Applied Genomics

Theoretical Foundations Handbook

Prepared for Advanced Genomics Study

August 15, 2025

Contents

1. Foundations of Genetics

1.1 Classical (Mendelian) Genetics

Overview

Classical genetics investigates how traits are inherited through generations. Initiated by Gregor Mendel, it describes how discrete genetic units (genes) segregate and assort independently.

Key Terms and Concepts

- **Gene:** A sequence of DNA that codes for a specific protein or functional RNA.
- **Allele:** A variant of a gene; individuals inherit one allele from each parent.
- **Genotype:** The genetic composition (e.g., AA, Aa, aa).
- **Phenotype:** The observable characteristic resulting from the genotype.
- **Homozygous:** Two identical alleles (AA or aa).
- **Heterozygous:** Two different alleles (Aa).
- **Dominant/Recessive:** Dominant alleles mask the expression of recessive ones.

Punnett Square Example

Cross: Aa × Aa (heterozygotes)

	A	a
A	AA	Aa
a	Aa	aa

Genotypic Ratio: 1 AA : 2 Aa : 1 aa
Phenotypic Ratio: 3 dominant : 1 recessive

Pedigree Analysis

Pedigrees trace inheritance through generations and help identify genetic disorders. Symbols include:

- Square = male, Circle = female
- Filled = affected, Empty = unaffected
- Horizontal line = mating, vertical line = offspring

Types of Inheritance:

- **Autosomal Dominant:** Affects every generation, equal sex ratio.
- **Autosomal Recessive:** Often skips generations, carriers are asymptomatic.
- **X-linked Recessive:** More males affected, no male-to-male transmission.

1.2 Molecular Genetics

The Central Dogma

The flow of genetic information:

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

Processes

- **Transcription:** Synthesis of mRNA from DNA, occurs in the nucleus.
- **Splicing:** Removal of introns from pre-mRNA.
- **Translation:** Ribosomes read mRNA to build polypeptides using tRNAs.

RNA Types

- mRNA – messenger RNA (template for protein)
- tRNA – transfer RNA (brings amino acids)
- rRNA – ribosomal RNA (structural component of ribosomes)

1.3 Population Genetics

Hardy-Weinberg Equilibrium

Defines conditions under which allele and genotype frequencies remain constant. Assumptions include:

- No mutation
- No migration
- No natural selection
- Large population
- Random mating

$$p^2 + 2pq + q^2 = 1 \tag{1.1}$$

Where p = freq. of dominant allele, q = freq. of recessive allele.

Example: If $p = 0.7$, then $q = 0.3$

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

Linkage Disequilibrium (LD)

LD describes the non-random association of alleles at different loci. Measured using:

- D' (standardized disequilibrium)
- r^2 (correlation coefficient between loci)

LD is critical in GWAS and for inferring haplotypes.

1.4 Quantitative Genetics

Polygenic Traits

Traits like height and intelligence are controlled by many genes and influenced by environment.

Normal Distribution: Polygenic traits often show a bell-shaped distribution in the population.

Heritability

$$h^2 = \frac{V_G}{V_P} \tag{1.2}$$

Where V_G is genetic variance and V_P is total phenotypic variance.

Interpretation:

- $h^2 \approx 1$: trait mostly genetic
- $h^2 \approx 0$: trait mostly environmental

1.5 Epigenetics and Gene Regulation

Epigenetic Modifications

Heritable changes in gene expression without altering the DNA sequence.

- DNA Methylation (usually represses transcription)
- Histone Modification (affects chromatin structure)
- Non-coding RNAs (e.g., miRNA, siRNA)

Regulatory Elements

- **Promoters:** Bind RNA polymerase to initiate transcription.
- **Enhancers:** Increase transcription levels; can be distant.
- **Silencers:** Repress transcription.

Imprinting and X-inactivation:

- Genomic imprinting: expression depends on parent of origin.
- X-inactivation: one X chromosome is silenced in females.

See also: Chapter 4 and Chapter 3 for epigenetics in gene expression and genome annotation.

2. DNA Sequencing Technologies

2.1 Introduction

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It is fundamental for genomics, enabling variant detection, genome assembly, transcriptomics, metagenomics, and personalized medicine. Over time, sequencing technologies have evolved through three major generations, each introducing new capabilities in terms of throughput, accuracy, and read length.

2.2 First-Generation Sequencing: Sanger Method

Principle

The Sanger method, or chain-termination sequencing, was developed by Frederick Sanger in 1977. It relies on selective incorporation of chain-terminating dideoxynucleotides (ddNTPs) by DNA polymerase during in vitro replication.

Protocol Overview

- A DNA template is combined with primers, DNA polymerase, deoxynucleotides (dNTPs), and fluorescently labeled ddNTPs.
- Incorporation of a ddNTP halts DNA synthesis.
- Resulting fragments are separated by capillary electrophoresis.
- A laser excites the fluorescent labels, and a detector records the sequence.

Applications and Limitations

- Read lengths: 500–1000 bp.
- High accuracy (>99.99%).
- Still used for validating mutations and sequencing single genes.
- Limited throughput and high cost make it unsuitable for large-scale projects.

2.3 Second-Generation Sequencing: NGS

2.3.1 Illumina Technology

Sequencing by Synthesis (SBS) Illumina platforms use a massively parallel approach where DNA fragments are immobilized on a flow cell and amplified by bridge PCR.

- Each cycle introduces four reversible terminator nucleotides with fluorescent labels.
- After incorporation, fluorescence is detected by high-resolution imaging.
- Terminator and fluorophore are cleaved before the next cycle.

Library Preparation

- DNA fragmentation (enzymatic or mechanical).
- Adapter ligation with index sequences (barcodes).
- Size selection and enrichment via PCR.

Output and Accuracy

- Read lengths: typically 100–300 bp (paired-end).
- Error rate: $<0.1\%$ (mainly substitution errors).
- High throughput: billions of reads per run.

