; ranges from 0 to 1.

Box 4.1: Summary of key quantitative genetics formulae and their interpretation

Pilot questions 4.4

1. Define the selection differential (S) and the response to selection (R) and state the breeder's equation.
2. A wheat breeding population has mean grain yield = 3,500 kg/ha; σ_p = 400 kg/ha; h^2 = 0.50; breeders select the top 10 percent as parents (i = 1.755). Calculate S and R.
3. Explain what happens to the response to selection over successive generations and why it tends to decrease.
4. Define QTL mapping and explain the principle underlying the association between a molecular marker and a quantitative trait.
5. Define marker-assisted selection (MAS) and state two situations in plant or animal breeding where MAS is particularly valuable.

Series Summary

This Series introduced the genetic basis of **quantitative (continuous) traits**, explaining how they differ from qualitative traits and why they require statistical rather than simple Mendelian analysis. You explored polygenic inheritance, the normal distribution, phenotypic and genetic variance, additive, dominance, and epistatic effects, as well as genotype by environment interactions. The Series then examined heritability, including broad-sense and narrow-sense heritability, methods of estimation, common misconceptions, and the breeder's equation for predicting response to selection. Finally, you were introduced to QTL mapping and marker-assisted selection as important tools in modern genetics and breeding.

By completing this Series, you should now be able to distinguish qualitative and quantitative traits, explain the genetic and environmental components of continuous variation, interpret heritability and predict selection response using appropriate genetic principles, and describe how QTL mapping and molecular markers support the improvement of economically important traits. These competencies collectively enable you to achieve all four **Series Learning Outcomes.**

Glossary of Terms

- **Additive genetic variance (V_a):** the component of genetic variance due to the average effects of individual alleles; determines response to selection and narrow-sense heritability; alleles act additively (heterozygote is exactly intermediate between homozygotes).
- **Breeder's equation:** $R = h^2 \times S$; the fundamental equation predicting response to selection (R) from narrow-sense heritability (h^2) and selection differential (S); equivalent form: $R = i \times h \times \sigma_a$.
- **Broad-sense heritability (H^2):** $H^2 = V^G/V_p$; the proportion of total phenotypic variance attributable to all genetic variance (additive + dominance + epistatic); includes effects not transmitted from parents to offspring; always ≥ narrow-sense heritability.
- **Dominance genetic variance (V^D):** the component of genetic variance due to within-locus allelic interactions (dominance, recessiveness, overdominance); the heterozygote deviates from the midpoint between homozygotes; contributes to H^2 but not h^2; not transmitted intact to offspring.

- **Epistatic (interaction) genetic variance (V^I):** the component of genetic variance due to between-locus interactions (epistasis); $V^I = V^G - V_a - V^D$; generally smaller than V_a for most agricultural traits.
- **Genotype-by-environment interaction (G×E):** the phenomenon in which different genotypes respond differently to changes in environment; the rank order of genotypes changes across environments (crossover interaction) or the magnitude of differences changes without rank change (scale interaction).
- **Marker-assisted selection (MAS):** the use of molecular markers linked to QTL to select individuals carrying favourable QTL alleles without requiring phenotypic measurement; particularly valuable for traits that are difficult, expensive, or destructive to measure.
- **Narrow-sense heritability (h^2):** $h^2 = V_a/V_p$; the proportion of phenotypic variance due to additive genetic variance; the key parameter determining response to artificial selection; estimated by parent-offspring regression, twin studies, half-sib analysis, or realised heritability.
- **Normal (Gaussian) distribution:** a symmetric, bell-shaped probability distribution completely described by its mean (μ) and variance (σ^2); approximately 95 percent of values fall within $\pm 2\sigma$ of the mean; explains the distribution of quantitative traits controlled by many loci.
- **Polygenic inheritance:** inheritance of a quantitative trait controlled by alleles at many loci (polygenes), each contributing a small, additive amount to the phenotype; produces continuous variation; first demonstrated by Nilsson-Ehle (1908) for wheat grain colour.
- **Quantitative trait:** a trait showing continuous variation in a population (not discrete classes); determined by many genes (polygenes) plus environmental effects; requires statistical analysis of population means and variances rather than individual genotype classification.
- **Quantitative trait locus (QTL):** a chromosomal region containing one or more genes that contribute to variation in a quantitative trait; identified by statistical association between molecular marker genotypes and trait phenotypes in segregating populations.
- **Realised heritability:** heritability estimated from an actual selection experiment: $h^2 = R/S$, where R is the observed response to selection and S is the selection differential applied; a direct measure of how much of the selection differential was transmitted to offspring.
- **Response to selection (R):** the difference between the mean phenotype of offspring of selected parents and the mean of the parental population before selection; $R = h^2 \times S$ (breeder's equation).

- **Selection differential (S):** the difference between the mean phenotype of selected parents and the mean of the population before selection; $S = \dot{x}_s - \dot{x} = i \times \sigma_p$; measures the superiority of selected parents.
- **Selection intensity (i):** the standardised selection differential in units of the phenotypic standard deviation: $i = S/\sigma_p$; increases as the proportion of the population selected decreases (more intense selection).
- **$V_p = V^G + V^E$:** the fundamental partitioning of total phenotypic variance into genetic and environmental components; $V_p = V_a + V^D + V^I + V^E$ (+ V^{GE}); the basis for heritability estimation.

Concept Check Questions [Series 4]

Objective Questions

1. Quantitative traits show continuous variation because they are _____ (controlled by many loci each with a small additive effect) and are also modified by the environment. (Linked to SLO 4.1)
2. For a polygenic trait controlled by two loci (n = 2), the number of phenotypic classes in the F2 = 2n + 1 = _____; the F2 phenotypic ratio is 1:4:6:4:1. (Linked to SLO 4.1)
3. Total phenotypic variance $V_p = V^G + V^E$; the component V_a is called _____ genetic variance and is the component that determines the response to artificial selection. (Linked to SLO 4.2)
4. Narrow-sense heritability $h^2 = V_a / V_p$; if $h^2 = 0$, no selection _____ is possible because all phenotypic variation is due to the environment and there is no additive genetic variation for selection to act on. (Linked to SLO 4.3)
5. The breeder's equation states that $R = h^2 \times S$, where S is the selection _____ (mean of selected parents minus population mean) and R is the response to selection. (Linked to SLO 4.4)

Short-Answer Questions

1. Distinguish between qualitative and quantitative traits with two examples of each. (Linked to SLO 4.1)
2. Explain the polygenic model of inheritance and predict the F2 phenotypic classes and ratio for a two-locus additive model. (Linked to SLO 4.1)
3. Distinguish between additive genetic variance (V_a) and dominance genetic variance (V^D) and explain why only V_a contributes to narrow-sense heritability. (Linked to SLO 4.2)
4. Define narrow-sense heritability and state two methods of estimating it. (Linked to SLO 4.3)
5. Apply the breeder's equation: a crop has $h^2 = 0.40$, $\sigma_p = 300$ kg/ha, and the top 10 percent are selected (i = 1.755). Calculate S and R. (Linked to SLO 4.4)

Long-Answer Questions

1. Distinguish qualitative from quantitative traits. Explain the polygenic model with Nilsson-Ehle's experiment. Describe continuous variation and the normal distribution. Explain variance partitioning ($V_p = V^G + V^E$), the types of genetic variance (V_a, V^D, V^I), and G×E interaction. (Linked to SLOs 4.1, 4.2)
2. Define broad-sense and narrow-sense heritability and describe methods of estimating them. Explain the misinterpretations of heritability (Lewontin's plant analogy; PKU). State the

breeder's equation and apply it to predict response to selection. Explain QTL mapping and marker-assisted selection. (Linked to SLOs 4.3, 4.4)

Answers to Concept Check Questions [Series 4]

