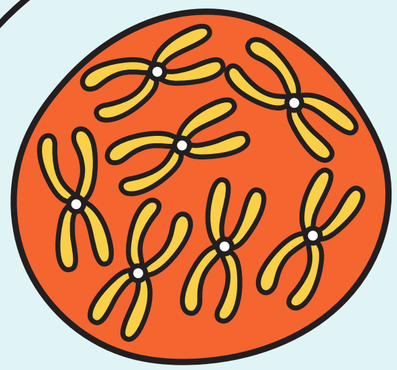
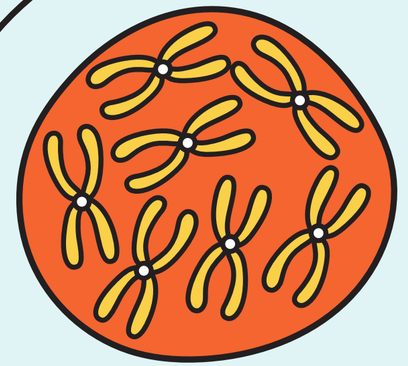
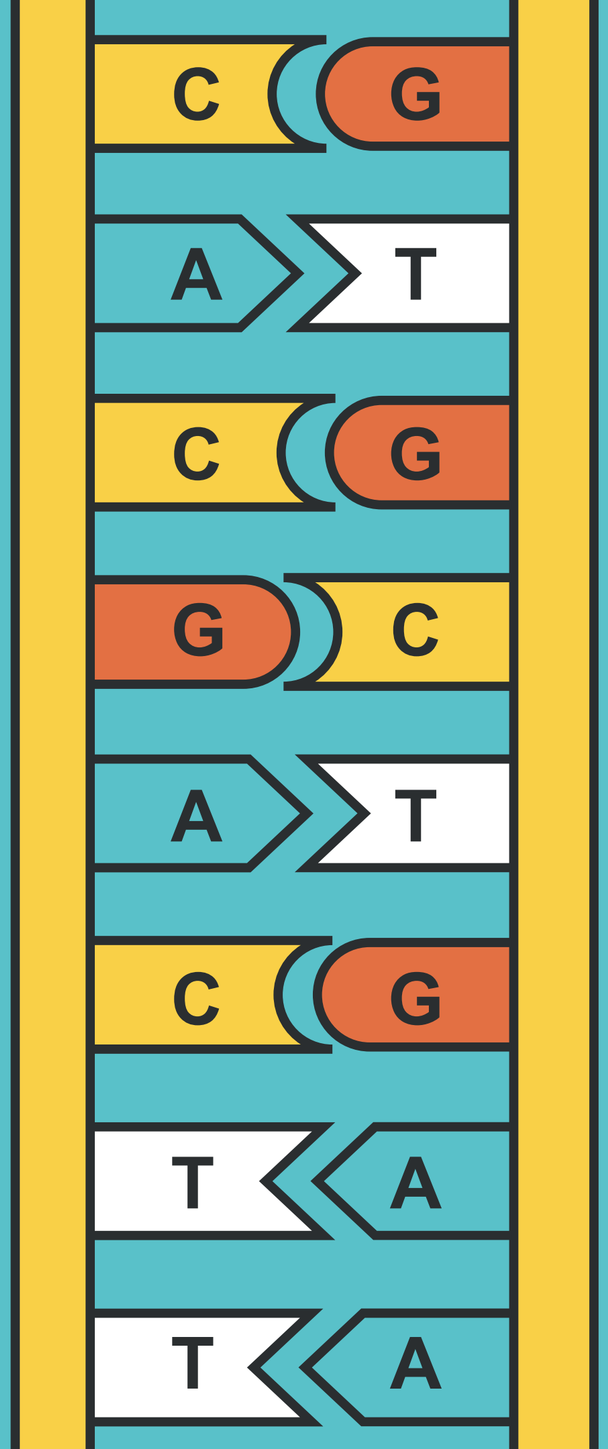


BREAST CANCER GENE ANALYSIS

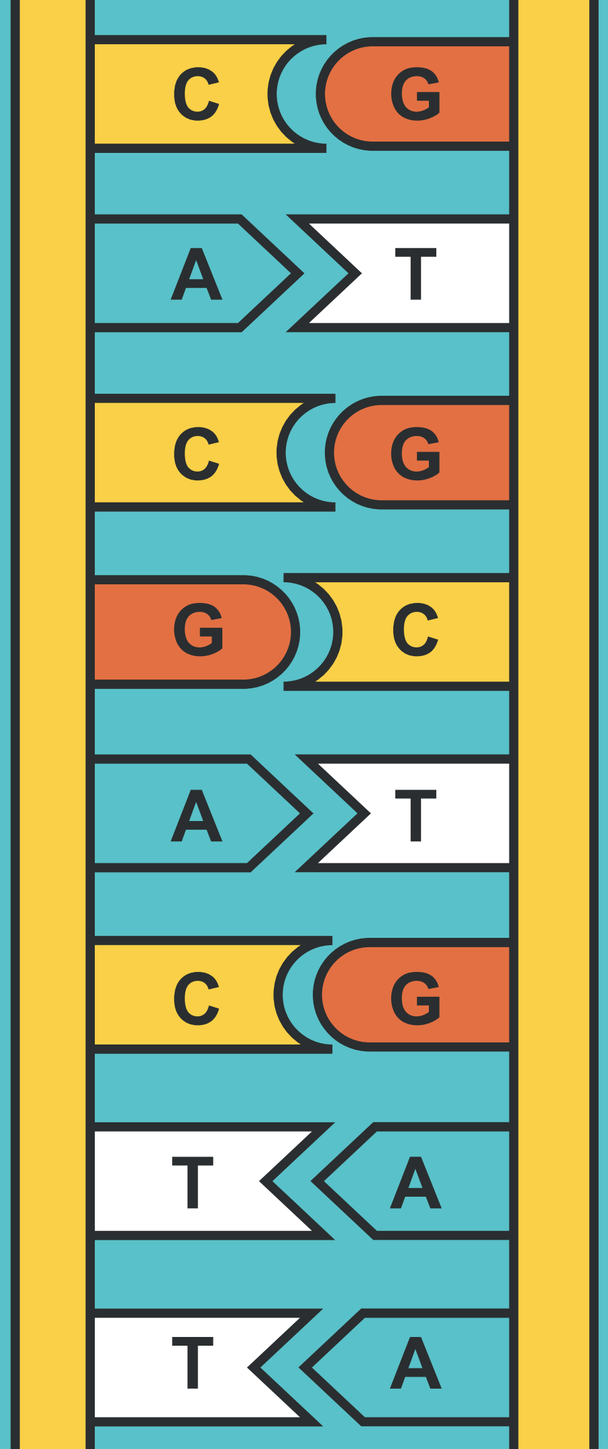
Team 1





AGENDA

1) Introduction
2) EDA
3) DEA
4) Dimension Reduction
5) Model Building
6) Conclusions
7) Deployment on Cloud



OBJECTIVES

1. Examine the differences in gene expression across various stages of breast cancer.
2. Predict the likelihood of survival for breast cancer patients based on gene expression.
3. Explore gene-gene interactions and their role in cancer progression.

LITERATURE REVIEW

Comparative evaluation of feature reduction methods for drug response prediction Expression Analysis (DEA) by Farzaneh Firoozbakht and others

What it is: study on feature reduction techniques for gene analysis.

How it was used: initial directions on techniques used.

Identification of Gene Expression in Different Stages of Breast Cancer with Machine Learning by Ali AbidalKareem and others

What it is: gene expression analysis by cancer stage

How it was used :Introduced Neighborhood Component Analysis (NCA) and provided insights on how it is used to isolate relevant genes or expression features.

DATA DESCRIPTION

Data source: TCGA human tumor data (transcriptomes) on breast cancer samples.

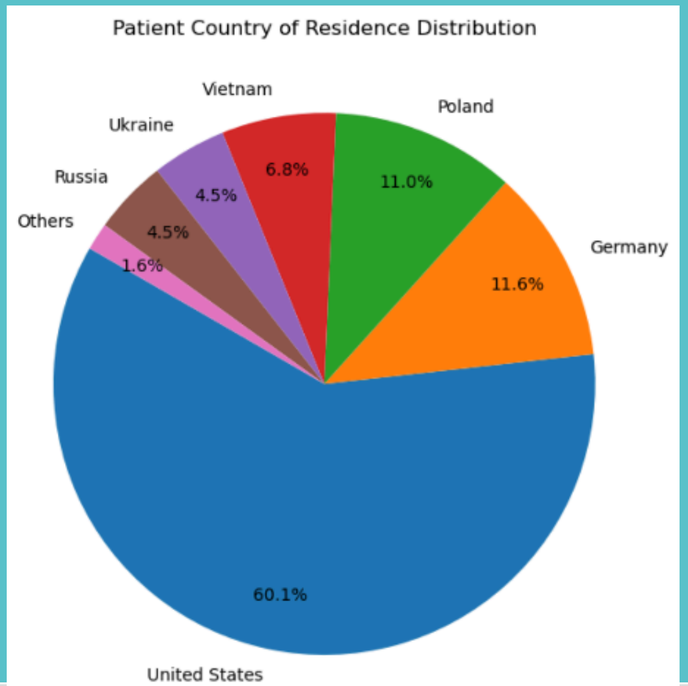
Gene Sample Size: 1231 obs. x 60.660 genes.

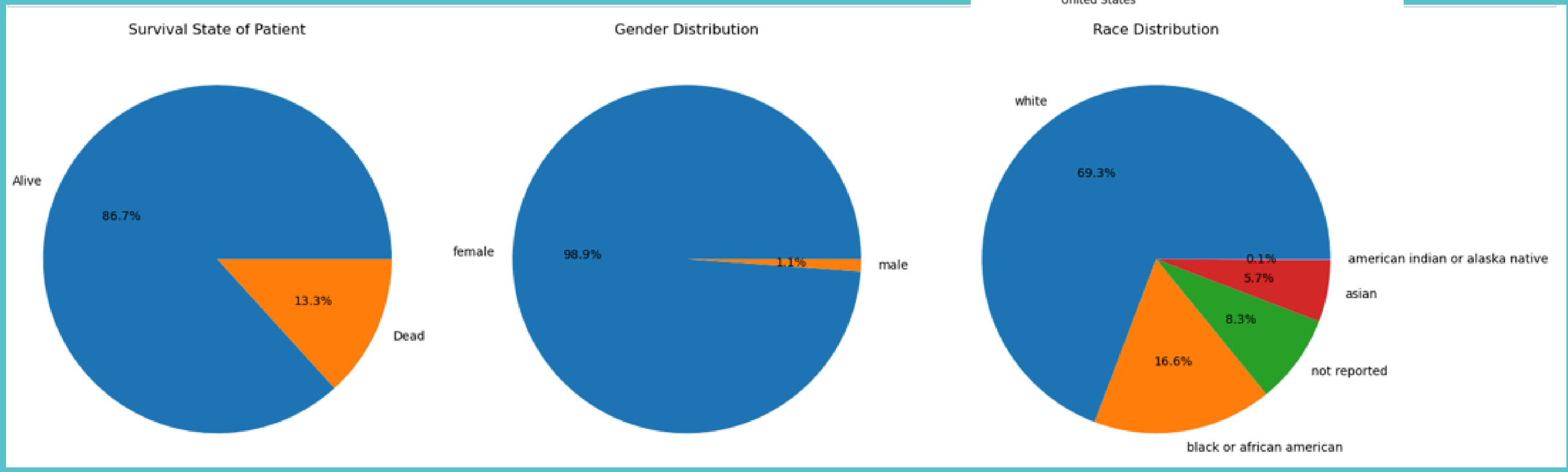
Metadata Collected: 81 features distributed in: Case, Diagnosis, Treatment, Follow Up, Demographic, Sample.

