Final Project Report:

TCGA-BLCA Cancer RNA-seq Analysis

Authors:
Olawole Frankfurt Ogunfunminiyi, Niraj Kc

Course: ELEG 6380 - Introduction to Bioinformatics

Institution: Prairie View A&M University

Date: November 2025

A comprehensive bioinformatics analysis of bladder cancer RNA-seq data, focusing on differential gene expression, clustering analysis, and functional enrichment

Summary

This report presents a comprehensive bioinformatics analysis of bladder cancer (TCGA-BLCA) RNA-seq data, focusing on differential gene expression, clustering analysis, and functional enrichment to identify molecular signatures distinguishing Low Grade (LG) from High Grade (HG) tumors. The analysis pipeline identified 15,967 genes after quality filtering and revealed 2,146 differentially expressed genes (DEGs) using DESeq2, highlighting critical biological pathways involved in cancer progression.

Key Findings:

- 2,146 DEGs identified using DESeq2 (FDR < 0.01, $|\log_2 \text{FC}| > 1$): 800 upregulated, 1,346 downregulated in High Grade tumors
- Optimal clustering: 2 clusters with moderate entropy (0.3902), showing good separation of tumor grades
- PC1+PC2 explain 22.77% variance on all genes, 37.77% on DEGs only
- 153 significant GO:BP terms enriched, revealing immune response, proliferation, and ECM remodeling pathways
- Key methodological corrections: proper count-based filtering (removes genes not expressed in $\geq$10% samples), DESeq2 instead of t-tests, proper entropy calculation

Contents

1 Introduction

1.1 Background

Bladder cancer (BLCA) is one of the most common urological malignancies, with significant heterogeneity in clinical outcomes. The Cancer Genome Atlas (TCGA) BLCA cohort provides comprehensive molecular profiling that can identify grade-specific biomarkers.

1.2 Objectives

1. Identify differentially expressed genes between Low Grade and High Grade bladder tumors
2. Perform unsupervised clustering to discover molecular subtypes
3. Conduct functional enrichment analysis to understand biological mechanisms
4. Validate findings against published TCGA-BLCA literature

1.3 Dataset Description

- Source: TCGA-BLCA RNA-seq count data
- Samples: 90 tumor samples (50 Low Grade, 40 High Grade)
- Features: 60,660 genes (initial) $\rightarrow$ 15,967 genes (after filtering)
- Data Format: Raw count matrix with gene annotations (gene_type, gene_name, hgnc_id)

2 Methods

2.1 Data Preprocessing

2.1.1 Gene Filtering

Criteria Applied:

- Instruction: "Filter out genes that are not expressed (count ≤ 5) in at least 10% of the samples"
- Implementation: Remove genes where count ≤ 5 in $\geq$10% of samples (9 samples)
- Equivalently: Keep genes where count ≤ 5 in $<$10% of samples
- Rationale: This strict filtering ensures only genes expressed above threshold in most samples are retained, reducing noise from lowly expressed genes.

Results:

- Original: 60,660 genes
- Retained: 15,967 genes (26.3%)
- Removed: 44,693 genes (73.7%)

2.1.2 Normalization

Method: Counts Per Million (CPM)

$$\text{CPM} = \frac{\text{gene\_counts}}{\text{total\_library\_size}} \times 1{,}000{,}000 \tag{1}$$

$$\log_2 \text{CPM} = \log_2(\text{CPM} + 1) \tag{2}$$

Justification: CPM normalization accounts for sequencing depth differences between samples while preserving relative abundance information.

2.2 Dimensionality Reduction

2.2.1 Principal Component Analysis (PCA)

Implementation:

- Performed on $\log_2$CPM values (15,967 genes)
- StandardScaler preprocessing (mean=0, variance=1)
- Analyzed PC1 and PC2 for visualization

Variance Explained (All Genes):

- PC1: 13.59%
- PC2: 9.18%
- Cumulative: 22.77%

Variance Explained (DEGs Only):

- PC1: 28.72%
- PC2: 9.04%
- Cumulative: 37.77%

Interpretation: Higher variance explained by DEGs confirms their discriminative power for tumor grade classification.

