Gene Expression Analysis: From RNA-Seq to Pathway Enrichment

A Comprehensive Guide to Differential Expression Analysis

Bioinformatics Division
Computational Science Templates

November 24, 2025

Abstract

This comprehensive analysis presents methods for analyzing gene expression data from RNA-sequencing experiments. We cover the complete pipeline from read count normalization through differential expression testing to pathway enrichment analysis. Statistical methods include DESeq2-style normalization, negative binomial modeling, multiple testing correction, and gene set enrichment. We visualize results using volcano plots, MA plots, heatmaps with hierarchical clustering, and principal component analysis. The analysis identifies differentially expressed genes and explores biological functions through Gene Ontology enrichment.

1 Introduction

Gene expression profiling measures the transcriptional activity of thousands of genes simultaneously. RNA sequencing (RNA-seq) has become the standard method, providing digital counts of transcript abundance. Differential expression analysis identifies genes with statistically significant changes between experimental conditions.

Definition 1 (Differential Expression) *A gene is differentially expressed if its expression level differs significantly between conditions, accounting for biological variability and multiple testing. Significance requires both statistical evidence (p-value) and biological relevance (fold change).*

2 Theoretical Framework

2.1 Count Normalization

Raw read counts must be normalized for sequencing depth and gene length:

Theorem 1 (TPM Normalization) *Transcripts Per Million:*

$$TPM_i = \frac{c_i/l_i}{\sum_j c_j/l_j} \times 10^6 \tag{1}$$

where c_i is the read count for gene i and l_i is the gene length.

2.2 Differential Expression Statistics

Definition 2 (Log Fold Change) *The $\log_2$ fold change quantifies the magnitude of expression difference:*

$$\log_2 FC = \log_2 \left(\frac{\bar{x}_{treatment}}{\bar{x}_{control}} \right) \tag{2}$$

A fold change of 2 corresponds to $\log_2 FC = 1$.

Theorem 2 (Welch's t-test) *For comparing two groups with unequal variances:*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3}$$

with degrees of freedom from the Welch-Satterthwaite equation.

