Lecture 2. Transcriptomics and Gene Expression Profiling

Aim of the lecture

To introduce the fundamental principles of transcriptomics and modern technologies for gene expression profiling, with emphasis on RNA sequencing, transcriptome data analysis, and their applications in functional genomics and plant biotechnology.

1. From genome sequence to genome function

Lecture 1 established:

Genome → Transcriptome → Proteome → Metabolome → Phenome

Now we focus specifically on the first transition:

Genome sequence → genomic variation → gene function → phenotype

The availability of a complete genome sequence provides ***a molecular "map,"*** but the next challenge is to determine:

1. **Where are the genes?**
2. **What do these genes encode?**
3. **What functions do they perform?**
4. **How are they regulated?**
5. **How do genes interact?**
6. **Which genomic variants influence phenotype?**

This is the central task of functional genomics.

Expected Learning Outcomes

Expected Learning Outcomes

1. Define transcriptomics and explain its role in postgenomic research.
2. Describe the major types of RNA and their biological functions.
3. Explain the principles of gene expression regulation and transcriptome dynamics.
4. Compare traditional and high-throughput gene expression profiling technologies.
5. Describe the main stages of RNA sequencing (RNA-seq), from RNA isolation to data analysis.
6. Explain differential gene expression analysis and functional annotation.
7. Identify applications of transcriptomics in plant development, stress responses, and crop improvement.
8. Recognize the importance of integrating transcriptomic data with other omics technologies.

Lecture Plan

1. Introduction to Transcriptomics: Definition, objectives, and significance in postgenomic biology.
2. RNA Diversity and Gene Expression Regulation: Types of RNA, transcription, RNA processing, and regulation.
3. Technologies for Gene Expression Profiling: RT-qPCR, DNA microarrays, and RNA sequencing.
4. RNA Sequencing Technologies: Principles, library preparation, sequencing platforms, and experimental design.
5. Bioinformatics Analysis of RNA-seq Data: Quality control, read alignment, transcript quantification, and differential expression.
6. Functional Interpretation of Transcriptomic Data: Gene ontology, pathway enrichment, and regulatory networks.
7. Applications in Plant Biotechnology: Plant development, environmental stress responses, and crop improvement.
8. Challenges and Future Perspectives: Technical limitations, single-cell transcriptomics, and integration with multi-omics approaches.

A molecular “map”

A molecular “map” is a representation of the **location and organization of genes, genetic markers, regulatory elements, and other DNA sequences within a genome or chromosome**.

A molecular map is a representation of the genome showing the positions and relationships of genes, genetic markers, and other DNA sequences along chromosomes.

- **Why do we call it a “map”?**
- Just as a geographical map shows the locations of cities and roads, a molecular map shows the positions of genomic features:

Chromosome → molecular markers → genes → regulatory regions → genomic variants

- One small clarification: in our earlier sentence, **“molecular map” was used rather broadly**. For greater scientific precision in Lecture 2, I would write:
- **The availability of a complete genome sequence provides a reference framework for locating genes and other genomic elements, but the next challenge is to determine their biological functions.**

“reference framework” here because *molecular map* can also have a more specific meaning in genetics, particularly **genetic linkage maps** and **physical maps**.

2. What is functional genomics?

Functional genomics is the genome-wide study of the **functions, regulation, and interactions of genes and other genomic elements**.

Unlike classical molecular genetics, which often investigates one or several genes individually, functional genomics uses **large-scale and high-throughput approaches**.

Classical genetics

One gene → one mutation → one phenotype

Functional genomics

Thousands of genes/variants → expression and interaction networks → multiple phenotypes

Thus, functional genomics represents a transition:

from studying individual genes to studying the genome as an integrated functional system.

3. Structural, comparative, and functional genomics

They are strongly interconnected:

Genome sequencing

↓

Genome assembly

↓

Genome annotation

↓

Comparative analysis

↓

Functional analysis

↓

Phenotypic interpretation

Field	Main question	Examples
Structural genomics	What is present in the genome?	Sequencing, assembly, gene location
Comparative genomics	How do genomes differ?	Orthologs, synteny, genome evolution
Functional genomics	What do genomic elements do?	Gene expression, knockout, CRISPR, functional screens

4. Genome annotation

After sequencing and assembling a genome, researchers must identify its functional elements.

