Journal of Big Data

RESEARCH Open Access

BEST: a web application for comprehensive biomarker exploration on large-scale data in solid tumors

Zaoqu Liu[1,10,11*†], Long Liu[3†], Siyuan Weng[1], Hui Xu[1], Zhe Xing[4], Yuqing Ren[5], Xiaoyong Ge[1], Libo Wang[3], Chunguang Guo[6], Lifeng Li[7], Quan Cheng[8], Peng Luo[9], Jian Zhang[9] and Xinwei Han[1,2*]

†Zaoqu Liu and Long Liu contributed equally to this work.

*Correspondence: liuzaoqu@163.com; fcchanxw@zzu.edu.cn

[1] Department of Interventional Radiology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China
Full list of author information is available at the end of the article

Abstract

Data mining from RNA-seq or microarray data has become an essential part of cancer biomarker exploration. Certain existing web servers are valuable and broadly utilized, but the meta-analysis of multiple datasets is absent. Most web servers only contain tumor samples from the TCGA database with only one cohort for each cancer type, which also means that the analysis results mainly derived from a single cohort are thin and unstable. Indeed, consistent performance across multiple independent cohorts is the foundation for an excellent biomarker. Moreover, the deeper exploration of specific biomarkers on underlying mechanisms, tumor microenvironment, and drug indications are missing in existing tools. Thus, we introduce BEST (Biomarker Exploration for Solid Tumors), a web application for comprehensive biomarker exploration on large-scale data in solid tumors. To ensure the comparability of genes between different sequencing technologies and the legibility of clinical traits, we re-annotated transcriptome data and unified the nomenclature of clinical traits. BEST delivers fast and customizable functions, including clinical association, survival analysis, enrichment analysis, cell infiltration, immunomodulator, immunotherapy, candidate agents, and genomic alteration. Together, our web server provides multiple cleaned-up independent datasets and diverse analysis functionalities, helping unleash the value of current data resources. It is freely available at https://rookieutopia.com/.

Keywords: Biomarker, Shiny, Pancancer, Survival, Web server, Immunotherapy

Introduction

Biomarker identification is an important goal of cancer research for clinicians and biologists. How to explore specific biomarkers that can distinguish tumoral from normal tissues, identify treatment-resistant patients, predict patient prognosis and recurrence, etc., are routine research tasks. Recently, immunotherapies represented by immune checkpoint inhibitors have opened a new era in cancer treatment, significantly improving the clinical outcomes of cancer patients [1]. However, only a small fraction of patients can generate considerable benefits from immunotherapies [2]. Exploring specific biomarkers

that can effectively predict immunotherapeutic efficacy is crucial for preventing under- or over-treatment.

With the advancement of bioinformatics techniques, researchers are inclined to explore cancer biomarkers using RNA-seq or microarray data [3, 4], and data mining has become an essential part of cancer research. However, these works may be difficult and inconvenient for clinicians and biologists without computational programming skills. Currently, several open-access web servers that allow users to analyze and visualize gene expression online directly are emerging, such as GEPIA [5], Xena [6], ExpressionAtlas [7], and HPA [8]. Although these web applications are valuable and broadly utilized, obtaining high confidence results in a specific tumor is difficult because their data sources are mainly derived from the TCGA database. Consistent performance across multiple independent datasets is the foundation for an excellent biomarker. In addition, the deeper exploration of specific biomarkers on underlying mechanisms, tumor microenvironment, and drug indications are missing in these tools.

To address these unmet needs, we have developed Biomarker Exploration for Solid Tumors (BEST), a web-based application for comprehensive biomarker exploration on large-scale data in solid tumors and delivering fast and customizable functionalities to complement existing tools.

Methods

Data collection

BEST is committed to identifying robust tumor biomarkers through large-scale data. Hence, we retrieved cancer datasets with both expression data and important clinical information (e.g., survival, therapy, etc.) as much as possible. Eligible datasets were mainly enrolled from five databases, including The Cancer Genome Atlas Program (TCGA, https://portal.gdc.cancer.gov), Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), International Cancer Genome Consortium (ICGC, https://dcc.icgc.org), Chinese Glioma Genome Atlas (CGGA, http://www.cgga.org.cn/), and ArrayExpress (https://www.ebi.ac.uk/arrayexpress/). In total, we included more than 50,000 samples from 64 datasets for 27 cancer types.

Data re-annotation and pre-processing

Raw expression data were extracted for subsequent processing (Fig. 1). Data were re-annotated if the original probe sequences were available based on the GRCh38 patch 13 sequences reference from GENCODE (https://www.gencodegenes.org/). For RNA-seq data, raw count read was converted to transcripts per kilobase million (TPM) and further log-2 transformed. The raw microarray data from Affymetrix®, Illumina®, and Agilent® were processed using the *affy* [9], *lumi* [10], and *limma* [11] packages, respectively. The normalized matrix files were directly downloaded for microarray data from other platforms. Gene expression was further transformed into z-score across patients in each dataset. To make it easier for users to interpret and present analysis results, we cleaned and unified the clinical traits. Take KRAS mutation as an example, GSE39084 [12] named it 'kras.gene.mutation.status', 'mutation' was labeled 'M' and 'wild type' was labeled 'WT'; whereas GSE143985 [13] named it 'kras_mutation', 'mutation' was labeled

