Applied Genomics – Multiple Choice Questions (MCQs)

Martina Castellucci

University of Bologna – Master in Bioinformatics

Contents

Instructions

Each question has four options (A–D). The correct answer is marked with **(*)** and a brief explanation is provided. References point to the course notes or lecture slides.

1 Classical Genetics and Mendelian Principles

Q1. Mendel's law of segregation states that:

(A) Each allele from a parent remains linked in the offspring

(B) Alleles segregate randomly during gamete formation **(*)**

(C) Genes on the same chromosome segregate independently

(D) Phenotype is always dominant over genotype

Explanation: Alleles of a gene separate during meiosis so that each gamete carries only one allele. *Ref:* Lecture 1, Slide 15.

Q2. In a dihybrid cross of two heterozygous individuals (AaBb x AaBb), the expected phenotypic ratio is:

(A) 3:1

(B) 9:3:3:1 **(*)**

(C) 1:2:1

(D) 2:1

Explanation: Independent assortment of two genes gives the 9:3:3:1 phenotypic ratio. *Ref:* AG_sbobins p. 21.

Q3. In a pedigree chart, a half-filled circle represents:

(A) A healthy male

(B) A healthy female

(C) A female carrier **(*)**

(D) An affected female

Explanation: By convention, circles = females; half-filled = carrier (heterozygote). *Ref:* Lecture 1, Pedigree slides.

2 Population Genetics

Q4. Hardy-Weinberg equilibrium assumes all of the following EXCEPT:

(A) Random mating

(B) Infinite population size

(C) Natural selection **(*)**

(D) No mutation

Explanation: H-W assumes no selection, mutation, migration, or drift. *Ref:* AG_sbobins, Population genetics section.

Q5. If allele A has a frequency of 0.6, what is the expected frequency of heterozygotes under H-W?

(A) 0.16

(B) 0.24

(C) 0.48 (*)

(D) 0.36

Explanation: Heterozygote frequency = 2pq = 2*0.6*0.4 = 0.48. *Ref:* Lecture 1, H-W example.

3 Molecular Genetics and Sequencing

Q6. The first-generation sequencing method is:

(A) Illumina sequencing

(B) Nanopore sequencing

(C) Sanger dideoxy sequencing (*)

(D) SOLiD sequencing

Explanation: Sanger sequencing was the classical 1st-generation method using chain terminators. *Ref:* Lecture 2, Sanger slides.

Q7. Which NGS technology detects changes in pH?

(A) Illumina

(B) PacBio

(C) Ion Torrent (*)

(D) Nanopore

Explanation: Ion Torrent detects H^+ release (pH change) during nucleotide incorporation. *Ref:* AG_sbobins p. 112.

Q8. A Phred score of 30 corresponds approximately to:

(A) 1/10 error rate

(B) 1/100 error rate

(C) 1/1000 error rate (*)

(D) 99% error rate

Explanation: Q = -10 $\log_{10}$P $\rightarrow$ Q30 = 0.001 error probability = 99.9% accuracy. *Ref:* Lecture 4, FASTQ slide.

4 NGS Data Analysis and QC

Q9. Which tool is most commonly used for initial quality control of FASTQ files?

(A) BWA

(B) FastQC (*)

(C) GATK

(D) SAMtools

Explanation: FastQC provides modular per-base and per-sequence quality plots. *Ref:* Lecture 4, QC module.

Q10. The correct order in a simple variant discovery pipeline is:

(A) Variant calling → Alignment → QC

(B) QC → Alignment → Variant calling (*)

(C) Alignment → QC → Variant calling

(D) QC → Variant calling → Alignment

Explanation: First check data quality, then align, then call variants. *Ref:* AG_sbobins, NGS pipeline.

5 Genome Assembly and Comparative Genomics

Q11. De Bruijn graphs in genome assembly use:

(A) Reads as nodes

(B) K-mers as edges (*)

(C) Reads as edges

(D) Contigs as nodes

Explanation: In a de Bruijn graph, (k-1)-mers are nodes and k-mers form edges. *Ref:* Lecture 6, assembly section.

Q12. Which strategy can improve assembly contiguity in repetitive regions?

(A) Using only short Illumina reads

(B) Increasing coverage with short reads

(C) Using long reads (PacBio/ONT) (*)

(D) Ignoring repeats

Explanation: Long reads span repeats and reduce fragmentation. *Ref:* AG_sbobins genome assembly notes.

6 Transcriptomics and Functional Genomics

Q13. Which RNA fraction is typically enriched for mRNA-Seq in eukaryotes?

(A) rRNA

(B) Poly-A RNA (*)

(C) tRNA

(D) snRNA

Explanation: Poly-A selection enriches mature mRNAs and removes most rRNA. *Ref:* Lecture 7, RNA-Seq library prep.

Q14. What is the main challenge in RNA-Seq read alignment?

(A) Sequencing errors

(B) Short read length

(C) Splice junctions (*)

(D) GC content

Explanation: Reads often span exon-exon junctions requiring splice-aware mappers. *Ref:* AG_sbobins transcriptomics.

Q15. Which tool is typically used for differential gene expression analysis?

(A) GATK

(B) DESeq2 (*)

(C) MAUVE

(D) OrthoFinder

Explanation: DESeq2 and edgeR are standard R packages for RNA-Seq DEGs. *Ref:* Lecture 8, RNA-Seq analysis.

7 GWAS, CNV, and Population Genomics

Q16. In a case-control GWAS, which plot is typically used to visualize significant associations?

(A) Synteny plot

(B) Manhattan plot (*)

(C) Phylogenetic tree

(D) PCA plot

Explanation: Manhattan plots show SNP p-values along the genome. *Ref:* Lecture 10, GWAS slide.

Q17. The fixation index (F_{ST}) measures:

(A) Allele frequency in one population

(B) Differentiation between populations (*)

(C) Inbreeding within a single individual

(D) Number of heterozygotes

Explanation: F_{ST} compares variance between vs. within populations. *Ref:* AG_sbobins, Pop Genomics.

Q18. Which technique is best to detect copy number variations genome-wide?

(A) ChIP-Seq

(B) aCGH (*)

(C) RNA-Seq

(D) ATAC-Seq

Explanation: Array-CGH compares hybridization signals to detect gains/losses. *Ref:* Lecture 11, CNV slides.

8 Applied Genomics and Bioinformatics Tools

Q19. Which tool is used for orthology and comparative genomics?

(A) STAR

(B) OrthoFinder (*)

(C) BUSCO

(D) DESeq2

Explanation: OrthoFinder clusters orthologs and infers species trees. *Ref:* Lecture 9, Comparative Genomics.

Q20. BUSCO is primarily used to:

(A) Detect SNPs

(B) Assess genome completeness (*)

(C) Identify CNVs

(D) Perform de novo assembly

Explanation: BUSCO searches for universal single-copy orthologs. *Ref:* Lecture 6, Assembly QC.

Q21. AntiSMASH is specialized for:

(A) Variant calling

(B) Biosynthetic gene cluster prediction (*)

(C) RNA-Seq alignment

(D) Contamination removal

Explanation: AntiSMASH annotates secondary metabolite clusters (BGCs). *Ref:* Lecture 9, Functional genomics.

9 Epigenomics and Chromatin Analysis

Q22. The main principle of ChIP-Seq is:

(A) Sequencing cDNA molecules

(B) Capturing DNA fragments bound by proteins of interest (*)

(C) Detecting methylated cytosines

(D) Measuring RNA expression

Explanation: ChIP-Seq uses antibodies to immunoprecipitate DNA-protein complexes and sequences the bound DNA. *Ref:* Lecture 10, ChIP-Seq slides.

Q23. Which chemical treatment is used in bisulfite sequencing to detect methylation?

(A) Formaldehyde

(B) Sodium bisulfite (*)

(C) Proteinase K

(D) DNase I

Explanation: Sodium bisulfite converts unmethylated cytosines to uracils, leaving 5-mC unchanged. *Ref:* AG_sbobins, Methyl-Seq notes.

Q24. In methyl-seq analysis, a C that remains a C after bisulfite treatment means:

(A) Conversion failed

(B) It was methylated (*)

(C) It was unmethylated

(D) It is a sequencing error

Explanation: Methylated cytosines resist bisulfite conversion. *Ref:* Lecture 10, DNA methylation slides.

10 Metagenomics and Environmental Genomics

Q25. The main advantage of shotgun metagenomics over 16S rRNA sequencing is:

(A) Lower cost

(B) It only targets bacteria

(C) Provides functional and taxonomic information (*)

(D) Requires pure cultures

Explanation: Shotgun metagenomics sequences all DNA to identify both species and gene functions. *Ref:* AG_sbobins, Metagenomics section.

Q26. Which step is NOT typical in a metagenomics workflow?

(A) DNA extraction from mixed sample

(B) Library preparation

(C) Taxonomic profiling

(D) Sanger capillary electrophoresis (*)

Explanation: Modern metagenomics uses NGS, not first-generation Sanger. *Ref:* Lecture 12, Environmental genomics.

Q27. The term "microbiome" refers to:

(A) Only bacterial DNA in soil

(B) The set of genes of a microbial community (*)

(C) Only culturable microorganisms

(D) Fungal spores in the environment

Explanation: Microbiome = collective genomes of all microbes in a niche. *Ref:* AG_sbobins p. 256.