Bias and Challenges

- GC bias due to PCR amplification.
- Short reads limit assembly of repetitive regions.
- Requires intensive computing for demultiplexing and basecalling.

2.3.2 Ion Torrent (Semiconductor Sequencing)

- Detects pH changes due to H^+ release when nucleotides are incorporated.
- Uses ion-sensitive field-effect transistor (ISFET) technology.
- Faster and cheaper setup compared to Illumina.
- Prone to indel errors in homopolymer regions.

2.4 Third-Generation Sequencing (TGS)

2.4.1 Pacific Biosciences (SMRT Sequencing)

Zero-Mode Waveguides (ZMWs)

- Single polymerase is immobilized in ZMW nanostructures.
- Fluorescently labeled nucleotides are detected in real time.
- Enables long reads (>10 kb; HiFi reads with circular consensus >99% accuracy).

Strengths

- Resolves structural variants, large insertions/deletions.
- Captures full-length isoforms in transcriptomics.
- Detects base modifications (methylation) without bisulfite treatment.

2.4.2 Oxford Nanopore Technologies (ONT)

Nanopore Principle

- DNA/RNA strands pass through biological nanopores embedded in a membrane.
- Changes in electrical current are measured and converted into base calls.
- No optical detection, minimal sample prep.

Features

- Real-time sequencing and analysis.
- Read lengths: 10 kb to >2 Mb (ultra-long reads).
- Error rate: 5–10% (improving with new basecallers and R10.4 pores).
- Direct RNA and DNA sequencing possible.

Devices

- MinION: portable USB-powered.
- GridION: scalable bench-top.
- PromethION: high-throughput.

2.5 Comparative Summary

Technology	Read Length	Accuracy	Speed	Best for
Sanger	500–1000 bp	>99.99%	Slow	Single genes
Illumina	50–300 bp	>99.9%	High	RNA-seq, WGS
Ion Torrent	400 bp	98–99%	High	Panels, amplicons
PacBio HiFi	10–25 kb	>99%	Moderate	Assembly, isoforms
Nanopore	>2 Mb	90–95%	Real-time	Ultra-long reads, fieldwork

2.6 Bioinformatic Outputs and Pre-processing

Output Formats

- **FASTQ:** Contains raw reads and quality scores.
- **BAM/CRAM:** Aligned reads to a reference.
- **VCF:** Variant call format for mutations/SNPs.

Pre-processing Steps

- Basecalling (especially for ONT)
- Demultiplexing (if multiplexed)
- Quality filtering and trimming (e.g., with `Fastp`, `Trimmomatic`)

2.7 Applications (without Assembly)

Key Use Cases

- Whole-genome/exome sequencing (WGS/WES)
- RNA-seq for transcriptome profiling
- Epigenomic studies via direct detection of modifications
- Metagenomics for microbial community analysis

See also: Assembly and annotation in Chapter 3.

3. Genome Assembly and Annotation

3.1 Introduction

Sequencing technologies produce millions of short or long DNA fragments (reads). However, these reads do not represent the genome in a linear format. Genome assembly is the process of reconstructing the original genome sequence from these reads, while annotation involves identifying and characterizing functional elements such as genes, regulatory regions, and repeats.

3.2 Assembly Strategies

De Novo vs Reference-Based Assembly

- **De novo:** Assembles the genome from scratch without using a reference. Useful for new species or highly divergent genomes.
- **Reference-based:** Maps reads to an existing reference genome to reconstruct the target genome. Faster, but can miss novel sequences.

Assembly Algorithms

Overlap-Layout-Consensus (OLC)

- Used in long-read assemblers (e.g., Canu, Flye).
- Compares all reads to find overlaps, builds a layout graph, and derives consensus sequence.
- Computationally intensive but effective for long noisy reads.

De Bruijn Graph (DBG)

- Used in short-read assemblers (e.g., SPAdes, Velvet).
- Breaks reads into k-mers and builds a graph where nodes represent k-1 mers.
- Fast and memory-efficient, but sensitive to sequencing errors.

String Graphs

- Advanced approach combining benefits of OLC and DBG.
- Used in assemblers like FALCON and wtdbg2.

3.3 Assembly Quality Metrics

Contiguity Metrics

- **N50:** Length of the shortest contig among the largest set that covers 50% of the total assembly.
- **L50:** Number of contigs that make up 50% of the genome.
- **NG50:** N50 relative to estimated genome size.

Completeness and Accuracy

- **BUSCO:** Assesses completeness using conserved single-copy orthologs.
- **LAI (LTR Assembly Index):** Measures assembly quality for repetitive regions.
- **Quast:** Compares assembly metrics and alignment to reference.

Common Errors

- Chimeric contigs (joining unrelated sequences).
- Collapse of repeats.
- Contig fragmentation.

3.4 Modern Assembly Techniques

Hybrid Assembly Combines long reads (for contiguity) and short reads (for accuracy). Tools include MaSuRCA, hybridSPAdes.

Polishing Refines assembly to correct errors:

- `Pilon`, `Racon`, `Medaka`

Scaffolding and Gap Closing Uses long-range information (e.g., Hi-C, optical mapping) to link contigs into scaffolds and close gaps.

3.5 Genome Annotation

Structural Annotation

Identification of genes and transcripts in the genome.

Approaches:

- **Ab initio:** Predicts genes based on intrinsic sequence features (e.g., AUGUSTUS, GeneMark).
- **Evidence-based:** Uses RNA-seq data or known proteins to support gene models (e.g., BRAKER, MAKER).

Functional Annotation

Assigns biological meaning to predicted genes.

Databases and Tools:

- Gene Ontology (GO), KEGG pathways, Pfam domains.
- `InterProScan`, `EggNOG-mapper`, `Blast2GO`, `CAZy` for carbohydrate-active enzymes.

Non-Coding Elements and Repeats

- `RepeatMasker`, `LTRharvest`, `Tandem Repeats Finder`.
- Identification of tRNAs (`tRNAscan-SE`), rRNAs (`Barrnap`).

