AG Notebook

Contents

1 Classical Genetics (Transmission Genetics or Formal Genetics) 18
- 1.1 Foundations of Classical Genetics . . . 18
- 1.2 Mendel's Laws and Inheritance Principles . . . 19
  - 1.2.1 Law of Segregation (Monohybrid Crosses) . . . 19
  - 1.2.2 Law of Independent Assortment (Dihybrid Crosses) . . . 19
- 1.3 Pedigrees and PLINK . . . 19
  - 1.3.1 Pedigrees . . . 19
  - 1.3.2 PLINK . . . 20
- 1.4 Cytogenetics and Chromosome Mapping . . . 20
  - 1.4.1 Cytogenetics . . . 20
  - 1.4.2 Linkage Disequilibrium and Haplotypes . . . 20
  - 1.4.3 Sex Chromosomes . . . 20
  - 1.4.4 Chromosomal Maps . . . 21

2 Molecular Genetics 23
- 2.1 Technical Foundations . . . 23
  - 2.1.1 Sanger Sequencing (First-Generation) . . . 24

3 Population Genetics 26
- 3.1 Allele and Genotype Frequencies . . . 26
- 3.2 Hardy-Weinberg Equilibrium . . . 27
- 3.3 Connections to Genomics . . . 27

4 Quantitative Genetics 29
- 4.1 Phenotype Composition . . . 29
- 4.2 Phenotypic and Genetic Variance . . . 30
- 4.3 Applications of Quantitative Genetics . . . 30

5 Genomics 33
- 5.1 Definition and Scope . . . 33
- 5.2 Historical Background . . . 33
- 5.3 Types of Genomes . . . 34
- 5.4 Big Data and Bioinformatics . . . 34

6 Next-Generation Sequencing and Data Analysis 37
- 6.1 Next-Generation Sequencing Overview . . . 37
- 6.2 NGS Technologies and Workflow . . . 38
  - 6.2.1 Ion Torrent Technology . . . 38
  - 6.2.2 454 Roche . . . 38

6.2.3 Illumina . . . 38
6.3 Before Sequencing: DNA Quality and Project Setup . . . 39
6.3.1 DNA Quality Assessment . . . 39
6.4 NGS Variant Discovery Workflow . . . 40
6.4.1 Alignment Considerations . . . 40
6.4.2 FASTQ and Phred Scores . . . 40
6.5 Quality Control and Tools . . . 41

7 Applied Genomics – Prof. Bovo: NGS Data Analysis 44

8 Next Generation Sequencing Technologies and Advanced Applications (Prof. Fontanesi) 47

9 Variant Discovery and Genome Assembly Pipeline (Bovo Lessons 6–7) 51
9.1 Introduction . . . 51
9.2 Quality Control and Trimming . . . 51
9.3 Alignment and SAM/BAM Files . . . 51
9.4 Filtering and Duplicate Removal . . . 52
9.5 Variant Calling and VCF Files . . . 52
9.6 Variant Annotation and IGV Inspection . . . 52
9.7 Genome Assembly and Algorithms . . . 53

10 Applied Genomics and Population Genetics 58
10.1 Sequencing Strategies for Population Analysis . . . 58
10.1.1 Pool-Seq and Cost-Efficient Population Variability Studies . . . 58
10.1.2 Targeted Sequencing and AmpliSeq Panels . . . 59
10.1.3 Exome Sequencing and Hybrid Capture . . . 59
10.1.4 Epigenomic and Transcriptomic Applications . . . 59
10.2 Custom Genotyping and NGS-Based Approaches . . . 60
10.2.1 SNP Chips and Hardy-Weinberg Equilibrium . . . 60
10.2.2 NGS-Based Genotyping . . . 60
10.3 Copy Number Variations (CNVs) . . . 60
10.3.1 Definition and Biological Significance . . . 60
10.3.2 Detection Methods . . . 61
10.4 Population Genomics and Evolutionary Insights . . . 61
10.4.1 Core Concepts . . . 61
10.4.2 Runs of Homozygosity (ROH) . . . 61
10.4.3 Fixation Index and Differentiation . . . 61
10.5 Genome-Wide Association Studies (GWAS) . . . 62
10.5.1 Study Design and Requirements . . . 62
10.5.2 Association Testing and Multiple Testing . . . 62
10.5.3 Interpretation and Linkage Disequilibrium . . . 62

Glossary of Key Terms in Classical Genetics

Allele
Alternative form of a gene at a given locus.

Autosome
Any chromosome that is not a sex chromosome.

Carrier
Individual carrying one copy of a recessive or disease-causing allele, without manifesting the phenotype.

Chromosomal Map
Mathematical representation of gene positions along a chromosome, based on recombination frequencies.

Codominance
Allelic relationship where heterozygotes display a combined or intermediate phenotype.

Crossing Over
Exchange of DNA segments between paired homologous chromosomes during meiosis.

Diploid
Organism or cell with two sets of chromosomes, one maternal and one paternal.

Gamete
Haploid reproductive cell (e.g., sperm or egg) carrying one allele per locus.

Genotype
Complete allelic composition of an individual.

Haplotype
Combination of alleles at adjacent loci inherited together.

Haplodiploidy
Sex determination system where males are haploid (from unfertilized eggs) and females are diploid (from fertilized eggs).

Homozygous
Having two identical alleles for a given gene.

Heterozygous
Having two different alleles for a given gene.

Independent Assortment
Mendelian law stating that alleles at different loci segregate independently during gamete formation (if unlinked).

Linkage
Physical association of loci on the same chromosome.

Linkage Disequilibrium (LD)
Non-random association of alleles at different loci in a population.

Locus
Physical position of a gene on a chromosome.

Monohybrid Cross
Cross between individuals differing for a single trait.

Pedigree
Diagram showing inheritance patterns across generations using standard symbols.

Phenotype
Observable traits of an organism, resulting from genotype and environment.

Pseudo-Autosomal Region (PARR)
Small region of X and Y chromosomes that undergoes recombination.

Recombination Frequency
Probability that two loci are separated by crossing over in meiosis, often measured in centiMorgan (cM).

Segregation
Separation of alleles during gamete formation, as stated by Mendel's first law.

Sex Chromosomes
Chromosomes involved in sex determination (e.g., X/Y in mammals, Z/W in birds).

Alternative Splicing
Process by which different mRNA molecules are produced from the same pre-mRNA by including or excluding exons.

Allele Frequency
Proportion of a specific allele among all alleles for a given gene in a population.

DNA Library
Collection of DNA fragments stored in host cells or vectors for sequencing or analysis.

Gene Expression
Process by which information from a gene is used to synthesize functional gene products (RNA, protein).

Gene Regulation
Mechanisms that control when and how genes are expressed.

Genotype Frequency
Proportion of individuals with a specific genotype in a population.

Hardy-Weinberg Equilibrium
Idealized state in which allele and genotype frequencies remain constant from generation to generation under defined assumptions.

Mutation Rate
Frequency at which new mutations appear in a genome.

NGS (Next-Generation Sequencing)
High-throughput DNA sequencing technologies allowing parallel analysis of millions of fragments.

Post-Translational Modification
Chemical modifications of a protein after its synthesis, affecting activity and localization.

PCR (Polymerase Chain Reaction)
Technique to amplify specific DNA sequences in vitro.

Recombinant DNA
DNA molecule created by joining genetic material from multiple sources.

Sanger Sequencing
First-generation DNA sequencing method based on chain termination using modified nucleotides.

Selection
Evolutionary process where certain alleles increase in frequency due to reproductive advantage.

Transcription
Synthesis of RNA from a DNA template.

Translation
Process by which ribosomes synthesize proteins using mRNA as a template.

Additive Genetic Effect
Cumulative effect of individual alleles on a quantitative trait.

Complex Trait
Trait influenced by multiple genes and often by environmental factors.

Dominance Effect
Genetic effect resulting from interactions between alleles at the same locus.

Environmental Effect
Influence of environmental factors (permanent or temporary) on phenotype.

Genetic Variance (Var(G))
Portion of total phenotypic variance attributable to genetic differences.

Heritability
Proportion of phenotypic variance explained by genetic variance within a population.

Interactive Genetic Effect (Epistasis)
Interaction between alleles at different loci influencing a trait.

Permanent Environmental Effect
Long-lasting environmental influence on phenotype.

Phenotypic Variance (Var(P))
Total variability observed in a trait in a population.

Quantitative Trait
Measurable trait with continuous variation influenced by many genes.

Quantitative Genetics
Study of inheritance of complex traits affected by many genes and environment.

Temporary Environmental Effect
Short-term environmental influence on phenotype.

Applied Genomics
Use of genomic technologies and data to study populations, traits, and gene functions.

Archaea
Domain of single-celled organisms living in extreme environments; includes thermophiles, halophiles, and methanogens.

Chloroplast Genome
Circular DNA of plant chloroplasts; plants typically have three genomes (nuclear, mitochondrial, chloroplast).

Comparative Genomics
Study of similarities and differences among genomes of different species.

Domain of Life
Highest taxonomic category: Bacteria, Archaea, Eukaryota.

Exabyte / Zettabyte
Units of digital storage; used to describe the huge scale of genomic big data.

Functional Genomics
Field studying gene functions and interactions using genome-wide approaches.

GenBank
Public database for DNA sequences with annotations and metadata.

Genome
Entire genetic content of an organism (all DNA in a cell).

Metadata
Descriptive information accompanying genomic data (e.g., species, sex, population).

Mitochondrial Genome
Small circular genome present in mitochondria, usually around 60 kbp.

Omics
Collective fields of large-scale biological study (genomics, transcriptomics, proteomics, metabolomics, phenomics).

Prokaryotic Genome
Typically circular DNA with plasmids, lacking a defined nucleus.

Shotgun Sequencing (WGS)
Sequencing strategy where random fragments are sequenced and later assembled computationally.

Variant Calling
Computational process of identifying genetic variants (SNPs, indels) from sequence data.

Amplicon
DNA fragment produced by PCR amplification for sequencing.

Barcoding
Adding unique DNA sequences to identify sample origin in multiplexed sequencing.

Base Calling
Process of converting raw signal into nucleotide sequence.

BED file
Tab-delimited file describing genomic regions or features.

BWA-MEM
Widely used Burrows-Wheeler-based algorithm for read alignment.

Burrows-Wheeler Transform
Data structure enabling efficient short-read mapping.

Clonal Amplification
Replication of a DNA fragment to increase detectable signal.

CNV (Copy Number Variation)
Genomic region present in abnormal copy number.

Coverage / Depth
Number of times a nucleotide is sequenced; expressed as X-fold (10x, 30x...).

Electrophoresis
Technique to separate DNA fragments by size using an electric field.

Emulsion PCR (emPCR)
PCR performed in water droplets in oil to amplify DNA on beads.

FASTQ file
Stores sequences with corresponding per-base quality scores.

FASTQC
Software for evaluating sequence data quality.

FASTA file
Plain text file storing nucleotide or protein sequences without quality.

Flow Cell
Solid surface with attached oligos where Illumina DNA clusters form.

Homopolymer
Sequence region with repeated identical nucleotides; prone to errors in NGS.

Indel
Small insertion or deletion mutation in a sequence.

Ion Torrent
NGS platform detecting H^+ ions from nucleotide incorporation.

Key Sequence / k-mer
Short sequence used to check run quality or evaluate coverage.

Library Preparation
Process of fragmenting, barcoding, and adapting DNA for sequencing.

Mapping Quality (MQ)
Score indicating confidence that a read is aligned to the correct location.

Missense Variant
Single base change resulting in amino acid substitution.

Paired-End Sequencing
Sequencing both ends of DNA fragments to improve mapping accuracy.

Phred Score (Q)
Logarithmic score representing probability of base call error.

Polyclonal Reads
Reads generated from wells with multiple DNA templates causing ambiguity.

Pyrosequencing
Sequencing method detecting light emission from pyrophosphate reaction.

QC (Quality Control)
Evaluation of sequence data integrity and accuracy.

Read
Short DNA fragment output from an NGS run.

SAM/BAM file
Sequence Alignment Map; BAM is the binary, compressed version.

Shotgun Sequencing
Random fragmentation and sequencing to reconstruct whole genome.

SNP (Single Nucleotide Polymorphism)
Single base variation present in greater or equal to 1 percent of population.

Translocation
Structural variation where a DNA segment moves to another genomic location.

Variant Calling
Process of detecting sequence variations from aligned reads.

VCF file
Variant Call Format; standard format for genomic variants.

Depth of Coverage (X)
Average number of times a nucleotide is sequenced.

Breadth of Coverage
Percentage of the genome covered by reads.

FASTQ File
Standard NGS raw data format with sequences and quality scores.

