Week 2 – Differential Expression Analysis (GEO2R)

A. Introduction

Transcriptomic techniques have revolutionized many areas of biology. By analyzing the transcriptome which is the complete set of RNA transcripts expressed in a cell or organism, we can better understand how genes are regulated across tissues and under different conditions, such as disease or in response to therapeutic interventions. This approach provides valuable insight into the underlying molecular mechanisms that drive these biological processes (Lowe *et al.*, 2017).

Several methods are available to study transcriptomics. Currently, RNA sequencing (RNA-seq) is one of the most widely used state-of-the-art technologies. It can be performed at different resolutions, including bulk RNA-seq, single-cell RNA-seq, and even spatial transcriptomics. The standard workflow typically involves RNA extraction, library preparation, sequencing, and subsequent data analysis (Wang *et al.*, 2024). Another well-established method is microarray analysis, which has also made significant contributions to the field. Although both techniques aim to quantify gene expression, they differ in their underlying technologies. Microarrays detect fluorescently labeled complementary DNA (cDNA) molecules through hybridization to complementary probes on a solid surface, with gene expression levels represented by fluorescence intensity. In contrast, RNA-seq utilizes next-generation sequencing (NGS) to sequence cDNA molecules directly, providing a digital and sequential readout of transcript abundance (Raplee *et al.*, 2025).

In downstream analysis, the identification of differentially expressed genes (DEGs) is a central step for both microarray and RNA-seq datasets. DEG analysis is typically performed between two or more experimental groups to identify genes that show statistically significant changes in expression under different conditions. These changes are commonly represented as log2 fold change (FC). Following DEG identification, further analyses such as pathway enrichment analysis are conducted to interpret the biological significance of the results and to uncover transcriptome-related alterations in molecular pathways (Rosati *et al.*, 2024).

Objectives

- Exploration on GEO database and to overlook public transcriptomics database
- Identification of genes that are drastically/differentially change in each condition compared
- Enrich us as researcher about potential biological meaning lead by DEG we found

B. Dataset Overview

Title: Systematic review of genome-wide expression studies in multiple sclerosis

Accession ID: GSE21942

Multiple sclerosis (MS) is a chronic and unpredictable autoimmune disease in which the immune system attacks the myelin sheath, leading to disrupted communication between the brain and the rest of the body. The etiology of MS is complex and involves an interplay between genetic susceptibility and environmental factors. Although numerous studies have identified sets of differentially expressed genes (DEGs) associated with MS patients, concerns remain regarding the reliability and reproducibility of these findings. The authors suspected the presence of false-positive results in previously published reports. Therefore, they conducted a systematic review of existing gene expression studies, focusing on immune cells from MS patients compared with healthy controls, which included seven independent microarray datasets. In addition, the authors performed their own microarray analysis to identify DEGs in MS using genome-wide expression profiling of peripheral blood mononuclear cells (PBMCs), aiming to strengthen the validity and robustness of the findings.

C. Method

Platforms

GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples and Grouping

Multiple Sclerosis (MS)	:	15 healthy unrelated female (mean age 54.2)
Control	:	12 female patients fulfilling Poser’s criteria for clinically definite MS (mean age 71.6)

GEO2R Parameter

P-values	:	Benjamini & Hochberg
Log transformation	:	Yes
Limma precision weights (vooma)	:	No
Force normalization	:	Yes
Significance level cut-off	:	0.05

Reasoning

- P-values (Benjamini & Hochberg): Benjamini and Hochberg (BH) provide a p-value adjustment that is not too stringent and is suitable for genomic studies. FDR correction is necessary because thousands of statistical tests are performed for all genes in a dataset. For example, without correction, if we test 20,000 genes using a p-value threshold of 0.05, about 5% of the results (approximately 1,000 genes) may be false positives. The BH method applies this 5% threshold to the DEGs identified during analysis by controlling the expected proportion of false positives among the declared significant genes.

- Log transformation (yes): this feature is important to reduce skewness (too high expression data among each genes), those too high genes will mask the lowest one. This will cause abnormal p-value, misleading DEG, asymmetrical volcano plot. Manually log-transformed data, we need to check if the range is already 0-15 (log transformed) or 0-50.000 (not log transformed). In this data, maximum (2.10) and minimum (66347.0) expression value were too much different (refer to original series matrix file) that will cause skewness if not log-transformed.

- Limma precision weights (no): this feature adds weight considering stability/precision on gene expression, so the noisy genes does not affect analysis.

