PCaDB - a comprehensive and interactive database for transcriptomes from prostate cancer population cohorts

Ruidong Li[1,2*], Zhenyu Jia[1,2*]

Affiliations:

[1] Department of Botany and Plant Sciences, University of California, Riverside, CA, USA

[2] Graduate Program in Genetics, Genomics, and Bioinformatics, University of California, Riverside, CA, USA

* To whom correspondence should be addressed: Ruidong Li at rli012@ucr.edu and Zhenyu Jia at arthur.jia@ucr.edu.

ABSTRACT

Prostate cancer (PCa) is a heterogeneous disease with highly variable clinical outcomes which presents enormous challenges in the clinical management. A vast amount of transcriptomics data from large PCa cohorts have been generated, providing extraordinary opportunities for the comprehensive molecular characterization of the PCa disease and development of prognostic signatures to accurately predict the risk of PCa recurrence. The lack of an inclusive collection and standard processing of the public transcriptomics datasets constrains the extensive use of the valuable resources. In this study, we present a user-friendly database, PCaDB, for a comprehensive and interactive analysis and visualization of gene expression profiles from 50 public transcriptomics datasets with 7,231 samples. PCaDB also includes a single-cell RNA-sequencing (scRNAseq) dataset for normal human prostates and 30 published PCa prognostic signatures. The advanced analytical methods equipped in PCaDB would greatly facilitate data mining to understand the heterogeneity of PCa and to develop prognostic signatures and machine learning models for PCa prognosis. PCaDB is publicly available at http://bioinfo.jialab-ucr.org/PCaDB/.

INTRODUCTION

As one of the most powerful approaches in oncology research, transcriptome profiling has been extensively used for understanding the molecular biology of cancer, drug target identification and evaluation, biomarker discovery for cancer diagnosis and prognosis, etc. over the past decade (1). PCa is the second most frequently diagnosed cancer in men worldwide with 1,414,259 new cases and 375,304 new deaths in 2020 (2). While The Cancer Genome Atlas Prostate Adenocarcinoma (TCGA-PRAD) project has produced valuable RNA sequencing (RNAseq) data based on the clinical samples from 498 PCa patients, most of these samples are primary tumor tissue or tumor-adjacent normal tissue (3, 4). Many other transcriptomics data of prostatic tissues from normal, primary tumor, and metastatic tumor samples from different prostate population cohorts have also been generated in recent years (5–8). The vast amount of the publicly available transcriptomics data for clinical samples provides extraordinary opportunities for the study of the heterogeneity in PCa, understanding the mechanisms of tumor initiation and progression, as well as identification and independent validation of prognostic signatures. However, because these transcriptomics datasets for PCa were generated using different technologies, processed with different bioinformatics pipelines, and deposited in different public data repositories, it remains challenging for researchers to leverage these valuable resources in their studies without a standard pipeline to download and process these datasets. Moreover, advanced programming skills and deep knowledge in bioinformatics and data science are also required for conducting a comprehensive analysis and visualization of the PCa transcriptomics data.

To fill this void, we present to the PCa research community a user-friendly database, PCaDB, for a comprehensive and interactive analysis and visualization of transcriptomics data from large PCa cohorts. A comprehensive bioinformatics pipeline is developed to download and process the gene expression data and metadata for 50 transcriptomics datasets with a total of 7,231 samples from public data repositories. A single-cell RNA-sequencing (scRNAseq) dataset for normal

human prostates has been included in PCaDB, allowing for the investigation of gene expression in different cell types of prostate (9). Moreover, PCaDB included 30 published PCa prognostic signatures which have been analyzed in a previous study for a comprehensive evaluation of the performances of machine learning models and prognostic signatures (10). A suite of advanced analytical and visualization tools is available for the comprehensive analysis of the transcriptomics data, such as characterization of a gene of interest across multiple bulk gene expression datasets and the scRNAseq dataset, functional characterization and evaluation of the published prognostic signatures to identify the most promising ones for further validations in prospective clinical studies, and whole-transcriptome data analysis for the identification of genes associated with tumor initiation and progression, discovery of biomarkers associated with clinical outcomes, as well as development and validation of prognostic signatures and models for PCa prognosis.