Total Unique Patients: 1098

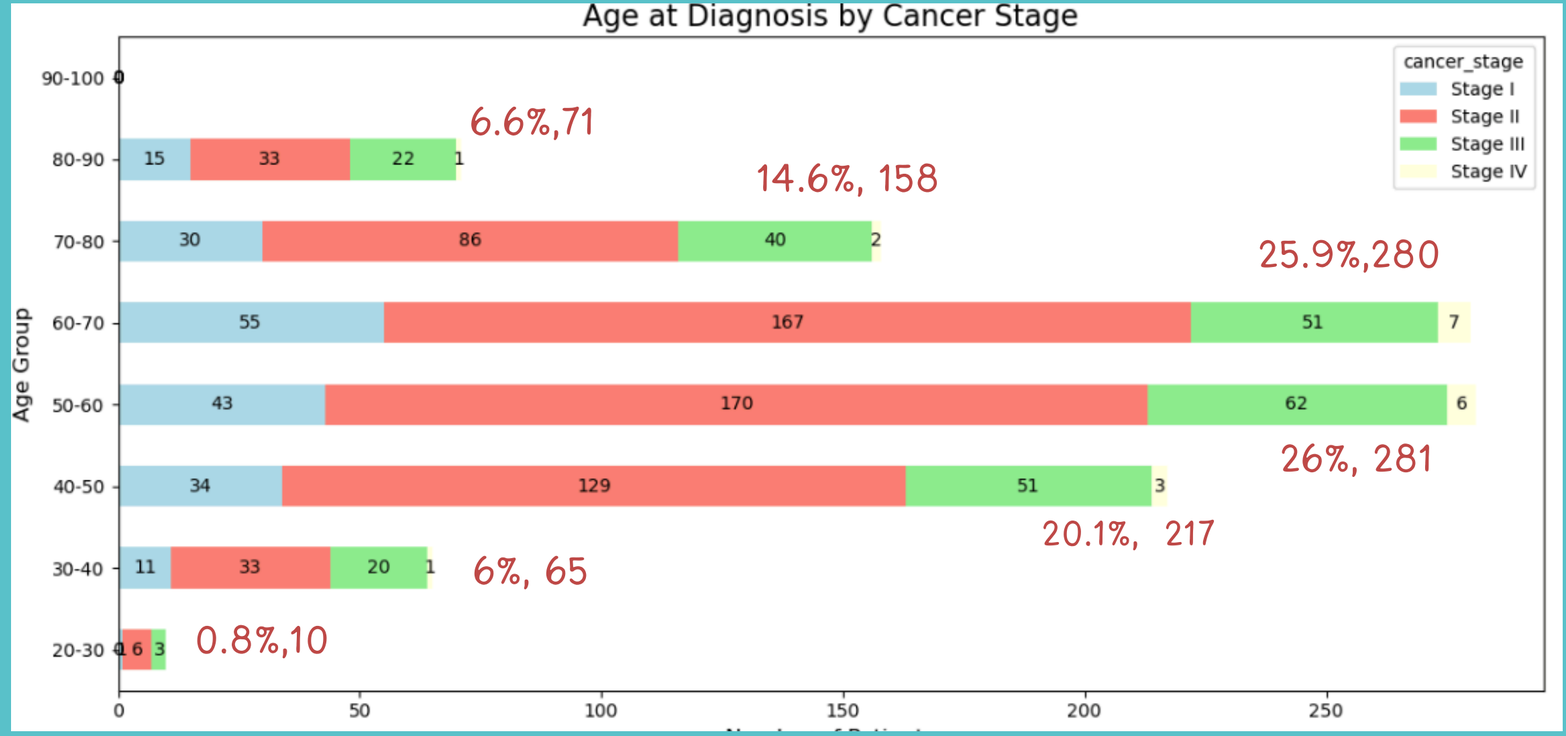
EDA

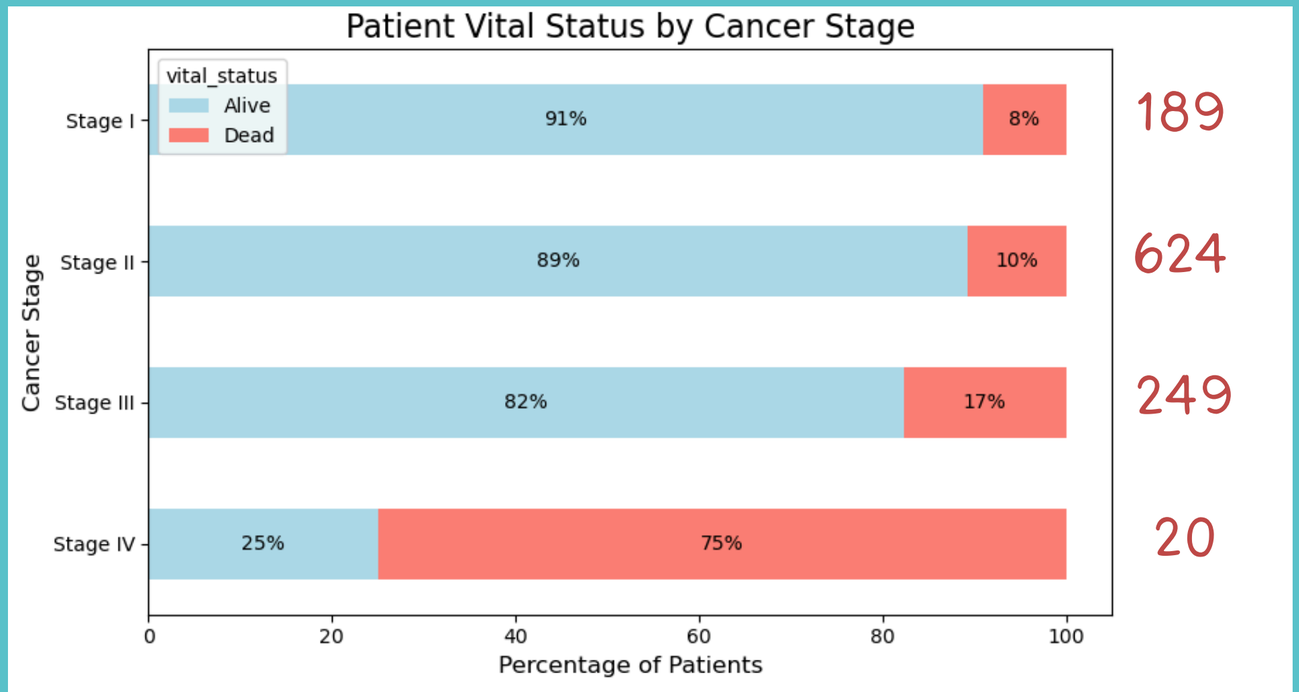
Total Unique Patients: 1098





EDA


Age at Diagnosis by Cancer Stage
cancer_stage
Stage I
Stage II
Stage III
Stage IV
Age Group
90-100
80-90
70-80
60-70
50-60
40-50
30-40
20-30
15
33
22
1
30
86
40
2
55
167
51
7
43
170
62
6
34
129
51
3
11
33
20
1
1
6
3
0
50
100
150
200
250
6.6%,71
14.6%, 158
25.9%,280
26%, 281
20.1%, 217
6%, 65
0.8%,10


Patient Vital Status by Cancer Stage
vital_status
Alive
Dead
Cancer Stage
Stage I
Stage II
Stage III
Stage IV
91%
8%
89%
10%
82%
17%
25%
75%
0
20
40
60
80
100
Percentage of Patients
189
624
249
20

DEA

Differential Expression Analysis (DEA)

What it is: technique that show statistically significant differences in expression levels between two or more conditions

Tools: rpy2 package in Python and DESeq2 tool in R.

Approach

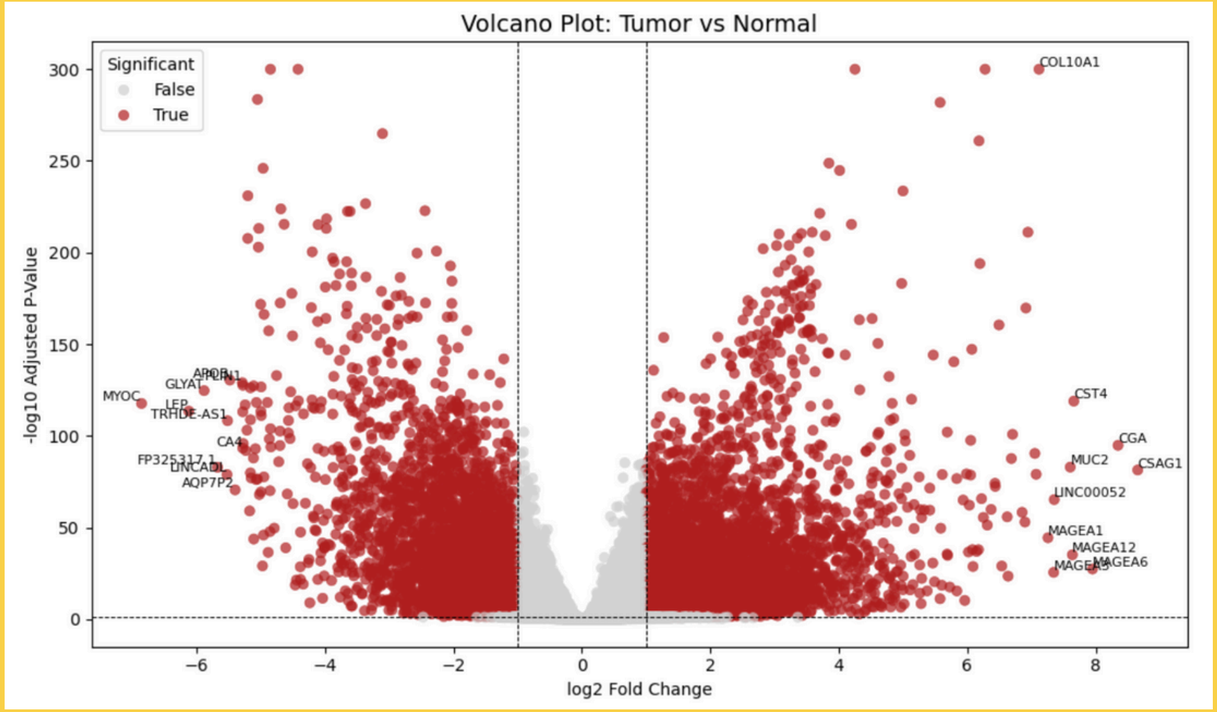
1) Volcano Plots to visualize upregulated or downregulated genes

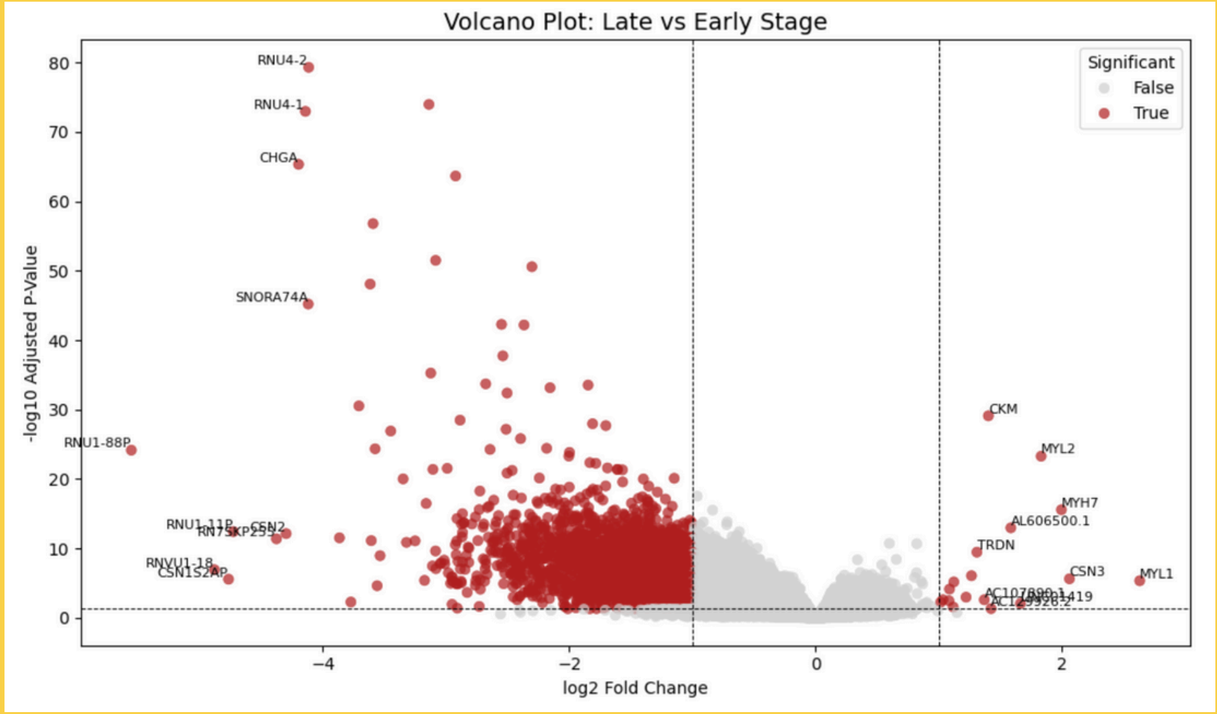
2) Identifying top DEGs

3) Plot heatmpas to identify clusters.