Objective Questions

1. Polygenic.
2. 5.
3. Additive.
4. Response.
5. Differential.

Short-Answer Questions

1. Qualitative traits: discrete non-overlapping phenotypic classes; few loci; Mendelian analysis; examples: ABO blood groups (four discrete classes); sickle-cell disease (affected vs normal). Quantitative traits: continuous phenotypic variation; many loci (polygenic); statistical analysis of populations required; examples: human height (130 to 220 cm range); crop grain yield.
2. Polygenic model: a quantitative trait is controlled by alleles at many loci, each contributing a small, equal, additive amount; no dominance at the basic level. For n = 2 loci (AABB x aabb cross): F1 = AaBb (intermediate); F2 phenotypic classes = 2n + 1 = 5; F2 ratio = 1:4:6:4:1 (from 0 to 4 contributing alleles).
3. Additive genetic variance (V_a): due to the average effects of individual alleles; the heterozygote is exactly intermediate between homozygotes; transmitted from parents to offspring; determines response to selection; contributes to h^2. Dominance genetic variance (V^D): due to within-locus allelic interactions (dominance); the heterozygote deviates from the midpoint between homozygotes; not transmitted intact from parents to offspring because offspring inherit alleles, not genotypic combinations; therefore V^D does not contribute to $h^2 = V_a/V_p$.
4. Narrow-sense heritability $h^2 = V_a/V_p$; the proportion of phenotypic variance due to additive genetic variance; determines the response to selection. Two estimation methods: parent-offspring regression (regression slope b of offspring on mid-parent value = h^2); realised heritability from selection experiments ($h^2 = R/S$).
5. $S = i \times \sigma_p = 1.755 \times 300 = 526.5$ kg/ha. $R = h^2 \times S = 0.40 \times 526.5 = 210.6$ kg/ha per generation.

Long-Answer Questions

1. Qualitative traits: discrete classes; few loci; Mendelian analysis; examples: ABO blood groups, sickle-cell disease. Quantitative traits: continuous variation; polygenic;

environmental influence; statistical analysis; examples: height, yield, milk production. Polygenic model (Nilsson-Ehle 1908): n loci; each contributing allele adds equal increment; AABB x aabb → F1 AaBb (intermediate red); F2 classes = 2n+1; for n=2: 5 classes in 1:4:6:4:1 ratio; as n increases, distribution approaches normal. Normal distribution: characterised by μ (mean) and σ^2 (variance); 68-95-99.7 rule; Central Limit Theorem explains why quantitative traits tend toward normality. Variance partitioning: $V_p = V^G + V^E$ (+ V^{GE}); $V^G = V_a + V^D + V^I$. V_a (additive): alleles add equally; transmitted to offspring; determines selection response; contributes to h^2. V^D (dominance): within-locus allelic interaction; heterozygote deviates from midpoint; not transmitted intact; contributes to H^2 but not h^2. V^I (epistatic): between-locus interactions; generally small. G×E: different genotypes respond differently to environments; scale (non-crossover) = ranks stable; crossover = ranks change across environments; detected by multi-environment trials; norm of reaction shows genotype responses; crossover G×E prevents a single universally best variety.

2. Heritability: $H^2 = V^G/V_p$ (broad-sense; all genetic effects); $h^2 = V_a/V_p$ (narrow-sense; additive only; determines selection response); $h^2 \leq H^2$; range 0 to 1; population and environment specific. Methods: parent-offspring regression ($b = h^2$ for mid-parent; $b = 0.5h^2$ for one parent); twin studies [$h^2 \approx 2(r^{MR} - r^{DR})$]; half-sib analysis ($h^2 = 4\sigma^{2A}/V_p$); realised heritability ($h^2 = R/S$); GREML/molecular methods. Misinterpretations: h^2 does not mean the trait cannot change with environment (PKU: high h^2 but phenotype normalised by diet; height secular trend); Lewontin's plant analogy: within-group $h^2 = 1.0$ in both batches yet the between-group difference is entirely environmental (soil quality); high within-group h^2 cannot be used to infer causes of between-group differences. Breeder's equation: $R = h^2 \times S$; $S = i \times \sigma_p$; annual gain = $(i \times h \times \sigma_a)/L$; worked example: milk yield $h^2 = 0.30$; $i = 1.400$; $\sigma_p = 800$ L; S = 1,120 L; R = 336 L/generation. Limits: V_a decreases as alleles fix; inbreeding depression; correlated responses. QTL mapping: identifies chromosomal regions harbouring genes for quantitative traits; principle: individuals with different genotypes at a marker linked to a QTL have different mean phenotypic values; methods: single-marker ANOVA; interval mapping (LOD scores, threshold LOD ≥ 3.0); CIM; GWAS. MAS: uses QTL-linked markers for selection without phenotyping; valuable for difficult-to-measure traits (carcass traits, disease resistance after challenge), late-expressing traits, and traits in one sex only.

Answers to Pilot Questions [Series 4]

Part 4.1

1. Qualitative traits: discrete, non-overlapping phenotypic classes; controlled by few loci; analysed by Mendelian ratios; examples: ABO blood groups (four discrete classes: A, B, AB, O); coat colour in Labrador retrievers (black, chocolate, yellow). Quantitative traits: continuous phenotypic variation with no discrete classes; controlled by many polygene loci plus environmental effects; require statistical analysis; examples: human height; crop grain yield.
2. Polygenic model (Nilsson-Ehle 1908): a quantitative trait is determined by alleles at multiple loci, each contributing a small, equal, additive increment to the phenotype; no dominance; loci assort independently. Wheat grain colour experiment: red (AABB) x white (aabb) gives intermediate F1 (AaBb); F2 from AaBb x AaBb yields five classes in a 1:4:6:4:1 ratio (0 to 4 contributing alleles); the 6/16 class (two contributing alleles) is intermediate red; 1/16 is dark red (AABB) and 1/16 is white (aabb).
3. For n = 3 loci: F2 phenotypic classes = 2(3) + 1 = 7. Extreme parental phenotypes (all contributing alleles = AABBCC; or none = aabbcc) each occur at frequency $(1/4)^3 = 1/64 \approx$ 1.6 percent.
4. Mean (μ): the arithmetic average of all values in the population; the central tendency of the distribution. Variance (σ^2): the average squared deviation from the mean; measures the spread of the distribution. 68-95-99.7 rule: approximately 68 percent of observations fall within ±1 standard deviation of the mean; approximately 95 percent within ±2 standard deviations; approximately 99.7 percent within ±3 standard deviations.
5. Two reasons: polygenic determination (many loci each contributing a small additive effect produce a large number of phenotypic classes that overlap and appear continuous; as n increases, extreme classes become rarer and the distribution approaches a normal curve) and environmental modification (even if a small number of loci controlled a trait, environmental variation adds additional variability to each genotypic class, causing the classes to overlap and merge into a continuous distribution).

Part 4.2

1. $V_p = V^G + V^E$; V_p = total phenotypic variance (the total variation in trait values observed in the population, measurable directly from phenotypic data). V^G = genetic variance (the portion of V_p due to genotypic differences among individuals; further partitioned into $V_a + V^D + V^I$). V^E = environmental variance (the portion of V_p due to non-genetic factors: nutrition, temperature, random developmental variation, measurement error).

2. Additive genetic variance (V_a): the component due to the average effects of individual alleles considered separately; each contributing allele adds the same increment (no dominance; heterozygote is exactly intermediate); transmitted from parents to offspring; determines response to selection; defines narrow-sense heritability ($h^2 = V_a/V_p$). Dominance genetic variance (V^D): the component due to within-locus allelic interactions (dominance, recessiveness, overdominance); the heterozygote phenotype deviates from the midpoint between homozygotes; does not contribute to h^2.
3. V^D does not contribute to selection response because offspring inherit individual alleles (one from each parent), not the parental genotypic combinations; the dominance effect (the deviation of the heterozygote from the midpoint) cannot be passed intact to offspring; only the additive effects of alleles (V_a) are reliably transmitted from parents to offspring and therefore respond to selection.
4. G×E interaction occurs when different genotypes respond differently to changes in environment; the effect of the genotype on the phenotype depends on the environment. Scale (non-crossover) interaction: the rank order of genotypes is the same in all environments but the magnitude of differences between genotypes varies across environments; the best genotype in one environment is still the best in another. Crossover interaction: the rank order of genotypes changes across environments; genotype A outperforms B in one environment but B outperforms A in another; the performance lines cross when plotted across environments.
5. Crossover G×E makes it impossible to identify a single genotype that is best for all target environments; a variety that performs best in one location may perform poorly in another. This forces breeders to conduct multi-environment trials and either develop specifically adapted varieties (optimised for particular environments) or select for stability (minimising the G×E interaction term for a variety); it greatly increases the cost and time of variety development and release.

