Glossary of Key Terms in Computational Biology Program

Genome
The complete set of genetic material in an organism.

Gene
A segment of DNA that codes for a protein or functional RNA.

Allele
A variant form of a gene.

Chromosome
A DNA molecule carrying genetic material.

Single Nucleotide Polymorphism (SNP)
A single base-pair variation in the genome.

Copy Number Variation (CNV)
A segment of DNA with variable copy numbers among individuals.

Variant Allele Frequency (VAF)
The proportion of sequencing reads showing a specific variant.

Mutation
A change in the DNA sequence.

Genotype
The genetic constitution of an individual.

Structural Variation
Large-scale alterations in chromosome structure.

Gene Amplification
An increase in the number of copies of a gene.

Gene Deletion
Loss of a DNA segment from the genome.

Promoter
A DNA region initiating transcription of a gene.

Enhancer
A DNA sequence that increases gene transcription levels.

Oncogene
A gene with the potential to cause cancer.

Tumor Suppressor Gene
A gene that protects cells from cancer formation.

Exon
A gene segment that codes for amino acids.

Intron
A non-coding segment within a gene.

Reference Genome
A standard sequence used for comparison in genomics.

Whole Genome Sequencing (WGS)
Sequencing the entire genome of an organism.

Transcriptome
The complete set of RNA transcripts in a cell.

RNA-Seq
A technique for analyzing the transcriptome using sequencing.

Differential Gene Expression
Changes in gene expression between conditions.

Reads Per Kilobase Million (RPKM)
Normalization metric for RNA-seq data.

Fragments Per Kilobase Million (FPKM)
Another RNA-seq normalization method.

Transcription Factor
A protein that regulates gene expression.

Splicing
Removing introns from RNA transcripts.

Non-coding RNA
RNA molecules not translated into protein.

miRNA (microRNA)
Small RNAs that regulate gene expression post-transcriptionally.

Gene Expression Matrix
A table of expression values for genes across samples.

Proteome
The full set of proteins expressed in a cell or organism.

Mass Spectrometry
A method to identify and quantify proteins.

Peptide
A short chain of amino acids.

Protein-Protein Interaction (PPI)
Physical contacts between proteins in a cell.

Post-translational Modification
Chemical changes to proteins after translation.

Western Blot
Technique to detect specific proteins in a sample.

Protein Quantification
Measuring the amount of protein in a sample.

Pathway Enrichment
Identifying biological pathways overrepresented in a dataset.

KEGG Pathway
Pathway database for biological interpretation of gene sets.

Reactome
A database of biological pathways and reactions.

Supervised Learning
ML with labeled input-output pairs.

Unsupervised Learning
ML that infers patterns from unlabeled data.

Deep Learning
Neural network-based learning for complex patterns.

Random Forest
An ensemble ML method using decision trees.

Support Vector Machine (SVM)
A classifier that separates data with a hyperplane.

Convolutional Neural Network (CNN)
Deep learning model for spatial data like images.

Feature Engineering
Creating input variables for ML from raw data.

Training Data
Data used to fit an ML model.

Model Evaluation
Assessing a model's performance.

Cross-Validation
Evaluating models by training/testing on different data splits.

Overfitting
Model performs well on training data but poorly on new data.

Underfitting
Model fails to capture the underlying data pattern.

ROC Curve
Graph showing performance of a classifier.

Accuracy
Correct predictions divided by total predictions.

Precision
True positives divided by predicted positives.

Recall
True positives divided by actual positives.

AUC (Area Under Curve)
Overall performance metric of classification model.

Feature Importance
Measure of how valuable each feature is to the model.

Label Encoding
Transforming labels into numerical form.

Normalization
Scaling data to a standard range.

Metastasis
The spread of cancer from one site to another.

Tumor Microenvironment
The environment around a tumor including surrounding cells and molecules.

Carcinogenesis
The formation of cancer.

Histology
Study of tissue structure under a microscope.

EGFR
A receptor often mutated in cancers.

HER2
A protein overexpressed in some breast cancers.

Hormone Receptor Status
Indicates hormone sensitivity of a tumor.

Tumor Grade
Describes how abnormal tumor cells look.

Tumor Stage
Extent of cancer spread.

Primary Tumor
The original site where cancer began.

Systems Biology
Study of complex interactions in biological systems.

Simulation
Using computational models to mimic biological processes.

ODE
Ordinary Differential Equations used in simulations.

Agent-Based Modeling
Simulates actions of individual agents in a system.

In Silico
Performed via computer simulation.

Sensitivity Analysis
Testing how changes in inputs affect model output.

Parameter Estimation
Finding model parameters that best fit data.

Biological Network
Graph-based representation of biological interactions.

Feedback Loop
Circular pathway where output feeds back as input.

Steady State
Condition where system variables remain constant.

FastQC
Tool for checking quality of sequencing reads.

GATK
Genome Analysis Toolkit for variant discovery.

bedtools
Suite of tools for genomic interval operations.

bcftools
Toolset for manipulating VCF/BCF files.

SAMtools
Tools for working with SAM/BAM sequencing data.

FeatureCounts
Tool for counting reads mapped to genomic features.

DESeq2
R package for differential expression analysis.

edgeR
Bioconductor tool for RNA-seq analysis.

Biopython
Python tools for bioinformatics.

Bioconductor
R-based platform for biological data analysis.

Annotation File
File containing gene or region information.

GTF/GFF/BED Format
File formats used to describe genomic features.

VCF
Format for storing genetic variants.

BAM/FASTQ File
Formats for aligned reads and raw sequencing reads.

Git & GitHub
Tools for version control and collaboration.

Docker
Tool for containerizing applications.

Snakemake
Workflow management system for data analysis.

Nextflow
Workflow manager for scalable bioinformatics.

Documentation
Instructions and descriptions for code and workflows.

Reproducibility
Ensuring others can repeat and verify analyses.