11 Hybrid Sequencing and Assembly Statistics

Q28. The main advantage of hybrid genome assembly is:

(A) It avoids sequencing errors

(B) Combines long reads for contiguity and short reads for accuracy (*)

(C) Requires no polishing

(D) It only uses Illumina data

Explanation: Hybrid assembly leverages ONT/PacBio long reads to span repeats and Illumina reads to polish errors. *Ref:* Lecture 6, Assembly slides.

Q29. N50 is defined as:

(A) The number of reads covering 50% of the genome

(B) The length at which 50% of the assembly is contained in contigs of that size or longer (*)

(C) The average read length

(D) The coverage of longest contig

Explanation: N50 is a measure of assembly contiguity: 50% of genome is in contigs N50 length. *Ref:* AG_sbobins assembly metrics.

Q30. L50 represents:

(A) 50% GC content

(B) The minimum number of contigs covering 50% of genome (*)

(C) Number of long reads over 50 kb

(D) Lowest base quality score

Explanation: L50 counts the fewest contigs needed to cover half the total genome size. *Ref:* Lecture 6, QC metrics.

12 k-mer Concepts and NGS Pipelines

Q31. In a de Bruijn graph, k-mer size affects:

(A) GC content

(B) Assembly resolution and repeat handling (*)

(C) Sequencing chemistry

(D) Only read length

Explanation: Larger k reduces ambiguity but requires higher coverage; smaller k can join repeats incorrectly. *Ref:* AG_sbobins genome assembly.

Q32. Which is the correct simplified order for an NGS variant calling pipeline?

(A) Alignment → Variant calling → QC

(B) QC → Alignment → Variant calling (*)

(C) QC → Variant annotation → Alignment

(D) Variant calling → Annotation → Alignment

Explanation: Raw data → QC → Align reads → Call variants → Annotate. *Ref:* Lecture 4, Variant analysis slide.

Q33. BUSCO evaluates:

(A) Structural variants

(B) Genome completeness using single-copy orthologs (*)

(C) Read duplication rate

(D) Gene expression levels

Explanation: BUSCO finds expected conserved genes to assess completeness of assemblies or annotations. *Ref:* Lecture 6, BUSCO.

13 Population and Comparative Genomics

Q34. ROH (Runs of Homozygosity) indicate:

(A) Recombination hotspots

(B) Segments identical by descent (*)

(C) Only heterozygous regions

(D) CNV duplications

Explanation: ROH = long homozygous stretches in the genome, linked to inbreeding. *Ref:* Lecture 11, Inbreeding section.

Q35. Which parameter measures population differentiation?

(A) F_{ST} (*)

(B) Heterozygosity

(C) Linkage disequilibrium

(D) LOD score

Explanation: F_{ST} compares genetic variance within vs. between populations. *Ref:* AG_sbobins, Population genomics.

Q36. A Manhattan plot in GWAS displays:

(A) Chromosome synteny

(B) SNP positions vs. –log10(p-value) (*)

(C) Gene expression levels

(D) Recombination frequency

Explanation: Peaks indicate loci significantly associated with traits. *Ref:* Lecture 10, GWAS example.

Applied Genomics – written test – examples.

Select the correct answer (one, and only one).

1. What does linkage disequilibrium describe?
a) The degree of similarity between two populations
b) The correlation between alleles of two SNPs within a population
c) The rate of linked contigs in the process of genome assembly
d) The rate of mutation in genetic markers

2. Which NGS technology is known for producing long reads, often used in de novo genome assembly?
a) Illumina
b) Ion Torrent
c) PacBio
d) Roche 454

3. In a Manhattan plot (GWAS analysis), what does the vertical axis typically represent?
a) The position of the SNP (bp) along the chromosome
b) The minor allele frequency (MAF) of each SNP
c) The significance level [$-\log_{10}(P)$] of each SNP association
d) The effect size (β) of each SNP on the phenotype

4. What is the purpose of a SAM file in NGS data analysis?
a) To store raw sequencing reads and quality scores
b) To store sequence alignments to a reference genome
c) To store variant calls and inferred genotypes
d) To store genome annotation information after genome assembly

5. What is aCGH?
a) A technology to re-sequence genome based on chip hybridization
b) A method based on microarray hybridization to identify CNV
c) A NGS approach based on pair end sequencing
d) An approach used for the advanced-evaluation of chromosomal genomic heterozygosity

$1 \cdot 10^8 \cdot 100 / 1 \cdot 10^9 = 10\times$
LN/G

Open questions (short answer):
1. What is the aim of a GWAS?
2. What is the primary purpose of ChIP-Seq?
3. Estimate the average depth of sequencing (e.g. 5x) of a genome of size 1Gbp ($=1\times10^9$) considering that a total of 100 million reads ($=1\times10^8$) of size 100bp have been obtained. Calculate and state.
4. Estimate if the population is in HWE considering 500 AA, 200 Aa, 300 aa individuals. Calculate and state.

$300^2 + 2 \cdot \quad + 300^2$

Got it! Here’s the mock exam in English with answers.

📘 Applied Genomics – Mock Exam

Part 1 – Multiple Choice Questions (25)

1. Which of the following is not a main branch of genetics?

A) Classical

B) Molecular

C) Quantitative

D) Phylogenetic

Answer: D

2. The Hardy-Weinberg law describes:

A) Ratio of dominant and recessive genes in a family

B) Equilibrium of allele frequencies in ideal populations

C) Probability of mutations in repetitive sequences

D) Mitochondrial inheritance

Answer: B

3. Ion Torrent technology detects:

A) Light emission

B) H+ production and pH changes

C) Pyrophosphate release

D) Fluorophore binding

Answer: B

4. In FASTQ format, base quality is encoded as:

A) Integer values 0–100

B) Binary codes

C) ASCII characters (Phred score)

D) FASTA sequences

Answer: C

5. The formula for average sequencing depth is:

A) G / (LN)

B) LN / G

C) L / (GN)

D) N / (LG)

Answer: B

6. N50 represents:

A) The average gene length

B) The GC content of a genome

C) The length of the shortest contig covering 50% of the assembly

D) The average sequencing depth

Answer: C

7. Which technology produces the longest reads on average?

A) Illumina

B) Ion Torrent

C) PacBio

D) ABI SOLiD

Answer: C

8. SMRT sequencing is associated with:

A) Illumina

B) Oxford Nanopore

C) PacBio

D) Ion Torrent

Answer: C

9. A Q score of 20 corresponds to an error rate of:

A) 1/10

B) 1/50

C) 1/100

D) 1/1000

Answer: C

10. A key advantage of Oxford Nanopore is:

A) High accuracy

B) Portability and field use

C) Short high-quality reads

D) No need for high-molecular-weight DNA

Answer: B

11. BAM format is:

A) Plain text

B) Binary compressed version of SAM

C) Identical to VCF

D) Protein-specific

Answer: B

12. Which software is commonly used for FASTQ quality control?

A) BWA

B) FASTQC

C) AUGUSTUS

D) PLINK

Answer: B

13. The most common assembly algorithm for NGS data is:

A) Greedy

B) Overlap-Layout-Consensus

C) De Bruijn Graph

D) String Graph

Answer: C

14. In a GFF file, the “strand” column indicates:

A) Contig length

B) Transcription direction (+ or –)

C) Gene quality

D) Alignment score

Answer: B

15. The principle of ChIP-seq is:

A) Sequencing total RNA

B) Using antibodies against DNA-bound proteins

C) Detecting SNPs

D) Identifying CNVs

Answer: B

16. Exon capture with biotinylated probes is typical of:

A) Pool-seq

B) RAD-seq

C) Whole Exome Sequencing

D) RNA-seq

Answer: C

17. Which of the following is a structural variant?

A) SNP

B) Indel

C) CNV

D) Substitution

Answer: C

18. IGV is used to:

A) Assemble genomes

B) Visualize alignments and variants

C) Detect contamination

D) Assess read quality

Answer: B

19. Bonferroni correction is used to:

A) Increase sequencing coverage

B) Correct GC bias

C) Reduce false positives in multiple testing

D) Improve de novo assembly

Answer: C

20. Which FST value indicates completely separated populations?

A) 0

B) 0.25

C) 0.5

D) 1

Answer: D

21. Bisulfite conversion is used to study:

A) CNVs

B) DNA methylation

C) Transcriptome

D) Alternative splicing

Answer: B

22. RAD-seq:

A) Sequences the entire genome at low coverage

B) Uses restriction enzymes to reduce genome complexity

C) Detects alternative transcripts

D) Is equivalent to RNA-seq

Answer: B

23. In genotyping assays, MAF stands for:

A) Frequency of the most common phenotype

B) Minor allele frequency

C) Mean FST values

D) Average contig size

Answer: B

24. Runs of Homozygosity (ROH) are used to:

A) Measure genomic inbreeding

B) Calculate GC content

C) Estimate sequencing coverage

D) Identify CNVs

Answer: A

25. Which tool estimates genome completeness using conserved genes?

A) FASTQC

B) GATK

C) BUSCO

D) Bowtie

Answer: C

Part 2 – Short Open Questions (5)

1. Explain the difference between de novo sequencing and resequencing.

Answer:

- De novo: genome assembly without a reference, needed for new species.
- Resequencing: aligning reads to an existing reference genome, cheaper and faster.

2. What are the main pros and cons of long-read vs. short-read technologies?

Answer:

- Pros: resolve repeats, detect structural variants, phase haplotypes.
- Cons: higher cost, require high-quality DNA, higher error rate (improved with PacBio CCS).