Part 4.3

1. Broad-sense heritability $H^2 = V^G/V_p = (V_a + V^D + V^I) / (V^G + V^E)$; the proportion of total phenotypic variance due to all genetic effects (additive + dominance + epistatic); ranges from 0 to 1. Narrow-sense heritability $h^2 = V_a/V_p$; the proportion of total phenotypic variance due to additive genetic variance only; determines response to selection; $h^2 \leq H^2$.
2. In the parent-offspring regression method, the mid-parent value (average of both parental phenotypes) is plotted on the x-axis and the mean offspring phenotype is plotted on the y-axis for many families; the slope b of the linear regression of offspring on mid-parent value equals

h^2 (the narrow-sense heritability); the steeper the slope, the higher h^2 and the more offspring resemble their parents.

3. $h^2 \approx 2(r^{MR} - r^{DR}) = 2(0.80 - 0.50) = 2 \times 0.30 = 0.60$.
4. $h^2 = 0.80$ for human height means that 80 percent of the phenotypic variance in height within a specific population is attributable to additive genetic differences among individuals. It does NOT mean that an individual's height is 80 percent fixed by their genes. The secular increase in average human height over the past 150 years in developed countries (caused by improved nutrition and healthcare) demonstrates that height is highly responsive to environmental change, even though within-generation variation is largely heritable.
5. Lewontin's plant analogy (1972): two batches of genetically diverse seeds (same genetic diversity) are grown, one batch in rich soil and one in poor soil; within each batch, all plants experience the same environment; all within-batch variation in plant height is genetic, so $h^2 = 1.0$ within each batch; yet the difference in mean height between the two batches is entirely due to soil quality (environment), not to genetic differences between the batches. This demonstrates that a high within-group heritability does not explain differences between groups; between-group differences may be entirely environmental even when within-group heritability is very high.

Part 4.4

1. Selection differential $S = \dot{x}_s - \dot{x}$ = the difference between the mean phenotype of selected parents and the mean of the whole parental population; it measures the superiority of selected parents. Response to selection $R = \dot{x}_o^{ff} - \dot{x}$ = the difference between the offspring mean and the pre-selection parental population mean; it measures the actual genetic gain achieved. Breeder's equation: $R = h^2 \times S$.
2. $S = i \times \sigma_p = 1.755 \times 400 = 702$ kg/ha. $R = h^2 \times S = 0.50 \times 702 = 351$ kg/ha per generation. The new population mean after one generation of selection = 3,500 + 351 = 3,851 kg/ha.
3. Response to selection tends to decrease across successive generations because: (1) additive genetic variance (V_a) decreases as favourable alleles increase in frequency toward fixation (the population becomes more homozygous for selected alleles; less genetic variation remains for selection to act on); (2) if inbreeding accompanies selection (small population size), inbreeding depression may reduce performance; (3) correlated responses in other traits may partially counteract gains; (4) alleles at low initial frequency take many generations to respond. Eventually, the population reaches a selection limit where all major QTL alleles are fixed and no further response is possible.

4. QTL mapping identifies chromosomal regions (quantitative trait loci) containing genes that contribute to variation in a quantitative trait. The principle: if a QTL influencing the trait is located close to a molecular marker on a chromosome, individuals with different genotypes at that marker will on average have different phenotypic values for the trait (because the marker genotype predicts the QTL genotype through linkage); statistical tests detect this association between marker genotype and trait phenotype in a segregating population.
5. Marker-assisted selection (MAS): the use of molecular markers linked to QTL to select individuals carrying favourable QTL alleles without directly phenotyping the trait. Two situations where MAS is particularly valuable: traits that are difficult, expensive, or impossible to measure in the individual being selected (e.g. carcass traits in livestock that can only be measured after slaughter; disease resistance requiring infection of the individual being tested); traits expressed in only one sex or late in life (e.g. milk production in dairy cattle expressed only in cows; selection on sires is based on daughters' performance, making early marker-based selection of bulls extremely valuable).

References and Attributions [Series 4]

Textual References (Books and External Sources)

- Falconer, D. S., & Mackay, T. F. C. (1996). Introduction to quantitative genetics (4th ed.). Longman, Harlow.
- Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., & Doebley, J. (2015). Introduction to genetic analysis (11th ed.). W. H. Freeman, New York.
- Lynch, M., & Walsh, B. (1998). Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland.
- Lewontin, R. C. (1972). The apportionment of human diversity. Evolutionary Biology, 6, 381–398.
- Lander, E. S., & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121, 185–199.

Figures, Plates, Tables, and Boxes

- Figure 4.1: Norms of reaction for four genotypes showing non-crossover and crossover G×E interaction. Attribution: AI generated. Suggested OER search string: "norm of reaction genotype environment interaction quantitative genetics OER".

- Box 4.1: Summary of key quantitative genetics formulae. Attribution: AI generated. Suggested OER search string: “quantitative genetics formulae heritability breeder’s equation QTL table OER”.

Course code: BIO 201

Course title: Genetics I

Course credit load: 2 Units

Series 5: Introduction to Population Genetics

- **Part 5.1: Genetic Variation in Populations**
  - Sub Part 5.1.1: What is a population and why study its genetics?
  - Sub Part 5.1.2: Types and measurement of genetic variation
  - Sub Part 5.1.3: Sources of genetic variation
- **Part 5.2: The Hardy-Weinberg Equilibrium**
  - Sub Part 5.2.1: The Hardy-Weinberg principle and its assumptions
  - Sub Part 5.2.2: Calculating allele and genotype frequencies
  - Sub Part 5.2.3: Testing for Hardy-Weinberg equilibrium
- **Part 5.3: Forces that Change Allele Frequencies**
  - Sub Part 5.3.1: Natural selection and changes in allele frequency
  - Sub Part 5.3.2: Genetic drift and effective population size
  - Sub Part 5.3.3: Mutation, migration, and non-random mating
- **Part 5.4: Molecular Population Genetics and Applications**
  - Sub Part 5.4.1: Molecular markers and their use in population genetics
  - Sub Part 5.4.2: Measures of population structure: F-statistics
  - Sub Part 5.4.3: Applications of population genetics

Introduction

Population genetics is the branch of genetics that studies the genetic composition of populations and how it changes over time. While Mendelian genetics focuses on transmission within families and quantitative genetics deals with traits influenced by many genes, population genetics operates at the level of entire populations, asking how allele and genotype frequencies change or remain stable across generations. It bridges genetics and evolutionary biology, providing the mathematical framework for understanding evolution at the genetic level.

This Series introduces genetic variation in populations and how it is measured; the Hardy-Weinberg equilibrium (HWE) as the null model of population genetics; the four major evolutionary forces (natural selection, genetic drift, mutation, and migration) that cause allele frequencies to change; and the molecular tools and applications of population genetics in conservation, forensics, epidemiology, and crop improvement. Together with Series 4, this Series completes the BIO 201 course.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 5.1** define a population in the genetic sense, describe the types of genetic variation found in populations, and explain the sources of genetic variation (Linked to Part 5.1),
- **SLO 5.2** state the Hardy-Weinberg principle and its assumptions, calculate allele and genotype frequencies from population data, and perform a chi-square test for Hardy-Weinberg equilibrium (Linked to Part 5.2),
- **SLO 5.3** explain how natural selection, genetic drift, mutation, and migration each change allele frequencies, and define effective population size (Linked to Part 5.3), and
- **SLO 5.4** describe the main categories of molecular markers used in population genetics, define F-statistics and explain their significance, and describe applications of population genetics in conservation, forensics, and crop improvement (Linked to Part 5.4).

5.1 Genetic Variation in Populations

5.1.1 What is a population and why study its genetics?

A population in the genetic sense is a group of individuals of the same species that occupy the same geographic area at the same time and interbreed with one another, sharing a common gene pool. The gene pool is the totality of all alleles at all loci in all individuals in the population at a given time. Population genetics is important because it provides the mathematical framework for understanding evolution (change in allele frequencies over time); it underpins conservation biology (genetic diversity in endangered species; risks of inbreeding and drift in small populations); it is essential for plant and animal breeding (maintaining diversity; estimating breeding values); it is required for human disease genetics (mapping disease genes; allele frequency distributions); it provides the basis for forensic DNA profiling (random match probabilities); and it informs human demographic history from patterns of genetic variation.

