Lecture 10. Multi-Omics Integration, Bioinformatics, and Systems Biology

Integration of genomic, transcriptomic, proteomic and metabolomic data; biological networks; databases; pathway analysis; PCA and clustering; machine learning and AI; systems biology.

Prof. Kenzhebayeva S.S. Dep. Biotecnology. Al-Farabi Kazakh national University

Aim of the lecture. Expected learning outcomes

Aim of the lecture

To introduce the principles and computational approaches of multi-omics integration, explain the use of biological databases, molecular networks, pathway analysis, multivariate statistics, machine learning, and artificial intelligence, and demonstrate their importance for understanding complex biological systems and identifying molecular mechanisms relevant to biotechnology.

Expected learning outcomes

- Explain the concept and importance of multi-omics integration.
- Distinguish between different strategies for integrating omics datasets.
- Identify major biological databases and their applications.
- Describe gene regulatory, protein interaction, and metabolic networks.
- Explain the principles of functional enrichment and pathway analysis.
- Interpret principal component analysis (PCA) and clustering results.
- Describe basic applications of machine learning and AI in omics research.
- Explain the main principles of systems biology.
- Discuss how integrated omics can support functional gene discovery, stress biology, and crop improvement.

Lecture plan

1. Introduction to multi-omics integration
2. Integration strategies and bioinformatics workflows
3. Biological databases and networks
4. Functional enrichment and pathway analysisPCA, clustering, and data visualization
5. Machine learning and AISystems biology and plant biotechnology applications
6. Challenges, conclusion, control questions, literature

Introduction to multi-omics integration

What is multi-omics integration?

Multi-omics integration is the combined analysis of two or more molecular data types to obtain a more comprehensive understanding of biological processes.

The principal data layers include:

- Genomics — genetic variants and genome structure.
- Transcriptomics — gene-expression patterns.
- Proteomics — protein abundance and regulation.
- Metabolomics — biochemical composition and metabolic changes.
- Phenomics — measurable biological characteristics.

Key concept: Integrated analysis can reveal relationships between molecular layers that may remain undetected when each dataset is studied independently.

Why is multi-omics integration necessary?

Biological processes are controlled by interconnected molecular mechanisms.

For example:

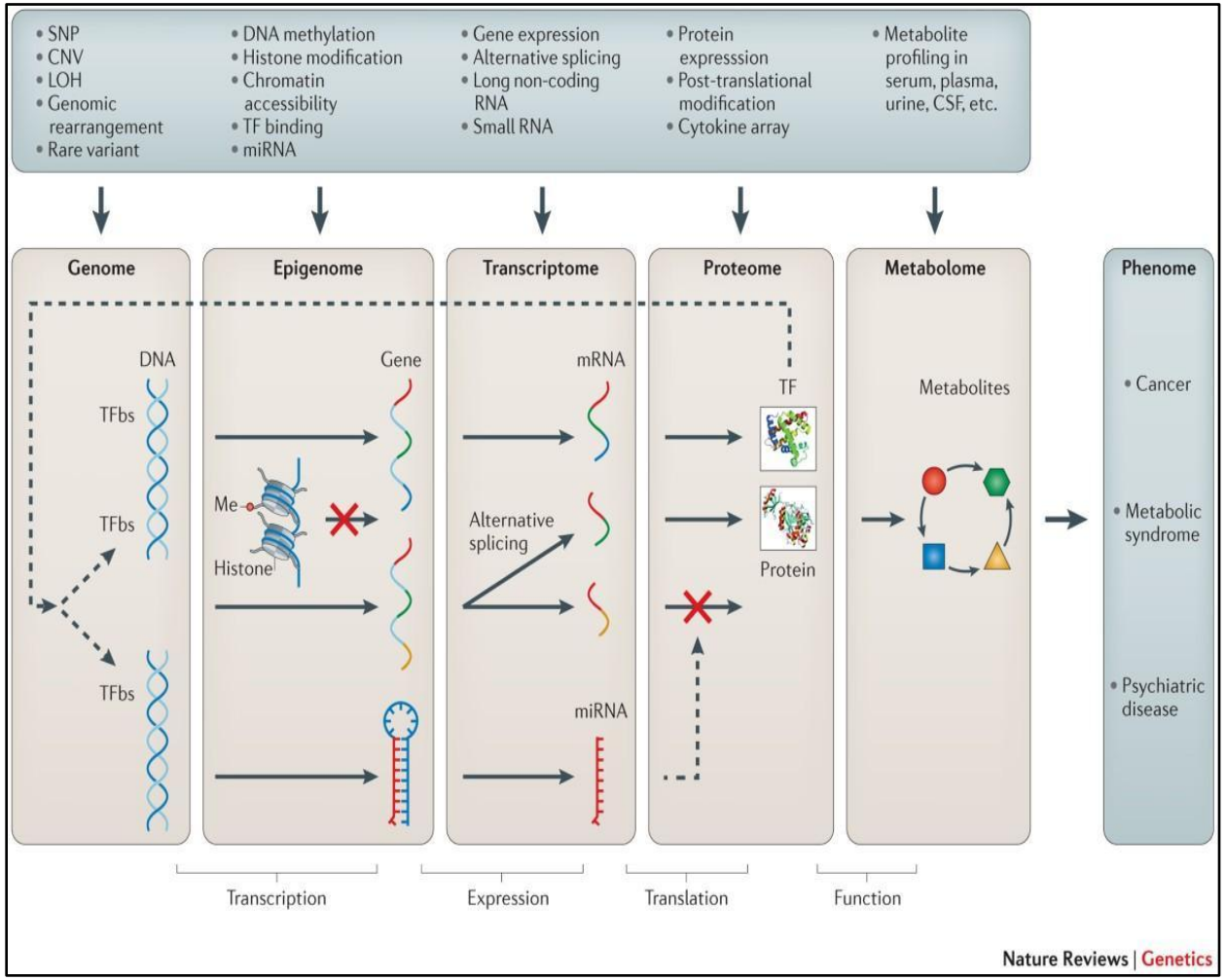
- A DNA variant may influence gene expression.
- Changes in transcription may alter protein abundance.
- Enzyme activity may affect metabolite concentrations.
- Metabolic regulation may influence physiological traits.
- However, these relationships are not always linear or directly proportional.

