An Interactive Shiny App for Biomarker Identification: Cell-Type-Specific HVGs & Functional Analysis in Single-cell RNA-Sequencing

Omid (Mohammad Reza) Mohajeri
Independent Bioinformatics Researcher | Teesside University (Research Collaboration)

Introduction

- Single-cell RNA-seq (scRNA-seq) reveals cell-level transcriptomic variation critical for understanding disease microenvironments.
- This application is developed and tested using kidney cancer scRNA-seq data (10x Genomics, ICB-treated); adaptable to other diseases with similar input structure (matrix, genes, barcodes, metadata).
- Identifying HVGs within subpopulations (e.g., non-responder, responder, tumor or immune cells) highlights genes involved in resistance mechanisms, immune activity, or treatment efficacy.
- Combined with other gene lists (e.g., from DEA), HVGs help isolate informative genes for building predictive ML/DL models.
- The tool simplifies metadata-driven analysis, enabling rapid insights into cell-type-specific biology relevant to drug discovery.

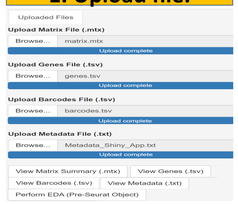
Objective

- To develop an interactive platform for dynamic identification of highly variable genes (HVGs) within selected cell types, enabling functional enrichment analysis (GSEA and ORA for GO terms and KEGG pathways) and biomarker discovery.
- To provide an accessible, real-time tool that supports both experimental and computational researchers through flexible input, intuitive design, export-ready outputs, and dynamic metadata integration. The app extracts metadata columns and their unique values (e.g., *ICB_Response*: ICB_PR, ICB_SD, ICB_PD), allowing users to subset cells of interest, identify HVGs, perform functional analyses (GO and KEGG via GSEA and ORA), iteratively compare gene sets and enriched pathways.

Methods and Materials

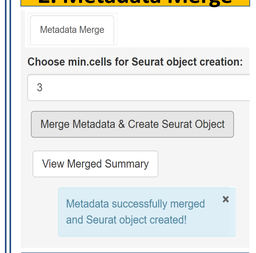
This Shiny app, built in R (**2,000+ code lines**, runs via RStudio), was developed and tested on **10x Genomics kidney cancer data** (**ICB therapy**). It covers six steps: from data upload to HVG identification and functional enrichment (**GO/KEGG** via **GSEA** & **ORA**). While tested on kidney cancer, it's **adaptable to other datasets and diseases**. Enables **cell-type-specific HVG detection** and **pathway analysis**.

1. Upload file.



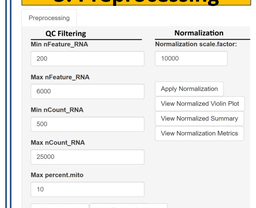
- Upload files:
  - matrix (.mtx),
  - genes (.tsv),
  - barcodes (.tsv),
  - metadata (.txt)
- View files

2. Metadata Merge



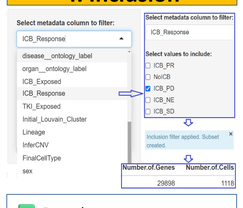
- Set min. cells per gene
- Create Seurat object
- Merge matrix & metadata
- Remove unmatched cells

3. Preprocessing



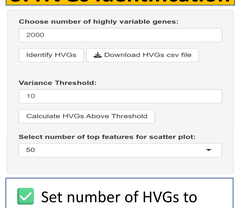
- Set filters
- Remove low-quality cells
- Normalize counts
- View summaries & plots

4. Inclusion



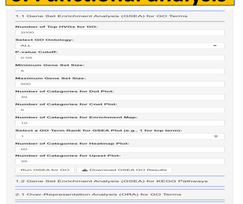
- Dropdown menu appears
- Lists all metadata columns
- User selects a column name
- Selects specific values
- Other cells: Excluded
- Analysis: only with selected cells

5. HVGs identification



- Set number of HVGs to identify
- Visualize & download HVGs (csv file)
- Show top features (e.g. top 50) (plot & table)
- Set variance threshold to count above

6. Functional analysis



- GSEA & ORA on GO/KEGG
- Set cutoffs & gene set size
- Download results & plots
- Visualize: dot, cnet, E-map, Heatmap, upset, bar, cloud, bubble, chord

Results

Data Import to HVG Identification (Step 1-5)

After uploading and integrating the scRNA-seq data, and metadata, quality control filtering and normalization reduced noise and improved data consistency. Interactive cell selection based on metadata allowed focused analysis on specific biological groups. From the filtered subset, highly variable genes were identified, representing key drivers of transcriptional variability across the selected cells and forming the basis for downstream functional analysis.