The distinction between genotype frequency and allele frequency is fundamental. Genotype frequency: the proportion of individuals with a specific genotype; f(AA) + f(Aa) + f(aa) = 1. Allele frequency: the proportion of all alleles at a locus of a specific type: p = f(AA) + ½ f(Aa); q = f(aa) + ½ f(Aa); p + q = 1. Alternatively: p = [2n(AA) + n(Aa)] / 2N; q = [2n(aa) + n(Aa)] / 2N, where n(AA), n(Aa), n(aa) are the numbers of each genotype and N is the total number of individuals.

Fill in the blank: A population in the genetic sense is a group of individuals of the same species occupying the same area and sharing a common _____ pool; allele frequency p = f(AA) + ½ f(Aa); p + q = 1 for two alleles at a biallelic locus.

5.1.2 Types and measurement of genetic variation

Types of genetic variation: (1) Single nucleotide polymorphisms (SNPs): the most common form; single base-pair differences; approximately 3 to 4 million per individual human genome; biallelic; used in GWAS and population structure analysis. (2) Insertion-deletion polymorphisms (indels): insertions or deletions of one or more base pairs; frameshift indels disrupt coding sequences. (3) Copy number variants (CNVs): variation in the number of copies of a genomic segment (1 kb to several Mb); affect gene dosage and phenotype. (4) Microsatellites (short

tandem repeats; STRs): 2 to 6 bp repeat motifs repeated in tandem; highly polymorphic (many alleles per locus); used in forensic profiling, paternity testing, and conservation genetics.

Measures of genetic variation: observed heterozygosity ($H_0 = n(Aa)/N$; proportion of heterozygotes in the sample); expected heterozygosity ($H^E = 2pq$ for a biallelic locus $= 1 - \Sigma p_i^2$; also called gene diversity; the expected proportion of heterozygotes under HWE); nucleotide diversity (π; average nucleotide differences per site between two randomly chosen sequences); allele richness (A; number of alleles per locus in the sample). Each measure captures a different aspect of diversity; H^E is the most widely reported.

Fill in the blank: Observed heterozygosity $H_0 = n(Aa)/N$; expected heterozygosity $H^E =$ _____ (for a biallelic locus) $= 1 - \Sigma p_i^2$; H^E is also called gene diversity; it is the proportion of heterozygotes expected under Hardy-Weinberg equilibrium given the observed allele frequencies.

5.1.3 Sources of genetic variation

(1) Mutation: the ultimate source of all new genetic variation; a permanent change in the DNA sequence; average point mutation rate in humans approximately 1 to 2×10^{-8} per base pair per generation, producing approximately 60 to 70 new mutations per person per generation; mutations include silent (synonymous), missense, and nonsense changes. (2) Recombination (crossing over): does not create new alleles but reshuffles existing alleles into new combinations, generating haplotypic diversity. (3) Gene flow (migration): movement of individuals or gametes between populations, introducing alleles from one population's gene pool into another; homogenises allele frequencies between populations. (4) Independent assortment: shuffles existing alleles into new genotypic combinations each generation; 2^{23} different gamete types possible from independent assortment alone in humans (23 chromosome pairs).

Maintenance of genetic variation: genetic variation is maintained by heterozygote advantage (overdominance; the Aa heterozygote has higher fitness than either homozygote, e.g. sickle-cell heterozygotes in malaria-endemic regions); frequency-dependent selection (rare alleles are favoured until they become common, maintaining a stable polymorphism); balancing selection; and neutral drift (most alleles are selectively neutral and drift at equal rates). Without these

mechanisms, directional selection and drift would eventually fix one allele and eliminate variation.

Fill in the blank: Mutation is the ultimate source of all new genetic variation; it occurs at approximately 1 to 2 × 10^{-8} per base pair per _____; recombination reshuffles existing alleles into new combinations but does not create new alleles; gene flow introduces alleles from other populations.

Pilot questions 5.1

1. Define a population in the genetic sense and explain what is meant by a gene pool.
2. In a population of 200 individuals, 80 are AA, 90 are Aa, and 30 are aa. Calculate the allele frequencies p and q.
3. Distinguish between observed heterozygosity (H_o) and expected heterozygosity (H^E) and state the formula for each.
4. State four sources of genetic variation in populations.
5. Explain why mutation rates are low yet mutation is described as the ultimate source of all genetic variation.

5.2 The Hardy-Weinberg Equilibrium

5.2.1 The Hardy-Weinberg principle and its assumptions

The Hardy-Weinberg principle (HWE), independently derived by G. H. Hardy and Wilhelm Weinberg in 1908, states: in a large, randomly mating population in the absence of mutation, migration, and natural selection, allele frequencies remain constant from generation to generation, and genotype frequencies reach the following equilibrium values after one generation of random mating: $f(AA) = p^2$; $f(Aa) = 2pq$; $f(aa) = q^2$, where $p + q = 1$ and $p^2 + 2pq + q^2 = 1$. The HWE is the fundamental null model of population genetics: it describes what is expected when no evolutionary forces are acting; any deviation from HWE indicates that one or more evolutionary forces are operating on the locus.

Assumptions of HWE: (1) Random mating (panmixia): every individual has an equal probability of mating with every other individual regardless of genotype. (2) Large population size: genetic

drift is negligible. (3) No natural selection: all genotypes have equal fitness. (4) No mutation: allele frequencies are not altered by new mutations. (5) No migration: no alleles enter or leave the population. When any assumption is violated, the population deviates from HWE. Testing for HWE deviation (using chi-square; see Part 5.2.3) tests whether any evolutionary force is operating on a locus.

Fill in the blank: HWE genotype frequencies are p^2, _____, and q^2; they are derived from the expansion of $(p + q)^2 = 1$; HWE is the null model of population genetics and any significant deviation from these frequencies indicates that evolutionary forces are acting on the population at that locus.

5.2.2 Calculating allele and genotype frequencies

Calculating allele frequencies from genotype counts: $p = [2n(AA) + n(Aa)] / 2N$; $q = [2n(aa) + n(Aa)] / 2N$; or equivalently $p = f(AA) + ½ f(Aa)$; $q = f(aa) + ½ f(Aa)$. Worked example: 500 individuals genotyped for MN blood group. Observed: 180 MM, 240 MN, 80 NN. Genotype frequencies: f(MM) = 0.36; f(MN) = 0.48; f(NN) = 0.16. Allele frequencies: p(M) = 0.36 + ½(0.48) = 0.60; q(N) = 0.16 + ½(0.48) = 0.40; p + q = 1.00. Expected HWE genotype frequencies: $f(MM) = p^2 = 0.36$; f(MN) = 2pq = 0.48; $f(NN) = q^2 = 0.16$. Expected numbers: E(MM) = 180; E(MN) = 240; E(NN) = 80. Observed = Expected: the MN locus is in HWE in this population.

For X-linked loci: males are hemizygous; allele frequency estimated separately from male and female genotype frequencies. For multiple alleles: allele frequencies p_1, p_2, ..., p^k must sum to 1; HWE genotype frequencies: $f(p_ip_i) = p_i^2$ (homozygote) and $f(p_ip_j) = 2p_ip_j$ (heterozygote) for $i \neq j$.

Fill in the blank: Allele frequency $p = f(AA) + ½ f(Aa)$; HWE expected genotype frequencies are p^2, 2pq, q^2; these expected _____ are multiplied by N to give expected numbers used in the chi-square goodness-of-fit test for HWE.

