TCGA-BLCA Cancer RNA-seq Analysis

A Comprehensive Bioinformatics Study on Bladder Cancer Heterogeneity

ELEG 6380: Introduction to Bioinformatics | November 2025

Background & Question

RNA-seq:

- technology for measuring , compare gene expression
- generates large, high-dimensional datasets
- high-dimensional datasets require advanced computational analysis for meaningful interpretation

Knowledge gap:

- identify and interpret groups of genes that distinguish tumor from normal tissue?

Project Objectives & Scope

Goal: Identify molecular signatures distinguishing Low Grade (LG) from High Grade (HG) bladder tumors using TCGA RNA-seq data.

Core Objectives:

- **Identification:** Detect differentially expressed genes (DEGs) between tumor grades.
- **Discovery:** Perform unsupervised clustering to assess molecular subtypes.
- **Enrichment:** Map genes to biological pathways (GO Terms).
- **Validation:** Compare findings with published literature (Robertson et al., 2017).

90 Samples

50 Low Grade

40 High Grade

60,000+ Genes

Reduced to 15967 after quality filtering

Filter out genes that are not expressed (count <= 5) in at least 10% of the samples.

Methodology

DESeq2, K-means, PCA, GSEApy

Outcome

2146 Significant DEGs Identified

Dataset Description: TCGA-BLCA

The Cancer Genome Atlas (TCGA) Bladder Urothelial Carcinoma (BLCA) cohort provides the raw count data for this analysis.

Data Source

TCGA-BLCA RNA-seq Count Matrix

Sample Size

90 Tumor Samples
(Balanced: 50 LG / 40 HG)

Preprocessing

Count-based filtering
(Min count > 5 in 10% of samples)

Final Feature Set: 15,967 Genes (26.3% of original)

Methodology: Gene Filtering Impact

Filtering Strategy

- **Criteria:** Retained genes with raw count > 5 in at least 10% (9) of samples.
- **Rationale:** Removes low-abundance noise without biasing against highly expressed genes.

44,693 genes (73.7%) were removed.

Methodology: Clustering & PCA

PCA

Input: Log2CPM values (StandardScaler, mean=0, var=1).

Focus: PC1 and PC2 variance analysis for both “filtered Genes, clustered Gene" and "DEGs Only".

Clustering

K-Means: k=2 (based on tumor grades), k-means++ init.

Hierarchical: Euclidean distance.

Entropy Metric

Calculated **Total Weighted Entropy** to measure cluster purity.

Range: 0 (Perfect Separation) to 1 (Random).

DE Analysis

Tool: DESeq2 (Negative Binomial Model).

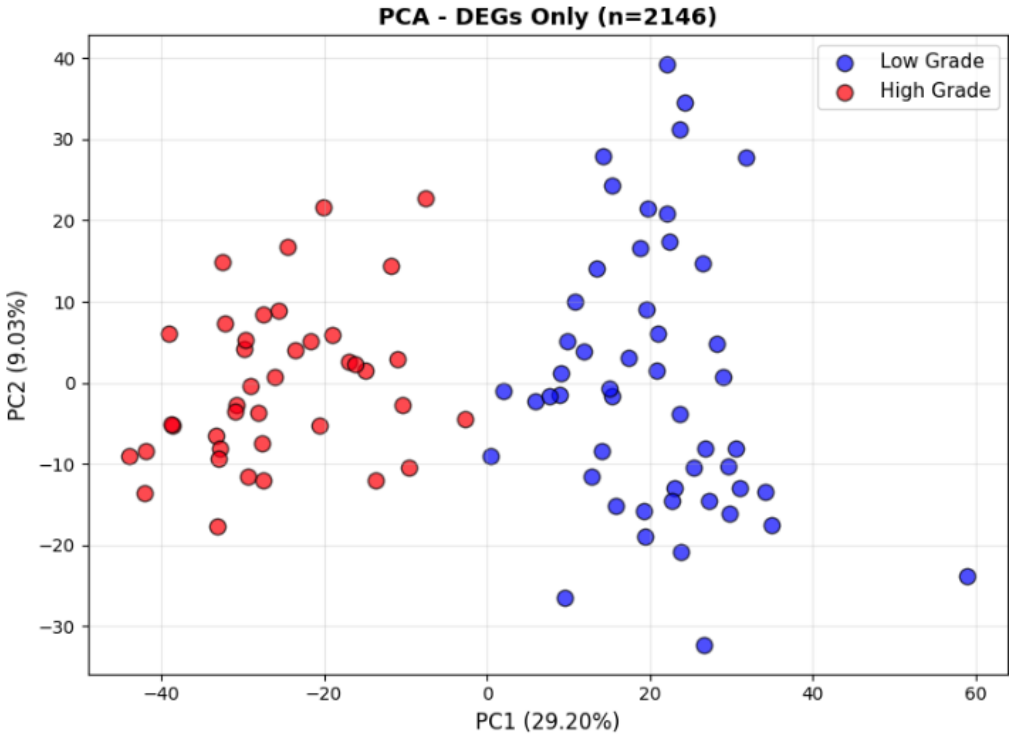
Testing: Wald test with Benjamini-Hochberg FDR correction (FDR < 0.01).