Cell Quality & Metadata overview (Step 1)

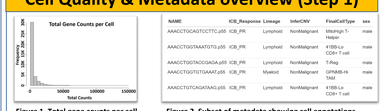


Figure 1. Total gene counts per cell

Figure 2. Subset of metadata showing cell annotations.

Merging metadata & matrix (Step 2)

Figure 3. Summary after merging metadata and matrix. Cells without metadata were removed; 34,325 matched cells retained.

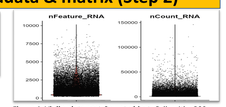


Figure 4. Violin plots post-Seurat object. Cells with <200 or >5,000 genes flagged for filtering (empty droplets/doublets).

QC filtering (Step 3)

Figure 5. Summary after QC filtering. 28,901 high-quality cells retained; gene count unchanged.

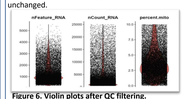


Figure 6. Violin plots after QC filtering. Outliers removed; distributions show cleaner profiles across key metrics.

Normalization (Step 3)

Figure 7. Summary after normalization. Gene and cell counts unchanged post-normalization.

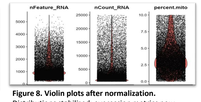


Figure 8. Violin plots after normalization. Distributions stabilized; expression metrics now comparable across cells.

Highly Variable Genes (Step 5)

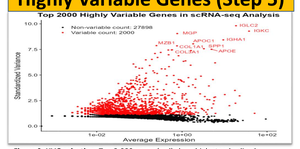


Figure 9. HVG selection. Top 2,000 genes (red) show high standardized variance, capturing key transcriptional diversity.

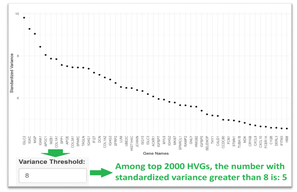


Figure 10. Top 50 HVGs by standardized variance. App plots top HVGs and counts those exceeding the user-set variance threshold.

Figure 11. Top 50 HVGs table. Alongside the plot, top HVGs are shown with key gene-level statistics.

Rank	Gene	Mean	Variance	VarianceStandardized
1	IGLC2	18.37	65292.78	9.82
2	IGKC	37.21	224272.35	9.28
3	MGP	0.77	36.79	9.05
4	IGHA1	9.25	19616.02	8.43
5	APOC1	1.82	69.00	8.05

Functional Enrichment Analysis (Step 6)

Functional analysis of HVGs in progressive disease (PD) cells revealed enriched GO terms and KEGG pathways linked to **extracellular matrix structure** and **immune processes**. GO terms included **collagen-containing extracellular matrix**, **structural molecule activity**, **complement activation**, and **humoral immune response**. KEGG pathways such as **ECM-receptor interaction**, **focal adhesion**, and **cytokine-cytokine receptor interaction** were also enriched. Notable HVGs associated with these pathways included **COL1A1**, **FN1**, **MMP2**, **C3**, **IGHM**, and **IL1B**.

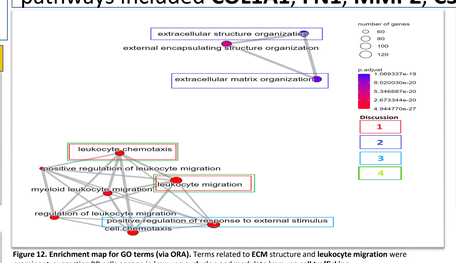


Figure 12. Enrichment map for GO terms (via ORA). Terms related to **ECM** structure and **leukocyte migration** were prominent, suggesting PD cells engage in **immune exclusion** and modulate **immune cell trafficking**.