Phred Score (Q)
Log-scaled probability that a base is called incorrectly.

Sliding Window Trimming
QC method removing low-quality bases across windows of reads.

GC Content
Proportion of guanine and cytosine bases in a DNA sequence.

Sequence Duplication Level
Number of times a specific sequence appears in the dataset.

Cluster Generation
Multiplication of DNA fragments on a flow cell to generate detectable signal in short-read sequencing (Illumina).

Barcoding
Addition of artificial sequences (tags) to DNA fragments to distinguish samples within the same sequencing run.

Barcode of Life
Project to classify species using specific mitochondrial DNA regions as natural sequence identifiers.

Paired-End Sequencing
Sequencing both ends of a DNA fragment to improve alignment and assembly of short reads.

Flow Cell
The surface used in Illumina sequencing for cluster generation and SBS chemistry.

4-/2-/1-Channel Chemistry
Illumina detection strategies: 4 colors for 4 bases, 2 colors (combinatorial), or 1 color (with chemical step separation).

Nanopore Sequencing
Long-read native DNA sequencing via electrical signal differences as DNA passes through a nanopore.

PacBio (SMRT)
Single-Molecule Real-Time sequencing technology generating accurate long reads after circular consensus correction.

Long-Read Sequencing
Sequencing of DNA fragments of several kb to Mb, improving assembly and detection of structural variants.

Structural Variation
Large genomic changes (insertions, deletions, inversions, translocations) detectable by long-read or paired-end sequencing.

Phase/Haplotype Resolution
Ability to assign variants to specific alleles using long reads.

FASTQC
Tool for NGS read quality control (per-base quality, GC content, duplication).

Trimming
Removal of low-quality bases from read ends.

Chimeric Read
Read resulting from the artificial fusion of two fragments during library preparation.

BWA / Bowtie
Alignment software for short reads.

SAM / BAM
Standard alignment formats; SAM is text-based, BAM is the compressed binary version.

Flag
Integer encoding the read status in SAM/BAM files.

CIGAR string
Compact string describing the read-to-genome alignment (`M, I, D, S`).

MAPQ
Mapping Quality, Phred-scaled score for alignment confidence.

PCR Duplicate
Read replicated during PCR, not providing new information.

VCF (Variant Call Format)
Standard file format for reporting sequence variants.

DP (Depth)
Number of reads supporting a position or variant.

AF (Allele Frequency)
Frequency of the alternative allele in the sample or population.

IGV
Integrative Genomics Viewer for visual inspection of BAM alignments and variants.

Contig
Continuous assembled sequence without gaps.

Scaffold
Ordered and oriented set of contigs, possibly separated by gaps.

N50
Length of the smallest contig in the set covering 50% of total assembly size.

BUSCO
Benchmarking single-copy orthologs, used for genome completeness assessment.

C-value
DNA content of a haploid genome (pg), used for genome size estimation.

K-mer
Subsequence of length K derived from reads.

Eulerian Path
Path in a graph that visits each edge exactly once (used in de Bruijn assembly).

Mate-Pair
Library where both ends of a long circularized fragment are sequenced.

Gap Size
Estimated length of an unsequenced region between contigs in a scaffold.

N50
Assembly metric representing the contig length at which 50% of the assembly length is covered.

Genome Coverage
Percentage of the expected genome size captured in the assembly.

Gene Coverage
Percentage of expected genes present in the assembly.

BUSCO
Benchmarking Universal Single-Copy Orthologs; tool for assessing assembly completeness.

ORF (Open Reading Frame)
Continuous stretch of codons from start to stop without termination.

Codon Bias
Unequal usage frequency of synonymous codons in coding sequences.

Exon–Intron Boundary
Junction between coding and non-coding regions recognized by splicing machinery.

Profile HMM
Hidden Markov Model capturing position-specific probabilities for sequence motifs.

GFF/GTF
Tabular formats describing genome features (coordinates, strand, attributes).

BED File
Simplified genomic interval format storing only chromosome, start, and end positions.

Long PCR
PCR strategy to amplify long DNA fragments (kb range), often for circular genomes.

Amplicon Library
Sequencing-ready library created from PCR-amplified DNA fragments.

mtDNA Haplotypes
Maternal lineages defined by mitochondrial DNA variants.

Pool-seq
Sequencing strategy in which DNA from multiple individuals is pooled prior to sequencing to estimate allele frequencies at the population level without maintaining individual identity.

AmpliSeq
Targeted amplicon sequencing platform using multiplex PCR to capture hundreds of predefined or custom genomic regions for variant detection.

Exome Sequencing
Selective sequencing of only the protein-coding regions (exons) of the genome, representing $\sim$2% of the human genome, to efficiently detect coding variants.

Bisulfite Conversion
Chemical treatment that converts unmethylated cytosines to uracil (read as thymine after PCR), while methylated cytosines remain unchanged; used for DNA methylation profiling.

ChIP-seq
Chromatin Immunoprecipitation followed by sequencing; maps protein-DNA interactions (e.g., transcription factor binding sites) across the genome.

Transcriptome
The complete set of RNA transcripts expressed in a cell, tissue, or organism at a specific time.

SNP Chip
DNA microarray containing thousands to millions of preselected SNPs for high-throughput genotyping.

Hardy-Weinberg Equilibrium (HWE)
Population genetics principle describing expected genotype frequencies under random mating: $p^2 + 2pq + q^2 = 1$.

CNV (Copy Number Variation)
Structural variant representing a DNA segment (usually ≥ 1 kb) present in variable copy number compared to the reference genome.

aCGH (Array Comparative Genomic Hybridization)
Microarray-based method for CNV detection by comparing sample and reference DNA hybridization intensities.

RAD-seq (Restriction-site Associated DNA sequencing)
Reduced-representation sequencing method based on restriction enzyme digestion for genotyping without requiring a reference genome.

Genetic Drift
Random fluctuations in allele frequencies across generations due to sampling effects in finite populations.

Bottleneck Effect
Reduction in population size leading to loss of genetic diversity and potential allele frequency shifts.

ROH (Run of Homozygosity)
Continuous stretch of homozygous genotypes across the genome, indicative of autozygosity or inbreeding.

$\mathbf{F}_{ROH}$
Genomic inbreeding coefficient estimated as the proportion of the autosomal genome contained within ROH segments.

$\mathbf{F}_{ST}$ Fixation Index; measures the degree of genetic differentiation between populations based on allele frequencies.

GWAS (Genome-Wide Association Study)
Large-scale analysis to identify statistical associations between genetic variants and phenotypic traits or diseases.

Manhattan Plot
Scatter plot of GWAS P-values across chromosomes, where peaks indicate loci associated with the phenotype.

Linkage Disequilibrium (LD)
Non-random association of alleles at different loci in a population.

Bonferroni Correction
Multiple testing correction dividing the significance threshold α by the number of independent tests.

False Discovery Rate (FDR)
Statistical method to control the expected proportion of false positives among declared significant results.

Key Formulas Across Genetics and Genomics

Allele Frequency

$$p = \frac{2 \cdot N_{AA} + N_{Aa}}{2N}, \quad q = 1 - p$$

- N_{AA} = homozygous dominant individuals
- N_{Aa} = heterozygotes
- N = total population size

Genotype Frequency

$$f(\text{genotype}) = \frac{\text{Number of individuals with genotype}}{\text{Total population}}$$

Hardy-Weinberg Equilibrium

$$p^2 + 2pq + q^2 = 1$$

Recombination Frequency (Morgan)

$$1 \text{ cM } = 1\% \text{ recombination } \approx 1/100 \text{ meioses}$$

Phenotypic Variance Decomposition

$$Var(P) = Var(G) + Var(E)$$

Detailed Phenotypic Variance

$$Var(P) = Var(A) + Var(D) + Var(I) + Var(E)$$

Heritability (Broad-Sense)

$$H^2 = \frac{Var(G)}{Var(P)}$$

Heritability (Narrow-Sense)

$$h^2 = \frac{Var(A)}{Var(P)}$$

Phenotype Equation

$$P = G + E$$

Sequencing Depth (Coverage)

$$X = \frac{L \times N}{G}$$

- L = average read length
- N = total number of reads

- G = haploid genome size

Breadth of Coverage

$$\text{Breadth} = \frac{\text{Covered bases}}{\text{Total target bases}} \times 100$$

Variant Allele Frequency (VAF)

$$\text{VAF} = \frac{\text{Reads supporting variant}}{\text{Total reads at locus}}$$

Phred Quality Score

$$Q = -10 \cdot \log_{10} P_{\text{error}}$$

- P_{error} = probability of base call error
- Higher $Q \rightarrow$ higher confidence

Average Read Quality (NGS-specific)

$$Q_{\text{avg}} = \frac{\sum Q_i}{\text{Read length}}$$

- Q_i = quality score of base i
- **Meaning:** Average confidence over the entire read

Genome Size from C-value

$$\text{Genome size (bp)} = \text{DNA content (pg)} \times 0.978 \times 10^9$$

- Conversion factor to approximate base pairs from picograms.

N50 Statistic

$N50 =$ length of the smallest contig such that $\sum$ contigs $\geq 50\%$ of the total assembly length

K-mer Genome Size Estimation

$$\text{Genome size} \approx \frac{\sum \text{K-mer counts}}{\text{Peak coverage depth}}$$

- Sum of K-mer counts = area under the K-mer frequency distribution
- Peak coverage = modal K-mer coverage (ignoring error peak)

Allele Balance in Heterozygotes

$$\text{AB} = \frac{\text{Alt reads}}{\text{Ref reads} + \text{Alt reads}}$$

GC Content

$$\text{GC\%} = \frac{G + C}{A + T + G + C} \times 100$$

Mapping Quality (Phred-scaled)

$$MAPQ = -10 \cdot \log_{10}(P_{\text{wrong mapping}})$$

N50 Calculation

$$\text{N50} = L_k \quad \text{where } \sum_{i=1}^{k} L_i \geq 0.5 \times L_{\text{total}}$$

Genome Coverage

$$\text{Coverage (\%)} = \frac{\text{Total Assembled Bases}}{\text{Expected Genome Size}} \times 100$$

Genome Size from C-Value

$$G_{\text{bp}} = \text{DNA content (pg)} \times 0.978 \times 10^9$$

Genome Size from K-mers

$$G_{\text{bp}} \approx \frac{\text{Total K-mer Count}}{\text{Average K-mer Depth}}$$

Expected Average Depth of Coverage

$$X = \frac{\text{Total Bases Sequenced}}{\text{Genome Size}}$$

Allele Frequency Difference for Pool-Seq

$$\Delta f = |f_{pop1} - f_{pop2}|$$

where f_{pop} = allele frequency in the pool of the population.

ROH-based Genomic Inbreeding Coefficient

$$F_{ROH} = \frac{\sum \text{Length of all ROHs (Mb)}}{\text{Total autosomal genome length (Mb)}}$$

Fixation Index (Population Differentiation)

$$F_{ST} = \frac{\text{Var between populations} - \text{Var within populations}}{\text{Var between populations}}$$

or equivalently:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

where H_T = total heterozygosity, H_S = subpopulation heterozygosity.

Bonferroni Threshold for GWAS

$$P_{threshold} = \frac{\alpha}{N_{tests}}$$

where α = significance level, N_{tests} = number of variants tested.

Hardy-Weinberg Expected Genotype Frequencies

$$\begin{aligned} f(AA) &= p^2, \\ f(Aa) &= 2pq, \\ f(aa) &= q^2 \end{aligned}$$

where $p + q = 1$.

Copy Number Log Ratio (aCGH or SNP Array)

$$\text{CNV signal} = \log_2 \frac{\text{Sample Intensity}}{\text{Reference Intensity}}$$

Linkage Disequilibrium (D)

$$D = P_{AB} - p_A p_B$$

where P_{AB} = frequency of haplotype AB, p_A and p_B = allele frequencies.

Manhattan Plot Significance

$$y = -\log_{10}(P_{association})$$

Used for visualizing GWAS peak significance.

Chapter 1

Classical Genetics (Transmission Genetics or Formal Genetics)

Classical genetics, also known as **transmission genetics** or **formal genetics**, is the study of how traits are transmitted from parents to offspring. It lays the foundation for understanding heredity and the behavior of genes within populations and across generations.

Historically, this field relied on **observable phenotypic differences** and **breeding experiments**, long before DNA was discovered. Today, even in the genomic era, the principles of classical genetics remain essential to interpret inheritance and plan genetic studies.

1.1 Foundations of Classical Genetics

Genes are arranged along **chromosomes**, which are the physical carriers of heredity. Each chromosome contains many **loci** (singular: locus), originally considered as units of inheritance of unknown molecular nature.