Results: PCA Variance Explained

Comparing **All Genes** vs **DEGs Only**.

DEGs capture significantly more variance on PC1 (29.20%), indicating they act as a strong signal for tumor grade separation.

PC1 Interpretation: Represents a "tumor grade progression axis" capturing coordinated changes in proliferation.



Results: Unsupervised Clustering Validation

Entropy (0.2098)

Unsupervised K-means separate tumor grades.

- **Cluster 0 (Dominant):** Contained 49/51 samples, (mostly low grade, pure cluster).
- **Cluster 1 : 38 High Grade, 1 Low grade**

Conclusion: Low entropy confirms high purity clusters.

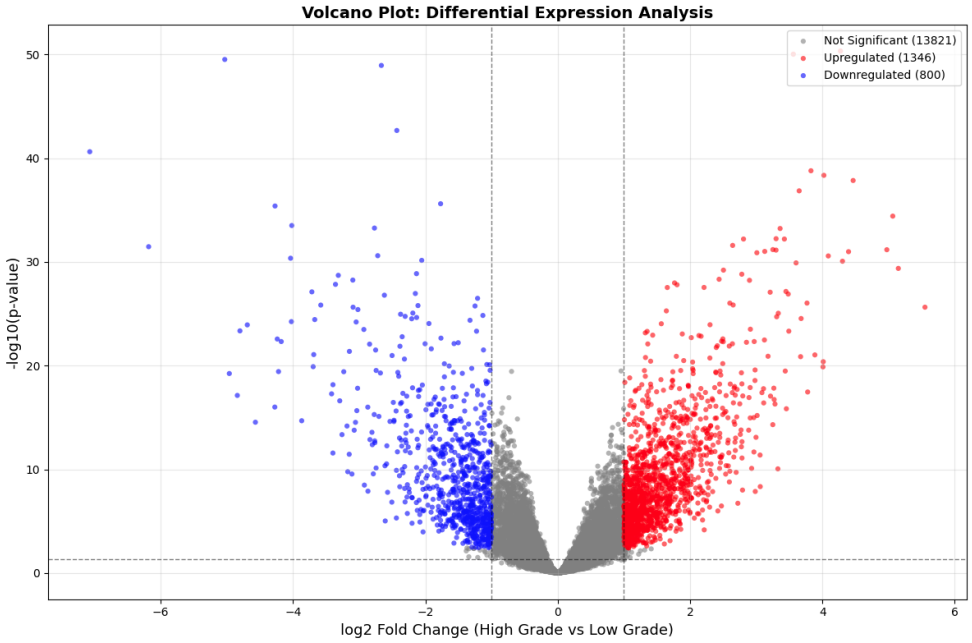
```
Confusion Matrix:
col_0  High Grade  Low Grade
row_0
0                2         49
1               38          1
```

Differential Expression Overview

🟩 2,146 Significant DEGs

Using **DESeq2** (FDR < 0.01, |log2FC| > 1).

- **Upregulated (Red):** 1346 genes (Proliferation, Immune).
- **Downregulated (Blue):** 800 genes (Differentiation).