3. How is RNA-seq used in genome annotation projects?

Answer:

RNA-seq identifies transcribed regions, exon–intron boundaries, and expression levels, improving structural and functional gene annotation.

4. How is the inbreeding coefficient (F-ped) calculated, and what are its limits?

Answer:

F-ped = probability that two alleles are identical by descent, derived from pedigrees.

Limits: requires full pedigree, assumes unrelated founders, ignores recombination randomness, may contain pedigree errors.

5. What does a Manhattan plot represent, and how is it interpreted?

Answer:

It shows GWAS results:

- X-axis = SNPs across genome

- Y-axis = –log10(p-value)

  Tall peaks = loci significantly associated with phenotype; nearby peaks often due to linkage disequilibrium.

Applied Genomics – Advanced Mock Exam (Detailed & Specific)

Part 1 – Multiple Choice Questions (25)

1. In Ion Torrent sequencing, what is the function of the key sequence?

A) Acts as an adapter for cluster formation

B) Quality control of the ionogram

C) Defines barcode uniqueness

D) Prevents incorporation errors

Answer: B

2. Which technology allows sequencing of RNA molecules directly, without cDNA synthesis?

A) Illumina

B) PacBio

C) Oxford Nanopore

D) 454 Roche

Answer: C

3. Which statement about paired-end sequencing is correct?

A) It doubles coverage without additional cost

B) It enables easier assembly by knowing both read ends

C) It prevents PCR duplication artifacts

D) It removes the need for reference genomes

Answer: B

4. Which error type is especially problematic for Ion Torrent?

A) Fluorescence miscalls

B) Homopolymer length estimation

C) Incorrect barcode assignment

D) Incorrect primer annealing

Answer: B

5. The Phred quality score (Q) is calculated as:

A) $Q = -10 \times \log10(e)$

B) $Q = \log2(1/e)$

C) $Q = -100 \times e$

D) $Q = \log10(e)$

(where e = probability of base error)

Answer: A

6. Which Illumina innovation reduces reagent costs while maintaining accuracy?

A) Rolling circle amplification

B) Two-color sequencing

C) Linear amplification

D) SMRT sequencing

Answer: B

7. Which library prep method in SOLiD sequencing allows paired-end reads?

A) Single DNA libraries

B) Fragment libraries

C) Circular consensus libraries

D) Long-read libraries

Answer: B

8. Which technology produces circular consensus sequencing (CCS) reads?

A) Illumina

B) PacBio

C) Oxford Nanopore

D) Ion Torrent

Answer: B

9. In a FASTQ file, which line encodes quality scores?

A) Line 1 (identifier)

B) Line 2 (raw sequence)

C) Line 3 (+ symbol)

D) Line 4 (ASCII characters)

Answer: D

10. Which software tool is widely used for alignment in variant discovery pipelines?

A) PLINK

B) BWA

C) FASTQC

D) BUSCO

Answer: B

11. Which graph algorithm is most commonly used in short-read genome assembly?

A) Greedy

B) De Bruijn graph

C) Overlap-layout-consensus

D) String graph

Answer: B

12. Why is GC-rich DNA problematic in Illumina sequencing?

A) GC bases interfere with polymerase incorporation

B) GC-rich fragments fragmentate more easily

C) GC regions prevent barcode detection

D) GC bias is not an issue in Illumina

Answer: A

13. The N50 of an assembly is 50 kb. What does this mean?

A) The average contig length is 50 kb

B) 50% of the genome is contained in contigs of at least 50 kb

C) Half the contigs are 50 kb long

D) The total genome size is 50 kb

Answer: B

14. Which parameter is most important to detect CNVs with NGS?

A) Read depth

B) Barcode errors

C) GC content

D) Phred score

Answer: A

15. Which step in annotation usually comes first?

A) Functional annotation

B) Repeat masking

C) RNA-seq alignment

D) Gene ontology assignment

Answer: B

16. AUGUSTUS is an example of which annotation method?

A) Intrinsic (ab initio)

B) Extrinsic (evidence-based)

C) Hybrid

D) Comparative annotation

Answer: A

17. BUSCO evaluates genome completeness based on:

A) GC content distribution

B) Highly conserved single-copy orthologs

C) Presence of barcoded sequences

D) SNP density

Answer: B

18. What does the CIGAR string “4S8M2I4M1D3M” indicate?

A) 4 bases trimmed, 8 matches, 2 insertions, 4 matches, 1 deletion, 3 matches

B) 4 mismatches, 8 matches, 2 insertions, 4 substitutions, 1 deletion, 3 duplications

C) Alignment quality <4

D) The read contains only soft-clipped bases

Answer: A

19. In a VCF file, the FILTER field indicates:

A) If the variant passed the user’s quality thresholds

B) The type of variant (SNP, indel, CNV)

C) The sequencing technology used

D) The depth of coverage

Answer: A

20. The International HapMap Project was designed to:

A) Annotate human reference genomes

B) Identify linkage disequilibrium patterns and common haplotypes

C) Develop Illumina sequencing

D) Detect protein-coding genes

Answer: B

21. A population bottleneck usually leads to:

A) Increased heterozygosity

B) Fixation of rare alleles

C) Reduced genetic diversity

D) Increased effective population size

Answer: C

22. Linkage disequilibrium (LD) occurs when:

A) Two genes are always on different chromosomes

B) Alleles are inherited together more often than expected

C) A population is in Hardy-Weinberg equilibrium

D) Sequencing errors inflate allele frequency

Answer: B

23. Runs of Homozygosity (ROH) are identified using:

A) RNA-seq data

B) SNP genotyping arrays

C) Proteomic datasets

D) FASTQ files

Answer: B

24. A Manhattan plot is typically used in:

A) RNA-seq differential expression

B) Genome-wide association studies (GWAS)

C) Variant calling pipelines

D) Genome annotation

Answer: B

25. The Bonferroni correction adjusts:

A) Sequencing quality scores

B) P-value thresholds in multiple testing

C) GC content bias

D) Genome coverage

Answer: B

Part 2 – Short Open Questions (5)

1. Describe the main differences between Illumina and Oxford Nanopore sequencing in terms of workflow, read length, and error profile.

Answer: Illumina: short reads (~150–300 bp), high accuracy, fluorescence detection, cluster amplification in flow cells. Oxford Nanopore: long reads (5–100 kb), portable device, high error rates (especially indels), direct DNA/RNA sequencing through protein nanopores.

2. Explain the concept of "breadth of coverage" and how it differs from "depth of coverage."

Answer: Depth = average number of times each base is sequenced (e.g., 30X). Breadth = proportion of the genome covered at least to a target depth (e.g., 95% of genome covered at ≥10X).

3. How are de novo repeats identified during genome annotation, and why is masking them important?

Answer: Identified using homology-based (e.g., RepeatMasker) or de novo (e.g., TEdenovo) tools. Masking avoids misannotation of repeats as genes and improves structural gene prediction.

4. What is the principle behind bisulfite sequencing for DNA methylation, and what is the main limitation of this method?

Answer: Bisulfite converts unmethylated cytosines to uracil (→ thymine after PCR), while methylated cytosines remain unchanged. Limitation: cannot easily distinguish true C→T SNPs from bisulfite-induced conversions.

5. In population genomics, how does the fixation index (FST) help measure population differentiation? Give an example of interpretation.

Answer: FST compares heterozygosity within vs. between populations. FST = 0 → no differentiation; FST = 1 → complete separation. Example: FST ~0.3 suggests moderate genetic divergence, often due to local adaptation.

📘 Applied Genomics – Full Coverage Mock Exam

Part 1 – Multiple Choice Questions (25)

Section A – Sequencing Technologies

1. Which technology detects pyrophosphate release coupled to light emission?

A) Illumina

B) 454 Roche

C) Ion Torrent

D) Oxford Nanopore

Answer: B

2. What is the average read length of Illumina sequencing?

A) 30–50 bp

B) 100–300 bp

C) 5–10 kb

D) >50 kb

Answer: B

3. PacBio CCS (circular consensus sequencing) improves accuracy by:

A) Using paired-end reads

B) Repeatedly sequencing the same circular DNA molecule

C) Trimming low-quality bases

D) Masking repeats before assembly

Answer: B

4. Oxford Nanopore sequencing errors are mostly due to:

A) Incorrect cluster formation

B) Difficulty in distinguishing homopolymer stretches

C) Misinterpretation of electric current signals

D) Fluorescent dye degradation

Answer: C

5. Which sequencing technology is best for degraded DNA (e.g., ancient samples)?

A) Oxford Nanopore

B) Illumina

C) PacBio

D) ABI SOLiD

Answer: B

Section B – Quality Control & File Formats

6. A DNA sample with 260/280 = 1.2 indicates:

A) Pure DNA

B) Protein contamination

C) Carbohydrate contamination

D) RNA contamination

Answer: B

7. The FASTQ quality string uses:

A) ASCII encoding of Phred scores

B) Hexadecimal error codes

C) Direct probability values

D) Binary encoding

Answer: A

8. In SAM format, the CIGAR field describes:

A) Read mapping quality

B) Alignment operations (matches, insertions, deletions)

C) Sequencing depth

D) Barcode identities

Answer: B

9. Which file contains annotated variants?

A) FASTQ

B) BAM

C) VCF

D) GFF

Answer: C

10. Which software checks GC distribution, duplication levels, and per-base quality?

A) IGV

B) GATK

C) FASTQC

D) AUGUSTUS

Answer: C

Section C – Genome Assembly & Annotation

11. Which graph approach underlies most short-read assemblers?

A) Overlap-Layout-Consensus (OLC)