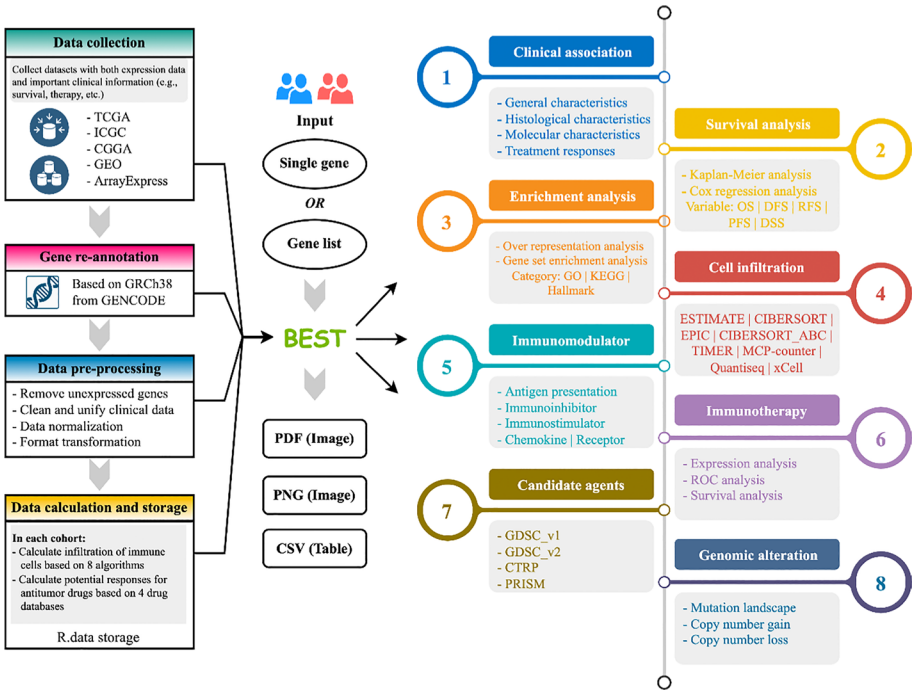

Fig. 1 Overview of the BEST analytical framework

'Y' and 'wild type' was labeled 'N'. We uniformly termed it 'KRAS', and 'mutation' was labeled 'Mut' and 'wild type' was labeled 'WT'.

Data calculation and storage

A tremendous amount of calculations are involved in BEST analysis, we thus have completed the time-consuming calculations in advance and used R.data for storage. Users can directly call these data, significantly reducing the user's waiting time and background computing pressure. Take colorectal cancer (CRC) as an example, we collected a total of 47 datasets. Drug assessment is an analysis module of BEST, which requires fitting ridge regression models for individual drugs based on drug responses and expression data of cancer cell lines from the Genomics of Drug Sensitivity in Cancer_v1 (GDSC_v1), Genomics of Drug Sensitivity in Cancer_v2 (GDSC_v2), The Cancer Therapeutics Response Portal (CTRP), and Profiling Relative Inhibition Simultaneously in Mixtures (PRISM) databases, and then predicting the sensitivity of each drug for CRC samples from all collected datasets. Apparently, if these results are not calculated in advance, users may have to wait more than 3 days. The pre-calculated content is displayed in Fig. 1.

Implementations

BEST is entirely free for users, built by the Shiny app and the HTML5, CSS, and JavaScript libraries for the client-side user interface. The Shiny app (version: 1.7.2) mainly executes data processing and analysis. The function of BEST is divided into eight tabs (Fig. 1): Clinical association, Survival analysis, Enrichment analysis, Cell infiltration, Immunomodulator, Immunotherapy, Candidate agents, and Genomic alteration.

Analysis results include images and tables, images can be downloaded in portable document format (PDF) and portable network graphics (PNG) format, and tables can be obtained in comma-separated value (CSV) format.

Results

Quick start

BEST offers a simple interactive interface. Users first select one cancer type and then determine the input category—single gene or gene list (Fig. 1). For the single-gene module, users can enter a gene symbol or an Ensembl ID in the 'Enter gene name' field to explore a gene of interest. The gene list module needs users to input a list of genes and pick a method to calculate the gene set score for each sample. The embedded methods include gene set variation analysis (GSVA) [14], single sample gene set enrichment analysis (ssGSEA) [15], z-score [16], pathway-level analysis of gene expression (PLAGE) [17], and the mean value. Users can customize the name of the gene set score.

Clinical association

In this module, users can explore the associations between the expression or score of the input variable and general characteristics (e.g., age, gender, alcohol, smoke, etc.), histological characteristics (e.g., tissue type, tumor site, stage, etc.), molecular characteristics (e.g., TP53 mutation, microsatellite instability, etc.) and treatment responses (e.g., chemotherapy and bevacizumab responses, etc.) (Fig. 2A). Whether to use parametric or nonparametric statistical tests for group comparisons based on the distribution of input variable [18]. For example, users can easily explore the differential expression of the input variable between tumor and normal tissues or find the associations between the input variable with smoke and alcohol. Our datasets also include abundant treatment responses, which might contribute to developing promising biomarkers in clinical settings. Importantly, analysis results tend to be displayed in multiple independent cohorts, which provides a reference for the stability power of a variable of interest. For instance, Fig. 2B illustrates that CRC tumors process a significantly higher expression of *COL1A2* than normal tissues in most CRC datasets with tissue type information.

Survival analysis

BEST performs survival analysis based on gene expression or gene set score. This module allows users to explore the prognostic significance for overall survival (OS), disease-free survival (DFS), relapse-free survival (RFS), progression-free survival (PFS), and disease-specific survival (DSS) (Fig. 2C). BEST generates Kaplan–Meier curves with log-rank test and forest plot with cox proportional hazard ratio and the 95% confidence interval information for various survival outcomes in multiple independent datasets (Fig. 2C, D). Kaplan–Meier analysis requires categorical variables, we thus provide five cutoff options for users to choose from, including 'median', 'mean', 'quantile', 'optimal', and 'custom'. For example, when investigating gene *COL1A2* in survival analysis of CRC, users can obtain Kaplan–Meier curves with a specific cutoff approach and a Cox forest plot for five survival outcomes across all CRC datasets with survival information.