3.6 Annotation Outputs and Formats

- **GFF3/GTF:** Gene feature formats describing coordinates and properties.
- **FASTA:** Nucleotide/protein sequences.
- **Protein/transcriptome files:** Derived from predicted coding sequences.
- **Functional tables:** Gene IDs with GO/KEGG/COG assignments.

3.7 Conclusion

Genome assembly and annotation are foundational for any downstream genomic analysis. A high-quality genome enables functional genomics, comparative studies, and biotechnological applications. Effective assembly strategies and accurate annotation pipelines are essential to transform raw sequencing data into meaningful biological knowledge.

4. Transcriptomics

4.1 Introduction

Transcriptomics is the study of the complete set of RNA transcripts produced by the genome under specific circumstances. Unlike the static genome, the transcriptome reflects dynamic gene expression patterns and regulatory mechanisms. The most common approach is RNA sequencing (RNA-seq), which enables the quantitative and qualitative analysis of RNA molecules.

4.2 RNA-Seq Overview

Experimental Workflow

- **RNA Extraction:** Total RNA is extracted from tissues or cells.
- **rRNA Depletion / Poly-A Enrichment:** rRNA is removed to enrich mRNA or non-coding RNAs.
- **Library Preparation:** RNA is fragmented, reverse transcribed to cDNA, and ligated to sequencing adapters.
- **Sequencing:** Libraries are sequenced using platforms such as Illumina, ONT, or PacBio.

Read Types

- **Single-end:** One end of each fragment is sequenced.
- **Paired-end:** Both ends of each fragment are sequenced; better for detecting isoforms and splicing.

4.3 Pre-processing and Mapping

Quality Control

- Performed using tools like `FastQC`, `MultiQC`.
- Includes filtering, trimming adapters, and low-quality bases.

Read Alignment

- Aligners: `STAR`, `HISAT2`, `TopHat2`.
- Splice-aware mappers align reads to a reference genome, considering exon-exon junctions.
- Outputs in BAM format.

Transcript Assembly (optional)

- Tools: `Cufflinks`, `StringTie`.
- Reconstruct transcript isoforms and quantify them.

4.4 Quantification and Normalization

Gene/Transcript Quantification

- Count reads overlapping genes or transcripts using `featureCounts`, `HTSeq`, or pseudoaligners like `Salmon`, `Kallisto`.
- Units: raw counts, TPM (Transcripts Per Million), FPKM (Fragments Per Kilobase Million).

Normalization

- Necessary to account for sequencing depth and gene length.
- Methods: TPM, DESeq2 normalization (median of ratios), TMM (edgeR).

4.5 Differential Expression Analysis

Purpose: Identify genes whose expression differs significantly between conditions (e.g., control vs treated).

Statistical Models:

- Tools: `DESeq2`, `edgeR`, `limma-voom`.
- Typically model raw counts using negative binomial distribution.
- Adjust p-values for multiple testing (FDR, Benjamini-Hochberg).

Outputs:

- Log2 fold change (LFC), adjusted p-value (padj), base mean expression.
- Volcano plots, heatmaps of differentially expressed genes (DEGs).

4.6 Transcript Isoforms and Splicing

Alternative Splicing

- Mechanism allowing one gene to produce multiple transcript variants.
- Types: exon skipping, intron retention, alternative donor/acceptor.
- Detected by tools like `rMATS`, `SUPPA2`.

Isoform Quantification

- Tools: `Kallisto`, `Salmon`, `FLAIR` (for long reads).
- Requires high-quality transcript models or reference annotation.

4.7 Functional Analysis of DEGs

Enrichment Analysis

- Identify overrepresented pathways or gene categories among DEGs.
- Databases: Gene Ontology (GO), KEGG, Reactome.
- Tools: `clusterProfiler`, `GSEA`, `DAVID`.

Visualization

- GO term enrichment barplots, KEGG pathway diagrams.
- Gene set enrichment plots, network graphs.

4.8 Special Considerations

- **Single-cell RNA-seq (scRNA-seq):** Resolves gene expression at individual cell level.
- **Long-read RNA-seq (Iso-seq, Nanopore):** Enables full-length transcript sequencing and discovery of novel isoforms.
- **Non-coding RNA profiling:** Includes miRNA-seq, lncRNA-seq.

4.9 Conclusion

Transcriptomics enables a deep understanding of gene expression and regulation in different biological contexts. RNA-seq remains the gold standard, with powerful bioinformatics tools available for processing, quantifying, and interpreting transcriptome data. Integration with genomics and proteomics leads to a comprehensive multi-omic view of biological systems.

5. Comparative and Functional Genomics

5.1 Introduction

Comparative genomics investigates the similarities and differences between the genomes of different organisms, revealing evolutionary relationships and functional conservation. Functional genomics, on the other hand, aims to understand gene function and interaction at a genome-wide scale. Combined, these approaches help elucidate the biological roles of genes, gene families, regulatory elements, and genome structure.

5.2 Orthology and Gene Relationships

Definitions:

- **Orthologs:** Genes in different species that evolved from a common ancestral gene by speciation; usually retain similar function.
- **Paralogs:** Genes related by duplication within a genome; may evolve new functions.
- **Homologs:** General term for genes with shared ancestry (includes orthologs and paralogs).

Tools for Orthology Inference:

- `OrthoFinder`, `Orthofinder2`, `EggNOG`, `OMA`.
- Input: proteomes of multiple species; Output: orthogroups and gene trees.

5.3 Phylogenomics and Evolution

Gene and Species Trees

- Gene trees represent the evolutionary history of gene families.
- Species trees reflect evolutionary relationships based on concatenated orthologs or core genes.

Phylogenetic Analysis

- Multiple sequence alignment: `MAFFT`, `Clustal Omega`.
- Tree building: `IQ-TREE`, `RAxML`, `FastTree`.
- Visualization: `iTOL`, `FigTree`.

5.4 Synteny and Genome Rearrangements

Synteny: Conserved order of genes or genomic segments between species. Helps in detecting genome duplications, rearrangements, and evolutionary conservation.