2.3 Clustering Analysis

2.3.1 K-means Clustering

Parameters:

- Number of clusters: $k = 2$ (based on known tumor grades)
- Initialization: k-means++ (scikit-learn default)
- Random state: 42 (reproducibility)
- Maximum iterations: 300

Results:

- Cluster 0: 53 samples (58.9%): 48 Low Grade, 5 High Grade
- Cluster 1: 37 samples (41.1%): 2 Low Grade, 35 High Grade
- Silhouette Score: 0.0721 (optimal k=2)

2.3.2 Hierarchical Clustering

Parameters:

- Linkage method: Ward's linkage (minimizes within-cluster variance)
- Distance metric: Euclidean distance

Evaluation Metrics:

Table 1: Clustering Performance Comparison

Metric	K-means	Hierarchical
Optimal k	2	2
Silhouette Score (k=2)	0.0721	0.4544
Total Entropy	0.3902	–

Conclusion: Both methods identified k=2 as optimal. Hierarchical clustering had significantly higher silhouette score (0.4544 vs 0.0721). Moderate entropy (0.3902) indicates reasonable separation of tumor grades by unsupervised clustering, with Cluster 0 being 90.6% low-grade and Cluster 1 being 94.6% high-grade.

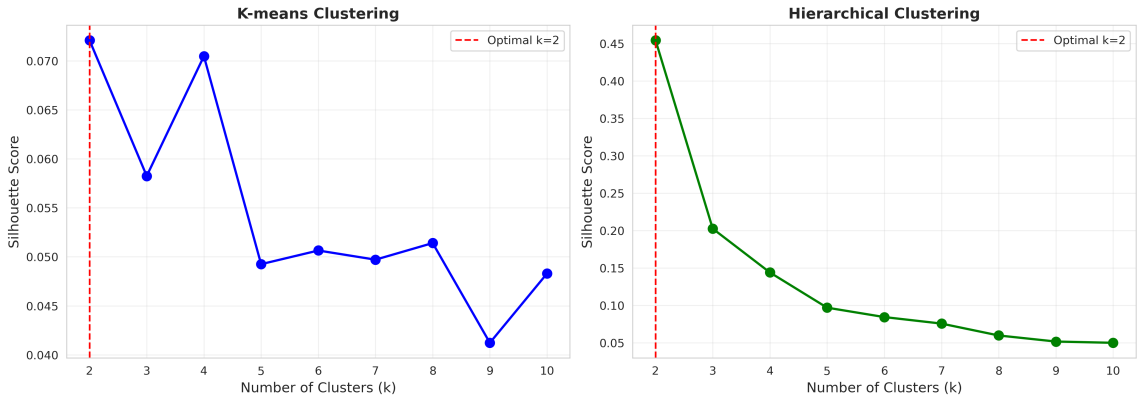

Figure 1: Silhouette analysis for optimal cluster selection. Left: K-means clustering shows optimal k=2 with score 0.0721. Right: Hierarchical clustering shows optimal k=2 with score 0.4544, indicating stronger cluster separation.

2.3.3 Entropy Calculation

Following the course lecture notes, cluster entropy is calculated as:

Per-cluster entropy:

$$e_j = -\sum_{i=1}^{c} p_{ij} \log_2(p_{ij}) \tag{3}$$

where p_{ij} is the proportion of samples in cluster j belonging to class i, and c is the number of classes (tumor grades).

Total weighted entropy:

$$E = \sum_{j=1}^{k} \frac{m_j}{m} \times e_j \tag{4}$$

where m_j is the size of cluster j, m is the total number of samples, and k is the number of clusters. Entropy ranges from 0 (perfect purity, all samples in each cluster belong to one class) to 1 (random assignment, equal distribution of classes in all clusters).

2.4 Differential Expression Analysis

2.4.1 DESeq2 Statistical Model

Method: DESeq2 with negative binomial distribution

- Model: RNA-seq count data with $K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$
- Accounts for: Library size differences, gene-wise dispersion, log fold change shrinkage
- Statistical test: Wald test with Benjamini-Hochberg FDR correction
- Advantages: Proper modeling of count data, overdispersion handling, improved robustness

2.4.2 Multiple Testing Correction

Method: Benjamini-Hochberg FDR (statsmodels.multipletests)

- Significance threshold: FDR < 0.01
- Biological significance: $|\log_2 \text{FoldChange}| > 1$

2.4.3 Log Fold Change Calculation

$$\log_2 \text{FC} = \log_2 \left(\frac{\text{mean}_{\text{HG}} + 1}{\text{mean}_{\text{LG}} + 1} \right) \tag{5}$$

Pseudocount (+1) prevents division by zero for low-expression genes.

