Identification of Prostate Cancer Recurrence Signatures Using Unsupervised Machine Learning

STAR-PREP
Science Training for Advancing biomedical Research

University of Maryland School of Medicine

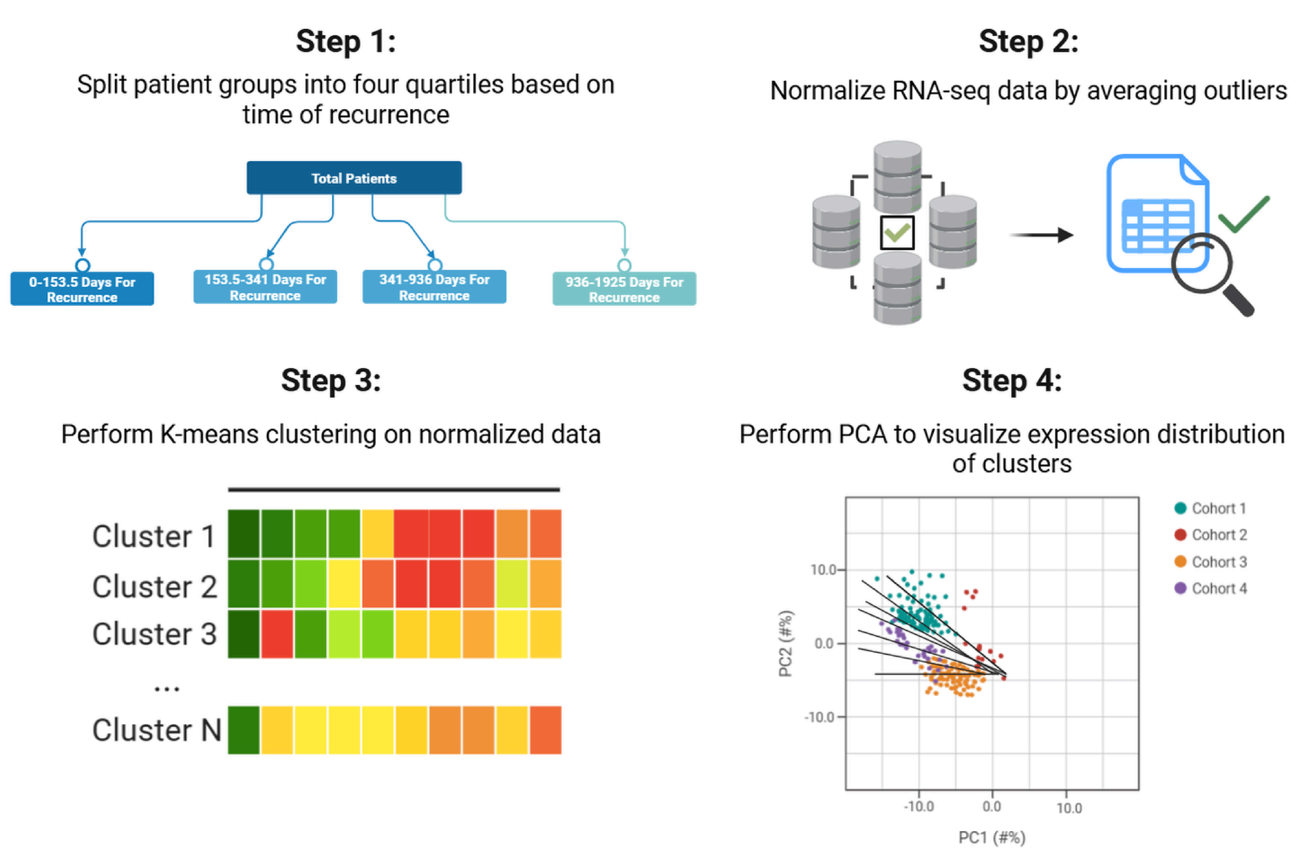
NCI Comprehensive Cancer Center
A Cancer Center Designated by the National Cancer Institute

SAFIULLAH RIFAI[1], AZIMULLAH RIFAI[2], ZUMAR MEHER[1], MOHAMMAD KHAN[2], LINBO WANG[2], WEI GUANG[2], ARIF HUSSAIN[2]

[1]UNIVERSITY OF MARYLAND SCHOOL OF MEDICINE, (BALTIMORE, MD, USA), [2]MARLENE AND STEWART GREENEBAUM CANCER CENTER, (BALTIMORE, MD, USA)

HYPOTHESIS

Unsupervised clustering methods can reveal transcriptomic patterns associated with early biochemical recurrence in prostate cancer.