Tools:

- `MCScanX`, `SyMAP`, `D-GENIES`, `minimap2` for whole-genome alignment.

Applications:

- Identifying ancient whole genome duplications (WGDs).
- Studying chromosome evolution and genome plasticity.

5.5 Pan-genome and Core-genome Concepts

Pan-genome: The entire set of genes found in all strains of a species.

- **Core genome:** Shared by all strains.
- **Accessory genome:** Present in some but not all strains (includes strain-specific genes).

Tools:

- `Roary`, `Panaroo`, `PPanGGOLiN`, `BPGA`.

Applications:

- Understanding pathogenicity, adaptation, and genomic diversity.
- Used in microbial genomics, agriculture, and epidemiology.

5.6 Functional Annotation and Enrichment

Gene Ontology (GO)

- Three categories: Biological Process (BP), Molecular Function (MF), Cellular Component (CC).
- Functional annotation databases: `GOA`, `InterPro`, `EggNOG`.

Pathways: KEGG and Reactome

- KEGG: Genes mapped to metabolic and signaling pathways.
- Reactome: Curated molecular pathways across multiple organisms.

Enrichment Analysis

- Identifies GO terms or pathways overrepresented in a gene set (e.g., DEGs).
- Tools: `clusterProfiler`, `GSEA`, `DAVID`, `g:Profiler`.
- Input: list of gene IDs; Output: adjusted p-values and functional categories.

Other Databases

- **COG:** Clusters of Orthologous Groups (bacteria and archaea).
- **Pfam:** Protein families based on HMMs.
- **CAZy:** Carbohydrate-active enzymes (glycoside hydrolases, transferases, etc.).
- **TCDB:** Transporter Classification Database.

5.7 Visualization and Interpretation

- GO/KEGG plots: barplots, bubble plots, dotplots.
- Synteny maps and circos plots.
- Heatmaps of gene presence/absence or functional categories.

5.8 Conclusion

Comparative and functional genomics provide key insights into genome evolution, gene function, and adaptation. Through the integration of orthology, synteny, pathway analysis, and pan-genomics, researchers can interpret complex datasets and identify meaningful biological patterns across species or strains.

6. Applied Genomics and Biotechnologies

6.1 Introduction

Applied genomics harnesses the power of genome data to solve real-world problems in medicine, agriculture, biotechnology, and environmental science. It builds upon foundational genomic technologies to develop practical solutions, improve biological systems, and engineer new capabilities in living organisms.

6.2 Environmental Genomics

Metagenomics:

- Sequencing DNA from entire microbial communities without culturing.
- Types: Shotgun metagenomics, 16S/18S rRNA amplicon sequencing.
- Tools: `QIIME2`, `MetaPhlAn`, `MEGAHIT`, `Kraken2`.

Applications:

- Microbiome profiling in soil, water, gut, compost.
- Environmental monitoring and bioremediation.
- Detection of pathogens and antimicrobial resistance genes.

6.3 Medical Genomics

Precision Medicine:

- Uses whole genome (WGS) or exome sequencing (WES) to tailor therapies.
- Identification of driver mutations in cancer and inherited diseases.

Pharmacogenomics:

- Studies how genetic variation affects drug response.
- Examples: CYP450 genes (drug metabolism), TPMT (thiopurines).
- Tools: `PharmGKB`, `ClinVar`.

Rare Disease Genomics:

- Trio-based WES/WGS for diagnostic yield in genetic disorders.
- Interpretation pipelines: `GATK`, `VEP`, `Exomiser`.

6.4 Agricultural and Industrial Genomics

Crop Genomics:

- Identifies genes for yield, drought resistance, disease tolerance.
- Marker-assisted selection (MAS) and genomic selection (GS).

Animal Genomics:

- Improves livestock productivity and health.
- Examples: QTL mapping, GWAS in cattle, poultry.

Industrial Biotechnology:

- Uses genomics to optimize microbes for enzyme production, biofuels.
- Synthetic pathways for biodegradable plastic, biosurfactants.

6.5 Genome Editing and Synthetic Biology

Genome Editing:

- Techniques: CRISPR-Cas9, TALENs, ZFNs.
- Applications: gene knock-out, knock-in, base editing.
- Delivery systems: plasmids, RNPs, viral vectors.

Synthetic Biology:

- Engineering new genetic circuits and pathways.
- Modular parts: promoters, ribosome binding sites, terminators.
- Tools: `Benchling`, `SBOL`, `SynBioHub`.

Genome Synthesis:

- Total synthesis of microbial genomes (e.g., *Mycoplasma mycoides JCVI-syn3.0*).
- Applications in chassis development and minimal cells.

6.6 Ethical, Legal, and Social Implications (ELSI)

Ethical Concerns:

- Germline editing and designer babies.
- Equity in access to genomic medicine.

Data Privacy and Consent:

- GDPR and international frameworks.
- Dynamic informed consent models.

Regulatory Landscape:

- GMO regulations, biosafety standards.
- Approval processes for gene-edited organisms and therapies.

6.7 Conclusion

Applied genomics is transforming biology from a descriptive to an engineering discipline. Whether improving crops, diagnosing diseases, cleaning environments, or designing synthetic organisms, it represents the future of biotechnology. Responsible integration of genomic innovations will be key to maximizing societal benefits while addressing ethical and regulatory challenges.

7. Applied Genomics and Biotechnologies

7.1 Introduction

Applied genomics harnesses the power of genome data to solve real-world problems in medicine, agriculture, biotechnology, and environmental science. It builds upon foundational genomic technologies to develop practical solutions, improve biological systems, and engineer new capabilities in living organisms.

7.2 Environmental Genomics

Metagenomics:

- Sequencing DNA from entire microbial communities without culturing.
- Types: Shotgun metagenomics, 16S/18S rRNA amplicon sequencing.
- Tools: `QIIME2`, `MetaPhlAn`, `MEGAHIT`, `Kraken2`.