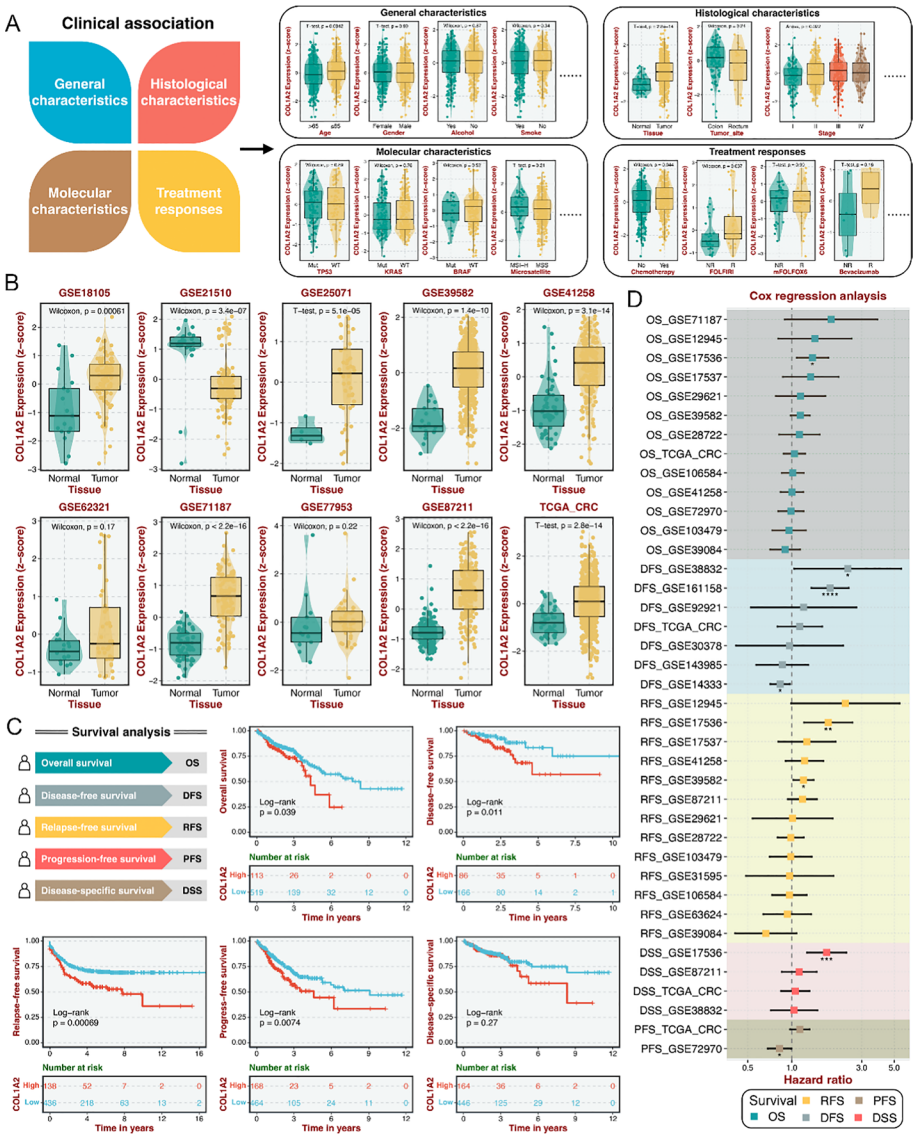

Fig. 2 Modules for clinical association and survival analysis. **A** Four categories of clinicopathologic information are mainly included in the clinical association module. **B** An example illustrates the differential expression of *COL1A2* in multiple CRC datasets between the tumor and normal groups. **C** Five categories of survival variables are utilized in the survival analysis module, and examples of five survival variables for Kaplan–Meier analysis. **D** An example displays the cox regression analysis of five survival variables in multiple CRC datasets

Enrichment analysis

BEST provides two enrichment frameworks: over-representation analysis (ORA) [19] and gene set enrichment analysis (GSEA) [20]. Users can select the top gene (self-defined number) most associated with the input variable to perform ORA and apply a ranked gene list based on the correlation between all genes and the input variable to carry out GSEA (Fig. 3A). Of note, the final correlation coefficient between the input variable and each gene is the average correlation of all datasets in specific cancer. The Pearson correlation was calculated between all genes and the input variable. If users input a gene list, which will be firstly calculated by one of the four provided algorithms, including gene