5.2.3 Testing for Hardy-Weinberg equilibrium

A chi-square goodness-of-fit test is used to test whether observed genotype frequencies differ significantly from HWE expectations. The procedure is identical to Series 3 methods with

expected values from HWE. Degrees of freedom for biallelic locus: df = 3 classes - 1 (total constraint) - 1 (p estimated from data) = 1. For k alleles: df = k(k-1)/2. Worked example: 1,000 individuals genotyped. Observed: n(AA) = 650, n(Aa) = 200, n(aa) = 150. Allele frequencies: p = [2(650)+200] / 2,000 = 0.75; q = 0.25. Expected: E(AA) = $0.75^2 \times 1{,}000 = 562.5$; E(Aa) = $2(0.75)(0.25) \times 1{,}000 = 375$; E(aa) = $0.25^2 \times 1{,}000 = 62.5$. Chi-square: $\chi^2 = [(650-562.5)^2/562.5] + [(200-375)^2/375] + [(150-62.5)^2/62.5] = 13.61 + 81.67 + 122.50 = 217.78$. df = 1; critical value 3.841 at $\alpha = 0.05$. Since 217.78 >> 3.841, REJECT H_0. Conclusion: significant deficiency of heterozygotes; population is not in HWE; suggests inbreeding or assortative mating.

Fill in the blank: For a biallelic HWE chi-square test, degrees of freedom = _____; if the calculated χ^2 exceeds the critical value of 3.841 ($\alpha = 0.05$), H_0 is rejected and the population is not in HWE; deficiency of heterozygotes indicates inbreeding; excess indicates heterozygote advantage or admixture.

Pilot questions 5.2

1. State the Hardy-Weinberg principle and list its five assumptions.
2. In a population of 400 individuals, 144 are AA, 192 are Aa, and 64 are aa. Calculate p, q, and the expected HWE genotype frequencies.
3. Using the data from question 2, perform a chi-square test and state whether the population is in HWE.
4. In a HWE chi-square test at a biallelic locus, why are the degrees of freedom 1 and not 2?
5. A population has $q^2 = 0.04$ for an autosomal recessive disorder. What is q? What is the carrier frequency (2pq)?

5.3 Forces that Change Allele Frequencies

5.3.1 Natural selection and changes in allele frequency

Natural selection is the differential survival and reproduction of individuals based on their genotype. It is the only evolutionary force that produces adaptive change. Fitness (w): the relative reproductive success of a genotype; the most fit genotype is set to w = 1. Selection coefficient (s = 1 - w): the reduction in relative fitness; s ranges from 0 (no selection) to 1 (lethal). For a recessive deleterious allele (a): AA w = 1; Aa w = 1; aa w = 1 - s (recessive selection). Change in allele frequency per generation: $\Delta q = -sq^2(1-q) / (1 - sq^2)$. Selection against

recessive alleles is slow when q is small because rare recessive alleles are sheltered in heterozygotes where selection cannot act on them; the frequency of aa (= q^2) is tiny when q is small.

Types of natural selection: (1) Directional selection: consistently favours one allele; the favoured allele increases toward fixation; reduces genetic variation. (2) Stabilising selection: favours intermediate phenotypes; reduces V_a without shifting the mean; most common type in stable environments. (3) Disruptive selection: favours extreme phenotypes over intermediates; can lead to population divergence. (4) Balancing selection: maintains multiple alleles by overdominance (heterozygote advantage), frequency-dependent selection, or spatially variable selection. Classic example: sickle-cell polymorphism; HbA/HbS heterozygotes have higher fitness than HbA/HbA (susceptible to malaria) or HbS/HbS (sickle-cell disease) in malaria-endemic areas; overdominance maintains both alleles at high frequency.

Fill in the blank: Natural selection is the only evolutionary force that produces _____ change; selection against a recessive allele is slow when q is small because rare recessive alleles are sheltered in heterozygotes; balancing selection (e.g. heterozygote advantage at the HbS locus) maintains multiple alleles at a locus.

5.3.2 Genetic drift and effective population size

Genetic drift is random change in allele frequencies from generation to generation due to sampling variation in a finite population. Effects: (1) Allele fixation or loss: in any finite population, drift will eventually cause every allele to be fixed (p = 1) or lost (p = 0). (2) Reduction of heterozygosity: $H^t = H_0 (1 - 1/2N^E)^t$; heterozygosity declines at rate $1/(2N^E)$ per generation, where N^E is the effective population size. (3) Genetic divergence: isolated populations drift independently, causing allele frequencies to diverge over time. Bottleneck effect: a drastic reduction in population size causes loss of rare alleles and random shifts in allele frequencies. Founder effect: a new population established by few individuals carries a non-representative sample of the source gene pool; explains high frequency of certain rare disorders in isolated human populations (e.g. Ellis-van Creveld syndrome in Old Order Amish).

Effective population size (N^E): the size of an idealised population that would experience the same rate of genetic drift as the actual population; N^E is almost always less than census size N because of: unequal sex ratios ($N^E = 4N^fN^f/(N^f + N^f)$); high variance in reproductive success; population size fluctuations (the harmonic mean across generations, strongly influenced by bottlenecks). A single severe bottleneck can reduce N^E to a small fraction of the long-term census size.

Fill in the blank: Genetic _____ reduces heterozygosity at rate $1/(2N^E)$ per generation; it is strongest in small populations and eventually causes alleles to be fixed or lost; the effective population size N^E is almost always smaller than the census size N because of unequal sex ratios, reproductive variance, and bottlenecks.

5.3.3 Mutation, migration, and non-random mating

Mutation: the ultimate source of new alleles; the point mutation rate is approximately 10^{-6} to 10^{-5} per locus per generation, making mutation the weakest force for changing allele frequencies in the short term. Mutation-selection balance: the equilibrium frequency of a recessive deleterious allele is $q^E \approx \sqrt{(\mu/s)}$, where μ is the mutation rate and s is the selection coefficient against aa. Migration (gene flow): movement of individuals or gametes between populations; the change in allele frequency per generation in the recipient population is $\Delta p = m(p^M - p)$, where m = migration rate and p^M = allele frequency in migrants; migration is the most powerful force homogenising allele frequencies between populations and greatly reduces $F^H{}_T$.

Non-random mating: (1) Inbreeding: mating between relatives; does NOT change allele frequencies but increases homozygosity and decreases heterozygosity; measured by the inbreeding coefficient F (probability that both alleles at a locus are identical by descent); genotype frequencies under inbreeding: $f(AA) = p^2(1-F) + pF$; $f(Aa) = 2pq(1-F)$; $f(aa) = q^2(1-F) + qF$; when F = 0 these reduce to HWE. Inbreeding depression: reduced fitness due to increased expression of deleterious recessive alleles. (2) Assortative mating: individuals mate preferentially with phenotypically similar (positive) or dissimilar (negative) partners; positive assortative mating increases homozygosity; negative assortative mating increases heterozygosity. (3) Sexual selection: mate choice based on specific traits; can cause rapid allele frequency change.