This process is called **genome annotation**.

Two major levels can be distinguished.

Structural annotation. Determines the location and organization of:

1. genes;
2. exons;
3. introns;
4. untranslated regions;
5. repetitive sequences;
6. regulatory regions;
7. non-coding RNAs.

Functional annotation. Attempts to determine:

1. **What biological function does this gene or genomic element perform?**

5. From sequence similarity to predicted gene function

One of the basic principles is:

Unknown gene

↓

Sequence comparison

↓

Identification of homologous sequences

↓

Protein-domain analysis

↓

Functional annotation

↓

Predicted biological function

For example, if an unknown wheat gene has strong sequence similarity to a characterized transporter in another plant, it may initially be annotated as a **transporter-like gene**.

But this is a crucial point: **Bioinformatic prediction is not equivalent to experimental proof of function.**

Experimental validation remains necessary.

6. Gene Ontology — GO

Gene Ontology (GO) provides standardized terminology for describing gene and protein functions.

- GO uses three major categories:

1. **Molecular Function**
2. **What does the molecule do?**

Examples:

1. **transporter activity;**
2. **kinase activity;**
3. **DNA-binding activity.**

Biological Process

Which biological process is it involved in?

Examples:

1. **response to oxidative stress;**
2. **photosynthesis;**
3. **iron homeostasis.**

Cellular Component

Where does it function?

Examples:

1. **nucleus;**
2. **chloroplast;**
3. **plasma membrane;**
4. **vacuole.**

- A diagram is:

Gene → Molecular function + Biological process + Cellular location

7. Genome-wide genetic variation

Individuals belonging to the same species do not possess completely identical genomes.

Major forms of genetic variation include:

- **SNPs** — single nucleotide polymorphisms;
- **indels** — small insertions and deletions;
- **CNVs** — copy-number variations;
- larger insertions/deletions;
- inversions;
- duplications;
- translocations;
- presence/absence variations.

Genome-wide genetic variation

SNP

For example:

Reference:

…A C G T A…

Variant:

…A C A T A…

A single nucleotide has changed:

G → A

SNPs are among the most extensively used genomic markers in modern genetics and breeding.

8. Functional consequences of genomic variants

Not every genomic variant produces a phenotypic effect.

A variant may occur in:

Coding region

→ **synonymous mutation**
→ **missense mutation**
→ **nonsense mutation**
→ **frameshift**

or in:

Regulatory region

→ altered gene expression

or in:

Non-coding region

→ potentially neutral or regulatory consequences.

Therefore:

DNA variant

↓

Possible change in RNA/protein

↓

Change in cellular function

↓

Possible phenotypic effect

This is an important bridge between **genomics and postgenomics**.

Forward Genetics: From Phenotype to Gene

Forward genetics is an experimental approach used to identify genes responsible for specific biological traits by starting with an observable phenotype and determining its underlying genetic basis.

Main principles:

- Begins with a phenotype of interest, such as drought tolerance, disease resistance, or altered plant growth.
- Uses natural genetic variation or experimentally induced mutations to generate phenotypic diversity.
- Identifies individuals exhibiting desirable or unusual traits through phenotypic screening.
- Applies genetic mapping, sequencing, and molecular markers to identify candidate genes.
- Uses genetic and functional validation to establish relationships between mutations and phenotypes.

Major approaches: Induced mutagenesis, genetic mapping, bulked segregant analysis (BSA), QTL mapping, and genome-wide association studies (GWAS).

Major applications:

- Plant biotechnology: *Identification of genes associated with drought tolerance, salinity resistance, disease resistance, yield, and grain quality.*
- Medical biotechnology: *Discovery of genes involved in hereditary disorders, disease susceptibility, and cellular dysfunction.*
- Microbial biotechnology: *Identification of mutations affecting microbial growth, enzyme production, and metabolite biosynthesis.*
- Industrial biotechnology: *Development of improved microbial strains for fermentation, biofuel production, and synthesis of valuable bioproducts.*
- Environmental biotechnology*: Identification of genes associated with pollutant degradation, heavy-metal tolerance, and environmental adaptation.*
- Pharmaceutical biotechnology: *Discovery of genes regulating biosynthesis of antibiotics and other bioactive compounds.*

10. Reverse genetics

Reverse genetics begins with a known gene and investigates what happens when its function is altered.