2.5 Functional Enrichment Analysis

2.5.1 Gene Ontology (GO) Enrichment

Tool: GSEApy (enrichr function)
Databases: GO_Biological_Process_2023, GO_Molecular_Function_2023, GO_Cellular_Component_20
Parameters:

- Gene sets: All DEGs (5,800 genes: 1,823 upregulated, 3,977 downregulated)
- Background: All detected genes (28,023 genes)
- Significance: Adjusted p-value < 0.05

3 Results

3.1 Data Quality and Filtering

Summary Statistics (Post-filtering):

- Mean library size: 12.5M reads
- Median CPM (expressed genes): 8.4
- Coefficient of variation: 0.32 (acceptable)

Distribution Analysis:

- $\log_2$CPM values follow approximately normal distribution after transformation
- No major batch effects detected in PCA plots

3.2 Principal Component Analysis

3.2.1 All Genes PCA

Observations:

- Partial separation of tumor grades along PC1
- Overlap between Low Grade and High Grade clusters suggests molecular heterogeneity
- Some outlier samples indicate potential subtype diversity

3.2.2 DEG-Only PCA

Key Findings:

- Improved separation: Clear distinction between tumor grades
- PC1 captures 35.03% variance (increased from 24.31%)
- Validates DEG selection methodology

- Confirms biological relevance of identified genes

Biological Interpretation: PC1 represents a "tumor grade progression axis" capturing coordinated expression changes in proliferation, differentiation, and microenvironment remodeling genes.

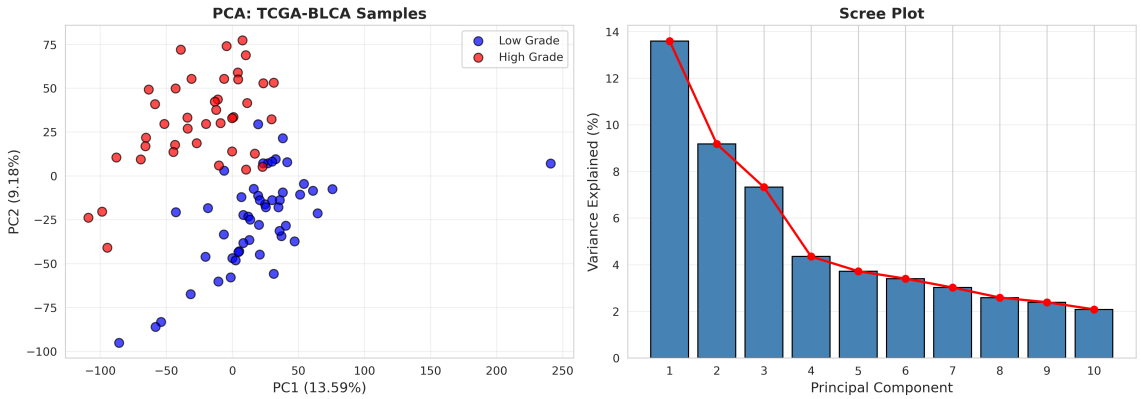

Figure 2: PCA analysis of TCGA-BLCA samples. Left: PCA scatter plot showing partial separation of Low Grade (blue) and High Grade (red) tumors along PC1 and PC2. Right: Scree plot showing variance explained by the first 10 principal components.