Applications:

- Microbiome profiling in soil, water, gut, compost.
- Environmental monitoring and bioremediation.
- Detection of pathogens and antimicrobial resistance genes.

7.3 Medical Genomics

Precision Medicine:

- Uses whole genome (WGS) or exome sequencing (WES) to tailor therapies.
- Identification of driver mutations in cancer and inherited diseases.

Pharmacogenomics:

- Studies how genetic variation affects drug response.
- Examples: CYP450 genes (drug metabolism), TPMT (thiopurines).
- Tools: `PharmGKB`, `ClinVar`.

Rare Disease Genomics:

- Trio-based WES/WGS for diagnostic yield in genetic disorders.
- Interpretation pipelines: `GATK`, `VEP`, `Exomiser`.

7.4 Agricultural and Industrial Genomics

Crop Genomics:

- Identifies genes for yield, drought resistance, disease tolerance.
- Marker-assisted selection (MAS) and genomic selection (GS).

Animal Genomics:

- Improves livestock productivity and health.
- Examples: QTL mapping, GWAS in cattle, poultry.

Industrial Biotechnology:

- Uses genomics to optimize microbes for enzyme production, biofuels.
- Synthetic pathways for biodegradable plastic, biosurfactants.

7.5 Genome Editing and Synthetic Biology

Genome Editing:

- Techniques: CRISPR-Cas9, TALENs, ZFNs.
- Applications: gene knock-out, knock-in, base editing.
- Delivery systems: plasmids, RNPs, viral vectors.

Synthetic Biology:

- Engineering new genetic circuits and pathways.
- Modular parts: promoters, ribosome binding sites, terminators.
- Tools: `Benchling`, `SBOL`, `SynBioHub`.

Genome Synthesis:

- Total synthesis of microbial genomes (e.g., *Mycoplasma mycoides JCVI-syn3.0*).
- Applications in chassis development and minimal cells.

7.6 Ethical, Legal, and Social Implications (ELSI)

Ethical Concerns:

- Germline editing and designer babies.
- Equity in access to genomic medicine.

Data Privacy and Consent:

- GDPR and international frameworks.
- Dynamic informed consent models.

Regulatory Landscape:

- GMO regulations, biosafety standards.
- Approval processes for gene-edited organisms and therapies.

7.7 Conclusion

Applied genomics is transforming biology from a descriptive to an engineering discipline. Whether improving crops, diagnosing diseases, cleaning environments, or designing synthetic organisms, it represents the future of biotechnology. Responsible integration of genomic innovations will be key to maximizing societal benefits while addressing ethical and regulatory challenges.

8. Full-Spectrum Revision: Integrated, Technical and Exam-Oriented Summary

8.1 Core Principle: Genome-Centered Biology

Everything in genomics begins and ends with the genome. The central logic is:

DNA → Information → Function → Interpretation → Engineering

Across organisms and omics layers (genomics, transcriptomics, functional, comparative), we decode, quantify, and modify genomic content.

8.2 Sequencing (Ch. 2)

Technologies

- **Illumina**: 150–300 bp, high accuracy (>99.9%), paired-end, $$.
- **ONT (Nanopore)**: 10–500 kb, lower accuracy (90–98%), portable, real-time.
- **PacBio HiFi**: 10–25 kb, high accuracy (>99.8%), ideal for de novo.

Quality Metrics

- Q-score: Q30 = 99.9% accuracy.
- Coverage: >30× for variant calling; >100× for de novo assembly.
- Duplication rate, GC bias, N content.

8.3 Genome Assembly and Annotation (Ch. 3)

Assembly

- **Approaches**: OLC (long-read), DBG (short-read), Hybrid.

- **Key Tools**: SPAdes, Flye, Canu, Unicycler.
- **Metrics to Remember**:
  - N50 (ideal > 1 Mb for fungi), L50, Total Length.
  - BUSCO completeness: >**95%** = high-quality.

Annotation

- **Structural**: Gene prediction via Augustus, BRAKER2.
- **Functional**:
  - BLASTp, InterProScan, EggNOG, Pfam, CAZy.
  - Outputs: GFF3, TSV, FASTA, GO, EC, KO.

File Formats

- FASTQ (raw), FASTA (sequences), GFF3 (features), BAM (alignments), VCF (variants).

8.4 Transcriptomics (Ch. 4)

Workflow and Tools

1. **QC**: FastQC, MultiQC.
2. **Trimming**: Trimmomatic, fastp.
3. **Mapping**: HISAT2, STAR.
4. **Quantification**: featureCounts, Salmon.
5. **DEGs**: DESeq2, edgeR.
6. **Visualization**: PCA, MA plot, volcano, heatmap.

Key Metrics

- TPM/FPKM: normalized expression.
- Padj < 0.05 for significance.
- log2FoldChange > 1 = upregulated; < -1 = downregulated.

Output

- Count matrix (genes × samples), DEG list, plots.

8.5 Functional & Comparative Genomics (Ch. 5)

Enrichment Analysis

- Tools: clusterProfiler (R), DAVID, GSEA.
- Inputs: Gene lists, background set.
- Outputs: Enriched GO terms (BP, MF, CC), KEGG pathways.
- Cutoffs: p.adjust < 0.05.

Orthology & Phylogeny

- Tools: OrthoFinder, MAFFT, FastTree, IQ-TREE.
- Concepts: Orthologs = conserved, Paralogs = duplicated.
- Phylogenetic trees from core genes.

Pan-genomics & Synteny

- Tools: Roary (bacteria), Panaroo, MCScanX.
- Core genome = common genes; accessory = niche adaptation.

8.6 Applied Genomics (Ch. 6)

Medical Genomics

- **WGS/WES**: Rare diseases, cancer mutations.
- **Biomarkers**: Expression (RNA), Methylation (bisulfite), Mutations (VCF).
- **Tools**: GATK, ClinVar, VarSome.

Agricultural and Industrial Genomics

- **GWAS**: SNP-trait associations (PLINK).
- **Microbial engineering**: for enzymes, biofuels, bioplastics.