ABSTRACT

Radical prostatectomy or radiation are the standard treatment for clinically localized prostate cancer. Although potentially curative, up to 15-30% of men undergoing definitive local therapy experience biochemical recurrence. The time to biochemical recurrence can serve as an important indicator of worse long term clinical outcomes. Therefore, understanding the factors driving early recurrence is critical. Our study examined transcriptomic data from 74 patients in the TCGA Prostate Adenocarcinoma dataset who experienced biochemical recurrence after radical prostatectomy for localized disease. We stratified patients into quartiles based on time to recurrence and analyzed their gene expression profiles using K-means clustering and Principal Component Analysis (PCA). K-means clustering revealed distinct gene expression patterns associated with earlier biochemical recurrence. Over-representation analysis across multiple Gene Ontology collections highlighted immune cell signatures as major contributors to these patterns. Principal Component Analysis identified epithelial splicing regulatory protein 1 (ESRP1) – a regulator of epithelial-mesenchymal transition – as significantly associated with early recurrence in a subset of patients. Future directions include deconvolution of the bulk RNA sequencing data to better characterize immune signatures across recurrence quartiles, and extension of our analysis to include methylation and protein expression data.

METHODS

- Source data from GDAC Firehose
- Patients with biochemical recurrence (BCR) were divided into four quartiles (17-18 patients per bucket) based on time of recurrence (Figure 3)
- Outlier gene expression values were replaced by the cohort-specific median for that gene (normalizes data)
- *e.g., If Patient 5 in Cohort 1 had an outlier GAPDH value, it was replaced with the GAPDH median from Cohort 1*
- Normalized RNA-seq data was used to perform K-means clustering
- Principal Component Analysis (PCA) was then performed on clusters of interest and gene vectors contributing most to variance were determined
- Hierarchical Clustering was performed to determine whether a given cluster could independently stratify patients into their respective quartile
- Gene set Over-representation Analysis on Gene Ontology collections was used to determine whether clusters mapped to previously characterized gene sets
- Firehose and SU2C Dream Team genomic and transcriptomic data was mined to search for prevalence of alterations in genes of interest in the primary and metastatic disease, as well as to determine prevalence of genes in clinical outcomes