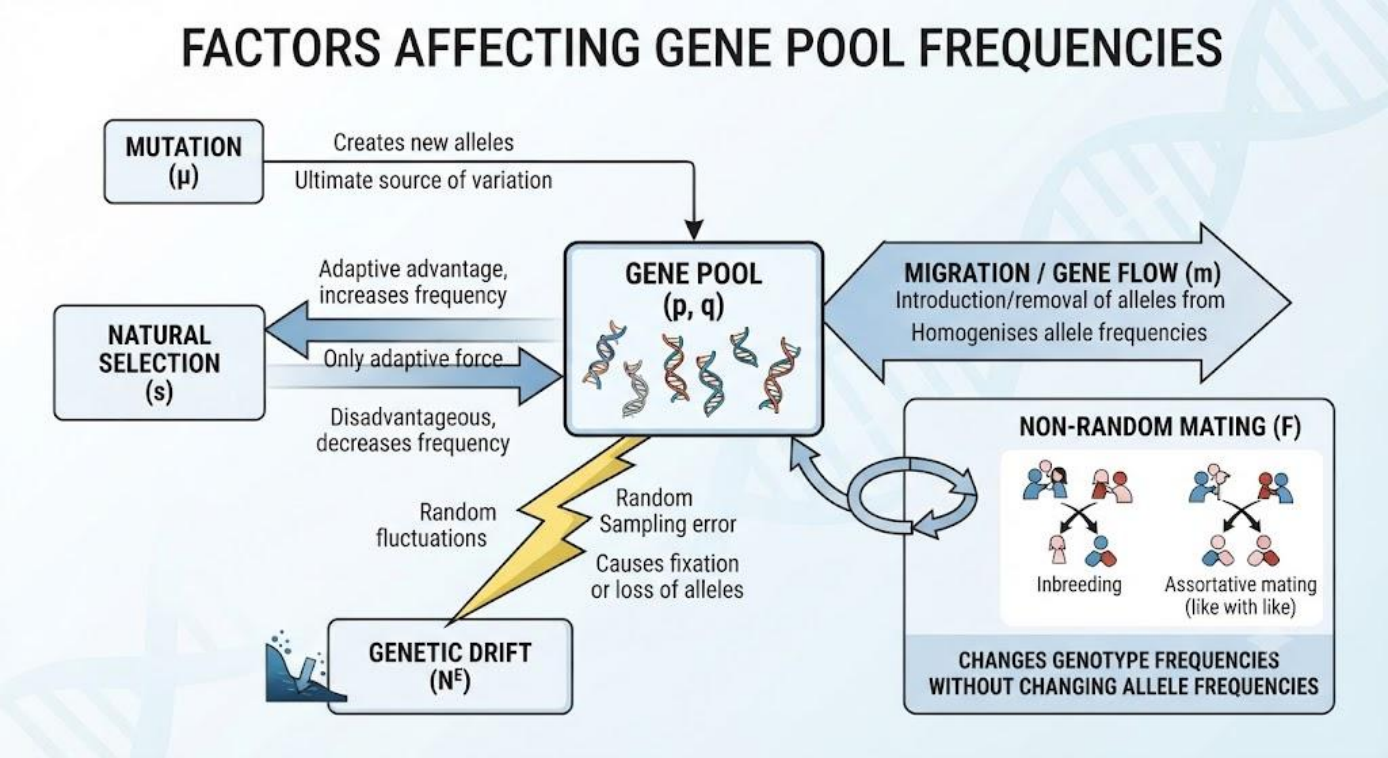

Figure 5.1: The four evolutionary forces (mutation, natural selection, genetic drift, and migration) and non-random mating acting on a population gene pool

Pilot questions 5.3

1. Define fitness (w) and selection coefficient (s) and explain why selection against recessive alleles is slow when the allele is rare.
2. Distinguish between directional, stabilising, and balancing selection with one example of each.
3. Define genetic drift and explain why it has a greater effect in small populations.
4. Define effective population size (N^E) and state two reasons why N^E is smaller than census size N.
6. Define the inbreeding coefficient F and write the genotype frequencies at a biallelic locus under inbreeding.

5.4 Molecular Population Genetics and Applications

5.4.1 Molecular markers and their use in population genetics

Characteristics of ideal molecular markers: codominant (all genotypes distinguishable; accurate allele frequency estimation); highly polymorphic (many alleles or high H^E; more information per locus); abundant throughout the genome; selectively neutral (variation reflects demography, not selection). Major marker types: (1) RFLPs (restriction fragment length polymorphisms): among the earliest molecular markers; based on restriction enzyme cut site variation; codominant; low polymorphism; largely replaced by newer types. (2) Microsatellites (STRs): 2 to 6 bp repeat units; highly polymorphic (H^E often above 0.70; many alleles per locus); codominant; typed by PCR and capillary electrophoresis; primary marker for forensic profiling, paternity testing, and conservation genetics. (3) SNPs (single nucleotide polymorphisms): most abundant (millions per genome); biallelic; lower polymorphism per locus than microsatellites; typed by microarray or next-generation sequencing; primary marker for GWAS and population structure analysis.

(4) AFLPs and RAPDs: dominant markers; cannot distinguish heterozygotes; used in plant population genetics and variety fingerprinting but limited for allele frequency estimation. (5) Mitochondrial DNA (mtDNA) and chloroplast DNA (cpDNA): maternally inherited; no recombination; evolves at predictable rates; used for phylogeography (geographic patterns of genetic lineages) and species identification; effective N^E = 1/4 of nuclear N^E in diploid organisms with equal sex ratios. The choice of marker depends on the question: STRs for forensics and conservation; SNPs for GWAS and demographic history; mtDNA for phylogeography and species identification.

Fill in the blank: Microsatellites (STRs) are highly _____ (many alleles per locus; H^E often above 0.70), codominant, and the primary marker for forensic profiling and conservation genetics; SNPs are more abundant (millions per genome) but less polymorphic per locus and are used for GWAS and population structure analysis.

5.4.2 Measures of population structure: F-statistics

Population structure is the subdivision of a species into genetically distinct subpopulations that differ in allele frequencies due to limited gene flow. Wright's F-statistics (fixation indices) measure inbreeding and genetic differentiation in hierarchical populations: F^{IH} (Individual within Subpopulation): within-subpopulation inbreeding coefficient; = 0 when subpopulations are in HWE. F^{HT} (Subpopulation within Total): standardised variance in allele frequencies among

subpopulations; $F^{H}_{T} = (H_T - H^H) / H_T$, where H_T = expected heterozygosity in the total population and H^H = average expected heterozygosity within subpopulations; ranges from 0 (no differentiation) to 1 (complete differentiation). F^{I}_{T} (Individual within Total): $F^{I}_{T} = 1 - (1-F^{IH})(1-F^{H}_{T})$.

Interpretation of F^{H}_{T}: 0 to 0.05 = little differentiation; 0.05 to 0.15 = moderate; 0.15 to 0.25 = large; above 0.25 = very large. Relationship between F^{H}_{T} and gene flow: under the island migration model, $F^{H}_{T} \approx 1/(1 + 4Nm)$, where Nm = number of migrants per generation; even one migrant per generation (Nm = 1) gives $F^{H}_{T} = 0.20$; Nm = 10 gives $F^{H}_{T} \approx 0.024$; small amounts of gene flow greatly reduce genetic differentiation. Alternative measures: R^{H}_{T} (for microsatellites); Jost's D (actual allelic differentiation independent of within-population heterozygosity).

Fill in the blank: $F^{H}_{T} = (H_T - H^H) / H_T$ measures genetic _____ among subpopulations; it ranges from 0 (no differentiation; high gene flow) to 1 (complete differentiation); under the island model $F^{H}_{T} \approx 1/(1+4Nm)$ so even one migrant per generation greatly reduces differentiation.

5.4.3 Applications of population genetics

(1) Conservation biology: measuring H^E, H_o, allele richness, and F^{H}_{T} to assess genetic diversity and inbreeding risk in threatened species; identifying management units (genetically distinct subpopulations managed separately to preserve local adaptation); estimating N^E from temporal allele frequency changes; designing captive breeding programmes to minimise inbreeding; detecting recent bottlenecks. (2) Forensic genetics: individual identification using panels of 15 to 20 highly polymorphic STR loci (e.g. CODIS in the USA; National DNA Database in the UK); random match probability (RMP) = product of genotype frequencies across all STR loci (assuming HWE at each locus and independence between loci); HWE allows genotype frequencies to be calculated from allele frequencies in population databases.

(3) Crop and livestock improvement: assessing genetic diversity in genebank accessions; identifying duplicate accessions; selecting diverse parents for crosses; marker-assisted selection (MAS); variety traceability and quality certification; parentage and pedigree verification. (4) Human population history and medical genetics: reconstructing demographic history

(bottlenecks, expansions, admixture) from SNP patterns; identifying population-specific alleles with clinical significance (pharmacogenomics); PCA and model-based clustering (STRUCTURE, ADMIXTURE) of SNP data reveal population structure and ancestry. (5) Disease epidemiology: identifying pathogen transmission routes and disease reservoirs; determining outbreak source (single vs multiple introductions); tracking spread of drug-resistant alleles.

Concept	Formula	Application
Allele Frequency	p=2NAA+NAa2N, q=1−p	Calculates frequency of alleles in a population; foundation of genetic structure analysis.
Hardy–Weinberg Genotype Frequencies	p2,2pq,q2	Predicts expected genotype frequencies under random mating and no evolutionary forces.
Hardy–Weinberg Chi-Square Test	$\chi 2=\sum(O-E)2E$	Tests whether observed genotype frequencies deviate significantly from HWE expectations.
Expected Heterozygosity	He=1−∑pi2	Measures genetic diversity; probability that two alleles drawn at random are different.
Selection Response	Δp=spq1−sq	Models allele frequency change under selection pressure.
Genetic Drift and Heterozygosity	Ht=H0(1−12N)t	Predicts loss of heterozygosity over generations due to finite population size.
Inbreeding Genotype Frequencies	AA=p2+Fpq, Aa=2pq(1−F), aa=q2+Fpq	Adjusts genotype frequencies for inbreeding coefficient F.
FST and Gene Flow	FST=HT−HSHT, Nm≈1−FST4FST	Quantifies population differentiation and estimates gene flow between populations.
Mutation–Selection Balance	q=μs (for recessive deleterious allele)	Predicts equilibrium frequency of harmful alleles maintained by mutation and selection.

