doi:10.1093/jmcb/mjab004
Published online January 25, 2021

Application Note

IPPC: an interactive platform for prostate cancer multi-omics data integration and analysis

Prostate cancer is a clinically heterogeneous disease and remains the most common non-skin malignancy in men worldwide (Abate-Shen and Shen, 2000; Litwin and Tan, 2017), which is often diagnosed through screening with digital rectal examinations and quantitation of serum levels of prostate-specific antigen (PSA). At the morphological aspect, the Gleason scoring system is regarded as the most reliable and predictive histological grading system (Welch and Albertsen, 2009; Heidenreich et al., 2011). Interpatient genomic heterogeneity in prostate cancer is well recognized; however, molecular stratification of prostate cancer to guide treatment selection based on predictive genomic biomarkers remains an unmet clinical need (Prensner et al., 2012; Topalian et al., 2016).

Developments in next-generation sequencing technologies have increased the speed and reduced the cost of sequencing for cancer samples. And many molecular characterization efforts such as The Cancer Genome Atlas (TCGA) database have unlocked opportunities to characterize the genomic and transcriptomic landscapes of cancer for basic science and clinical oncology research (Ceram'i et al., 2012). Besides, an enormous amount of omics data from independent prostate cancer studies have been deposited into the NCBI Gene Expression Omnibus (GEO) database (Barrett et al., 2013). The integration of these databases provides an unprecedented opportunity for prostate cancer genomics research. However, an overwhelming amount of multi-omics data from various technical platforms make it increasingly challenging to perform data exploration, analytics, and visualization, especially for scientists without a computational background. Hence, an user-friendly integrated database must be urgently established for the retrieval, integration, and analysis of big data in prostate cancer studies. In this study, we curated the multi-omics data from various data resources and integrated the customized analysis tools to develop an interactive platform, called IPPC, to provide mutations profile, survival probability, gene expression, co-expression, immune infiltration, miRNA–target association, and single-cell sequencing analyses for query genes. We collected a mass of data from a total of 14134 samples in 66 datasets under strict quality control and uniform processing. The IPPC web-server will facilitate the exploration of multidimensional genomics data of prostate cancer by allowing analysis and visualization across genes, samples, and data types.

IPPC web-server collected and integrated multi-omics data from public databanks by using the keywords: 'PCa', 'PC', and 'prostate cancer' (before October 30, 2020). To ensure the quality and reliability of data warehousing, we screened, collected, and filtered the data with the following criteria: (i) gene microarray or high-throughput sequencing data extracted from prostate cancer or adjacent normal tissues; (ii) the sample size of each dataset is no less than 50; (iii) for multi-duplicated or same annotated samples, we adapted the quality control to delete the redundant samples; and (iv) the clinical information of samples contains key features: age, race, tumor site, Gleason score, PSA, clinical status, pathologic status, and sample type. In total, we obtained 66 independent datasets including 44 datasets from GEO, 2 datasets from TCGA, and 20 datasets of cBioProtal (Figure 1). We also incorporated the 329 KEGG pathways and 18176 gene ontology terms for gene annotation, and 380639 miRNA–target interactions for gene–miRNA expression association analysis. The dataset screening needs to be affirmed again by another researcher. The quality control was performed on each of the 44 GEO datasets by using the 'simpleaffy' and 'affyPLM' packages of R software (https://www.r-project.org/). The raw data for all the datasets were normalized, summarized, and log-transformed through the robust multi-array average function of 'affy' R package. The probe-based expression of genes was converted into gene expression profiles, and the gene containing multiple probes reserved the probes with the largest interquartile ranges while giving up others. For the RNA–seq data, all the samples of the remaining datasets were normalized with fragments per kilobase million (FPKM) or reads per kilobase million (RPKM) and then log-transformed. The somatic mutation data of prostate cancer were filtered to exclude the inappropriate molecular consequences including intergraded variants, non-coding variants, or intron variants and to remain nonsynonymous mutations like missense variants. We compiled a series of clinical features,