During **germline formation in diploid organisms**, maternal and paternal chromosomes pair and can exchange segments via **crossing over**. **The probability of recombination increases with genetic distance.**

Classical genetics studied this arrangement using *Drosophila melanogaster*, leading to the creation of the first **genetic maps** based on recombination frequencies.

Summary:

- Genes are located on chromosomes and passed to offspring.
- Crossing over reshuffles loci; distance influences recombination.
- Only loci with phenotypic variation were initially detectable.

1.2 Mendel's Laws and Inheritance Principles

1.2.1 Law of Segregation (Monohybrid Crosses)

Gregor Mendel discovered that traits are determined by discrete hereditary elements (now called **genes**) with **two alleles** in diploid organisms.

- **Genotype:** allele combination (invisible)
- **Phenotype:** observable trait, genotype + environment
- Alleles can be **dominant**, **recessive**, or **codominant**.

Example:

$$\text{P: } TT \times tt \quad \Rightarrow \quad \text{F1: all Tt (tall)}$$

$$\text{F2: 3 tall : 1 dwarf (phenotype), } 1:2:1 \text{ (genotype)}$$

Summary:

- Alleles segregate randomly during gamete formation.
- Monohybrid F2 ratio: phenotype 3:1, genotype 1:2:1.

1.2.2 Law of Independent Assortment (Dihybrid Crosses)

When considering two traits located on **different chromosomes**, alleles segregate independently:

$$\text{F2 phenotype ratio: } 9:3:3:1$$

Linked genes (on the same chromosome) deviate from this ratio, a principle later clarified by Morgan.

Summary:

- Independent assortment applies to unlinked genes.
- Classical 2-factor ratio: 9:3:3:1.

1.3 Pedigrees and PLINK

1.3.1 Pedigrees

Pedigrees are **diagrams** showing inheritance across generations:

- Circle = Female, Square = Male, Rhombus = Unknown
- Empty = Healthy, Filled = Affected, Half-filled = Carrier

Useful for detecting carriers and studying **inbreeding** (alleles **identical by descent**).

1.3.2 PLINK

PLINK is a software to handle **high-throughput genotyping data**. A basic .ped file contains:

1. Family ID
2. Individual ID
3. Paternal ID (0 if unknown)
4. Maternal ID (0 if unknown)
5. Sex (1=male, 2=female, 0=unknown)
6. Phenotype (1=control, 2=case, 0/-9=missing)

Additional columns store **alleles for each locus**, linkable to a chromosome map.

Summary:

- Pedigrees visualize inheritance.
- PLINK formalizes pedigree + genotype + phenotype for computation.

1.4 Cytogenetics and Chromosome Mapping

1.4.1 Cytogenetics

Study of **chromosomes under the microscope**, typically during metaphase. It provided the first **karyotypes and chromosome identification** before sequencing.

1.4.2 Linkage Disequilibrium and Haplotypes

Linkage disequilibrium (LD): non-random association of alleles at different loci. Physically close loci form **haplotypes** and are inherited together.

Centimorgan (cM):

- 1 cM = 1% recombination
- ≥ 50 cM = loci behave as unlinked

1.4.3 Sex Chromosomes

- **Mammals:** XX (female), XY (male), recombination only in Pseudo-Autosomal Regions (PARR)
- **Other systems:** XO (insects), ZW (birds), haplodiploidy (bees), environment-dependent (turtles, fish)

1.4.4 Chromosomal Maps

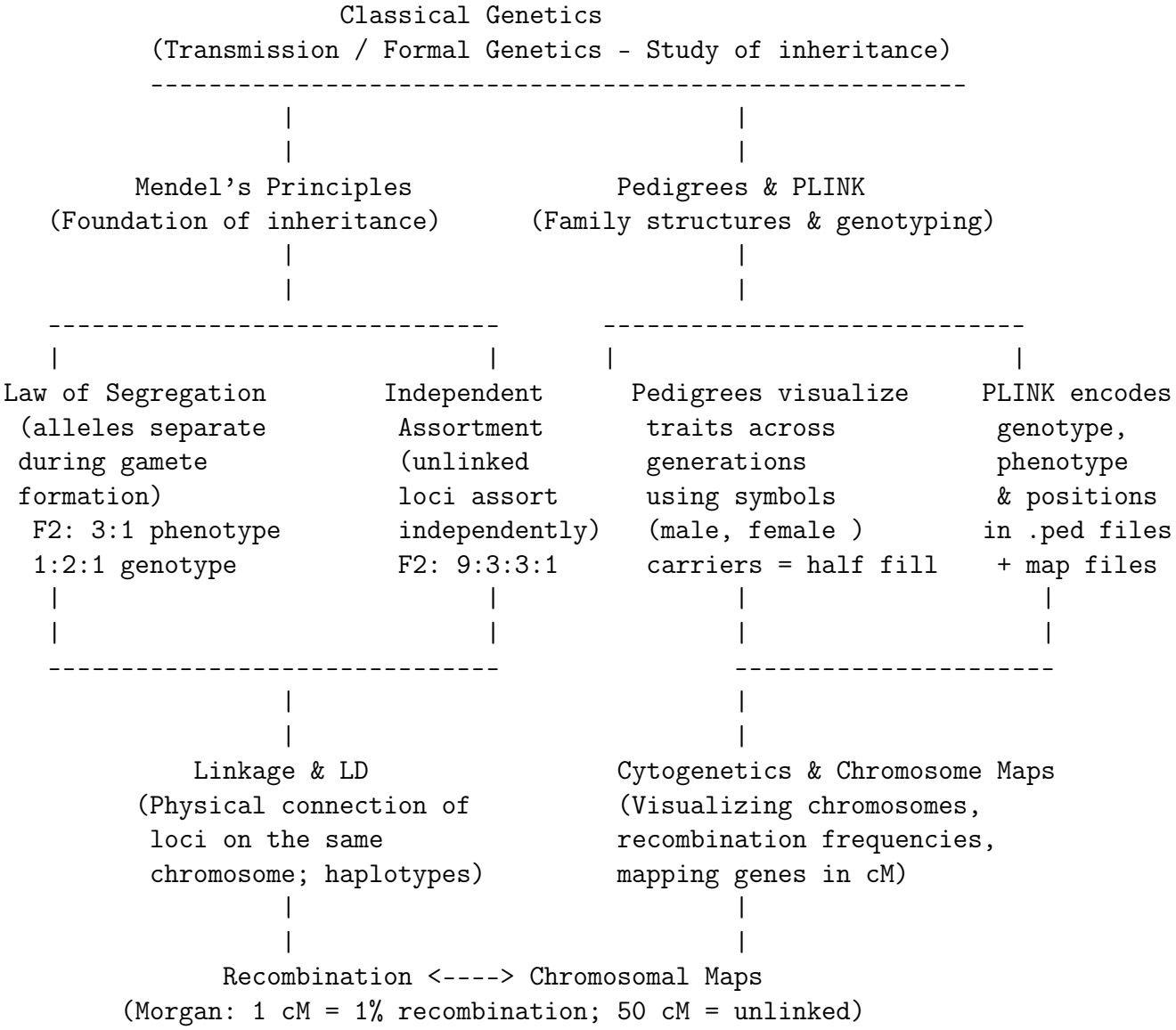
Maps built by summing **distances from recombination frequencies**. Close loci ⇒ low recombination, distant loci ⇒ high recombination.

Summary:

- LD shows non-random allele associations.
- Sex chromosomes have unique recombination patterns.
- Chromosome maps rely on recombination frequencies (Morgan).

==========

Classical Genetics Concept Map (Detailed Text Version)



Chapter 2

Molecular Genetics

Molecular genetics focuses on the **chemical nature of genes**, their structure, function, and the mechanisms by which genetic information is **encoded, replicated, expressed, and regulated**.

It explores:

- How genetic information flows: DNA → RNA → Protein
- Cellular processes: **Replication, Transcription, Translation**
- **Gene regulation**: controlling when and how genes are expressed
- **Functional diversity**: one gene can produce multiple proteins via alternative splicing and post-translational modifications

Summary:

- Molecular genetics = gene structure, function, and expression.
- Central dogma: DNA → RNA → Protein
- Proteome diversity enhanced by alternative splicing and modifications.

2.1 Technical Foundations

Modern molecular genetics relies on **laboratory techniques** that allow the manipulation and analysis of nucleic acids:

- **Recombinant DNA technology** – cloning and vector construction
- **DNA sequencing** – 1st generation (Sanger) and NGS
- **PCR amplification** – exponential in vitro DNA replication
- **Hybridization and gel electrophoresis** – fragment separation and detection

2.1.1 Sanger Sequencing (First-Generation)

Sanger sequencing is based on **chain-termination**:

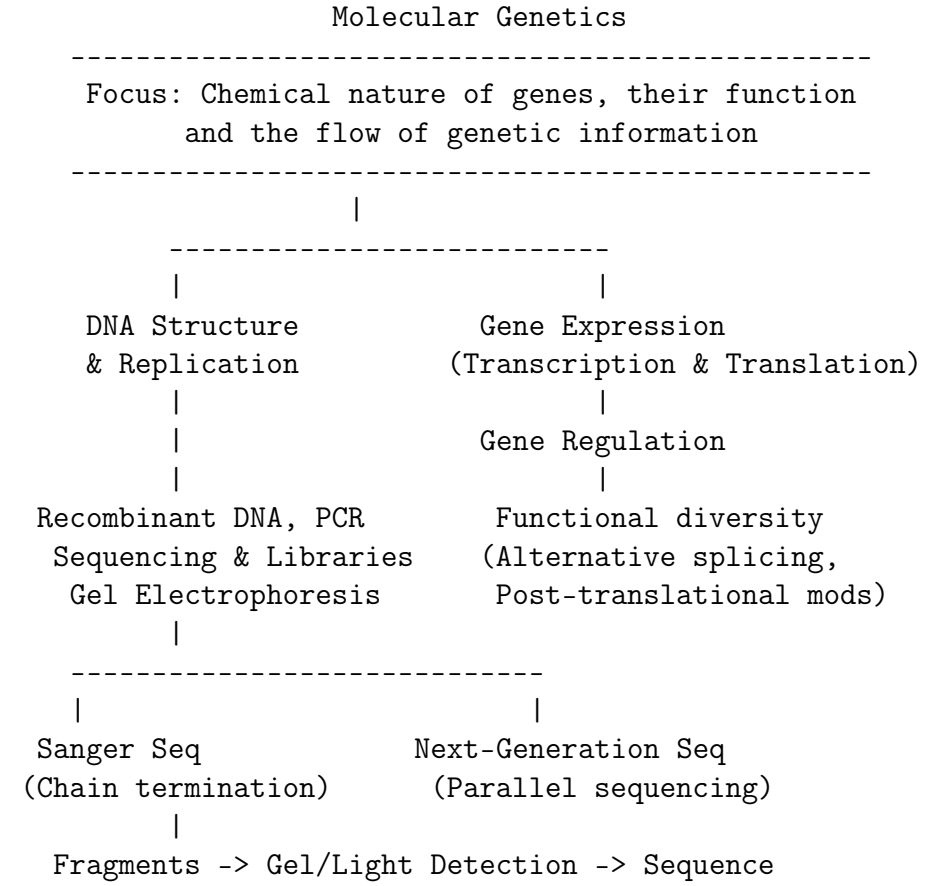
- Requires a **DNA template**, **primer**, **dNTPs**, **DNA polymerase** and **modified nucleotides** (ddNTPs) that terminate elongation.
- Linear amplification produces many fragments of different lengths.
- Fragments are **separated by electrophoresis** and detected by **color or fluorescence**.
- Sequence is deduced from the **order of terminated fragments**.

Heterozygous diploid sequences can produce **overlapping peaks** at variable positions because two alleles differ at that nucleotide.

Summary:

- Molecular genetics uses recombinant DNA, PCR, sequencing.
- Sanger sequencing: chain-termination + electrophoresis = sequence.
- NGS allows high-throughput, parallelized sequencing.

Molecular Genetics Concept Map (Text Version)



Key Takeaways:

- Central Dogma: DNA -> RNA -> Protein
- Sequencing allows direct gene analysis
- Molecular tools enable applied genomics

Chapter 3

Population Genetics

Population genetics studies the **genetic composition of populations** and how it changes over time or across geographic regions. It connects **inheritance principles** with **evolutionary dynamics**.

Key concepts:

- Populations consist of **gene pools** with allele diversity.
- **Allele and genotype frequencies** describe population structure.
- Changes in these frequencies reflect **evolutionary forces**.

