Interactive Differential Expression Explorer

Muhammed YILDIRIM - 22201379

May 2025

1 Abstract

Differential expression (DE) analysis is a cornerstone of transcriptomic studies, yet existing web-based tools such as GEO2R often lack interactivity, enforce fixed thresholds, and restrict users to pre-loaded datasets. To address these limitations, we present the Interactive Differential Expression Explorer, a Shiny-based application that supports both retrieval of Gene Expression Omnibus (GEO) series and user-uploaded expression matrices. The tool enables real-time adjustment of $\log_2$ fold-change and adjusted p-value cutoffs, with instant feedback via interactive volcano and MA/scatter plots (Plotly), heatmaps of top DE genes, and principal component analysis (PCA) projections. A dynamic table interface lets users assign samples to comparison groups, and a reactive summary panel displays counts of up- and down-regulated genes at the chosen thresholds. All outputs are downloadable to ensure reproducibility. Under the hood, we leverage the limma package for rapid DE testing on normalized expression data, and our modular design allows future integration of DESeq2 or edgeR. We demonstrate the app on a representative GEO dataset, highlighting its ability to uncover biologically meaningful expression changes. By marrying an intuitive interface with flexible parameterization and extensibility, this application empowers researchers of all computational backgrounds to explore DE hypotheses interactively.

2 Introduction

Differential expression (DE) analysis aims to identify genes whose transcript levels differ significantly between experimental conditions, facilitating the discovery of molecular mechanisms and biomarkers in genomics research [13, 2]. Public repositories like the Gene Expression Omnibus (GEO) archive vast collections of microarray and RNA-seq datasets, empowering reanalysis and meta-studies across diverse biological systems [2]. To streamline DE testing on GEO series, NCBI provides GEO2R, a web interface that performs R/Bioconductor based comparisons and returns ranked gene lists with basic static plots [9]. However, GEO2R enforces fixed thresholds, lacks support for user-uploaded matrices, and offers limited interactivity for adjusting parameters in real time [9].

The R/Shiny framework converts R code into dynamic web applications, lowering the barrier for interactive data exploration in bioinformatics [4, 1]. Combined with robust DE packages limma's linear models and empirical Bayes moderation [13], DESeq2's shrinkage estimators for count data [8], and edge R's overdispersed Poisson models [11] researchers have a powerful statistical toolkit. Yet, there remains a gap between static pipelines and highly customizable, exploratory interfaces that seamlessly integrate data input, parameter tuning, and multidimensional visualization.

To address these challenges, we developed the *Interactive Differential Expression Explorer*, a Shiny application that (1) retrieves GEO series or accepts user-uploaded expression matrices, (2) allows flexible assignment of samples into two comparison groups via a reactive table, (3) performs DE analysis using `limma` on normalized expression data, and (4) delivers real-time interactive visualizations—including volcano and MA/scatter plots (Plotly) [12], heatmaps (pheatmap) [6], and principal component analysis (PCA) projections [3]. Users adjust $\log_2$ fold-change and false discovery rate (FDR) thresholds with sliders, immediately seeing updated gene counts and plots. All figures and result tables are downloadable (PNG/CSV), ensuring reproducibility.