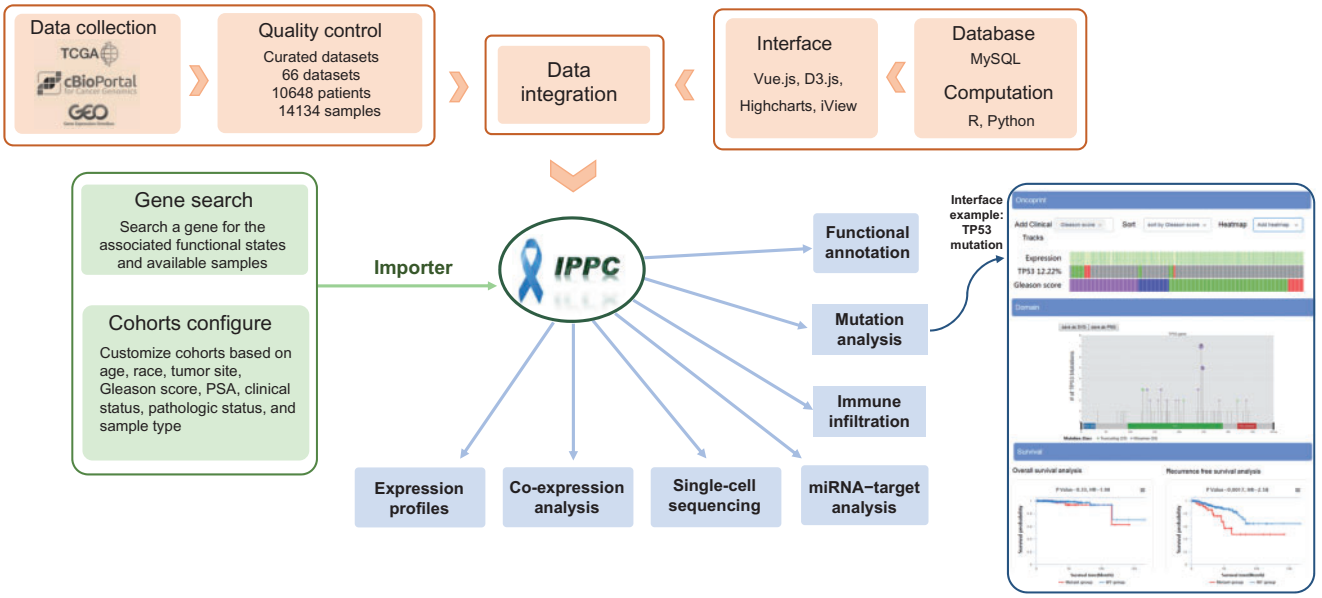

Figure 1 Illustration of study workflow. The flowchart of data collection, data curation, database implementation, analysis modules, and interface example in this work.

i.e. risk factors along with each sample for subgrouping that contain age, Gleason score, PSA, race, clinical status, pathologic status, tumor site, and sample type with each subgroup corresponding to the detailed, accepted standard stage.

The website interface IPPC comprises five sections, 'Home', 'Search', 'Help', 'FAQ', and 'Contact us', and is designed to be a gene-centered platform that provides seven analysis modules: gene functional analysis (including KEGG pathway and gene ontology annotation), genomic mutation profile analysis, expression profile analysis, co-expression analysis, tumor immune-infiltration analysis, miRNA regulation analysis based on miRNA–target interactions, and single-cell sequencing data analysis (Figure 1). IPPC webserver provides a user-friendly and concise interface consisting of four steps: inputting gene name, customizing patient cohorts, searching, and showing detailed results. A gene symbol or gene alias could be input in the text field where we provide a fuzzy matching function. Then the users could customize their personalized subgroups of sample cohorts by using single or combined clinical features. IPPC provides a total of 8 clinical features including age, Gleason score, PSA, race, clinical status, pathologic status, tumor site, and sample type. Then the third step, the IPPC will provide the list of all the available datasets for the user to select. Finally, the results will display seven analysis modules corresponding with graphical and tabular presentation and analysis results.

The IPPC web-server was specifically designed to lower the barriers of access to the complex multi-omics data of prostate cancer and thereby accelerate the translation of genomic data into new biological insights, therapies, and clinical trials. It is a free, intuitive, and interactive tool for tapping the full potential of publicly available prostate cancer genomics data, which enables biologists and clinicians without any programming experience to obtain ready-to-use multi-omics data and perform a diverse range of data analyses. In the future, we will not only continuously update multi-omics data from both prostate cancer and normal samples but also plan to add several new features, including complete support for protein and phosphoprotein data, interactive analysis for methylation value, summary reports for individual and customized cancer studies, and further extensions to the multi-genes query analysis. We believe that IPPC webserver would facilitate a better understanding of molecular data and clinical attributes from large-scale cancer genomics projects and empower researchers and clinicians to translate these rich datasets into biologic insights and clinical applications.

[The IPPC is freely accessible at http://ippc.bio-it.cn/. This research was supported by the Center for Health Statistics and Information, National Health Commission of China (2127000141). X.Y. and F.P. performed the analysis and prepared the manuscript. J.L., H.Z., W.C., and H.W. helped to perform the analysis and design the system architecture. P.Z. and X.H. conceived the study, supervised the project, and revised the manuscript.]

Xiongjun Ye[1,†], Fujun Peng[2,3,†], Jun Liu[1], Haiyue Zhao[1], Weinan Chen[1], Huanrui Wang[1], Peng Zhang[4,*], and Xiaobo Huang[1,*]

[1]Urology and Lithotripsy Center, Peking University People's Hospital, Beijing 100034, China
[2]Center for Systems Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China
[3]Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences, Suzhou 215123, China
[4]University of Maryland School of Medicine, Baltimore, MD 21201, USA
[†]These authors contributed equally to this work.

*Correspondence to: Peng Zhang, E-mail: peng.zhang@ihv.umaryland.edu; Xiaobo Huang, E-mail: huang6299@sina.com

Edited by Luonan Chen

References

Abate-Shen, C., and Shen, M.M. (2000). Molecular genetics of prostate cancer. Genes Dev. *14*, 2410–2434.

Barrett, T., Wilhite, S.E., Ledoux, P.L., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. *41*, D991–D995.

Cerami, E., Gao, J., Dogrusoz, U., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Heidenreich, A., Bellmunt, J., Bolla, M., et al. (2011). EAU guidelines on prostate cancer. Part 1: screening, diagnosis, and treatment of clinically localised disease. Eur. Urol. *59*, 61–71.

Litwin, M.S., and Tan, H.-J. (2017). The diagnosis and treatment of prostate cancer: a review. JAMA *317*, 2532–2542.

Prensner, J.R., Rubin, M.A., Wei, J.T., et al. (2012). Beyond PSA: the next generation of prostate cancer biomarkers. Sci. Transl. Med. *4*, 127rv3.

Topalian, S.L., Taube, J.M., Anders, R.A., et al. (2016). Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. Nat. Rev. Cancer *16*, 275.

Welch, H.G., and Albertsen, P.C. (2009). Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986–2005. J. Natl Cancer Inst. *101*, 1325–1329.