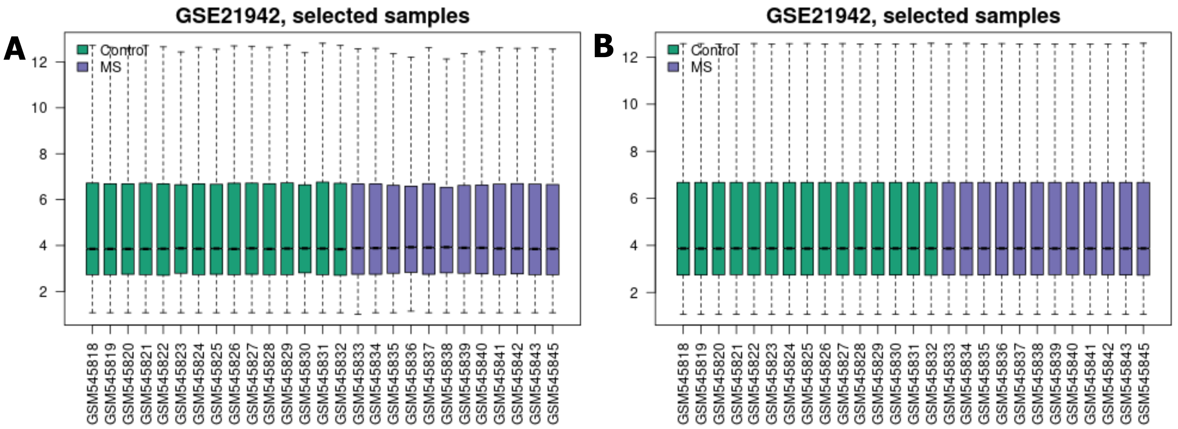


Figure 1. Distribution of GSE21942 selected samples. Boxplots show normalized expression values for each sample in the Control (green) and MS (purple) groups. (A) Expression distributions before normalization. (B) Expression distributions after normalization. The alignment of medians and interquartile ranges in (B) indicates successful normalization and improved comparability across samples.

- Force normalization (yes): This feature forces all samples to have the same expression distribution before analysis. It is important when samples appear different due to batch effects (e.g., differences in technology, processing time, or other technical factors), when the data have not been normalized, or when the expression ranges between samples vary substantially. In this analysis (Figure 1), some datasets showed slightly different expression ranges across samples. Therefore, force normalization was applied to standardize the distribution before analysis and reduce potential batch effects.

- Significance level cut-off (0.05): it is a threshold that will help us say whether our findings are due to coincidence or not. In this analysis, in our whole analysis the chance of us mistakenly determining DEG is 5%. The lower P-value is, the more stringent the DEG available and the more high-impact our findings.

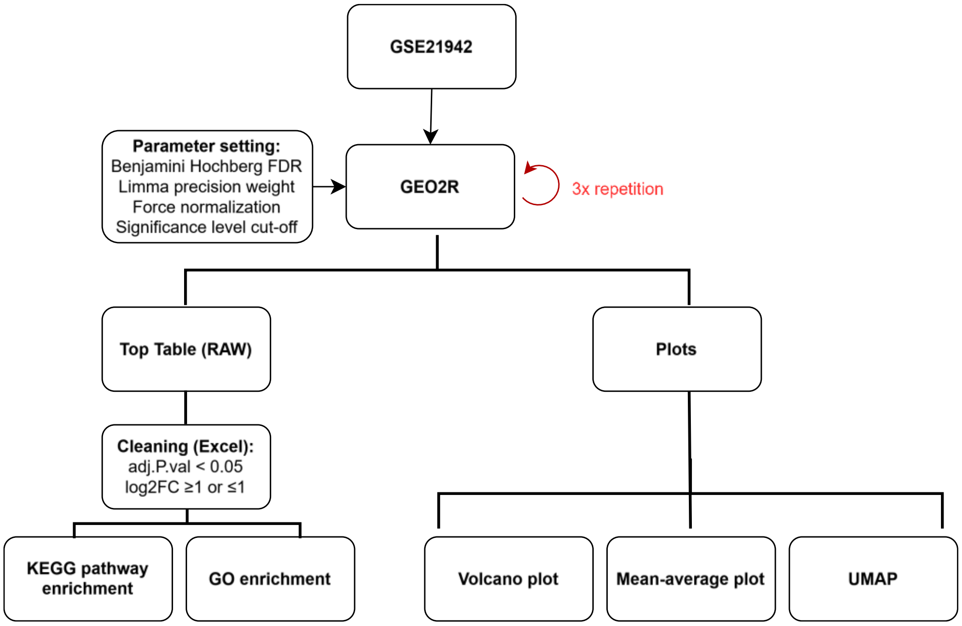
Correction Method for Multiple Testing

Benjamini & Hochberg (False Discovery Rate)

Significancy Criteria

- adj.P.val < 0.05
- log2FC ≥1 or ≤1

Replication Scheme



D. Results and Discussion

Differentially Expressed Genes (DEGs)