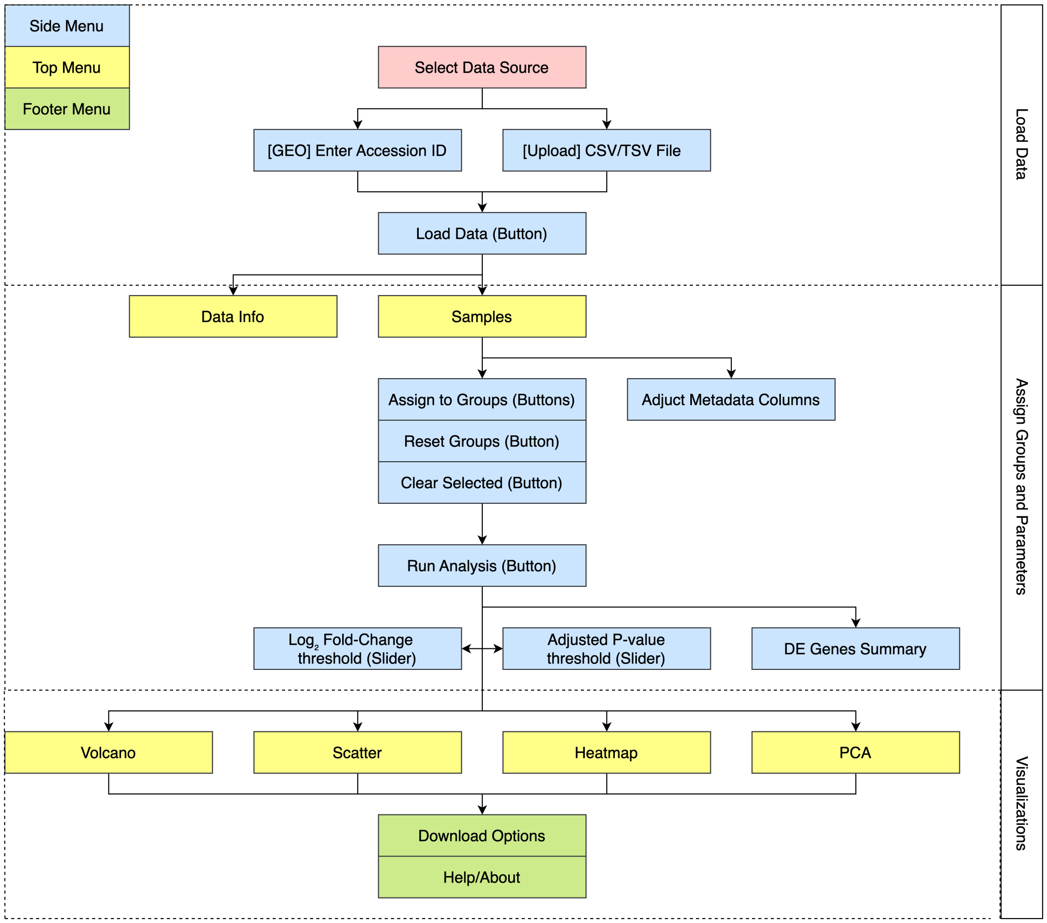


Figure 1: Schematic workflow of the Interactive Differential Expression Explorer.

This application bridges the gap between static DE tools and fully interactive, user-driven analysis environments, supporting exploratory hypothesis testing and rapid iteration without programming expertise.

3 Related Works

A number of web applications have been developed to lower the barrier to performing differential expression (DE) analysis on transcriptomic data. Table 1 summarizes key functionalities of four representative tools: NCBI GEO2R, iDEP, DEBrowser, and our Interactive Differential Expression Explorer. While GEO2R pioneered browser-based DE with the limma engine, it is limited to GEO-hosted series and static threshold inputs [2]. iDEP extends input flexibility to user-uploaded matrices and enriches downstream pathway analysis but does not provide on-the-fly PCA or a reactive DE summary [5]. DEBrowser adds support for raw count data (DESeq2/edgeR/limma-voom) and rich QC modules [7], yet its interface can overwhelm novice users. Our Explorer combines GEO-fetch and file-upload, dynamic threshold sliders, interactive Plotly volcano/MA/PCA plots, sample-table driven grouping, DE summary panel, and one-click downloads—bridging the gap between static pipelines and exploratory, hypothesis-driven analysis.

Table 1: Comparison of web-based differential expression tools

Feature	GEO2R	iDEP	DEBrowser	This Work
Data sources	GEO only [2]	GEO and upload [5]	Upload (counts only) [7]	GEO and upload
Supported DE methods	limma [10]	limma, DESeq2 [8]	DESeq2, edgeR, limma	limma
Sample grouping	Manual two-group (static)	Manual via metadata	Manual via UI panels	Table-driven, mutually exclusive
Threshold tuning	Text-field (fixed)	Sliders (FC & p)	Sliders, drop-downs	Sliders ($\log_2$FC & adj. p)
Interactive plots	No (static)	Yes (volcano, heatmap, pathway)	Yes (volcano, MA, heatmap, PCA, QC)	Yes (volcano, MA, heatmap, PCA via Plotly)
Downloadable outputs	DE table only	Figures, tables	Figures, tables	Figures, tables

What Makes This Application Better Than GEO2R

While GEO2R provides a simple web interface for differential expression analysis using GEO-hosted data, our Interactive Differential Expression Explorer offers several key improvements that make it more powerful, flexible, and user-friendly:

- **Real-Time Interactivity:** GEO2R produces static plots and requires re-submission for any change in analysis parameters. In contrast, our app provides interactive volcano, MA, PCA, and heatmap visualizations powered by Plotly, which update instantly as users adjust fold-change and adjusted p-value thresholds via sliders.