Genome Editing and Synthetic Biology

- CRISPR-Cas9, base editors, prime editing.
- Design: Benchling, SnapGene, Geneious.
- Chassis: *E. coli*, *S. cerevisiae*, minimal genome (JCVI-syn3.0).

8.7 Ethical and Legal Aspects

- GDPR: genomic data = sensitive.
- Germline editing: banned in humans (many countries).
- Consent: dynamic and re-consent models.
- Synthetic biology: dual-use, biosecurity.

8.8 Master Workflow Overview

From Raw Data to Functional Insight and Application:

Sequencing → QC → Assembly → Annotation → Expression → DEG → Enrichment → Comparative → Editing/Application

8.9 Essential Concepts to Memorize

- **Coverage**: >30× (variant calling), >100× (assembly).
- **N50**: median contig length; high = better.
- **BUSCO**: >95% = complete assembly.
- **DEG threshold**: padj < 0.05 and $|\log2FC| > 1$.
- **CRISPR**: Cas9 + guide RNA = cut + HDR/NHEJ.

Ultimate Review Checklist

1. Workflow: Can I explain step-by-step from DNA to application?
2. Files: FASTQ, BAM, GFF3, VCF – what are they and when?
3. Tools: Which software for each task?
4. Concepts: Assembly vs Annotation? Ortholog vs Paralog?
5. Numbers: N50? Coverage? TPM? Fold change?
6. Pitfalls: Contamination? Pseudogenes? Annotation errors?
7. Applications: How is this used in real genomics (human, fungal, microbial)?

Project Integration: Theory Meets Practice

The applied genomics project on *Purpureocillium lilacinum* PLA-C1 served as a comprehensive implementation of the concepts, tools, and workflows covered across all theoretical chapters. From raw data to biological interpretation, each step mirrored foundational principles in sequencing, assembly, transcriptomics, and comparative genomics.

Sequencing Technologies. The project employed hybrid sequencing combining ONT long reads (MinION) and Illumina short reads (paired-end 150 bp). This choice reflected the strengths of both platforms: Nanopore enabled scaffolding and resolution of repeats (ultra-long reads >20 kb), while Illumina ensured base-level accuracy (Q30 > 90%). This strategy exemplifies the hybrid sequencing paradigms discussed in Chapter 2.

Genome Assembly. Assembly was performed via the Flye assembler (long-read), polished with `Medaka` and short-read tools like `Pilon`. The result: 21 contigs, N50 = 3.2 Mb, total genome size ∼40 Mb, and BUSCO completeness = **98.1%**, aligning with the quality expectations outlined in Chapter 3. Contig L50 = 5. Assembly metrics were benchmarked using `QUAST` and `BUSCO` with the `sordariomycetes_odb10` lineage dataset.

Genome Annotation. Structural annotation integrated ab initio (`Augustus`) and evidence-based (`BRAKER2`) strategies using ONT-derived transcriptomic alignments. Functional annotation incorporated `EggNOG-mapper`, `InterProScan`, and `CAZy`, yielding 11,582 predicted protein-coding genes, with GO and KEGG terms assigned to over 70%. Notably, 356 CAZymes were identified, including GH, CE, and AA families involved in PLA degradation (see Chapter 3 and 5).

Transcriptomics and DEGs. RNA-seq data (Illumina PE150) from PLA and control conditions were quality-checked with `FastQC`, trimmed with `Trimmomatic`, and mapped with `HISAT2`. Expression was quantified via `featureCounts`, normalized (TPM and DESeq2), and differential expression analysis yielded 1,213 DEGs ($|\log_2 FC| > 1$, padj < 0.05). Volcano plots and heatmaps visualized key DEGs upregulated in PLA exposure, supporting Chapter 4 principles.

Functional and Comparative Genomics. DEGs were functionally enriched using `clusterProfiler` and KEGG Mapper, revealing overrepresentation in xenobiotic degradation, peroxisome activity, and oxidative stress. Comparative genomics with related fungi via `OrthoFinder` and synteny with `MCScanX` confirmed expansion of detoxification-related gene families. Phylogenetic analysis (IQ-TREE, GTR+G model) placed PLA-C1 within the Ophiocordycipitaceae with strong support (bootstrap $> 95\%$). This directly illustrates the use of orthology, phylogenomics, and gene family analysis discussed in Chapter 5.

From Theory to Impact. The project epitomizes applied genomics: integrating multi-omic data, leveraging advanced tools, and translating theoretical principles into actionable biological insight. The identification of enzymes and transcriptomic signatures associated with PLA degradation has potential for future biotechnological exploitation. This capstone experience demonstrates not only technical mastery but also conceptual understanding, fully aligned with the pedagogical goals outlined throughout this handbook.

Experimental Design and Sample Metadata. The fungal isolate *P. lilacinum* PLA-C1 was cultivated under PLA-enriched and control conditions to investigate transcriptional responses to synthetic polymer exposure. Three replicates per condition were sequenced using Illumina PE150, generating ∼45 million reads/sample post-trimming.

Assembly Optimization and Curation. Long reads from ONT were assembled using `Flye` with default minimum overlap length (1,000 bp) and polished using both `Medaka` (ONT consensus) and `Pilon` (short-read correction). After manual removal of mitochondrial and contaminant contigs (via BLASTn vs NCBI nt), the final assembly consisted of 21 contigs with no ambiguous bases (Ns) and GC content of 52.6%.

Gene Family Annotation and CAZyme Profiling. `dbCAN2` and `HMMER3` were used to annotate carbohydrate-active enzymes. A total of 356 putative CAZymes were found:

- 132 Glycoside Hydrolases (GHs) – notably GH3, GH5, GH16
- 71 Auxiliary Activities (AAs) – including AA9 (LPMOs), AA3
- 49 Carbohydrate Esterases (CEs) – including CE1 and CE5

These families are commonly associated with plant biomass degradation and, by analogy, plastic depolymerization. Signal peptide analysis via `SignalP` indicated extracellular secretion potential for 42% of the CAZymes.

