UALCAN: An update to the integrated cancer data analysis platform ☆

Darshan Shimoga Chandrashekar [1,*]; Santhosh Kumar Karthikeyan [1]; Praveen Kumar Korla [1]; Henalben Patel [1]; Ahmedur Rahman Shovon [2]; Mohammad Athar [3,4]; George J. Netto [1,3]; Zhaohui S. Qin [6]; Sidharth Kumar [2]; Upender Manne [1,3]; Chad J. Crieghton [8,#]; Sooryanarayana Varambally [1,3,5,7,#]

[1] Department of Pathology, University of Alabama at Birmingham, Birmingham, AL, USA
[2] Department of Computer science, University of Alabama at Birmingham, Birmingham, AL, USA
[3] O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL, USA
[4] Department of Dermatology, University of Alabama at Birmingham, Birmingham, AL, USA
[5] Informatics Institute, University of Alabama at Birmingham, Birmingham, AL, USA
[6] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA
[7] Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI, USA
[8] Department of Medicine and Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA

Abstract

Cancer genomic, transcriptomic, and proteomic profiling has generated extensive data that necessitate the development of tools for its analysis and dissemination. We developed UALCAN to provide a portal for easy exploring, analyzing, and visualizing these data, allowing users to integrate the data to better understand the gene, proteins, and pathways perturbed in cancer and make discoveries. UALCAN web portal enables analyzing and delivering cancer transcriptome, proteomics, and patient survival data to the cancer research community. With data obtained from The Cancer Genome Atlas (TCGA) project, UALCAN has enabled users to evaluate protein-coding gene expression and its impact on patient survival across 33 types of cancers. The web portal has been used extensively since its release and received immense popularity, underlined by its usage from cancer researchers in more than 100 countries. The present manuscript highlights the task we have undertaken and updates that we have made to UALCAN since its release in 2017. Extensive user feedback motivated us to expand the resource by including data on a) microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and promoter DNA methylation from TCGA and b) mass spectrometry-based proteomics from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). UALCAN provides easy access to pre-computed, tumor subgroup-based gene/protein expression, promoter DNA methylation status, and Kaplan-Meier survival analyses. It also provides new visualization features to comprehend and integrate observations and aids in generating hypotheses for testing. UALCAN is accessible at http://ualcan.path.uab.edu

Neoplasia (2022) 000, 1–9

Keywords: UALCAN, Cancer gene expression, Bioinformatics, Transcriptomics, Patient survival, lncRNA, Proteomics, Correlation, Differential expression

[*] Corresponding authors at: O'Neal Comprehensive Cancer Center, WTI 420B, University of Alabama at Birmingham, Birmingham, AL 35233, USA
E-mail addresses: dshimogachandrasheka@uabmc.edu (D.S. Chandrashekar), svarambally@uabmc.edu (S. Varambally).

☆ Funding/support: Department of Pathology, UAB

[#] Share Senior Authorship

Received 24 November 2021; received in revised form 5 January 2022; accepted 10 January 2022; Available online xxx

https://doi.org/10.1016/j.neo.2022.01.001

Introduction

Cancer is a complex and heterogeneous disease, rarely detected at its initial stages [1]. Current methods rely on biomarker panels that contain several genes rather than a single specific biomarker to quickly and accurately detect cancers [2–4]. The advent of high-throughput technologies, such as whole/targeted exome sequencing, whole-genome sequencing, large-scale RNA sequencing, and chromatin immunoprecipitation followed by sequencing (ChIP-Seq) and mass spectrometry-based proteomics, has accelerated cancer research and has resulted in a large volume of publicly available data. Clinicians and cancer researchers involved in detecting, discovering, and validating cancer biomarkers and treatments, find it difficult to access, process, integrate, and interpret high-throughput data. Hence, easy-to-use, web-based/standalone tools enable cancer researchers/clinicians to access omic data and perform multilevel analyses. With this objective, we developed UALCAN [5]. UALCAN enables researchers to access Level 3 RNA-seq data from The Cancer Genome Atlas (TCGA) and perform gene expression and survival analysis on about 20,500 protein-coding genes in 33 different tumor types. Since its publication, UALCAN has become a highly used web portal for cancer researchers around the world. Since its release in 2017, the web portal has been accessed more than 750,000 times from more than 100 countries. We have recently performed several additional data analyses and data integration and upgraded the UALCAN web portal. We have expanded the scope of UALCAN by including RNA-seq data analysis related to non-coding genes and adding new analysis features such as gene correlation, pan-cancer analysis, and promoter methylation analysis. The current manuscript highlights the additions and upgrades that we have incorporated into the latest version of UALCAN. Its web portal is available at http://ualcan.path.uab.edu.

Methods

Collection of non-coding gene expression data

Level 3 miRNA-seq data were downloaded from TCGA using TCGA-assembler pipeline [6]. DownloadmiRNASeqData() was utilized to download gene expression values (reads per million, RPM) for 1871 pre-miRNAs in 32 cancer types. Normal tissue and primary tumor RPM values were downloaded separately. RPM values of primary tumor samples were categorized based on the demography of patients and clinicopathological features such as race, age, gender, tumor grade, tumor stage, and molecular subtype.

TCGA gene expression values for long non-coding RNAs were downloaded from the Genome Data Commons (GDC) portal (https://gdc.cancer.gov/), which facilitates downloading, for each sample, a gene expression quantification file generated via "HTSeq-FPKM" workflow (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/). Each file contains an Ensembl gene id and a Fragments Per Kilobase of transcript per Million mapped reads (FPKM) value for all genes, including 14,076 lncRNAs.

Collection of DNA methylation data

TCGA-assembler pipeline was used to download TCGA DNA methylation data generated using the Illumina Infinium HumanMethylation450 BeadChip. At first, the Download MethylationData() function was used to download Infinium HumanMethylation450 BeadChip data for primary tumors and matched normal samples of each cancer type as tab-separated text files. Downloaded data were further processed to calculate an average methylation (beta) value for each gene, considering CpG sites located in the promoter region of the gene (1500 bases upstream and 200 bases downstream of transcription start sites [TSS]) via the CalculateSingleValueMethylationData() in TCGA-assembler.

