

APPENDIX C

Glossary of Key Terms

- Admixture:** ancestry derived from two or more historically distinct source populations, carried within a single individual as a result of inter-population mating over recent generations (Chapter 14).
- Allele:** one of the alternative forms of a gene at a given locus.
- Assortative mating:** non-random mating based on phenotypic similarity (positive) or dissimilarity (negative) between mates.
- Burden test:** a rare-variant association test that collapses all qualifying rare variants within a gene into a single summary score per individual before testing for association (Chapter 20).
- Candidate-gene study:** an association study testing one or a small number of genes selected in advance for biological plausibility, prior linkage evidence, or previous literature reports (Chapter 12).
- Coalescence:** the event, traced backward in time, at which two gene-copy lineages are found to derive from a single common ancestral copy.
- Cryptic relatedness:** unrecorded close biological relatedness among individuals in a sample that is nominally

treated as consisting of unrelated individuals (Chapter 14).

Effective population size (N_e): the size of an idealised Wright–Fisher population that would experience the same rate of genetic drift as the actual population under study.

FBAT (Family-Based Association Test): a general statistical framework, of which the TDT is a special case, testing genetic association using within-family transmission information, robust to population stratification (Chapter 13).

Founder: an individual in a pedigree with no parents recorded, treated as unrelated to other founders unless otherwise specified (Chapter 6).

Genome-wide significance: the stringent p-value threshold (conventionally 5×10^{-8}) used in GWAS to correct for testing very large numbers of genetic markers simultaneously (Chapter 18).

Genomic control (λ_{GC}): the ratio of the median observed test statistic to its expected null value across many genome-wide markers, used to detect and correct for systematic inflation, often due to population stratification (Chapter 14, Chapter 18).

Genotype imputation: statistical estimation of unobserved genotypes at untyped markers, using linkage disequilibrium with genotyped markers and a reference haplotype panel (Chapter 19).

Hardy–Weinberg equilibrium: the stable genotype frequencies (p^2 , $2pq$, q^2) reached after one generation of random mating at a locus, absent mutation, migration, and selection.

Heritability (narrow-sense, h^2): the proportion of total phenotypic variance attributable to additive genetic variance.

- High-throughput (next-generation) sequencing:** technology that reads DNA sequence directly, base by base, enabling detection of novel and rare variants beyond the reach of genotyping arrays (Chapter 20).
- Identical by descent (IBD):** alleles that are copies of the same single ancestral allele, traceable through a known pedigree.
- Inbreeding coefficient (F):** the probability that an individual's two alleles at a locus are identical by descent.
- Kinship coefficient (ϕ):** the probability that a randomly chosen allele from one individual and a randomly chosen allele from another (or the same) individual, at the same locus, are identical by descent; $\phi = r/2$ (Chapter 6).
- Likelihood:** the probability of the observed data, viewed as a function of unknown model parameters.
- Linkage disequilibrium (LD):** non-random statistical association between alleles at different (often linked) loci.
- LOD score:** logarithm (base 10) of the likelihood ratio comparing a trial recombination fraction θ to $\theta = 0.5$ (no linkage).
- Missing heritability** — the gap between the heritability of a trait estimated from family/twin data and the (typically much smaller) heritability explained by genome-wide-significant variants identified by GWAS (Chapter 20).
- Penetrance:** the probability that an individual with a given genotype expresses the corresponding phenotype.
- Pedigree:** a diagram summarising the pattern of a trait's occurrence across the members of a family, over multiple generations, using standardised symbols (Chapter 6).

- Population stratification:** the presence of two or more genetically distinct subgroups within a sample, which can confound association tests if case/control ancestry proportions and tested-marker allele frequencies both differ across subgroups (Chapter 14).
- Principal component analysis (PCA), in genetics:** a statistical technique applied to genome-wide genotype data to extract the major axes of ancestry variation within a sample, used to detect and adjust for population stratification (Chapter 14).
- Proband (propositus):** the affected individual through whom a family is first identified for a genetic study.
- Rare variant:** conventionally, a genetic variant with minor allele frequency below approximately 1–5% in the population studied (Chapter 20).
- Recombination fraction (θ):** the probability that a gamete from a doubly heterozygous individual is recombinant at two loci.
- Reference panel:** a densely genotyped or sequenced set of haplotypes used as the basis for genotype imputation (Chapter 19).
- Segregation analysis:** the use of family data to test hypothesised modes of inheritance and estimate transmission parameters (Chapter 7).
- SKAT (Sequence Kernel Association Test):** a variance-component rare-variant association test, robust to a mixture of risk-increasing and risk-decreasing variants within a gene (Chapter 20).
- Tag SNP:** a SNP chosen to represent, via linkage disequilibrium, a set of other correlated SNPs, reducing the number of markers that must be directly genotyped (Chapter 12, Chapter 15).

TDT (Transmission Disequilibrium Test): a family-based association test comparing, among heterozygous parents of affected offspring, the number of times each allele is transmitted versus not transmitted (Chapter 13).

Variant Call Format (VCF): the standard file format for reporting genetic variants identified from sequencing data (Chapter 20).

Wahlund effect: the mechanical deficit of heterozygotes (relative to Hardy–Weinberg expectation computed from the pooled allele frequency) that results from combining two or more subpopulations with differing allele frequencies (Chapter 14).