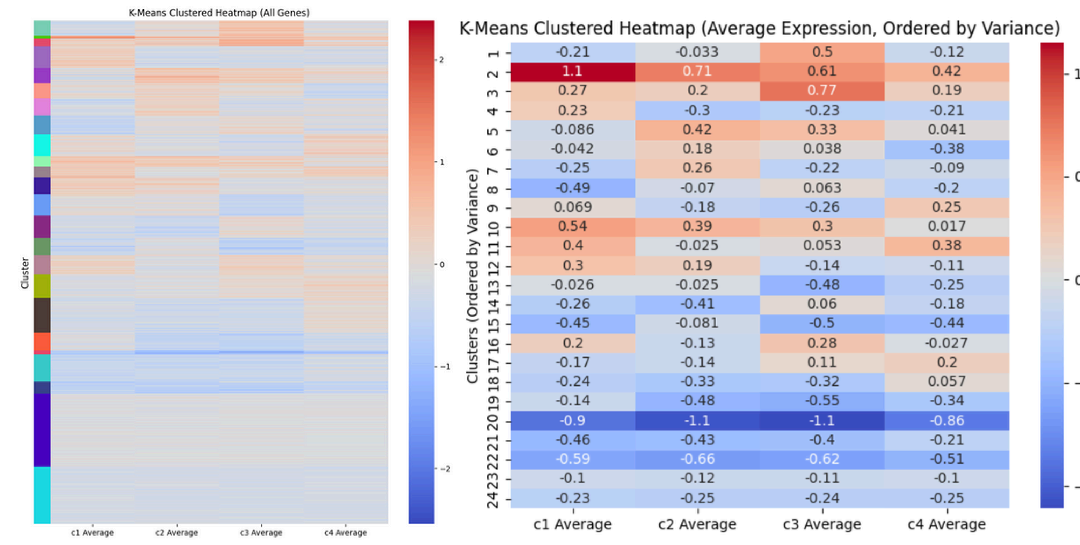


Figure 2: K-means heatmap for the 24 clusters

Figure 3: Quartile Split for Biochemically Recurrent Patients

RESULTS

- Cluster 2 (95 genes) was identified as a gene set of interest due to its step-wise change in expression and overexpression in patients with early biochemical recurrence
- Hierarchical Clustering using the 95-gene signature of Cluster 2 suggests that it is a predictor of early biochemical recurrence
- PCA on patients using the Cluster 2 gene signature showed a modest stratification (~32% variance captured on Principal Component 1)
- The top 10 genes vectors contributing to PC1 were dramatically amplified in Firehose and Dream Team datasets and alterations led to worse outcomes

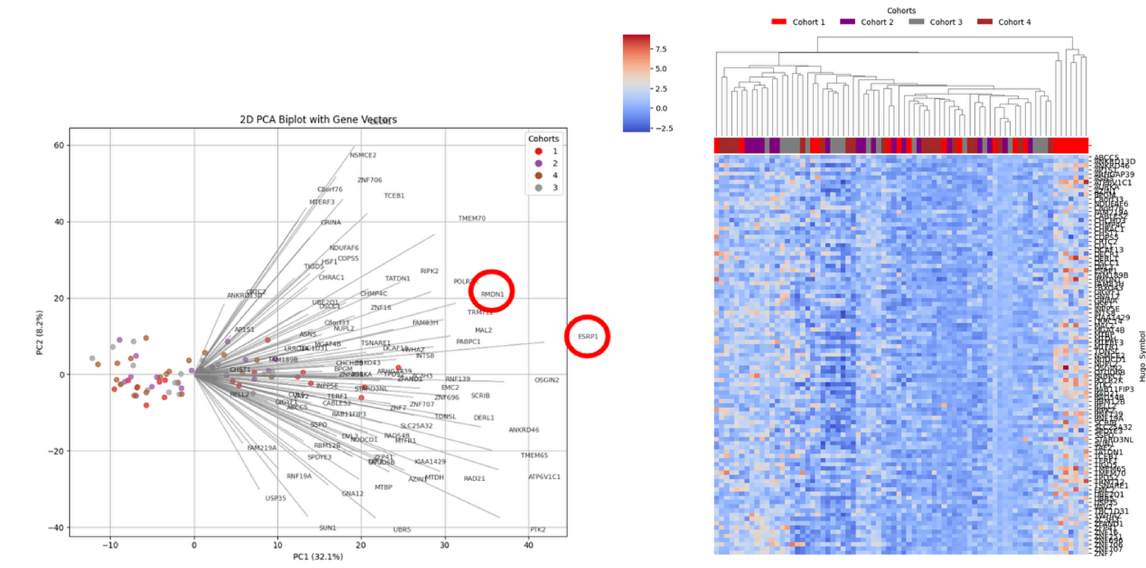


Figure 4: PCA Graph of Cluster 2. Circled are known genes of significance, **ESRP1** (driver of EMT) and **RMDN1** (associates with oncogenic Myc)

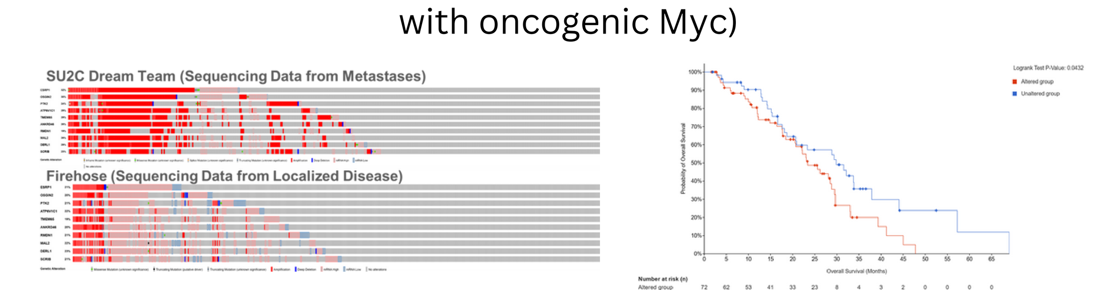


Figure 5: Gene amplification/overexpression plot (left) and KM plot (right) displaying alteration vs no alteration of the top 10 genes contributing to PC1 (SU2C Dream Team data); time to death shows poor prognosis for patients with alterations

