

RESEARCH ARTICLE

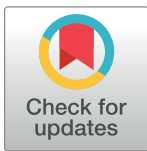
PESSA: A web tool for pathway enrichment score-based survival analysis in cancer

Hong Yang^{1,2‡}, Ying Shi^{1,3‡}, Anqi Lin^{1‡}, Chang Qi⁴, Zaoqu Liu^{5,6}, Quan Cheng^{7,8*}, Kai Miao^{9,10*}, Jian Zhang^{1*}, Peng Luo^{1*}

1 Department of Oncology, Zhujiang Hospital, Southern Medical University, Haizhu District, Guangzhou, Guangdong, China, **2** The First School of Clinical Medicine, Southern Medical University, Baiyun District, Guangzhou, Guangdong, China, **3** The Second School of Clinical Medicine, Southern Medical University, Baiyun District, Guangzhou, Guangdong, China, **4** Institute of Logic and Computation, TU Wien, Austria, **5** State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing, China, **6** State Key Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Department of Pathophysiology, Peking Union Medical College, Beijing, China, **7** Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha, Hunan, China, **8** National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China, **9** Cancer Centre and Institute of Translational Medicine, Faculty of Health Sciences, University of Macau, Macau SAR, China, **10** MoE Frontiers Science Center for Precision Oncology, University of Macau, Macau SAR, China

‡ These authors contributed equally to this work and share first authorship.

* chengquan@csu.edu.cn (QC); kaimiao@um.edu.mo (KM); zhangjian@i.smu.edu.cn (JZ); luopeng@smu.edu.cn (PL)



OPEN ACCESS

Citation: Yang H, Shi Y, Lin A, Qi C, Liu Z, Cheng Q, et al. (2024) PESSA: A web tool for pathway enrichment score-based survival analysis in cancer. *PLoS Comput Biol* 20(5): e1012024. <https://doi.org/10.1371/journal.pcbi.1012024>

Editor: Jason M. Haugh, North Carolina State University, UNITED STATES

Received: October 7, 2023

Accepted: March 26, 2024

Published: May 8, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pcbi.1012024>

Copyright: © 2024 Yang et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The datasets supporting the conclusions of this article are included within the article and its additional files. The results of PESSA's built-in data analysis are

Abstract

The activation levels of biologically significant gene sets are emerging tumor molecular markers and play an irreplaceable role in the tumor research field; however, web-based tools for prognostic analyses using it as a tumor molecular marker remain scarce. We developed a web-based tool PESSA for survival analysis using gene set activation levels. All data analyses were implemented via R. Activation levels of The Molecular Signatures Database (MSigDB) gene sets were assessed using the single sample gene set enrichment analysis (ssGSEA) method based on data from the Gene Expression Omnibus (GEO), The Cancer Genome Atlas (TCGA), The European Genome-phenome Archive (EGA) and supplementary tables of articles. PESSA was used to perform median and optimal cut-off dichotomous grouping of ssGSEA scores for each dataset, relying on the survival and survminer packages for survival analysis and visualisation. PESSA is an open-access web tool for visualizing the results of tumor prognostic analyses using gene set activation levels. A total of 238 datasets from the GEO, TCGA, EGA, and supplementary tables of articles; covering 51 cancer types and 13 survival outcome types; and 13,434 tumor-related gene sets are obtained from MSigDB for pre-grouping. Users can obtain the results, including Kaplan–Meier analyses based on the median and optimal cut-off values and accompanying visualization plots and the Cox regression analyses of dichotomous and continuous variables, by selecting the gene set markers of interest. PESSA (<https://smuonco.shinyapps.io/PESSA/> OR <http://robinl-lab.com/PESSA>) is a large-scale web-based tumor survival analysis tool covering a large amount of data that creatively uses predefined gene set activation levels as molecular markers of tumors.

available here: <https://data.mendeley.com/datasets/ph3d6sn75t/1> (DOI:10.17632/ph3d6sn75t.1).

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Author summary

The activation levels of biologically significant gene sets are emerging tumor molecular markers and play an irreplaceable role in the tumor research field; however, web-based tools for prognostic analyses using it as a tumor molecular marker remain scarce.

PESSA is an open-access web tool for visualizing the results of tumor prognostic analyses using gene set activation levels. A total of 238 datasets from the GEO, TCGA, EGA, and supplementary tables of articles; covering 51 cancer types and 13 survival outcome types; and 13,434 tumor-related gene sets are obtained from MSigDB for pre-grouping. Users can obtain the results, including Kaplan–Meier analyses based on the median and optimal cut-off values and visualization plots and the Cox regression analyses of dichotomous and continuous variables, by selecting the gene set markers of interest.

Freely available on the web at <https://smuonco.shinyapps.io/PESSA/> OR <http://robinlab.com/PESSA>. Website implemented in R Shiny, with major browsers supported.

Introduction

Cancer is a major public health problem worldwide. A total of 609,820 people are predicted to die from cancer in the United States in 2023 [1]. To reduce the threat posed by cancer, medical researchers have been exploring potentially effective therapeutic approaches and technologies to improve patient benefit from targeted therapies and immunotherapies via biomolecular markers as emerging strategies [2–4]. The widespread use of biomolecular markers in targeted cancer therapies and immunotherapy has significantly improved the clinical benefits for patients, and mortality rates for serious malignancies such as lung cancer [5], kidney cancer [6], leukaemia [7], and melanoma [2] have declined dramatically compared with those observed in the past decade, bringing hope to both oncology patients and medical researchers [1].

Studies have revealed that signalling pathways play important roles in human physiology and pathology, especially in oncology, where abnormalities in signalling pathways are closely associated with cancer cell activities, including tumorigenesis, growth and proliferation, invasion and metastasis, and apoptosis [8–12]. Currently, the level of signalling pathway activity is widely used in tumor diagnosis [13], molecular classification [14], prognostic evaluation [15] and treatment [16]. The use of overall activation levels of signal transduction pathways/gene sets as molecular markers for tumors is an emerging trend building on monogenic tumor markers [17–20]. Comprehensive analysis of the expression of multiple related genes within a signal transduction pathway/gene set can more sensitively identify subtle—but indicative of obvious—consistent functional changes in the expression level of each gene, thus addressing the limitation that conclusions based on single-gene tumor markers may overlook the overall biological role played by the pathway containing the gene [21]. In a study by Cao et al. on human bladder cancer (BLCA), it was found that EMT-related genes may be valid prognostic candidates for BLCA after considering the results of gene set variation analysis (GSVA), gene set enrichment analysis (GSEA), and survival analyses [22]; then, based on risk scores, it was concluded that patients with low risk scores had higher survival rates than those with high risk scores, consistent with the findings of Wang et al [23]. On the other hand, compared with single-gene prediction, the overall activation level of the gene set as a marker has the advantages of having lower volatility and producing more stable prediction results [24]. Based on this observation, overall gene set activation levels calculated using pathway enrichment analysis have been widely used in the prediction and assessment of cancer occurrence, progression,

and prognosis, as well as in drug discovery and development [25] and treatment [24,26–28]. This approach greatly assists clinicians and researchers in screening tumor patients, guiding the monitoring of tumor patients, and performing decision-making for personalized treatment modalities.