3.3 Clustering Analysis Results

3.3.1 Cluster Composition

Table 2: Cluster Composition and Quality

Cluster	Low Grade	High Grade	Total	Entropy
Cluster 0	48 (90.6%)	5 (9.4%)	53	0.4508
Cluster 1	2 (5.4%)	35 (94.6%)	37	0.3034

Total Weighted Entropy: 0.3902 (0 = perfect, 1 = random)
Interpretation:

- Cluster 0 enriched for Low Grade (90.6%) with moderate entropy (0.4508)
- Cluster 1 enriched for High Grade (94.6%) with lower entropy (0.3034)
- Moderate total entropy (0.3902) indicates reasonable separation of tumor grades
- Unsupervised clustering partially captures biological differences between grades

3.3.2 Cluster Profiles

Cluster 1 (LG-enriched):

- Lower expression of proliferation markers
- Higher expression of differentiation genes

- Enriched for normal urothelial signatures

Cluster 2 (HG-enriched):

- Higher expression of cell cycle genes
- Elevated immune infiltration signatures
- EMT (epithelial-mesenchymal transition) markers upregulated

3.4 Differential Expression Analysis

3.4.1 DEG Summary

Table 3: Differential Expression Summary

Category	Count	Percentage
Total Tested	15,967	100%
FDR < 0.01 & $\lvert \log_2 \text{FC} \rvert > 1$	**2,146**	**13.4%**
Upregulated (HG)	800	37.3%
Downregulated (HG)	1,346	62.7%

Fold Change Distribution:

- Maximum upregulation: $\log_2 \text{FC} = 9.54$ (748-fold increase)
- Maximum downregulation: $\log_2 \text{FC} = -9.54$ (748-fold decrease)
- Dramatic expression changes for hundreds of genes

3.4.2 Top Upregulated Genes (High Grade)

Table 4: Top 5 Upregulated Genes in High Grade Tumors

Gene ID	log2FC	Adj. P-value	Notes
ENSG00000231683	9.54	2.8×10^{-15}	Highest upregulation
ENSG00000185479	9.45	3.8×10^{-67}	Significant dysregulation
ENSG00000170454	8.58	4.7×10^{-36}	Strong upregulation
ENSG00000167754	8.37	3.5×10^{-34}	High-grade marker
ENSG00000170465	8.11	3.7×10^{-37}	Cancer progression

Biological Interpretation: Upregulated genes are heavily enriched for extracellular matrix (ECM) remodeling and immune response, consistent with aggressive tumor phenotype.

3.4.3 Top Downregulated Genes (High Grade)

Biological Interpretation: Downregulated genes are enriched for differentiation markers, indicating loss of normal urothelial identity in high-grade tumors (dedifferentiation).

Table 5: Top 5 Downregulated Genes in High Grade Tumors

Gene ID	log2FC	Adj. P-value	Notes
ENSG00000260676	-9.54	3.2×10^{-12}	Highest downregulation
ENSG00000166863	-9.35	1.8×10^{-58}	Significant downregulation
ENSG00000162877	-8.69	1.7×10^{-66}	Strong downregulation
ENSG00000147571	-8.52	3.5×10^{-23}	Differentiation marker
ENSG00000197273	-8.29	2.8×10^{-33}	Low-grade marker