B) De Bruijn graph

C) String graph

D) Greedy algorithms

Answer: B

12. N50 is defined as:

A) Average read length

B) Median genome coverage

C) Contig length at which 50% of assembly is contained

D) The number of contigs covering 50% of the genome

Answer: C

13. BUSCO analysis measures:

A) GC content

B) Read quality

C) Presence of conserved single-copy orthologs

D) Mapping quality

Answer: C

14. In gene annotation, repeat masking is important because:

A) Repeats mimic coding sequences

B) Repeats increase Q scores

C) Repeats lower sequencing coverage

D) Repeats improve contiguity

Answer: A

15. AUGUSTUS is an example of:

A) Homology-based annotation

B) Ab initio gene prediction

C) Variant discovery

D) Population genomics tool

Answer: B

Section D – Variant Discovery & Functional Genomics

16. SNPs are defined as variants with frequency:

A) ≥ 0.1%

B) ≥ 0.5%

C) ≥ 1%

D) ≥ 5%

Answer: C

17. Which alignment tool uses Burrows-Wheeler Transform?

A) Bowtie

B) BWA

C) PLINK

D) BUSCO

Answer: B

18. Which sequencing strategy enriches coding regions using biotinylated probes?

A) RNA-seq

B) ChIP-seq

C) Whole Exome Sequencing

D) RAD-seq

Answer: C

19. In bisulfite sequencing, unmethylated cytosines are read as:

A) C

B) T

C) G

D) A

Answer: B

20. ChIP-seq identifies:

A) DNA methylation sites

B) Protein-DNA interaction regions

C) Copy number variations

D) Alternative splicing events

Answer: B

Section E – Population Genomics & GWAS

21. The Hardy-Weinberg formula for genotype frequencies is:

A) $p + q = 1$

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

C) $2pq$ = heterozygosity only

D) FST = (HT – HS) / HT

Answer: B

22. A population bottleneck typically results in:

A) Increased heterozygosity

B) Loss of rare alleles and reduced variability

C) Higher mutation rates

D) Increased effective population size

Answer: B

23. FST = 1 means:

A) Populations are identical

B) Populations share alleles at equal frequencies

C) Populations are completely differentiated

D) No recombination occurred

Answer: C

24. A Manhattan plot is generated in:

A) RNA-seq

B) GWAS

C) Variant annotation

D) Genome assembly

Answer: B

25. Runs of Homozygosity (ROH) are mainly used to:

A) Measure sequencing coverage

B) Estimate inbreeding levels

C) Detect methylated cytosines

D) Assemble highly heterozygous genomes

Answer: B

Part 2 – Short Open Questions (5)

1. Explain the concept of depth vs. breadth of coverage in sequencing and why both are important.

Answer:

- Depth: average number of times each base is sequenced (e.g., 30X). Important for detecting SNPs reliably.
- Breadth: proportion of genome covered at least to target depth. Important for ensuring completeness (e.g., 95% covered at ≥10X).

---

2. What are the main challenges of de novo genome assembly in polyploid or highly heterozygous organisms?

Answer:

High heterozygosity leads to divergent haplotypes that assemblers may collapse or discard, resulting in fragmented assembly. Polyploidy increases complexity, requiring higher coverage and careful separation of homologous sequences.

---

3. Describe how RAD-seq reduces sequencing costs while still enabling population studies.

Answer:

Restriction enzymes cut DNA, generating a reduced subset of fragments. Only these are sequenced, lowering costs but still capturing informative SNPs for population-level analyses.

---

4. What is linkage disequilibrium (LD), and why is it important in GWAS?

Answer:

LD = alleles at nearby loci are inherited together more often than expected. Important in GWAS because SNPs in LD with causal variants appear associated with traits, enabling detection of disease loci.

---

5. How does RNA-seq contribute to functional annotation and transcriptome analysis?

Answer:

RNA-seq reveals expressed genes, exon-intron boundaries, isoforms, and expression levels. It validates computational predictions and identifies functional elements beyond protein-coding genes (e.g., lncRNAs, alternative splicing).

📘 Applied Genomics – Detailed Cheat Sheet

🔬 1. Sequencing Technologies & Strategies

Ion Torrent (ThermoFisher)

- Detects H+ ions released during nucleotide incorporation → pH change.
- Uses nanowells on a semiconductor chip, each with a bead carrying clonally amplified DNA.
- Workflow:
    1. DNA fragmentation (200–300 bp) + adapters.
    2. Emulsion PCR → each droplet contains DNA fragment + bead.
    3. Nucleotide flows → incorporation detected by voltage change.
- Output: Ionogram (bar height = number of bases incorporated).
- Weakness: homopolymers miscalled.
- Quality: avg Q ~20.

454 Roche (discontinued)

- “Mother” of Ion Torrent. Detects pyrophosphate + luciferase → light.
- Expensive (optical detection).

ABI SOLiD (dying)

- Based on ligation of fluorescent probes (dinucleotides).
- Very accurate, very short reads (~50 bp).

Illumina (dominant)

- Short reads (100–300 bp), very accurate.
- Workflow:
    1. DNA fragmentation + adapters (including barcode).
    2. Flow cell: DNA binds via capture sequences.
    3. Bridge amplification → clusters.
    4. Sequencing by synthesis: nucleotides with reversible terminator + fluorophore.
    5. Camera detects color → base call.
- Paired-end sequencing: both ends read → easier assembly, better variant calling.
- Innovations: two-color, one-color chemistries to reduce cost.

Oxford Nanopore

- DNA passes through protein nanopore, current disruption measured.
- Direct sequencing of DNA or RNA (no PCR, no cDNA).
- Very long reads (5–100 kb).
- Device is portable (MinION).
- Weakness: higher error rate (especially indels), requires intact DNA.

PacBio SMRT (Single Molecule Real-Time)

- DNA circularized with adapters → polymerase sits at bottom of well.
- Nucleotides with fluorescent tag at phosphate tail incorporated → signal detected.
- CCS (circular consensus sequencing) reduces error by reading the same molecule multiple times.
- Long reads (>20 kb), very accurate.
- Can detect base modifications (polymerase pausing).

Strategies

- De novo sequencing: No reference, requires long reads, hybrid approaches common.
- Resequencing: Align reads to existing reference (cheaper).
- Targeted sequencing: PCR/hybridization to enrich specific regions.
- Exome sequencing: Capture exons (1–2% of genome, most functional variants).
- Pool-seq: Mix DNA from many individuals, reduces cost but loses individual genotypes.
- RNA-seq: Sequence transcriptome, measure expression, support annotation.
- Methyl-seq: Bisulfite treatment converts C→T if unmethylated (methylated Cs remain).
- ChIP-seq: Identify protein–DNA interactions with immunoprecipitation.
- RAD-seq: Restriction enzymes reduce genome complexity, cost-efficient for SNP discovery.

🧩 2. Genome Assembly

Challenges:

- Large genome size.
- High heterozygosity (allelic variation).
- Repeats (collapse or branching in graphs).
- Polyploidy → assembly fragmentation.

Graph algorithms:

- Greedy (obsolete): extend reads one by one.
- Overlap-Layout-Consensus (OLC): all-vs-all overlaps, used in long-read assembly.
- De Bruijn Graph (DBG): short-reads → k-mers as nodes, overlaps as edges. Efficient, widely used.
- String graph: variation of OLC.

Mate-pair sequencing: Generates longer distance links (2–15 kb) to span repeats.

Assembly quality metrics:

- N50: contig/scaffold length at which 50% of total assembly length is contained.
- Coverage (depth): LN / G (read length × number of reads ÷ genome size).
- Completeness: % of expected genome recovered.
- BUSCO: checks presence of universal single-copy orthologs.

Gap-filling & scaffolding:

- Optical mapping (BioNano).
- Linked reads (10X Chromium).
- Hi-C (chromatin contacts).

Computational resources:

- Large genomes (1 Gb diploid) → ~96 CPUs, 1 TB RAM.

🧬 3. Genome Annotation

Steps:

1. Repeat identification & masking
    - Tools: RepeatMasker, REPBASE, TEdenovo.
    - Mask repeats as "N" to avoid misannotation.
2. Structural annotation
    - Ab initio tools: AUGUSTUS (predict ORFs using models).
    - Extrinsic evidence: RNA-seq, protein homology.
    - Hybrid approach: combine both.
3. Functional annotation
    - Assign function via homology, machine learning, or GO terms.
    - Annotate non-coding RNAs, pseudogenes too.
4. Manual curation
    - Human review important for accuracy.

Annotation file formats:

- GFF (General Feature Format): 9 columns (seqname, source, feature, start, end, score, strand, frame, attributes).
- GTF: variant of GFF.
- BED: simpler, chrom, start, end (+ optional).
- EMBL/GenBank: submission formats.

📂 4. File Formats in NGS

FASTQ:

- 4 lines per read:

1. @identifier
2. sequence
3. + (optional same ID)
4. ASCII-encoded quality string.

BAM/SAM:

- SAM = human-readable, BAM = binary compressed.
- Header: reference info.
- Fields: read name, flag, reference, position, mapping quality, CIGAR, sequence, qualities.
- FLAG field: encodes mapping info (paired, unmapped, duplicate, etc.).
- CIGAR: alignment string (M=match, I=insertion, D=deletion, S=soft-clip). Example: 10M1I5M.