DEA GRAPH

Volcano Plot





Tumor : 1080 samples

Normal :111 samples

Early (Stage 1 & Stage 2):897 samples

Late (Stage 3 & Stage 4) : 294 samples

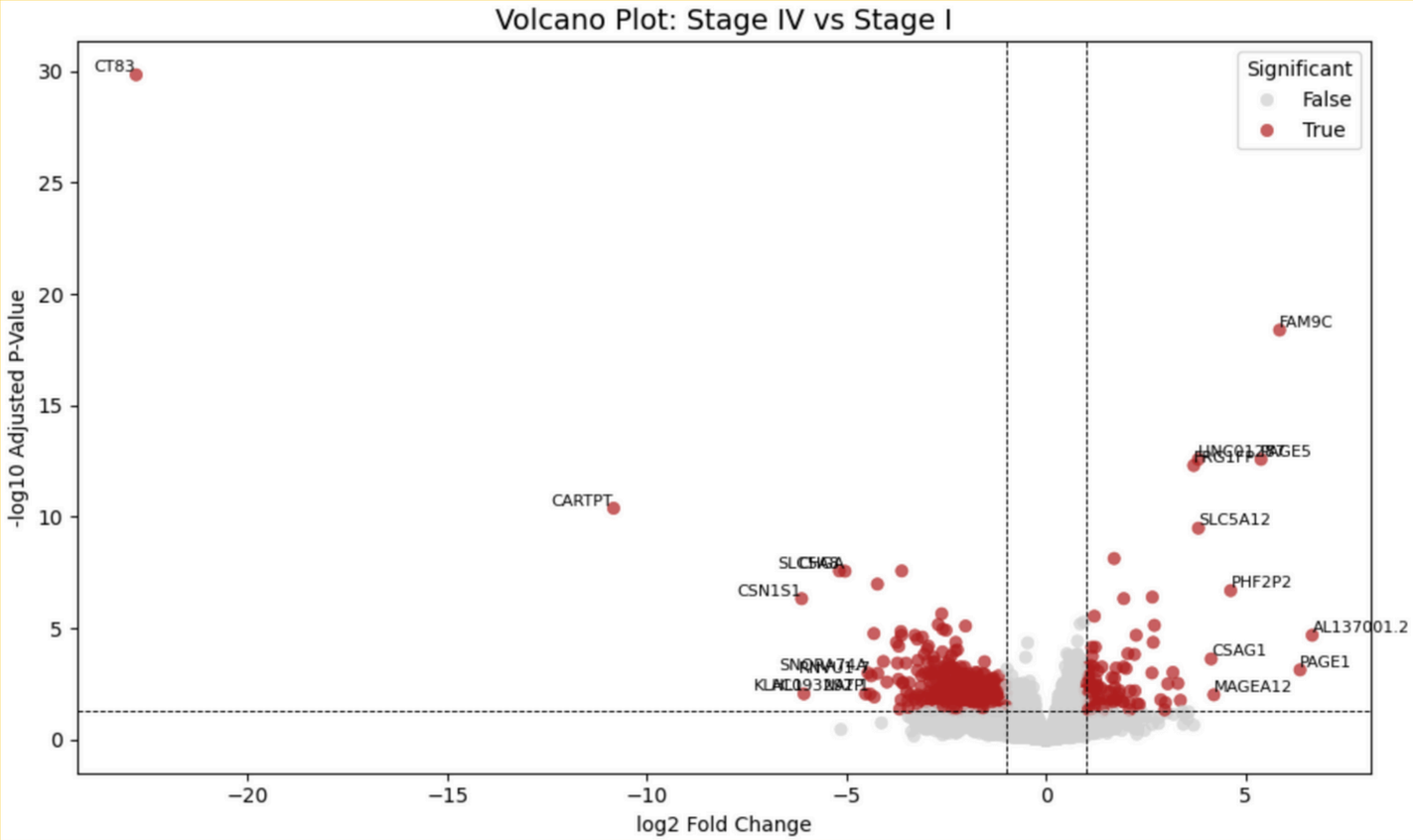
Fold change threshold >1 & P- adjusted < 0.05

Fold change 1 = 2-fold increase

padj < 0.05 = highlighting changes that are statistically reliable

DEA GRAPH

Volcano Plot



Stage I : 204 samples

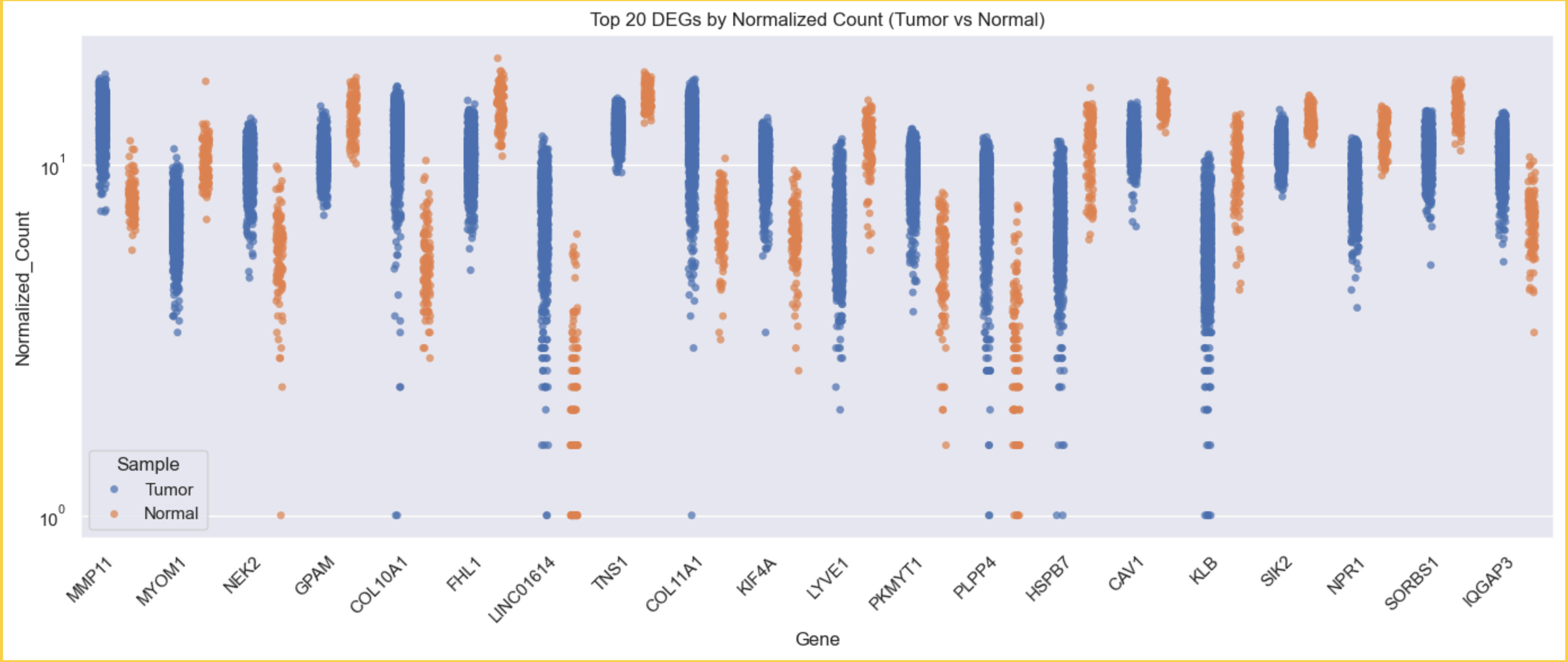
Stage IV : 22 samples

The upregulated genes, like MAGEA12 and PAGE1, are often associated with cancer progression, immune escape, or poor prognosis.

The downregulated genes, like CT83 and CARTPT, could potentially early-stage markers

DEA GRAPH

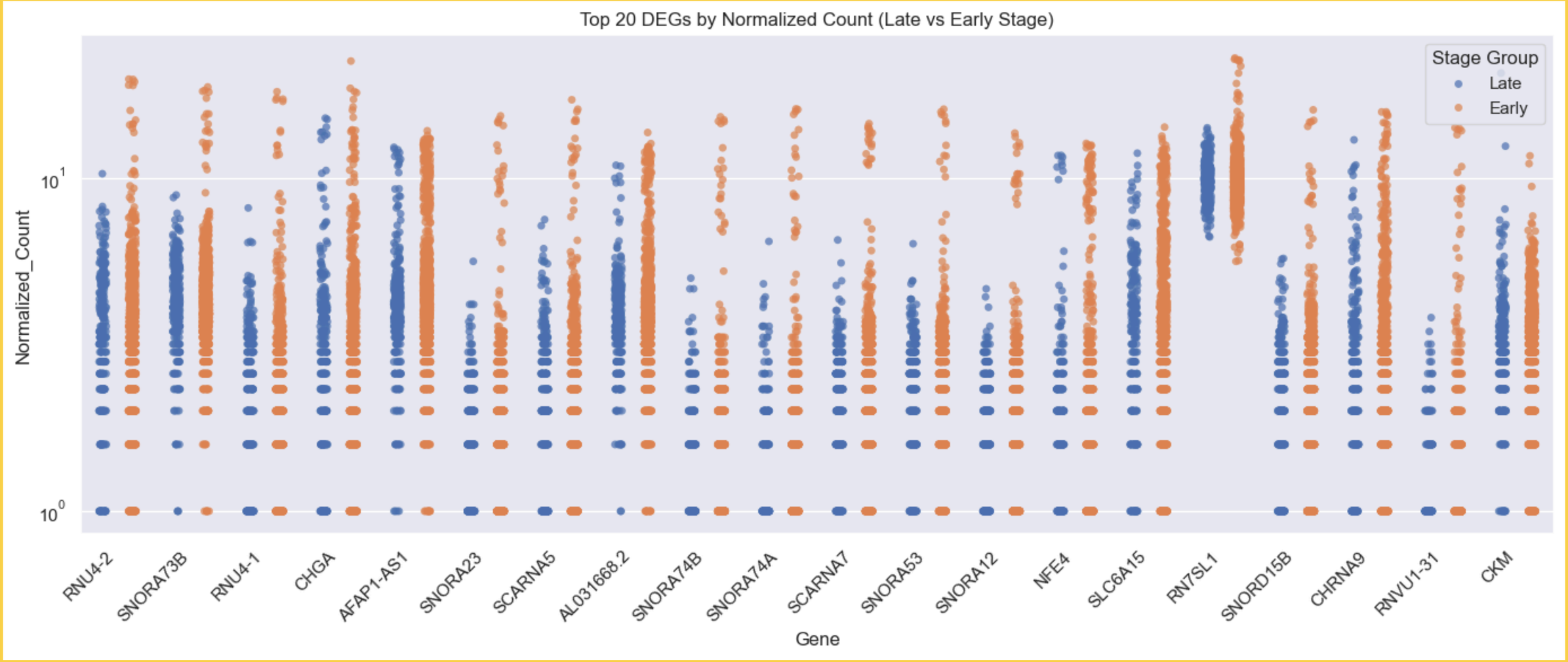
Top 20 DEG



This pattern suggest these gene could serve as potential biomarkers for tumor detection