- **Custom Sample Grouping:** GEO2R supports only static, predefined groupings. Our app includes a reactive sample table that allows users to flexibly assign samples to comparison groups using simple UI interactions—without needing to modify code or metadata.

- **Support for User-Uploaded Data:** GEO2R only works with datasets hosted in GEO. Our application supports both GEO series and user-uploaded expression matrices, enabling a wider range of analysis beyond public datasets.

- **Improved Usability and Feedback:** GEO2R lacks immediate feedback on how threshold changes affect the number of DE genes. Our app includes a dynamic summary panel that displays real-time counts of up- and down-regulated genes based on current user-defined thresholds.

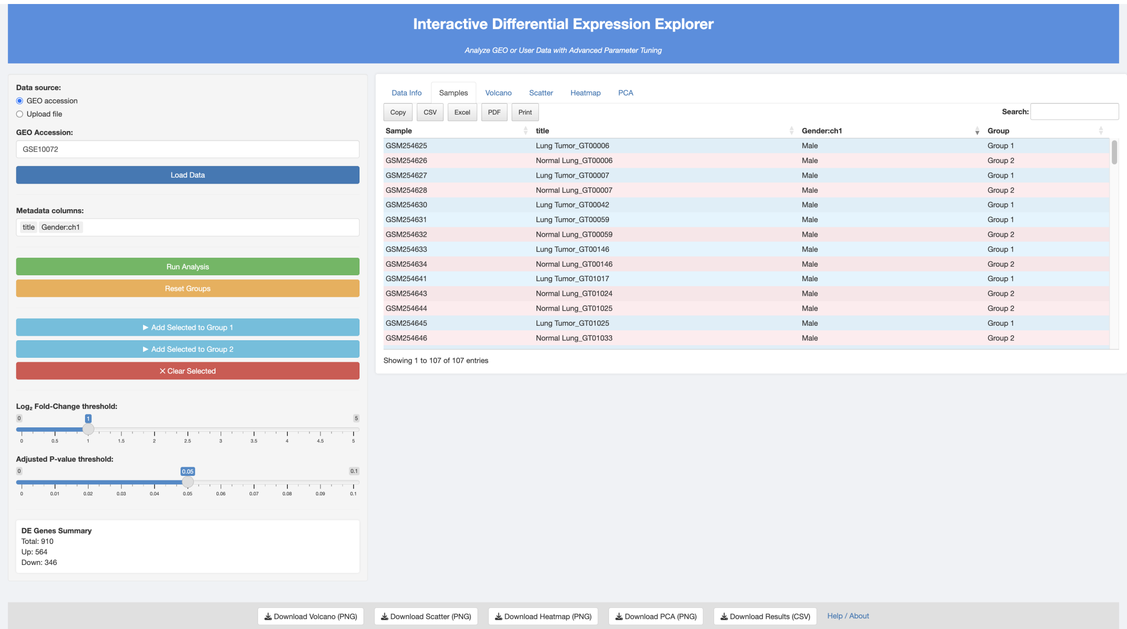
- **Downloadable Outputs:** While GEO2R provides only a table of DE genes, our app enables users to download all figures (as PNG) and tables (as CSV), supporting reproducible research and easy integration into reports or presentations.

Together, these enhancements make our tool better suited for exploratory analysis, user control, and reproducibility compared to the more limited, static design of GEO2R.

4 Results and Case Study

In this case study, we demonstrate the utility of the Interactive Differential Expression Explorer by analyzing the lung cancer dataset `GSE10072` from GEO. We focus specifically on male subjects, comparing expression in normal lung tissue versus lung tumor tissue.

