

# VU LEARNING HUB

[vulearninghub.xyz](https://vulearninghub.xyz)

**HANDOUT**

## **BIF502**

Genome Informatics

**BIF502 by VU Learning HUB**

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

## BIF502 - Genome Informatics

A structured study handout compiled from the course's lecture topics. Read each topic's objectives first, study the key concepts, then use the summary and exam tips for revision.

# BIF502

## Genome Informatics

Based on 45 lecture topics

Compiled for VU students as a structured study guide

*Concepts are summarized in the writer's own words;  
use alongside your lecture videos for best results.*

# Contents

|                                                             |           |
|-------------------------------------------------------------|-----------|
| <b>Unit 1: Genomes and Data</b>                             | <b>3</b>  |
| Topic 1: What Is Genome Informatics . . . . .               | 3         |
| Topic 2: DNA, Genes, and Genome Organization . . . . .      | 4         |
| Topic 3: Sequencing Technologies . . . . .                  | 4         |
| Topic 4: Sequence Formats and Databases . . . . .           | 5         |
| Topic 5: Genome Projects and References . . . . .           | 6         |
| <b>Unit 2: Sequence Alignment</b>                           | <b>7</b>  |
| Topic 6: Why Alignment Matters . . . . .                    | 7         |
| Topic 7: Dynamic Programming Alignment . . . . .            | 7         |
| Topic 8: BLAST and Heuristic Search . . . . .               | 8         |
| Topic 9: Read Mapping . . . . .                             | 9         |
| Topic 10: Multiple Sequence Alignment . . . . .             | 10        |
| <b>Unit 3: Assembly and Annotation</b>                      | <b>10</b> |
| Topic 11: Genome Assembly . . . . .                         | 10        |
| Topic 12: Gene Prediction . . . . .                         | 11        |
| Topic 13: Functional Annotation . . . . .                   | 12        |
| Topic 14: Variant Calling . . . . .                         | 13        |
| Topic 15: RNA-seq and Expression Analysis . . . . .         | 13        |
| <b>Unit 4: Comparative Genomics and Applications</b>        | <b>14</b> |
| Topic 16: Comparative Genomics . . . . .                    | 14        |
| Topic 17: Phylogenetics . . . . .                           | 15        |
| Topic 18: Metagenomics and Microbiomes . . . . .            | 16        |
| Topic 19: Genomics in Medicine and Agriculture . . . . .    | 17        |
| Topic 20: Ethics, Privacy, and Future of Genomics . . . . . | 17        |
| <b>Glossary of Key Terms</b>                                | <b>18</b> |

# Course Introduction

Genome informatics is where biology became a data science. A single human genome holds 3 billion letters; modern sequencers produce terabytes daily; and making sense of this flood - assembling genomes, finding genes, comparing species, linking variants to disease - is fundamentally computational.

This course covers the algorithms and tools of genome analysis: sequence formats and databases, alignment algorithms (the engines of bioinformatics), genome assembly, gene prediction, variant calling, and comparative genomics. It also introduces the statistics of high-throughput data and the ethical dimensions of genomic information.

No prior biology beyond basics is assumed; the emphasis is on the informatics - the data structures, algorithms, and practical pipelines that turn raw reads into biological knowledge.

**How to use this handout:** Each topic lists its learning objectives, explains the key concepts in simple words, defines important terms, gives an example where helpful, and closes with a summary plus exam tips. Topics marked **Very important** deserve extra revision.

## Unit 1: Genomes and Data

### Topic 1: What Is Genome Informatics

#### Learning Objectives

- Define genome informatics
- Explain the data deluge
- List core analysis tasks

#### Key Concepts

Genome informatics applies computation to genome-scale biological data: DNA sequences, gene annotations, variation catalogs, and functional genomics measurements. It sits within bioinformatics, focused on whole genomes rather than single genes or proteins.

The data deluge: sequencing cost fell faster than Moore's law (\$3B for the first human genome; under \$1000 today), producing petabyte-scale archives (SRA, ENA). Analysis - not sequencing - is now the bottleneck.