Collection of proteomic data

High-throughput mass spectrometry data related to breast cancer, colon cancer, ovarian cancer, clear cell renal carcinoma, uterine corpus endometrial carcinoma, lung adenocarcinoma, and pediatric brain cancer were obtained from the Clinical Proteomic Tumor Analysis Consortium (CPTAC), which provides expression data for approximately 10,000 proteins. Integration and analysis of these data have been reported [7,8]. In brief, protein expression values downloaded from the CPTAC data portal were log2 normalized in each sample. Then a Z-value for each sample for each protein was calculated as standard deviations from the median across samples.

Collection of ChIP-seq data

We have now acquired and incorporated into UALCAN, ChIP-sequencing data related to a) activating or repressing histone modifications (H3K27me3, H3K27ac, H3K9ac, and H3K9me3) in breast cancer cell lines from the Gene Expression Omnibus (GEO) [GSE85158], b) transcription factors (AR, ERG, EZH2, SUZ12, and BRD4), and activating or repressing histone modifications (H3K27me3 and H3K27ac) in prostate cancer cell lines from GEO [GSE73616, GSE83653, GSE55062, GSE135623, GSE83860, GSE39459, GSE137207, GSE114737] [9–16].

Database integration of non-coding RNA targets

For facilitating integrated analysis of miRNAs or lncRNAs, several non-coding RNA target databases were downloaded and embedded in the UALCAN server. This effort includes miRNA target databases such as TargetScan, comprised of target predictions based on conserved sites in miRNA seed regions [17]; microRNA.org, in which target predictions are based on a miRanda algorithm [18]; and miRDB, which includes miRNA targets predicted by the MirTarget tool [19]. LncRNA target databases such as LncREG, which provides validated lncRNA-gene associations from published research [20]; and LncRNA2Target, a database of genes altered on lncRNA knockdown or overexpression [21]. Data were downloaded as tab-separated files.

Data analyses

Analyses of gene expression/promoter methylation

miRNA and lncRNA expression values for each tumor subgroup or normal samples were displayed as box-whisker plots, similar to protein-coding gene expression values presented previously. For each tumor subgroup, the box-whisker plots present interquartile ranges (IQRs), including minimum, 1st quartile, median, 3rd quartile, and maximum values. In terms of statistics, we used the descriptive PERL module to calculate IQR values after filtering outliers. Welch's T-test estimated the significance of differences in expression levels between normal and primary tumors or tumor subgroups based on clinicopathological features. We followed similar methods to display promoter DNA methylation status and measure the significance of hypo-/hyper-methylation status.

Top 100 over-/under-expressed lncRNAs in specific cancer were selected, considering lncRNAs a) with median expression value (i.e FPKM) greater than 1 in either tumor samples or normal samples, b) with higher ratio between mean tumor expression value and mean normal expression value and c) statistical significance of 0.001 or less between normal and tumor gene expression values. Same procedure was followed to identify top 50 over-/under-expressed miRNAs.