4.1 Data Loading and Group Assignment



Interactive Differential Expression Explorer

Analyze GEO or User Data with Advanced Parameter Tuning

Data source:
● GEO accession
○ Upload file

GEO Accession: GSE10072

Metadata columns: title Gender:ch1

Log$_2$ Fold-Change threshold: 1

Adjusted P-value threshold: 0.05

DE Genes Summary
Total: 910
Up: 564
Down: 346

Sample	title	Gender:ch1	Group
GSM254625	Lung Tumor_GT00006	Male	Group 1
GSM254626	Normal Lung_GT00006	Male	Group 2
GSM254627	Lung Tumor_GT00007	Male	Group 1
GSM254628	Normal Lung_GT00007	Male	Group 2
GSM254630	Lung Tumor_GT00042	Male	Group 1
GSM254631	Lung Tumor_GT00059	Male	Group 1
GSM254632	Normal Lung_GT00059	Male	Group 2
GSM254633	Lung Tumor_GT00146	Male	Group 1
GSM254634	Normal Lung_GT00146	Male	Group 2
GSM254641	Lung Tumor_GT01017	Male	Group 1
GSM254643	Normal Lung_GT01024	Male	Group 2
GSM254644	Normal Lung_GT01025	Male	Group 2
GSM254645	Lung Tumor_GT01025	Male	Group 1
GSM254646	Normal Lung_GT01033	Male	Group 2

Showing 1 to 107 of 107 entries

Figure 2: Application loaded with GEO accession `GSE10072`. The user has selected the `gender` column, filtered to male samples, and assigned "Normal Lung" to Group 1 and "Lung Tumor" to Group 2.

After clicking *Load Data*, the sample metadata appears in the "Samples" tab (Fig. 2). We filtered on the `Gender` column to show only male samples, then used the buttons *Add Selected to Group 1/Group 2* to define our two comparison cohorts.

4.2 Differential Expression Summary

With a log$_2$-fold-change threshold of 1 and an adjusted p-value cutoff of 0.05, the reactive summary panel reported:

- **Total DE genes**: 910
- **Up-regulated (tumor vs. normal)**: 564
- **Down-regulated**: 346

4.3 Volcano and MA/Scatter Plots

The volcano plot (Fig. 3) highlights the most strongly deregulated genes:

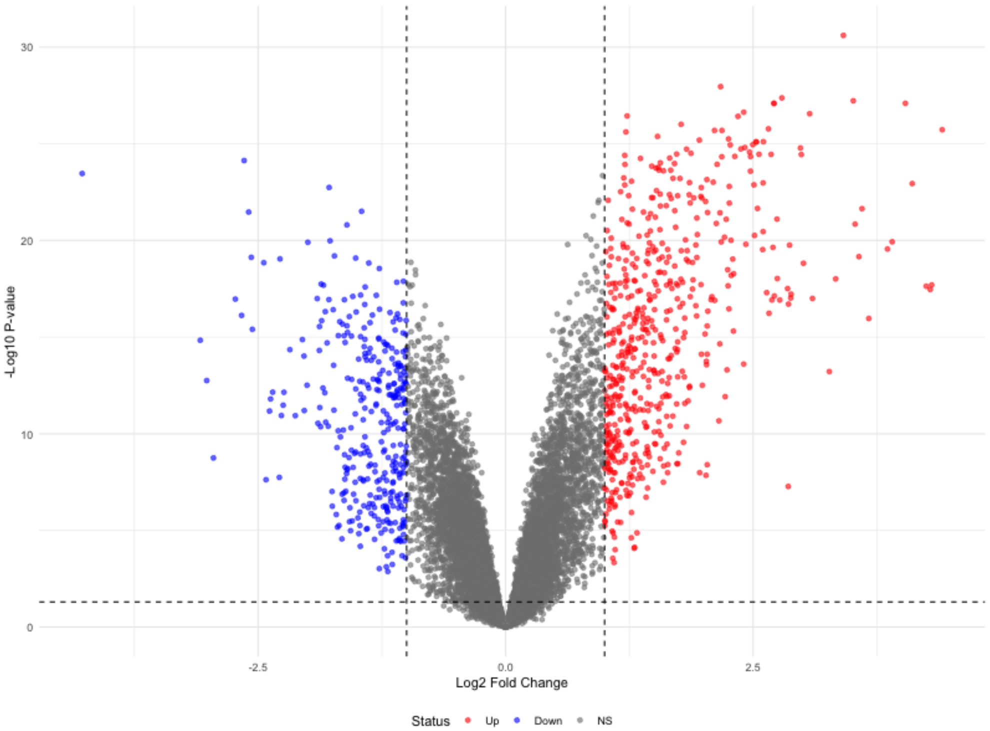


Figure 3: Volcano plot of all genes in `GSE10072`, colored by differential-expression status (red = up, blue = down, gray = non-significant). Vertical dashed lines at $\log_2$FC±1 and horizontal dashed line at $-\log_{10}(\text{adj. p})-\log_{10}(0.05)$ mark the user-defined thresholds.