Core tasks: read alignment, genome assembly, variant calling, annotation (finding genes), comparative genomics, and linking genotype to phenotype. Each is an algorithmic challenge with dedicated tools.

#### Key Points

- Genome informatics: computation on genome-scale data.
- Sequencing costs collapsed; analysis is the bottleneck.
- Core tasks: align, assemble, call variants, annotate, compare.

#### Key Terms

**Genome:** An organism's complete DNA sequence.

**Bioinformatics:** Computation applied to biological data.

**High-throughput sequencing:** Massively parallel DNA sequencing.

### Example

- *The 1000 Genomes Project cataloged human variation computationally.*

**Summary:** Genomes are big data; informatics makes them knowledge.

### Exam Tips

- ★ Key point: the cost-collapse consequence.
- ★ Remember core task list.

## Topic 2: DNA, Genes, and Genome Organization

### Learning Objectives

- Review DNA structure
- Explain genes and regulation
- Describe genome features

### Key Concepts

DNA: double helix of A/T/G/C bases; genes are transcribed to RNA, translated to proteins (central dogma). Eukaryotic genes have exons (coding) interrupted by introns (spliced out); regulatory regions (promoters, enhancers) control expression.

Genome organization: humans have 3 billion bases, ~20,000 protein-coding genes (only ~1.5% coding!), plus non-coding RNAs, regulatory elements, and vast repetitive sequences (transposons, ~50% of the genome). Bacterial genomes are compact; plant genomes can be enormous.

Informatics implication: finding the functional 1.5% in 3 billion bases is a signal-detection problem - the motivation for gene prediction and comparative methods.

### Key Points

- Central dogma: DNA → RNA → protein; exons/introns structure.
- Only ~1.5% of human genome codes for proteins.
- Repeats and regulation dominate genome landscapes.

### Key Terms

**Exon:** Coding segment retained in mature RNA.

**Intron:** Non-coding segment spliced out.

**Transposon:** A mobile repetitive DNA element.

### Example

- *Finding genes among repeats is like finding sentences in static.*

**Summary:** Genomes are mostly non-coding; informatics finds the signals.

### Exam Tips

- ★ Very important: the 1.5% figure.
- ★ Remember exon/intron.

## Topic 3: Sequencing Technologies

### Learning Objectives

- Compare sequencing generations

- Explain read types
- Discuss error profiles

## Key Concepts

Sanger sequencing (1st gen): accurate, long, but one fragment at a time - sequenced the first genomes slowly. NGS (2nd gen, Illumina): millions of short reads (50-300 bp), cheap, with substitution errors - the workhorse generating most data.

Third generation (PacBio, Oxford Nanopore): long reads (10kb-1Mb+), higher error rates (improving), enabling complete assemblies and direct RNA/methylation detection. Nanopore devices are pocket-sized - sequencing in the field.

Informatics consequences: short reads need assembly and struggle with repeats; long reads resolve repeats but need error correction. Tool choice follows technology.

## Key Points

- Sanger: accurate but slow; Illumina: cheap short reads; long-read: resolving repeats.
- Error profiles differ: substitutions vs. indels.
- Technology determines algorithmic strategy.

## Key Terms

**Read:** A sequenced DNA fragment.

**NGS:** Next-Generation (short-read) Sequencing.

**Long-read sequencing:** Technologies producing 10kb+ reads.

## Example

- *Nanopore sequenced Ebola genomes in field hospitals.*

**Summary:** Sequencing tech shapes every downstream algorithm.

## Exam Tips

- ★ Key point: compare the three generations.
- ★ Remember read-length consequences.

# Topic 4: Sequence Formats and Databases

## Learning Objectives

- Read FASTA/FASTQ
- Explain quality scores
- Navigate genome databases

## Key Concepts

FASTA: sequences with headers (>name, then letters) - the universal format. FASTQ: adds per-base quality scores (Phred:  $Q = -10 \log_{10}(\text{error prob})$ ; Q30 = 1/1000 error) - essential for filtering unreliable bases.