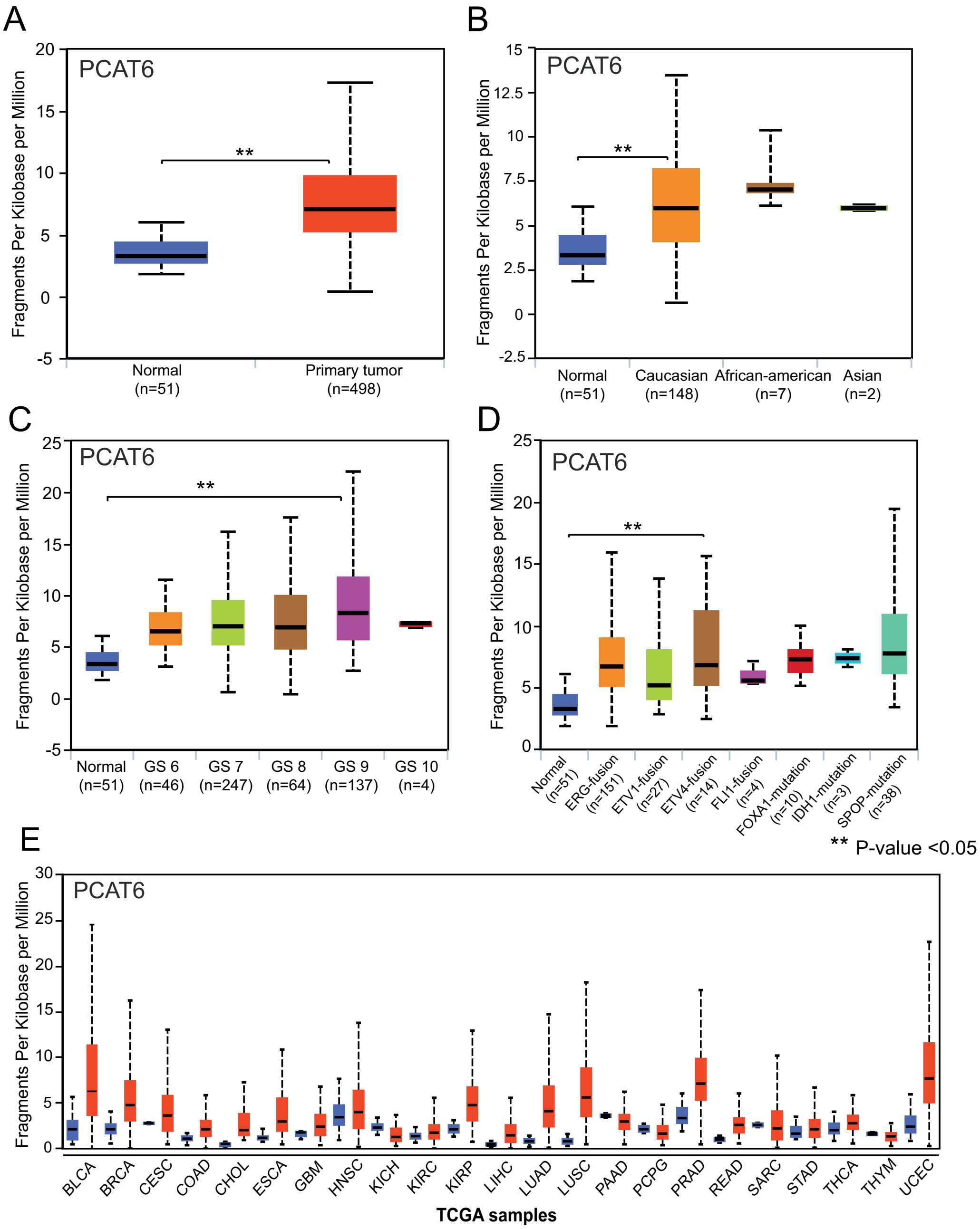


Fig. 1. LncRNA expression profile in prostate adenocarcinoma. (A-D) Graphs showing expression level analysis of PCAT6 using UALCAN web-portal in normal prostate, all primary tumors and subgroups based on patient race (B), Gleason score(C) and molecular subtypes (D). (E) Pan cancer gene expression profile of PCAT6. Red boxplot depicts expression level in primary tumors, while blue boxplot indicate expression in normal samples.

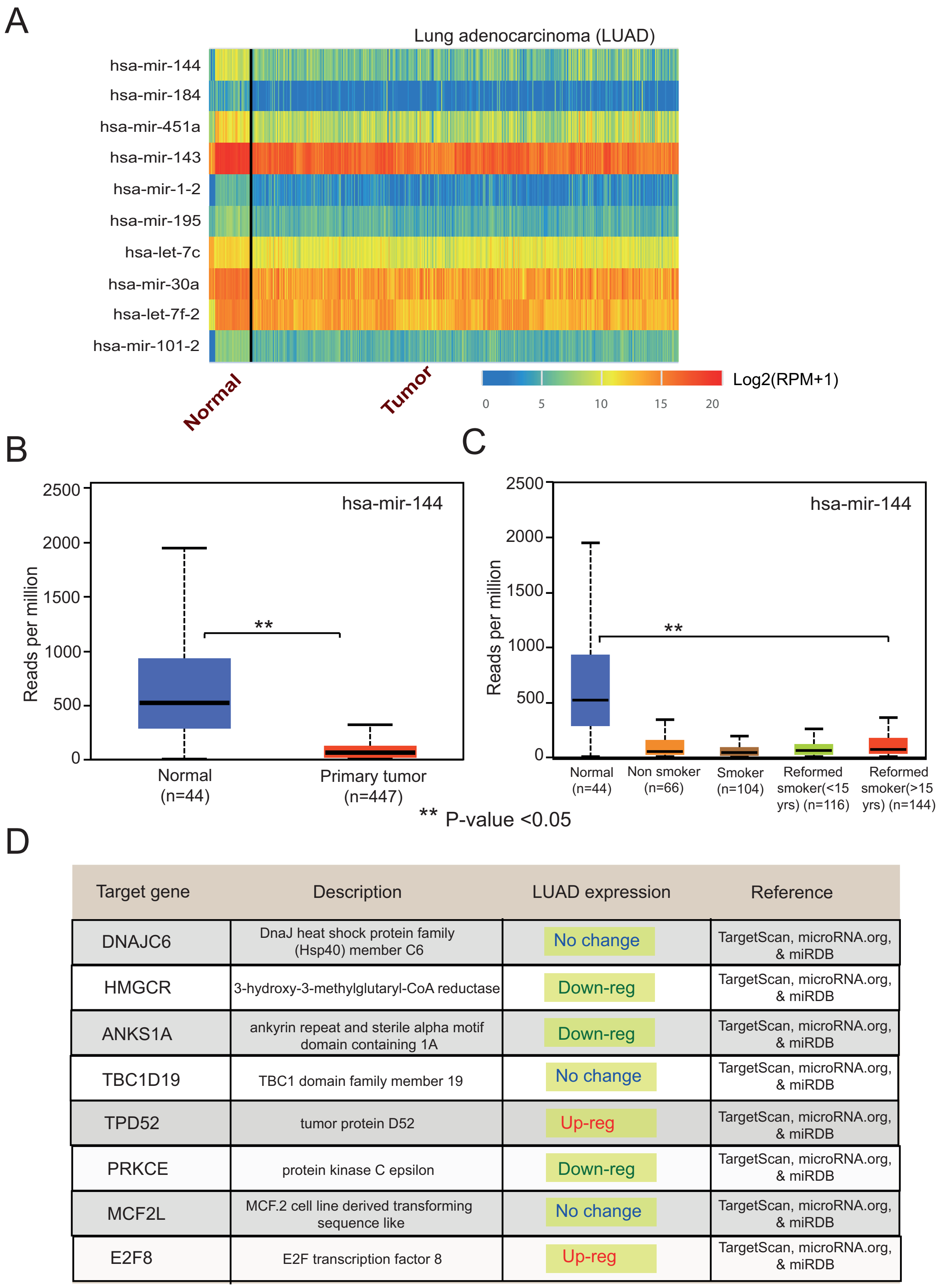
Survival analyses

Kaplan-Meier analyses are a prominent feature of UALCAN. We performed additional univariate and multivariate survival analyses to aid users in evaluating the effect of non-coding RNA expression levels on the overall survival of patients. As mentioned in our previous manuscript [5], from TCGA patient survival data, we considered ‘day_to_last_follow_up’ if the patient was alive and ‘days_to_death’ if the patient was dead. Primary tumor samples were divided into a ‘High expression’ group (i.e., samples with gene expression values equal to or more than the 3rd quartile value) and a ‘Low/Medium expression’ group (i.e., samples with gene expression values less than the 3rd quartile).