VCF (Variant Call Format):

- Header: describes fields.
- Columns: CHROM, POS, ID (RS#), REF, ALT, QUAL, FILTER, INFO, FORMAT, sample genotypes.
- FILTER indicates if variant passed thresholds.

GFF/GTF: structural annotation.

BED: simpler coordinates for genome browsers.

📊 5. Quality Control & Metrics

DNA quality:

- 260/280 ~1.8 (pure DNA, proteins contaminate if lower).
- 260/230 ~2–2.2 (carbohydrate contamination lowers it).
- Integrity: gel electrophoresis.

Sequencing QC:

- FASTQC: per-base quality (boxplots), GC distribution, sequence duplication.
- Prinseq: additional trimming.

Trimming strategies:

- Fixed cutoff (trim until Q > threshold).
- Sliding window average Q.

Coverage:

- Depth (X): average times a base is read.
- Breadth: % genome covered at least to target depth.

N50: contiguity metric.

BUSCO: genome completeness.

IGV: visual exploration of BAM/VCF.

Graphs:

- Quality plots (boxplot of Q vs. position).
- GC distribution curve (should be bell-shaped).
- Manhattan plot (GWAS results).

🌍 6. Population Genomics & GWAS

Hardy-Weinberg equilibrium:

- $p + q = 1$
- Genotypes: $p^2 + 2pq + q^2 = 1$
- Assumes no selection, drift, migration, mutation, non-random mating.

Genetic drift & bottleneck:

- Small populations lose rare alleles, diversity drops.

Inbreeding:

- F-ped = pedigree-based.
- F-ROH = fraction of genome in ROH (from SNP data).

Linkage Disequilibrium (LD):

- Alleles at nearby loci inherited together.
- D, D', r^2 measure LD strength.

Fixation index (FST):

- FST = (HT – HS) / HT
- 0 = identical, 1 = fully separated.

GWAS:

- Large sample size.
- Requires phenotype definition, correction for population structure.
- Output: Manhattan plot.

Applications in breeding:

- Marker-assisted selection.
- Genomic selection.
- Parentage verification.

Technology	Read length	Accuracy	Detection principle	Strengths	Weaknesses	Best use
Illumina	100–300 bp	Very high (>99.9%)	Fluorescent reversible terminator nucleotides	High throughput, cheap per base, paired-end	Short reads, assembly in repeats difficult	Resequencing, RNA-seq, GWAS
Ion Torrent	100–400 bp	Moderate (Q~20)	pH change (H+ release)	Compact, fast	Homopolymer errors	Small genomes, targeted sequencing
454 Roche	400–700 bp	Good	Pyrophosphate → luciferase light	Longer than Illumina	Costly, discontinued	Historical use

• Technology	Read length	Accuracy	Detection principle	Strengths	Weaknesses	Best use
SOLiD	~50 bp	Very high	Ligation with fluorescent dinucleotide probes	Accuracy	Very short reads, expensive, discontinued	Niche legacy
PacBio SMRT	10–20 kb (CCS >20 kb)	High with CCS	Fluorescent label at phosphate tail	Long reads, structural variants, methylation detection	Expensive, needs good DNA	De novo assembly, epigenetics
Oxford Nanopore	5–100 kb	Lower (improving)	DNA through nanopore, current disruption	Ultra-long reads, portable, direct RNA	High indel rate, DNA integrity required	Field sequencing, metagenomics

Algorithm	Principle	Use case	Pros	Cons
Greedy	Extend overlaps one by one	Early sequencing	Simple	Fails with repeats
OLC (Overlap-Layout-Consensus)	Find overlaps → build consensus	Long reads (PacBio, Nanopore)	Accurate	Computationally heavy
De Bruijn Graph	K-mers → nodes, overlaps → edges	Short reads (Illumina)	Efficient	Errors in k-mers → false branches
String Graph	Simplified OLC	Hybrid approaches	Balanced	Complex implementation

Step	Tools	How they work
Repeat masking	RepeatMasker, TEdenovo	Library or de novo recognition of transposons/repeats
Structural annotation	AUGUSTUS (ab initio): predicts ORFs by models; RNA-seq evidence: align transcripts	Predict exons, splice sites
Functional annotation	BLAST, InterProScan, GO terms	Assign function via homology & conserved domains
Completeness check	BUSCO	Detects conserved single-copy orthologs

Format	Purpose	Key columns
FASTQ	Raw reads + quality	ID, sequence, +, quality string (ASCII Phred)
SAM/BAM	Read alignments	QNAME, FLAG, RNAME, POS, MAPQ, CIGAR, SEQ, QUAL
VCF	Variants	CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT, sample genotypes
GFF/GTF	Annotation	seqname, source, feature, start, end, score, strand, frame, attributes

Format	Purpose	Key columns
BED	Simple genome coordinates	chrom, start, end (+ optional annotations)

Metric	What it measures	Good value	Interpretation
260/280 ratio	DNA purity vs proteins	~1.8	<1.6 = protein contamination
260/230 ratio	DNA vs carbohydrates	2–2.2	Low = contamination (plants)
Coverage depth (X)	Avg bases read	≥30X for human	LN/G
Coverage breadth	% genome covered	≥95%	Uniformity
N50	Assembly contiguity	Larger = better	Median contig length
BUSCO score	Completeness	>90%	Missing or duplicated genes = assembly issue
Phred Q score	Base call quality	Q30 = 1 error/1000 bases	ASCII encoded

Concept	Definition	Formula / Graph	Use
Hardy–Weinberg equilibrium	Allele/genotype frequencies remain constant absent evolutionary forces	$p^2 + 2pq + q^2 = 1$	Baseline population genetics
Genetic drift	Allele loss in small pops	—	Explains bottlenecks
Inbreeding (F-ped, F-ROH)	Homozygosity by descent	Runs of Homozygosity	Breeding, conservation
Linkage Disequilibrium (LD)	Alleles co-inherited more than expected	D, D', r^2	GWAS signal interpretation
FST	Population differentiation	(HT – HS)/HT	0=identical, 1=separated
Manhattan plot	SNP associations in GWAS	x = SNPs, y = –log10(p)	Identify loci for traits

Exercises

A. Hardy–Weinberg

A population has allele A (p=0.7) and allele a (q=0.3).

- Expected genotype frequencies?

Solution:

- $AA = p^2 = 0.49$
- $Aa = 2pq = 0.42$
- $aa = q^2 = 0.09$

B. Coverage Calculation

Genome = 3 Gb, reads = 600M, read length = 150 bp.

- What is coverage?

Solution:

Coverage = $(L \times N)/G = (150 \times 600M)/3Gb = (90\ Gb / 3\ Gb) = 30X$.

C. Manhattan Plot Interpretation

You see a peak on chromosome 6 with $-\log10(p) = 12$.

- What does this mean?

Solution:

- Variant strongly associated with phenotype.
- Likely causal locus nearby (but nearby SNPs in LD may also appear).

D. ROH (Runs of Homozygosity)

If 20% of a genome is in ROH, what does this suggest?

Solution: High inbreeding → individual likely from small or consanguineous population.

E. N50 Example

Assembly contigs: 100 kb, 80 kb, 50 kb, 20 kb. Total = 250 kb.

- N50?

Solution: Half genome = 125 kb.

Contigs sorted: 100+80=180 kb (>125). N50 = 80 kb.

Chapter 1. Genetics Foundations

Genetics is the basis of genomics, and without understanding inheritance, variation, and population structure, the massive data from sequencing would have no biological meaning.

Classical genetics describes inheritance patterns. Mendelian laws and pedigree analyses helped identify dominant and recessive traits, while recombination frequencies enabled genetic maps. Different species have different sex-determination systems (XY in humans, ZW in birds, haplodiploidy in bees), all relevant for interpreting inheritance patterns.

Molecular genetics added the ability to study DNA directly through cloning, restriction enzymes, and PCR. Sequencing technologies grew out of these approaches.

Population genetics focuses on how alleles behave across generations. At equilibrium, under no evolutionary forces, allele frequencies follow Hardy–Weinberg equilibrium:

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

where p and q are allele frequencies. Deviations point to selection, drift, migration, mutation, or non-random mating.

Quantitative genetics extends this to complex traits, governed by many loci and influenced by the environment. Here concepts such as heritability and polygenic scores are critical.

Applied genomics is where all of these branches converge: sequencing gives us the raw data, assembly and annotation give us reference genomes, and population and quantitative genetics allow us to interpret variation in terms of traits and evolution.

Chapter 2. Sequencing Technologies

Modern genomics was revolutionized by Next Generation Sequencing (NGS) platforms. Each differs in chemistry, read length, error profile, and cost.

Illumina dominates with short, accurate reads. DNA is fragmented, ligated to adapters, captured on a flow cell, and amplified into clusters by bridge amplification. Sequencing occurs cycle by cycle: a fluorescent reversible terminator nucleotide is added, the base is imaged, then the terminator is cleaved. Reads are typically 100–300 bp with >99.9% accuracy. Paired-end sequencing reads both ends of a fragment, aiding assembly and variant discovery. The limitations are short reads and GC bias, which can hinder assembly.

Ion Torrent detects H+ ions released during DNA synthesis. Each well contains a bead with amplified DNA, and voltage changes reflect nucleotide incorporations. The advantage is speed and cost, but homopolymer runs (AAAA…) are hard to distinguish, leading to errors.

454 Roche, a precursor, used pyrophosphate release detected by luciferase (light). It gave longer reads than early Illumina but was costly and is now discontinued.