3.1 Allele and Genotype Frequencies

$$\text{Allele frequency} = \frac{\#\text{ copies of an allele}}{\text{total alleles in the population}}$$

$$\text{Genotype frequency} = \frac{\#\text{ individuals with a genotype}}{\text{total individuals}}$$

Example:

- AA: $8/21 \approx 0.38$
- AB: $6/21 \approx 0.28$
- BB: $7/21 \approx 0.33$
- Allele A: $\frac{8*2+6}{21*2} = 0.52$, Allele $B = 0.48$

Summary:

- Allele frequency $\rightarrow$ fraction of alleles
- Genotype frequency $\rightarrow$ fraction of individuals

3.2 Hardy-Weinberg Equilibrium

A population is in **Hardy-Weinberg equilibrium (HWE)** when allele and genotype frequencies remain stable across generations if:

1. Organisms are **diploid and reproduce sexually**
2. Generations are **non-overlapping**
3. Mating is **random**
4. Population is **infinitely large** (no drift)
5. No **migration, mutation, or selection**
6. Allele frequencies are equal in both sexes

For a **bi-allelic locus** (alleles A and B):

$$p = \text{frequency of } A, \quad q = \text{frequency of } B$$

$$p + q = 1 \quad \Rightarrow \quad p^2 + 2pq + q^2 = 1$$

- p^2: expected frequency of AA
- $2pq$: expected frequency of AB
- q^2: expected frequency of BB

Deviations from HWE indicate **small population size, inbreeding, or selection**.

Summary:

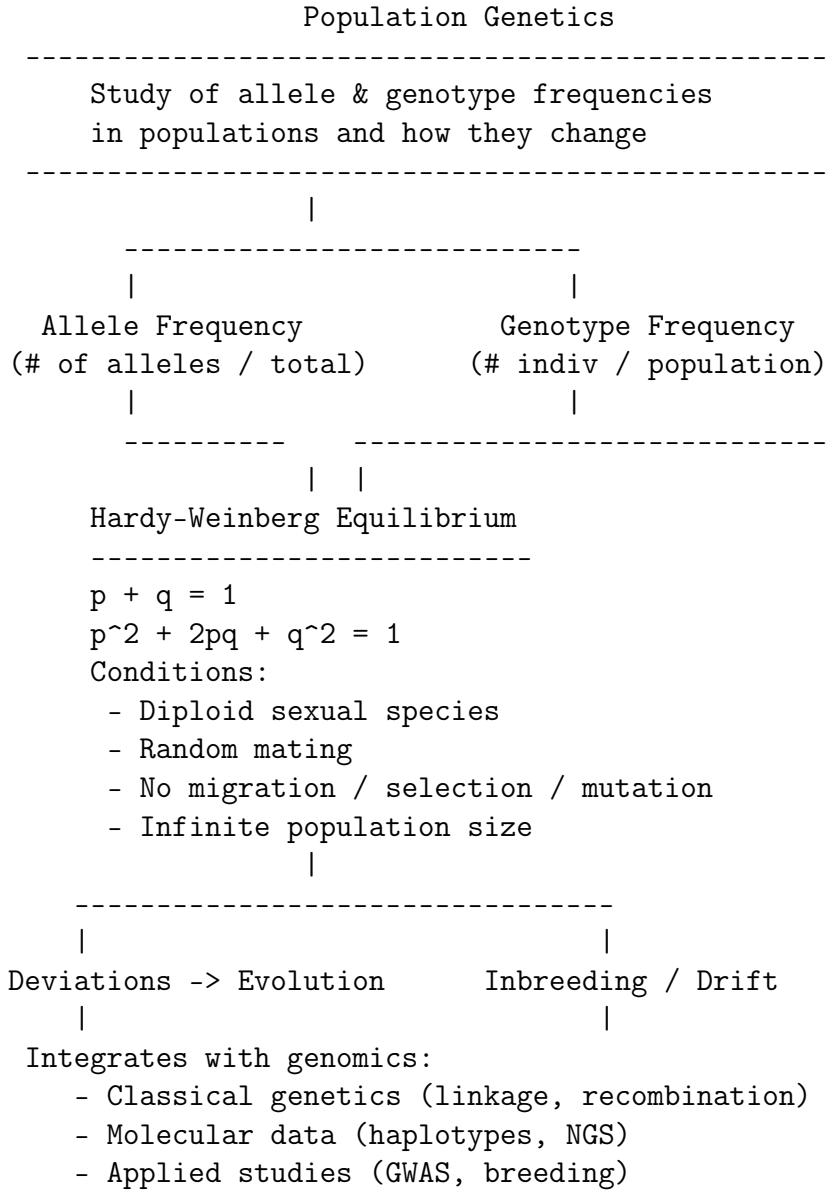
- HWE provides a **null model** for allele/genotype frequencies.
- $p^2 + 2pq + q^2 = 1$ for two-allele systems.
- Deviations suggest **evolutionary or demographic processes**.

3.3 Connections to Genomics

Population genetics integrates:

- **Classical genetics**: linkage and recombination
- **Molecular data**: haplotypes, allele frequencies from sequencing
- **Applied genomics**: GWAS, evolutionary inference, breeding strategies

Population Genetics Concept Map (Text Version)



Chapter 4

Quantitative Genetics

Quantitative genetics, also known as **genetics of complex traits**, studies traits influenced by **many genes and environmental factors**. Examples include:

- Stature (height)
- Eye color
- Litter size or survival (binary traits with polygenic basis)
- Newborn weight or adult weight

4.1 Phenotype Composition

The phenotype (P) of an individual is the result of:

$$P = G + E$$

Where:

- G: **Genetic effect** (heritable component)
- E: **Environmental effect** (permanent + temporary environmental factors)

The genetic effect can be subdivided into:

$$G = A + D + I$$

Where:

- A: **Additive effect** – sum of individual allele contributions
- D: **Dominance effect** – interaction of alleles at the same locus
- I: **Interaction (epistatic) effect** – interaction of alleles across loci

Summary:

- Quantitative traits are influenced by multiple genes and environment.

- Phenotype = Genotype + Environment.
- Genetic variance can be decomposed into additive, dominance, and interaction components.

4.2 Phenotypic and Genetic Variance

The total variance of a trait is:

$$Var(P) = Var(G) + Var(E)$$

Heritability (h^2) measures the proportion of phenotypic variance explained by genetic factors:

$$h^2 = \frac{Var(G)}{Var(P)}$$

- $h^2 < 0.1$: Low heritability
- $0.1 \leq h^2 \leq 0.4$: Medium heritability
- $h^2 > 0.4$: High heritability

Heritability is only meaningful if we can assess **genetic relationships between individuals**. Otherwise, the variance appears to be purely environmental.

Summary:

- Phenotypic variance is the sum of genetic and environmental variance.
- Heritability quantifies how much of phenotype is explained by genetics.
- Medium to high heritability implies potential for selection or improvement.

4.3 Applications of Quantitative Genetics

Quantitative genetics is crucial in:

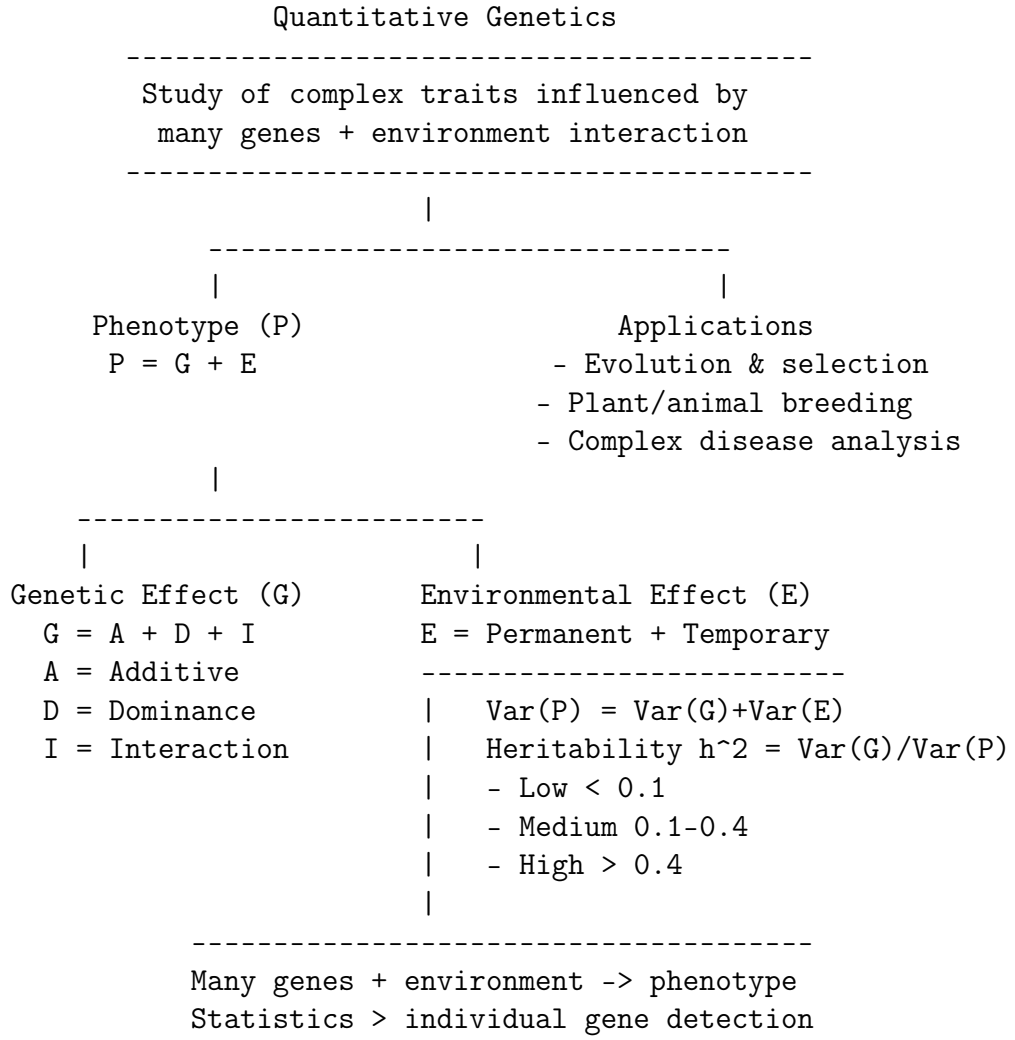
- **Evolutionary biology**: Understanding variation among relatives, trait heritability, and population response to selection.
- **Animal and plant breeding**: Predicting genetic gain and improving complex traits.
- **Complex disease studies**: Identifying heritable components in multifactorial disorders.

Founders of the discipline recognized that, for polygenic traits, it is **impractical to identify the effect of individual genes**, hence the focus on statistical models and heritability.

Summary:

- Quantitative genetics links genes, environment, and trait variability.
- Essential for breeding, evolutionary studies, and understanding complex disease.
- Individual gene effects are rarely detectable; the model is statistical.

Quantitative Genetics Concept Map (Text Version)



Chapter 5

Genomics

5.1 Definition and Scope

Genomics is the study of **genomes**, the complete set of genetic material of an organism. Originally defined by Hans Winkler in 1920 as the **collection of genes in a haploid set of chromosomes**, today a genome includes **all DNA in a cell**.

In 1986, Thomas Roderick introduced the term **genomics** to describe:

- Mapping, sequencing, and characterizing genomes
- Functional genomics, transcriptomics, proteomics, metabolomics, and phenomics (collectively, **Omics**)

Genomics is a **marriage of genetics, molecular biology, robotics, and computing**, enabling large-scale analysis of genetic information.

Summary:

- Genome = entire genetic content of an organism.
- Genomics studies structure, function, and comparison of genomes.
- Omics approaches (functional, transcriptomic, proteomic, etc.) expand the field.

5.2 Historical Background

The field grew rapidly with the **Human Genome Project** (HGP):

- Goal: sequence **3 billion base pairs** of the human genome.
- Required **automated sequencing and new technologies**.
- Led to **comparative genomics**, sequencing simpler genomes (bacteria, yeast) first.

Applications:

- Identify genes affecting phenotypes

- Characterize population structures
- Compare annotated and unannotated genomes

Summary:

- HGP triggered high-throughput sequencing and big data era.
- Comparative genomics compares functions across species.

5.3 Types of Genomes

Genomes vary by organism:

1. **Viral genomes**: DNA or RNA, sometimes segmented.
2. **Prokaryotic genomes**: circular pseudonuclear genome with plasmids; no fixed reference.
3. **Archaea and Monera**: single-celled, some extreme-environment specialists.
4. **Eukaryotic genomes**:
    - Nuclear genome: in most cells (except RBCs, etc.)
    - Mitochondrial genome: small circular DNA (∼60 kbp)
    - Chloroplast genome: present in plants (plants have three genomes)

Domains of life:

- **Bacteria**
- **Archaea** (thermophiles, halophiles, methanogens)
- **Eukaryota**

Summary:

- Organisms may have multiple genomes (nuclear, organelle).
- Understanding genome type is crucial for sequencing and annotation.