We conducted survival analyses and generated Kaplan-Meier plots using the R packages ‘survival’ and ‘survminer’. P-values from log rank tests were derived to determine the significance of survival analyses. Multivariate survival analyses were also performed to assess the combinatorial effect of non-coding RNA expression and race/gender on patient survival.

Pan-cancer analysis

To facilitate analysis of gene expression patterns across TCGA cancer data, we provided a “Pan Cancer View” since expression values (TPM or FPKM) for a gene could vary substantially between cancer types. These were log-normalized as log2(TPM+1) or as log2(FPKM+1), and interquartile ranges (IQRs) were derived for each cancer type and projected as box-whisker plots.



Target gene	Description	LUAD expression	Reference
DNAJC6	DnaJ heat shock protein family (Hsp40) member C6	No change	TargetScan, microRNA.org, & miRDB
HMGCR	3-hydroxy-3-methylglutaryl-CoA reductase	Down-reg	TargetScan, microRNA.org, & miRDB
ANKS1A	ankyrin repeat and sterile alpha motif domain containing 1A	Down-reg	TargetScan, microRNA.org, & miRDB
TBC1D19	TBC1 domain family member 19	No change	TargetScan, microRNA.org, & miRDB
TPD52	tumor protein D52	Up-reg	TargetScan, microRNA.org, & miRDB
PRKCE	protein kinase C epsilon	Down-reg	TargetScan, microRNA.org, & miRDB
MCF2L	MCF.2 cell line derived transforming sequence like	No change	TargetScan, microRNA.org, & miRDB
E2F8	E2F transcription factor 8	Up-reg	TargetScan, microRNA.org, & miRDB

Fig. 2. miRNA expression profile in Lung adenocarcinoma. (A) Heatmap generated using UALCAN showing top miRNAs under expressed in lung adenocarcinoma [LUAD]. (B-C) Boxplot showing expression level of hsa-miR-144 in normal, primary tumors and tumor subgroups based on patient's smoking status. (D) Predicted targets of hsa-miR-144 and their expression status is lung adenocarcinoma.