DATA COLLECTION AND PROCESSING

Collection of bulk transcriptomics data, single-cell RNAseq data, gene expression-based prognostic signatures, and gene annotation data

To obtain a complete list of transcriptomics data of PCa, we conducted a comprehensive search in public data repositories, including the National Cancer Institute (NCI) Genomic Data Commons (GDC) (11), cBioportal (12), National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (13), and ArrayExpress (14). The identification of datasets in GDC and cBioPortal were straightforward since the relevant data were grouped by cancer types. The keywords ‘prostate cancer’, ‘prostate tumor’, or ‘PCa’, AND ‘gene expression’, ‘mRNA’, ‘RNAseq’, ‘transcriptomics’, or ‘transcriptome’ were used to search the GEO and ArrayExpress databases. Many other datasets were identified in previous publications (15–18).

The following criteria were used for dataset selection: (*i*) the samples in the dataset must be collected from human PCa patients; (*ii*) the number of samples

should be greater than 15; and (*iii*) the dataset must be generated using a genome-wide gene expression profiling platform. A total of 50 PCa transcriptomics datasets with 7,231 samples were identified based on the three criteria. The 7,231 samples included 5,179 (71.6%) primary tumor samples, 879 (12.2%) samples from healthy or tumor-adjacent normal tissues, 716 (9.9%) metastatic samples, and 457 (6.3%) other types of samples such as stromal and whole blood samples, etc.

A single-cell RNAseq dataset for cell types isolated by fluorescence-activated cell sorting (FACS) from normal human prostates (9) was downloaded from the GenitoUrinary Development Molecular Anatomy Project (GUDMAP) database (19) and integrated into the PCaDB database, allowing for the investigation of gene expression at the single-cell level.

The gene lists of 30 published prognostic signatures for PCa, which were comprehensively evaluated in a previous study (10), were included in the PCaDB database for a more detailed functional characterization and prognostic performance comparison in each of the transcriptomics datasets.

We collected the gene annotation from the Ensembl (20), GENCODE (21), HUGO Gene Nomenclature Committee (HGNC) (22), and NCBI Entrez Gene (23) databases and developed a pipeline to map the different types of the gene identifiers including the Ensembl ID, HGNC approved gene symbol and alias symbol, as well as Entrez ID to facilitate the harmonization of transcriptomics data collected from different resources and the query of genes in PCaDB.

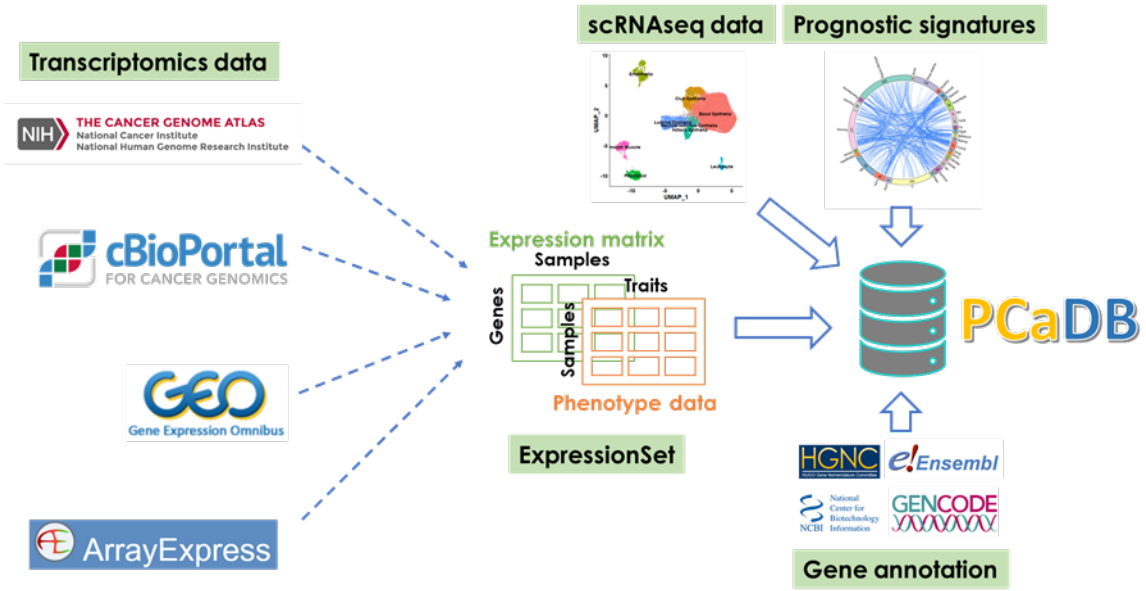

Figure 1. Overview of the collection of bulk transcriptomics data, scRNAseq data, gene expression prognostic signatures, and gene annotation data for PCaDB.