5.4 Big Data and Bioinformatics

Modern genomics produces **astronomical amounts of data**:

- Units: Terabyte → Petabyte → Exabyte → Zettabyte

- **Variant calling** and **assembly** require large-scale computation.
- Often, **data mining** is more valuable than producing new data.

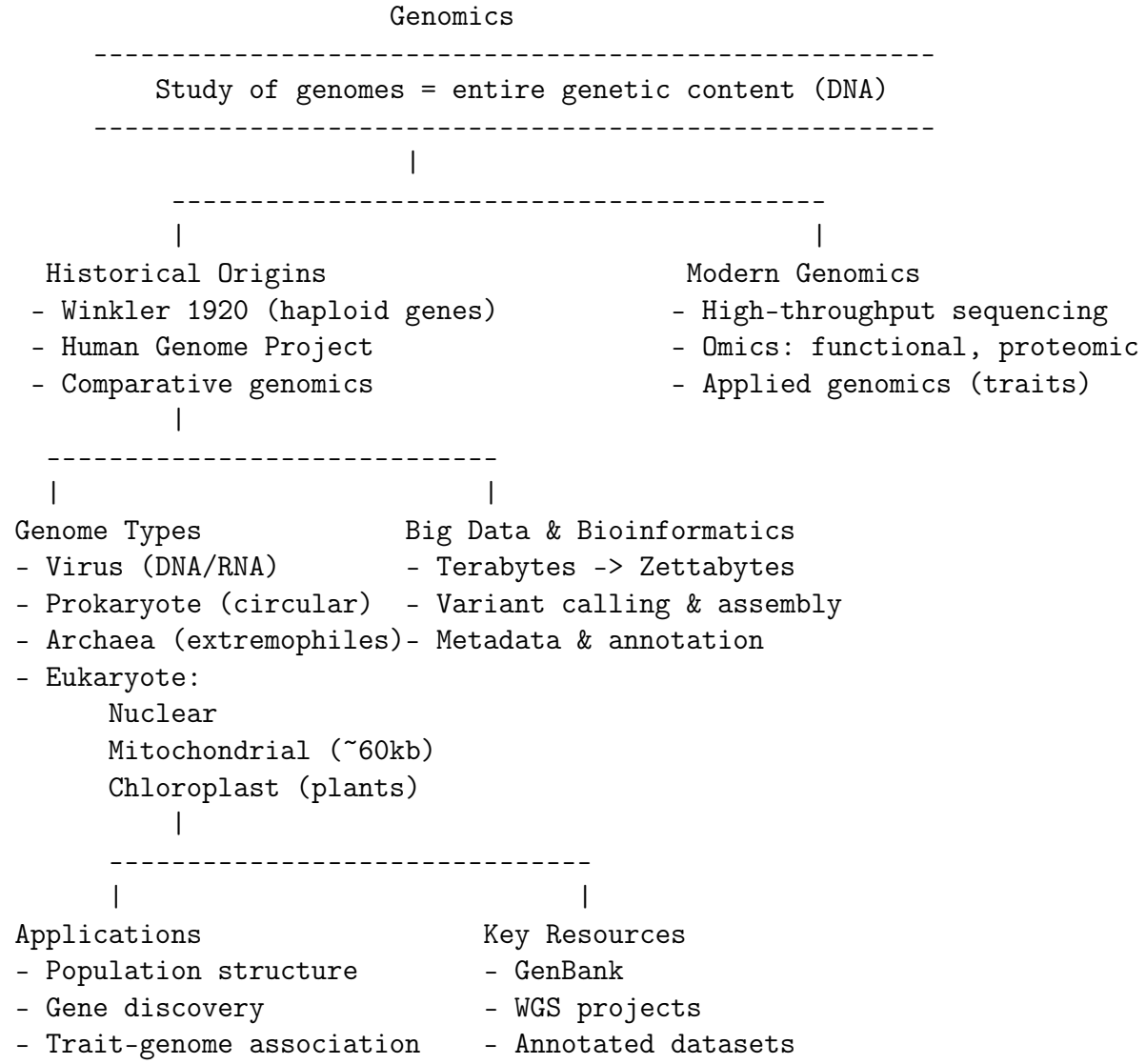
GenBank stores sequence data with:

- Number of nucleotides
- Annotation and metadata (species, origin, authors)
- Project type: **WGS** (whole genome shotgun) or complete

Summary:

- Genomics relies on sequencing + big data + computing.
- Metadata is critical for interpretation and comparative analysis.
- WGS and variant calling are core pipelines in applied genomics.

Genomics Concept Map (Text Version)



Chapter 6

Next-Generation Sequencing and Data Analysis

6.1 Next-Generation Sequencing Overview

Next-Generation Sequencing (NGS) represents a paradigm shift in genomics:

- From single DNA target sequencing (Sanger)
- To **massively parallel sequencing** of millions of fragments

Sanger Sequencing:

- Single amplicon per lane (max ∼48 lanes/run)
- High cost per base and low throughput
- Sequence content known in advance (targeted PCR)

NGS:

- Millions of sequences simultaneously (short reads)
- Unknown sequences are discovered after analysis
- Cost per Mb $<$ 0.01 USD; complete human genome $<$ 1000 USD
- Broke **Moore's Law** in cost reduction and throughput

Experimental design now focuses on:

1. **Planning** (most critical)
2. **Data production** (cheap and fast)
3. **Data analysis** (time-consuming)

Summary:

- Sanger = slow, expensive, known targets.
- NGS = high throughput, low cost, unknown targets.
- Main effort shifts to planning and analysis.

6.2 NGS Technologies and Workflow

NGS requires:

1. **Library Preparation**: DNA fragmentation (sonication or enzymatic), adapter ligation, barcoding.
2. **Template Amplification**: Clonal amplification via emulsion PCR or bridge amplification.
3. **Sequencing Reaction**: Signals generated per nucleotide incorporation.
4. **Signal Detection**: Ion current, light emission, or chemical detection.

6.2.1 Ion Torrent Technology

- Sequencing on **chips with millions of nano-wells**.
- Incorporation of nucleotide releases H^+ detected as pH change $\rightarrow$ converted to voltage.
- Requires **clonal fragments** for detectable signal.
- **Homopolymer regions** introduce errors (signal not linear with repeats).

Workflow:

1. DNA fragmentation $\rightarrow$ library with adapters
2. Barcoding for multiplexing up to 300 samples
3. Emulsion PCR to amplify single fragments on beads
4. Sequencing-by-synthesis with programmed **flow** of nucleotides (TACG...)
5. Data analysis: base calling, filtering, alignment

6.2.2 454 Roche

- Similar to Ion Torrent, but detects **pyrophosphate** and uses **luciferase/light capture**.
- Longer reads than Ion Torrent but higher complexity and cost.

6.2.3 Illumina

- Uses **bridge amplification** on flow cells to form clonal clusters.
- Sequencing-by-synthesis with **fluorescently labeled, reversible terminators**.
- Lower error rates than Ion Torrent; slower due to stop-and-read cycles.
- Flow cells: new generations have **mapped clusters** to improve signal separation.

Summary:

- Ion Torrent: detects protons, cheap, errors in homopolymers.
- 454 Roche: light-based detection, longer reads, expensive.
- Illumina: fluorescence-based, low error, standard for genomics.

6.3 Before Sequencing: DNA Quality and Project Setup

Before any bioinformatic analysis, a bioinformatician must understand:

- **Project design and goals**
- **Sequencing technology used** (Illumina, Ion Torrent...)
- **Library preparation type** (PCR-free, barcoded)
- **Expected coverage and depth of sequencing**

6.3.1 DNA Quality Assessment

Key parameters:

- **260/280 ratio** ~ 1.8: indicates pure DNA
- **260/230 ratio** $\sim 2.0 - 2.2$: absence of contaminants
- **Electrophoresis integrity**: thin upper band = good; smear = degraded DNA

Coverage metrics:

$$\text{Depth (X)} = \frac{L \times N}{G}$$

- L: read length
- N: number of reads
- G: haploid genome size

Summary:

- DNA quality affects library prep and read length.
- Coverage and depth determine variant detection power.

6.4 NGS Variant Discovery Workflow

Typical pipeline:

1. **Sequencing** → FASTQ files
2. **Read Alignment** to reference genome → BAM files
3. **Variant Calling** → VCF files (SNPs, Indels)
4. **Variant Annotation** → functional impact

Variants include:

- **SNPs**: single-base polymorphisms
- **Indels**: small insertions or deletions
- **CNVs**: copy number variations
- **Inversions and translocations**: structural variants

6.4.1 Alignment Considerations

Aligners:

- BWA-MEM (Burrows-Wheeler Transform-based)
- Bowtie2 (hash-based)

Mapping Quality (MQ):

- Probability that a read is correctly aligned
- Low MQ → potential false positives

6.4.2 FASTQ and Phred Scores

FASTQ format:

1. @identifier
2. sequence (ACGT...)
3. +
4. quality string (ASCII-encoded Phred scores)

$$Q = -10 \log_{10} P_{error}$$

- $Q20$ = 1% error
- $Q30$ = 0.1% error

6.5 Quality Control and Tools

Essential QC steps:

- **FASTQC**: per-base and per-read quality
- **Read filtering and trimming**
- **Check duplicates and polyclonal reads**
- **Coverage evaluation**: breadth and depth

Paired-end sequencing produces two FASTQ files (/1 and /2). QC ensures correct read orientation and removal of adapters or low-quality regions.

Summary:

- Variant discovery = alignment + variant calling + annotation
- Phred score quantifies confidence in base calls
- QC prevents propagation of errors to variant discovery

Next-Generation Sequencing and Data Analysis Concept Map (Text Version)

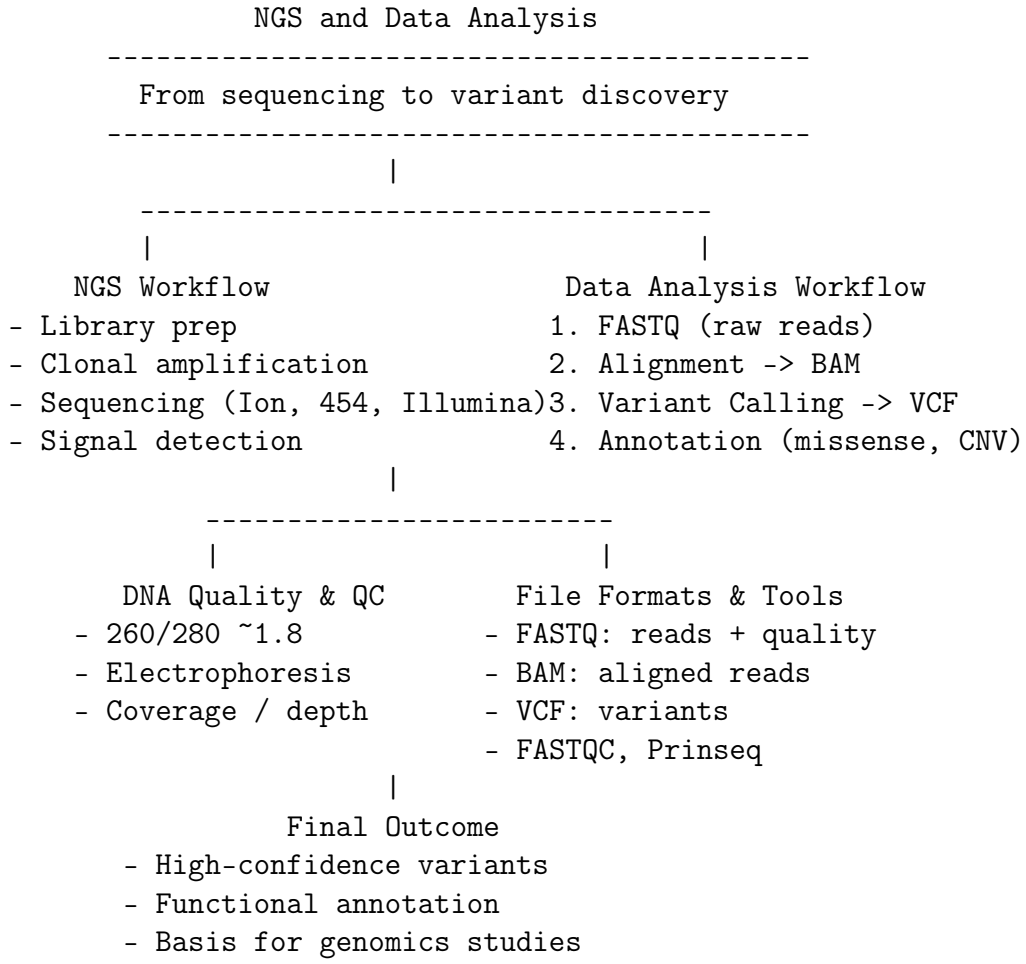


Table 6.1: Comparison of main sequencing technologies (compact)

Feature	**Sanger**	**Ion Torrent**	**454**	**Illumina**
Detection	Chain term.	pH / H^+	Light (PPi)	Fluorescence
Read length	500–1000 bp	200–400 bp	400–700 bp	75–300 bp
Throughput	Very low	Medium	Medium	Very high
Main error	None (targeted)	Homopolymers	Homopolymers	Random, low
Run time	Hours	1–3 h	∼10 h	12–48 h
Cost/Mb	> \$500	Low	Medium	< \$0.01
Typical use	Target seq.	Amplicons	Microbiology	WGS, RNAseq

Table 6.2: Common NGS file formats and usage

Format	**Content**	**Use**
FASTQ	Reads + quality	Raw data from sequencer
FASTA	Sequences only	Reference genomes
SAM/BAM	Reads + alignments	Mapping to reference
VCF	Variants	SNP/Indel annotation
BED	Genomic intervals	Visualization / tracks

Table 6.3: Phred quality scores and base call accuracy

Q Score	**Error probability**	**Accuracy (%)**
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

Chapter 7

Applied Genomics – Prof. Bovo: NGS Data Analysis

It is not recommended to start working directly on your data without understanding how they were produced. Knowing the laboratory workflow allows a better choice of tools and correct interpretation of results.