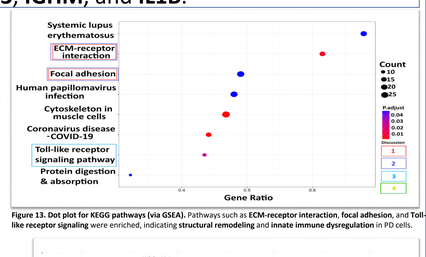


Figure 13. Dot plot for KEGG pathways (via GSEA). Pathways such as **ECM-receptor interaction**, **focal adhesion**, and **Toll-like receptor signaling** were enriched, indicating **structural remodeling** and **innate immune dysregulation** in PD cells.

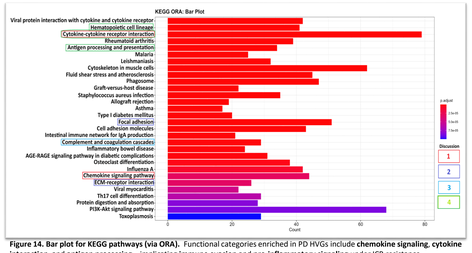


Figure 14. Bar plot for KEGG pathways (via ORA). Functional categories enriched in PD HVGs include **chemokine signaling**, **cytokine interaction**, and **antigen processing**—implicating **immune evasion** and **pro-inflammatory signaling** under ICB resistance.

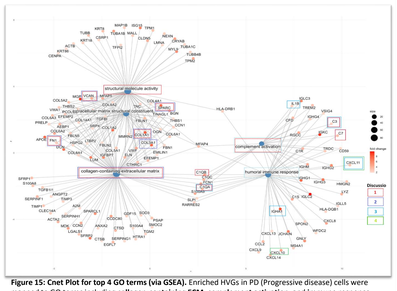


Figure 15: Cnet Plot for top 4 GO terms (via GSEA). Enriched HVGs in PD (Progressive disease) cells were mapped to GO terms including **collagen-containing ECM**, **complement activation**, and **immune response**. Notable genes (e.g., **COL1A1**, **C3**, **IL1B**) suggest **matrix remodeling** and **immune dysfunction**.

Discussion

- High-variance genes (HVGs) enriched in PD cells suggest an **immune-evasive** and **fibrotic tumor environment** [(1)].
- GO and KEGG analyses consistently revealed **extracellular matrix remodeling** (e.g., **COL1A1**, **FN1**, **SPARC**) via terms like **ECM-receptor interaction** and **focal adhesion**, pointing to enhanced **structural barriers** against immune infiltration [(2)].
- **Complement-related** and **humoral immune genes** (**C3**, **C1QA**, **IGHA1**) imply **chronic inflammation** and **immune dysfunction**, while **chemokine and cytokine signaling** terms suggest **altered immune cell trafficking** [(3)].
- Notably, **leukocyte migration** and **antigen presentation pathways** (**CXCL10**, **HLA-DRB1**) indicate ongoing **immune activity**, yet potentially ineffective, consistent with ICB resistance [(4)].
- These insights align with mechanisms of **immune exclusion** and **resistance** in progressive kidney cancer.

Note: [(1)]–[(4)] refer to color-coded terms in all functional analysis plots.

Conclusions

This app streamlines scRNA-seq preprocessing, QC, normalization, and HVG identification via an interactive interface. By integrating metadata with the count matrix, it enables **dynamic cell-type subsetting**, **functional enrichment** (GO/KEGG via GSEA and ORA), and **biomarker discovery**—supporting experimental and computational researchers with real-time, flexible analysis.

Future Work

- **Adapting and evaluating** the app across additional disease contexts beyond ICB-treated kidney cancer
- Adding capabilities for **differential expression analysis**, **HVG-DEG comparison**, and **functional analysis of shared genes**

Contact Information
Omid Mohajeri
Bioinformatician, Independent Researcher
omidmoh1980@gmail.com

Scan to watch 4-min demo (of the Shiny app)

My Portfolio Website:

References

1. Bi et al. (2021). Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. Cancer Cell, 39(5):649–661.e5.
2. Chen & Mellman (2017);
3. Mariathasan et al. (2018);
4. Roumenina et al. (2019);
5. Kuen et al. (2017)

Acknowledgements

I sincerely thank **Professor Claudio Angione** (Professor of AI and Co-Lead of the Centre for Digital Innovation, and recipient of a Turing Network Development Award funded by The Alan Turing Institute in 2022 and 2023) and **Dr. Annalisa Occhipinti** (Associate Professor), Teesside University, for their guidance and inspiration during our collaborative project on scRNA-seq analysis in kidney cancer, which laid the foundation for this work.