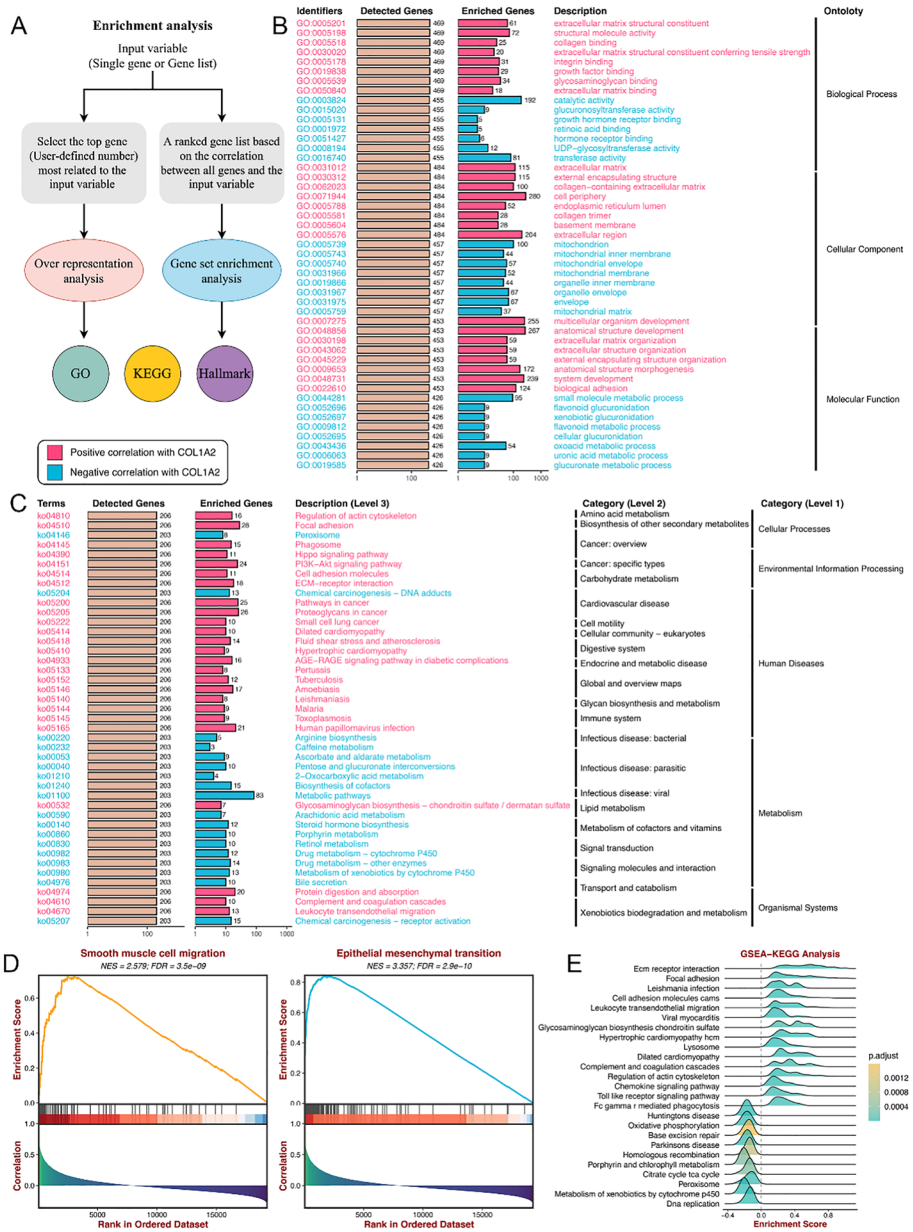

Fig. 3 Enrichment analysis module. **A** Enrichment analysis module includes two enrichment frameworks: over-representation analysis (ORA) and gene set enrichment analysis (GSEA). **B**, **C** GO (**B**) and KEGG (**C**) bar charts for the ORA framework. **D**, **E** GSEA-Plot (**D**) and Ridge-Plot (**E**) examples for the GSEA framework

set variation analysis (GSVA), single sample gene set enrichment analysis (ssGSEA), z-score, pathway-level analysis of gene expression (PLAGE), and the mean value. The output forms of ORA are GO and KEGG bar charts (Fig. 3B, C). The 'Detected Genes' are all top gene most related to the input variable, which are also existed in GO or KEGG gene sets. The 'Enriched Genes' are the top gene within the specific biological pathway. Also, GSEA results are exhibited using GSEA-Plot (Fig. 3D) and Ridge-Plot (Fig. 3E) images. The GO, KEGG, and Hallmark gene sets for GSEA are obtained from Molecular Signatures Database (MSigDB). Similarly, users could select single gene or gene list as

input variable. The specific biological term of GO, KEGG, and Hallmark gene set could be shown as GSEA-Plot, or a series of biological terms could be displayed as Ridge-Plot.

Cell infiltration and immunomodulator

BEST offers eight prevalent algorithms to estimate immune cell infiltration in the tumor microenvironment (TME) (Fig. 4A), including CIBERSORT [21], CIBERSORT ABS [21], EPIC [22], ESTIMATE [23], MCP-counter [24], Quantiseq [25], TIMER [26], and xCell [27]. To avoid time-consuming calculations for users and save computing resources, these eight algorithms have been executed in advance across all datasets, and the resulting data have been stored in the website background. Additionally, BEST provides five immunomodulator categories: antigen presentation, immunoinhibitors, immunostimulators, chemokines, and receptors (Fig. 4A). Users can generate heatmap and correlation scatter plots from these two analysis modules. The heatmaps illustrate the correlations of

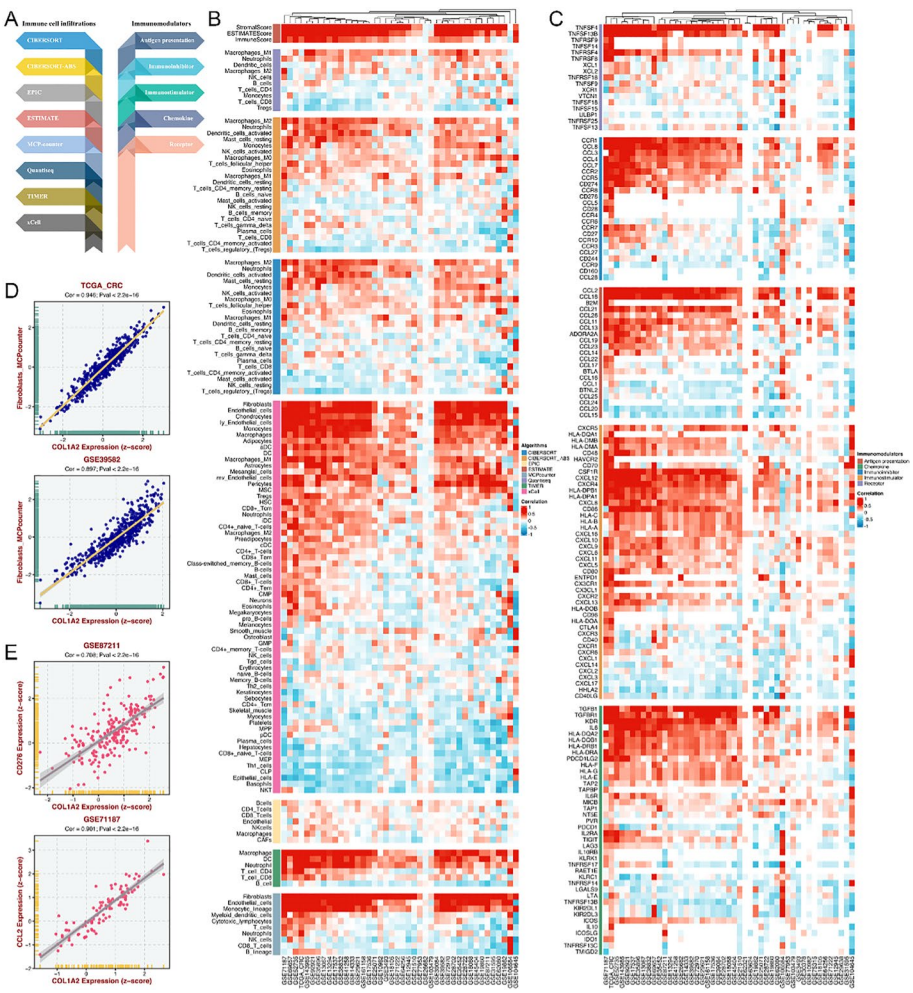

Fig. 4 Modules for cell infiltration and immunomodulator analysis. **A** BEST offers eight immune infiltration assessment algorithms and five categories of immunomodulators. **B**, **C** Heatmaps illustrate the correlations of the input variable with each immune cell (**B**) or immunomodulator (**C**) across all CRC datasets. **D**, **E** Correlation scatters plots indicate the correlation of the input variable and an immune cell (**D**) or immunomodulator (**E**) in a specific dataset