Basic Requirements

- **Good DNA quality is essential:** correct concentration, purity, and integrity.
- **Purity check:** performed using the 260/280 absorbance ratio ($\sim$1.8 indicates good DNA quality).
- **Integrity check:** usually performed with electrophoresis to verify fragment length.

Coverage Metrics

- **Depth of Coverage:** Average number of times a single nucleotide is covered by reads.
- **Breadth of Coverage:** Percentage of the reference genome covered by reads.

Different sequencing technologies provide different outputs in terms of total sequenced bases, which impacts coverage.

Sequencing the human genome ($\sim 3\,Gb$) with an Ion 110 chip ($\sim 1\,Gb$ output) results in depth $< 1x$, meaning that on average each base is covered less than once.

Variant Discovery and Consequences

- Variants can occur in exons, introns, regulatory regions, or intergenic regions.
- Variants in exons often affect proteins directly, potentially altering:
  - Active/binding sites
  - Transmembrane domains

  - Protein-protein interaction domains

- Variants in regulatory regions can change transcription factor binding and gene expression.

Standardized **ontologies** exist to describe the effects of variants, allowing easier annotation and sharing.

NGS Basic Pipeline

1. **Sequencing** → generation of raw reads (FASTQ format)
2. **Mapping / Alignment** → reads aligned to a reference genome (BAM file)
3. **Variant Calling** → SNPs, indels (VCF file)
4. **Variant Annotation** → classification by location and effect

The **GATK Best Practices** provide guidelines to ensure reliable variant discovery.

FASTQ Format and Quality Scores

FASTQ files consist of four lines per read:

1. @ Read identifier
2. Nucleotide sequence
3. + Optional description
4. ASCII-encoded Phred quality scores (one per base)

The **Phred score** is calculated as:

$$Q = -10 \cdot \log_{10} P_{\text{error}}$$

A score of 20 corresponds to 99% accuracy per base.

Quality Control and Filtering

Quality assessment ensures reliable downstream analysis:

- **Per-base quality plots:** Check read quality along positions.
- **Per-sequence quality plots:** Assess average quality per read.
- **GC content:** Detect contamination or biases in library preparation.
- **Sequence duplication:** Identify PCR artifacts or over-represented fragments.

Trimming: Remove low-quality bases using a threshold (e.g., Q20) or a sliding window approach to improve dataset quality.

Importance of Quality Filtering

- Retaining low-quality reads increases false positive variant calls.
- Removing duplicates and low-quality reads ensures high-confidence SNP discovery.

Software like **FASTQC** and **Prinseq** is commonly used for QC and trimming.

NGS Data Analysis Workflow (Prof. Bovo)

DNA Extraction → Library Preparation → Sequencing (FASTQ) → Quality Control (FASTQC) → Trimming / Filtering → Alignment (BAM) → Variant Calling (VCF) → Variant Annotation → Biological Interpretation

Chapter 8

Next Generation Sequencing Technologies and Advanced Applications (Prof. Fontanesi)

Introduction

In the previous lessons, we explored the foundations of Next Generation Sequencing (NGS) technologies, focusing on their principles and applications. NGS technologies have revolutionized genomics by enabling high-throughput, cost-effective, and precise DNA sequencing. In this chapter, we will recap the main sequencing platforms, their characteristics, and their specific advantages and limitations, with a particular focus on **Illumina**, **Ion Torrent**, **Nanopore**, and **PacBio** technologies.

NGS platforms share a common principle: sequencing is performed by reconstructing a DNA sequence from smaller units, using signals generated either by the incorporation of nucleotides (sequencing-by-synthesis) or by detecting the native DNA passing through sensors. Each technology differs in chemistry, detection method, read length, and error profiles, which determine its optimal applications.

Ion Torrent Technology

Ion Torrent sequencing is based on the detection of **ions** produced during DNA polymerization. When a nucleotide is incorporated into the growing complementary DNA strand, a **proton (H^+)** is released. The system measures this change in pH within millions of tiny wells on a silicon chip, converting it into an electrical signal.

- **Strengths:** Fast and cost-effective sequencing without labeled nucleotides.
- **Weaknesses:** Vulnerable to errors in **homopolymeric regions**, where multiple identical nucleotides in a row (e.g., AAAA) generate ambiguous signals.

Illumina Sequencing: Short-Read Sequencing

Illumina has become the **dominant short-read sequencing technology**. It relies on **Sequencing by Synthesis (SBS)** using **fluorescently labeled nucleotides**. Each incorporated nucleotide emits light, which is detected by high-resolution cameras.

Cluster Generation and Barcoding

Before sequencing, DNA fragments are **engineered with adapters** to allow:

1. **Attachment to the flow cell** (surface coated with complementary oligos),
2. **Cluster generation**, where fragments are amplified to enhance signal detection,
3. **Barcoding**, the addition of unique sequence tags to distinguish samples in multiplex runs.

Illumina Chemistry (1-, 2-, 4-Channel)

- **4-channel:** Each nucleotide has a unique color. Highest accuracy, but slower and more expensive.
- **2-channel:** Uses only two colors (Red = C, Green = T, Yellow = A, no signal = G). Faster and cheaper.
- **1-channel:** All nucleotides detected with one color using sequential chemical steps. Reduces cost but increases chemistry complexity.

Short-read sequencing is highly accurate and works well on **degraded DNA** and for **high-throughput applications**, but aligning and assembling short reads can be challenging, especially in repetitive genomes.

Paired-End Sequencing

Paired-end sequencing reads both ends of a DNA fragment. This approach improves:

- **Genome assembly**, by bridging repetitive or complex regions.
- **Structural variant detection**, by linking distant genomic locations.

For example, a 1000 bp fragment can yield two 200 bp reads from the ends, leaving a central gap. This linkage allows bioinformaticians to reconstruct scaffolds and fill gaps in de novo genome assembly.

Long-Read Sequencing Technologies

Nanopore Sequencing

Nanopore sequencing differs fundamentally from SBS approaches:

- **Native single-strand DNA** passes through a **protein nanopore** embedded in a membrane.
- **Voltage changes** caused by nucleotides blocking the pore are measured as electric signals.
- Long reads up to **10-20 kb routinely, with potential for Mb-sized fragments**.

- **Challenges:** High error rate (5–10%) and interpretation complexity.

Nanopore devices (e.g., MinION) are **portable and low-cost**, enabling **on-site sequencing** and compliance with the **Nagoya protocol** by sequencing DNA in the country of origin without sample export.

PacBio (SMRT Sequencing)

Pacific Biosciences developed **Single Molecule Real-Time (SMRT)** sequencing:

- DNA is **circularized**, allowing the polymerase to repeatedly sequence the same molecule.
- Repeated passes generate a **consensus sequence** that corrects the originally high random error rate (20%).
- Produces **highly accurate long reads** (*HiFi reads*) suitable for structural variant detection and phasing.

Advantages of Long-Read Technologies

- Improved genome assembly in **repetitive or complex genomes**.
- Detection of **structural variants** and **phased haplotypes**.
- Reduced need for computational scaffolding compared to short-read assemblies.

Applications and Strategic Considerations

Choice of sequencing technology depends on:

1. **Sample quality** (long DNA required for long-read sequencing),
2. **Project aim** (variant calling vs. de novo assembly),
3. **Budget and throughput requirements**.

In practice:

- Illumina remains the **workhorse for high-throughput short-read sequencing**.
- Ion Torrent is **cost-effective but limited by homopolymers**.
- Nanopore and PacBio **unlock structural and phasing information with long reads**, at the cost of lower throughput.

Comparison of NGS Platforms

Table 8.1: Comparison of Major NGS Technologies (Prof. Fontanesi)

Platform	Read Length	Error Rate	Throughput	Key Notes
Illumina	150-300 bp	∼0.1%	Very High	Short reads; SBS; high accuracy
Ion Torrent	200-400 bp	1-2%	Medium	pH detection; homopolymer errors
454 Roche	400-700 bp	1%	Low	Light detection; obsolete
Nanopore	10 kb-1 Mb	5-10%	Medium	Native DNA; portable; long reads
PacBio	10-20 kb	∼1% (HiFi)	Medium	SMRT; accurate long reads

Table 8.2: Strengths and Weaknesses of Major NGS Technologies

Platform	Strengths	Weaknesses
Illumina	High accuracy; High throughput; Cost-effective for large projects; Works with degraded DNA	Short reads; More difficult genome assembly; Requires expensive infrastructure
Ion Torrent	Simple chemistry; Fast runs; Medium cost	Homopolymer errors; Moderate throughput
Nanopore	Portable; Real-time sequencing; Ultra-long reads (structural variants, phasing)	Higher error rate; Lower throughput; DNA quality critical
PacBio	Accurate long reads (HiFi); Excellent for genome assembly; Detects structural variants	Higher cost; Complex library prep; Lower throughput

Table 8.3: Compact workflow from DNA to annotated variants

Step	Output	Key Point
DNA Extraction	Genomic DNA	Purity & integrity critical
Library Prep & Barcoding	DNA library	PCR or PCR-free
Cluster/Template Generation	Clonal templates	Signal amplification
Sequencing Run	FASTQ files	Tech-specific chemistry
QC & Trimming	Clean FASTQ	Remove low-quality
Alignment	BAM/SAM	Check mapping quality
Variant Calling	VCF	SNPs/Indels identified
Annotation	Annotated VCF	Biological interpretation

Chapter 9

Variant Discovery and Genome Assembly Pipeline (Bovo Lessons 6–7)

9.1 Introduction

In this chapter, we describe the first part of the variant discovery pipeline and the principles of genome assembly and annotation. Starting from raw FASTQ reads, the workflow moves through quality control, alignment, variant calling, and finally genome assembly if no reference is available.

—

9.2 Quality Control and Trimming

Raw NGS data are delivered in **FASTQ** format, with 4 lines per read: ID, sequence, separator, and Phred quality string. **FASTQC** is used to check:

- Per-base quality and quality distribution
- GC content and sequence duplication
- Overrepresented sequences and possible contaminants

If the second half of reads shows low quality, **trimming** is applied:

1. **End-trimming:** cut 3' bases until reaching Q20.
2. **Sliding-window trimming:** remove windows whose average Phred score is below the threshold.

Short or chimeric reads are discarded.

—

9.3 Alignment and SAM/BAM Files

Reads are aligned to a reference genome using software such as **BWA** or **Bowtie2**. The alignment output is stored in:

- **SAM**: text alignment format

- **BAM**: binary and compressed SAM

Key SAM elements:

1. Read ID and **Flag** (encodes mapping status)
2. Reference chromosome and mapping position
3. **MAPQ**: mapping quality (Phred-scaled)
4. **CIGAR string**: compact representation of alignment (`M,I,D,S`)

Example: `4S8M2I4M1D` = 4 soft-clipped, 8 match, 2 insertion, 4 match, 1 deletion.

—

9.4 Filtering and Duplicate Removal

- Reads with **MAPQ = 0** (multi-mapped) are often discarded.
- **PCR duplicates** are removed using tools like **Picard**.
- Flags help identify unmapped reads (4), secondary alignments (256), and duplicates (1024).