Data processing for the bulk transcriptomics data and scRNAseq data

A comprehensive pipeline was created to download and process the transcriptomics data and the associated metadata from the public data repositories, including GDC, cBioPortal, GEO, and ArrayExpress. Generally, gene expression data generated by RNAseq and Affymetrix microarray platforms were reprocessed if raw data (*i.e.*, FASTQ or CEL files) were available. Otherwise, the processed data such as FPKM values for RNAseq and normalized intensities for microarray data were downloaded directly from the data repositories and a simple log2 transformation may be performed if this hasn't been applied to the original data. The Ensembl gene identifiers were used for all the gene expression data. If multiple probes/genes matched to the same Ensembl ID, only the one with the maximum interquartile range (IQR) for the gene expression was used for this Ensembl ID. Metadata associated with the samples were obtained directly from the public data repositories and harmonized using a custom script followed by a careful manual curation. We created a comprehensive list of 31 field names, including sample id, patient id, tissue, batch, sample type, age at diagnosis,

ethnicity, race, clinical stages, pathology stages, preoperative PSA, Gleason score, overall survival, relapse-free survival, treatment, etc., for each sample. The complete list of the field names can be found in the 'Pipeline' page of PCaDB. An additional column called 'pcadb_group' was added for the standardized sample type information. For example, the 'tumor', 'tumour', 'primary', 'localized' samples, etc. were all labeled as 'Primary', whereas the 'normal', 'adjacent', 'benign' samples, etc. were all labeled as 'Normal'. The *ExpressionSet* object was created for each dataset and deposited into the PCaDB database. Details about the pipeline for downloading and processing data generated from different platforms or obtained from different repositories was describe below.

The TCGA-PRAD data were downloaded and processed using a series of functions in the R package *GDCRNATools* (24). The raw HTSeq-Counts data was normalized using the Trimmed Mean of M values (TMM) method implemented in the R package *edgeR* (25). Clinical characteristics, such as preoperative PSA level, which were not available in GDC were retrieved from Broad GDAC Firehose (https://gdac.broadinstitute.org/).

The expression data and clinical data can be downloaded directly from cBioPortal. The cBioPortal data were usually not used in PCaDB, unless the raw data were not available in any of the other data repositories. This is because only the normalized data are provided in cBioPortal, which may be incompatible with some downstream analytical tools. For example, the FPKM/RPKM values are not recommended for differential expression analysis. The processing of the expression data from cBioPortal was relatively simple. Usually the log2 transformation may be performed on the RPKM/FPKM values if it hasn't been done and the gene symbols or Entrez IDs were converted to Ensembl IDs.

Most of the microarray data in PCaDB were obtained from GEO. All the Affymetrix microarray data were reprocessed starting from the raw .CEL files. The annotation R packages for the probes were downloaded from the Brainarray database (GENCODEG, Version 24) (26). The expression data were normalized using the Robust Multichip Average (RMA) algorithm implemented in the R

package *oligo* (27). The RNAseq data in GEO were downloaded from Sequence Read Archive (SRA) if the raw FASTQ files were available. We used *fasterq-dump* in the SRA Toolkit (version 2.10.8) to download the raw sequencing data. The *STAR* (version 2.7.2a) (28) was used for sequence alignment and *featureCounts* (version 2.0.0) (29) was used for gene expression quantification. Similar to the TCGA-PRAD data, the count data was normalized using the TMM method implemented in the R package *edgeR* (25). The normalized intensities for microarray data generated using other technologies were downloaded directly from GEO and a simple log2 transformation may be performed. Metadata in the series matrix files were downloaded using the R package *GEOquery* (30).

Microarray gene expression data and metadata from ArrayExpress were downloaded using bash command lines. The gene expression data can be processed using the same pipeline as those from GEO, and the metadata were harmonized using a custom script followed by manual curation.

For the scRNAseq dataset, the *Seurat* object can be downloaded directly from the GUDMAP (https://www.gudmap.org/chaise/record/#2/RNASeq:Study/RID=W-RAHW). The normalized gene expression matrix, cell type annotation, as well as the TSNE and UMAP coordinates were included in an R list object.