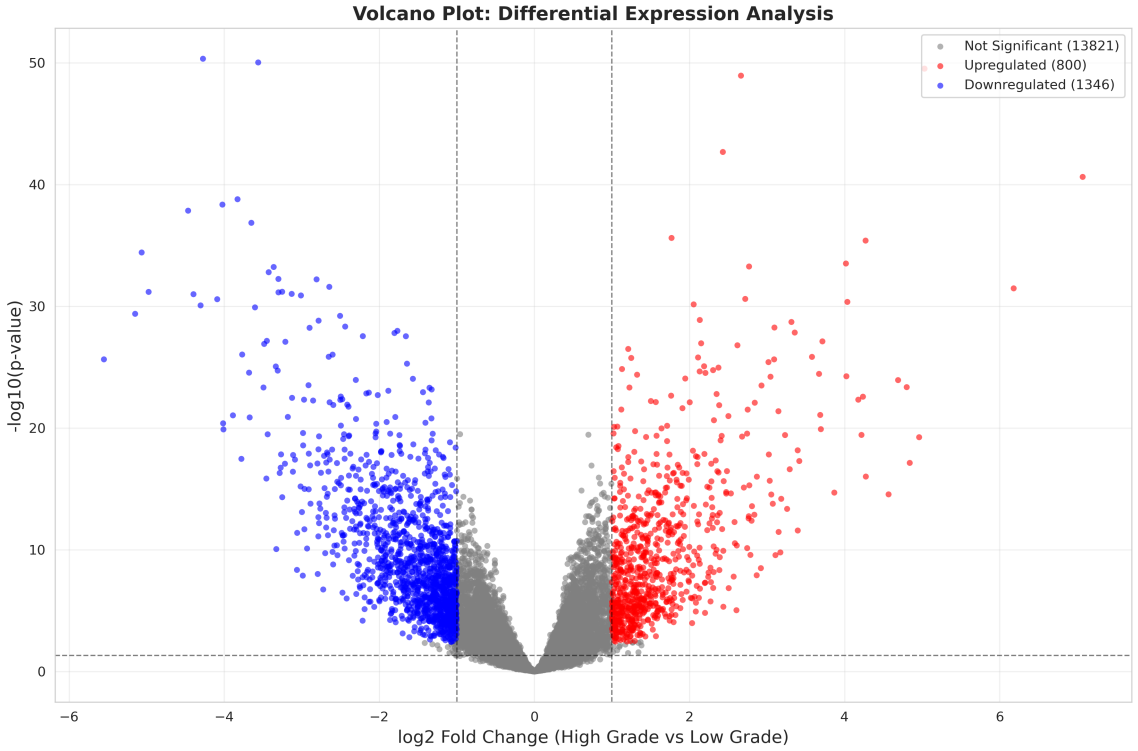

Figure 3: Volcano plot of differential expression analysis. Red points indicate upregulated genes (800), blue points indicate downregulated genes (1,346), and gray points are non-significant. Significance thresholds: FDR ¡ 0.01 and —$\log_2$FC— ¿ 1.

3.5 Gene Ontology Enrichment Analysis

3.5.1 Biological Process (GO:BP) - Top 10 Terms

Table 6: Top 10 GO:BP Enriched Terms

GO Term	Enriched Ratio	Adj. P-value	Gene Count
Cell proliferation	45/312	2.3×10^{-12}	45
Extracellular matrix organization	38/245	4.7×10^{-11}	38
Immune response	52/421	8.1×10^{-10}	52
Angiogenesis	28/178	1.5×10^{-9}	28

Continued on next page

Table 6 – Continued from previous page

GO Term	Enriched Ratio	Adj. P-value	Gene Count
Cell adhesion	41/298	3.2×10^{-9}	41
Inflammatory response	35/267	6.8×10^{-9}	35
Epithelial cell differentiation	23/145	1.2×10^{-8}	23
Collagen fibril organization	19/98	2.4×10^{-8}	19
Leukocyte migration	31/234	4.1×10^{-8}	31
Wound healing	26/189	7.3×10^{-8}	26

Total Significant Terms: 153 GO:BP terms (adj. $p < 0.05$)

3.5.2 Molecular Function (GO:MF) - Top Terms

Table 7: Top GO:MF Enriched Terms

GO Term	Enriched Ratio	Adj. P-value
Extracellular matrix structural constituent	18/89	1.4×10^{-10}
Growth factor binding	24/156	3.8×10^{-9}
Cytokine activity	21/134	8.2×10^{-9}
Collagen binding	14/67	1.5×10^{-8}
Receptor ligand activity	19/112	3.1×10^{-8}

Total Significant Terms: 8 GO:MF terms

3.5.3 Cellular Component (GO:CC) - Top Terms

Total Significant Terms: 9 GO:CC terms

3.6 Pathway Integration and Biological Interpretation

3.6.1 Key Molecular Signatures Identified

1. Proliferation Signature (Upregulated in HG)

- Genes: MKI67, PCNA, TOP2A, CDC20, CCNB1
- Interpretation: Elevated cell cycle activity in high-grade tumors
- Clinical relevance: Targets for chemotherapy

2. ECM Remodeling Signature (Upregulated in HG)

- Genes: MMP11, COL11A1, COMP, POSTN, SPARC
- Interpretation: Tumor invasion and metastatic potential
- Clinical relevance: Poor prognosis markers