DEA GRAPH

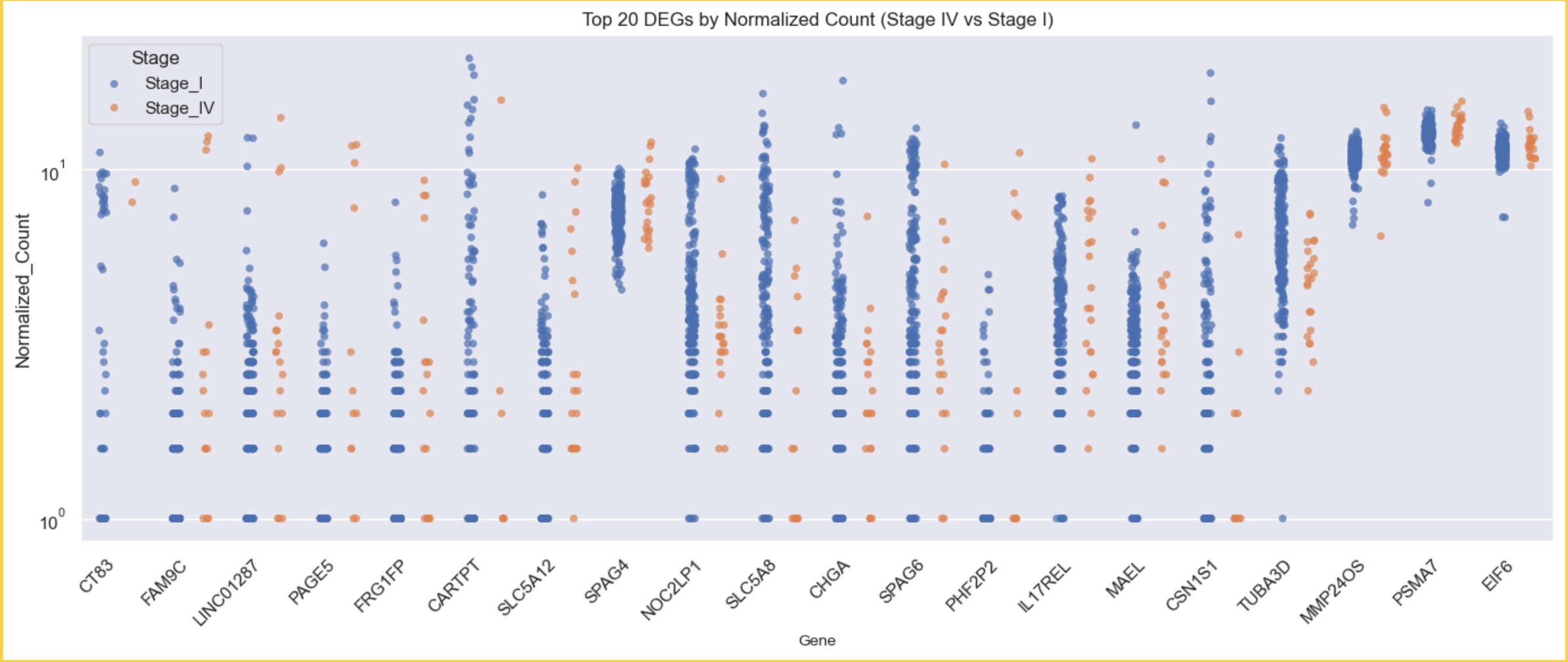
Top 20 DEG



The clear split implies strong stage-specific regulation in the transcriptome

DEA GRAPH

Top 20 DEG



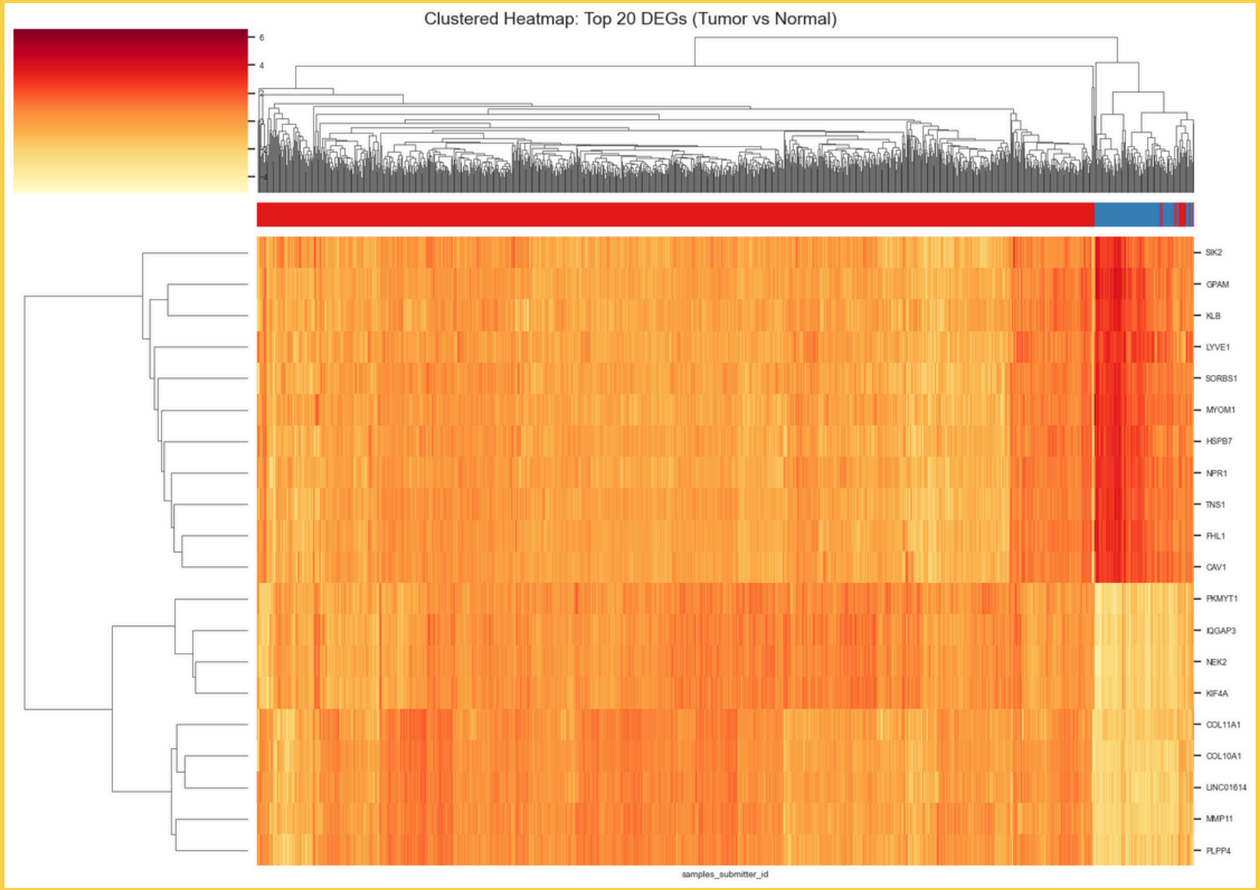
The pattern is more heterogeneous or mixed expression trends, could reflect tumor adaption, therapy resistance or metastasis potential

DEA GRAPH

Heatmap

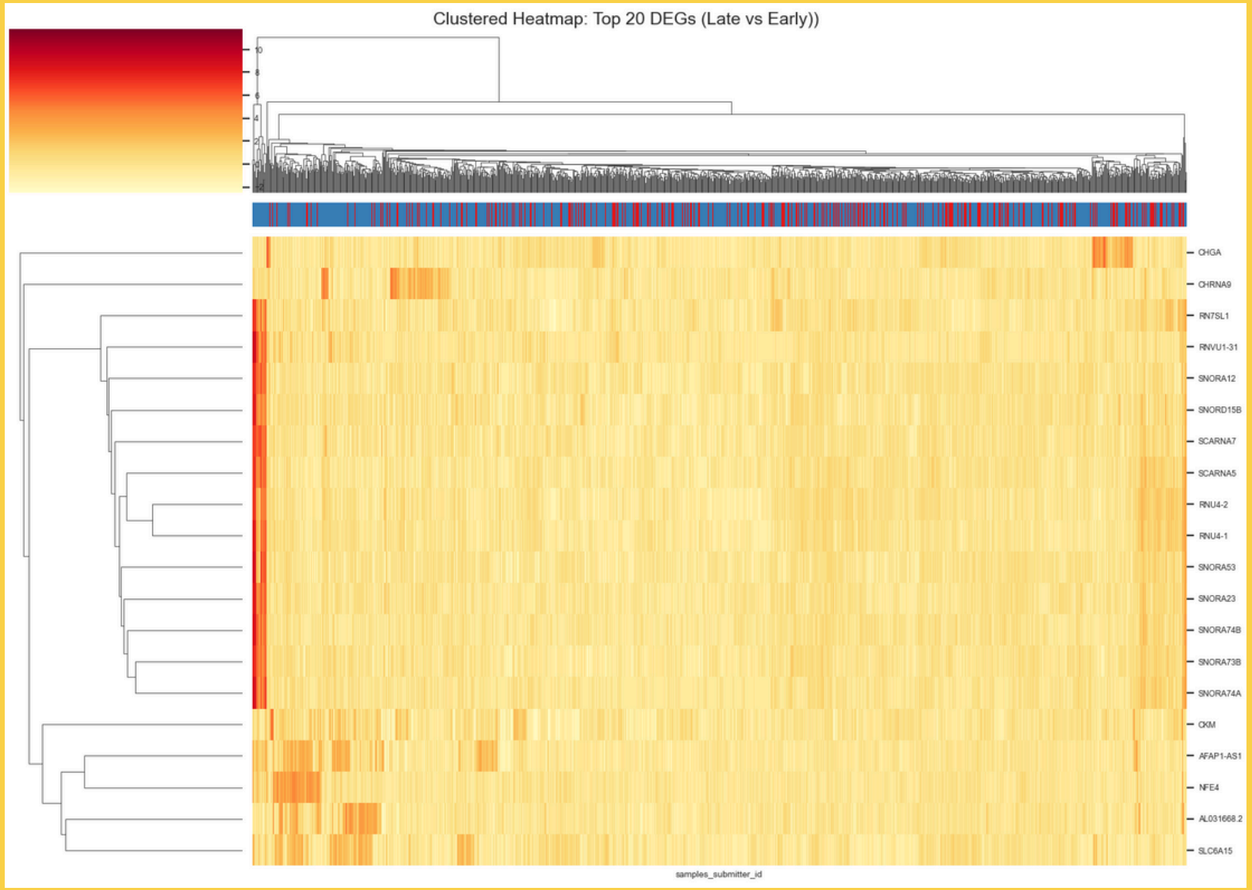
Legend :

1. Rows = Genes (based on expression similarity across samples
2. Columns = Samples (based on similarity accross gene expression profile)



Key Takeaway :

1. The pattern validates that DEG's are biologically meaningful
2. Good clustering = strong candidates for further discovery

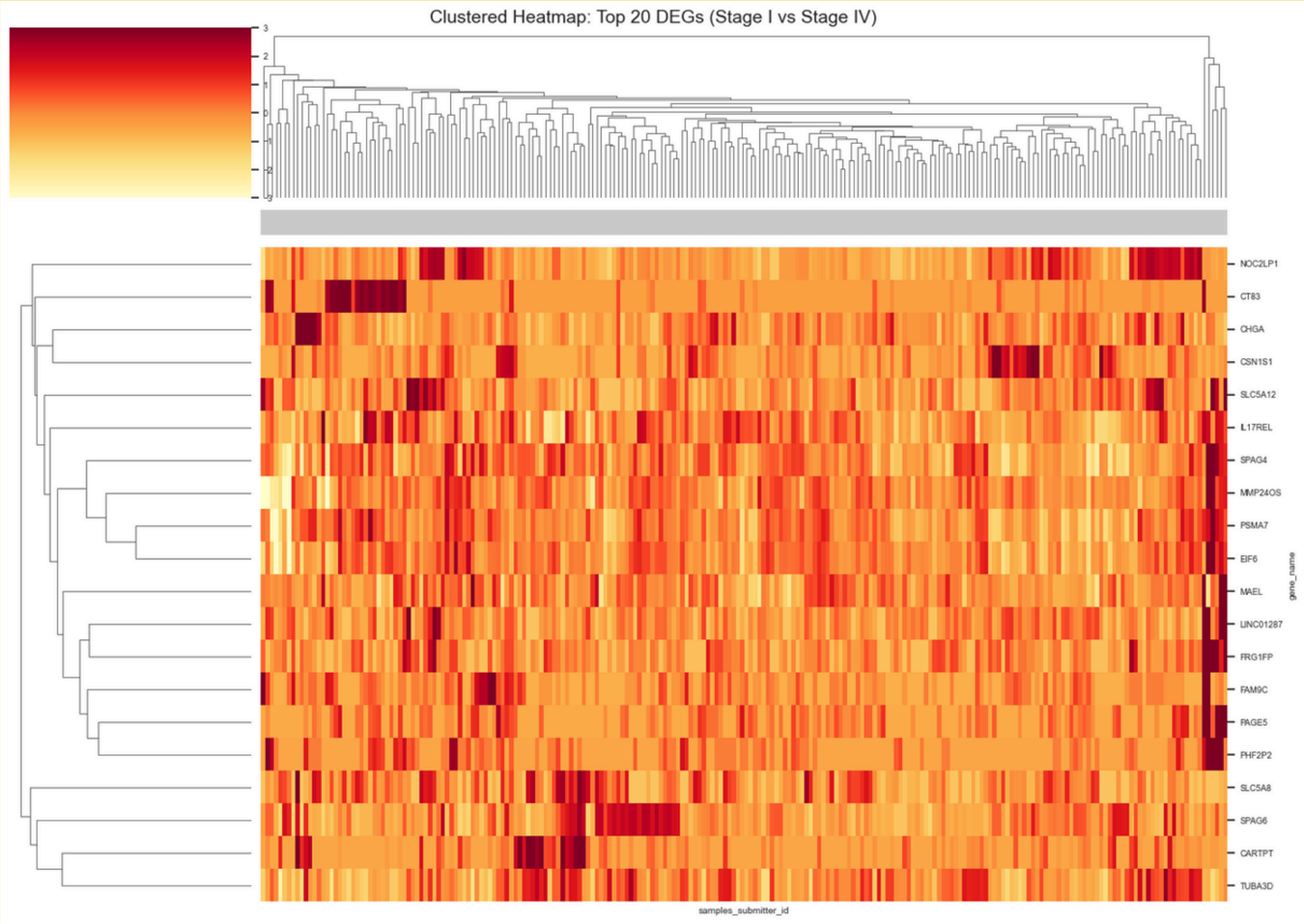


Key Takeaway :