Transcriptomic Insights into PLA Response. RNA-seq analysis revealed 1,213 differentially expressed genes (DEGs):

- 684 upregulated, 529 downregulated in PLA vs. control
- Top upregulated genes included: catalase-peroxidase, cutinase-like lipase, and hydrolase domain proteins.

- Key enriched GO terms: "oxidoreductase activity", "lipid metabolic process", "response to oxidative stress"
- KEGG pathways enriched: "Peroxisome" (ko04146), "Fatty acid degradation" (ko00071)

Comparative Genomics and Phylogeny. Using `OrthoFinder`, 16,204 orthogroups were identified across 12 related fungal genomes. PLA-C1 had 267 unique orthogroups, including expansions in genes encoding catalase, superoxide dismutase, and cytochrome P450s. Core genome alignment (2,311 single-copy orthologs) was used to construct a maximum-likelihood tree with `IQ-TREE` under the GTR+G4 model, confirming evolutionary placement near *Purpureocillium lilacinum* strain 36-1 with 100% bootstrap support.

Custom Scripts and Workflow Management. All steps were orchestrated with Snakemake and reproducible environments using Conda. Custom Python scripts were used for:

- Filtering CAZymes based on domain count and EC number
- Extracting DEGs intersecting with annotated CAZymes
- Generating BED files from GFF3 for visualization in IGV

Output Summary.

- **Assembly**: 21 contigs, N50 = 3.2 Mb, 98.1% BUSCO
- **Gene prediction**: 11,582 genes
- **DEGs**: 1,213 (684 up, 529 down)
- **CAZymes**: 356 total (132 GH, 71 AA, 49 CE)
- **Enriched pathways**: 7 KEGG, 22 GO terms
- **Unique orthogroups**: 267

Biological Interpretation. These findings support the hypothesis that *P. lilacinum* PLA-C1 mounts a transcriptional response involving peroxisomal -oxidation, oxidative stress defenses, and extracellular enzyme production when exposed to synthetic bioplastics. The co-occurrence of CAZymes and oxidative enzymes suggests a concerted mechanism of extracellular degradation and intracellular detoxification, aligning with emerging models in fungal plastic biodegradation.

Theory-to-Project Mapping: Structured Keywords and Concepts

Objective: Provide a fast-access, exam-oriented reference that links each theoretical concept to project steps using keywords, tools, parameters, and outputs.

1. Sequencing

- **Keywords:** ONT, Illumina, Q-score, Coverage, Read length, Base calling, Adapter trimming
- **Tools:** MinION (ONT), NovaSeq (Illumina), Guppy, FastQC, NanoPlot, Porechop
- **Parameters:** Q30 $> 90\%$, Coverage $> 100\times$, Read length: ONT (10–100 kb), Illumina PE150
- **Project Links:** Raw data QC, dual-platform sequencing for resolution + accuracy
- **File Types:** FASTQ, summary stats (HTML)

2. Genome Assembly

- **Keywords:** Hybrid assembly, OLC, DBG, Polishing, Contig, Scaffold, N50, L50
- **Tools:** Flye, Minimap2, Medaka, Pilon, QUAST, BUSCO
- **Parameters:** N50 = 3.2 Mb, Total length = 40.1 Mb, GC content = 52.6%, BUSCO = 98.1%
- **Project Links:** Hybrid approach used to combine accuracy (Illumina) and contiguity (ONT)
- **File Types:** FASTA, AGP, BUSCO summary, QUAST report

3. Structural and Functional Annotation

- **Keywords:** ab initio, evidence-based, gene prediction, CDS, exon, domain, ontology, EC number
- **Tools:** BRAKER2, Augustus, InterProScan, EggNOG-mapper, dbCAN2, SignalP
- **Parameters:** 11,582 coding genes, 356 CAZymes, 267 unique orthogroups, 42% signal peptides
- **Project Links:** Transcript evidence from ONT used to guide BRAKER2; CAZymes and GO terms extracted
- **File Types:** GFF3, protein/nucleotide FASTA, annotation tables (TSV)

4. Transcriptomics and DEG Analysis

- **Keywords:** QC, trimming, mapping, quantification, normalization, statistical testing, log2FC, padj
- **Tools:** Trimmomatic, HISAT2, SAMtools, featureCounts, DESeq2, R scripts, MultiQC
- **Parameters:** 1,213 DEGs, log2FC > 1, padj < 0.05, 45M reads/sample, 85–90% mapping rate
- **Project Links:** PLA condition vs control, DEGs with known hydrolytic/oxidative function identified
- **Visuals/Outputs:** Volcano plot, heatmap, PCA, count matrix (CSV)

5. Functional Enrichment and CAZy Profiling

- **Keywords:** Gene Ontology (GO), KEGG, CAZy, enrichment, peroxisome, secretion, detoxification
- **Tools:** clusterProfiler, KEGG Mapper, dbCAN2, SignalP, PfamScan
- **Parameters:** Top GO terms: "oxidoreductase activity", "response to oxidative stress", KEGG: Peroxisome (ko04146)
- **Project Links:** Identified upregulated CAZymes (GH5, AA9, CE1), many with extracellular localization
- **Visuals/Outputs:** Bar plots, dotplots, CAZy tables (CSV), GO term tables

6. Comparative Genomics and Phylogeny

- **Keywords:** Ortholog, paralog, core genome, pan-genome, phylogeny, gene expansion, bootstrap
- **Tools:** OrthoFinder, IQ-TREE, MAFFT, MCScanX, iTOL
- **Parameters:** 16,204 orthogroups, 2,311 single-copy core genes, 100% bootstrap support
- **Project Links:** PLA-C1 had expanded detox gene families (catalase, P450); positioned near P. lilacinum 36-1
- **Visuals/Outputs:** Phylogenetic tree (newick + PNG), collinearity plots