Currently, large publicly available databases such as The Cancer Genome Atlas (TCGA) [29] and Gene Expression Omnibus (GEO) [30] store large amounts of pancancer multiomics data. Among the included data are gene expression profiling data accompanied by detailed clinical data. However, the complexity of the database interfaces and the large number of datasets, which are difficult to select, make it difficult to fully exploit these data. On the other hand, the commonly used survival analysis methods and pathway enrichment analysis (e.g., over-representation analysis (ORA), gene set enrichment analysis (GSEA), and single-sample gene set enrichment analysis (ssGSEA)) for assessing the activation levels of gene sets are challenging for clinicians and researchers who lack a bioinformatics background [31]. Therefore, an accurate and easily accessible web-based analysis tool that provides the ability to screen tumor prognostic markers based on pathway enrichment analysis is urgently needed. Although many web tools for single-gene tumor prognostic marker exploration are available, there are a limited number of existing web tools that use the overall activation level of a gene set as a tumor marker, and these tools have some obvious limitations. The tool represented by the CeNet Omnibus application [32] breaks new ground by using a self-developed algorithm, liquid association [33], to assess the overall activation level of a gene set for prognostic analysis. However, this tool can only analyse the sequencing results of ceRNAs and miRNAs and does not cover the full range of RNA types. Moreover, tools such as ESurv [34] and BRCA-Pathway [35] use conventional Gene Ontology (GO) [36] or Kyoto Encyclopedia of Genes and Genomes (KEGG) [37] methods for pathway enrichment analysis. The ssGSEA algorithm, which is capable of performing pathway enrichment analyses on single samples with high confidence, has not been fully utilized. Finally, the above web tools either incorporate only some of the TCGA tumor datasets or support users only in uploading their data, thus failing to fully exploit the rich resources in the public databases. Thus, a web tool that incorporates pancancer transcriptome data from multiple large databases and predicts tumor markers from pathway activation levels calculated by the ssGSEA algorithm would more effectively meet the needs of oncology researchers and clinicians.

Therefore, we developed PESSA, a web tool for prognostic analysis based on the R Shiny framework that is easy to access by all biomedical scholars and provides pathway-enriched gene set activation levels as molecular markers for tumor prognosis. We plan to incorporate tumor transcriptome data from several large databases, covering a wide range of tumor types and survival outcomes. In PESSA, ssGSEA algorithms will be used to assess the level of pathway activation in individual samples, and multiple survival analysis methods and custom visualizations will be available on the associated web pages. We hope that PESSA will become a useful tool for oncology researchers and clinicians.

Materials and methods

Transcriptome data collection

We collected human tumor-related transcriptome data from three large publicly available databases, GEO, TCGA, and EGA; and supplementary tables of articles. We obtained microarray datasets from GEO via GEOmirror (0.2.0) [38] and GEOquery (2.60.0) [31] and retrieved high-throughput sequencing datasets via direct download from the GEO webpage. We obtained the high-throughput sequencing dataset from TCGA via the UCSC Xena Browser. All of the above datasets had accompanying detailed clinical information and survival

information for the represented patients. We excluded datasets with incomplete clinical information related to survival outcome and survival time and those with a sample size that was too small (<30) and finally included 180 GEO (S1 Table), 33 TCGA (S2 Table), and a total of 25 immune checkpoint inhibitor therapy-related pathway datasets from sources such as EGA, GEO, supplementary tables of articles, and others (S3 Table) datasets. All expression profiling data were subjected to $\log_2(\text{exp}+1)$ transformation and normalization (between-array normalization) via the `normalizeBetweenArrays` function in the `limma` package (3.48.3) [39]; this function is used for batch effect removal and between-sample normalization, which removes between-group differences. Standardization is adjusted from zero before conversion to a logarithmic scale to avoid missing values or large variance. Studies have shown that standardization in the median range allows for reconciliation between noise and bias [40]. For survival data, survival times were standardized to be in months. Samples missing corresponding survival information were removed before proceeding to the next step of the analysis.

ssGSEA evaluation

The Molecular Signatures Database (MSigDB; <https://www.gsea-msigdb.org/gsea/msigdb>) [41] is one of the most widely used and comprehensive gene set databases for gene set enrichment analyses. Since the creation of MSigDB, the cancer-associated gene sets have been continuously updated and expanded, and MSigDB now includes >10,000 gene sets. These gene sets more completely represent a wider range of biological processes and diseases [42]. The R package `msigdb` was used to obtain data on predefined gene sets in MSigDB, including hallmark gene sets (Hallmarks), conventional biological pathway gene sets (C2CP, which includes relevant gene sets sourced from the BioCarta, KEGG, Pathway Interaction Database, Reactome, and WikiPathways databases), and a GO database-sourced tumor-related gene set (C5GO). Furthermore, the R package `GSEA` [43] and `GSEABase` were used for pathway enrichment analysis of the obtained transcriptome sequencing data by the ssGSEA [44] method. Gene expression values for a given sample were normalized by sorting, and enrichment scores were generated using the empirical cumulative distribution function (ECDF) for the genes in the signature and the remaining genes. We finally converted the transcript expression levels into ssGSEA scores to derive the activation level of each gene set in individual samples of different types of tumors.