1. Scattered and less obvious separation
2. Stage progression has more subtle transcriptomic changes

DEA GRAPH

Heatmap

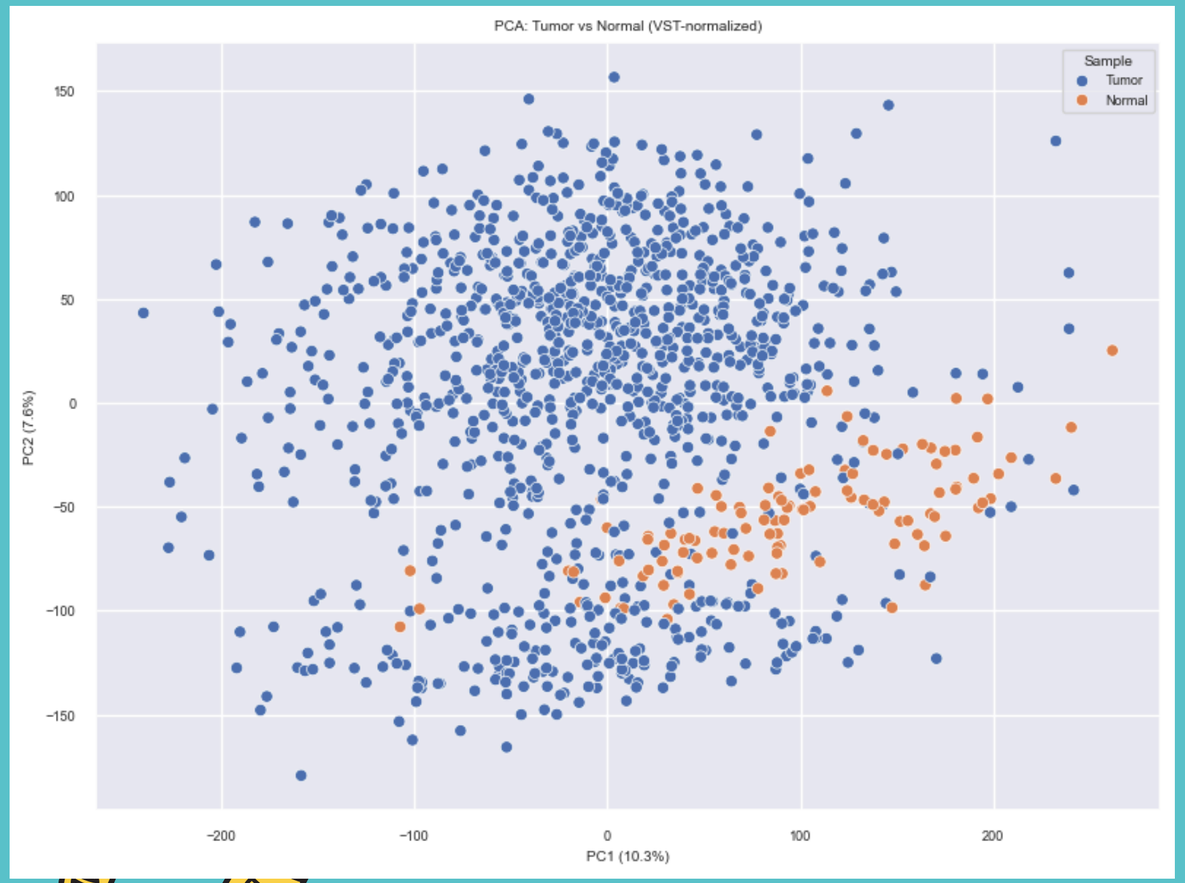


Key Takeaway :

1. Less sharply divided / no absolute distinct on gene expression across stage
2. There may be biological overlap or gradual progression, rather than a binary shift between Stage I and Stage IV.

DIMENSION REDUCTION

PCA



Observations = All genes

Variance = 17.9%

Number of PC : 2 PC's

Observations = Top Genes - DEG Filtered

Variance = 69.22%

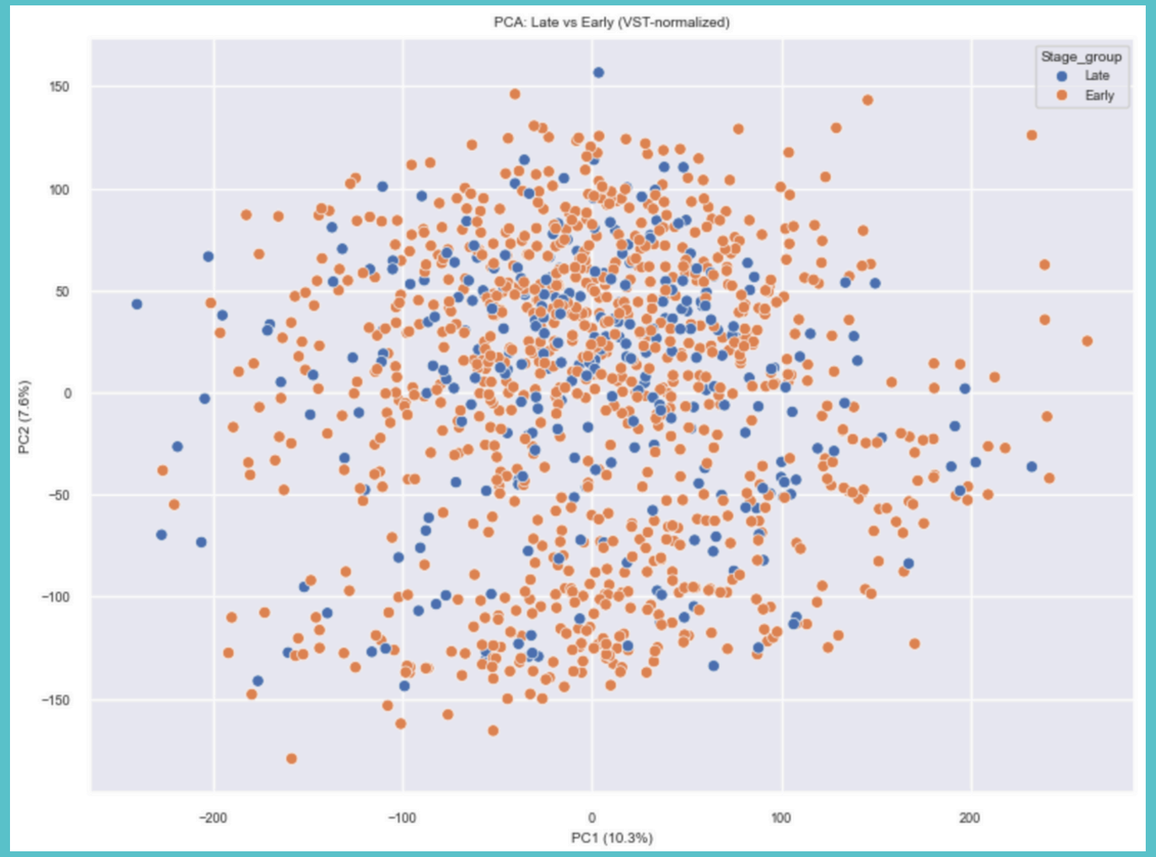
Number of PC : 2 PC's

DIMENSION REDUCTION

PCA

DIMENSION REDUCTION

PCA



Observations = All genes

Variance = 17.9%

Number of PC : 2 PC's

Observations = Top Genes - DEG Filtered

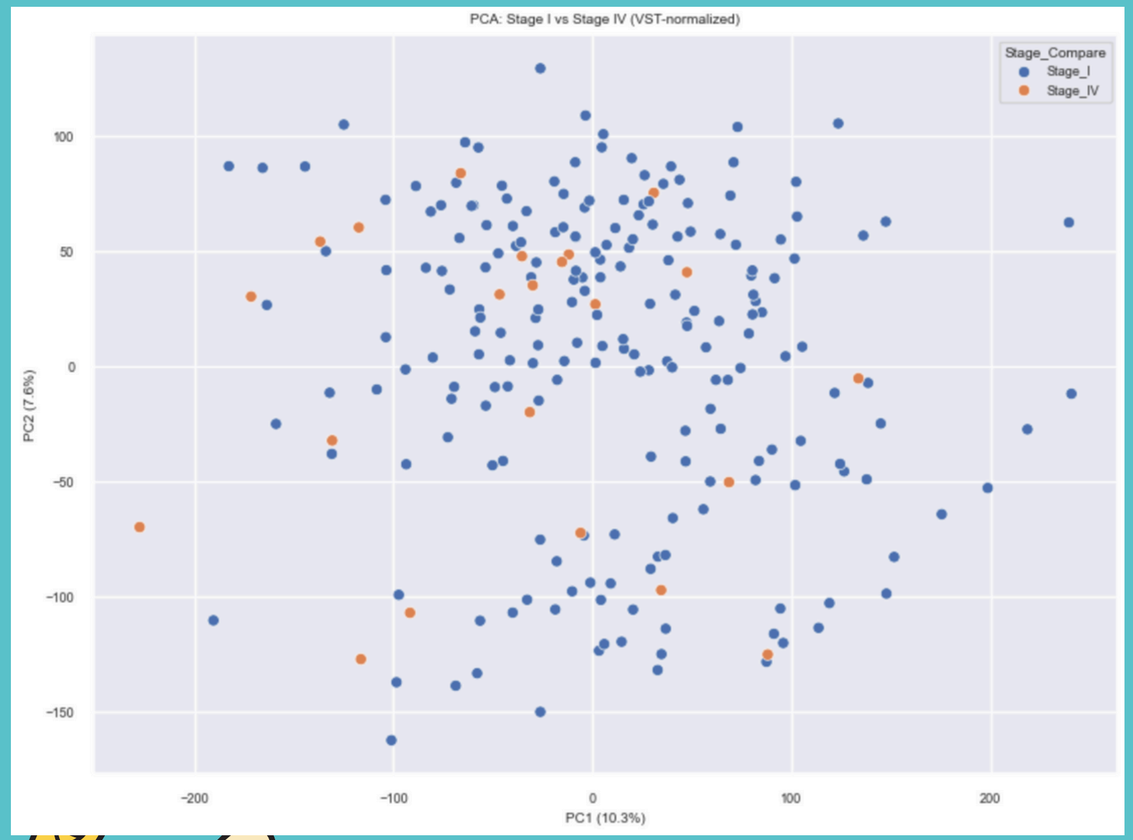
Variance = 45.64%

Number of PC : 2 PC's

The gene expression differences between early and late cancer stages may not be strong or linear enough for clear classification in this reduced-dimensional space.