After applying the selection criteria (|log2FC| ≥ 1 and adjusted p-value < 0.05) and removing blank array results and duplicate entries, a total of 1,278 DEGs were identified in MS patients compared to controls. Among these, 488 genes were upregulated and 790 were downregulated. Clustering analysis using UMAP (Figure 2C) showed that most MS and control samples formed distinct clusters based on their gene expression profiles. However, three datasets displayed proximity to the control cluster. This may indicate biological variability or the presence of batch effects that require further investigation. The volcano plot (Figure 2A) illustrates gene distribution according to log2FC and statistical significance (−log10 p-value). Genes in the upper right quadrant (red) represent upregulated genes, whereas those in the upper left quadrant (blue) represent downregulated genes. The top 10 DEGs are listed in Table 1, and the complete DEG list is available in the GitHub repository.

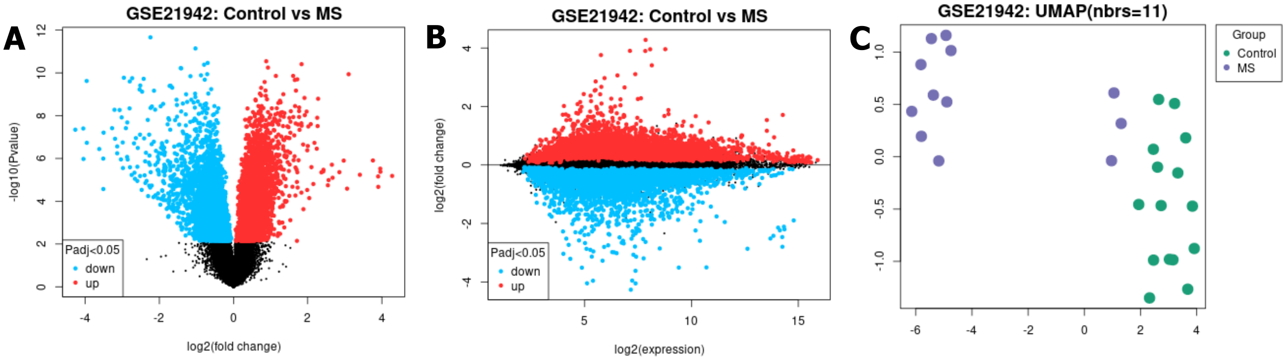


Figure 2. Differential expression and sample clustering analysis of GSE21942 (Control vs MS). Volcano plot showing differentially expressed genes between Control and MS samples. Red dots represent significantly upregulated genes, blue dots represent significantly downregulated genes, and black dots indicate non-significant genes (Padj < 0.05). (A) MA plot displaying the relationship between log$_2$ fold change and mean log$_2$ expression levels, highlighting significantly upregulated (red) and downregulated (blue) genes (B). UMAP projection (n_neighbors = 11) illustrating sample clustering based on global gene expression profiles. Control samples (green) and MS samples (purple) form distinct clusters, indicating clear transcriptomic differences between groups (C).

Table 1. Top differentially expressed genes between Control and MS samples in GSE21942.

Gene symbol	**Log FC**	**adj.P.value**
Upregulated		
HBD	-427	0.00121
HBG2	-405	0.00108
ALAS2	-405	0.00913
LTF	-396	0.000059
BOD1L1	-351	0.00892
MALAT1	-350	0.00152
HBM	-328	0.00105
SLC25A37	-321	0.000349
CLC	-320	0.00233
SAMSN1	-314	0.00388
Downregulated		
EIF5A	341	0.028
HINT3	310	0.000052
ALYREF	228	0.000196
LILRA5	227	0.000907
ARF6	226	0.000551
SNX20	218	0.00305
GM2A	210	0.0163
CASP2	208	0.000814
FCAR	202	0.000912
MOB3A	200	0.00379