Integration of gene annotation data

It's very critical to map different types of gene identifiers from different databases for the data analysis in a transcriptome atlas and for the integrated analysis of multiple datasets, especially when the data were generated using different platforms, processed with different bioinformatics pipelines, or collected from different resources. The Ensembl gene annotation (Release 98) was downloaded using the R package *biomaRt* (31). The GTF file for the GENCODE gene annotation (Release 32) was downloaded from the FTP site (http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_32/). The two files hgnc_complete_set.txt and withdrawn.txt for the HGNC gene annotation were downloaded from the FTP site

(http://ftp.ebi.ac.uk/pub/databases/genenames/new/tsv/), and the two files gene_info.gz and gene_history.gz for the NCBI gene annotation were downloaded from the FTP site (https://ftp.ncbi.nih.gov/gene/DATA/) on 12/30/2019. A comprehensive pipeline was developed to map the gene IDs and the source code is publicly available at https://github.com/rli012/PCaDB and on the 'PCaDB Pipeline' page in the database.

DATABASE CONTENT AND USAGE

PCaDB provides a user-friendly interface and a suite of analytical and visualization tools for the comprehensive analysis of gene expression data at three levels: (*i*) query an individual gene of interest, (*ii*) characterization of prognostic signatures, and (*iii*) whole-transcriptome data analysis (Figure 2). Multiple analytical and visualization functions can be implemented in each module.

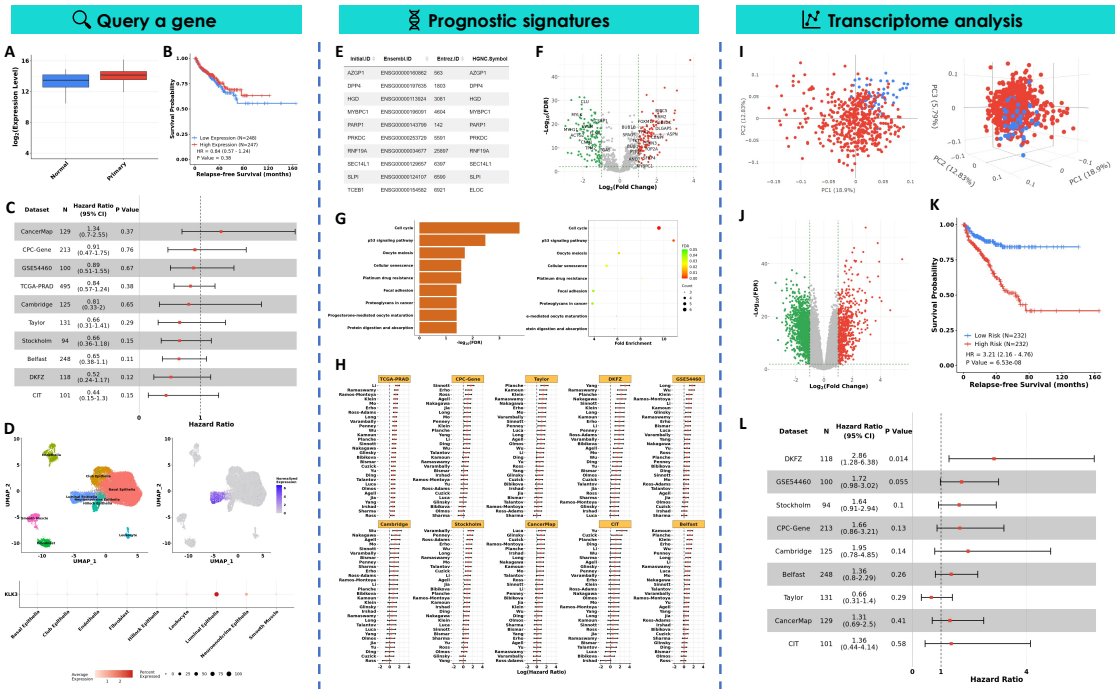

Figure 2. Comprehensive analysis of gene expression data at three levels in PCaDB, including the query of an individual gene of interest, characterization of prognostic signatures, and whole-transcriptome data analysis. (**A**) The box plot to visualize the differential expression of the gene of interest in different sample types. (**B**) The KM survival curve of RFS for the selected gene in an individual dataset.