SAM/BAM: aligned reads with positions, mapping quality, CIGAR strings; VCF: variant calls with genotypes and annotations. Mastering these formats is practical literacy.

Databases: GenBank/ENA/DDBJ (raw sequences), RefSeq (curated references), Ensembl/UCSC Genome Browser (annotations and visualization), dbSNP/ClinVar (variants). Genome browsers turn

coordinates into explorable biology.

### Key Points

- FASTA: sequences; FASTQ: + quality scores (Phred).
- SAM/BAM: alignments; VCF: variants.
- GenBank, Ensembl, UCSC: the reference resources.

### Key Terms

**Phred score:** Log-scaled base-call quality.

**VCF:** Variant Call Format.

**Genome browser:** Visualization of annotations on coordinates.

### Example

- *O30 bases are trusted: O10 bases are filtered out.*

**Summary:** Formats and databases are the working language of genomics.

### Exam Tips

- ★ Very important: FASTA vs. FASTQ.
- ★ Remember Phred meaning.

## Topic 5: Genome Projects and References

### Learning Objectives

- Summarize landmark projects
- Explain reference genomes
- Discuss pangenomes

### Key Concepts

Landmarks: Human Genome Project (2003 draft), 1000 Genomes (human variation), ENCODE (functional elements), Earth BioGenome (all species), and countless pathogen projects (transforming epidemiology).

Reference genomes: a representative assembly all analysis maps against (e.g. GRCh38/hg38 for human). References carry biases (one ancestry) - motivating pangenomes: graph structures representing many genomes' variation, improving mapping for diverse populations.

Versions matter: coordinates differ between builds (liftOver converts); always record which reference was used - reproducibility depends on it.

### Key Points

- HGP, 1000 Genomes, ENCODE: foundational resources.
- Reference genomes anchor analysis but carry bias.
- Pangenome graphs represent population variation.

### Key Terms

**Reference genome:** The standard assembly for a species.

**Pangenome:** The full gene/variant set of a species or population.

**liftOver:** Converting coordinates between genome builds.

### Example

- *GRCh38 coordinates differ from hg19 - analyses must specify.*

**Summary:** References organize genomics; pangenomes generalize them.

### Exam Tips

- ★ Key point: reference bias.
- ★ Remember pangenome idea.

## Unit 2: Sequence Alignment

### Topic 6: Why Alignment Matters

#### Learning Objectives

- Define sequence alignment
- Explain its applications
- Distinguish alignment types

#### Key Concepts

Alignment arranges sequences to reveal similarity: matching bases imply shared ancestry (homology) or function. It underlies nearly everything: mapping reads, finding genes, comparing species, building trees.

Types: pairwise (two sequences) vs. multiple; global (end-to-end, Needleman-Wunsch) vs. local (best-matching substrings, Smith-Waterman); DNA vs. protein (proteins use substitution matrices like BLOSUM reflecting chemical similarity).

Scoring: matches rewarded, mismatches/gaps penalized; optimal alignment maximizes score. Statistics (E-values, BLAST) distinguish real homology from chance - essential when searching huge databases.

#### Key Points

- Alignment reveals homology and function.
- Global vs. local; pairwise vs. multiple.
- Scoring + statistics separate signal from chance.

#### Key Terms

**Homology:** Similarity from shared ancestry.

**Global alignment:** End-to-end sequence alignment.

**Local alignment:** Best-substring alignment.

### Example

- *BLAST-ing an unknown gene finds its likely function via homologs.*

**Summary:** Alignment is bioinformatics' fundamental operation.

### Exam Tips

- ★ Very important: global vs. local.
- ★ Remember homology concept.