DIMENSION REDUCTION

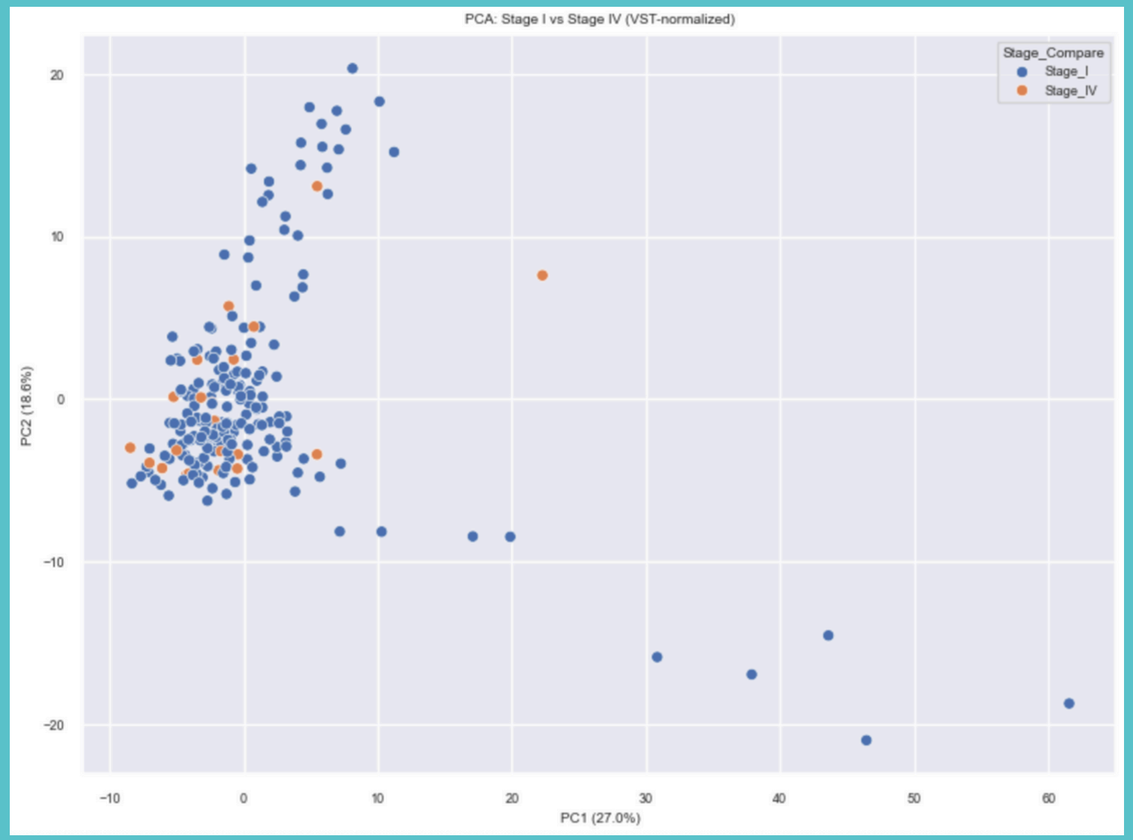
PCA



Observations = All genes

Variance = 17.9%

Number of PC : 2 PC's



Observations = Top Genes - DEG Filtered

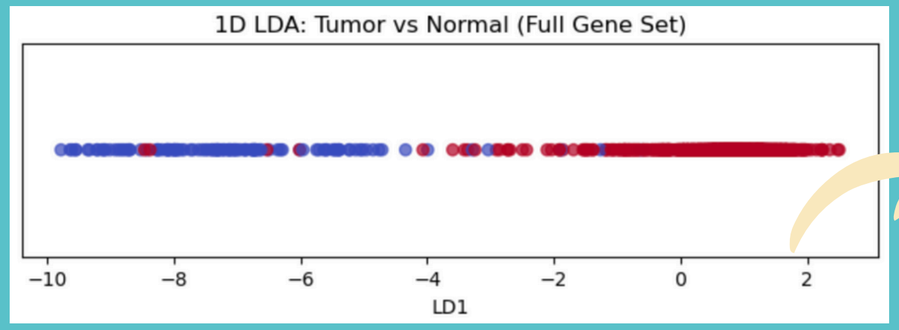
Variance = 45.64%

Number of PC : 2 PC's

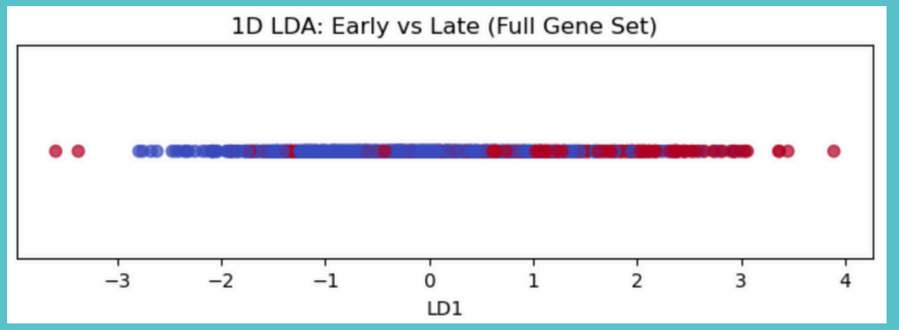
The lack of separation is primarily due to sample imbalance—Stage I samples far outnumber Stage IV.
PCA after DEG filtering does show mild improvement in group differentiation.

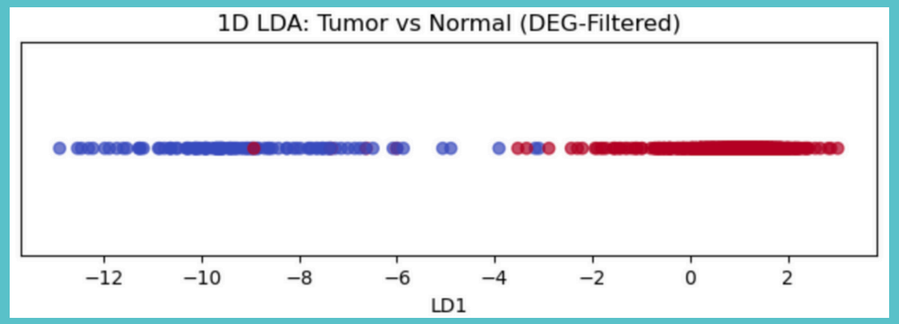
DIMENSION REDUCTION

LDA

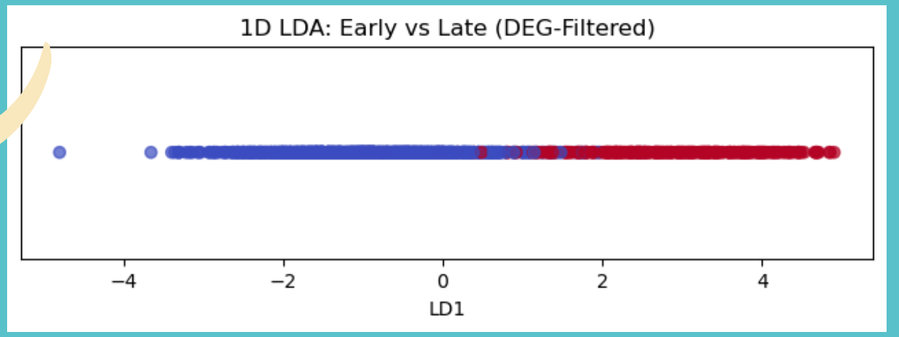


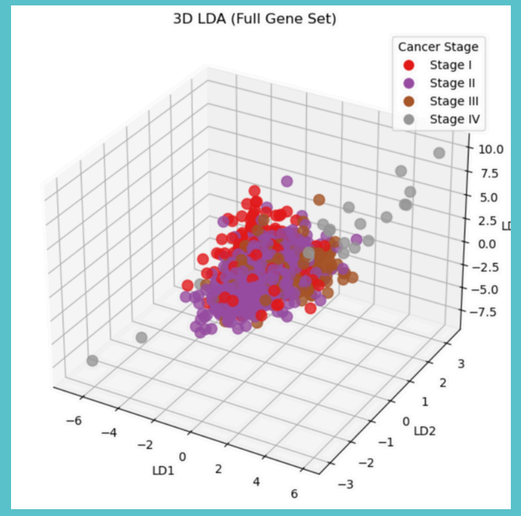
Using DEG-filtered set improved class separation (reduced overlap and clearer clustering) between Tumor and Normal samples in the LDA projection.

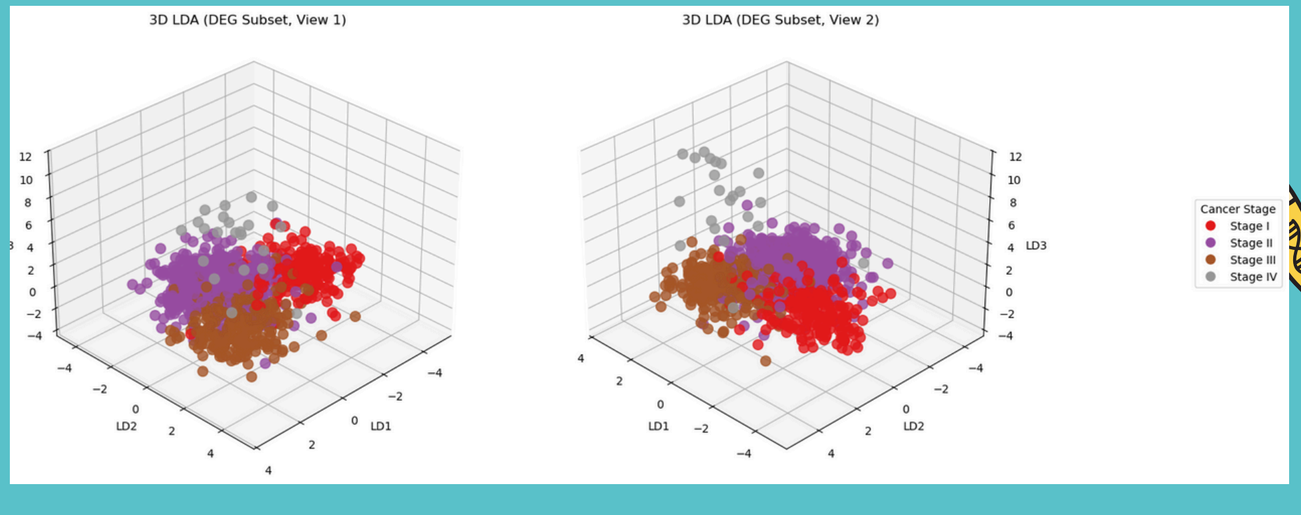




This suggests that the DEGs identified via differential expression analysis offer stronger biological signals for staging compared to the full gene set, which includes noisy or less relevant features.







MODEL BUILDING

Performance Comparison:

2 Datasets (Full & Reduced) + 4 Dimensionality Reduction + 3 Classifiers

Pipeline:

Full Gene Set Pipeline (Baseline)

> **Raw RNA-seq counts → VST/log2 normalization → PCA/LDA/NCA/PCoA → Classifier**

DEG-Filtered Gene Set Pipeline

> **Raw RNA-seq counts → DEG filtering (DESeq2) → VST/log2 normalization → PCA/LDA/NCA/PCoA → Classifier**

MODEL BUILDING

(1) Normal vs. Tumor Classification: Benchmarking

Model	Accuracy	Precision	Recall	F1-score	Confusion Matrix
PCA + RF on full gene set	0.979079	0.930153	0.947633	0.938671	[[20, 2], [3, 214]]
PCA + RF on DEG-filtered gene set	0.987448	0.954207	0.972664	0.963202	[[21, 1], [2, 215]]
PCA + XGB on full gene set	0.987448	0.954207	0.972664	0.963202	[[21, 1], [2, 215]]
PCA + XGB on DEG-filtered gene set	0.987448	0.954207	0.972664	0.963202	[[21, 1], [2, 215]]
PCA + LogReg on full gene set	0.987448	0.940000	0.993088	0.964605	[[22, 0], [3, 214]]
PCA + LogReg on DEG-filtered gene set	0.987448	0.940000	0.993088	0.964605	[[22, 0], [3, 214]]
LDA + RF on full gene set	**0.991632**	0.958333	**0.995392**	**0.975946**	[[22, 0], [2, 215]]
LDA + RF on DEG-filtered gene set	0.983264	0.923077	0.990783	0.953682	[[22, 0], [4, 213]]
LDA + XGB on full gene set	0.987448	0.940000	0.993088	0.964605	[[22, 0], [3, 214]]
LDA + XGB on DEG-filtered gene set	0.983264	0.923077	0.990783	0.953682	[[22, 0], [4, 213]]
LDA + LogReg on full gene set	0.983264	0.923077	0.990783	0.953682	[[22, 0], [4, 213]]
LDA + LogReg on DEG-filtered gene set	0.983264	0.923077	0.990783	0.953682	[[22, 0], [4, 213]]

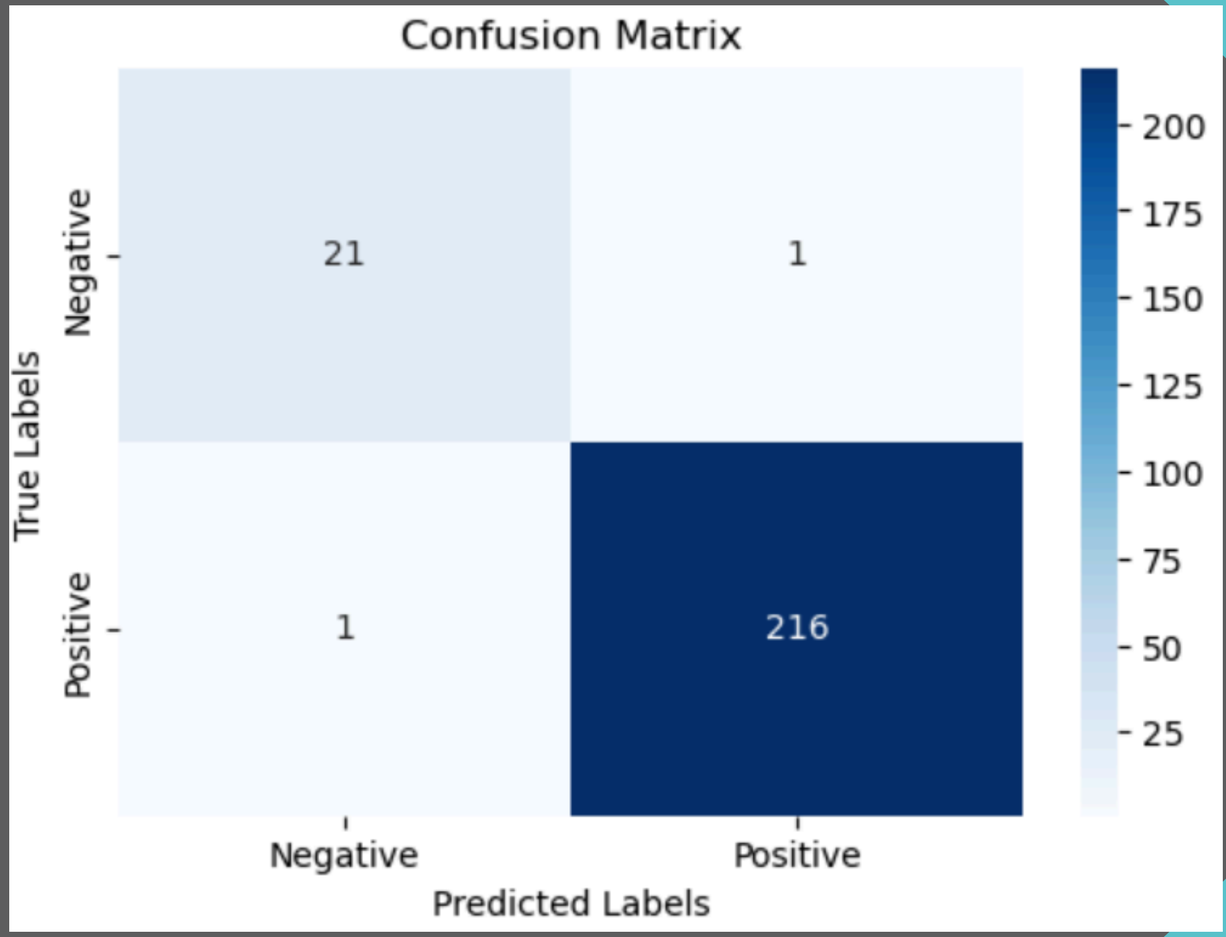
Model	Accuracy	Precision	Recall	F1-score	Confusion Matrix
NCA + RF on full gene set	0.983264	0.949937	0.949937	0.949937	[[20, 2], [2, 215]]
NCA + RF on DEG-filtered gene set	0.987448	0.954207	0.972664	0.963202	[[21, 1], [2, 215]]
NCA + XGB on full gene set	0.987448	0.954207	0.972664	0.963202	[[21, 1], [2, 215]]
NCA + XGB on DEG-filtered gene set	**0.991632**	**0.974969**	0.974969	0.974969	[[21, 1], [1, 216]]
NCA + LogReg on full gene set	0.987448	0.940000	0.993088	0.964605	[[22, 0], [3, 214]]
NCA + LogReg on DEG-filtered gene set	**0.991632**	0.958333	**0.995392**	**0.975946**	[[22, 0], [2, 215]]
PCoA + RF on full gene set	0.899582	0.453586	0.495392	0.473568	[[0, 22], [2, 215]]
PCoA + RF on DEG-filtered gene set	0.895397	0.453390	0.493088	0.472406	[[0, 22], [3, 214]]
PCoA + XGB on full gene set	0.866109	0.451965	0.476959	0.464126	[[0, 22], [10, 207]]
PCoA + XGB on DEG-filtered gene set	0.887029	0.452991	0.488479	0.470067	[[0, 22], [5, 212]]
PCoA + LogReg on full gene set	0.786611	0.447619	0.433180	0.440281	[[0, 22], [29, 188]]
PCoA + LogReg on DEG-filtered gene set	0.811715	0.449074	0.447005	0.448037	[[0, 22], [23, 194]]