Main objective: To identify coordinated molecular changes and generate testable hypotheses about biological mechanisms

From genotype to phenotype

- Genomics — DNA variants
- Transcriptomics — RNA expression
- Proteomics — proteins and enzymes
- Metabolomics — biochemical processes
- Phenomics — measurable traits

Feedback regulation, developmental processes, and environmental factors influence all levels



Integration of genomic, transcriptomic, proteomic, and metabolomic information provides a comprehensive framework for understanding biological functions and phenotypic variation

Single-omics versus multi-omics approaches

Multi-omics analysis complements rather than replaces well-designed single-omics experiments.

Feature	Single-omics	Multi-omics
Data	One molecular layer	Multiple molecular layers
Biological perspective	Layer-specific	Integrated
Main advantage	Focused analysis	Broader mechanistic interpretation
Complexity	Usually lower	Higher
Applications	Gene or protein profiling	Network and systems-level analysis

Major objectives of multi-omics research

1. Identify coordinated molecular responses.
2. Discover candidate genes and regulatory pathways.
3. Characterize molecular interactions.
4. Investigate genotype–phenotype relationships.
5. Identify potential biomarkers.
6. Develop predictive models.
7. Support functional validation and biotechnology applications.

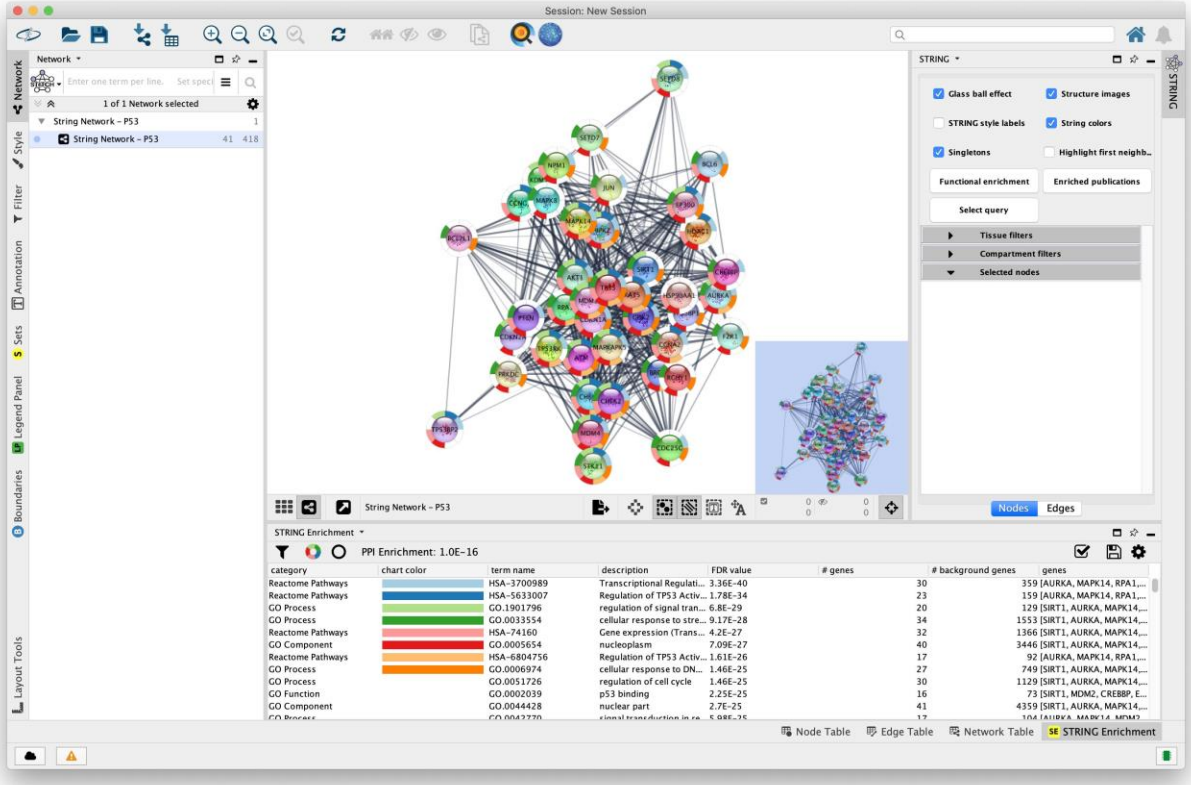
Integration strategies and workflows

Main strategies of multi-omics integration

Three common strategies are:

1. Early integration: Combine processed features from multiple omics layers into one dataset before modeling.
2. Intermediate integration: Extract shared or complementary representations using joint statistical or computational models.
3. Late integration: Analyze each omics layer separately and subsequently combine the results or predictions.

Each strategy has advantages depending on sample size, data structure, and research objectives.