### Topic 7: Dynamic Programming Alignment

## Learning Objectives

- Explain Needleman-Wunsch
- Explain Smith-Waterman
- Analyze complexity

## Key Concepts

Needleman-Wunsch (global): fill a matrix where cell (i,j) = best score aligning prefixes; recurrence takes max of match/mismatch (diagonal), gap in one (up), gap in other (left); traceback from bottom-right yields the alignment.  $O(nm)$  time and space.

Smith-Waterman (local): same but floor scores at 0 and traceback from the maximum anywhere - finds the best local match, ideal for database search and domain finding.

Complexity limits genome-scale use (3B x 3B impossible) - hence heuristics (BLAST) and indexed methods for large problems; DP remains the exact gold standard for moderate sizes.

## Key Points

- DP recurrence: max of diagonal/up/left; traceback recovers alignment.
- Needleman-Wunsch: global; Smith-Waterman: local (floor at 0).
- $O(nm)$  - exact but too slow for whole genomes.

## Key Terms

**Dynamic programming:** Solving via optimal subproblems.

**Traceback:** Recovering the alignment from the DP matrix.

**Gap penalty:** Cost of insertions/deletions.

## Example

- *Alianina two 1kb aenes exactlv with Needleman-Wunsch.*

**Summary:** Dynamic programming gives optimal alignments at quadratic cost.

## Exam Tips

- ★ Very important: the recurrence and the two algorithms' difference.
- ★ Remember  $O(nm)$ .

## Topic 8: BLAST and Heuristic Search

### Learning Objectives

- Explain BLAST's strategy
- Interpret E-values
- Choose BLAST variants

### Key Concepts

BLAST (Basic Local Alignment Search Tool): heuristic for searching huge databases fast. Strategy: find short exact 'seed' matches (words), extend them into high-scoring pairs, then evaluate statistically. Seeds make it fast; extension keeps sensitivity.

E-value: expected number of chance hits with this score -  $E=0.001$  means trustworthy;  $E=10$  means noise. Bit scores normalize across databases. Always judge by E-value, not raw score or percent identity alone.

Variants: blastn (DNA), blastp (protein), blastx (translated DNA vs protein - finds distant homologs), tblastn, PSI-BLAST (iterative profiles for remote homology). Choose by question and evolutionary distance.

### Key Points

- BLAST: seed, extend, evaluate - fast heuristic search.
- E-value judges significance; small is good.
- Match variant to molecule type and distance.

### Key Terms

**E-value:** Expected chance hits at a given score.

**Seed:** Short exact match starting extension.

**PSI-BLAST:** Iterative profile search for remote homologs.

### Example

- *blastx finds a gene's function across distant species.*

**Summary:** BLAST trades exactness for database-scale speed, with statistics.

### Exam Tips

- ★ Very important: E-value interpretation.
- ★ Remember seed-and-extend.

## Topic 9: Read Mapping

### Learning Objectives

- Explain short-read mapping
- Describe indexing strategies
- Discuss mapping challenges

### Key Concepts

Read mapping places millions of short reads onto a reference genome: the foundation of variant calling, RNA-seq, and CHIP-seq. Tools (BWA, Bowtie2) index the genome (Burrows-Wheeler transform + FM-index) for fast exact seed lookup, then extend with DP.

Challenges: repeats (reads map to multiple places - mapping quality flags ambiguity), sequencing errors, true variants vs. reference differences, and spliced reads (RNA-seq needs splice-aware aligners like STAR).

Paired-end reads rescue ambiguity: the pair's expected distance/orientation disambiguates placements and reveals structural variants. Mapping quality (MAPQ) quantifies placement confidence.

### Key Points

- Index (BWT/FM) + seed-extend maps millions of reads fast.
- Repeats cause ambiguous mapping; MAPQ flags it.
- Paired ends disambiguate and reveal structure.

### Key Terms

**Read mapping:** Placing reads on a reference.

**Burrows-Wheeler transform:** Reversible permutation enabling fast search.

**MAPQ:** Mapping quality score.

### Example

- *BWA-MEM mapping 100M reads for a human genome.*

**Summary:** Mapping turns read piles into positioned evidence.

### Exam Tips

- ★ Key point: repeat challenge and MAPQ.
- ★ Remember BWT indexing idea.