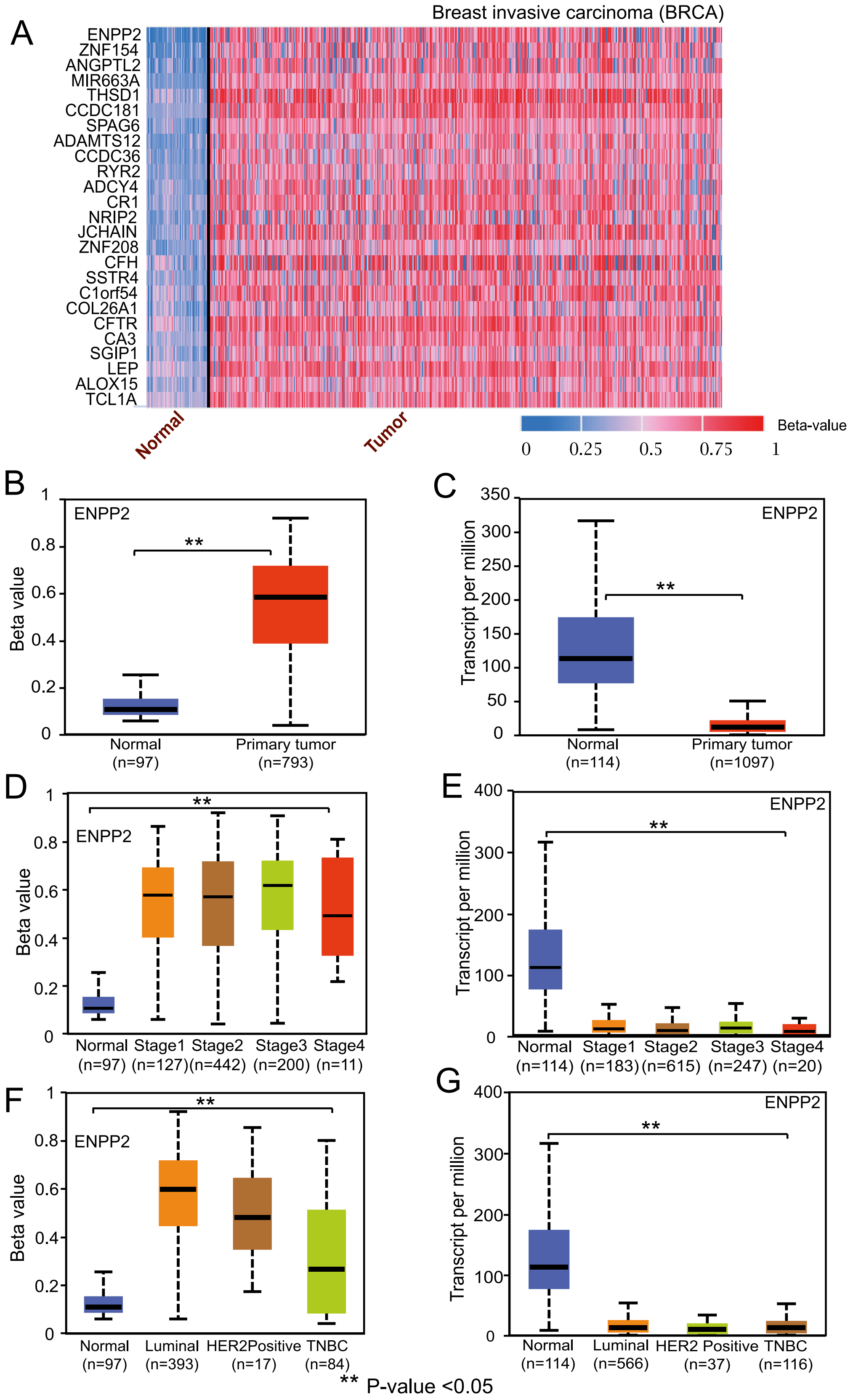
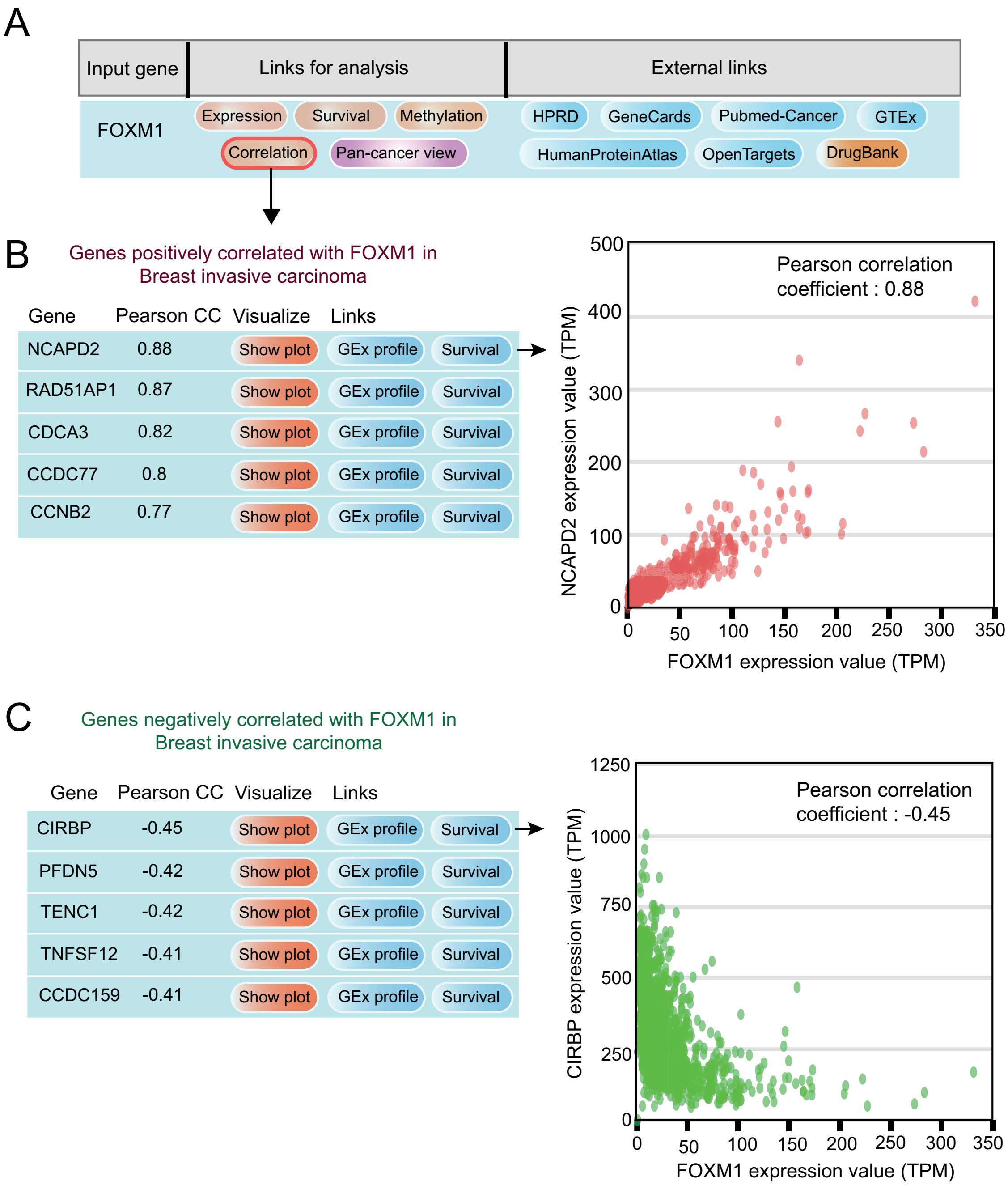


Fig. 3. Promoter methylation of DNA information from TCGA Illumina bead chip data in UALCAN. (A) UALCAN generated heatmap showing top 25 genes with hyper-methylated promoter DNA in breast invasive carcinoma. (B, C) Boxplots showing inverse relation between promoter methylation status and gene expression profile of ENPP2 in TCGA breast invasive carcinoma [BRCA]. (D-G) Boxplots showing promoter methylation level and mRNA expression level of ENPP2 in different stages of breast cancer (D,E) and different subclasses of breast cancer (F, G).


A
B
C
Pearson correlation coefficient : 0.88
NCAPD2 expression value (TPM)
FOXM1 expression value (TPM)
0
100
200
300
400
500
0
50
100
150
200
250
300
350
Pearson correlation coefficient : -0.45
CIRBP expression value (TPM)
FOXM1 expression value (TPM)
0
250
500
750
1000
1250
0
50
100
150
200
250
300
350

Input gene	Links for analysis	External links
FOXM1	Expression Survival Methylation Correlation Pan-cancer view	HPRD GeneCards Pubmed-Cancer GTEx HumanProteinAtlas OpenTargets DrugBank

Genes positively correlated with FOXM1 in Breast invasive carcinoma

Gene	Pearson CC	Visualize	Links	
NCAPD2	0.88	Show plot	GEx profile	Survival
RAD51AP1	0.87	Show plot	GEx profile	Survival
CDCA3	0.82	Show plot	GEx profile	Survival
CCDC77	0.8	Show plot	GEx profile	Survival
CCNB2	0.77	Show plot	GEx profile	Survival

Genes negatively correlated with FOXM1 in Breast invasive carcinoma

Gene	Pearson CC	Visualize	Links	
CIRBP	-0.45	Show plot	GEx profile	Survival
PFDN5	-0.42	Show plot	GEx profile	Survival
TENC1	-0.42	Show plot	GEx profile	Survival
TNFSF12	-0.41	Show plot	GEx profile	Survival
CCDC159	-0.41	Show plot	GEx profile	Survival

Fig. 4. Gene expression correlation analysis of FOXM1 in Breast invasive carcinoma [BRCA]. (A) Gene query result page in UALCAN directs user to gene correlation analysis page. (B, C) UALCAN generated list of positively [Pearson Correlation Coefficient > 0.3] and negatively [Pearson Correlation Coefficient < -0.3] correlated genes of FOXM1 in BRCA. Scatter plots are provided to visualize correlation for each gene pair.