the input variable with each immune cell/immunomodulator across all cohorts (Fig. 4B, C), and the correlation scatters plots indicate the correlation of the input variable and an immune cell/immunomodulator in a specific dataset (Fig. 4D, E).

Immunotherapy

To further investigate the clinical significance of the input variable in immunotherapies, we retrieved 19 immunotherapeutic cohorts with expression data and immunotherapy information (e.g., CAR-T, anti-PD-1, anti-CTLA4, etc.) (Fig. 5A). Based on gene expression or gene set score in these datasets, users can conduct differential expression analysis (DEA) between response and non-response groups (Fig. 5B), receiver operating

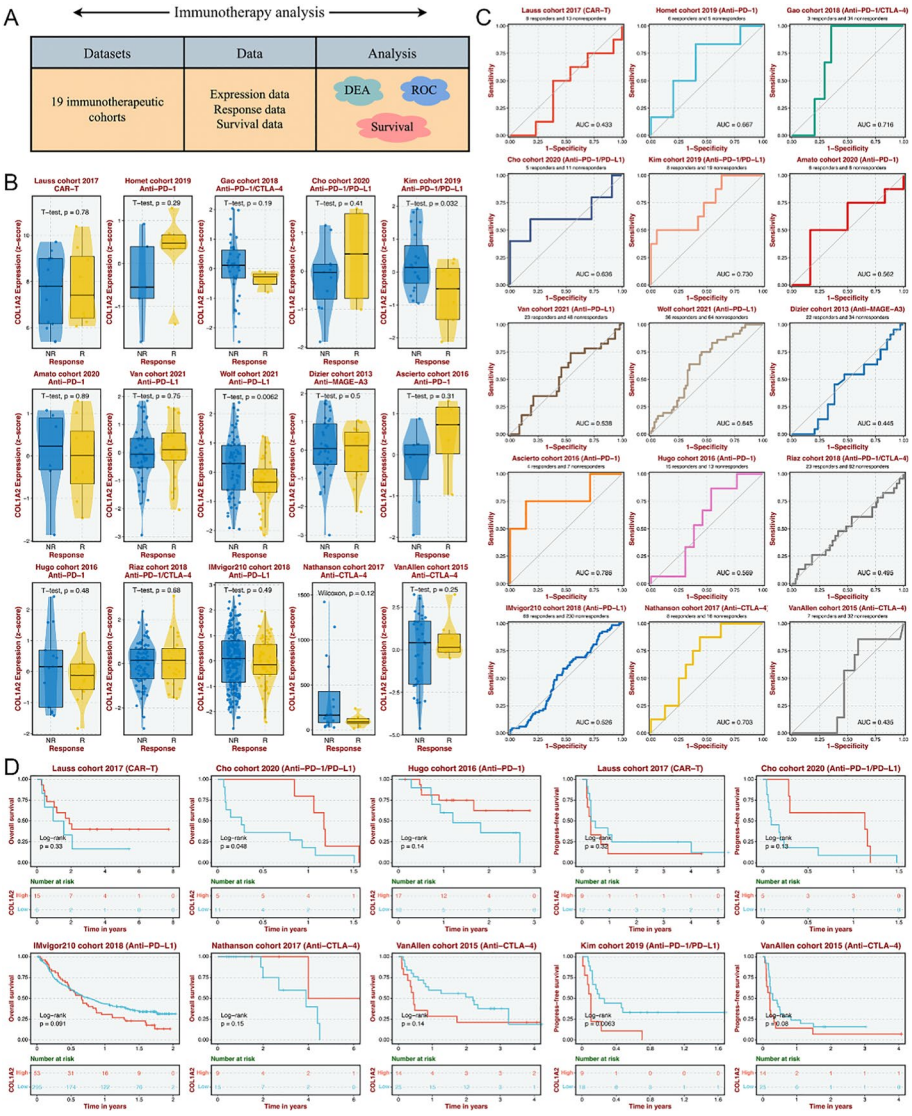

Fig. 5 Immunotherapy analysis module. **A** Schema describing data details and analysis for the immunotherapy module. **B** Boxplots indicate the differential expression of *COL12A* between response and non-response groups. **C** Receiver operating characteristic (ROC) curves evaluate the performance of the input variable in predicting the immunotherapeutic efficacy. **D** Kaplan–Meier curves assess the impact of the input variable on survival (OS and PFS) in immunotherapeutic cohorts that have undergone immunotherapies

characteristic (ROC) curve to evaluate the performance of the input variable in predicting the immunotherapeutic efficacy (Fig. 5C), and survival analysis to assess the impact of the input variable on survival (OS and PFS) in immunotherapeutic cohorts that have undergone immunotherapies (Fig. 5D).