(**C**) The forest plot to visualize the association of a gene of interest with RFS across multiple datasets. (**D**) Visualization of the gene expression in different prostate cell types using the UMAP plot and the bubble plot. (**E**) The list of genes in the selected prognostic signature. (**F**) Differential gene expression analysis of genes in all the prognostic signatures between tumor and normal samples. (**G**) Visualization of the functional enrichment analysis results using the bar plot and bubble plot. (**H**) Comprehensive evaluation of the performances of all the prognostic signatures across multiple datasets based a selected training dataset and a selected machine learning algorithm. (**I**) 2D and 3D interactive visualization of the principal component analysis result for a transcriptomics dataset. (**J**) Differential gene expression analysis for the whole transcriptome profiling data between tumor and normal samples. (**K**) A new gene expression-based prognostic model for RFS prediction can be developed by selecting a training dataset and a survival analysis algorithm, and a KM survival curve is used to evaluate the performance of the model in the training dataset. (**L**) Validation of the prognostic model leveraging gene expression data from multiple independent cohorts

Query an individual gene of interest

Users can query a gene of interest by typing the Ensembl ID, Entrez ID, or HGNC approved symbol and alias symbol in the 'Search a gene' field and selecting the gene from the dropdown list. The general information about the gene and some useful external links to the databases, such as ENSEMBL, HGNC, and NCBI for more detailed description of the gene, Genotype-Tissue Expression (GTEx) (32) and Human Protein Atlas (HPA) (33) for the gene expression pattern in different human tissues, and Kyoto Encyclopedia of Genes and Genomes (KEGG) (34) for the pathways that the gene involves in are provided. A suite of advanced analyses and visualizations can be interactively performed for the selected gene, including differential expression analysis between different types of samples, RFS survival analysis, and gene expression analysis in different prostate cell types at the single-cell level.

The box plot is used to visualize the gene expression in different sample types, such as healthy, tumor-adjacent normal, primary tumor, or metastatic tumor in different tissues, depending on the availability of the data in the selected dataset. Kaplan Meier (KM) survival analysis of RFS can be performed in the 10 datasets with 1,754 primary tumor samples from PCa patients with the data of BCR status and follow up time after RP. A forest plot with the information from the survival analysis, including the numbers of samples, hazard ratios (HRs), 95% confidence intervals (CIs), and p values across all the datasets, will be generated, and the KM survival curve for each dataset will also be plotted. The expression pattern of the selected gene in different cell types from normal human prostate, including basal, luminal, neuroendocrine (NE), club, and hillock epithelia, endothelia, leukocyte, fibroblast, and smooth muscle, can be visualized using the pre-calculated t-distributed stochastic neighbor embedding (t-SNE) plot and uniform manifold approximation and projection (UMAP) plot, violin plot, and bubble plot with the average gene expression and percent of cells expressed for each cell type.

Characterization of prognostic signatures

Gene expression-based prognostic signatures have been proven to be useful to predict the aggressiveness or clinical outcomes of PCa. Many prognostic signatures have been developed for PCa prognosis and some of them have already been used in clinical practice. A comprehensive evaluation of the prognostic performances of 30 published signatures was performed in a previous study and we included all those signatures in the PCaDB database, allowing for a more detailed characterization of the signatures, including DE analysis of the signature genes, KM survival analysis of RFS, functional enrichment analysis, and evaluation of the prognostic performances of the signatures.

On the ‘Prognostic Signatures page, the list of genes in a signature of interest can be viewed by selecting the signature from the dropdown list. The DE analysis of the signature genes between primary tumor and tumor-adjacent normal samples can be performed using the R package *limma* (35). A data table will be created to list the differentially expressed genes (DEGs) and a volcano plot will be generated

to visualize these DEGs. The KM survival analysis of RFS for the signature genes can be performed in each of the 10 datasets with RFS information. A forest plot for the common genes in 3 or more signatures and a data table with the survival analysis result for all the signature genes are generated. Functional enrichment analysis of the signature genes can be performed using the R package *clusterProfiler* (36). Many pathway/ontology knowledgebases including KEGG (34), Gene Ontology (GO) (37), Reactome (38), Disease Ontology (DO) (39), Network of Cancer Gene (NCG) (40), DisGeNET (41), and Molecular Signatures Database (MSigDB) (42) are leveraged for the functional analysis. A data table is produced to summarize the significantly enriched pathways/ontologies. A bar plot and a bubble plot are used to visualize the top enriched pathways/ontologies based on the p values adjusted by the Benjamini-Hochberg (BH) method. A more detailed evaluation of the performances of the signatures can be performed in PCaDB using different survival analysis algorithms and different training and test datasets. Users can select 'All Signatures' from the dropdown list to perform a comprehensive analysis to compare and rank the signatures in each test set based on three metrics, including concordance index (C-index), time-dependent receiver operating characteristics (ROC) curve, and hazard ratio (HR) estimated by the Kaplan Meier (KM) survival analysis. If a given signature is selected, a prognostic model can be developed using the expression data of the signature genes in the selected training dataset and the selected survival analysis method. The risk score of each patient in the test datasets will be computed based on the model, and the C-indexes, the area under the ROC curves (AUCs), and the HRs are calculated to assess the prognostic power for the signature based on the independent test cohorts. Forest plots are used to visualize the results, while data tables with more detailed results are also provided.