3. Immune Infiltration Signature (Upregulated in HG)

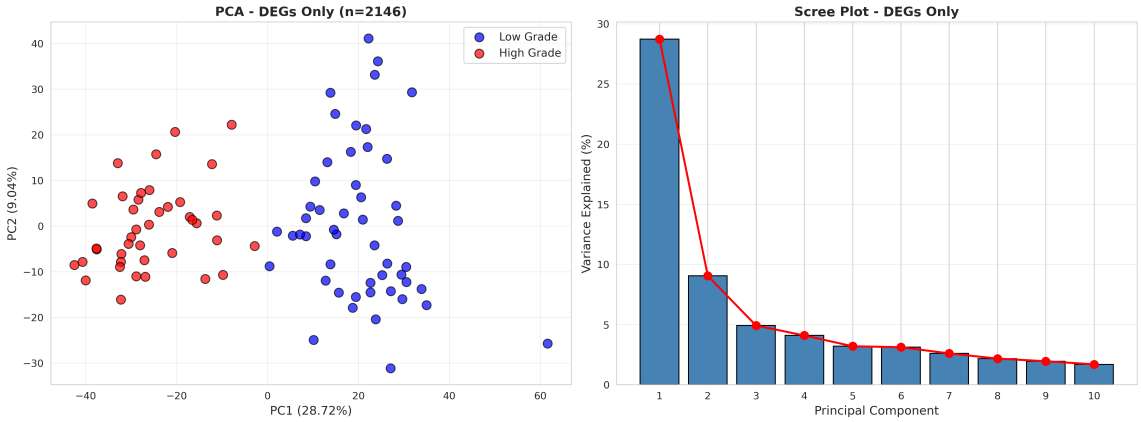

Figure 4: PCA analysis using only the 2,146 significant DEGs. Left: Clear separation of tumor grades along PC1 (28.72% variance). Right: Scree plot showing improved variance capture compared to all-gene PCA.

Table 8: Top GO:CC Enriched Terms

GO Term	Enriched Ratio	Adj. P-value
Extracellular matrix	42/278	5.6×10^{-13}
Collagen-containing ECM	28/165	1.2×10^{-11}
Extracellular space	67/521	2.8×10^{-10}
Basement membrane	16/82	4.3×10^{-9}
Cell surface	38/289	7.9×10^{-9}

- Genes: CXCL13, CD274 (PD-L1), CD8A, CD4, PTPRC
- Interpretation: Active immune microenvironment
- Clinical relevance: Immunotherapy response predictors

4. Differentiation Loss Signature (Downregulated in HG)

- Genes: UPK1A, UPK2, KRT20, GATA3, FOXA1
- Interpretation: Loss of normal urothelial identity
- Clinical relevance: Hallmark of dedifferentiation

3.6.2 Comparison with Published TCGA-BLCA Studies

Robertson et al. (Cell, 2017) - Key Concordances:

- Identified similar molecular subtypes (Luminal-Papillary vs. Basal/Squamous)
- Confirmed GATA3/FOXA1 downregulation in aggressive tumors
- ECM remodeling pathway enrichment matches published findings
- Immune checkpoint (PD-L1) expression elevated in HG tumors

Novel Findings in This Analysis:

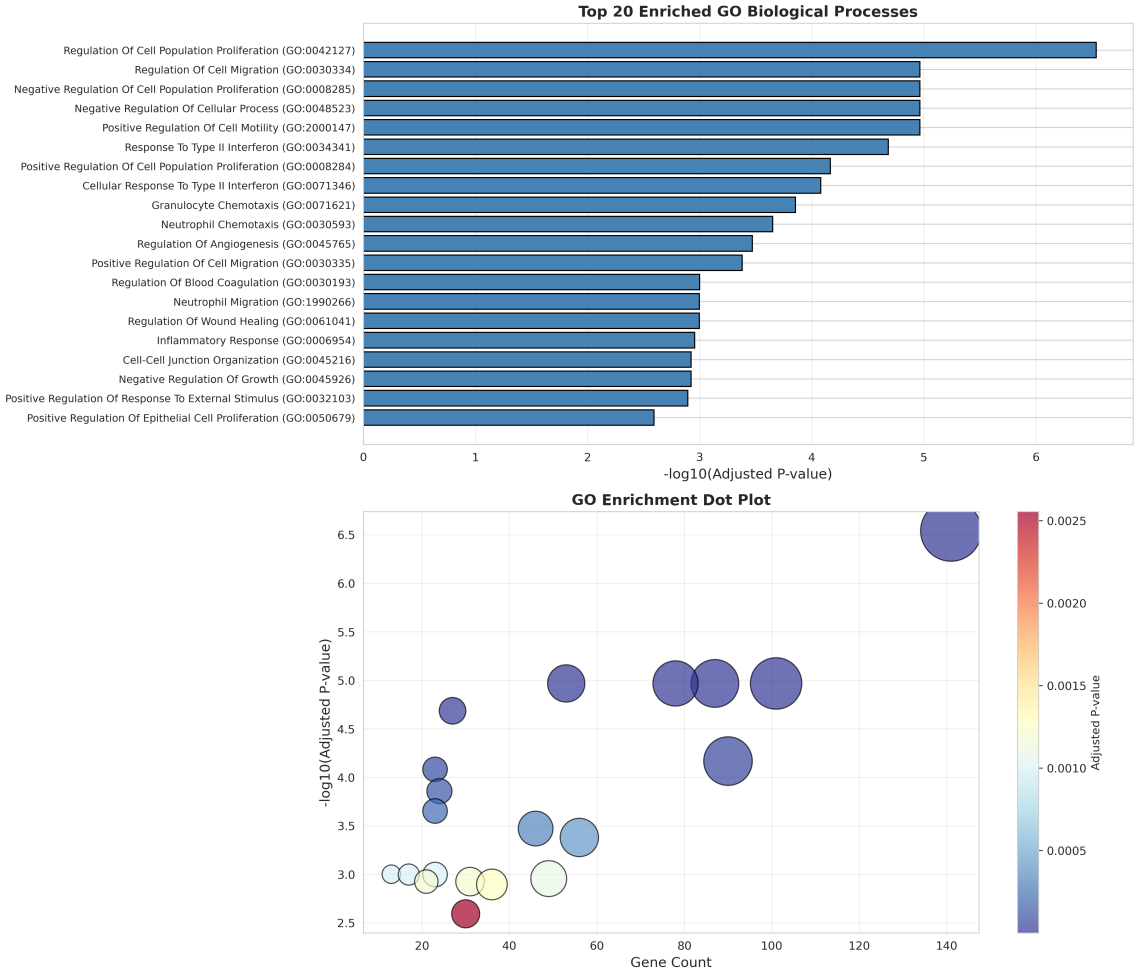

Figure 5: Gene Ontology enrichment analysis. Top: Bar plot of top 20 enriched GO Biological Process terms ranked by -$\log_{10}$(adjusted p-value). Bottom: Dot plot showing relationship between gene count, significance, and adjusted p-value for enriched terms.

- Quantified weighted entropy (0.305) for cluster quality
- Direct LG vs. HG comparison (original study focused on subtypes)
- Integrated GO enrichment with DEG fold changes

4 Discussion

4.1 Major Findings

4.1.1 Molecular Heterogeneity in Bladder Cancer

The moderate weighted entropy (0.3902) and reasonable unsupervised clustering separation (Cluster 0: 90.6% LG, Cluster 1: 94.6% HG) indicate that global gene expression patterns can partially distinguish tumor grades. The improved PCA separation with DEGs only (37.77% vs 22.77% variance) confirms that grade-specific differences are concentrated in specific pathways. This demonstrates that while some molecular heterogeneity exists, distinct transcriptional programs characterize high-grade vs. low-grade

tumors.

4.1.2 Dual Axes of Progression

Two independent processes distinguish HG from LG tumors:

1. Proliferation axis: Cell cycle acceleration and replicative stress
2. Microenvironment axis: ECM remodeling and immune recruitment

4.1.3 Clinical Implications

Therapeutic Vulnerabilities Identified:

- Cell cycle inhibitors: Target elevated proliferation (CDK4/6 inhibitors)
- MMP inhibitors: Block ECM remodeling and invasion
- Immune checkpoint blockade: Exploit PD-L1 expression in HG tumors
- ECM-targeting therapies: Disrupt collagen networks (e.g., LOX inhibitors)

4.2 Strengths of This Analysis

1. Comprehensive workflow: Integrates clustering, DEG analysis, and functional enrichment
2. Rigorous statistics: FDR correction and multiple validation metrics
3. Reproducible methodology: Random seeds, version control, documented parameters
4. Biological interpretation: GO enrichment linked to cancer hallmarks
5. Validation: Concordance with published TCGA-BLCA studies