Gene correlation analyses

Correlation analysis of genes using RNA expression values helps researchers to identify interacting or co-expressing genes. Thus, we performed Pearson correlation analysis using expression values of all protein-coding genes. We carried out analyses using an in-house PERL script that utilizes the "Statistics::Basic" module. Gene pairs showing Pearson correlation coefficients of 0.3 or above were considered as positive correlations, and those showing Pearson correlation coefficients of -0.3 or below were considered as negative correlations.

ChIP-seq data analysis

Raw sequence read files (fastq) were downloaded using the fastq-dump tool of the NCBI SRA toolkit (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). Downloaded data were cleaned using Trim galore [https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/]. Trimmed reads were mapped to the human reference genome (hg38) using the Burrows-Wheeler Aligner [bwa mem] [22]. Mapped reads were sorted, and duplicate reads were marked using Picard tools [https://broadinstitute.github.io/picard/]. For visualization purposes, a BigWig file for each BAM

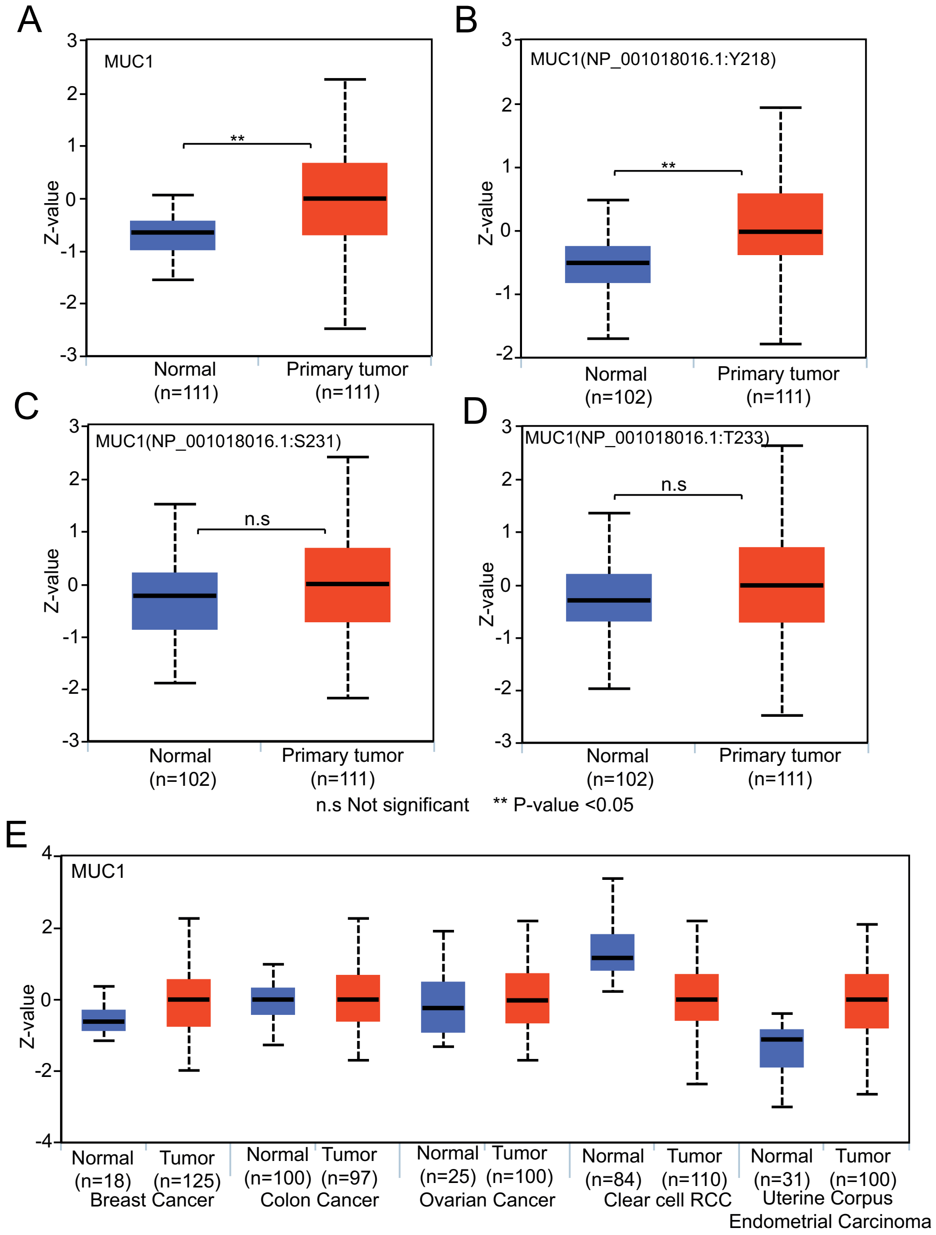


Fig. 5. Total protein and phosphoprotein expression pattern of MUC1 in lung adenocarcinoma. (A) Boxplot generated using UALCAN showing total protein expression of MUC1 in lung adenocarcinoma. (B-D) Boxplots depicting expression level of MUC1 phosphoproteins in lung adenocarcinoma. (E) Pan cancer view of MUC1 total protein expression in breast, colon, ovarian, renal and endometrial cancer.

file was generated using the bamCoverage module of deepTools [23]. Peak calling was performed with MACS2 [24].