Candidate agents

In this analysis tab, BEST performs drug assessment in bulk samples based on drug responses and expression data of cancer cell lines from the GDSC_v1, GDSC_v2, CTRP, and PRISM databases (Fig. 6A). Given the inherent differences between bulk samples

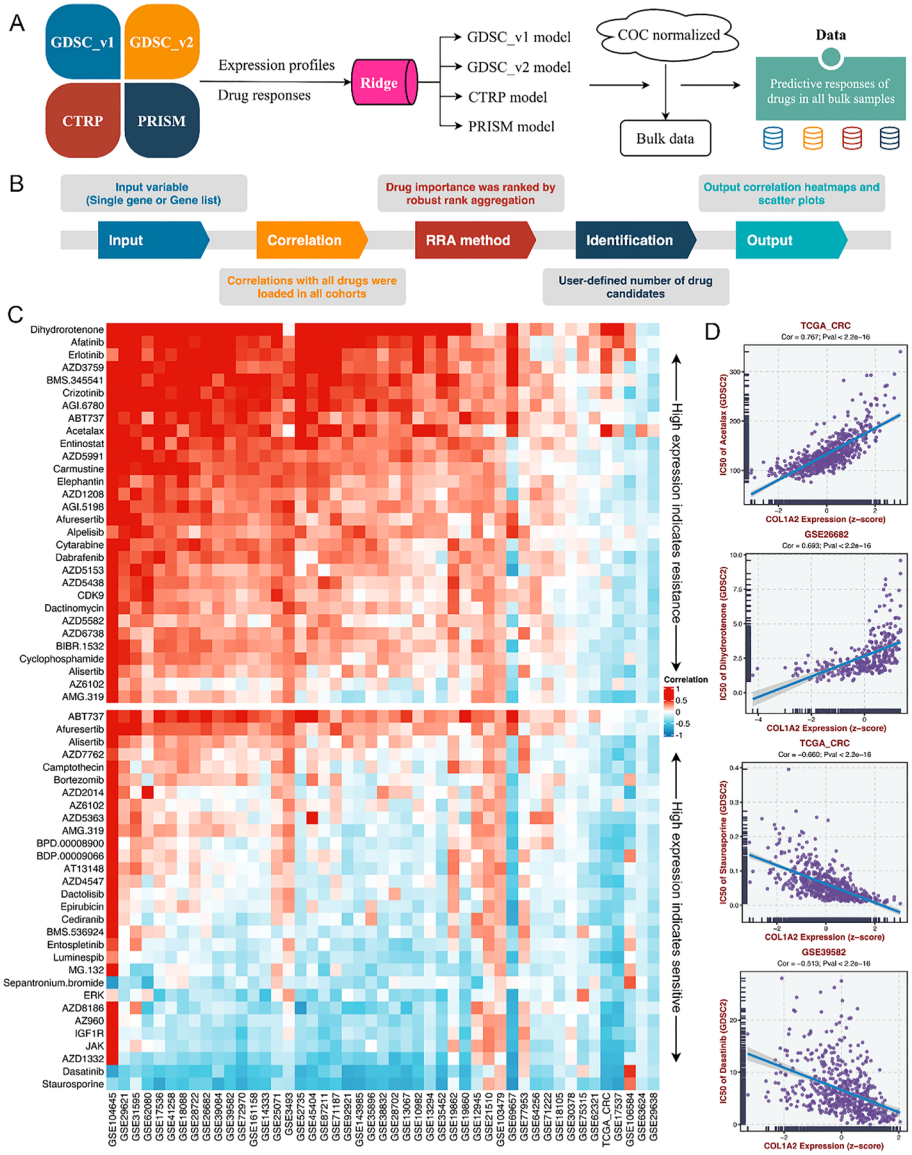

Fig. 6 Candidate agents module. **A** Overview of BEST performing drug assessment. **B** Details of obtaining drugs that are significantly associated with the input variable. **C** Heatmaps illustrate the correlations of the input variable with each drug across all cohorts. Higher-ranked drugs indicate that high levels of the input variable predict drug resistance and vice versa. **D** Correlation scatters plots indicate the correlation of the input variable and a drug in a drug database and a tumor dataset

and cell line cultures, we introduced a correlation of correlations framework [28] to retain genes presenting analogical co-expression patterns in bulk samples and cell lines. As previously reported [29], the model used for predicting drug response was the ridge regression algorithm implemented in the *oncoPredict* package [30]. This predictive model was trained on transcriptional expression profiles and drug response data of cancer cell lines with a satisfied predictive accuracy were evaluated by default 10-fold cross-validation, thus allowing the estimation of clinical drug response using only the expression data of bulk samples (Fig. 6A). Modeling and prediction works have been completed, and drug assessments of all tumor samples based on four databases have been stored in the website background. BEST will calculate the correlations between all drugs and the input variable in all cohorts. According to the correlation rank of each drug across all datasets, we applied the robust rank aggregation (RRA) [31] to determine drug importance related to the input variable (Fig. 6B). Users can select the top drugs (self-defined number) to display the heatmap that illustrates the correlations of the input variable with each drug across all cohorts. Higher-ranked drugs indicate that high levels of the input variable predict drug resistance and vice versa. For example, high expression of *COL1A2* might suggest Afatinib resistance and Dasatinib sensitivity based on the GDSC_v2 database (Fig. 6C). Also, users can select a drug database, a tumor dataset, and a specific drug to generate a correlation scatter plot (Fig. 6D).

Genomic alteration

In this module, BEST has pre-processed mutation and copy number variation data from the TCGA database using *maftools* [32] and *GISTIC2.0* [33], respectively. Users can obtain a heatmap indicating genomic alterations as the input variable increase. The right panel of heatmap also displays the proportion of genomic alteration and statistical differences between the high and low groups. For example, with the rise in *COL1A2* expression, the genomics landscape of the TCGA-CRC dataset is illustrated in Fig. 7. We found that the loss of chromosome segment 1p13.2 was more frequent in the high expression group.

Discussion

As an interactive web tool, BEST aims to explore the clinical significance and biological functions of cancer biomarkers through large-scale data. Therefore, data richness is the foundation of BEST. From data collection, re-annotation, pre-processing, and pre-calculation to storage, we provide a tidy and uniform pan-cancer database, allowing users to call and interpret data quickly. BEST offers prevalent analysis modules to enable researchers without computational programming skills to conduct various bioinformatics analyses. Compared with other available tools [5–8, 34–36], BEST has more datasets and more diverse analysis options, which complements well with them (Table 1).