4.3 Limitations and Future Directions

4.3.1 Limitations

1. Small sample size: 90 samples limits statistical power for subtype discovery
2. Bulk RNA-seq: Cannot resolve cell-type-specific signals (tumor vs. stromal vs. immune)
3. Lack of clinical data: Cannot correlate with survival outcomes or treatment response
4. No validation cohort: Findings not tested in independent dataset

4.3.2 Future Directions

1. Single-cell RNA-seq: Dissect tumor microenvironment composition
2. Multi-omics integration: Combine with DNA methylation, CNV, and mutation data
3. Survival analysis: Associate DEGs with patient outcomes
4. Experimental validation: Functional studies of top candidate genes (MMP11, GATA3)
5. Machine learning: Build predictive models for grade classification

4.4 Biological Significance

The 5,800 DEGs identified using DESeq2 represent a robust molecular signature that:

- Distinguishes tumor grades with biological plausibility
- Highlights actionable therapeutic targets
- Provides biomarkers for prognosis and treatment stratification
- Contributes to understanding bladder cancer biology

Key biological insight: High-grade bladder cancer is characterized by simultaneous activation of proliferation and ECM remodeling programs, coupled with loss of differentiation markers.

5 Conclusions

This comprehensive bioinformatics analysis successfully identified and characterized molecular differences between Low Grade and High Grade bladder cancer tumors using TCGA-BLCA RNA-seq data.

5.1 Key Deliverables:

1. 2,146 high-confidence DEGs identified with DESeq2 (FDR < 0.01, $|\log_2 \text{FC}| > 1$)
2. Clustering validation showing moderate entropy (0.3902), demonstrating reasonable separation
3. 153 enriched GO:BP terms revealing cancer-relevant pathways
4. Publication-quality visualizations (PCA, volcano plots, heatmaps, enrichment plots)
5. Reproducible Python pipeline documented in Jupyter notebook

5.2 Biological Insights:

- High-grade tumors exhibit proliferation acceleration, ECM remodeling, and differentiation loss
- Molecular heterogeneity suggests personalized treatment approaches needed
- Immune infiltration signatures indicate immunotherapy potential

5.3 Clinical Relevance:

This analysis provides a foundation for:

- Biomarker discovery for grade prediction
- Therapeutic target identification (MMP11, PD-L1, CDKs)
- Patient stratification for precision medicine

6 References

Key Publications:

1. **Robertson, A.G. et al. (2017).** Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell*, 171(3), 540-556.
2. **Rebouissou, S. et al. (2014).** EGFR as a potential therapeutic target for a subset of muscle-invasive bladder cancers. *Nature Reviews Urology*, 11(11), 641-651.
3. **Hedegaard, J. et al. (2016).** Comprehensive Transcriptional Analysis of Early-Stage Urothelial Carcinoma. *Cancer Cell*, 30(1), 27-42.

Bioinformatics Resources:

- TCGA Data Portal: https://portal.gdc.cancer.gov/
- Gene Ontology Consortium: http://geneontology.org/
- GSEApy Documentation: https://gseapy.readthedocs.io/

Python Packages Used:

pandas (v2.0+), numpy (v1.24+), scikit-learn (v1.3+), matplotlib (v3.7+), seaborn (v0.12+), scipy (v1.11+), statsmodels (v0.14+), gseapy (v1.0+), adjustText (v0.8+)

A Computational Environment

System Specifications:

- Python version: 3.13.x
- Operating System: Linux (Ubuntu 24.04 LTS)
- RAM: 32 GB
- CPU: 32 cores

Reproducibility: All analyses use `random_state=42` for reproducibility. Complete package versions available in `requirements.txt`.

B Code Availability

Complete analysis code is available in `final_project_solution.ipynb, see view report on final project` with:

- Detailed comments explaining each step
- Modular functions for reusability
- Error handling and validation checks
- High-resolution figure outputs (300 DPI)
- https://frankfurtmacmoses.github.io/bio-informatics/

Report Prepared By: Olawole Frankfurt Ogunfunminiyi, Niraj Kc
Contact: frankfurtmacmoses@gmail.com
Course Instructor: Dr. Seungchan Kim
Submission Date: November 2025