Box 5.1: Summary of key population genetics formulae and their applications

Pilot questions 5.4

6. State three properties of an ideal molecular marker for population genetics.
7. Distinguish between microsatellites (STRs) and SNPs as population genetics markers, comparing polymorphism, abundance, and primary applications.
8. Define $F^{H}T$ and state what values of 0, 0.10, and 0.50 indicate about genetic differentiation.
9. Explain how HWE is used in forensic DNA profiling to calculate the random match probability.
10. Describe two applications of population genetics in conservation biology.

Series Summary

This Series introduced the principles of **population genetics**, focusing on how genetic variation is measured, maintained, and changes within populations over time. You explored populations and allele frequencies, types and sources of genetic variation, and the **Hardy-Weinberg Equilibrium (HWE)** as the foundation for studying population genetics. The Series also explained how to calculate allele and genotype frequencies, test for HWE using the chi-square test, and interpret deviations from equilibrium.

Building on these concepts, the Series examined the major evolutionary forces that shape genetic variation, including natural selection, genetic drift, mutation, migration, and non-random mating. You also studied molecular markers, F-statistics, and the practical applications of population genetics in conservation, forensics, crop improvement, disease genetics, and epidemiology. By completing this Series, you should now be able to analyse genetic variation within populations, apply Hardy-Weinberg principles, evaluate the effects of evolutionary processes on allele frequencies, and explain the practical uses of population genetics, thereby achieving all four **Series Learning Outcomes (SLOs 5.1 to 5.4)**.

Glossary of Terms

- **Allele frequency:** the proportion of all alleles at a locus in a population that are of a specific type; $p = f(AA) + ½\ f(Aa)$; $p + q = 1$ for two alleles.
- **Balancing selection:** natural selection that maintains two or more alleles at a locus; mechanisms include heterozygote advantage, frequency-dependent selection, and spatially variable selection.
- **Bottleneck effect:** a drastic reduction in population size causing loss of rare alleles and random shifts in allele frequencies; reduces heterozygosity and increases genetic drift.
- **Effective population size (N^E):** the size of an idealised population experiencing the same rate of genetic drift as the actual population; almost always smaller than the census size N.
- **Expected heterozygosity (H^E):** the expected proportion of heterozygotes under HWE: $H^E = 2pq$ (biallelic) $= 1 - \Sigma p_i^2$; a standard measure of genetic diversity; also called gene diversity.

- **$F^{H}{}_{T}$ (fixation index):** Wright's F-statistic measuring genetic differentiation among subpopulations: $F^{H}{}_{T} = (H_T - H^{H})/H_T$; ranges from 0 (no differentiation) to 1 (complete differentiation); ≈1/(1+4Nm) under the island model.
- **Fitness (w):** the relative reproductive success of a genotype relative to the most fit genotype (w = 1); selection coefficient s = 1 - w.
- **Founder effect:** a bottleneck in which a new population is founded by a small number of individuals; the new gene pool is a non-representative sample of the source population.
- **Gene pool:** all alleles at all loci in all individuals in a population at a given time; the unit of study in population genetics.
- **Genetic drift:** random change in allele frequencies due to sampling variation in finite populations; reduces H^{E} at rate $1/(2N^{E})$ per generation; strongest in small populations.
- **Hardy-Weinberg equilibrium (HWE):** the state of a large, randomly mating population with no selection, mutation, or migration; genotype frequencies = p^2, 2pq, q^2; the null model of population genetics.
- **Inbreeding coefficient (F):** the probability that the two alleles at a locus are identical by descent; increases homozygosity; heterozygosity under inbreeding = 2pq(1-F).
- **Microsatellite (STR):** a highly polymorphic marker (2 to 6 bp repeat motifs); many alleles per locus; codominant; used in forensics, paternity testing, and conservation genetics.
- **Mutation-selection balance:** the equilibrium at which ongoing mutation replenishes deleterious recessive alleles removed by selection; $q^{E} \approx \sqrt{(\mu/s)}$.
- **Natural selection:** differential survival and reproduction based on genotype; the only evolutionary force producing adaptive change; types: directional, stabilising, disruptive, balancing.
- **Population (genetic):** a group of individuals of the same species sharing a common gene pool; the unit of study in population genetics.
- **Population structure:** the subdivision of a species into genetically distinct subpopulations differing in allele frequencies due to limited gene flow; measured by $F^{H}{}_{T}$.
- **SNP (single nucleotide polymorphism):** the most common form of genetic variation; a single base-pair difference; biallelic; used in GWAS and population structure analysis.

Concept Check Questions [Series 5]

Objective Questions

6. A population in the genetic sense is a group of individuals sharing a common _____ pool; allele frequency p = f(AA) + ½ f(Aa); p + q = 1 for two alleles. (Linked to SLO 5.1)
7. HWE predicts genotype frequencies of p^2, 2pq, and q^2; this equilibrium is maintained only when the population is large, randomly mating, and has no _____, mutation, or migration. (Linked to SLO 5.2)
8. Genetic _____ is random change in allele frequencies in finite populations; it reduces heterozygosity at rate $1/(2N^E)$ per generation and is strongest in small populations. (Linked to SLO 5.3)
9. Inbreeding changes genotype frequencies but does NOT directly change _____ frequencies; heterozygosity under inbreeding is 2pq(1-F). (Linked to SLO 5.3)
10. F^H_T measures genetic _____ among subpopulations; $F^H_T = 0$ means no differentiation; $F^H_T = 1$ means subpopulations are fixed for different alleles. (Linked to SLO 5.4)

Short-Answer Questions

11. In a population of 500 individuals, 200 are AA, 200 are Aa, and 100 are aa. Calculate p, q, and the HWE expected genotype numbers. (Linked to SLO 5.2)
12. State the five assumptions of HWE and state the evolutionary force violated when each assumption fails. (Linked to SLO 5.2)
13. Explain why selection against a recessive allele is slow when the allele is rare. (Linked to SLO 5.3)
14. Define N^E and state two reasons it is smaller than census size N. (Linked to SLO 5.3)
15. Give two applications of population genetics in conservation and two in forensics. (Linked to SLO 5.4)

Long-Answer Questions

13. Describe genetic variation in populations: types, measures, and sources. Explain the Hardy-Weinberg principle, its assumptions, calculation of allele and genotype frequencies, and the chi-square test for HWE. (Linked to SLOs 5.1, 5.2)
14. Explain how natural selection, genetic drift, mutation, migration, and non-random mating each affect allele or genotype frequencies. Describe molecular markers used in population genetics, define F-statistics, and outline applications of population genetics. (Linked to SLOs 5.3, 5.4)

Answers to Concept Check Questions [Series 5]

Objective Questions

7. Gene.
8. Selection.
9. Drift.
10. Allele.
11. Differentiation.

Short-Answer Questions

1. p = [2(200)+200]/(2×500) = 600/1,000 = 0.60; q = 0.40. Expected numbers: E(AA) = 0.36×500 = 180; E(Aa) = 0.48×500 = 240; E(aa) = 0.16×500 = 80.
2. Large population (violated → drift); random mating (violated → inbreeding/assortative mating changes genotype frequencies); no selection (violated → allele frequencies shift toward fitter genotype); no mutation (violated → new alleles created); no migration (violated → allele frequencies altered by incoming alleles).
3. Rare recessive alleles occur predominantly in Aa heterozygotes; selection only acts on aa homozygotes (frequency = q^2); when q is very small, q^2 is tiny so very few individuals are exposed to selection each generation and Δq is negligible.
4. N^E is the idealised population size experiencing the same drift rate as the actual population. Two reasons $N^E < N$: unequal sex ratios ($N^E = 4N^fN^f/(N^f+N^f)$); high variance in reproductive success (a few individuals contribute disproportionately many offspring, reducing the effective number of breeders).
5. Conservation (two): genetic diversity assessment (H^E and F^HT to determine inbreeding risk); estimating N^E from temporal allele frequency changes to assess extinction risk. Forensics (two): individual identification via STR profiles (product rule match probability assuming HWE); parentage testing (matching STR genotypes of putative parent and child).