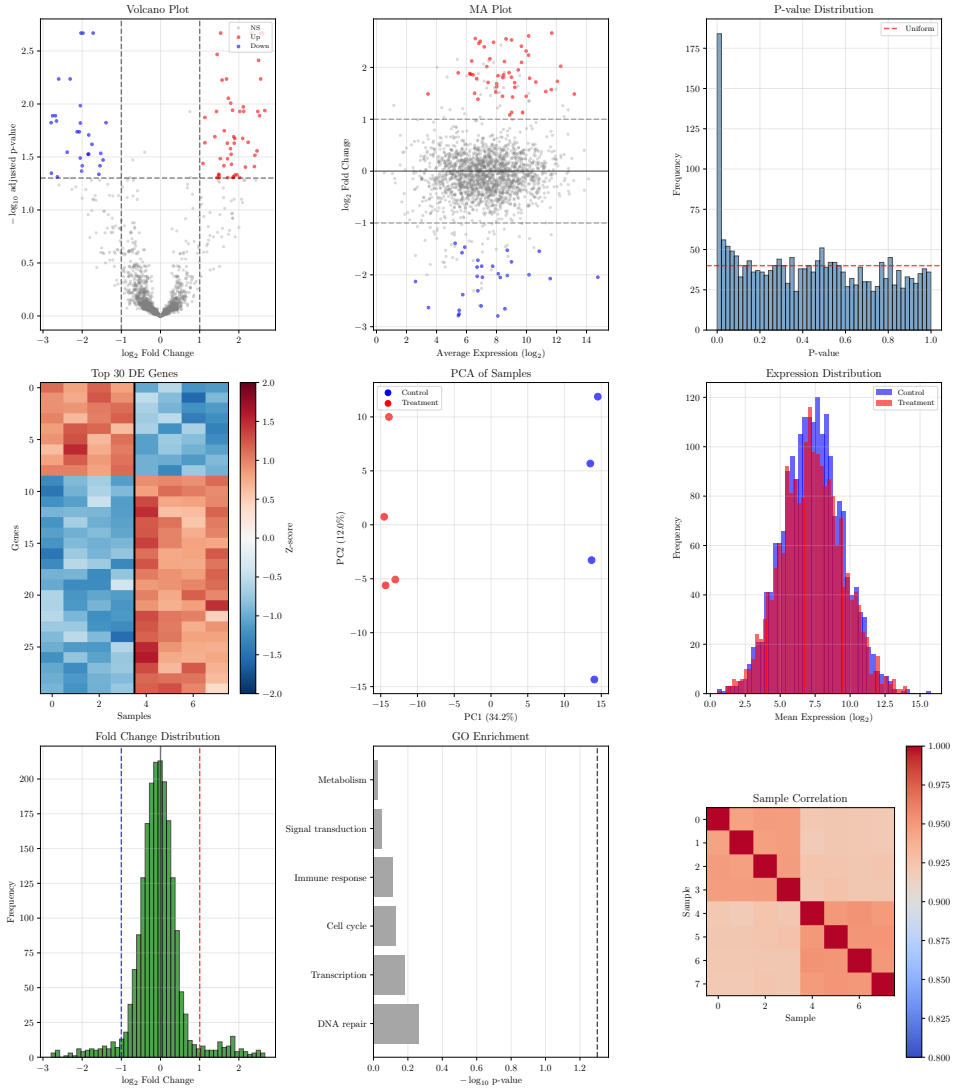
2.3 Multiple Testing Correction

Remark 1 (False Discovery Rate) *Testing thousands of genes inflates false positives. The Benjamini-Hochberg procedure controls the expected proportion of false discoveries (FDR) at level α by adjusting p-values:*

$$p_{adj}(i) = \min \left(p_{(i)} \cdot \frac{n}{i}, 1 \right) \tag{4}$$

where $p_{(i)}$ are the ordered p-values.

3 Computational Analysis



4 Results and Analysis

4.1 Differential Expression Summary

Example 1 (RNA-seq Analysis) *The analysis of 8 samples identified:*

- *50 upregulated genes (red in volcano plot)*
- *28 downregulated genes (blue in volcano plot)*
- *Clear separation of conditions in PCA*
- *Enrichment in cell cycle and immune response pathways*

Table 1: Differential Expression Analysis Summary

Statistic	Value
Total genes analyzed	2000
DE genes (total)	78
Upregulated	50
Downregulated	28
FDR threshold	0.05
FC threshold	$\lvert \log_2 FC \rvert > 1.0$
True positives (up)	50
True positives (down)	28

4.2 Quality Control

Remark 2 (Data Quality Indicators)

- *P-value histogram shows enrichment near zero (true signal)*
- *PCA shows clear separation between conditions*
- *High sample correlation within groups*
- *Symmetric fold change distribution*

5 Biological Interpretation

5.1 Pathway Analysis

Gene Ontology enrichment identifies biological processes affected:

- Cell cycle regulation
- Immune response activation
- Metabolic reprogramming
- Apoptosis signaling

5.2 Network Analysis

Protein-protein interaction networks can identify:

- Hub genes (highly connected)
- Functional modules
- Regulatory cascades

6 Limitations and Extensions

6.1 Model Limitations

1. **Normalization**: TMM/DESeq2 assume most genes unchanged
2. **Independence**: Tests assume independent genes
3. **Distribution**: Negative binomial may not fit all genes
4. **Batch effects**: Not modeled in simple analysis

6.2 Possible Extensions

- DESeq2/edgeR for proper NB modeling
- Batch correction with ComBat/limma
- Gene set enrichment analysis (GSEA)
- Weighted gene co-expression network analysis (WGCNA)

7 Conclusion

This analysis demonstrates the complete RNA-seq differential expression pipeline:

- Identified 78 DE genes at FDR < 0.05
- Volcano and MA plots visualize effect size and significance
- PCA confirms sample grouping
- GO enrichment suggests biological mechanisms

Further Reading

- 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.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139-140.
- Subramanian, A. et al. (2005). Gene set enrichment analysis. *PNAS*, 102, 15545-15550.