7. Application: PLA Degradation Potential

- **Keywords:** Bioplastics, hydrolase, cutinase, lipase, oxidative metabolism, fungal secretion
- **Tools:** Custom Python scripts, BLASTp, SignalP, eggNOG, KEGG pathways
- **Findings:** Upregulation of catalase-peroxidase, cutinase-like enzymes, lipid metabolism; 42% predicted secreted
- **Project Links:** Enzymatic basis for PLA degradation elucidated; candidate genes proposed
- **Output:** Functional gene list, annotated CAZymes, integration with DEG/GO data

Extra Keywords to Remember

2

- FASTQ, GFF3, BAM, VCF, TSV
- N50, BUSCO, TPM, padj, log2FC
- Hybrid assembly, ab initio prediction
- CAZy: GH, AA, CE, CBM, PL
- GO terms: BP, MF, CC
- OrthoFinder, IQ-TREE, DESeq2
- KEGG IDs, Pfam domains, EC numbers
- SignalP, dbCAN2, clusterProfiler

Practice Questions and Answers for Exam Review

Multiple Choice (Sample Questions)

1. **What does linkage disequilibrium (LD) 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 genome assembly
   d) The mutation rate in microsatellites

2. **Which sequencing platform generates the longest reads?**
   a) Illumina
   b) Ion Torrent
   c) PacBio ()
   d) SOLiD

3. **What is the y-axis in a Manhattan plot?**
   a) Chromosome position
   b) Minor allele frequency
   c) $-\log_{10}$(p-value) ()
   d) Beta coefficient (effect size)

4. **Which file stores aligned reads to a reference genome?**
   a) FASTQ
   b) VCF
   c) GFF3
   d) SAM/BAM ()

5. **What does BUSCO assess?**
   a) Genome contamination
   b) Repeat content
   c) Completeness based on conserved orthologs ()
   d) SNP frequency

6. **What is a CIGAR string used for?**
   a) Describes methylation of CpGs
   b) Describes how a read aligns to the reference genome ()
   c) Annotates protein domains
   d) Stores sequencing quality scores

7. **What does the VCF format store?**
   a) Read alignments
   b) SNP and variant calls ()
   c) Gene annotations
   d) Expression levels

8. **How does SOLiD sequencing work?**
   a) Synthesis-based sequencing of fluorescent nucleotides
   b) Sequencing by ligation using 2-base encoding ()
   c) Nanopore-based single molecule detection
   d) Pyrosequencing with light signal

Short Answer (Sample Questions)

- **Depth of Coverage Calculation:**
  Coverage = (Number of Reads × Read Length) / Genome Size.

- **N50 Calculation:**
  Sort contigs by length, sum total, find length where 50% of assembly is reached.

- **Bisulfite Sequencing:**
  Converts unmethylated C to U → T in sequencing; preserves methylated C.

- **C-value Method:**
  Genome size (Gb) = C-value (pg) $\times$ 0.978.
- **ChIP-Seq:**
  Maps protein-DNA binding via chromatin immunoprecipitation + sequencing.
- **de Bruijn Graph:**
  k-mer graph for short read assembly: nodes = k-1-mers; edges = shared k-mers.
- **Hardy-Weinberg Deviations:**
  Indicate population structure, selection, inbreeding, or error.
- **GWAS Aim:**
  Identify associations between SNPs and phenotype using large cohort data.
- **Population Stratification:**
  Differences in ancestry that may bias association studies.
- **PCA:**
  Reduces dimensionality to visualize population structure or batch effects.

Mini Case Exercises (Project-Based)

Case 1: Contamination Check

- **Q:** You assembled a fungal genome. How do you identify and remove bacterial contamination?
- **A:** Use Kraken2 to classify scaffolds. Visualize with BlobTools. Filter scaffolds assigned to bacteria. Check read coverage and GC content.

Case 2: Identifying DEGs in RNA-Seq

- **Q:** You want to identify genes upregulated in PLA-grown cultures vs glucose.
- **A:** Use STAR for mapping, featureCounts for quantification, DESeq2 for differential analysis. Set cutoffs: padj < 0.05 and $\log_2$FC > 1.

Case 3: Assembly Evaluation

- **Q:** How do you assess the quality of your hybrid assembly?
- **A:** Use Quast (N50, total length, GC), BUSCO (conserved genes), and read mapping statistics.

Case 4: Functional Annotation of CAZymes

- **Q:** How do you identify carbohydrate-active enzymes in your genome?
- **A:** Use dbCAN pipeline: HMMER against CAZy, filtered by coverage/e-value. Annotate genes with associated functions (e.g., cutinase, esterase).

Case 5: GO Enrichment from DEGs

- **Q:** After DEG analysis, how do you extract biological insight?
- **A:** Use topGO with DEGs to test enrichment for GO terms (MF/BP/CC). Visualize enriched categories (FDR corrected).

Case 6: Detection of Horizontal Gene Transfer (HGT)

- **Q:** How do you investigate if some genes were horizontally transferred from bacteria?
- **A:** Use BLASTp against NCBI nr, build phylogenetic trees, check GC content/anomalies, validate taxonomic origin.

Case 7: Variant Calling in Hybrid Assembly

- **Q:** How do you identify SNPs/indels from reads aligned to your hybrid genome?
- **A:** Use BWA-MEM or Minimap2 for mapping, then FreeBayes or GATK HaplotypeCaller. Output in VCF format.

Case 8: Transcript Isoform Analysis

- **Q:** How can you distinguish transcript isoforms in your RNA-Seq data?
- **A:** Use tools like StringTie or IsoSeq3 (PacBio). Annotate alternative splicing events. Compare with GFF3.

Case 9: Detecting Gene Family Expansion

- **Q:** How do you determine if a gene family has expanded in your strain?
- **A:** Use OrthoFinder or CAFE. Compare orthogroup sizes across species. Analyze duplication events.

Case 10: Functional Prediction of Novel Genes

- **Q:** You find unannotated ORFs in your genome. How do you infer their function?
- **A:** Predict domains with InterProScan, use EggNOG for orthology, check GO terms and KEGG pathways.