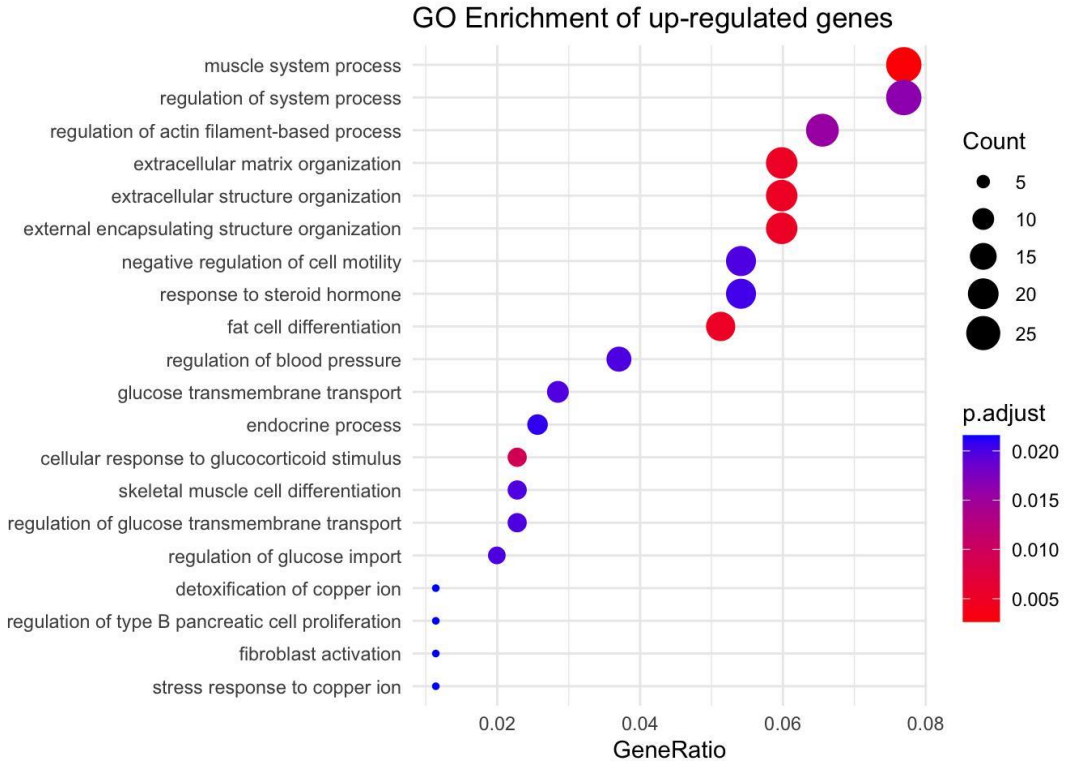


Biological networks represent molecular components as nodes and their interactions or functional associations as edges. Network analysis helps identify functional modules and highly connected components.

Experimental design for multi-omics studies

1. Reliable integration begins with an appropriate experimental design.
2. Important considerations:
3. Clear biological hypothesis.
4. Comparable biological samples.
5. Adequate replication.
6. Consistent sampling conditions.
7. Developmental stage and tissue specificity.
8. Randomization and batch control.
9. Availability of relevant phenotypic measurements.

Where possible, molecular layers should be measured using matched biological samples.



Gene Ontology enrichment analysis identifies biological functions statistically overrepresented among selected genes or proteins

Data preprocessing and normalization

Different omics platforms generate data with different units, distributions, and technical biases.

Typical preprocessing includes:

1. Quality control.
2. Removal of low-quality features.
3. Handling missing values.
4. Appropriate normalization.
5. Batch-effect assessment.
6. Feature filtering.
7. Data transformation and scaling.

Important: A single normalization procedure is not necessarily suitable for every omics dataset.

General multi-omics workflow

Biological samples and experimental design

Genomics + Transcriptomics + Proteomics + Metabolomics

Quality control and preprocessing

Data integration and statistical analysis

Network and pathway interpretation

Candidate mechanisms and experimental validation

Biological databases and molecular networks

Biological databases in multi-omics research

Biological databases provide reference sequences, molecular annotations, interaction information, and metabolic pathways.

Database	Main information
NCBI	Genome sequences, genes, transcripts, proteins
Ensembl	Genome annotation and comparative genomics
UniProt	Protein sequences and functional annotation
Gene Ontology / QuickGO	Gene and protein functional terms
KEGG	Biological pathways and molecular functions
Metabolomics Workbench	Metabolomics datasets and metadata

Biological networks: concept and significance

A biological network is a representation of relationships among molecular or cellular components.

Network elements:

- Nodes: genes, proteins, metabolites, or other biological entities.
- Edges: interactions, regulatory relationships, reactions, or statistical associations.

Biological networks help investigate molecular organization and identify potential regulatory components.

Gene regulatory networks (GRNs)

Gene regulatory networks describe relationships among transcription factors, regulatory elements, and target genes.

A simplified example:

Stress signal → transcription factor → target genes → physiological response

GRNs are useful for investigating:

1. Developmental regulation.
2. Cell differentiation.
3. Environmental stress responses.
4. Gene-expression coordination.

Computationally inferred regulatory links require experimental validation.

Protein–protein interaction networks

Protein–protein interaction (PPI) networks represent physical interactions or functional associations between proteins.

Applications include:

- Identification of protein complexes.
- Investigation of signaling pathways.
- Functional annotation of poorly characterized proteins.
- Identification of highly connected proteins.

The STRING database is a useful resource for exploring protein association networks.

Checking your browser - reCAPTCHA

Metabolic networks

Metabolic networks represent biochemical reactions connecting substrates, intermediates, products, and enzymes.

For example:

- Glucose
- Glycolysis
- Pyruvate
- Acetyl-CoA and central carbon metabolism

Metabolic network analysis supports investigation of carbon metabolism, amino-acid biosynthesis, energy metabolism, and responses to stress.