(1) Normal vs. Tumor Classification: Top Performer

NCA + XGBoost + DEG

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	22
1	1.00	1.00	1.00	217
accuracy			0.99	239
macro avg	0.97	0.97	0.97	239
weighted avg	0.99	0.99	0.99	239



MODEL BUILDING

(2) Early vs. Late Stage Classification: Benchmarking

Model	Accuracy	Precision	Recall	F1-score	Confusion Matrix
PCA + RF on full gene set	0.753138	0.376569	0.500000	0.429594	[[180, 0], [59, 0]]
PCA + RF on DEG-filtered gene set	**0.774059**	**0.751794**	0.559463	0.548109	[[177, 3], [51, 8]]
PCA + XGB on full gene set	0.753138	0.376569	0.500000	0.429594	[[180, 0], [59, 0]]
PCA + XGB on DEG-filtered gene set	0.757322	0.652093	0.554049	0.546282	[[172, 8], [50, 9]]
PCA + LogReg on full gene set	0.694561	0.552453	0.540866	0.541797	[[152, 28], [45, 14]]
PCA + LogReg on DEG-filtered gene set	0.673640	0.569490	0.572552	0.570777	[[139, 41], [37, 22]]
LDA + RF on full gene set	0.682008	0.550107	0.543927	0.545445	[[147, 33], [43, 16]]
LDA + RF on DEG-filtered gene set	0.686192	0.558889	0.552401	0.554337	[[147, 33], [42, 17]]
LDA + XGB on full gene set	0.748954	0.608852	0.525706	0.497899	[[174, 6], [54, 5]]
LDA + XGB on DEG-filtered gene set	0.711297	0.585490	0.569068	**0.573049**	[[153, 27], [42, 17]]
LDA + LogReg on full gene set	0.623431	0.542524	0.550612	0.541560	[[125, 55], [35, 24]]
LDA + LogReg on DEG-filtered gene set	0.640167	0.557399	0.567420	0.558093	[[128, 52], [34, 25]]

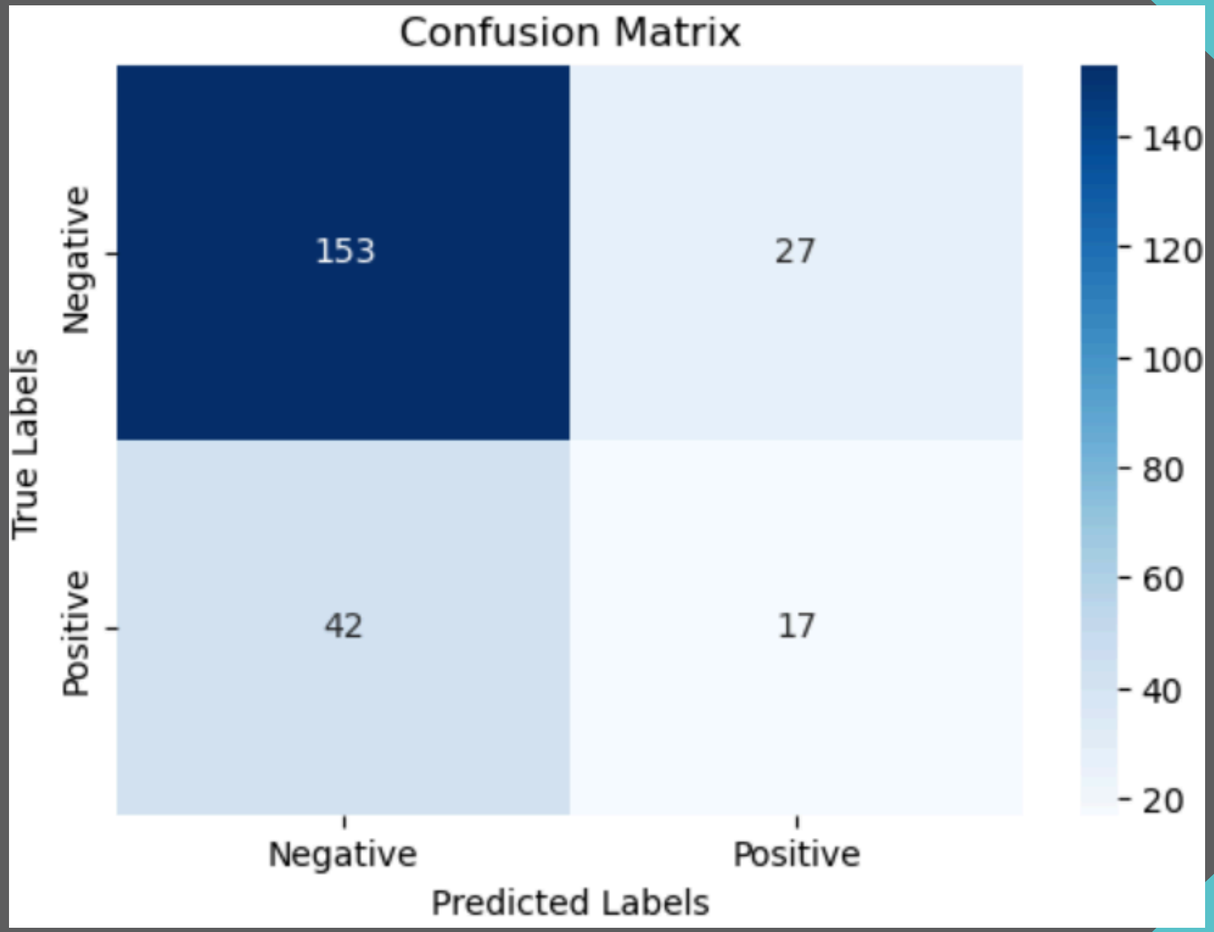
Model	Accuracy	Precision	Recall	F1-score	Confusion Matrix
NCA + RF on full gene set	0.728033	0.475546	0.494727	0.449523	[[172, 8], [57, 2]]
NCA + RF on DEG-filtered gene set	0.740586	0.544928	0.508757	0.468508	[[174, 6], [56, 3]]
NCA + XGB on full gene set	0.753138	0.376569	0.500000	0.429594	[[180, 0], [59, 0]]
NCA + XGB on DEG-filtered gene set	0.744770	0.375527	0.494444	0.426859	[[178, 2], [59, 0]]
NCA + LogReg on full gene set	0.564854	0.555400	**0.574388**	0.530097	[[100, 80], [24, 35]]
NCA + LogReg on DEG-filtered gene set	0.539749	0.538690	0.552024	0.507678	[[95, 85], [25, 34]]
PCoA + RF on full gene set	0.753138	0.376569	0.500000	0.429594	[[180, 0], [59, 0]]
PCoA + RF on DEG-filtered gene set	0.748954	0.543785	0.502919	0.444014	[[178, 2], [58, 1]]
PCoA + XGB on full gene set	0.753138	0.376569	0.500000	0.429594	[[180, 0], [59, 0]]
PCoA + XGB on DEG-filtered gene set	0.732218	0.446429	0.491808	0.437482	[[174, 6], [58, 1]]
PCoA + LogReg on full gene set	0.518828	0.463133	0.452684	0.447260	[[105, 75], [40, 19]]
PCoA + LogReg on DEG-filtered gene set	0.556485	0.491311	0.489077	0.481286	[[112, 68], [38, 21]]

(2) Early vs. Late Stage Classification: Top Performer

LDA + XGBoost + DEG

Classification Report:

	precision	recall	f1-score	support
0	0.78	0.85	0.82	180
1	0.39	0.29	0.33	59
accuracy			0.71	239
macro avg	0.59	0.57	0.57	239
weighted avg	0.69	0.71	0.70	239



MODEL BUILDING

(3) Stage I to IV Classification: Benchmarking

Model	Accuracy	Precision	Recall	F1-score
PCA + RF on full gene set	0.581590	0.145397	0.250000	0.183862
PCA + RF on DEG-filtered gene set	**0.589958**	0.396624	0.259091	0.202384
PCA + XGB on full gene set	0.560669	0.227438	0.246501	0.195741
PCA + XGB on DEG-filtered gene set	0.552301	0.325533	0.252697	0.216399
PCA + LogReg on full gene set	0.510460	0.307295	0.299676	0.301224
PCA + LogReg on DEG-filtered gene set	0.460251	0.317548	0.332194	0.317244
LDA + RF on full gene set	0.552301	0.355230	0.318499	0.322705
LDA + RF on DEG-filtered gene set	0.485356	0.375596	0.350781	0.359502
LDA + XGB on full gene set	0.569038	0.355676	0.289269	0.280703
LDA + XGB on DEG-filtered gene set	0.502092	**0.427189**	0.365021	**0.384566**
LDA + LogReg on full gene set	0.443515	0.295378	0.305414	0.297183
LDA + LogReg on DEG-filtered gene set	0.405858	0.347581	0.351481	0.339937