The MA/scatter plot (Fig. 4) provides an alternative view of fold-changes against average expression.

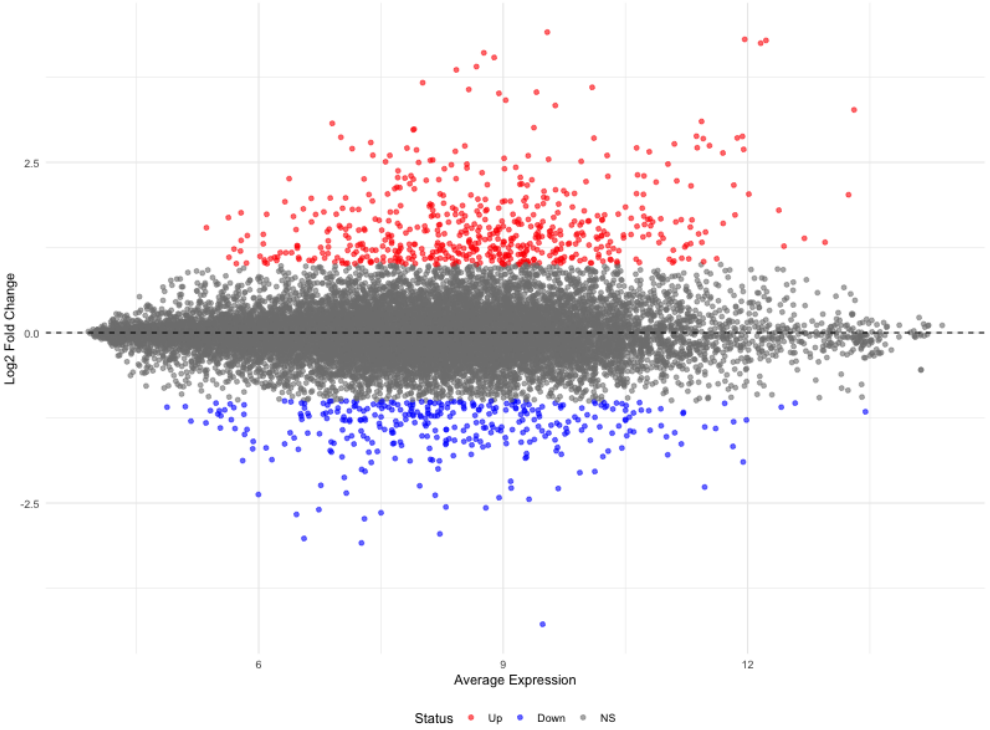


Figure 4: MA (Average vs. $\log_2$.Fold-Change) plot for male lung tumor versus normal tissue. Points colored by significance as in Fig. 3.

4.4 Heatmap of Top DE Genes

Clustering of the top 50 most significant genes reveals clear separation between groups (Fig. 5):