Core functionality: Survival analysis

The ssGSEA score allows functional gene set activation levels to be used as molecular markers of tumor prognosis for survival analysis, and the results are presented to the user in an interactive visual format that is the core functionality of PESSA. First, the various gene set ssGSEA scores obtained by preprocessing each dataset are used for dichotomous grouping by the median (or optimal cut-off) value. We remove gene sets with extremely unbalanced groupings (sample size below 5% of the overall population or above 95% of the overall population under either grouping scheme) to avoid errors in the analysis caused by a sample size that is too small. The implementation of the optimal cut-off grouping in [45] is performed via the `surv_cutpoint` function in the R package `survminer` [46] and specifies that the sample size under either grouping scheme is not less than 5% of the overall population. The realization of the optimal cutoff value grouping relies on the method of selecting the most significant p-value [45,47]. The implementation of the survival analysis is performed via the `survfit` function within the `survival` package [48], which performs a log-rank test of survival outcome under the different ssGSEA score pregrouping schemes. We also use the `coxph` function for single-factor Cox proportional hazards regression analysis. For single-factor Cox proportional hazards

regression analysis, we perform analyses based on dichotomization of the ssGSEA scores as described above and analyses based on the ssGSEA score as a continuous variable. In addition, hazard ratios (HRs), 95% confidence intervals (95% CIs), p values as well as the p-value of the Schoenfeld residual test of Cox analysis limited to continuous variables [49], based on the `cox.zph` function in survival package [48]; calculated under different grouping schemes and via different survival analysis methods are calculated accordingly. By integrating the above analyses, we provide a table of survival analysis results focusing on the activation level of each gene set in the respective dataset as well as interactive, visualized Kaplan–Meier survival analysis results, which describe and compare the distribution of survival times between different groups unlike regression analyses to assess the effects of risk factors.

Application implementation

PESSA (<https://smuonco.shinyapps.io/PESSA/> OR <http://robinl-lab.com/PESSA>) is powered by Shiny [50] and is available online via shinyapps.io. PESSA uses the tidyverse [51] family of R packages (e.g., dplyr [52], purrr [53], and tibble) [54] to provide support for internal tabular data transformations, the DT package [55] to provide users with interactive data tables, and the survminer package [46] to present visualizations of survival analysis results. In addition, programming languages such as HTML5, CSS, and JavaScript are used to assist in web page construction and beautification. Fig 1 summarizes the PESSA build process.

Results

Data overview

PESSA is an open-access web tool for survival data visualization based on the activation levels of pathway-enriched gene sets as molecular markers of tumor prognosis. PESSA (1.0.0) currently incorporates a total of 6,360,737 records distributed across 180 GEO datasets, 33 TCGA datasets, and 25 datasets from GEO, EGA, article schedules for treatment with immunosuppressive agents, which cover data for 51 types of cancers and 13 different survival outcomes. A total of 13,434 tumor-related gene sets are obtained from MSigDB (S5 Table), and the transcriptome expression profile of each sample in the GEO, TCGA, and patients treated with immunosuppressants datasets is transformed based on the ssGSEA score.

Detailed information on all included datasets can be quickly queried in the DATA tab via an interactive table; the available data include cancer type, dataset platform, survival outcome, intervention method, and treatment details (drugs used, etc.), as well as links to the corresponding datasets. The results can be saved locally, including copying to the clipboard and saving locally in various formats such as CSV, Excel, or PDF or printing, for further research and analysis.

Documentation

Users with any questions about functions, needs, or suggestions as well as uploading customized data for analysis can provide effective feedback by the message box under the About tab and sending an email to the technician's detailed email address. All comments and suggestions will be thoughtfully considered, and we will update PESSA accordingly to meet the needs of our users. Moreover, the updated records will be synchronized with the homepage and displayed on this page. Under the FAQ section, we provide detailed answers to user queries. For example, users can check the list of all predefined gene sets covered by PESSA and their source web addresses or detailed descriptions of the 13 survival outcomes. This functionality will more helpfully assist users in conducting oncology research with PESSA.