Biological Interpretation

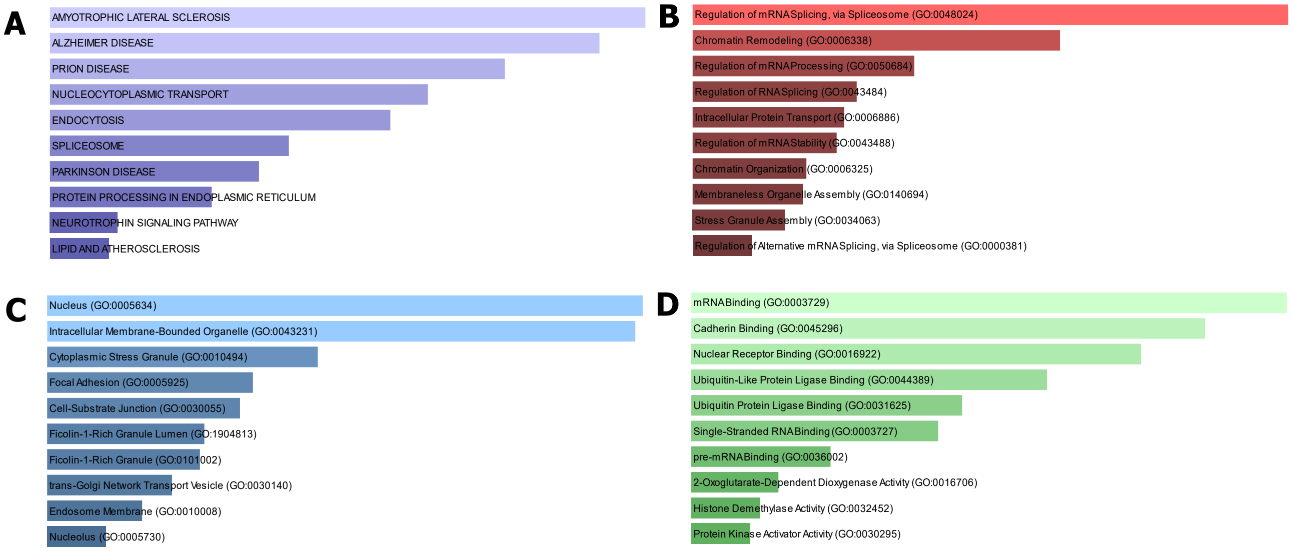


Figure 3. Functional enrichment analysis of differentially expressed genes (DEGs) in GSE21942. Bar plots show the top significantly enriched based on Gene Ontology (GO) and pathway analysis (A). Enriched categories include Biological Process (BP) (B), Cellular Component (CC) (C), and Molecular Function (MF) (D).

Gene Ontology (GO) enrichment analysis was performed to identify the biological function categories represented by the identified DEGs. In addition, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was conducted to reveal gene–gene interactions and their involvement in molecular pathways using the KEGG database. In this study, the Enrichr comprehensive database was used for both analyses. The results showed that the DEGs were mainly enriched in pathways related to Amyotrophic Lateral Sclerosis (ALS) and Alzheimer's disease. ALS is known to share certain pathway similarities with MS. GO enrichment analysis further indicated that these DEGs were involved in:

Biological Process

- Regulation of mRNA splicing, via Spliceosome
- Chromatin remodeling
- Regulation of mRNA processing

Cellular Component

- Nucleus
- Intracellular Membrane-Bounded Organelle
- Cytoplasmic Stress Granule

Molecular Function

- mRNA binding
- Cadherin binding
- Nuclear receptor binding

E. Conclusion

In conclusion, the present analysis demonstrates that the multiple sclerosis condition is associated with significant transcriptional alterations, characterized by both upregulated and downregulated genes compared to the control group. The identified differentially expressed genes (DEGs) suggest that multiple sclerosis induces structured molecular changes rather than random transcriptional variation.

Functional enrichment analysis further revealed that these DEGs are significantly associated with pathways related to amyotrophic lateral sclerosis, indicating potential shared molecular mechanisms or overlapping neurodegenerative processes between the two conditions. Moreover, Gene Ontology analysis highlighted a strong enrichment in biological processes related to mRNA splicing regulation, particularly within the intranuclear compartment. This finding suggests that dysregulation of RNA processing and post-transcriptional modification may play a critical role in the pathophysiology of multiple sclerosis.

F. Reference

Kemppinen, A. K., Kaprio, J., Palotie, A., & Saarela, J. (2011). Systematic review of genome-wide expression studies in multiple sclerosis. BMJ Open, 1(1), e000053–e000053. https://doi.org/10.1136/bmjopen-2011-000053

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. PLOS Computational Biology, 13(5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Raplee, I. D., Borkar, S. A., Yin, L., Venturi, G. M., Shen, J., Chang, K.-F., Upasana Nepal, Sleasman, J. W., & Goodenow, M. M. (2025). The Role of Microarray in Modern Sequencing: Statistical Approach Matters in a Comparison Between Microarray and RNA-Seq. BioTech, 14(3), 55–55. https://doi.org/10.3390/biotech14030055

Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E., & Giordano, A. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A Review. Computational and Structural Biotechnology Journal, 23. https://doi.org/10.1016/j.csbj.2024.02.018

Wang, H., Xu, Y., Zhang, Z., Zhang, G., Tan, C., & Ye, L. (2024). Development and application of transcriptomics technologies in plant science. Crop Design, 3(2), 100057. https://doi.org/10.1016/j.cropd.2024.100057