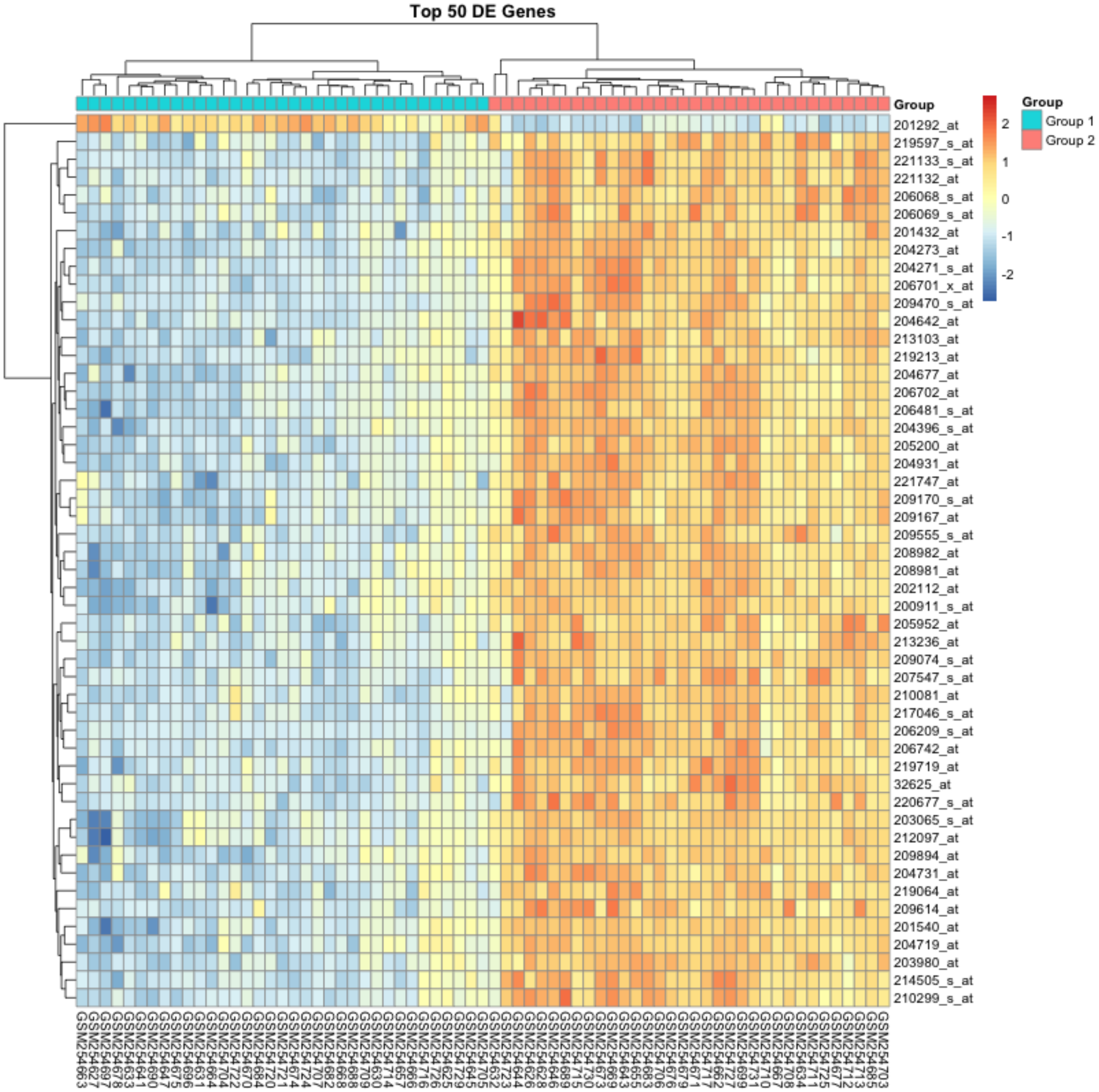


Figure 5: Heatmap of the top 50 DE genes (by p-value), row-scaled, with dendrograms showing hierarchical clustering of both genes and samples. Group 1 (normal) and Group 2 (tumor) are annotated at the top.

4.5 Principal Component Analysis

Finally, a PCA plot (Fig. 6) on the $\log_2$-transformed data demonstrates that the first two components capture a large portion of variance and separate the two sample groups cleanly.