In BEST web application, users can identify cancer biomarkers associated with critical clinical traits (e.g., stage and grade), prognosis, and immunotherapy. Moreover, the underlying mechanisms of these biomarkers could be further explored using the enrichment, cell infiltration, and immunomodulator analysis modules. Users can also apply the candidate agent analysis tab to investigate high levels of cancer biomarkers that might indicate which drugs are resistant and which are sensitive to specific cancer.

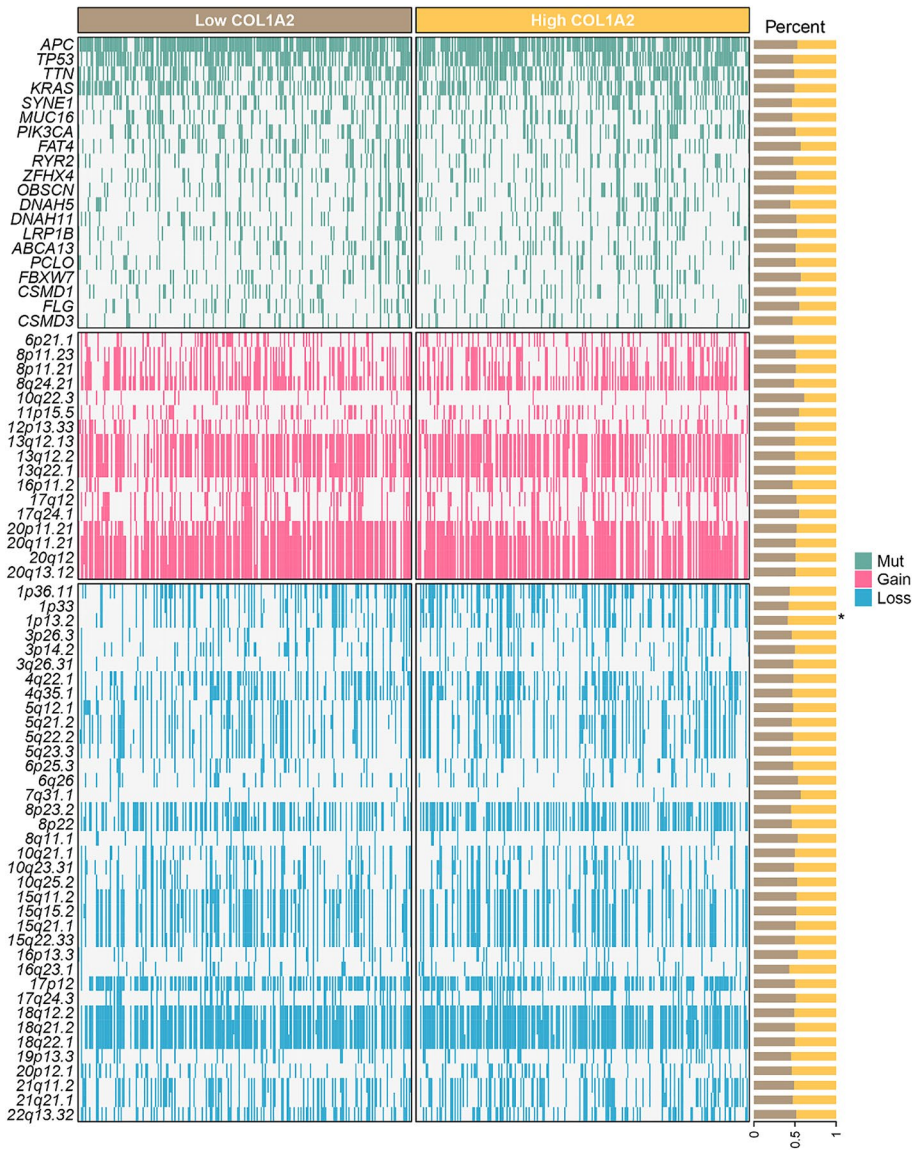

Fig. 7 Genomic alteration analysis module

Taken together, BEST provides a curated database and innovative analytical pipelines to explore cancer biomarkers at high resolution. It is an easy-to-use and time-saving web tool that allows users, especially clinicians and biologists without background knowledge of bioinformatics data mining, to comprehensively and systematically explore the clinical significance and biological function of cancer biomarkers. With constant user feedback and further improvement, BEST is promising to serve as an integral part of routine data analyses for researchers.

Abbreviations

BEST	Biomarker Exploration for Solid Tumors
CRC	Colorectal cancer
PDF	Portable document format
PNG	Portable network graphics

Table 1 Comparison of BEST with other tools

Tools	BEST	GEPIA	Xena	ExpressionAtlas	HPA	UALCAN	PrognoScan	PROGgeneV2
Pancancer	✓	✓	✓	✓	✓	✓	✓	✓
Multiple datasets	✓	✗	✓	✓	✗	✗	✓	✓
Analytic target								
Single gene	✓	✓	✓	✓	✓	✓	✓	✓
Gene list	✓	✓	✗	✗	✗	✗	✗	✓
Clinical association								
Tidy data	✓	✗	✗	✗	✗	✓	✗	✗
General characteristics	✓	✗	✓	✓	✓	✓	✗	✗
Histological characteristics	✓	✓	✓	✓	✓	✓	✗	✗
Molecular characteristics	✓	✗	✓	✗	✗	✓	✗	✗
Treatment responses	✓	✗	✓	✗	✗	✗	✗	✗
Survival analysis								
Multiple outcomes	✓	✓	✓	✗	✗	✗	✓	✓
Multiple cutoff options	✓	✓	✗	✗	✗	✗	✗	✓
Cox regression analysis	✓	✓	✗	✗	✗	✗	✓	✓
Kaplan-Meier analysis	✓	✓	✓	✗	✓	✓	✓	✓
Enrichment analysis								
ORA	✓	✗	✗	✗	✗	✗	✗	✗
GSEA	✓	✗	✗	✗	✗	✗	✗	✗
Cell infiltration								
CIBERSORT	✓	✓	✗	✗	✗	✗	✗	✗
CIBERSORT_ABS	✓	✗	✗	✗	✗	✗	✗	✗
EPIC	✓	✓	✗	✗	✗	✗	✗	✗
ESTIMATE	✓	✗	✗	✗	✗	✗	✗	✗
MCP-counter	✓	✗	✗	✗	✗	✗	✗	✗
Quantiseq	✓	✓	✗	✗	✗	✗	✗	✗
TIMER	✓	✗	✗	✗	✗	✗	✗	✗
xCell	✓	✗	✗	✗	✗	✗	✗	✗
Immunomodulator								
Antigen presentation	✓	✗	✗	✗	✗	✗	✗	✗
Immunoinhibitor	✓	✗	✗	✗	✗	✗	✗	✗
Immunostimulator	✓	✗	✗	✗	✗	✗	✗	✗
Chemokine	✓	✗	✗	✗	✗	✗	✗	✗
Receptor	✓	✗	✗	✗	✗	✗	✗	✗
Immunotherapy								
DEA	✓	✗	✗	✗	✗	✗	✗	✗
ROC analysis	✓	✗	✗	✗	✗	✗	✗	✗
Survival analysis	✓	✗	✗	✗	✗	✗	✗	✗
Drug analysis	✓	✗	✗	✗	✗	✗	✗	✗
Genomic analysis	✓	✗	✓	✗	✗	✗	✗	✗