To facilitate interactive visualization of the results of ChIP-seq data analysis, “igv.js” JavaScript, developed by the Integrative Genome Viewer (IGV) team, was embedded into UALCAN [https://github.com/igvteam/igv.js/].

Results

Incorporation of new data and analysis features in UALCAN

The UALCAN web portal is hosted on a powerful server with a Linux operating system (CentOS) and an Apache web server. The front end of the web portal was designed using PERL CGI, CSS, and JavaScripts from Highcharts and IGV.

Long non-coding RNA expression and patient survival analysis

Cancer genome transcribes many non-coding RNAs apart from protein coding genes. These include long non-coding RNAs. The UALCAN analysis page allows users to enter one or more lncRNAs [as official gene symbol(s)] and select one of 33 TCGA cancers. On submission, the user is directed to i) a web page showing the expression pattern of the lncRNA (as a box-whisker plot) in normal and primary tumor samples [Fig. 1A], in tumor subgroups based on race, molecular subtypes, and by other clinicopathologic features [Fig. 1B-D]; ii) a web page presenting Kaplan-Meier plots depicting the effect of the lncRNA on overall survival of patients [Supplementary Fig. 1]; or iii) a web page providing a pan-cancer view, i.e., the expression pattern of the lncRNA across TCGA cancers [Fig. 1E]. In addition, users can obtain a list of target genes and their mRNA expression pattern by clicking the “Target genes” button at the bottom of the box-whisker plot on the gene expression page.

The analysis page also allows the user to obtain a list of the top 100 over-/under-expressed genes for a specific tissue-based cancer type or a specific molecular subtype of cancer. The gene list can be visualized and downloaded as a heatmap.

MicroRNA expression and patient survival analysis using TCGA transcriptome sequencing data

In the updated UALCAN, we have also integrated the small non-coding RNAs called microRNAs in UALCAN. The UALCAN analysis page allows users to i) retrieve a list of the top 50 over-/under-expressed miRNAs for a specific cancer type [Fig. 2A] and ii) query an miRNA of interest to obtain its expression in normal tissue, primary tumors, and tumor subgroups [Fig. 2B-C], as well as view expression-based overall survival plots and a list of predicted target genes [Fig. 2D].

Promoter DNA methylation analysis

Epigenetic modifications are involved in the regulation of gene expression and controlling many cellular process in both normal and cancer cells. To evaluate the potential role of DNA promoter methylation and histone modifications in regulating the expression of cancer genes, we incorporated DNA promoter methylation analysis into UALCAN (Fig. 3A). The protein-coding gene expression page includes a link to a web page providing promoter DNA methylation levels in primary tumor samples and their subgroups [Fig. 3B, D, F]. This feature allows cancer researchers to assess the direct influence of promoter DNA methylation on protein-coding gene expression in tumor samples [Fig. 3C, E, G]. The methylation page also provides a list of the top 100 hyper-/hypo-methylated genes for each cancer type.

ChIP-seq data analysis

We have now included selected public ChIP-seq datasets, integrated into UALCAN, facilitating analysis of activating (H3K9Ac/H3K27Ac) or repressing (H3K9me3/H3K27me3) histone modifications along upstream or gene body regions for breast cancer and prostate cancer. In addition, ChIP-seq data for transcription factors and polycomb group complex members (ERG, EZH2, AR, and SUZ12) are included for prostate cancer cell lines. ChIP-seq results are displayed as interactive genome visualization for easy interpretation [Supplementary Fig. 2]. We are presently extending this feature to other cancers.

Identification of cancer-specific correlated genes

Most genes do not function in isolation but have networks that facilitate gene function. There are master regulatory genes and the expression of some genes regulates numerous other genes. Thus, to identify associated gene expression, we have now incorporated correlated gene expressions. The expression-based gene correlations are useful in constructing and understanding gene interaction/regulatory networks. With this in mind, we provide a list of positively and negatively correlated genes for each query gene in each cancer type [Fig. 4A]. Scatter plots for each gene pair in cancers can be visualized and downloaded [Fig. 4B, C].

We have initiated the incorporation of protein expression and post-translational modification analyses using CPTAC data (https://cptac-data-portal.georgetown.edu/). With the integration of high-throughput mass spectrometry data, UALCAN facilitates the analysis of the expression of total protein [Fig. 5A] and phosphorylated proteins [Fig. 5B-D] for seven cancer types. The pan-cancer view also helps visualize relative expression levels of proteins across breast, colon, ovarian, renal, and uterine cancers [Fig. 5E].

Discussion

The recent advances in DNA and RNA sequencing technology and advanced proteomic technology have enhanced cancer research and facilitated patient treatment. These technologies also produce extensive data that can potentially be utilized to perform various analyses and identify new biomarkers and therapeutic targets. However, big data acquisition, management, curation, analysis, and sharing remain a bottle-neck. Cancer biologists and researchers need to identify biomarkers and cancer-related biological associations, discover therapeutic targets, and re-purpose existing drugs. To provide easy access and analysis to researchers, we developed the UALCAN portal [5], which researchers worldwide have extensively used since release of the first version in 2017. Since then, we have embarked on adding new data analyses and datasets and upgraded the portal to enhance its usage and user experience. We have included non-coding gene expression and DNA methylation from TCGA, protein expression data from CPTAC, and ChIP-Seq data from NCBI GEO. New features, including gene correlation analyses, pan-cancer views, and target gene listings for non-coding RNAs, have elevated the utility of this web portal.

The updates and upgrades that we have made in UALCAN have enhanced the portal’s functionality and now allow researchers to cross-compare the data for protein coding gene expression with both non-coding RNAs and proteomic changes. Furthermore, the inclusion of epigenetic data, including promoter methylation, enables researchers to identify potential regulators of gene expression by these mechanisms. The inclusion of microRNA and lncRNA expression and survival analysis adds another dimension to biomarker discovery and analysis of gene expression regulation.