SOLiD used sequencing by ligation of fluorescent probes. Accuracy was high, but reads were short, and the platform is obsolete.

PacBio SMRT (Single Molecule Real-Time) attaches a DNA polymerase to a zero-mode waveguide (tiny well). Each nucleotide has a fluorescent label on the phosphate tail; when incorporated, the

flash is recorded. DNA molecules are circularized so the same template can be read multiple times (circular consensus sequencing, CCS). PacBio generates reads >20 kb, highly accurate with CCS, and can detect base modifications like methylation.

Oxford Nanopore passes DNA or RNA through a protein nanopore in a membrane, measuring electrical current changes. It is unique in being portable (MinION) and can directly sequence RNA without conversion to cDNA. Reads can exceed 100 kb, but the error rate (especially indels) is higher.

Sequencing Strategies

- De novo sequencing: assembling a genome without reference; requires long reads.
- Resequencing: aligning to a known reference; Illumina excels here.
- Targeted sequencing: enriching specific loci with PCR or hybridization.
- Whole exome sequencing (WES): enriches coding regions (1–2% of genome).
- RNA-seq: transcriptome profiling, revealing expression and isoforms.
- ChIP-seq: finds DNA–protein binding sites.
- Methyl-seq: bisulfite converts unmethylated cytosines to uracil (→ thymine).
- RAD-seq: restriction enzymes reduce complexity; cost-effective for SNPs.
- Pool-seq: pools DNA from individuals, giving allele frequencies at low cost.

Chapter 3. Genome Assembly

Sequencing reads are fragments; assembly reconstructs the genome.

Challenges include repeats (which collapse short reads), heterozygosity (alternative haplotypes confuse assemblers), and polyploidy (multiple genome copies).

Algorithms:

- Greedy: extend reads sequentially; obsolete.
- Overlap–Layout–Consensus (OLC): all-vs-all overlaps → layout graph → consensus; used with long reads.

- De Bruijn Graph (DBG): k-mers as nodes, overlaps as edges; efficient for short reads (Illumina). Errors in reads introduce false branches.
- String Graph: refined OLC, avoids redundancy.

Metrics:

- Coverage depth: \frac{L \times N}{G}, where L = read length, N = reads, G = genome size.
- Breadth of coverage: proportion of genome covered.
- N50: contig length such that 50% of assembly lies in contigs ≥ that length.
- BUSCO: evaluates completeness by searching for conserved orthologs.

Improvement strategies include mate-pair sequencing (longer links), PacBio/Nanopore long reads, and scaffolding with BioNano optical maps, 10X linked reads, or Hi-C contact maps.

Chapter 4. Genome Annotation

An assembled genome is just sequence; annotation gives it meaning.

Steps:

1. Repeat masking: detect repeats with tools like RepeatMasker or TEdenovo; masking prevents mis-annotation.
2. Structural annotation:
    - Ab initio (e.g., AUGUSTUS) uses models to predict exons, introns, and coding sequences.
    - Evidence-based uses RNA-seq and protein homology to support gene models.
    - Hybrid approaches combine both.
3. Functional annotation: BLAST or InterProScan assign gene functions, pathways, GO terms. Non-coding RNAs and pseudogenes also annotated.
4. Manual curation: human review improves accuracy.

Annotation formats:

- GFF/GTF: 9-column format (chrom, source, feature, start, end, score, strand, frame, attributes).

- BED: simple coordinates.
- GenBank/EMBL: submission standards.

Chapter 5. NGS File Formats & QC

FASTQ stores raw reads: identifier, sequence, plus line, quality string encoded as ASCII Phred scores.

SAM/BAM store alignments. Fields include read name, flag (mapping info), reference name, position, mapping quality, and CIGAR string (e.g., 8M1I4M = 8 matches, 1 insertion, 4 matches). BAM is compressed binary.

VCF stores variants: columns include CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT, and sample genotypes. FILTER indicates whether thresholds were passed.

GFF/GTF describe annotation; BED provides simple coordinates.

QC metrics:

- 260/280 ratio ~1.8 (DNA purity).
- 260/230 ratio ~2.0 (low values = polysaccharide contamination).
- Phred Q scores: Q30 = 1 error in 1000 bases.
- FASTQC reports base quality, GC content, duplication.
- Prinseq trims and filters.
- IGV visualizes reads and variants.

Chapter 6. Population Genomics & GWAS

Hardy–Weinberg Equilibrium provides a null expectation: $p^2 + 2pq + q^2 = 1$. Deviations reveal forces like selection, drift, or inbreeding.

Genetic drift and bottlenecks reduce diversity by random allele loss. Inbreeding increases homozygosity, measurable via F-statistics or Runs of Homozygosity (ROH).

Linkage Disequilibrium (LD): alleles at nearby loci inherited together. LD is essential for GWAS: even if the causal SNP isn't genotyped, nearby SNPs in LD will show association.

Fixation Index (FST): measures population differentiation. FST = 0 means identical, FST = 1 means completely distinct.

GWAS (Genome-Wide Association Studies): genotype many individuals at millions of SNPs, correlate with phenotype. Requires large samples, correction for multiple testing (Bonferroni). Results are visualized in Manhattan plots: x-axis = SNPs along genome, y-axis = $-\log_{10}(p)$. Peaks show candidate loci.

Applications include identifying disease loci in humans, traits in crops, and selection signatures in domestic animals.

Chapter 7. Key Metrics & Graphs

- Phred Q Score: $Q = -10 \log_{10}(e)$. Q30 = 1 error/1000 bp.
- Coverage depth & breadth: ensure confidence and completeness.
- N50: measures contiguity.
- BUSCO: completeness by conserved genes.
- GC plots: detect bias.
- Manhattan plots: GWAS results.
- IGV screenshots: show alignments, coverage, variants.

Applied Genomics – Unified Compendium (Block 1)

Genomics as a discipline is built on the foundations of genetics, because the very reason we sequence genomes is to understand how variation at the DNA level translates into inheritance, traits, and evolution. Classical genetics, originating with Mendel, showed us that traits are passed in discrete units, alleles, and that segregation and independent assortment determine their distribution in offspring. This classical view allowed the construction of linkage maps, where recombination frequency between loci was used as a proxy for distance on a chromosome. Already here we see the beginnings of genomics: the genome is an encoded map, and its decoding requires both experimental and statistical tools.

But genetics is more than transmission; it also concerns populations. Population genetics describes how allele frequencies behave across generations. The Hardy–Weinberg principle is the baseline: if allele A has frequency p and allele a has frequency q, then under random mating and in the absence of evolutionary forces, the genotype frequencies will stabilize as p^2 for AA, 2pq for Aa, and q^2 for aa. This equilibrium provides a null model. Whenever real populations deviate, it tells us that something is happening — drift, selection, migration, mutation, or non-random mating. Quantitative genetics adds another layer, where traits are not binary but continuous, influenced by many loci plus the environment. Here the concept of heritability becomes central: it quantifies how much of the phenotypic variance is due to genetic variance. This is where genomics becomes essential, because identifying the genetic contribution to complex traits requires scanning across the entire genome.

However, classical and population genetics could only hypothesize about genetic material until molecular genetics revealed the physical basis: DNA. Once Watson and Crick described its double helix structure, and molecular biology provided tools such as restriction enzymes and PCR, it became possible to directly manipulate and analyze DNA. Yet, to fully grasp genetic variation, one needs to sequence — to read the nucleotide letters in order. That is where genomics departs from genetics: sequencing does not just study inheritance indirectly; it decodes the text of life.

The first sequencing method was Sanger sequencing, based on chain termination by dideoxynucleotides. It could only process fragments hundreds of bases long, but it laid the foundation for the first human genome project. Applied genomics today, however, is powered by high-throughput sequencing, known as Next Generation Sequencing (NGS). Here multiple technologies coexist, each exploiting a different physical principle but all aiming at the same goal: reading DNA at scale.

One of the most widely used is Illumina sequencing, which combines parallelization with accuracy. The process begins by fragmenting DNA and ligating adapters. These fragments are fixed to a glass flow cell coated with complementary oligonucleotides. Each fragment bends over, hybridizing to a nearby oligo, forming a “bridge,” and then is replicated by polymerase. This “bridge amplification” repeats until a dense cluster of identical molecules forms, producing enough signal to detect a single base incorporation. Sequencing proceeds cycle by cycle: all four nucleotides are added, but each has a fluorescent label and a reversible terminator that prevents more than one incorporation. A camera records the color of the incorporated base across millions of clusters simultaneously, after which the terminator and dye are chemically cleaved, and the next cycle begins. This ensures one base at a time is read, giving short but highly accurate reads. With paired-end sequencing, both ends of a fragment are read, effectively providing spatial information that helps resolve repeats and structural variants. Illumina’s weakness is its short reads (100–300 bp) and biases in GC-rich regions, but its accuracy (>99.9%) makes it ideal for resequencing, RNA-seq, and genome-wide association studies.

Other platforms pursue different strategies. Ion Torrent dispenses with optics altogether and measures hydrogen ions released during nucleotide incorporation. Each DNA fragment is attached to a bead and amplified by emulsion PCR, then deposited into a well on a semiconductor chip. When polymerase incorporates a nucleotide, a proton is released, causing a pH change detected as voltage. The strength of the signal reflects the number of bases incorporated. This eliminates the need for lasers and cameras, making the system compact and fast, but it struggles with homopolymer stretches, where distinguishing six A’s from seven becomes error-prone.