Whole-transcriptome data analysis

More advanced and comprehensive analyses can be performed at the whole-transcriptome level in PCaDB, allowing users to identify DEGs associated with tumor initiation and progression, identify biomarkers associated with clinical

outcomes (*i.e.*, BCR), as well as develop and validate gene expression-based signatures and models for PCa prognosis.

A transcriptome dataset of interest can be selected on the 'Transcriptome Analysis' page, and the summary of the dataset including platform, data processing pipeline, and the available metadata, such as sample type, preoperative PSA, Gleason score, BCR status, and time to BCR, will be displayed automatically. Principal component analysis (PCA) can be performed using the highly expressed genes in the selected dataset, and a 2D or 3D interactive plot based on the first two or three principal components, respectively, will be generated for visualization. The DE analysis using the whole-transcriptome data allows users to identify DEGs associated with tumor initiation or progression by comparing the case and control groups, *i.e.*, primary tumor vs. tumor-adjacent normal, or metastatic tumor vs. primary tumor, etc. The R package *limma* (35) is used to identify DEGs in PCaDB. Both the univariate Cox proportional hazards (CoxPH) and KM survival analyses of RFS can be performed at the whole-transcriptome level to identify biomarkers associated with clinical outcome of PCa in a selected dataset of interest. The biomarkers that are significant across multiple datasets may be used alone or in combination with other biomarkers to derive a prognostic signature for PCa. In PCaDB, users can provide a list of genes and select any survival analysis method, such as CoxPH, Cox model regularized with ridge penalty (Cox-Ridge), or lasso penalty (Cox-Lasso) (43), to develop a prognostic model using the selected dataset as a training set. Risk scores for the patients in the training set are calculated and the median value is used as the threshold to dichotomize these patients into low- and high-risk groups. A KM survival curve is generated to show the prognostic performance of the signature in the training set. Similarly, risk scores are calculated for patients in each of the nine remaining datasets with RFS data, and a forest plot is generated to validate the prognostic model in independent cohorts.

Data download

All the processed data, including the 50 public PCa transcriptomics datasets, the scRNAseq data for normal human prostates, the summary and the gene lists of the 30 published prognostic signatures, and the integrated gene annotation data can be downloaded easily on the 'Download' page of PCaDB. The summary of the transcriptomics datasets including the GEO, ArrayExpress, or EGA accession number, the gene expression profiling platform, the bioinformatics pipeline that was used to process the data, the original publication of each dataset, etc. is also available for downloading. The *ExpressionSet* class is used for the gene expression data and metadata of the transcriptomics datasets, and the data can be downloaded in the RDS format. The *Seurat* object of the scRNAseq data and the gene annotation data are also available for users to download in the RDS format. The summary of the transcriptomics datasets, and the summary and the gene lists of the prognostic signatures are provided as Excel files for downloading in PCaDB.

IMPLEMENTATION

PCaDB has been developed using R Shiny (https://CRAN.R-project.org/package=shiny), which provides an elegant and powerful web framework for building interactive web applications using the R language (https://www.R-project.org/). The major advantage of Shiny is that a lot of R/Bioconductor packages such as *limma* (35), *clusterProfiler* (36), *Biobase* (44), *ggplot2* (45), etc. can be used for the advanced bioinformatics analyses and visualization. In PCaDB, the majority of the visualizations are based on the R package *ggplot2*, interactive tables are generated using the R package *DT* (https://CRAN.R-project.org/package=DT), allowing users to filter, sort, copy, and download the data, and interactive plots are made using the R package *plotly* (46). The blue gradient theme in the *dashboardthemes* package (https://CRAN.R-

project.org/package=dashboardthemes) is used with some modifications. The web application is deployed on Amazon Web Services (AWS).

SUMMARY

PCaDB is a comprehensive database for transcriptomes of prostate cancer cohorts with a total of 7,231 samples from 50 public datasets. A suite of well-designed functions is provided in PCaDB for the interactive analysis and visualization of the transcriptomics data.