Network visualization and analysis

Network analysis commonly investigates:

- Degree: number of connections associated with a node.
- Betweenness centrality: extent to which a node lies on shortest paths.
- Network modules: groups of closely associated components.
- Hubs: highly connected nodes.

Cytoscape is widely used for visualizing and analyzing biological networks.

Important: High network connectivity does not automatically establish biological essentiality or causal regulation.

Functional enrichment and pathway analysis.

Functional enrichment analysis

Functional enrichment analysis determines whether particular biological functions or pathways are represented more frequently than expected in a selected gene or protein set.

Common analyses include:

- Gene Ontology enrichment.
- KEGG pathway enrichment.
- Reactome pathway analysis.
- Plant-specific metabolic pathway analysis.

Gene Ontology (GO) analysis

Gene Ontology classifies gene products into three major categories.

Biological Process

What biological process is involved?

Stress response, photosynthesis

Molecular Function

What molecular activity is performed?

Enzyme activity, DNA binding

Cellular Component

Where is the gene product located?

Chloroplast, nucleus

Pathway enrichment analysis

Pathway enrichment identifies biological pathways associated with selected molecular features.

Example: analysis of drought-responsive wheat genes may reveal enrichment of:

1. ABA signaling.
2. Proline biosynthesis.
3. Antioxidant defense.
4. Photosynthetic processes.
5. Carbohydrate metabolism.

Enrichment indicates statistical overrepresentation, not necessarily activation of the entire pathway.

Integrating genes, proteins, and metabolites into pathways

1. Differentially expressed genes
2. Differentially abundant proteins
3. Differential metabolites
4. Pathway mapping and integrated interpretation
5. Candidate regulatory and metabolic mechanisms

This approach can identify concordant or discordant changes across molecular layers.

PCA, clustering, and data visualization.

Principal component analysis (PCA)

Principal component analysis is an unsupervised statistical method that reduces high-dimensional datasets to a smaller number of components representing major sources of variation.

Applications:

1. Visualization of sample similarities.
2. Detection of outliers.
3. Assessment of batch effects.
4. Exploration of treatment-related variation.
5. Illustrative PCA scores plot

Conceptual example with simulated sample coordinates; not experimental data.

Interpretation of PCA results

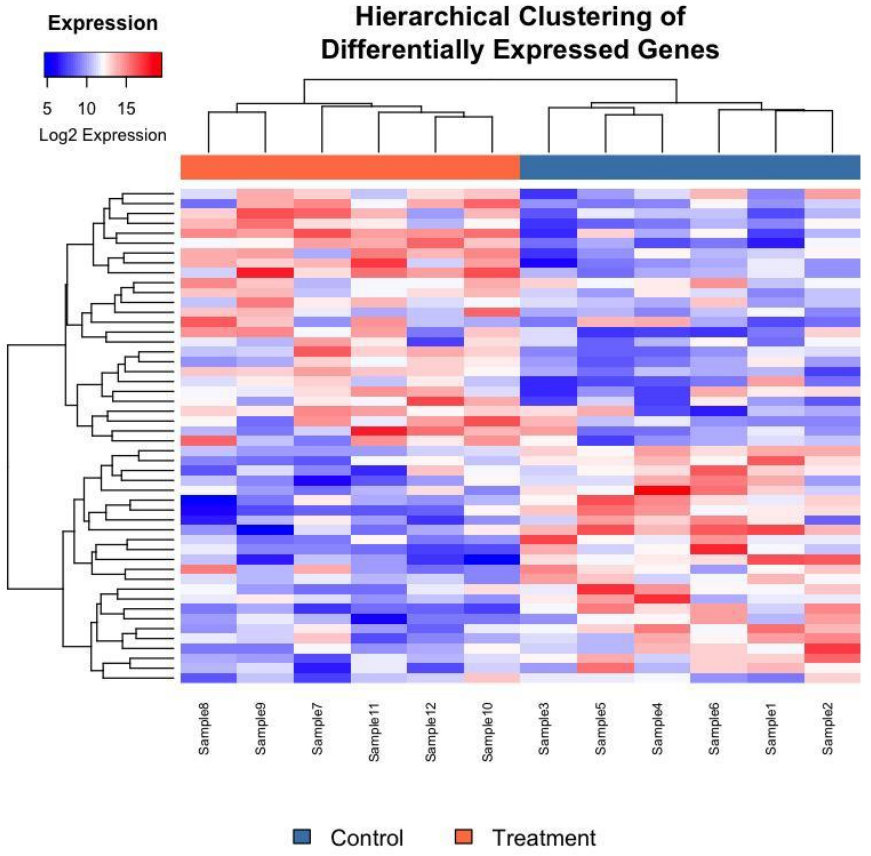
PCA produces:

Scores: positions of samples in principal-component space.