## Topic 10: Multiple Sequence Alignment

### Learning Objectives

- Explain MSA
- Describe progressive alignment
- Discuss applications

### Key Concepts

Multiple sequence alignment aligns 3+ sequences simultaneously, revealing conserved columns (functional sites) and evolutionary relationships. Exact DP is exponential - heuristics rule.

Progressive alignment (Clustal): align most-similar pairs first, progressively add sequences following a guide tree. Fast and popular; errors early propagate ('once a gap, always a gap'). Iterative refiners (MUSCLE, MAFFT) improve; consistency-based methods (T-Coffee) weigh evidence.

Applications: phylogenetic trees, protein family analysis, conserved motif discovery, and primer design. Alignment quality determines all downstream results - inspect alignments, don't blindly trust them.

### Key Points

- MSA reveals conservation across many sequences.
- Progressive: pairs first, then grow; fast but error-propagating.
- Quality determines downstream results - inspect.

### Key Terms

**Multiple sequence alignment:** Simultaneous alignment of 3+ sequences.

**Guide tree:** Rough phylogeny ordering progressive alignment.

**Conserved column:** A position unchanged across sequences.

### Example

- *Albion hemoalobin across species shows conserved oxvaen-binding sites.*

**Summary:** MSA exposes evolution's conserved core.

### Exam Tips

- ★ Key point: progressive strategy and its flaw.
- ★ Remember applications.

## Unit 3: Assembly and Annotation

### Topic 11: Genome Assembly

## Learning Objectives

- Define assembly
- Compare assembly paradigms
- Evaluate assemblies

## Key Concepts

Assembly reconstructs genomes from reads: overlap-layout-consensus (OLC: find read overlaps, layout, derive consensus - suits long reads) vs. de Bruijn graphs (chop reads into k-mers, build graph, find Eulerian paths - suits short reads).

Repeats are the enemy: identical repeats collapse, fragmenting assemblies. Long reads resolve repeats - the key to telomere-to-telomere genomes. Coverage depth (30-50x) provides error correction.

Metrics: N50 (contig length such that half the assembly is in pieces this size or larger), contig/scaffold counts, BUSCO (conserved gene completeness). High N50 with low BUSCO means long but wrong.

## Key Points

- OLC for long reads; de Bruijn graphs for short reads.
- Repeats fragment assemblies; long reads resolve them.
- N50 measures contiguity; BUSCO measures completeness.

## Key Terms

**Contig:** A contiguous assembled sequence.

**N50:** Contiguity metric for assemblies.

**k-mer:** A substring of length k.

## Example

- *T2T human genome finished with long reads.*

**Summary:** Assembly puzzles genomes together; repeats are the hard pieces.

## Exam Tips

- ★ Very important: N50 meaning.
- ★ Remember OLC vs. de Bruijn.

## Topic 12: Gene Prediction

### Learning Objectives

- Explain gene finding
- Describe statistical signals
- Compare ab initio and evidence-based

### Key Concepts

Gene prediction locates genes in raw DNA. Signals: start/stop codons, splice sites (GT-AG), codon bias (coding DNA uses codons unevenly), and conservation across species.

Ab initio predictors (Genscan, Augustus) use hidden Markov models combining signals statistically - good for novel genomes but error-prone. Evidence-based annotation adds RNA-seq (shows transcribed regions) and protein homology - far more accurate.

Eukaryotic complexity (introns, alternative splicing) makes this hard; pipelines (Ensembl, NCBI) combine methods with manual curation. Annotation quality varies - treat predicted genes skeptically.

### Key Points

- Signals: codons, splice sites, codon bias, conservation.
- Ab initio (HMMs) vs. evidence-based (RNA-seq, homology).
- Eukaryotic genes are hard; curation matters.