—

9.5 Variant Calling and VCF Files

After QC, alignment, and filtering, **variants** are detected:

- **VCF file** stores: CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT + samples.
- INFO fields include **DP** (depth) and **AF** (allele frequency).
- Multi-sample calling is preferred for population studies.

Low-quality variants and those in **homopolymeric regions** are filtered to reduce false positives.

—

9.6 Variant Annotation and IGV Inspection

Variants are annotated using **ENSEMBL**, **dbSNP**, and **UniProt** to determine:

- Exonic or intronic location
- Synonymous or nonsynonymous changes
- Regulatory or splicing impacts

Visual inspection with **IGV** confirms whether SNPs or indels are well-supported.

—

9.7 Genome Assembly and Algorithms

When no reference is available, **de novo genome assembly** is required.

Pipeline for Assembly

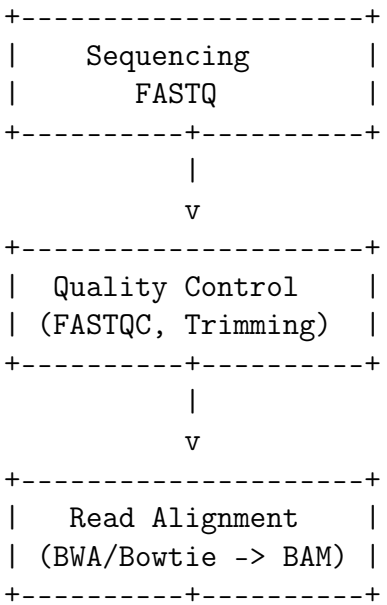
1. Estimate genome size (C-value or K-mer frequency)
2. Extract high-quality, high-molecular-weight DNA
3. Select sequencing technology (Illumina / PacBio / Nanopore / Hybrid)
4. Prepare library (single-end, paired-end, mate-pair)
5. Assemble reads into **contigs**, then **scaffolds**
6. Polish, validate, and compute statistics (N50, coverage)
7. Mask repeats and perform annotation

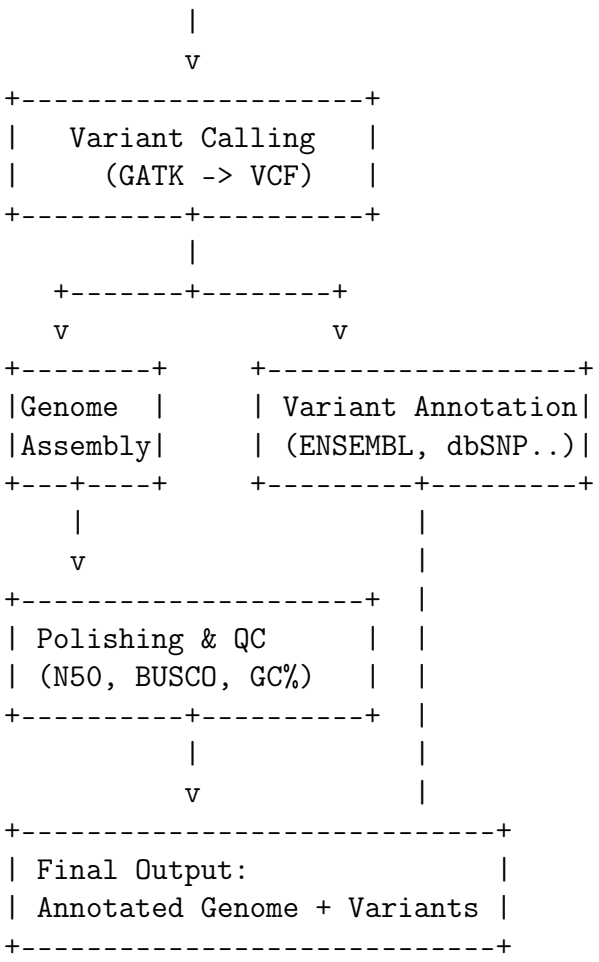
Assembly Algorithms

- **Greedy:** joins best overlaps iteratively (Sanger)
- **Overlap-Layout-Consensus (OLC):** builds graph of overlaps and consensus
- **deBruijn Graph:** represents reads as K-mers for efficient NGS assembly

Key Assembly Metrics

- **Coverage (X)** = $\frac{L \times N}{G}$
- **N50**: contig length where cumulative length = 50% of assembly
- **BUSCO score**: completeness via single-copy orthologs





Lesson 8: Statistics Recap, Genome Annotation, and Variant Identification

Assembly Statistics

N50 Metric The N50 is a widely used metric to assess the contiguity of a genome assembly. It is defined as the length of the smallest contig such that when contigs are ranked from longest to shortest, the sum of their lengths up to this contig reaches 50% of the total assembly length.

A higher N50 indicates longer contigs and, consequently, a more contiguous assembly. However, N50 does **not measure correctness**—it simply summarizes contig lengths.

Interpretation:

- If $N50$ exceeds the median gene length, most genes will fit within a single contig or scaffold.
- Small $N50$ values lead to fragmented gene models and complicate annotation.

Genome Coverage Genome coverage quantifies the proportion of the expected genome size that is actually assembled:

$$\text{Genome Coverage (\%)} = \frac{\text{Total Assembled Bases}}{\text{Expected Genome Size}} \times 100$$

Two approaches can be used to estimate genome size:

1. **C-value method (cytometry)**: genome size (in base pairs) is

$$G_{\text{bp}} = \text{DNA content (pg)} \times 0.978 \times 10^9$$

2. **K-mer frequency distribution**: Decompose reads into k-mers, compute their frequency distribution, and estimate genome size as:

$$G_{\text{bp}} \approx \frac{\text{Total K-mer count}}{\text{Mean K-mer depth}}$$

Gene Coverage Gene coverage measures the proportion of expected genes that are fully or partially present in the assembly. It often exceeds genome coverage because repetitive regions are typically gene-poor.

Assessment tools: BUSCO is widely used to estimate completeness by detecting highly conserved single-copy orthologs.

Genome Annotation Pipeline

Genome annotation is the process of assigning biological meaning to each nucleotide. It includes:

1. **Annotation of repetitive elements (REs)**
2. **Gene annotation:**
   - **Structural annotation:** prediction of ORFs, exons, introns
   - **Functional annotation:** assign names, functions, and ontology terms
3. **Data submission, maintenance, and updates**

Repetitive Elements (REs)

- Low-complexity sequences (homopolymeric runs)
- Transposable elements (TEs), including:
  - LINEs (Long Interspersed Nuclear Elements)
  - SINEs (Short Interspersed Nuclear Elements)
  - Viral insertions

Detection tools:

- Homology-based (`RepeatMasker`, `Dfam`)
- De novo (e.g., `TEdenovo`, HMM profile-based)

Gene Annotation Approaches

1. **Intrinsic / Ab-initio**: relies solely on sequence statistics (e.g., codon bias, ORFs, exon-intron boundaries)
2. **Extrinsic**: aligns known transcripts or proteins to the genome
3. **Combiners**: integrate intrinsic + extrinsic evidence (e.g., AUGUSTUS with RNA-seq)

Curation and Data Submission

- Manual curation validates and corrects automatic predictions.
- Submissions are made to ENA, GenBank, or EMBL for reproducibility and public access.
- Typical output formats: **GFF/GTF** for features, **BED** for intervals, **FASTA** for sequences.

Variant Identification from WGS Data

Mitochondrial genome sequencing (mtDNA) is often used to identify haplotypes and study maternal lineages, traits, or breed-specific markers.

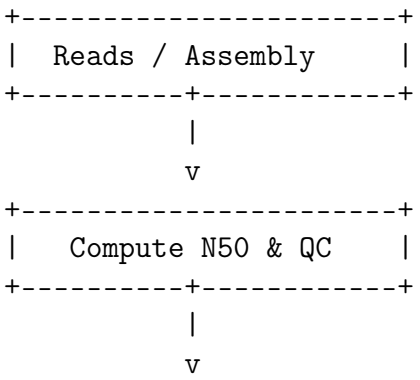
Key steps:

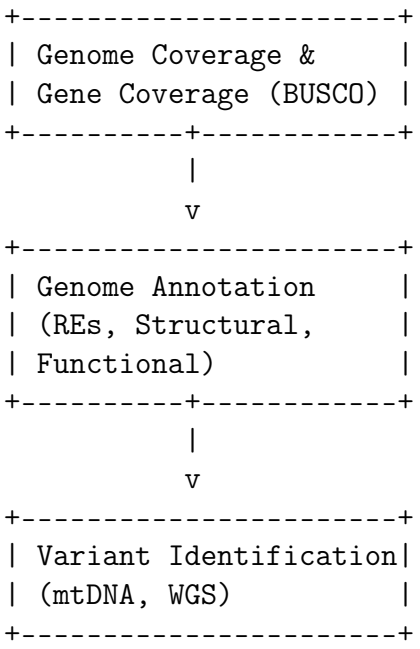
1. Amplify mtDNA using long-PCR with overlapping primer pairs
2. Sequence with NGS platforms (Ion Torrent, Illumina)
3. Map reads to the mtDNA reference genome
4. Inspect depth and variant distribution
5. Beware of **overlapping PCR amplicons** $\rightarrow$ double coverage regions

Careful inspection of the experimental design (primers, fragment sizes, overlaps) is critical to correctly interpret coverage spikes and avoid false structural variant calls.

Workflow Schematic

Below is a simplified summary of the Lesson 8 workflow:




Genome Coverage &
Gene Coverage (BUSCO)
Genome Annotation
(REs, Structural,
Functional)
Variant Identification
(mtDNA, WGS)

Chapter 10

Applied Genomics and Population Genetics

This chapter explores advanced applications of next-generation sequencing (NGS), custom genotyping approaches, copy number variation (CNV) analysis, population genomics, and genome-wide association studies (GWAS). It integrates methodological details, statistical considerations, and biological interpretations for applied genomics projects.

10.1 Sequencing Strategies for Population Analysis

10.1.1 Pool-Seq and Cost-Efficient Population Variability Studies

In population genomics, individual-level sequencing can be prohibitively expensive. When the main goal is to detect **allele frequency differences** rather than individual genotypes, **DNA pooling** can be used:

- DNA from multiple individuals is extracted and combined in equimolar concentrations.
- The pool is sequenced as a single sample.
- Reads are mapped to a reference genome to identify **single-nucleotide polymorphisms (SNPs)**.
- Allele frequencies are estimated from the read counts supporting each allele.

Advantages:

- Significant reduction in cost per individual.
- Suitable for detecting population-level differences in allele frequencies.

Limitations:

- Individual genotypes cannot be recovered.
- Unequal DNA contributions can bias allele frequency estimates.

This strategy is ideal for **comparisons between extreme phenotypes**, e.g., resistant vs. susceptible populations, or red vs. yellow bird populations, where causative loci produce strong allele frequency differences.

10.1.2 Targeted Sequencing and AmpliSeq Panels

Targeted DNA sequencing focuses on **predefined genomic regions**. Typical applications include:

- Clinical panels for disease-associated genes.
- Candidate gene studies for specific traits.
- Sequencing of regulatory or non-coding regions.

AmpliSeq provides off-the-shelf and customizable panels. The workflow generally involves:

1. PCR amplification or hybrid capture of selected regions.
2. Library preparation and NGS sequencing.
3. High-depth coverage of targeted loci, reducing false negatives.

Increasing the number of samples reduces per-sample depth, while fewer samples allow **ultra-deep coverage**, which improves variant detection.

10.1.3 Exome Sequencing and Hybrid Capture

Exome sequencing targets only the **protein-coding regions (exons)**, representing $\sim 2\%$ of the human genome. This dramatically reduces data volume and cost:

- Whole-genome sequencing at $20\times$ coverage may require ~ 60 GB/sample.
- Exome sequencing achieves high coverage with ~ 4–5 GB/sample.

Hybrid capture workflow:

1. Fragment genomic DNA.
2. Use thousands of biotinylated probes to capture exonic fragments.
3. Wash and retain hybridized fragments.
4. Sequence enriched fragments and call variants.

Exome sequencing is widely used to identify **rare variants** and **causative mutations** in Mendelian and complex diseases.

10.1.4 Epigenomic and Transcriptomic Applications

NGS can extend beyond the static genome to study **regulatory dynamics**:

- **Methyl-seq:** Detects cytosine methylation via bisulfite conversion, which changes unmethylated cytosines to uracil, allowing inference of DNA methylation patterns across the genome.
- **ChIP-seq:** Identifies DNA regions bound by transcription factors or histone marks, revealing regulatory landscapes.
- **RNA-seq:** Profiles the transcriptome to quantify gene expression, discover isoforms, and detect non-coding RNAs.