The logic is:

Gene → phenotype

A gene can be:

1. **knocked out;**
2. **knocked down;**
3. **overexpressed;**
4. **edited;**
5. **mutated.**

- The resulting phenotype is then analyzed.

Thus:

Candidate gene

↓

Alter gene function

↓

Analyze molecular/physiological changes

↓

Analyze phenotype

↓

Determine gene function

11. Comparison of Forward versus reverse genetics

Forward genetics	Reverse genetics
Starts with phenotype	Starts with gene
Phenotype → gene	Gene → phenotype
Identifies unknown genes	Tests candidate genes
Mutant screening	**Targeted gene manipulation**
Mapping/sequencing	CRISPR, RNAi, knockout, overexpression

Both approaches are now integrated with genome-wide technologies.

12. Mutagenesis as a functional-genomics tool

This section is particularly valuable for plant biotechnology.

Mutagenesis generates genetic variation that can reveal gene function.

Approaches include:

1. **spontaneous mutations;**
2. **chemical mutagenesis;**
3. **physical mutagenesis;**
4. **insertional mutagenesis;**
5. **transposon-based mutagenesis;**
6. **targeted genome editing**.

For example:

Mutagenesis

↓

Mutant population

↓

Phenotypic screening

↓

Interesting mutant

↓

Genome analysis

↓

Candidate mutation.

↓

Functional validation

This connects **classical mutation breeding with modern functional genomics**.

13. TILLING

TILLING — Targeting Induced Local Lesions IN Genomes — combines induced mutagenesis with molecular screening to **identify mutations in particular genes.**

General principle of **TILLING**:

Mutagenesis → mutant population → DNA screening/sequencing → mutation in target gene → phenotype

TILLING is particularly useful in crop species because it can generate ***allelic series*** containing different mutations within the same gene.

14. Genome-wide association studies — GWAS

GWAS identifies statistical associations between **genetic variants and phenotypic traits** across a population.

The fundamental idea is:

Genetic diversity + Phenotypic diversity

↓

Genome-wide markers

↓

Statistical association analysis

↓

Significant marker–trait associations

↓

Candidate genomic regions/genes

15. Basic GWAS workflow. A simple workflow:

Population selection

↓

Genotyping

↓

Phenotyping

↓

Quality control

↓

Population structure / relatedness analysis

↓

Genome-wide association analysis

↓

Significant SNPs

↓

Candidate genes

↓

Functional validation

16. Manhattan plot in GWAS

A **Manhattan plot** is a graphical representation of the results of a **genome-wide association study (GWAS)**.

It shows the statistical association between thousands or millions of genetic variants, usually **SNPs**, and a particular phenotype or trait.

How to read the plot

The **X-axis** represents the genomic position of SNPs arranged according to chromosomes:

Chr 1 → Chr 2 → Chr 3 → …

The **Y-axis** represents the statistical significance of the association, usually expressed as:

−log10(P)-\log_{10}(P)

Each **dot represents one genetic variant (SNP)**.

Therefore:

higher point → smaller P-value → stronger statistical evidence of association with the trait.

For example:

$P = 0.01 \rightarrow -\log_{10}(P) = \mathbf{2}$

$P = 0.0001 \rightarrow -\log_{10}(P) = \mathbf{4}$

$P = 10^{-8} \rightarrow -\log_{10}(P) = \mathbf{8}$

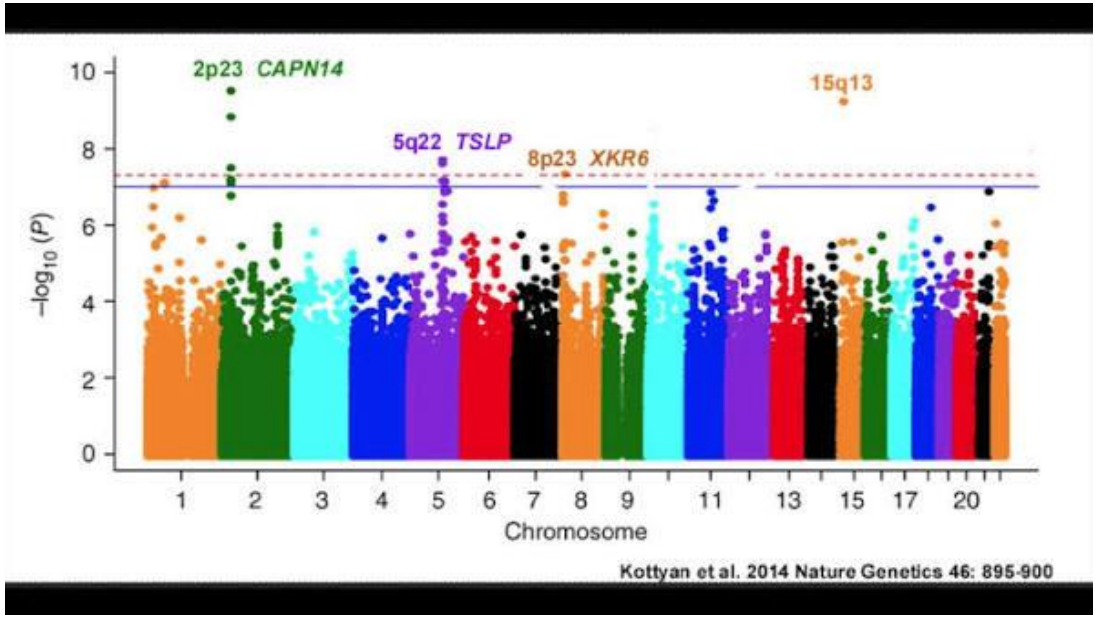
A horizontal line is commonly used to indicate the selected **genome-wide significance threshold**.

Why is it called a "Manhattan" plot?

When several genomic regions contain strongly associated variants, they form tall groups of points or **peaks**, visually resembling the skyline of Manhattan.

- **Manhattan plot** visualizes g**Manhattan plot** visualizes genome-wide associations between genetic variants and a phenotype.
- Each point represents a SNP; chromosomes/genomic positions are displayed along the X-axis, while the Y-axis shows $-\log_{10}(P)$.
- Peaks above the genome-wide significance threshold indicate genomic regions containing variants significantly associated with the investigated trait.
- Significant associations require subsequent identification and functional validation of candidate genes.

Manhattan plot



- Genome-wide associations between genetic variants and a phenotype. Each point represents a SNP;
- chromosomes/genomic positions are displayed along the X-axis, while the Y-axis shows $-\log_{10}(P)$.
- Peaks above the genome-wide significance threshold indicate genomic regions containing variants significantly associated with the investigated trait. Significant associations require subsequent identification and functional validation of candidate genes.

Manhattan plot in GWAS

What does a peak mean?

Suppose we perform GWAS for **drought tolerance in wheat**.

Phenotypic variation in drought tolerance
↓
Genome-wide SNP genotyping
↓
Association analysis
↓
Manhattan plot
↓
Significant peak on a chromosome
↓
Candidate genomic region
↓
Search for candidate genes
↓
Expression/functional analysis
↓
Experimental validation

17. GWAS and population structure

One important problem in GWAS is **false association**.

Individuals within populations may be genetically related or belong to different subpopulations.

Therefore, GWAS models commonly account for:

1. ***population structure;***
2. ***kinship;***
3. ***allele frequency;***
4. ***multiple testing.***

This illustrates an important general principle:

Large datasets require rigorous statistical analysis before biological conclusions can be made.

16. Manhattan plot

Manhattan plot, because they will encounter it frequently in genomics papers.

X-axis: genomic position/chromosomes

Y-axis: statistical significance, usually **$-\log_{10}(P)$**

High peaks indicate genomic regions containing variants strongly associated with the investigated trait.

For example:

SNP → significant association → genomic region → candidate gene

However:

A statistically significant SNP does not necessarily identify the causal gene itself.

Further investigation is required.