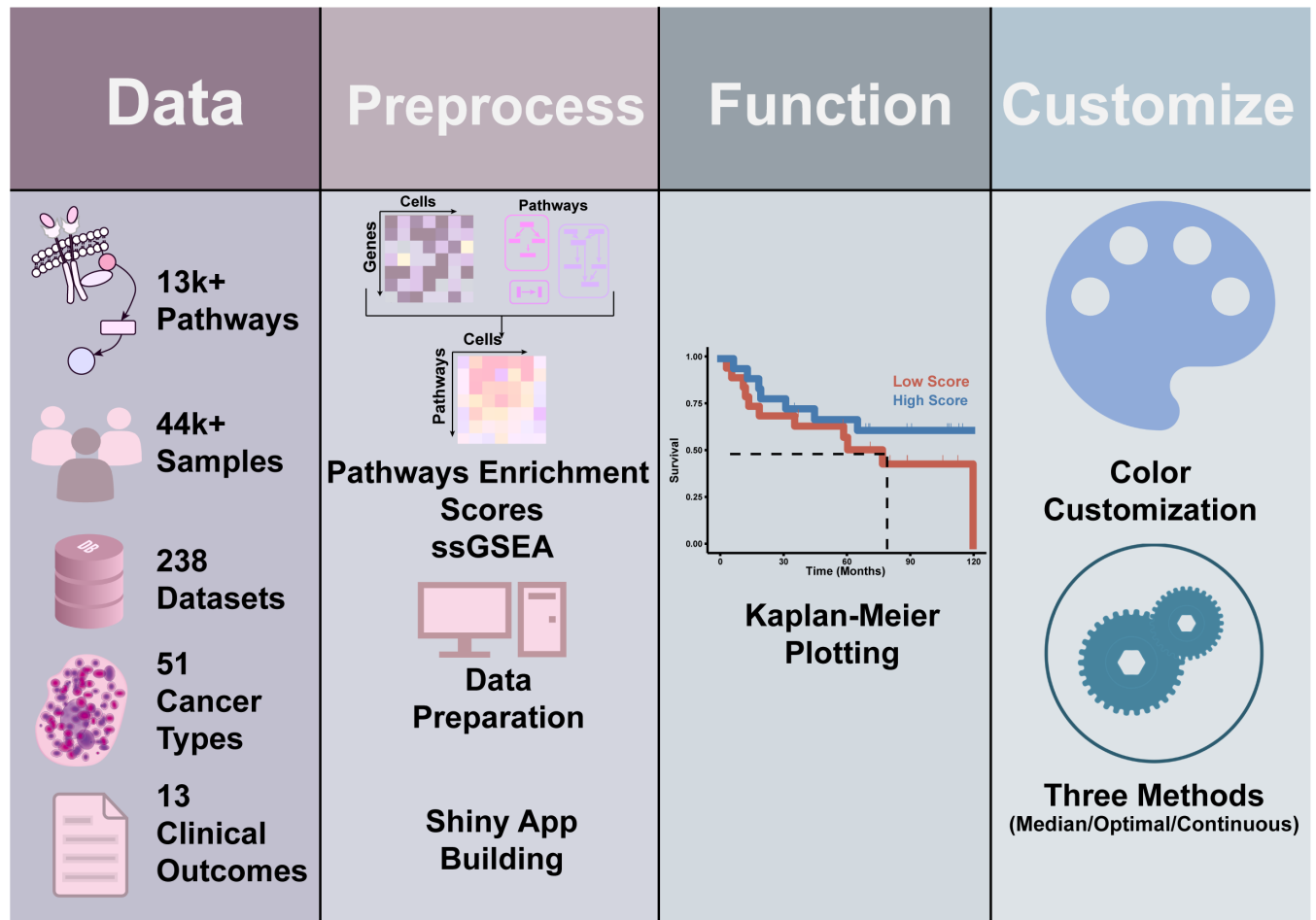


Fig 1. PESSA workflow. PESSA (v1.0.0) currently incorporates a total of 238 datasets from TCGA and GEO and a total of 13,434 predefined tumor-related gene sets from MSigDB. After data preprocessing in the background, PESSA provides the user with attractive and informative Kaplan–Meier curves and allows the user to customize the settings. Abbreviations: GEO: the Gene Expression Omnibus; MSigDB: the Molecular Signatures Database; TCGA: The Cancer Genome Atlas. Fig 1 cliparts are from Openclipart.

<https://doi.org/10.1371/journal.pcbi.1012024.g001>

Survival analysis

PESSA is a robust and powerful survival analysis tool that enables survival analysis using gene set activation levels in the pancancer domain by transforming transcriptome sequencing data into ssGSEA scores.

On the PLOT tab, PESSA guides the user through a three-part analysis process by interactively opening a collapsible box. First, the user selects/inputs the gene set of interest; second, based on the summary table of gene set-related survival analysis results displayed in the box, the user selects the target of interest and feeds it back into PESSA for visualization; and third, PESSA provides the user with interactively downloadable Kaplan–Meier plots that meet the requirements for submission to a wide range of journals.

An example of the use of PESSA is the exploitation of the activation level of the Reactome database-sourced gene set TGF-beta Receptor Activation Levels in Emt Epithelial to Mesenchymal Transition (TGF-beta Receptor Signaling in Emt Epithelial to Mesenchymal Transition), as a molecular marker for potential BLCA tumors. Survival analyses were performed using the TCGA-sourced BLCA dataset to analyse the survival outcomes of patients

dichotomized under the median ssGSEA score grouping scheme based on this pathway. The PESSA results showed that the overall survival (OS) time was significantly longer in the group with low TGF- β receptor EMT-related gene set activation levels than in the group with high corresponding levels, consistent with the results of Chen et al [56], who found that TGF- β -induced translation promotes BLCA metastasis by regulating epithelial-mesenchymal transition and invasive filopodia formation (Fig 2).

For the analysis after immunosuppressant treatment, the results obtained by PESSA were also as described previously [57]. We analyzed patients with human bladder cancer (BLCA) treated with PD1 inhibitors using HALLMARK—DNA REPAIR as the target DNA Damage Repair (DDR) pathway. The PESSA results demonstrated that patients with BLCA treated with PD1 inhibitors who exhibited a high degree of activation of the DDR pathway. Further, we analyzed the PESSA melanoma dataset using PD1 inhibitors with the same results. (Fig 3)

Discussion

PESSA is a large, interactive, and easily accessible web-based tool for clinical researchers to creatively analyse whether gene set activation levels can be used as a prognostic biomarker for cancer patients. PESSA provides a convenient opportunity for oncology researchers and clinicians to use the enormous amount of tumor transcriptome data available in GEO, TCGA, and other publicly available datasets without the need for programming skills, to annotate the predefined gene sets and ssGSEA scores in the MSigDB, to perform accurate survival analysis and to obtain attractive and informative Kaplan–Meier curves.

PESSA has advantages in many aspects, such as the breadth of predefined gene sets incorporated, the abundance of cancer types incorporated, the size of the dataset, the diversity of survival outcome information, the comprehensiveness of the analysis methods, and the user-friendliness of the customization functions. PESSA complements the scarcity of the previous survival analysis tools using the activation levels of predefined gene sets and innovatively provides a means for accurate and precise survival analysis by ssGSEA scoring of the genes that encompass the mRNA, miRNA, and long noncoding RNA (lncRNA) transcriptomes, thus expanding the data used by previous such analyses [32]. S4 Table shows a side-by-side comparison of PESSA with current common survival analysis visualization tools. It should be noted that although UCSC Xena has more advantages than PESSA in terms of data breadth and analysis function, PESSA is able to provide survival analysis based on specific gene sets, which is not available in UCSC Xena. Gene set-based analysis can help researchers to find specific pathways related to tumors, which can help researchers to discover specific biological functions of tumors from another angle. Moreover, ssGSEA has the feature of generating pathway activation scores for each sample, allowing for finer, personalized pathway associations with survival to explain heterogeneous activation between patients. Furthermore, compared to enrichment analysis methods such as GSEA [21], GSEA [43] and ORA [58], ssGSEA scores have been shown to better generalize known biology and validate experimental assays of pathway activity [44,59] as well as continuous ssGSEA scores allow for more robust survival modeling using the scores as a predictor variable as compared to dichotomous enrichment calls. The ssGSEA score has been proven by previous authors to be useful as a biomarker for survival prognosis in patients with tumors, which is helpful for tumor molecular classification [60] and treatment. For example, Zhang et al [61] applied the ssGSEA score to assess the relationship between autophagy and immune cell infiltration in hepatocellular carcinoma, innovatively suggesting that induction or inhibition of autophagy in combination with immunotherapy could be a prospective treatment strategy. Wei et al [62] suggested that autophagy inhibition could become an effective treatment strategy for hepatocellular carcinoma. Indeed, their study