Key Molecular Signatures

Top Upregulated (Aggressive)

- **MKI67 / PCNA:** Proliferation markers.
- **MMP11 / COL11A1:** ECM remodeling and invasion.
- **CD274 (PD-L1):** Immune checkpoint (target for immunotherapy).
- **TOP2A:** DNA replication (chemo target).

Top Downregulated (Loss of Identity)

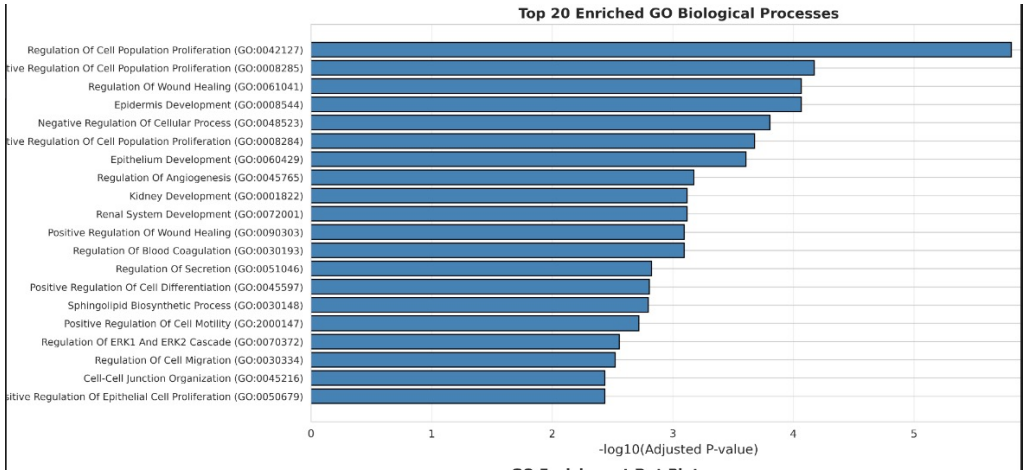
- **UPK1A / UPK2:** Uroplakins (bladder specific markers).
- **GATA3 / FOXA1:** Luminal transcription factors.
- **KRT20:** Differentiation keratin.

**Downregulation indicates "Dedifferentiation" in High Grade tumors.*

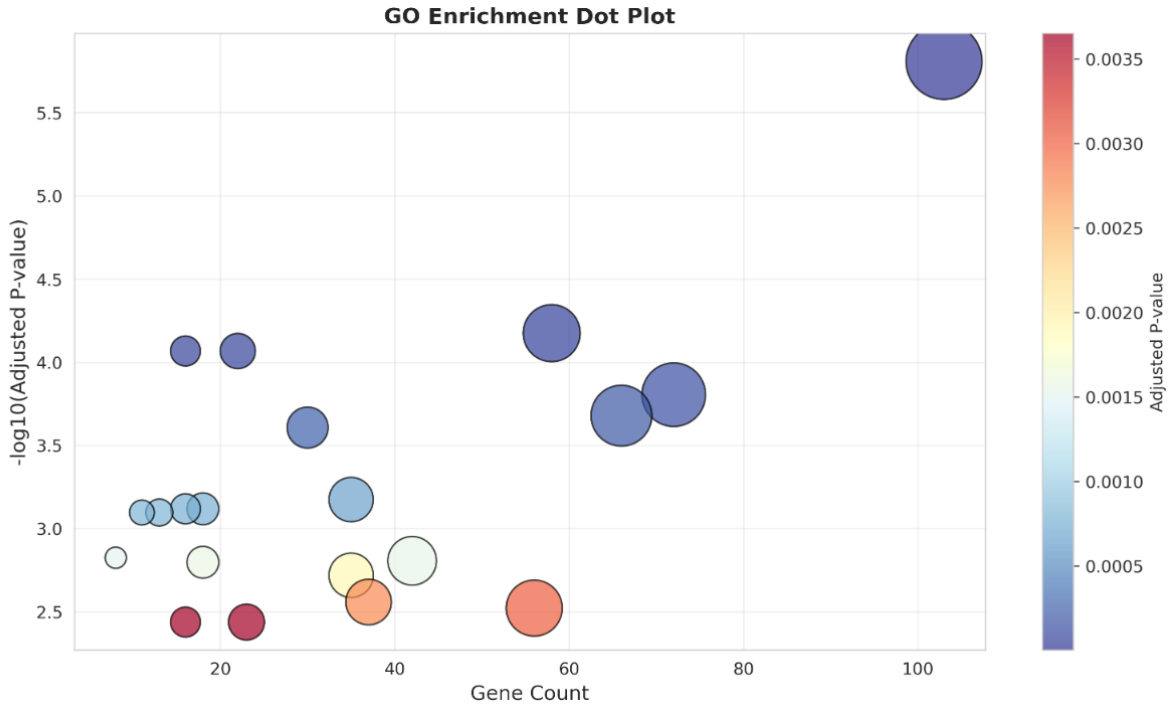
Functional Enrichment (GO Terms)