Major objectives of postgenomic research

Postgenomic technologies aim to:

1. **Identify gene functions.**
2. **Determine gene-expression patterns under different conditions.**
3. **Identify regulatory regions and epigenetic mechanisms.**
4. **Characterize the complete set of proteins produced by cells and tissues.**
5. **Determine post-translational modifications of proteins.**
6. **Identify protein–protein and other molecular interactions.**
7. **Characterize metabolites and metabolic pathways.**
8. **Discover molecular markers and biomarkers.**
9. **Establish relationships between molecular variation and phenotype.**
10. **Integrate multiple molecular datasets into biological networks.**
11. **Predict biological responses using computational approaches.**
12. **Functionally validate candidate genes and molecular pathways.**

7. Major “omics” technologies

Omics field	Object studied	Main question
Genomics	**DNA/genome**	**What genetic information is present?**
Epigenomics	**DNA/chromatin modifications**	**How is genome activity regulated?**
Transcriptomics	**RNA**	**Which genes are expressed?**
Proteomics	**Proteins**	**Which proteins are present and active?**
Metabolomics	**Metabolites**	**What metabolic processes are occurring?**
Interactomics	**Molecular interactions**	**How do biological molecules interact?**
Phenomics	**Phenotypic traits**	**How is genotype expressed as phenotype?**
Single-cell omics	**Individual cells**	**How do individual cells differ?**
Spatial omics	**Molecules within tissues**	**Where are molecular processes occurring?**

8. Functional genomics

Functional genomics investigates the functions and interactions of genes at a genome-wide level. Its objectives include:

1. identification of gene functions;
2. analysis of gene-expression patterns;
3. identification of regulatory elements;
4. investigation of gene interactions;
5. association of genes with phenotypic traits.

Modern functional genomics combines:

Genome sequencing + RNA-seq + gene perturbation + CRISPR/Cas + phenotyping + bioinformatics

- This represents an important bridge between genome information and biological function.

9. Transcriptomics

The transcriptome is the complete set of RNA molecules expressed in a cell, tissue, or organism under particular conditions.

Unlike the genome, the transcriptome is highly dynamic.

It changes according to:

1. **developmental stage;**
2. **tissue;**
3. **environmental conditions;**
4. **stress;**
5. **nutrition;**
6. **disease;**
7. **genotype.**

9. Transcriptomics. RNA sequencing

- RNA-seq is now one of the major technologies for transcriptome analysis.
- Typical workflow:
- RNA isolation → library preparation → sequencing → read processing/mapping → quantification → differential expression → pathway analysis
- For example:Control wheat vs drought-stressed whe
- at
- ↓
- RNA-seq
- ↓
- Identification of differentially expressed genes (DEGs)
- ↓
- Identification of drought-responsive pathways.

Conclusion

- Transcriptomics provides a comprehensive understanding of gene expression and its regulation under different biological and environmental conditions.
- RNA sequencing (RNA-seq) enables high-throughput identification and quantification of transcripts, including differentially expressed genes.
- RT-qPCR, microarrays, and RNA-seq are complementary approaches for investigating gene expression.
- Bioinformatics tools are essential for RNA-seq quality control, transcript quantification, differential expression analysis, and functional annotation.
- Transcriptomic profiling helps identify candidate genes and regulatory pathways involved in development, metabolism, and stress responses.
- Integration of transcriptomics with genomics, proteomics, and metabolomics strengthens functional interpretation and supports biotechnology applications.

Key message: Transcriptomics connects genomic information with gene activity and provides essential insights into molecular regulation and biological responses.

Control Questions

1. What is transcriptomics, and what are its main objectives?
2. What are the major types of RNA and their biological functions?
3. How is gene expression regulated at transcriptional and post-transcriptional levels?
4. What are the differences between RT-qPCR, microarrays, and RNA-seq?
5. What are the main stages of the RNA-seq workflow?
6. Why are biological replicates and RNA quality important in transcriptomic experiments?
7. What is differential gene expression analysis?
8. What are the functions of Gene Ontology and pathway enrichment analysis?
9. How can transcriptomics contribute to candidate gene identification in forward genetics?
10. What are the applications of transcriptomics in plant, medical, microbial, and industrial biotechnology?
11. How can RNA-seq be used to investigate drought-stress responses in wheat?
12. What are the main challenges and future directions of transcriptomic research?

Recommended Literature

- 1. Conesa, A., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biology, 17, 13. https://doi.org/10.1186/s13059-016-0881-8
- 2. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, 550. https://doi.org/10.1186/s13059-014-0550-8
- 3. Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. Nature Reviews Genetics, 20, 631–656. https://doi.org/10.1038/s41576-019-0150-2