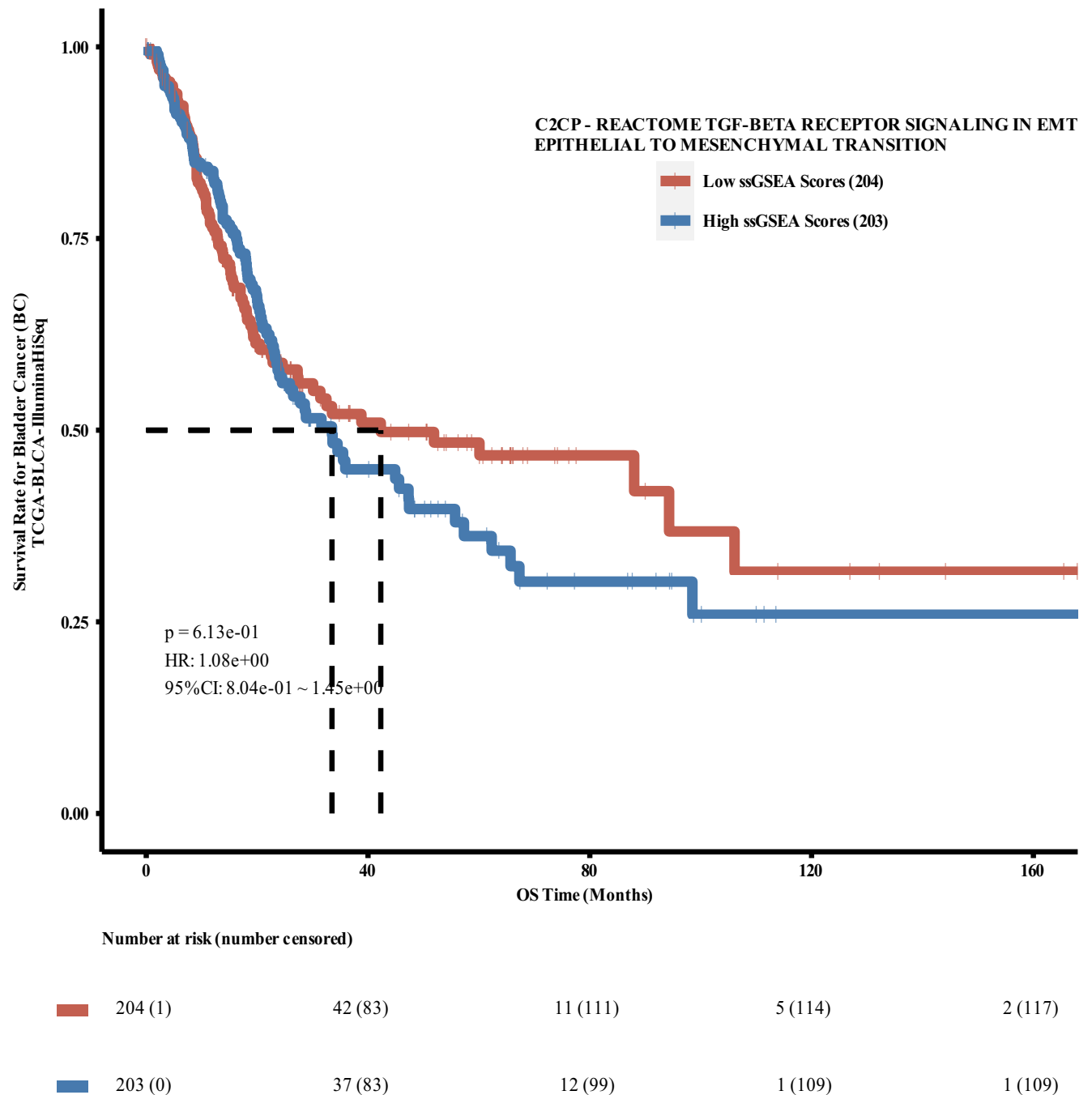


Fig 2. Survival analysis feedback Kaplan–Meier curve. Survival analysis of patients in the median-grouped TCGA-BLCA dataset using the gene set TGF-beta Receptor Activation Levels in Emt Epithelial to Mesenchymal Transition (TGF-beta Receptor Activation Levels in Emt Epithelial to Mesenchymal Transition) sourced from the Reactome database. The results from the TCGA-BLCA dataset indicate that BLCA patients with high levels of activation of the TGF-beta Receptor EMT-related gene set (TGF-beta Receptor Signalling in Emt Epithelial to Mesenchymal Transition) had significantly shorter OS times than patients with low levels of activation. Abbreviations: BLCA: Bladder Cancer; OS: Overall Survival; TCGA: The Cancer Genome Atlas.

<https://doi.org/10.1371/journal.pcbi.1012024.g002>

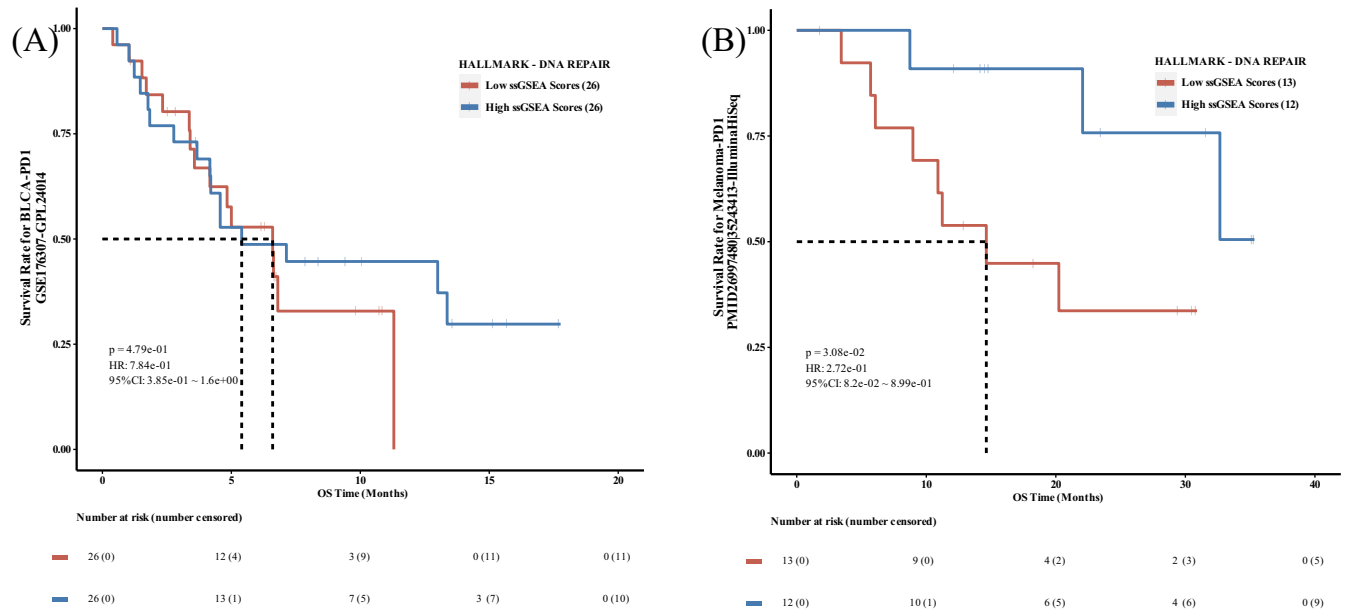


Fig 3. Feedback Kaplan-Meier curves for survival analysis based on immunosuppressive therapy data. A. Using the median dichotomous group-based dataset of patients with bladder cancer after PD1 inhibitor treatment, GSE176307, the results of the survival analysis demonstrated that the highly activated DNA damage repair pathway possessed a longer overall survival time. B. Examined using a built-in dataset derived from PMID 26997480 and 35243413 based on median dichotomous subgroups of PD1 inhibitor-treated melanoma patients, the results of the survival analyses demonstrated that the highly activated DNA damage repair pathway possessed a longer overall survival time.