- Gene set Over-representation Analysis (**Figure 6**) mapped Cluster 2 strongly to chromosome 8 (blue), specifically 8q24 (near the Myc gene), which is known to be amplified in aggressive localized PC. Cluster 2 also mapped to immune signatures (orange), particularly colonic immunosuppressive tumor-associated macrophages (TAMs)

Collection	Term	Overlap	P-value	Adjusted P-value	Odds Ratio	Combined Score	Genes
Immune Signature	GSE18804_BRAIN_VS_COLON_TUMORAL_MACROPHAGE_DN	18/175	9.92E-19	3.14E-15	27.491019	1139.627988	TRMT12;PABPC1;YWHAZ;SLC25A32;RNF139;UBR5;NUDCD1;COPS5;C8orf76;DSCC1;ZNF696;RAD21;NDUFAF6;SCRIB;DCAF13;NSMCE2;INTS8;CHMP4C
Immune Signature	GSE45365_CTRL_VS_MCMV_INFECTION_NK_CELL_DN	15/167	7.33E-15	1.16E-11	22.907993	745.569052	TRMT12;PTK2;TERF1;RNF139;DSCC1;C8orf76;COPS5;RAD21;SCRIB;NDUFAF6;DCAF13;NSMCE2;INTS8;DERL1;CHMP4C
Immune Signature	GSE45365_CD8A_DC_VS_CD11B_DC_IFNAR_KO_DN	12/172	8.50E-11	8.96E-08	16.915235	392.244267	TATDN1;ZFAND1;RNF139;RAD21;ZNF16;NDUFAF6;DCAF13;NSMCE2;TAF2;PTK2;TBC1D31;CHMP4C
Immune Signature	GSE18804_SPLEEN_MACROPHAGE_VS_BRAIN_TUMORAL_MACROPHAGE_DN	7/172	3.07E-05	2.42E-02	9.283886	96.483922	ESRP1;TERF1;RNF139;ANKRD46;RAB11FIP3;DSCC1;NSMCE2
Chromosome Position	chr8q24	32/154	1.45E-43	3.18E-42	75.929777	7489.915268	ZFP41;C8orf33;GRINA;DSCC1;LRRC14;TAF2;TRMT12;TATDN1;FAM83H;TSNARE1;ZNF696;NSMCE2;ZNF7;TIGD5;MAL2;PTK2;C8orf76;ZNF16;TBC1D31;DERL1;ZNF251;RNF139;ARHGAP39;ZNF707;CHRAC1;RAD21;SCRIB;ZC3H3;MTBP;TMEM65;HSF1;TONSL
Chromosome Position	chr8q22	19/60	9.04E-30	9.94E-29	112.123081	7498.339676	MTDH;UBR5;NDUFAF6;FBXO43;PABPC1;RAD54B;RNF19A;POLR2K;DCAF13;MTERF3;SLC25A32;ANKRD46;AZIN1;RBM12B;ATP6V1C1;ZNF706;INTS8;ESRP1;YWHAZ
Chromosome Position	chr8q21	10/63	8.49E-13	6.23E-12	41.874898	1163.884152	DECR1;TERF1;ZFAND1;OSGIN2;OTUD6B;TPD52;TMEM70;RMDN1;RIPK2;CHMP4C
Chromosome Position	chr8q23	2/25	2.65E-03	1.46E-02	36.210141	214.789285	NUDCD1;EMC2
Cellular Component	GOCC_CHROMOSOMAL_REGION	10/350	1.32E-05	3.45E-03	6.475951	72.755969	SUN1;TERF1;DSCC1;AURKA;CHRAC1;RAD21;MTBP;NSMCE2;HSF1;CHMP4C
Cellular Component	GOCC_CHROMOSOME_CENTROMERIC_REGION	7/225	1.66E-04	2.16E-02	7.011403	61.029922	DSCC1;AURKA;CHRAC1;RAD21;MTBP;HSF1;CHMP4C
Cellular Component	GOCC_DERLIN_1_RETROTRANSLOCATION_COMPLEX	2/25	5.45E-04	4.74E-02	88.918328	668.154589	DERL1;RNF139
Biological Processes	GOBP_CELL_CYCLE	22/1518	7.26E-06	5.62E-03	3.436488	40.662402	RAB11FIP3;OSGIN2;DSCC1;CUL1;COPS5;RMDN1;TAF2;ASNS;FBXO43;RAD54B;AURKA;DCAF13;CABLES2;NSMCE2;SUN1;ZNF16;CHMP4C;TERF1;RAD21;SCRIB;MTBP;HSF1
Biological Processes	GOBP_ORGANELLE_FISSION	12/460	4.34E-06	5.62E-03	5.957149	73.556874	SUN1;TERF1;RAD54B;AURKA;MTBP;RAD21;RMDN1;DCAF13;MTFR1;NSMCE2;CHMP4C;FBXO43
Biological Processes	GOBP_MEIOTIC_CELL_CYCLE	8/262	6.36E-05	2.18E-02	6.886455	66.538924	SUN1;TERF1;RAD54B;OSGIN2;AURKA;RAD21;DCAF13;FBXO43
Biological Processes	GOBP_MEIOTIC_CELL_CYCLE_PROCESS	7/196	7.03E-05	2.18E-02	8.097357	77.43727	SUN1;TERF1;RAD54B;AURKA;RAD21;DCAF13;FBXO43