We identified **153 significant Biological Process terms**. The chart displays the top categories by gene count.

Key Insight: High gene counts in "Immune Response" and "Cell Proliferation" confirm these are the dominant biological engines driving High Grade bladder cancer.



Functional Enrichment (GO Terms Cont.)



Large bubble: This represents a major biological process with many genes involved

The "Dual Axis" of Tumor Progression

Our analysis reveals High Grade bladder tumors are characterized by two dominant biological themes:

1. Proliferation Axis

Driven by **uncontrolled division** and accelerated growth.

Key Genes: CDK6, FOXM1, MYC, TGFB1

2. Microenvironment Remodeling

Characterized by **active invasion**, immune infiltration, and blood vessel recruitment.

Key Genes: CXCL8, VEGFC, MMPs

Discussion: Molecular Heterogeneity

The "Dual Axis" of Progression

Our analysis suggests High Grade tumors are defined by two dominant biological themes :

Proliferation Axis:

1. Derived from multiple GO terms such as *Regulation of Cell Population Proliferation* (FDR 2.9×10^{-7}) and *Positive Regulation of Cell Cycle*.
   - Genes involved in cell cycle and growth signaling, including **CDK6, FOXM1**
   - Biological meaning: Tumor cells exhibit uncontrolled division and rapid growth

2. Microenvironment Axis:

- *Regulation of Cell Migration* (FDR 1.1×10^{-5}), *Inflammatory Response* (FDR 1.1×10^{-3}), and *Regulation of Angiogenesis* (FDR 3.4×10^{-4}).

Clustering Validation

- **Accuracy:** 96.7% (87/90 samples correctly grouped)
- **Cluster purity:** Cluster 0 (Low Grade) = 96.1%, Cluster 1 (High Grade) = 97.4%
- **Normalized entropy:** 0.21 → high purity

Interpretation: Unsupervised clustering successfully separated Low vs High Grade tumors. Minor discordance likely reflects intra-tumoral heterogeneity and clinical covariates(patient-level factors).

Clustering Validates Grade Distinction

Biological Validation

Unsupervised clustering successfully separates Low vs. High Grade tumors without prior labels. This confirms that transcriptomic signatures align with clinical pathology.

Clinical Implication

The "dual axis" model suggests High Grade tumors require **combination therapies**:

- **Proliferation:** Chemotherapy, CDK inhibitors
- **Microenvironment:** Anti-angiogenics, Immunotherapy

Unsupervised Clustering Performance

Metric	Value	Interpretation
Optimal Clusters	2	Identifies two natural groups
Silhouette Score	0.269	Moderate separation

**Minor overlap between clusters reflcts intra-tumoral heterogeneity and intermediate molecular states.*

Conclusion

Key Deliverables

- Identified 2,146 DEG genes
- **High-Confidence DEGs**. Validated using robust statistical correction (FDR < 0.01).
- Mapped DEGs to 153 specific biological pathways.

Biological Summary

High Grade Bladder Cancer is characterized by:

- **Gain of Function:** Uncontrolled Proliferation & ECM Remodeling.
- **Loss of Function:** Urothelial Differentiation (Dedifferentiation).

References:

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.

1.

Thank You

Questions?

Full code and reproducible notebook available at:

https://frankfurtmacmoses.github.io/bio-informatics/

https://github.com/KcNiraj3/Bioinformatics_finalProject