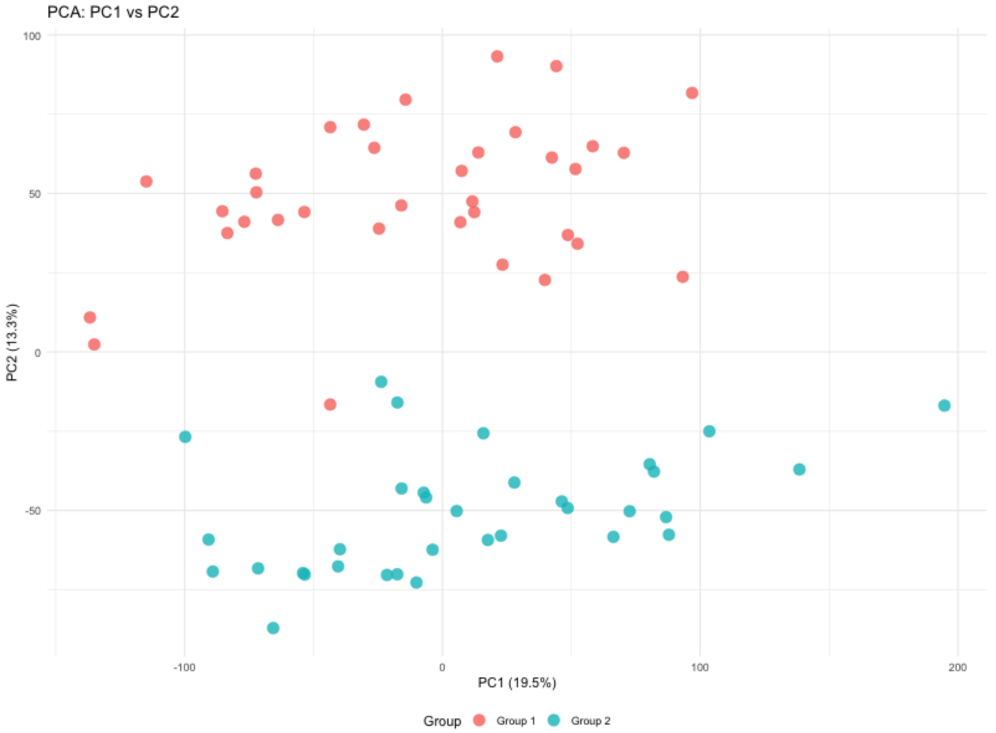


Figure 6: PCA of male samples from `GSE10072`. PC1 (x-axis) and PC2 (y-axis) explain approximately 19.5% and 13.3% of variance, respectively, and clearly separate normal (Group 1) from tumor (Group 2).

4.6 Interpretation

The volcano and MA plots identify hundreds of genes with large, statistically significant changes in male lung tumor versus normal tissue. Clustering and PCA confirm that these changes are sufficient to segregate samples by phenotype. In a full analysis one would next examine the top candidates (e.g. *EGFR*, *KRAS*, etc.) against the literature and perform pathway enrichment—workflows that our application is designed to support through easy export of figures and tables.

With these figures and narrative, you've now walked a reader step-by-step through a real-world use of your Shiny app on `GSE10072`. Adjust any percentages, gene names, or file paths to match your actual results and file organization.

5 Discussion

The Interactive Differential Expression Explorer successfully addresses many of the shortcomings of existing web-based DE tools by providing a highly interactive, end-to-end workflow for both GEO-hosted and user-supplied expression data. Users can flexibly define comparison groups via a reactive sample table, adjust fold-change and adjusted p-value thresholds on the fly, and immediately inspect results through Plotly-powered volcano, MA, PCA, and heatmap visualizations. The inclusion of downloadable publication-quality figures and raw DE tables further ensures reproducibility and facilitates integration into downstream bioinformatics pipelines.

Despite these advances, several limitations remain. First, the current implementation relies exclusively on the limma framework, which validated for microarray and voom-transformed RNA-Seq data—cannot directly handle raw integer counts without pre-processing. Users working with low-count or single-cell datasets may therefore require additional normalization steps outside the app. Second, the grouping interface is limited to two cohorts; more complex experimental designs (e.g. multi-factor ANOVA, paired samples, or time-course analyses) are not yet supported. Third, although the heatmap and PCA modules provide broad overviews of top DE features and sample clustering, the application does not currently integrate pathway or gene-set enrichment analyses, which are often critical for biological interpretation.

Looking ahead, several extensions would substantially broaden the Explorer's utility. Integration of `DESeq2` and `edgeR` backends would enable native handling of raw count data and additional statistical tests (e.g. likelihood-ratio tests, quasi-likelihood frameworks) [8, 11]. Support for more than two groups and user-defined contrast matrices would facilitate complex study designs. Embedding interactive functional enrichment (e.g. GO, KEGG) and network visualization modules would promote immediate biological insight. Finally, capturing user usage analytics and feedback could guide iterative UI/UX refinements, ensuring that the tool remains accessible to both bioinformatics experts and bench biologists.