Moving forward, we intend to obtain, analyze, and incorporate additional publicly available transcriptome sequencing datasets so that they can be used as validation datasets for the observations made with TCGA transcriptome sequencing data. We will also incorporate additional proteomic data for the analysis. We will incorporate multiple relevant ChIP-Seq data and the dot-pot feature to help identify the expression of outlier genes. We will also continue to incorporate the suggestions from end-users as and when they are applicable and feasible.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Darshan Shimoga Chandrashekar: Conceptualization, Methodology, Formal analysis, Software, Validation, Project administration, Writing – original draft. **Santhosh Kumar Karthikeyan:** Formal analysis. **Praveen Kumar Korla:** Formal analysis, Validation. **Henalben Patel:** Formal analysis, Validation. **Ahmedur Rahman Shovon:** Formal analysis. **Mohammad Athar:** Writing – review & editing. **George J. Netto:** Writing – review & editing. **Zhaohui S. Qin:** Writing – review & editing. **Sidharth Kumar:** Formal analysis. **Upender Manne:** Writing – review & editing. **Chad J. Crieghton:** Resources, Formal analysis, Writing – review & editing. **Sooryanarayana Varambally:** Conceptualization, Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision.

Acknowledgments

This study was supported by the UAB Department of Pathology, the UAB O’Neal Comprehensive Cancer Center, and the UAB Heersink School of Medicine. S.V. and U.M were supported by the funding from U54

CA118948. S.V. is also supported by DOD funding, W81XWH-19-1-0588. C.J.C. is supported by grant CA125123. We thank Ms. Meena Varambally from the UAB College of Arts and Sciences for her assistance in validation of statistical analysis and testing of front-end web page functionality. We thank High-Charts (https://www.highcharts.com/) for the graphic scripts. We acknowledge Dr. Donald Hill from UAB O'Neal Comprehensive Cancer Center for the help in editing this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neo.2022.01.001.

References

[1] Blackadar CB. Historical review of the causes of cancer. *World J Clin Oncol* 2016;**7**(1):54–86.

[2] Gan X, Wang T, Chen ZY, Zhang KH. Blood-derived molecular signatures as biomarker panels for the early detection of colorectal cancer. *Mol Biol Rep* 2020;**47**(10):8159–68.

[3] Sopyllo K, Erickson AM, Mirtti T. Grading Evolution and Contemporary Prognostic Biomarkers of Clinically Significant Prostate Cancer. *Cancers (Basel)* 2021;**13**(4).

[4] Srivastava A, Gupta A, Patidar S. Review of biomarker systems as an alternative for early diagnosis of ovarian carcinoma. *Clin Transl Oncol* 2021;**23**(10):1967–78.

[5] Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi B, et al. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* 2017;**19**(8):649–58.

[6] Zhu Y, Qiu P, Ji Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods* 2014;**11**(6):599–600.

[7] Chen F, Chandrashekar DS, Varambally S, Creighton CJ. Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. *Nat Commun* 2019;**10**(1):5679.

[8] Monsivais D, Vasquez YM, Chen F, Zhang Y, Chandrashekar DS, Faver JC, et al. Mass-spectrometry-based proteomic correlates of grade and stage reveal pathways and kinases associated with aggressive human cancers. *Oncogene* 2021;**40**(11):2081–95.

[9] Xu K, Wu ZJ, Groner AC, He HH, Cai C, Lis RT, et al. EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* 2012;**338**(6113):1465–9.

[10] Asangani IA, Dommeti VL, Wang X, Malik R, Cieslik M, Yang R, et al. Therapeutic targeting of BET bromodomain proteins in castration-resistant prostate cancer. *Nature* 2014;**510**(7504):278–82.

[11] Malinen M, Niskanen EA, Kaikkonen MU, Palvimo JJ. Crosstalk between androgen and pro-inflammatory signaling remodels androgen receptor and NF-kappaB cistrome to reprogram the prostate cancer cell transcriptome. *Nucleic Acids Res* 2017;**45**(2):619–30.

[12] Kedage V, Selvaraj N, Nicholas TR, Budka JA, Plotnik JP, Jerde TJ, et al. An Interaction with Ewing's Sarcoma Breakpoint Protein EWS Defines a Specific Oncogenic Mechanism of ETS Factors Rearranged in Prostate Cancer. *Cell Rep* 2016;**17**(5):1289–301.

[13] Bose R, Karthaus WR, Armenia J, Abida W, Iaquinta PJ, Zhang Z, et al. ERF mutations reveal a balance of ETS factors controlling prostate oncogenesis. *Nature* 2017;**546**(7660):671–5.

[14] Franco HL, Nagari A, Malladi VS, Li W, Xi Y, Richardson D, et al. Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. *Genome Res* 2018;**28**(2):159–70.

[15] Singh AA, Schuurman K, Nevedomskaya E, Stelloo S, Linder S, Droog M, et al. Optimized ChIP-seq method facilitates transcription factor profiling in human tumors. *Life Sci Alliance* 2019;**2**(1):e201800115.

[16] Jain P, Ballare C, Blanco E, Vizan P, Di Croce L. PHF19 mediated regulation of proliferation and invasiveness in prostate cancer cells. *Elife* 2020;**9**.

[17] Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 2015;**4**.

[18] Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and expression. *Nucleic Acids Res* 2008;**36**:D149–53 (Database issue).

[19] Wang X. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* 2008;**14**(6):1012–17.

[20] Zhou Z, Shen Y, Khan MR, Li A. LncReg: a reference resource for lncRNA-associated regulatory networks. *Database (Oxford). 2015* 2015.

[21] Jiang Q, Wang J, Wu X, Ma R, Zhang T, Jin S, et al. LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. *Nucleic Acids Res* 2015;**43**:D193–6 (Database issue).

[22] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**(14):1754–60.

[23] Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;**44**(W1):W160–5.

[24] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**(9):R137.