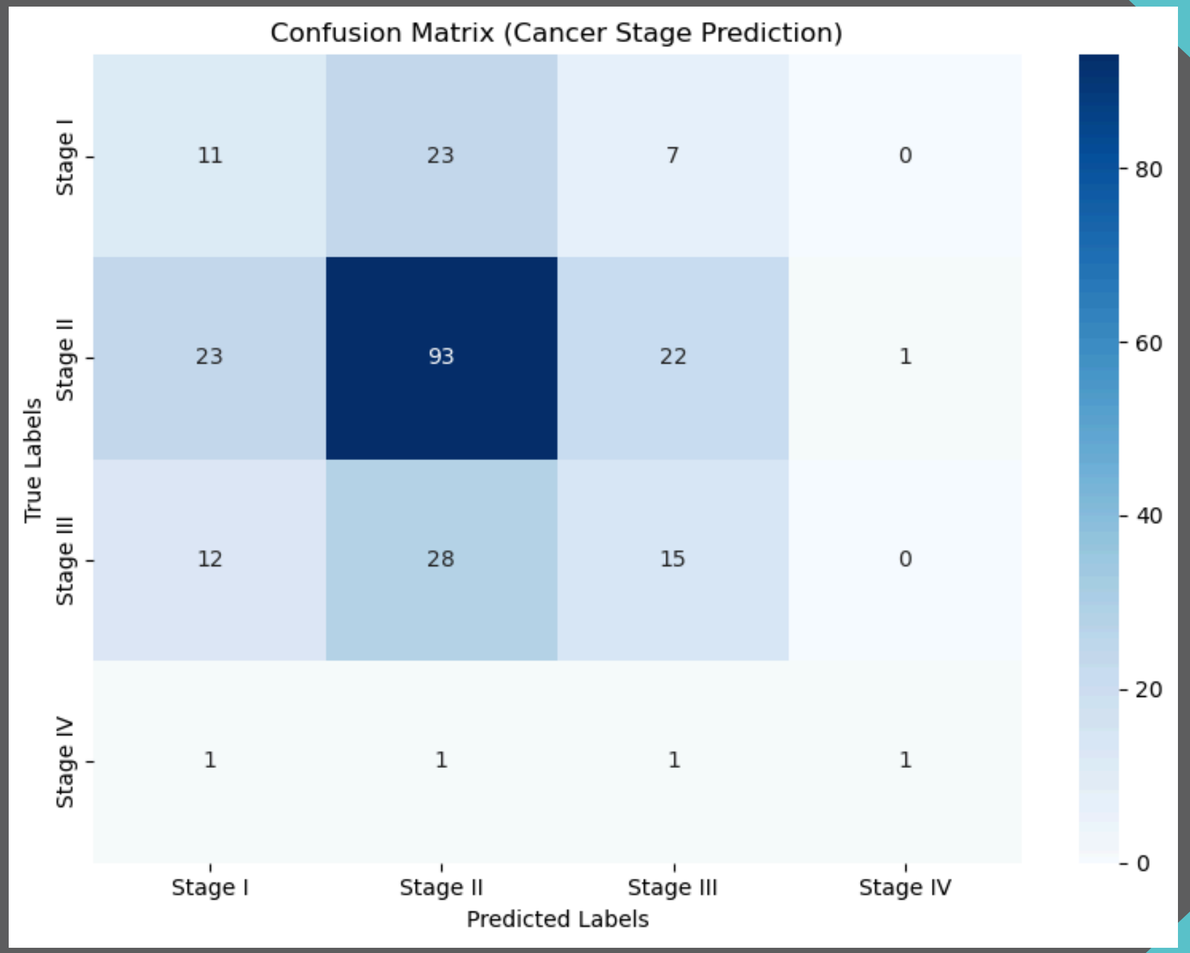
Model	Accuracy	Precision	Recall	F1-score
NCA + RF on full gene set	0.577406	0.223026	0.256442	0.207001
NCA + RF on DEG-filtered gene set	0.560669	0.283301	0.253547	0.213158
NCA + XGB on full gene set	0.581590	0.209936	0.252747	0.193461
NCA + XGB on DEG-filtered gene set	0.577406	0.144958	0.248201	0.183024
NCA + LogReg on full gene set	0.280335	0.286440	0.298953	0.238857
NCA + LogReg on DEG-filtered gene set	0.238494	0.252667	0.276311	0.212630
PCoA + RF on full gene set	0.577406	0.144958	0.248201	0.183024
PCoA + RF on DEG-filtered gene set	0.577406	0.144958	0.248201	0.183024
PCoA + XGB on full gene set	0.573222	0.346476	0.258943	0.216399
PCoA + XGB on DEG-filtered gene set	0.577406	0.144958	0.248201	0.183024
PCoA + LogReg on full gene set	0.263598	0.283793	**0.417818**	0.231312
PCoA + LogReg on DEG-filtered gene set	0.142259	0.211213	0.120622	0.127239

(3) Stage I to IV Classification: Top Performer

LDA + XGBoost + DEG

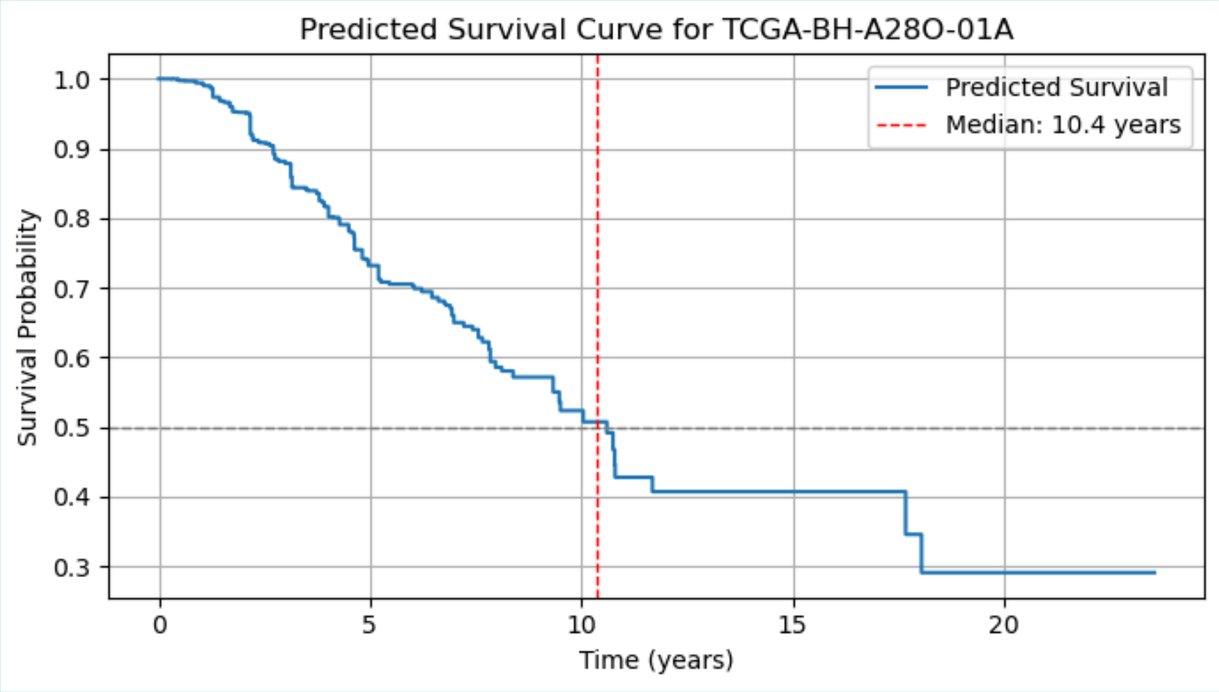
Classification Report:

	precision	recall	f1-score	support
0	0.23	0.27	0.25	41
1	0.64	0.67	0.65	139
2	0.33	0.27	0.30	55
3	0.50	0.25	0.33	4
accuracy			0.50	239
macro avg	0.43	0.37	0.38	239
weighted avg	0.50	0.50	0.50	239



MODEL BUILDING

(4) Survival Analysis Using Random Survival Forest



Survival Prediction Summary

==========

Sample ID: TCGA-BH-A28O-01A

C-index (Concordance Index): 0.571

Median Predicted Survival Time: 3785 days (~10.4 years)

Interpretation:

- The C-index of 0.571 reflects the model's ability to rank patients by survival risk. A value closer to 1.0 indicates stronger predictive performance.
- A median survival time of 3785 days suggests that 50% of similar patients are expected to survive longer than approximately 10.4 years.

CONCLUSIONS

- Tumor vs. Normal classification achieved the highest performance, with an overall accuracy of 99%, using the NCA + XGBoost model on the DEG-filtered gene set (11,883 genes). This model will be confidently deployed in our Streamlit application.

- Early vs. Late Stage classification using the LDA + XGBoost model on the DEG-filtered gene set (2,634 genes) reached 71% accuracy, with Precision, Recall, and F1-score around 57%. While promising, this model should be used with caution, and predictions should be presented alongside confidence levels in the application.

- Stage I vs. Stage II vs. Stage III vs. Stage IV classification showed the weakest performance using LDA + XGBoost on the DEG-filtered gene set (2,077 genes). Difficulty in separating intermediate stages suggests that further refinement is needed, potentially by incorporating clinical or histopathological data to enhance resolution.

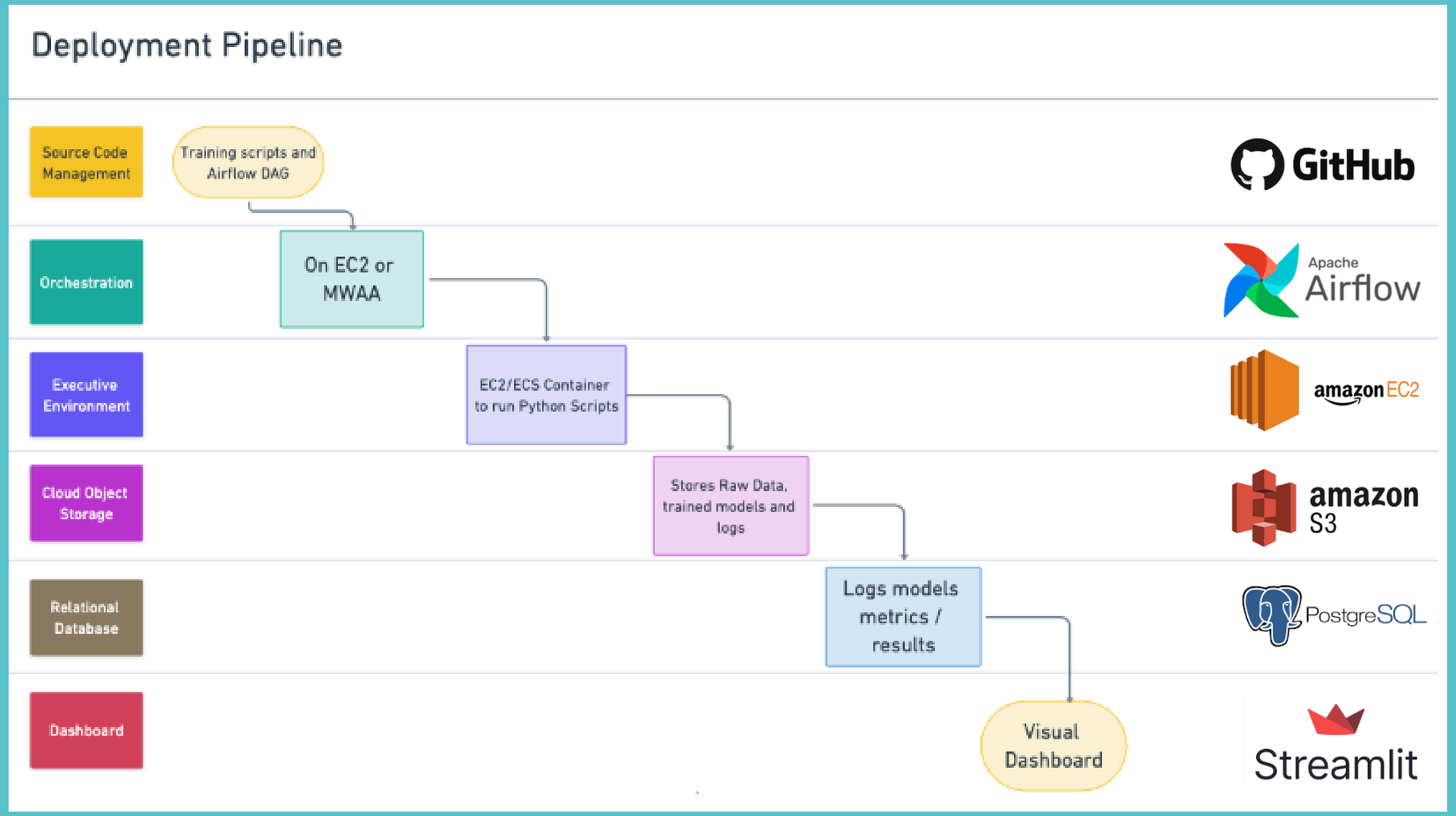
CONCLUSIONS

- Across all three classification tasks, top-performing models consistently appeared in the DEG-filtered set, while models trained on the full gene set produced fewer standout results—highlighting the effectiveness of this feature selection strategy.

- XGBoost classifier consistently delivered the best results, and supervised dimensionality reduction techniques such as NCA and LDA outperformed unsupervised methods, highlighting the advantage of label-aware feature selection for gene expression data.

- Survival Prediction: Using Random Survival Forests on the DEG-filtered gene set, we achieved a C-index of 0.571, indicating moderate predictive performance. While not highly discriminative, the model is still informative for generating individual survival curves and estimating median survival times, which we plan to include in our Streamlit app for exploratory purposes.

DEPLOYMENT


Deployment Pipeline
Source Code Management
Training scripts and Airflow DAG
GitHub
Orchestration
On EC2 or MWAA
Apache Airflow
Executive Environment
EC2/ECS Container to run Python Scripts
amazon EC2
Cloud Object Storage
Stores Raw Data, trained models and logs
amazon S3
Relational Database
Logs models metrics / results
PostgreSQL
Dashboard
Visual Dashboard
Streamlit

NEXT STEPS

- Improve model performance, particularly:
  - Early vs. Late Stage Classification - increase accuracy, recall, etc.
  - Stage I vs. II vs. III vs. IV Classification - increase accuracy, recall, etc.
  - Random Survival Forest - increase C-index from 0.571 to near 1

- Enhance the dashboard to include more features:
  - Risk Analysis

- Explore alternative feature selection methods to reduce even more the amount of genes used to identify express genes in cancer diagnosis.