CONCLUSIONS

1. K-means clustering effectively separated patients with varying time to biochemical recurrence via unique transcriptomic signatures
2. PCA and Hierarchical Clustering on Cluster 2 independently verified stratification of early recurrence cases
3. Analysis of clinical genomic, transcriptomic, and outcomes data for localized and metastatic disease confirmed worse outcomes for patients with a Cluster 2-like phenotype
4. Cluster 2 was shown to contain known oncogenes such as ESRP1 and RMDN1, as well as novel oncogenes
5. Cluster 2 mapped strongly to chromosome 8, near Myc, a known early driver of PC, as well as tumor-associated macrophage immune signatures

SIGNIFICANCE

- Transcriptomic data analysis was able to identify signatures of aggressive disease that correlate with early biochemical recurrence
- Genes that were identified as important in biochemical recurrence were also found to be overexpressed in metastatic prostate cancer
- The mapping of Quartile 1 to overexpression of Cluster 2, which includes genes on Chromosome 8 and immune cell related genes, could inform the use of more targeted therapies
- The gene contributing most to PC1 was Epithelial Splicing Regulatory Protein 1 (ESRP1), which regulates RNA splicing of several genes in epithelial cells and plays a role in epithelial-mesenchymal transition
- Chromosome 8 contains about 700 genes, most notably relevant in PC is an over-expression of MYC (a regulator of cell growth and current PC target of interest), although the oncogenic role of many of these genes has yet to be characterized (below are the top 10 genes contributing to PC1 and their recorded functions)

ESRP1	Splicing regulator promoting epithelial characteristics; often downregulated during prostate cancer epithelial-to-mesenchymal transition (EMT).
OSGIN2	Oxidative stress-induced gene with limited prostate-specific data; may signal cellular stress responses in tumor microenvironments.
PTK2 (FAK)	Focal adhesion kinase driving cell migration and invasion; frequently associated with advanced prostate cancer and metastasis.
ATP6V1C1	V-ATPase subunit involved in pH regulation; may contribute to acidic tumor microenvironments and castration resistance.
TMEM65	Mitochondrial membrane protein; its dysregulation may reflect altered energy metabolism in prostate cancer cells.
ANKRD46	Ankyrin-repeat protein with poorly characterized function in prostate cancer; possible novel biomarker or regulatory node.
RMDN1	Regulator of microtubule dynamics and apoptosis; may impact prostate tumor cell survival under stress.
MAL2	Lipid raft trafficking protein; overexpression could affect membrane signaling and endocytosis in prostate cancer progression.
DERL1	ER-associated degradation component; linked to unfolded protein response activation in aggressive prostate cancer subtypes.
SCRIB	Polarity protein often mislocalized in tumors; its loss or misregulation is associated with increased invasion and poor prognosis.

This work was supported by:

- **National Cancer Institute grants**
  - **5R25GM113262-09 (Gregory Carey)**
- **Cigarette Restitution Fund (CRF) Program**
- **Department of Health**
- **State of Maryland**

Maryland

University of Maryland Marlene and Stewart Greenebaum Comprehensive Cancer Center