### Key Terms

**Ab initio prediction:** Gene finding from sequence signals alone.

**Hidden Markov model:** Statistical model of sequence states.

**Alternative splicing:** One gene producing multiple proteins.

### Example

- *Auauustus predictina aenes in a newlv sequenced funaus.*

**Summary:** Gene finding blends statistical signals with biological evidence.

### Exam Tips

- ★ Key point: signal types.
- ★ Remember ab initio vs. evidence-based.

## Topic 13: Functional Annotation

### Learning Objectives

- Explain functional annotation
- Describe GO terms
- Discuss annotation transfer

### Key Concepts

Beyond locating genes: what do they do? Functional annotation assigns functions via homology transfer ('similar sequence, similar function' - with caution), protein domains (Pfam/InterPro signatures), and Gene Ontology (GO) terms - a controlled vocabulary of molecular function, biological process, cellular component.

Enrichment analysis: given a gene list (e.g. differentially expressed), which GO terms are overrepresented? - revealing the biology behind experiments. Statistics (hypergeometric tests, multiple-testing correction) are essential.

Caveats: homology transfer errors propagate through databases; electronic annotations lack experimental backing - check evidence codes.

### Key Points

- Assign function via homology, domains, GO terms.
- Enrichment analysis interprets gene lists.
- Beware propagated annotation errors; check evidence codes.

### Key Terms

**Gene Ontology:** Controlled vocabulary of gene functions.

**Enrichment analysis:** Finding overrepresented functions in gene lists.

**Evidence code:** Annotation reliability indicator.

### Example

- *Enrichment shows stress-response genes active in drought data.*

**Summary:** Functional annotation turns gene lists into biological stories.

### Exam Tips

- ★ Key point: GO's three aspects.
- ★ Remember enrichment logic.

## Topic 14: Variant Calling

### Learning Objectives

- Define variant types
- Explain calling pipelines
- Discuss clinical relevance

### Key Concepts

Variants: SNPs (single-base changes), indels (small insertions/deletions), structural variants (large rearrangements), copy-number variants. Calling: map reads, identify positions differing from reference, filter artifacts - GATK's best-practices pipeline is the standard.

Challenges: distinguishing true variants from errors (quality/depth filters, VQSR), repetitive regions, and tumor heterogeneity in cancer. Annotation (VEP, ANNOVAR) predicts effects: synonymous vs. missense vs. loss-of-function.

Clinical: variant interpretation for rare disease and cancer (ACMG guidelines), pharmacogenomics (drug response), and population screening - where informatics meets medicine, with strict accuracy demands.

### Key Points

- SNPs, indels, SVs, CNVs - called from mapped reads.
- Filtering separates true variants from artifacts.
- Annotation + guidelines enable clinical interpretation.

### Key Terms

**SNP:** Single Nucleotide Polymorphism.

**Indel:** Insertion/deletion variant.

**VQSR:** Variant Quality Score Recalibration.

### Example

- *GATK pipeline calling variants for rare-disease diagnosis.*

**Summary:** Variant calling finds meaningful differences among billions of bases.

### Exam Tips

- ★ Very important: variant types.
- ★ Remember filtering necessity.

## Topic 15: RNA-seq and Expression Analysis

## Learning Objectives

- Explain RNA-seq workflow
- Describe quantification
- Introduce differential expression

## Key Concepts

RNA-seq measures gene expression: sequence RNA, map to genome/transcriptome, count reads per gene. Normalization (TPM, DESeq2 size factors) corrects for depth and gene length - raw counts mislead.

Differential expression: which genes change between conditions? Statistical models (negative binomial in DESeq2/edgeR) handle biological replicates and variance; multiple-testing correction (FDR) controls false discoveries among 20,000 genes.

Beyond: alternative splicing analysis, single-cell RNA-seq (cell-type resolution, huge sparse matrices), and spatial transcriptomics. Expression links genome to function.

## Key Points

- RNA-seq: sequence, map, count, normalize.
- DESeq2/edgeR model counts; FDR controls false discoveries.
- Single-cell adds cell-type resolution.