Loadings: contributions of original variables to the components.

Important considerations:

1. Components are ordered by explained variance.
2. Separation may reflect treatment, genotype, batch, or other sources.
3. PCA does not independently establish statistical significance.
4. Scaling and normalization can strongly influence results.



Clustering methods

Clustering groups samples or molecular features according to their similarities.

Common approaches:

1. Hierarchical clustering.
2. K-means clustering.
3. Density-based clustering.
4. Model-based clustering.

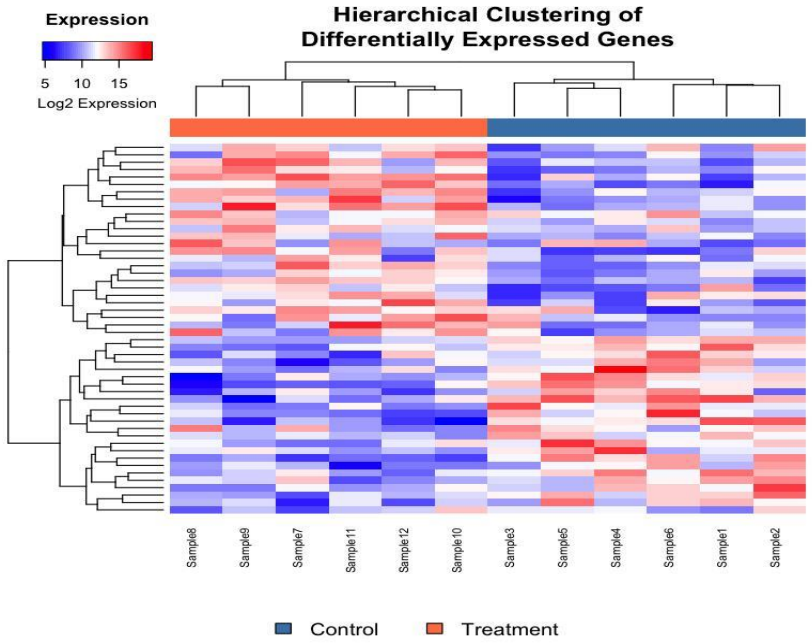
Applications include identifying co-expressed genes, metabolite groups, and biological sample subpopulations.

Heatmaps and hierarchical clustering

A heatmap displays numerical values using a color scale, often accompanied by clustering dendrograms.

Illustrative gene-expression heatmap

1. C1
2. C2
3. T1
4. T2
5. G1
6. G2
7. G3



1. G4
2. G5
3. G6
4. Lower
5. Higher

Simulated standardized expression values. C = control; T = treatment.

Heatmaps are useful for visualizing coordinated molecular responses but do not establish causal relationships.