10.2 Custom Genotyping and NGS-Based Approaches

10.2.1 SNP Chips and Hardy-Weinberg Equilibrium

SNP chips provide high-throughput genotyping for a predefined set of SNPs. Design considerations:

- Minor Allele Frequency (MAF)
- Hardy-Weinberg equilibrium (HWE)
- Avoidance of repetitive regions

Hardy-Weinberg equilibrium:

- $p^2 + 2pq + q^2 = 1$
- p, q = allele frequencies

Deviation from HWE may indicate:

- Technical errors in genotyping.
- Structural variants like copy number variations.
- True biological phenomena (selection, inbreeding).

10.2.2 NGS-Based Genotyping

Instead of genotyping with fixed arrays, NGS allows targeted or reduced-representation sequencing approaches:

- **Amplicon sequencing:** Multiplex PCR amplifies target regions for ultra-deep coverage.
- **Hybrid capture genotyping:** Captures selected regions before sequencing.
- **RAD-seq:** Uses restriction enzymes to target reproducible regions near restriction sites, useful in species without reference genomes.

These methods balance cost, depth, and genomic coverage.

10.3 Copy Number Variations (CNVs)

10.3.1 Definition and Biological Significance

Copy Number Variations (CNVs) are DNA segments ≥ 1 kb that are present in variable copy numbers across individuals of the same species. CNVs can affect:

- Gene dosage
- Expression regulation
- Phenotypic traits like coat color or disease susceptibility

They are often **tandem duplications or deletions** on the same haplotype.

10.3.2 Detection Methods

1. **Array CGH (aCGH):** Compares hybridization intensity of sample vs. reference DNA on microarrays.
   - $\log_2(\text{Sample/Reference}) > 0$ indicates gain
   - $\log_2(\text{Sample/Reference}) < 0$ indicates loss
2. **SNP array intensity analysis (PennCNV):** Infers CNVs using contiguous SNPs with similar intensity shifts.
3. **NGS-based methods:**
   - Read depth (coverage) analysis
   - Read-pair and split-read mapping
   - Local de novo assembly

High-depth sequencing improves confidence in CNV detection.

10.4 Population Genomics and Evolutionary Insights

10.4.1 Core Concepts

Population genomics studies the distribution of genetic variation across individuals, populations, and species. It informs:

- Genetic drift and bottleneck effects
- Natural and artificial selection
- Inbreeding and effective population size
- Historical demography and migration

10.4.2 Runs of Homozygosity (ROH)

ROHs are contiguous genomic regions that are fully homozygous. They are used to compute the **genomic inbreeding coefficient**:

- $F_{ROH} = \frac{\sum \text{Length of all ROH}}{\text{Total autosomal genome length}}$

ROH islands are hotspots where many individuals share homozygous regions, often reflecting selection for advantageous alleles.

10.4.3 Fixation Index and Differentiation

The **fixation index** (F_{ST}) measures population differentiation:

- $F_{ST} = \frac{\text{Var between populations} - \text{Var within populations}}{\text{Var between populations}}$
- $F_{ST} \approx 0$ indicates no differentiation
- $F_{ST} \approx 1$ indicates complete separation

Sliding-window F_{ST} analysis across the genome can reveal regions under divergent selection between populations.

10.5 Genome-Wide Association Studies (GWAS)

GWAS links **genetic variants** to **phenotypic traits**.

10.5.1 Study Design and Requirements

- Well-defined phenotype (binary or quantitative)
- Adequate sample size for statistical power
- Dense genotyping or whole-genome sequencing
- Correction for population stratification

10.5.2 Association Testing and Multiple Testing

- Each SNP is tested for allele frequency differences between cases and controls.
- Association is measured with a P-value.
- Multiple testing corrections:
  - Bonferroni: $P_{threshold} = \frac{\alpha}{N_{tests}}$
  - False Discovery Rate (FDR)
  - Genome-wide significance: $P < 5 \times 10^{-8}$

10.5.3 Interpretation and Linkage Disequilibrium

Significant peaks in a **Manhattan plot** often include multiple SNPs in **linkage disequilibrium (LD)**, reflecting the inheritance of haplotype blocks rather than single variants. Fine-mapping and functional annotation are needed to pinpoint causative genes.

—

Key Takeaways

- Sequencing strategies can be tailored for cost, depth, and resolution.
- CNVs and SNPs jointly contribute to phenotypic diversity.
- ROHs and F_{ST} enable population history and selection inference.
- GWAS requires careful phenotype definition, population control, and robust statistical thresholds.

Conceptual Core of Applied Genomics

Applied genomics integrates **sequencing technologies**, **population genetics**, and **bioinformatics** to understand:

- Genetic **variation**: SNPs, INDELs, CNVs
- **Gene function** and **regulatory mechanisms**
- **Population structure**, selection, and evolutionary history

Projects may focus on:

- Variant discovery (disease or trait causative alleles)
- Population variability (allele frequencies, drift, bottlenecks)
- Functional genomics (expression, methylation, TF binding)

Technique	**Scope**	**Main Use**
WGS	Whole genome	Variant discovery, CNV detection
WES	Exons (∼2%)	Rare variants, Mendelian disease
AmpliSeq / Targeted	Selected genes	Clinical genotyping, disease screening
RNA-Seq	Transcriptome	Expression profiling, isoforms
Methyl-Seq	CpG methylation	Epigenetic regulation studies
ChIP-Seq	Protein-DNA binding	Regulatory
TF mapping		
SNP-Chip	Predefined loci	GWAS, ROH, pop. genetics
RAD-Seq / GBS	Reduced genome	Pop. genomics, no full ref.
Pool-Seq	Mixed DNA	Allele freq. screening, cost-effective

Table 10.1: Compact summary of key sequencing and genotyping techniques.

Key Databases & Repositories

- **Variation**: dbSNP, gnomAD, ClinVar
- **Population & Evolutionary**: 1000 Genomes, Ensembl, UCSC Genome Browser
- **Functional**: GWAS Catalog, GEO, ENCODE, Roadmap Epigenomics

Key Computational Tools

- **Variant Calling**: BWA, Bowtie2, GATK, Samtools/BCFtools
- **Population Genomics & GWAS**: PLINK, PennCNV, VCFtools
- **Annotation**: ANNOVAR, SnpEff, Ensembl VEP

Key Analytical Concepts

- **Genetic Variation**: SNPs, INDELs, CNVs, structural variants
- **Population Structure**: Allele frequency, HWE, drift, bottleneck, admixture
- **Inbreeding & ROH**: Homozygosity segments indicate recent inbreeding or selection

- **Genome-Wide Association (GWAS)**: Links phenotypes to causative loci
- **Linkage Disequilibrium (LD)**: Variants inherited together in haplotype blocks
- **Copy Number Variants (CNVs)**: aCGH, SNP-chip intensity, NGS read depth
- **Functional Genomics**: RNA-Seq (expression), ChIP-Seq (TF binding), Methyl-Seq (epigenetics)

Integrated Workflow

1. **Sample & Phenotype Collection** (metadata, pedigree)
2. **Sequencing / Genotyping** (WGS, WES, SNP-chip, Pool-Seq)
3. **Data Processing** (Alignment → Variant Calling → Annotation)
4. **Population Analysis** (Frequencies, HWE, LD, Fst, ROH, CNV)
5. **GWAS & Functional Analysis** (association tests, peak detection)
6. **Result Interpretation & Biological Insight** (candidate loci, evolutionary patterns)

Applied Genomics – Exam Simulation with Answers

Multiple Choice (25 Questions)

1. **Allele definition**
   Answer: One of two or more alternative forms of a gene.

2. **SOLiD sequencing principle**
   Answer: Sequencing by ligation using color-space coding.

3. **FST definition**
   Answer: Measure of population differentiation due to genetic structure.

4. **Ion Torrent**
   Answer: NGS technology based on pH/ion release detection during DNA polymerization.

5. **LD (Linkage Disequilibrium)**
   Answer: Non-random association of alleles at two or more loci in a population.

6. **ROH relation to inbreed degree**
   Answer: Long Runs of Homozygosity indicate higher inbreeding coefficients.

7. **Illumina size of reads**
   Answer: Typically 100–300 bp (short-read technology).

8. **GWAS aim**
   Answer: Identify associations between genetic variants (SNPs) and traits/diseases.

9. **Population not in HW eq: selection**
   Answer: Indicates deviation due to selection, inbreeding, or population stratification.

10. **Population stratification with MDS**
    Answer: Visualizes population substructure to correct confounding in GWAS.

11. **Paired-end sequencing**
Answer: Sequencing from both ends of a DNA fragment to improve mapping.

12. **VCF file**
Answer: Stores variant calls, positions, genotypes, and annotations.

13. **FASTQ file**
Answer: Stores raw reads with nucleotide sequences and quality scores.

14. **PacBio SMRT sequencing**
Answer: Single Molecule Real-Time sequencing producing long reads.

15. **CIGAR meaning**
Answer: Encodes how a read aligns to a reference (matches, insertions, deletions, clips).

16. **Functional and structural annotation**
Answer: Structural: identifies gene/exon positions. Functional: assigns gene functions.

17. **de Bruijn graph algorithm (Eulerian path)**
Answer: Used in de novo assembly: traverses k-mer overlaps to reconstruct contigs.

18. **Completeness of a genome: BUSCO**
Answer: Assesses genome completeness using conserved single-copy orthologs.

19. **Equimolar DNA pool meaning**
Answer: DNA from each sample is pooled at equal molar concentration for sequencing.

20. **aCGH**
Answer: Microarray hybridization method to detect Copy Number Variations (CNVs).

21. **What does LD describe?**
Answer: (b) The correlation between alleles of two SNPs within a population.

22. **Which NGS produces long reads?**
Answer: (c) PacBio.

23. **Manhattan plot Y-axis (GWAS)**
Answer: (c) Significance level $-\log_{10}(P)$ of each SNP association.

24. **Purpose of a SAM file**
Answer: (b) Store sequence alignments to a reference genome.

25. **Primary purpose of ChIP-Seq**
Answer: Identify DNA regions bound by proteins (e.g., transcription factors).

Short Answer (5 Questions with Solutions)

1. **Depth of coverage calculation**
   $Depth = \frac{\text{Total bases sequenced}}{\text{Genome size}}$
   Example: 100 M reads × 100 bp/1 Gbp = 10× coverage.

2. **N50 calculation example**
   Sort contigs by size, sum lengths until 50% of the total assembly is reached; the length of the last contig added = N50.

3. **Explain bisulfite sequencing**
   Converts unmethylated cytosines to uracil (read as T), methylated cytosines remain C → used to study DNA methylation.

4. **Estimate genome size (C-value method)**
   $Genome\,size(bp) = DNA\,content(pg) \times 0.978 \times 10^9$.

5. **Explain and draw a Manhattan plot**
   Plot $-\log_{10}(P)$ vs chromosome position; peaks indicate loci associated with traits.

Multiple Choice Questions (Select one answer)

1. What does **linkage disequilibrium (LD)** describe?
   Answer: b) The correlation between alleles of two SNPs within a population

2. Which NGS technology is known for producing **long reads** often used in de novo genome assembly?
   Answer: c) PacBio

3. In a **Manhattan plot (GWAS analysis)**, what does the vertical axis typically represent?
   Answer: c) The significance level $[-\log_{10}(P)]$ of each SNP association

4. What is the purpose of a **SAM file** in NGS data analysis?
   Answer: b) To store sequence alignments to a reference genome

5. What is **aCGH**?
   Answer: b) A method based on microarray hybridization to identify CNV (copy number variants)

Open Questions with Solutions

1. **Aim of GWAS:** Identify statistical associations between genetic variants (e.g., SNPs) and phenotypic traits to discover candidate loci for traits or diseases.

2. **Primary purpose of ChIP-Seq:** Detect protein-DNA interactions, such as transcription factor binding sites or histone modification locations.

3. **Compute average sequencing depth:**
   Genome size $G = 1 \times 10^9$ bp
   Total bases sequenced $B = 1 \times 10^8$ reads × 100 bp $= 1 \times 10^{10}$ bp
   Depth $D = \frac{B}{G} = \frac{1\times10^{10}}{1\times10^9} = 10\times$ coverage

4. **Check Hardy-Weinberg Equilibrium (HWE):**
   Given: 500 AA, 200 Aa, 300 aa individuals ($N = 1000$)
   Allele freq: $p = \frac{2*500+200}{2000} = 0.6$, $q = 0.4$
   Expected: $p^2 = 0.36$, $2pq = 0.48$, $q^2 = 0.16$
   Observed: 0.5, 0.2, 0.3 $\Rightarrow$ **Not in HWE**