## Key Terms

**TPM:** Transcripts Per Million: normalized expression.

**Differential expression:** Statistically significant change between conditions.

**FDR:** False Discovery Rate.

## Example

- *Finding immune genes activated by infection via DESeq2.*

**Summary:** RNA-seq quantifies the genome in action.

## Exam Tips

- ★ Key point: why normalization matters.
- ★ Remember FDR.

# Unit 4: Comparative Genomics and Applications

## Topic 16: Comparative Genomics

### Learning Objectives

- Explain comparative approaches
- Describe synteny
- Discuss evolutionary insights

### Key Concepts

Comparing genomes reveals function: conserved sequences are likely functional (purifying selection preserves them). Phylogenetic footprinting finds regulatory elements conserved across species.

Synteny: conserved gene order across species - revealing rearrangements, duplications, whole-genome duplications (plants!). Dot plots and alignment chains visualize large-scale comparisons.

Insights: human-chimp comparison (~98.8% similar) localizes human-specific changes; across mammals, conserved non-coding elements point to regulatory innovation; pathogen comparisons track virulence evolution.

### Key Points

- Conservation implies function - compare to find it.
- Synteny tracks genome rearrangements.
- Comparisons reveal species-specific innovations.

### Key Terms

**Synteny:** Conserved gene order across species.

**Purifying selection:** Evolution removing harmful changes.

**Phylogenetic footprinting:** Finding functional elements via conservation.

### Example

- *Mammalian comparisons revealing conserved enhancers.*

**Summary:** Evolution's experiments, compared, annotate genomes.

### Exam Tips

- ★ Key point: conservation-function logic.
- ★ Remember synteny.

## Topic 17: Phylogenetics

### Learning Objectives

- Explain tree building
- Compare methods
- Interpret trees critically

### Key Concepts

Phylogenetic trees depict evolutionary relationships from sequence data. Methods: distance-based (neighbor-joining - fast), parsimony (fewest changes), maximum likelihood and Bayesian (statistical models of evolution - most rigorous).

Models matter: substitution models (Jukes-Cantor to GTR) correct for multiple hits; rate variation across sites; bootstrapping quantifies branch support - low-support branches are hypotheses, not facts.

Applications: pathogen outbreak tracing (COVID variants), species identification (DNA barcoding), gene family evolution. Trees need critical reading: support values, model choice, and data quality.

### Key Points

- Methods: NJ (fast), parsimony, ML/Bayesian (rigorous).
- Substitution models correct for hidden changes; bootstrap measures support.
- Applications: outbreaks, barcoding, gene families.

### Key Terms

**Phylogeny:** Evolutionary tree.

**Bootstrap:** Resampling support measure.

**Substitution model:** Statistical model of sequence evolution.

### Example

- *SARS-CoV-2 variant trees tracking the pandemic.*

**Summary:** Trees reconstruct evolution - with quantified uncertainty.

### Exam Tips

- ★ Key point: method comparison.
- ★ Remember bootstrap meaning.

## Topic 18: Metagenomics and Microbiomes

### Learning Objectives

- Define metagenomics
- Explain analysis workflows
- Discuss applications

### Key Concepts

Metagenomics sequences whole microbial communities directly - no culturing needed (most microbes don't culture). 16S rRNA profiling identifies 'who's there'; shotgun metagenomics reveals 'what they can do' (gene content).

Workflows: quality-filter reads, assign taxonomy (k-mer classifiers like Kraken), assemble contigs, bin genomes (MAGs - metagenome-assembled genomes), annotate functions. Contamination and host-DNA removal are critical.

Applications: human microbiome-disease links, environmental monitoring, pathogen discovery, antibiotic-resistance surveillance. The microbiome is a second genome - analyzed with genome informatics.

### Key Points

- Sequence communities directly; 16S for identity, shotgun for function.
- Binning recovers genomes (MAGs) from mixtures.
- Applications: health, environment, resistance surveillance.