<https://doi.org/10.1371/journal.pcbi.1012024.g003>

demonstrated that the m7 regulatory factor-associated ssGSEA score was an independent prognostic factor for lung adenocarcinoma (LUAD) and liver hepatocellular carcinoma (LIHC) [62]. More importantly, regarding survival analysis methods, PESSA comprehensively provides Kaplan–Meier survival analyses based on subgrouping by the median and optimal cut-off values as well as regression analyses of the abovementioned subgroups and continuous variables with a one-factor Cox proportional hazards model. This is undoubtedly one of PESSA’s major strengths. Moreover, the availability of tabulated results and the interactively customizable publisher-compliant Kaplan–Meier curve generator will satisfy the research needs of a wide range of oncology researchers and clinicians for analysing tumor prognosis.

Undeniably, there are still areas of improvement for PESSA, and we plan to complete these enhancements in subsequent releases. First, there are some limitations in the data used in PESSA. The original transcriptome sequencing data were derived only from microarray and high-throughput sequencing data in GEO, TCGA, EGA, and supplementary tables of articles; the predefined gene sets cover the commonly used gene sets related to tumors included in MSigDB. However, a large amount of potential target data—for example, data in supplementary information of published oncology articles; the Sequence Read Archive (SRA), European Bioinformatics Institute (EBI), Surveillance, Epidemiology, and End Results (SEER), and other databases; the continuously updated GEO database; and the remaining predefined gene sets included in MSigDB—have not yet been incorporated in PESSA. We plan to update these data regularly in the future. Second, PESSA cannot analyse data uploaded by users, presenting an obstacle for users with unique samples, and it is thus likely to miss key discoveries in oncology. We are currently only able to encourage users to send data via the message board to our back office, which is processed and loaded into PESSA. We will endeavor to overcome this limitation in the future. Third, PESSA has not yet been able to satisfy the need for user-defined cut-

off values and currently can only use set cut-off values preprocessed in the background, a limitation that weakens the personalized features of PESSA. Fourth, due to the lack of information on tumor grading and treatment measures in some of the sample sets, PESSA is currently unable to satisfy subgroup analysis based on tumor grading, treatment measures, and tumor histology. This may cause the emergence of analysis heterogeneity to a certain extent, which affects the accuracy of the analysis. PESSA has currently demonstrated the treatment information under Data Tabs, and we will try to break through this limitation in the future.

Currently, the data available are still limited, and we plan to actively maintain and update PESSA, regularly replenish the database-derived data, and continuously improve the functionality of PESSA. In addition, our team plans to develop additional types of web tools for tumor survival analysis, including proteomics and epigenomics tools, to help researchers and clinicians in oncology to more comprehensively and efficiently explore the molecular level of tumors at the key junctures of possible diagnosis and treatment.

In summary, PESSA is an easy-to-access and easy-to-use analysis tool that can meet the needs of biomedical practitioners in exploring the survival prognoses of different cancer types and survival outcomes using specific predefined gene set activation levels as markers. We believe that PESSA will become a stable and reliable oncological tool with continuous improvement.

Supporting information

S1 Table. The 180 GEO datasets incorporated.

(XLSX)

S2 Table. The 33 TCGA datasets incorporated.

(XLSX)

S3 Table. The 25 ICIs datasets incorporated.

(XLSX)

S4 Table. A side-by-side comparison of PESSA with current common survival analysis visualization tools.

(XLSX)

S5 Table. Gene sets included in PESSA with corresponding brief description and detail pages.

(XLSX)

Author Contributions

Conceptualization: Quan Cheng, Kai Miao, Jian Zhang, Peng Luo.

Formal analysis: Hong Yang, Ying Shi, Anqi Lin.

Resources: Hong Yang, Ying Shi, Anqi Lin, Chang Qi.

Supervision: Quan Cheng, Kai Miao, Jian Zhang, Peng Luo.

Visualization: Hong Yang, Ying Shi, Anqi Lin.

Writing – original draft: Hong Yang, Ying Shi, Anqi Lin.

Writing – review & editing: Hong Yang, Ying Shi, Anqi Lin, Chang Qi, Zaoqu Liu, Quan Cheng, Kai Miao, Jian Zhang, Peng Luo.