Long-Answer Questions

1. Genetic variation: SNPs (most common; biallelic; GWAS); STRs (highly polymorphic; forensics); indels; CNVs. Measures: H_o = n(Aa)/N; H^E = 2pq; nucleotide diversity π; allele richness A. Sources: mutation (ultimate source; approximately 60 to 70 new mutations/person/generation); recombination (reshuffles alleles); gene flow (homogenises frequencies; introduces alleles); independent assortment (2^{23} gamete types in humans). HWE: p^2, 2pq, q^2 after one generation of random mating; five assumptions (large N; random mating; no selection; no mutation; no migration); calculation: p = f(AA)+½ f(Aa); expected genotypes =

p^2N, 2pqN, q^2N; chi-square goodness-of-fit test (df = 1; critical value 3.841; deficiency of heterozygotes = inbreeding; excess = heterozygote advantage or admixture).

2. Natural selection: w = fitness; s = 1-w; directional (one allele to fixation); stabilising (intermediate favoured; reduces V_a); disruptive (extremes favoured); balancing (multiple alleles maintained; sickle-cell overdominance). Genetic drift: random; $H^I = H_0(1-1/2N^E)^I$; bottleneck; founder effect; $N^E < N$. Mutation: very slow; mutation-selection balance $q^E \approx \sqrt{(\mu/s)}$. Migration: $\Delta p = m(p^M-p)$; homogenises frequencies; reduces $F^H{}_T$. Non-random mating: inbreeding (F = P(IBD); f(Aa) = 2pq(1-F)); assortative mating; sexual selection. Molecular markers: STRs (polymorphic; forensics); SNPs (abundant; GWAS; population structure); mtDNA (phylogeography). F-statistics: F^{IH} (within-subpopulation inbreeding); $F^H{}_T$ (0 to 1; $\approx 1/(1+4Nm)$); $F^I{}_T$ (individual vs total). Applications: conservation (diversity; N^E; management units; captive breeding); forensics (STR RMP); crop improvement (MAS; genebank diversity); human history and disease genetics; pathogen epidemiology.

Answers to Pilot Questions [Series 5]

Part 5.1

1. A population in the genetic sense is a group of individuals of the same species occupying the same geographic area and sharing a common gene pool. The gene pool is the totality of all alleles at all loci in all individuals in the population.
2. p = [2(80)+90]/(2×200) = 250/400 = 0.625; q = [2(30)+90]/400 = 150/400 = 0.375; check: p + q = 1.00.
3. H_0 = n(Aa)/N: the observed proportion of heterozygotes in the sample. H^E = 2pq (biallelic) = 1-Σp_i^2: the proportion of heterozygotes expected under HWE given the observed allele frequencies.
4. Mutation (creates new alleles; ultimate source). Recombination (reshuffles existing alleles). Gene flow/migration (introduces alleles from other populations). Independent assortment (generates new genotypic combinations each generation).
5. Mutation rates are very low (approximately 10^{-8} per base pair per generation) so mutation changes allele frequencies very slowly each generation; however, it is the ultimate source because it is the only process that creates entirely new alleles that have never existed before in the population.

Part 5.2

1. HWE states that in a large, randomly mating population with no selection, mutation, or migration, genotype frequencies are p^2, 2pq, q^2 and allele frequencies are constant across generations. Five assumptions: large population; random mating; no natural selection; no mutation; no migration.
2. p = [2(144)+192]/(2×400) = 480/800 = 0.60; q = 0.40. HWE: f(AA) = 0.36; f(Aa) = 0.48; f(aa) = 0.16. Expected numbers (N=400): E(AA)=144; E(Aa)=192; E(aa)=64.
3. Observed = (144, 192, 64); Expected = (144, 192, 64); $\chi^2 = 0$; do not reject H_0; the population is in perfect HWE.
4. Three genotypic classes give 3 data points; df = 3 - 1 (total constraint) - 1 (p estimated from data; q = 1-p is not independent) = 1.
5. $q = \sqrt{0.04} = 0.20$; p = 0.80; carrier frequency 2pq = 2×0.80×0.20 = 0.32 (32 percent carriers).

Part 5.3

1. Fitness w = relative reproductive success of a genotype; selection coefficient s = 1 - w. Selection against recessive alleles is slow when q is small because rare recessive alleles occur mostly in Aa heterozygotes; only aa individuals (frequency q^2) are selected against; when q is tiny, q^2 is negligible and Δq per generation is very small.

2. Directional: consistently favours one allele; example: antibiotic resistance alleles increasing under antibiotic treatment. Stabilising: favours intermediate phenotypes; example: human birth weight (very large or very small babies have higher mortality). Balancing: maintains multiple alleles; example: HbA/HbS heterozygotes have higher fitness than either homozygote in malaria-endemic areas.
3. Genetic drift is random change in allele frequencies due to sampling variation in finite populations. It has a greater effect in small populations because the gametes forming the next generation are a smaller sample; sampling error is proportionately larger; rate of heterozygosity loss = $1/(2N^E)$ is larger when N^E is small.
4. N^E is the size of an idealised population experiencing the same rate of genetic drift as the actual population. Two reasons $N^E < N$: unequal sex ratios ($N^E = 4N^fN^f/(N^f+N^f)$); high variance in reproductive success (some individuals contribute many offspring, others few, reducing effective number of breeders).
5. F = the probability that both alleles at a locus in an individual are identical by descent (IBD). Genotype frequencies under inbreeding: $f(AA) = p^2(1-F) + pF$; $f(Aa) = 2pq(1-F)$; $f(aa) = q^2(1-F) + qF$; when F = 0 these reduce to HWE frequencies.

Part 5.4

1. Codominant (all genotypes distinguishable; accurate allele frequency estimation); highly polymorphic (many alleles or high H^E; more information per locus); selectively neutral (variation reflects demographic history, not selection).
2. Microsatellites (STRs): highly polymorphic (H^E often above 0.70; many alleles); codominant; primary use in forensic profiling, paternity testing, and conservation genetics. SNPs: biallelic; less polymorphic per locus; millions per genome; primary use in GWAS and population structure analysis with genome-wide data.
3. $F^H_T = (H_T - H^H)/H_T$. $F^H_T = 0$: no differentiation; high gene flow; subpopulations have identical allele frequencies. $F^H_T = 0.10$: moderate differentiation; some allele frequency differences between subpopulations. $F^H_T = 0.50$: very large differentiation; subpopulations differ substantially; very limited gene flow.
4. In forensic profiling, the genotype frequency at each STR locus is calculated assuming HWE (heterozygote frequency = $2p_ip_j$; homozygote = p_i^2) using allele frequencies from a population database; the random match probability (RMP) is the product of genotype frequencies across all STR loci (product rule for independent loci).
5. Genetic diversity assessment: measuring H^E, H_o, and F^H_T to quantify inbreeding risk and guide gene flow management in threatened species. Identifying management units: using molecular

markers to delineate genetically distinct subpopulations that should be conserved and managed separately to preserve local adaptation.

References and Attributions [Series 5]

Textual References (Books and External Sources)

- Hartl, D. L., & Clark, A. G. (2007). Principles of population genetics (4th ed.). Sinauer Associates, Sunderland.
- Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., & Doebley, J. (2015). Introduction to genetic analysis (11th ed.). W. H. Freeman, New York.
- Nei, M. (1987). Molecular evolutionary genetics. Columbia University Press, New York.
- Wright, S. (1951). The genetical structure of populations. Annals of Eugenics, 15, 323–354.

Figures, Plates, Tables, and Boxes

- Figure 5.1: Evolutionary forces acting on a population gene pool. Attribution: AI generated. Suggested OER search string: “evolutionary forces allele frequency population genetics OER”.
- Box 5.1: Summary of key population genetics formulae. Attribution: AI generated. Suggested OER search string: “population genetics formulae Hardy-Weinberg FST table OER”.