Earlier, 454 Roche sequencing used pyrosequencing: incorporation of a nucleotide released pyrophosphate, which triggered a luciferase reaction producing light. This method generated longer reads than early Illumina, but the high cost and competition led to its discontinuation. Another discontinued system, SOLiD, used sequencing by ligation with fluorescently labeled probes. Its accuracy was extremely high, but read lengths were very short (~50 bp), and its complex chemistry limited adoption.

Modern long-read platforms addressed the shortcomings of short-read methods. PacBio Single Molecule Real-Time (SMRT) sequencing attaches a DNA polymerase to the bottom of a zero-mode waveguide, a tiny well illuminated only at its base. Each nucleotide carries a fluorescent label on its phosphate tail; when incorporated, the label flashes before being cleaved. Because each DNA molecule is circularized, the polymerase can read it multiple times, and consensus can be built (CCS, circular consensus sequencing). This yields reads tens of kilobases long, highly accurate after correction, and capable of detecting modifications like methylation. The challenge is that high molecular weight DNA is required, and the cost is higher than short-read sequencing.

Oxford Nanopore Technologies took a radical step further by forgoing synthesis altogether. DNA or RNA is threaded through a protein pore embedded in a membrane. As each nucleotide passes, it changes the ionic current in a characteristic way. Machine learning translates these current fluctuations into base calls. Nanopore devices range from small, portable MinIONs to larger

PromethIONs, enabling sequencing anywhere, even in the field. Reads can exceed 100 kb, sometimes reaching megabase length, capturing entire chromosomes or transcripts in a single pass. Nanopore can also directly sequence RNA, preserving modifications like methylation. Its limitation is accuracy, particularly with insertions and deletions, but with depth and polishing algorithms, this is improving rapidly.

Different technologies naturally suit different sequencing strategies. When building a genome without a reference, de novo sequencing requires long reads to resolve repeats and assemble contiguously. Resequencing projects, by contrast, aim to compare an individual's genome to an existing reference; short Illumina reads suffice, because alignment to a known scaffold simplifies interpretation. Targeted sequencing enriches particular loci using PCR or hybrid capture, while whole exome sequencing focuses on coding regions (~1–2% of the genome), reducing cost for medical applications. RNA-seq sequences transcripts to measure gene expression and alternative splicing, ChIP-seq identifies protein–DNA binding sites, and methyl-seq uses bisulfite conversion to distinguish methylated cytosines. In ecological and evolutionary studies, reduced representation approaches like RAD-seq sequence only fragments near restriction sites, giving a cost-effective snapshot of genetic variation across populations. Pool-seq takes this further by mixing DNA from many individuals and sequencing them together, providing allele frequency estimates without the cost of individual genotyping.

Thus, from the basic need to study inheritance, we arrive at a suite of sequencing strategies and technologies, each designed to balance read length, accuracy, throughput, and cost. And this is where the next challenge emerges: once we have the reads, how do we put them back together into a genome?

Applied Genomics – Unified Compendium (Block 2)

Sequencing technologies give us billions of short or long fragments, but by themselves these are like sentences cut into pieces and shuffled. The task of genome assembly is to reconstruct the original book from those fragments. This is not trivial, because genomes are filled with repeated elements, heterozygosity, and complex structures.

At its heart, assembly is about overlaps. If one read ends with ATCGT and another begins with TCGTA, then they likely came from adjacent positions in the genome. Early assemblers simply extended overlaps greedily, but this breaks down when repeats create multiple possible continuations. Modern algorithms instead use graphs. For long reads, the Overlap–Layout–Consensus (OLC) approach dominates: every read is compared to every other to find overlaps,

these overlaps are used to build a layout graph that orders the reads, and then a consensus sequence is derived. For short reads, where comparing every pair is computationally infeasible, the De Bruijn graph is used. Here, reads are chopped into k-mers (short words of length k). Each k-mer becomes a node in the graph, and edges connect nodes that overlap by k–1 bases. Traversing this graph reconstructs the genome. But errors in reads produce false branches, and repeats collapse into a single path, making assembly ambiguous. To address this, assemblers apply error correction, coverage filtering, and paired-end linking information.

Another refinement is the String Graph, which is similar to OLC but avoids redundant overlaps and is efficient for large long-read datasets. Still, assembly is never perfect. Repetitive transposons, segmental duplications, and highly heterozygous regions often fragment contigs or create misassemblies.

To evaluate assemblies, several metrics are used. The most fundamental is coverage depth, calculated as \frac{L \times N}{G}, where L is read length, N the number of reads, and G the genome size. High coverage increases confidence, but beyond ~30× for short reads, the returns diminish. Equally important is breadth of coverage, the fraction of the genome covered by at least one read. Another key metric is N50, defined as the contig length at which 50% of the genome assembly is contained in contigs of at least that length. A higher N50 suggests greater contiguity, though it does not guarantee correctness. To assess completeness, tools like BUSCO (Benchmarking Universal Single-Copy Orthologs) search for sets of conserved genes expected to be present in nearly all members of a lineage. Missing BUSCOs imply incomplete assembly, while duplicated BUSCOs may reflect assembly errors or genuine gene duplications.

Because assemblies often remain fragmented, additional strategies are employed for scaffolding. Mate-pair libraries, which have long insert sizes, provide information about the relative position and orientation of contigs. Long-read technologies like PacBio and Nanopore can span repeats and close gaps. Optical mapping (BioNano), linked-read technologies (10X Genomics), and Hi-C contact maps all add long-range information that helps order and orient contigs into chromosome-scale scaffolds. Thus, assembly is not a single step but an iterative process: initial contigging, scaffolding, polishing with short reads to correct errors, and finally quality assessment.

Yet an assembled genome is still just a string of nucleotides. To extract biological meaning, it must be annotated. Annotation is the process of identifying which regions correspond to genes, regulatory elements, repeats, and functional units.

The first step is repeat annotation and masking. Repeats can make up the majority of a eukaryotic genome, especially in plants. If left unmasked, they can be misidentified as protein-coding genes. Tools like RepeatMasker, using libraries such as REPBASE, scan for known repetitive elements. De novo repeat finders like TEdenovo build custom libraries from the genome itself. Masked

repeats are typically replaced with “N” or lowercased bases to prevent confusion in downstream analyses.

Next comes structural annotation, which predicts the architecture of genes. There are three main approaches. Ab initio prediction uses statistical models trained on known gene structures to identify open reading frames, splice sites, start and stop codons, and coding potential. AUGUSTUS is a classic tool in this category. Evidence-based annotation aligns external data — RNA-seq reads, ESTs, or proteins from related species — to provide experimental support for gene models. Finally, hybrid pipelines combine both: ab initio predictions are refined with evidence, producing more accurate results.

Once gene structures are defined, functional annotation assigns meaning. Sequence similarity searches with BLAST identify homologs of known genes. Domain databases like InterProScan reveal conserved motifs, while Gene Ontology terms classify biological processes, molecular functions, and cellular components. Pathway databases link genes to metabolism, signaling, or regulatory networks. Non-coding RNAs, pseudogenes, and other features are also annotated, ensuring that the genome reflects not only protein-coding genes but the full spectrum of genetic elements.

Annotation is not purely computational. Automated pipelines are powerful but prone to false positives and missed features. Manual curation by experts — inspecting alignments in genome browsers, confirming gene models, correcting splice boundaries — remains the gold standard for high-value genomes like model organisms or clinical references.

To manage and share annotations, standardized file formats are essential. GFF and GTF files record features in nine columns: sequence ID, source, feature type (e.g., exon, CDS), start, end, score, strand, frame, and attributes (such as gene ID or transcript ID). BED files provide a simpler, three-column format (chromosome, start, end) used for genome browsers. Richer formats like GenBank or EMBL include both sequence and annotation metadata, suitable for public database submission.

Thus, the journey from raw reads to a functional genome passes through assembly and annotation, two steps that transform data into knowledge. Sequencing provides fragments, assembly rebuilds the genome, and annotation interprets it. And yet, all of this work relies on the assumption that our data are accurate — which is why the next concern in genomics is quality control and file integrity.

Applied Genomics – Unified Compendium (Block 3)

Once genomes are assembled and annotated, a fundamental question arises: how reliable are our data and predictions? Quality control (QC) is the backbone of applied genomics, because sequencing errors, contamination, or format inconsistencies can propagate and distort biological conclusions.

The process begins at the very start, with DNA extraction. The purity of DNA is measured spectrophotometrically: the ratio of absorbance at 260 nm and 280 nm (A260/280) should be close to 1.8 for pure DNA. Lower values indicate protein contamination. The A260/230 ratio should be ~2.0–2.2; lower values signal contamination with polysaccharides or phenolic compounds, common in plant extracts. These simple numbers already tell us whether sequencing will succeed or fail.

When DNA is sequenced, the first data produced are FASTQ files, the raw format of reads. Each entry has four lines: an identifier (with instrument and run information), the sequence itself, a plus line (often just "+"), and a quality string. The quality string encodes Phred scores, which are logarithmic measures of error probability:

$$Q = -10 \log_{10} e$$

where e is the probability of error. A Q30 score means 1 error in 1000 bases, considered the gold standard for Illumina data. Quality scores are stored as ASCII characters, so "I" might mean Q40, while "#" means Q2, depending on the encoding offset.

Before assembly or mapping, FASTQ files are checked with tools like FASTQC, which plots base quality per cycle, GC content, sequence duplication, and overrepresented sequences (often adapters or contaminants). Trimming and filtering are performed with tools like Prinseq or Trimmomatic, removing low-quality bases or adapter contamination.