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin.* 2023 Jan; 73(1):17–48. <https://doi.org/10.3322/caac.21763> PMID: 36633525
2. Berk-Krauss J, Stein JA, Weber J, Polsky D, Geller AC. New Systematic Therapies and Trends in Cutaneous Melanoma Deaths Among US Whites, 1986–2016. *Am J Public Health [Internet].* 2020 Mar 19; 110(5):731–3. Available from: <https://doi.org/10.2105/AJPH.2020.305567> PMID: 32191523
3. Muthusamy B, Patil PD, Pennell NA. Perioperative Systemic Therapy for Resectable Non–Small Cell Lung Cancer. *Journal of the National Comprehensive Cancer Network [Internet].* 2022; 20(8):953–61. Available from: <https://doi.org/10.6004/jnccn.2022.7021> PMID: 35948038
4. Carlino MS, Larkin J, Long G V. Immune checkpoint inhibitors in melanoma. *The Lancet [Internet].* 2021 Sep 11; 398(10304):1002–14. Available from: [https://doi.org/10.1016/S0140-6736\(21\)01206-X](https://doi.org/10.1016/S0140-6736(21)01206-X).
5. Howlader N, Forjaz G, Mooradian MJ, Meza R, Kong CY, Cronin KA, et al. The Effect of Advances in Lung-Cancer Treatment on Population Mortality. *New England Journal of Medicine [Internet].* 2020 Aug 12; 383(7):640–9. Available from: <https://doi.org/10.1056/NEJMoa1916623> PMID: 32786189
6. Bukavina L, Bensalah K, Bray F, Carlo M, Challacombe B, Karam JA, et al. Epidemiology of Renal Cell Carcinoma: 2022 Update. *Eur Urol [Internet].* 2022; 82(5):529–42. Available from: <https://www.sciencedirect.com/science/article/pii/S0302283822026094>. <https://doi.org/10.1016/j.eururo.2022.08.019> PMID: 36100483
7. Kantarjian HM, Keating MJ, Freireich EJ. Toward the potential cure of leukemias in the next decade. *Cancer [Internet].* 2018 Nov 15; 124(22):4301–13. Available from: <https://doi.org/10.1002/cncr.31669> PMID: 30291792
8. Vaghari-Tabari M, Ferns GA, Qujeq D, Andevani AN, Sabahi Z, Moein S. Signaling, metabolism, and cancer: An important relationship for therapeutic intervention. Vol. 236, *Journal of Cellular Physiology.* John Wiley and Sons Inc; 2021. p. 5512–32.
9. Porta C, Paglino C, Mosca A. Targeting PI3K/Akt/mTOR signaling in cancer. Vol. 4 APR, *Frontiers in Oncology.* Frontiers Research Foundation; 2014.
10. Samatar AA, Poulidakos PI. Targeting RAS-ERK signalling in cancer: Promises and challenges. Vol. 13, *Nature Reviews Drug Discovery.* Nature Publishing Group; 2014. p. 928–42.
11. Yang CM, Ji S, Li Y, Fu LY, Jiang T, Meng FD. β -Catenin promotes cell proliferation, migration, and invasion but induces apoptosis in renal cell carcinoma. *Onco Targets Ther.* 2017 Feb 20; 10:711–24.
12. Jang GB, Kim JY, Cho SD, Park KS, Jung JY, Lee HY, et al. Blockade of Wnt/ β -catenin signaling suppresses breast cancer metastasis by inhibiting CSC-like phenotype. *Sci Rep.* 2015 Jul 23; 5.
13. Ye Z, Zhang H, Kong F, Lan J, Yi S, Jia W, et al. Comprehensive Analysis of Alteration Landscape and Its Clinical Significance of Mitochondrial Energy Metabolism Pathway-Related Genes in Lung Cancers. *Oxid Med Cell Longev.* 2021;2021. <https://doi.org/10.1155/2021/9259297> PMID: 34970420
14. Ashkani J, Naidoo KJ. Glycosyltransferase Gene Expression Profiles Classify Cancer Types and Propose Prognostic Subtypes. *Sci Rep.* 2016 May 20; 6. <https://doi.org/10.1038/srep26451> PMID: 27198045
15. McCubrey JA, Steelman LS, Chappell WH, Abrams SL, Wong EWT, Chang F, et al. Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance. Vol. 1773, *Biochimica et Biophysica Acta—Molecular Cell Research.* 2007. p. 1263–84. <https://doi.org/10.1016/j.bbamcr.2006.10.001> PMID: 17126425
16. Majumder S, Crabtree JS, Golde TE, Minter LM, Osborne BA, Miele L. Targeting Notch in oncology: the path forward. Vol. 20, *Nature Reviews Drug Discovery.* Nature Research; 2021. p. 125–44. <https://doi.org/10.1038/s41573-020-00091-3> PMID: 33293690
17. Wang D, Wei G, Ma J, Cheng S, Jia L, Song X, et al. Identification of the prognostic value of ferroptosis-related gene signature in breast cancer patients. *BMC Cancer [Internet].* 2021; 21(1):645. Available from: <https://doi.org/10.1186/s12885-021-08341-2> PMID: 34059009
18. Yang X, Weng X, Yang Y, Zhang M, Xiu Y, Peng W, et al. A combined hypoxia and immune gene signature for predicting survival and risk stratification in triple-negative breast cancer [Internet]. Vol. 13. 2021. Available from: www.aging-us.com.
19. Jin Y, Wang Z, He D, Zhu Y, Chen X, Cao K. Identification of novel subtypes based on ssGSEA in immune-related prognostic signature for tongue squamous cell carcinoma. *Cancer Med [Internet].* 2021 Dec 1; 10(23):8693–707. Available from: <https://doi.org/10.1002/cam4.4341> PMID: 34668665
20. Chong W, Shang L, Liu J, Fang Z, Du F, Wu H, et al. m6A regulator-based methylation modification patterns characterized by distinct tumor microenvironment immune profiles in colon cancer. *Theranostics [Internet].* 2021; 11(5):2201–17. Available from: <https://www.thno.org/v11p2201.htm>. <https://doi.org/10.7150/thno.52717> PMID: 33500720

21. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles [Internet]. 2005. Available from: www.pnas.org/cgi/doi/10.1073/pnas.0506580102.
22. Cao R, Yuan L, Ma B, Wang G, Qiu W, Tian Y. An EMT-related gene signature for the prognosis of human bladder cancer. *J Cell Mol Med*. 2020 Jan 1; 24(1):605–17. <https://doi.org/10.1111/jcmm.14767> PMID: 31657881
23. Wang H, Mei Y, Luo C, Huang Q, Wang Z, Lu GM, et al. Single-Cell Analyses Reveal Mechanisms of Cancer Stem Cell Maintenance and Epithelial–Mesenchymal Transition in Recurrent Bladder Cancer. *Clinical Cancer Research* [Internet]. 2022 Nov 15; 27(22):6265–78. Available from: <https://doi.org/10.1158/1078-0432.CCR-20-4796>.
24. Yang T, Hui R, Nouws J, Sauler M, Zeng T, Wu Q. Untargeted metabolomics analysis of esophageal squamous cell cancer progression. Vol. 20, *Journal of Translational Medicine*. BioMed Central Ltd; 2022. <https://doi.org/10.1186/s12967-022-03311-z> PMID: 35287685
25. Hennessy BT, Smith DL, Ram PT, Lu Y, Mills GB. Exploiting the PI3K/AKT pathway for cancer drug discovery. Vol. 4, *Nature Reviews Drug Discovery*. 2005. p. 988–1004. <https://doi.org/10.1038/nrd1902> PMID: 16341064
26. Levantini E, Maroni G, Del Re M, Tenen DG. EGFR signaling pathway as therapeutic target in human cancers. Vol. 85, *Seminars in Cancer Biology*. Academic Press; 2022. p. 253–75.
27. Greally M, Kelly CM, Cercek A. HER2: An emerging target in colorectal cancer. Vol. 42, *Current Problems in Cancer*. Mosby Inc.; 2018. p. 560–71.
28. Rosario SR, Long MD, Affronti HC, Rowsam AM, Eng KH, Smiraglia DJ. Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. *Nat Commun*. 2018 Dec 1;9(1). <https://doi.org/10.1038/s41467-018-07232-8> PMID: 30552315
29. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine* [Internet]. 2016 Sep 22; 375(12):1109–12. Available from: <https://doi.org/10.1056/NEJMp1607591> PMID: 27653561
30. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* [Internet]. 2002 Jan 1; 30(1):207–10. Available from: <https://doi.org/10.1093/nar/30.1.207> PMID: 11752295
31. Davis S. Get data from NCBI Gene Expression Omnibus (GEO) [Internet]. [cited 2023 Apr 15]. Available from: <https://bioconductor.org/packages/release/bioc/html/GEOquery.html>.
32. Wen X, Gao L, Song T, Jiang C. CeNet Omnibus: an R/Shiny application to the construction and analysis of competing endogenous RNA network. *BMC Bioinformatics*. 2021 Dec 1; 22(1).
33. Wen X, Gao L, Hu Y. LAceModule: Identification of Competing Endogenous RNA Modules by Integrating Dynamic Correlation. *Front Genet*. 2020 Mar 18;11. <https://doi.org/10.3389/fgene.2020.00235> PMID: 32256525
34. Pak K, Oh SO, Goh TS, Heo HJ, Han ME, Jeong DC, et al. A user-friendly, web-based integrative tool (ESurv) for survival analysis: Development and validation study. *J Med Internet Res*. 2020 May 1; 22(5). <https://doi.org/10.2196/16084> PMID: 32369034
35. Kim I, Choi S, Kim S. BRCA-Pathway: A structural integration and visualization system of TCGA breast cancer data on KEGG pathways. *BMC Bioinformatics*. 2018 Feb 19;19.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Vol. 25, *Nature Genetics*. 2000. p. 25–9.
37. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. Vol. 28, *Nucleic Acids Research*. 2000. Available from: <http://www.genome.ad.jp/kegg/>
38. Zeng J, Xia. GEOmirror [Internet]. [cited 2023 Apr 15]. Available from: <https://github.com/jmzeng1314/GEOmirror>.
39. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015; 43(7). <https://doi.org/10.1093/nar/gkv007> PMID: 25605792
40. Shi W, Oshlack A, Smyth GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res*. 2010 Dec; 38(22). <https://doi.org/10.1093/nar/gkq871> PMID: 20929874
41. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011 Jun; 27(12):1739–40. <https://doi.org/10.1093/bioinformatics/btr260> PMID: 21546393
42. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst*. 2015 Dec 23; 1(6):417–25.

43. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data [Internet]. Vol. 14, BMC Bioinformatics. 2013. Available from: <http://www.biomedcentral.com/1471-2105/14/7http://www.bioconductor.org/Background>.
44. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009 Nov 5; 462(7269):108–12. <https://doi.org/10.1038/nature08460> PMID: 19847166
45. Nagy Á, Lánczky A, Menyhárt O, Györfy B. Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Sci Rep* [Internet]. 2018; 8(1):9227. Available from: <https://doi.org/10.1038/s41598-018-27521-y> PMID: 29907753
46. Alboukadel K, Marcin K, Przemyslaw B. survminer: Drawing Survival Curves using “ggplot2” [Internet]. [cited 2023 Apr 8]. Available from: <https://cran.r-project.org/web/packages/survminer/index.html>.
47. Mizuno H, Kitada K, Nakai K, Sarai A. PrognoScan: A new database for meta-analysis of the prognostic value of genes. *BMC Med Genomics*. 2009;2.
48. Terry T. survival: Survival Analysis [Internet]. [cited 2023 Apr 8]. Available from: <https://cran.r-project.org/web/packages/survival/index.html>.
49. SCHOENFELD D. Partial residuals for the proportional hazards regression model. *Biometrika* [Internet]. 1982 Apr 1; 69(1):239–41. Available from: <https://doi.org/10.1093/biomet/69.1.239>.
50. Chang, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. shiny: Web Application Framework for R [Internet]. [cited 2023 Apr 15]. Available from: <https://cran.r-project.org/web/packages/shiny/index.html>.
51. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019 Nov 21; 4(43):1686.
52. Hadley W, Romain, Lionel H, Kirill M, Davis V. dplyr: A Grammar of Data Manipulation [Internet]. [cited 2023 Apr 8]. Available from: <https://cran.r-project.org/web/packages/dplyr/index.html>.
53. Hadley W, Lionel H. purrr: Functional Programming Tools [Internet]. [cited 2023 Apr 8]. Available from: <https://cran.r-project.org/web/packages/purrr/index.html>.
54. Kirill M, Hadley W. tibble: Simple Data Frames [Internet]. [cited 2023 Apr 8]. Available from: <https://cran.r-project.org/web/packages/tibble/index.html>.
55. Yihui X, Joe C, Xianying T. DT: A Wrapper of the JavaScript Library “DataTables” [Internet]. [cited 2023 Apr 8]. Available from: <https://cran.r-project.org/web/packages/DT/index.html>.
56. Chen Z, He S, Zhan Y, He A, Fang D, Gong Y, et al. TGF- β -induced transgelin promotes bladder cancer metastasis by regulating epithelial-mesenchymal transition and invadopodia formation. *EBioMedicine*. 2019 Sep 1; 47:208–20.
57. Zhou C, Lin A, Cao M, Ding W, Mou W, Guo N, et al. Activation of the DDR Pathway Leads to the Down-Regulation of the TGF β Pathway and a Better Response to ICIs in Patients With Metastatic Urothelial Carcinoma. *Front Immunol*. 2021 Jun 18; 12.
58. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* [Internet]. 2012 Mar 28; 16(5):284–7. Available from: <https://doi.org/10.1089/omi.2011.0118> PMID: 22455463
59. Mubeen S, Tom Kodamullil AT, Hofmann-Apitius M, Domingo-Fernández D. On the influence of several factors on pathway enrichment analysis. Vol. 23, *Briefings in Bioinformatics*. Oxford University Press; 2022.
60. Lin D, Fan W, Zhang R, Zhao E, Li P, Zhou W, et al. Molecular subtype identification and prognosis stratification by a metabolism-related gene expression signature in colorectal cancer. *J Transl Med*. 2021 Dec 1; 19(1). <https://doi.org/10.1186/s12967-021-02952-w> PMID: 34193202
61. Zhang H, Xia P, Liu J, Chen Z, Ma W, Yuan Y. Atic inhibits autophagy in hepatocellular cancer through the akt/foxo3 pathway and serves as a prognostic signature for modeling patient survival. *Int J Biol Sci*. 2021; 17(15):4442–58. <https://doi.org/10.7150/ijbs.65669> PMID: 34803509
62. Wei W, Liu C, Wang C, Wang M, Jiang W, Zhou Y, et al. Comprehensive pan-cancer analysis of N7-methylguanosine regulators: Expression features and potential implications in prognosis and immunotherapy. *Front Genet*. 2022 Oct 21; 13. <https://doi.org/10.3389/fgene.2022.1016797> PMID: 36339001