A scRNAseq dataset for normal human prostates is also included, allowing for the investigation of gene expression at the single-cell level. Detailed characterization and evaluation of 30 published prognostic signatures can also be performed in PCaDB to identify the most promising ones for further validations in prospective clinical studies. All the data are processed with a comprehensive pipeline and the data can be easily downloaded from the database. The pipelines can also be used by the users to process the datasets of interests from the public data repositories. While PCaDB is diligently serving the prostate cancer research community, new datasets and analytical methods will be included in PCaDB as soon as they are available. We expect that PCaDB would become a valuable online resource for a comprehensive analysis of PCa transcriptomics data to understand the molecular mechanisms of tumor initiation and progression, and to identify and validate biomarkers and signatures for PCa prognosis.

AVAILABILITY

The web interface to PCaDB is publicly available at http://bioinfo.jialab-ucr.org/PCaDB/. All the processed data can be downloaded on the 'Download' page of the database. The pipelines used to process the data are available at https://github.com/rli012/PCaDB and on the 'PCaDB Pipeline' page in the PCaDB database.

FUNDING

This work was supported by Z.J.'s UC Riverside Faculty Start-up Fund and UC Cancer Research Coordinating Committee Competition Award.

DISCLOSURE DECLARATION

The authors declare that they have no competing interests.

REFERENCES

1. Lovén,J., Orlando,D.A., Sigova,A.A., Lin,C.Y., Rahl,P.B., Burge,C.B., Levens,D.L., Lee,T.I. and Young,R.A. (2012) Revisiting Global Gene Expression Analysis. *Cell*, **151**, 476–482.

2. Sung,H., Ferlay,J., Siegel,R.L., Laversanne,M., Soerjomataram,I., Jemal,A. and Bray,F. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, **71**, 209–249.

3. Abeshouse,A., Ahn,J., Akbani,R., Ally,A., Amin,S., Andry,C.D., Annala,M., Aprikian,A., Armenia,J., Arora,A., *et al.* (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, **163**, 1011–1025.

4. Liu,J., Lichtenberg,T., Hoadley,K.A., Poisson,L.M., Lazar,A.J., Cherniack,A.D., Kovatich,A.J., Benz,C.C., Levine,D.A., Lee,A.V., *et al.* (2018) An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, **173**, 400-416.e11.

5. Stelloo,S., Nevedomskaya,E., Kim,Y., Schuurman,K., Valle-Encinas,E., Lobo,J., Krijgsman,O., Peeper,D.S., Chang,S.L., Feng,F.Y.-C., *et al.* (2018) Integrative epigenetic taxonomy of primary prostate cancer. *Nat Commun*, **9**, 4900.

6. Sinha,A., Huang,V., Livingstone,J., Wang,J., Fox,N.S., Kurganovs,N., Ignatchenko,V., Fritsch,K., Donmez,N., Heisler,L.E., *et al.* (2019) The Proteogenomic Landscape of Curable Prostate Cancer. *Cancer Cell*, **35**, 414-427.e6.

7. Gerhauser,C., Favero,F., Risch,T., Simon,R., Feuerbach,L., Assenov,Y., Heckmann,D., Sidiropoulos,N., Waszak,S.M., Hübschmann,D., *et al.* (2018) Molecular Evolution of Early-

Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer Cell*, **34**, 996-1011.e8.

8. Abida,W., Cyrta,J., Heller,G., Prandi,D., Armenia,J., Coleman,I., Cieslik,M., Benelli,M., Robinson,D., Van Allen,E.M., *et al.* (2019) Genomic correlates of clinical outcome in advanced prostate cancer. *Proc Natl Acad Sci U S A*, **116**, 11428–11436.

9. Henry,G.H., Malewska,A., Joseph,D.B., Malladi,V.S., Lee,J., Torrealba,J., Mauck,R.J., Gahan,J.C., Raj,G.V., Roehrborn,C.G., *et al.* (2018) A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. *Cell Reports*, **25**, 3530-3542.e5.

10. Li,R. and Jia, Z. (2021) Comprehensive evaluation of machine learning models and gene expression signatures for prostate cancer prognosis using large population cohorts. *bioRxiv*.

11. Jensen,M.A., Ferretti,V., Grossman,R.L. and Staudt,L.M. (2017) The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, **130**, 453–459.

12. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,S.O., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E., *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*, **6**, pl1.

13. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207–210.

14. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G., *et al.* (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **31**, 68–71.