Once reads are aligned, results are stored in SAM or BAM files. SAM (Sequence Alignment/Map) is a tab-delimited text file, while BAM is its compressed binary form. Each line represents one read alignment, including fields such as:

- QNAME: read name.
- FLAG: a bitwise code describing mapping status (paired, mapped, reversed, etc.).
- RNAME: reference sequence name.
- POS: 1-based position on the reference.
- MAPQ: mapping quality.
- CIGAR: compact encoding of alignment (e.g., 8M1I4M means 8 matches, 1 insertion, 4 matches).
- SEQ/QUAL: the read sequence and quality.

Flags and CIGAR strings make SAM/BAM powerful but dense; genome browsers like IGV (Integrative Genomics Viewer) visualize them, showing alignments across loci, coverage depth, and mismatches.

Variants discovered from alignments are stored in VCF (Variant Call Format). A VCF has mandatory columns: chromosome, position, ID, reference allele, alternate allele(s), quality, filter, and INFO. Additional columns describe genotype calls for each sample. The FILTER field indicates whether a variant passed quality thresholds, while INFO can encode allele frequency, depth, or functional annotation. VCFs allow large-scale analysis of SNPs and indels across populations.

Annotation itself is stored in GFF or GTF files, as we saw, while BED files provide lightweight coordinates for display in genome browsers. These file formats are the lingua franca of genomics: FASTQ → SAM/BAM → VCF → GFF, a pipeline that underlies nearly every project.

But applied genomics is rarely about one genome in isolation. The real power comes from comparing individuals and populations. Here, population genomics connects sequencing to evolutionary and biomedical questions.

At the population level, the Hardy–Weinberg equilibrium serves as a null hypothesis. If allele A has frequency p and allele a has frequency q, then $p^2 + 2pq + q^2 = 1$. Deviations indicate forces at work. Genetic drift is one such force: in small populations, random sampling causes allele frequencies to fluctuate, sometimes leading to fixation or loss. A bottleneck, where a population is drastically reduced, permanently lowers genetic diversity. In contrast, migration introduces new alleles, while selection biases survival in favor of particular genotypes.

Inbreeding increases homozygosity beyond Hardy–Weinberg expectations. It can be quantified through pedigree-based coefficients or directly from the genome using Runs of Homozygosity (ROH) — long continuous stretches of homozygous genotypes. The proportion of the genome in ROHs reflects the inbreeding coefficient (F_ROH).

Another central concept is Linkage Disequilibrium (LD), the non-random association of alleles at different loci. If alleles A and B co-occur more often than expected from their individual frequencies, they are in LD. LD is crucial for association studies because even if a causal variant is not genotyped, nearby SNPs in LD with it will show a signal. LD decays with recombination distance, so it also provides information about population history and effective size.

Population differentiation is measured with FST, defined as (HT – HS) / HT, where HT is total heterozygosity and HS is average subpopulation heterozygosity. An FST close to 0 means populations are genetically similar, while values approaching 1 indicate strong differentiation, possibly reproductive isolation.

These principles culminate in Genome-Wide Association Studies (GWAS), which combine population genomics with quantitative genetics. In a GWAS, thousands of individuals are genotyped at millions of SNPs, and each SNP is tested for association with a trait. Because millions of tests are performed, correction for multiple testing is essential — commonly by Bonferroni correction or False Discovery Rate (FDR) methods. Population structure must also be accounted for; otherwise, differences between subpopulations may create spurious associations.

Results are visualized in the Manhattan plot: the x-axis represents genomic coordinates, typically grouped by chromosome, and the y-axis is –log10(p-value). Most SNPs hover near the baseline, but significant associations tower upward like skyscrapers, hence the name. The horizontal line marks the genome-wide significance threshold. Peaks highlight regions of interest that harbor candidate genes or linked variants influencing the trait.

GWAS has revealed loci for human diseases like diabetes and cancer, but also for agricultural traits such as yield, disease resistance, and drought tolerance. In livestock, it guides breeding programs by pinpointing markers linked to milk production, fertility, or disease resistance. Thus, the flow from sequencing → assembly → annotation → QC → population genomics culminates in GWAS, where genomic data finally connect back to phenotypes and real-world applications.

Applied Genomics – Unified Compendium (Block 4 – Synthesis)

Applied genomics is best understood not as a set of disconnected techniques, but as a continuous pipeline where each step provides the foundation for the next. The story begins with the biological question — inheritance, variation, evolution — and ends with actionable insights, whether they are gene functions, evolutionary patterns, or associations with traits.

At the base lies genetics, the study of how alleles are transmitted. Hardy–Weinberg equilibrium gives us the expectation for genotype frequencies in a neutral, random-mating population, and deviations point to forces such as selection, drift, or inbreeding. Quantitative genetics extends this to complex traits, introducing heritability and the idea that many loci of small effect shape continuous phenotypes. This conceptual framework justifies why we need dense genome-wide data in the first place: only by scanning the entire genome can we capture the polygenic architecture of traits.

Sequencing technologies provide that data. Short-read Illumina sequencing, with its bridge amplification and fluorescent terminators, yields billions of highly accurate reads, perfect for resequencing, RNA-seq, or GWAS. Long-read platforms like PacBio SMRT and Oxford Nanopore overcome the limitations of short reads by spanning repeats and structural variants, crucial for de novo assembly. Other systems like Ion Torrent detect protons, while historical platforms like 454 and SOLiD paved the way for today's methods. Each technology has its own read length, error profile, and niche application, but together they form a toolkit that can be tailored to any project.

Once reads are generated, assembly reconstructs the genome. Short reads are best handled by de Bruijn graphs, where k-mers form nodes connected by overlaps, while long reads use overlap-layout-consensus or string graphs to directly align fragments. Assembly metrics such as coverage depth, N50, and BUSCO completeness quantify how good a reconstruction is. Yet contigs rarely reach chromosome scale without additional scaffolding technologies like mate pairs, optical maps, or Hi-C. Still, the goal remains the same: to move from a soup of fragments to a coherent representation of the genome.

Annotation then transforms this raw sequence into biology. Repeats are identified and masked to prevent false positives. Gene models are predicted using ab initio algorithms like AUGUSTUS, refined with evidence from RNA-seq or homologous proteins, and then functionally annotated with BLAST and InterProScan. Functional terms, domains, and pathways are assigned, and manual curation polishes the models. The result is a catalog of coding and non-coding genes, regulatory elements, and genomic landmarks. Annotation formats like GFF, GTF, BED, or GenBank ensure this information is portable and interpretable by genome browsers and downstream tools.

All along, quality control ensures that errors do not undermine interpretation. DNA purity is checked with spectrophotometric ratios. Raw reads in FASTQ are inspected with Phred quality scores and FASTQC. Alignments are stored in SAM/BAM with flags and CIGAR strings encoding mapping details. Variants are described in VCF with fields for alleles, quality, and genotypes. These file formats structure the flow of data: from sequencing (FASTQ) to mapping (BAM) to variation (VCF) to annotation (GFF). QC is not an optional step but a continuous safeguard at every stage.

With annotated genomes in hand, genomics broadens to populations. Sequencing multiple individuals reveals allele frequency distributions, deviations from Hardy–Weinberg, and signatures of drift, selection, or migration. Inbreeding is quantified through Runs of Homozygosity, and population differentiation is captured by FST. Linkage disequilibrium, the non-random association of alleles at different loci, both complicates and enables analysis — complicating because nearby markers are not independent, enabling because even untyped causal variants can be detected via markers in LD.

These principles come together in Genome-Wide Association Studies, which scan the genome for statistical associations between SNPs and traits. By testing millions of markers across thousands of individuals, GWAS identifies loci underlying complex phenotypes. The results are distilled in Manhattan plots, where skyscraper-like peaks represent significant associations. Statistical rigor — multiple testing correction, population structure control — is essential to avoid false positives. GWAS closes the loop between sequence and phenotype, returning to the genetic questions that motivated genomics in the first place.

In this way, applied genomics is a cycle: genetics defines the questions, sequencing provides the data, assembly and annotation create the reference, QC maintains integrity, population genomics interprets variation, and GWAS connects it back to traits. Each tool, file format, and metric is part of this chain. Illumina, Nanopore, PacBio, De Bruijn graphs, AUGUSTUS, BLAST, FASTQ, BAM, VCF, BUSCO, N50, Hardy–Weinberg, FST, Manhattan plots — they are not isolated concepts, but links in a single continuum. To master applied genomics is to see not just the individual techniques, but the logic that binds them together: from molecules to populations, from data to meaning.

AG written exam

Multiple question (25 questions):

- Allele definition
- SOLiD sequencing principle
- FST definition
- Ion torrent
- LD
- ROH relation to inbreed degree
- Illumina size of reads
- GWAS aim
- What does it means if a population is not in HW eq: selection
- Population stratification with MDS
- Paired-end seq: sequencing the same fragment from both sides
- VCF file
- FASTQ file
- PACBIO: smrt sequencing
- CIGAR meaning
- Functional and structural annotation, what it means
- de Brujin graph algorithm: eulerian path
- Completeness of a genome: BUSCO
- Meaning of equimolar DNA pool

Short answer (5 questions):

- Depth of coverage calculation
- N50 calculation giving a the length of some contigs
- Explaining bisulfide sequencing
- How to estimate genome size before sequencing: c-value
- Explaining and drawing a Manhattan plot