CSV	Comma-separated value
TME	Tumor microenvironment
OS	Overall survival
DFS	Disease-free survival
RFS	Relapse-free survival
PFS	Progression-free survival
DSS	Disease-specific survival
GDSC	Genomics of Drug Sensitivity in Cancer
CTRP	The Cancer Therapeutics Response Portal
PRISM	Profiling Relative Inhibition Simultaneously in Mixtures

Acknowledgements
Not applicable.

Author contributions
ZQL contributed study design, data analysis, and paper writing. XWH contributed project oversight and paper revisiting. LL, SYW, HX, ZX, XYG, LBW, and CGG collected samples and generated data. LBW and L.L performed and interpreted trail assays. YQR, LFL, QC, PL, and JZ contributed paper revisiting.

Funding
This study was supported by The Collaborative Innovation Major Project of Zhengzhou (Grant No. 20XTZX08017), The National Natural Science Foundation of China (Grant No. 82002433), and Science and Technology Project of Henan Provincial Department of Education (Grant No. 21A320036).

Data availability
BEST is available at https://rookieutopia.com/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
We have obtained consents to publish this paper from all the participants of this study.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Interventional Radiology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China. [2]Interventional Treatment and Clinical Research Center of Henan Province, Zhengzhou, Henan, China. [3]Department of Hepatobiliary and Pancreatic Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shanxi, China. [4]Department of Neurosurgery, The Fifth Affiliated Hospital of Zhengzhou University, Henan, China. [5]Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China. [6]Department of Endovascular Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China. [7]Cancer center, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, Henan, China. [8]Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha, China. [9]Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, China. [10]State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing, China. [11]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.

Received: 7 July 2022 Accepted: 16 October 2023
Published online: 01 November 2023

References

1. Hamilton PT, Anholt BR, Nelson BH. Tumour immunotherapy: lessons from predator-prey theory. Nat Rev Immunol. 2022. https://doi.org/10.1038/s41577-022-00719-y.
2. Vesely MD, Zhang T, Chen L. Resistance mechanisms to anti-PD cancer immunotherapy. Annu Rev Immunol. 2022;40:45–74.
3. Liu Z, Liu L, Weng S, Guo C, Dang Q, Xu H, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. Nat Commun. 2022;13(1):816.
4. Liu Z, Guo C, Dang Q, Wang L, Liu L, Weng S, et al. Integrative analysis from multi-center studies identities a consensus machine learning-derived lncRNA signature for stage II/III colorectal cancer. EBioMedicine. 2022;75:103750.
5. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45(W1):W98–102.
6. Goldman MJ, Craft B, Hastie M, Repecka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675–8.

7. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 2016;44(D1):D746-752.
8. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347(6220): 1260419.
9. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15.
10. Du P, Kibbe WA, Lin SM. Lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24(13):1547–8.
11. 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):e47.
12. Kirzin S, Marisa L, Guimbaud R, De Reynies A, Legrain M, Laurent-Puig P, et al. Sporadic early-onset colorectal cancer is a specific sub-type of cancer: a morphological, molecular and genetics study. PLoS ONE. 2014;9(8): e103159.
13. Shinto E, Yoshida Y, Kajiwara Y, Okamoto K, Mochizuki S, Yamadera M, et al. Clinical significance of a gene signature generated from tumor budding grade in colon cancer. Ann Surg Oncol. 2020;27(10):4044–54.
14. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14: 7.
15. 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;462(7269):108–12.
16. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol. 2008;4(11): e1000217.
17. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinform. 2005;6:225.
18. le Cessie S, Goeman JJ, Dekkers OM. Who is afraid of non-normal data? Choosing between parametric and non-parametric tests. Eur J Endocrinol. 2020;182(2):E1–E3.
19. Tokar T, Pastrello C, Jurisica I. GSOAP: a tool for visualization of gene set over-representation analysis. Bioinformatics. 2020;36(9):2923–5.
20. 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. Proc Natl Acad Sci USA. 2005;102(43):15545–50.
21. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
22. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife. 2017;6: 6.
23. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
24. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17(1):218.
25. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11(1):34.
26. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17(1):174.
27. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220.
28. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21(11):1350–6.
29. Yang C, Chen J, Li Y, Huang X, Liu Z, Wang J, et al. Exploring subclass-specific therapeutic agents for hepatocellular carcinoma by informatics-guided drug screen. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbaa295.
30. Maeser D, Gruener RF, Huang RS. oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbab260.
31. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012;28(4):573–80.
32. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28(11):1747–56.
33. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4): R41.
34. Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi B, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia. 2017;19(8):649–58.
35. Mizuno H, Kitada K, Nakai K, Sarai A. PrognoScan: a new database for meta-analysis of the prognostic value of genes. BMC Med Genom. 2009;2:18.
36. Goswami CP, Nakshatri H. PROGgeneV2: enhancements on the existing database. BMC Cancer. 2014;14: 970.

Publisher's Note