15. Li,R., Wang,S., Cui,Y., Qu,H., Chater,J.M., Zhang,L., Wei,J., Wang,M., Xu,Y., Yu,L., *et al.* (2021) Extended application of genomic selection to screen multiomics data for prognostic signatures of prostate cancer. *Briefings in Bioinformatics*, **22**.

16. Yang,L., Roberts,D., Takhar,M., Erho,N., Bibby,B.A.S., Thiruthaneeswaran,N., Bhandari,V., Cheng,W.-C., Haider,S., McCorry,A.M.B., *et al.* (2018) Development and Validation of a 28-gene Hypoxia-related Prognostic Signature for Localized Prostate Cancer. *EBioMedicine*, **31**, 182–189.

17. You,S., Knudsen,B.S., Erho,N., Alshalalfa,M., Takhar,M., Al-Deen Ashab,H., Davicioni,E., Karnes,R.J., Klein,E.A., Den,R.B., *et al.* (2016) Integrated Classification of Prostate Cancer Reveals a Novel Luminal Subtype with Poor Outcome. *Cancer Res*, **76**, 4948–4958.

18. Luca,B.-A., Moulton,V., Ellis,C., Connell,S.P., Brewer,D.S. and Cooper,C.S. (2020) Convergence of Prognostic Gene Signatures Suggests Underlying Mechanisms of Human Prostate Cancer Progression. *Genes (Basel)*, **11**, E802.

19. Harding,S.D., Armit,C., Armstrong,J., Brennan,J., Cheng,Y., Haggarty,B., Houghton,D., Lloyd-MacGilp,S., Pi,X., Roochun,Y., *et al.* (2011) The GUDMAP database--an online resource for genitourinary research. *Development*, **138**, 2845–2853.

20. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., *et al.* (2020) Ensembl 2020. *Nucleic Acids Research*, **48**, D682–D688.

21. Frankish,A., Diekhans,M., Jungreis,I., Lagarde,J., Loveland,J.E., Mudge,J.M., Sisu,C., Wright,J.C., Armstrong,J., Barnes,I., *et al.* (2021) GENCODE 2021. *Nucleic Acids Research*, **49**, D916–D923.

22. Tweedie,S., Braschi,B., Gray,K., Jones,T.E.M., Seal,R.L., Yates,B. and Bruford,E.A. (2021) Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research*, **49**, D939–D946.

23. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, **33**, D54–D58.

24. Li,R., Qu,H., Wang,S., Wei,J., Zhang,L., Ma,R., Lu,J., Zhu,J., Zhong,W.-D. and Jia,Z. (2018) GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. *Bioinformatics*, **34**, 2515–2517.

25. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

26. Dai,M., Wang,P., Boyd,A.D., Kostov,G., Athey,B., Jones,E.G., Bunney,W.E., Myers,R.M., Speed,T.P., Akil,H., *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, **33**, e175–e175.

27. Carvalho,B.S. and Irizarry,R.A. (2010) A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, **26**, 2363–2367.

28. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

29. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

30. Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.
31. Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, **4**, 1184–1191.
32. The GTEx Consortium. (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
33. Uhlen,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S., *et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol*, **28**, 1248–1250.
34. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, **45**, D353–D361.
35. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47–e47.
36. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, **16**, 284–287.
37. The Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, **49**, D325–D334.
38. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R., *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res*, **48**, D498–D503.
39. Schriml,L.M., Mitraka,E., Munro,J., Tauber,B., Schor,M., Nickle,L., Felix,V., Jeng,L., Bearer,C., Lichenstein,R., *et al.* (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research*, **47**, D955–D962.
40. Repana,D., Nulsen,J., Dressler,L., Bortolomeazzi,M., Venkata,S.K., Tourna,A., Yakovleva,A., Palmieri,T. and Ciccarelli,F.D. (2019) The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol*, **20**, 1.
41. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J., Ronzano,F., Centeno,E., Sanz,F. and Furlong,L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, **48**, D845–D855.

42. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

43. Simon,N., Friedman,J., Hastie,T. and Tibshirani,R. (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw*, **39**, 1–13.

44. Huber,W., Carey,V.J., Gentleman,R., Anders,S., Carlson,M., Carvalho,B.S., Bravo,H.C., Davis,S., Gatto,L., Girke,T., *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*, **12**, 115–121.

45. Wickham,H. (2009) ggplot2: Elegant Graphics for Data Analysis Springer-Verlag, New York.

46. Sievert,C. (2020) Interactive Web-Based Data Visualization with R, plotly, and shiny CRC Press.