### Key Terms

**Metagenomics:** Sequencing whole microbial communities.

**MAG:** Metagenome-Assembled Genome.

**16S rRNA:** Bacterial identification marker gene.

### Example

- *Gut microbiome studies linking bacteria to disease.*

**Summary:** Metagenomics reads entire ecosystems' genomes.

### Exam Tips

- ★ Key point: 16S vs. shotgun.

- ★ Remember MAGs.

## Topic 19: Genomics in Medicine and Agriculture

### Learning Objectives

- Describe clinical genomics
- Explain agricultural genomics
- Discuss implementation challenges

### Key Concepts

Medicine: rare-disease diagnosis (exome/genome sequencing ends diagnostic odysseys), cancer genomics (tumor profiling guiding therapy), pharmacogenomics (dosing by genotype), and population screening. Precision medicine runs on variant interpretation pipelines.

Agriculture: genomic selection accelerates breeding (predict breeding values from markers), disease-resistance genes, and crop improvement for climate resilience - feeding billions.

Challenges: variant interpretation uncertainty (VUS - variants of uncertain significance), data infrastructure, clinician training, cost and equity (who benefits?), and returning results responsibly.

### Key Points

- Clinical: rare disease, cancer, pharmacogenomics.
- Agriculture: genomic selection, resilient crops.
- Challenges: VUS, equity, infrastructure, counseling.

### Key Terms

**Precision medicine:** Treatment tailored to genetics.

**VUS:** Variant of Uncertain Significance.

**Genomic selection:** Breeding by marker-predicted values.

### Example

- *Exome sequencing diagnosis of a child's rare disease.*

**Summary:** Genomics transforms medicine and breeding - with interpretation challenges.

### Exam Tips

- ★ Key point: application areas.
- ★ Remember VUS.

## Topic 20: Ethics, Privacy, and Future of Genomics

### Learning Objectives

- Discuss genomic privacy
- Explain consent issues
- Survey future directions

### Key Concepts

Privacy: DNA identifies you and your relatives - de-identification fails for genomes (re-identification demonstrated); forensic genealogy and data breaches raise stakes. Policies (GINA in the US) partly protect against discrimination.

Consent: broad consent for future research vs. specific; returning incidental findings; Indigenous data sovereignty. Ethical frameworks evolve with technology.

Future: single-cell and spatial genomics, long-read clinical sequencing, pangenome references, AI variant interpretation, and genome editing (CRISPR) - informatics underpins each. The field's trajectory: more complete, more personal, more predictive - demanding matching ethical maturity.

## Key Points

- Genomes are inherently identifying; privacy needs active protection.
- Consent, incidental findings, data sovereignty are live issues.
- Future: single-cell, pangenomes, AI interpretation, editing.

## Key Terms

**Re-identification:** Linking 'anonymous' genomes to people.

**Incidental findings:** Unexpected clinically-relevant results.

**CRISPR:** Genome editing technology.

## Example

- *Golden State Killer case via forensic genealogy - power and concern.*

**Summary:** Genomic power demands ethical and privacy maturity.

## Exam Tips

- ★ Very important: why genomes can't be anonymized.
- ★ Remember future directions.

# Glossary of Key Terms

**Genome:** An organism's complete DNA sequence.

**Read:** A sequenced DNA fragment.

**FASTA/FASTQ:** Sequence formats; FASTQ adds quality scores.

**BLAST:** Fast heuristic sequence-database search tool.

**E-value:** Expected chance hits at a given alignment score.

**Dynamic programming:** Optimal alignment via subproblem recurrence.

**Contig:** A contiguous assembled sequence.

**N50:** Assembly contiguity metric.

**SNP:** Single Nucleotide Polymorphism.

**VCF:** Variant Call Format.

**RNA-seq:** Sequencing RNA to measure gene expression.

**Synteny:** Conserved gene order across species.

**Phylogeny:** Evolutionary tree of sequences or species.

**Metagenomics:** Sequencing whole microbial communities.

**Pangenome:** The full variant/gene set of a population or species.