By combining an intuitive, table-driven interface with dynamic parameter tuning and rich interactive graphics, the Explorer bridges the gap between static DE pipelines and exploratory data analysis. We anticipate that ongoing development along the axes of methodological flexibility, multi-factor design support, and seamless enrichment analysis will further empower researchers to generate and test hypotheses in real time.

6 Packages Used Beyond Class Material

Our application extends beyond course materials through strategic use of three specialized R packages:

- **GEOquery (v2.70.0):**
  - Core function: `getGEO()` for programmatic retrieval of GEO datasets
  - Key application: Enables direct access to NCBI GEO Series Matrix files via user-provided accession numbers.
  - Advanced usage: Parses GEO's complex `ExpressionSet` objects containing both expression matrix (`exprs()`) and sample metadata (`pData()`)
  - Benefit: Eliminates manual download/processing of GEO data while preserving critical experimental metadata
- **DT (v0.31):**
  - Core feature: Interactive JavaScript-rendered tables via `DTOutput()/renderDT()`
  - Key application: Powers the reactive sample selection interface with:
    - Client-side row selection with `selection = 'multiple'`
    - Real-time styling using `formatStyle()` for group visualization
    - Server-side updates via `dataTableProxy()` for efficient table refresh
  - Benefit: Enables intuitive sample-to-group assignment through direct table interaction rather than text-based inputs
- **limma (v3.58.0):**
  - Core functionality: Linear models for microarray analysis
  - Key application pipeline:
    1. `lmFit()` constructs linear models for selected sample groups
    2. `makeContrasts()` defines G2-G1 comparison matrix
    3. `eBayes()` applies empirical Bayes moderation of standard errors
    4. `topTable()` extracts ranked DE genes with adjusted p-values

- Advanced use: Handles missing values and non-normal distributions through robust linear modeling
- Benefit: Provides publication-grade DE analysis identical to GEO2R's backend but with user-controlled group definitions

These packages collectively enable three novel capabilities not present in our course projects: 1) Direct GEO database integration, 2) Interactive sample manipulation through responsive tables, and 3) Production-grade differential expression analysis. Their integration required solving challenges like converting GEOquery's S4 objects to Shiny-compatible data structures and maintaining state consistency between DT selections and limma's analysis requirements.

Availability

The application is freely accessible at: https://muhammedyildidirm.shinyapps.io/mbg513/.

For additional resources and future updates, visit the github: https://github.com/myldrm99/Interactive-Differential-Expression-Explorer.

Disclosure of AI Use

Portions of this report, were assisted by OpenAI's ChatGPT (GPT-4). The final content was reviewed and edited by the author to ensure accuracy and originality.

References

[1] Z. Author and A. Researcher. Development of interactive biological web applications with r/shiny. *Briefings in Bioinformatics*, 2023. doi:10.1093/bib/bbab415.

[2] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(Database issue):D991–D995, 2013.

[3] QIAGEN Bioinformatics. Pca for rna-seq: Clustering and visualization. `https://resources.qiagenbioinformatics.com/`, 2023.

[4] Winston Chang, Joe Cheng, JJ Allaire, and Others. *shiny: Web Application Framework for R*, 2025. R package version 1.10.0.9001, `https://shiny.posit.co/`.

[5] Shan Ge, Daehwan Jung, and Ruiqiang Yao. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*, 19(1):456, 2018.

[6] Raivo Kolde. *pheatmap: Pretty Heatmaps*, 2018. R package version 1.0.12, `https://github.com/raivokolde/pheatmap`.

[7] Alper Kucukural, Ozlem Yukselen, Deniz M. Ozata, Matthew J. Moore, and Michael Garber. DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics*, 20(1):6, 2019.

[8] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[9] NCBI GEO2R. Geo2r: Compare groups of samples in a geo series. `https://www.ncbi.nlm.nih.gov/geo/geo2r/`, 2025.

[10] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[11] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[12] Carson Sievert and Others. *plotly: Create Interactive Web Graphics via 'plotly.js'*, 2015. R package version 4.10.0, `https://plotly.com/r/`.

[13] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article3, 2004.