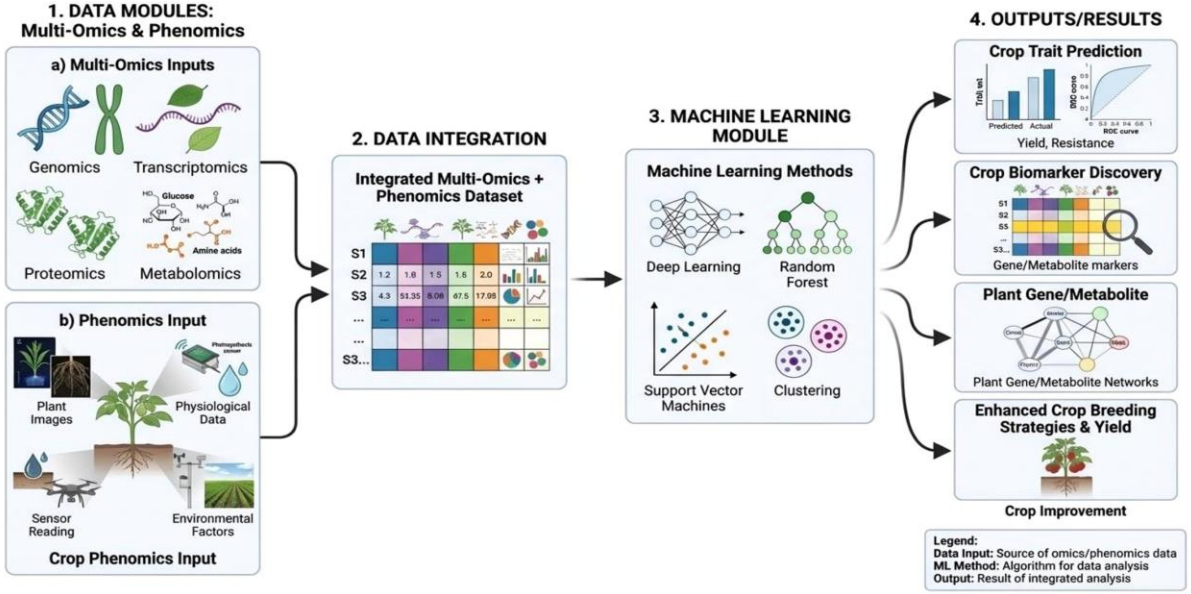
Heatmap with dendrogram

Machine learning and artificial intelligence (*Machine learning in omics research)*

Machine learning (ML) uses computational algorithms to identify patterns and make predictions from data.

1. Major categories:
2. Supervised learning.
3. Unsupervised learning.
4. Deep learning.

Applications include biomarker discovery, phenotype prediction, molecular classification, and analysis of complex biological datasets.



Machine-learning approaches integrate molecular features to predict biological traits and identify candidate biomarkers. Independent validation is necessary to evaluate model performance

Supervised versus unsupervised learning

Feature	Supervised learning	Unsupervised learning
Training data	Labeled outcomes	No outcome labels required
Objective	Predict a known outcome	Discover underlying patterns
Methods	Random forest, SVM, regression	PCA, clustering
Applications	Stress-tolerance prediction	Molecular subgroup discovery

Artificial intelligence and deep learning

AI-based approaches can analyze complex molecular datasets and identify nonlinear relationships.

Examples:

1. Neural networks for phenotype prediction.
2. Deep learning for sequence analysis.
3. Protein structure prediction.
4. Integration of imaging and molecular datasets.
5. Prediction of molecular interactions.

Model training and validation

Reliable machine-learning analysis requires:

1. Data preprocessing and feature selection.
2. Training, validation, and independent test datasets.
3. Appropriate cross-validation.
4. Prevention of data leakage.
5. Model performance assessment.
6. Biological interpretation and external validation.

For plant datasets, validation across genotypes, environments, and experimental batches is particularly important.

Systems biology and plant biotechnology

(*Systems biology: concept and principles)*

Systems biology investigates biological systems as interconnected networks of molecular and cellular components.

Its main principles include:

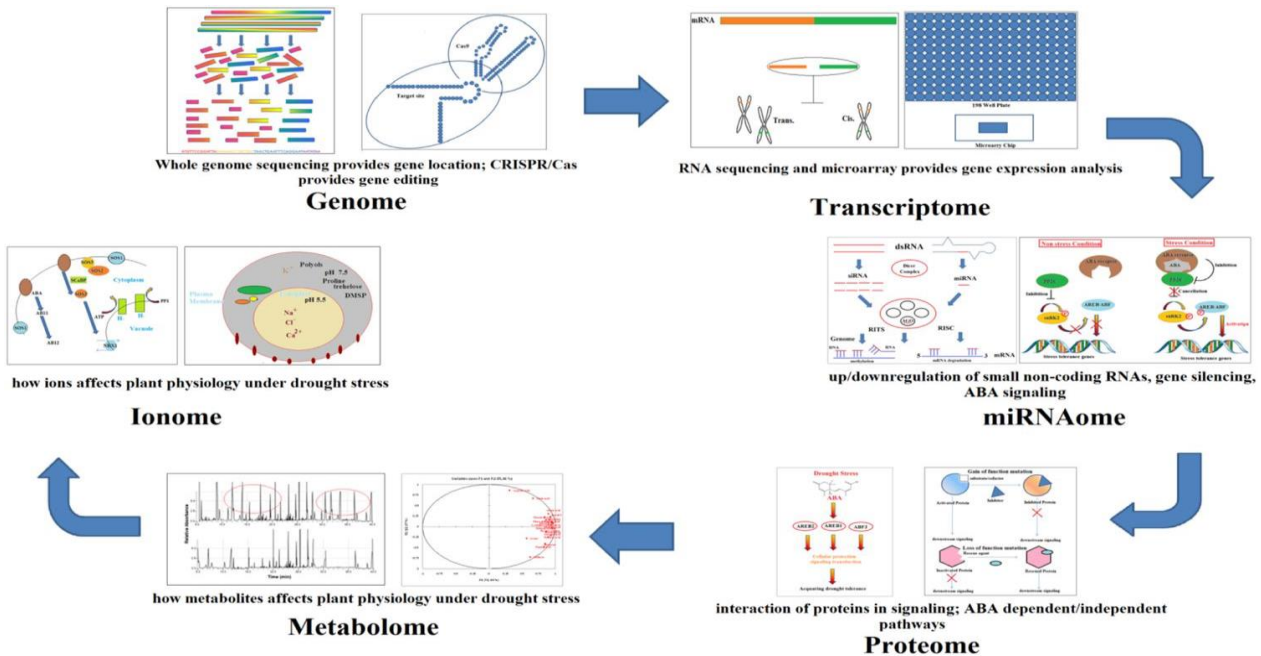
1. Integration of experimental data.
2. Network-based interpretation.
3. Mathematical and computational modeling.
4. Investigation of feedback regulation.
5. Prediction and experimental testing.

Systems biology moves beyond studying individual genes toward understanding coordinated biological functions.

Case study: multi-omics investigation of wheat drought tolerance

Integrated experimental framework

1. Genomics
2. SNPs and candidate loci
3. Transcriptomics
4. Stress-responsive genes
5. Proteomics
6. Antioxidant and metabolic proteins
7. Metabolomics
8. Proline, sugars, organic acids
9. Integrated pathway and network analysis
10. Phenotyping: photosynthesis, water status, growth and yield



Candidate mechanisms → functional validation → breeding applications

This is a proposed research framework. Multi-omics associations help prioritize candidate mechanisms but do not independently demonstrate causality.

Challenges and future perspectives

Major challenges:

- Differences in omics data structure and scale.
- Missing values and incomplete molecular coverage.
- Limited sample sizes relative to feature numbers.
- Batch effects and technical variability.
- Difficulties in linking statistical associations to mechanisms.
- Model overfitting and limited generalizability.
- Need for experimental validation.

Future directions include single-cell multi-omics, spatially resolved integration, interpretable AI, and predictive systems biology.

Conclusion

Multi-omics integration combines genomic, transcriptomic, proteomic, and metabolomic information to provide a systems-level understanding of biological processes.

Biological databases, molecular networks, pathway analysis, PCA, clustering, and machine learning support the interpretation of complex datasets.

Integration with phenotypic measurements and experimental validation creates opportunities for identifying functional genes, investigating stress-response mechanisms, and improving crops through biotechnology.

Control questions

1. What is multi-omics integration?
2. Why is multi-omics analysis important in postgenomic biology?
3. What are early, intermediate, and late integration?
4. Which databases are commonly used in omics research?
5. What is a biological network?
6. How do gene regulatory networks differ from protein interaction networks?
7. What is functional enrichment analysis?

1. What information does pathway analysis provide?
2. What is PCA, and how are its results interpreted?
3. What is the purpose of clustering?
4. How do supervised and unsupervised learning differ?
5. What is the role of AI in multi-omics research?
6. What are the main principles of systems biology?
7. How can integrated omics contribute to wheat drought-tolerance research?

Recommended literature

- Hasin, Y.; Seldin, M.; Lusis, A. Multi-omics approaches to disease. Genome Biology, 2017, 18, 83. DOI: 10.1186/s13059-017-1215-1.
- Argelaguet, R. et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 2018, 14, e8124. DOI: 10.15252/msb.20178124.
- Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Computational Biology, 2017, 13, e1005752. DOI: 10.1371/journal.pcbi.1005752.
- Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research, 2023, 51, D638–D646. DOI: 10.1093/nar/gkac1000.
- Pang, Z. et al. MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. Nucleic Acids Research, 2024, 52, W398–W406.