CBioExplorer: a web and standalone application for screening, validation, and annotation of cancer survival related biomarkers from molecular level to clinical settings

Xiao-Ping Liu[1,2,3], Hongjie Shi[2,3], Sheng Li[1,2,3#], Xing-Huan Wang[1,2,3#]

1. Department of Urology, Zhongnan Hospital of Wuhan University, Wuhan, Hubei province, China. 430071.

2. Cancer Precision Diagnosis and Treatment and Translational Medicine Hubei Engineering Research Center, Wuhan 430071, China.

3. Department of Biological Repositories, Zhongnan Hospital of Wuhan University, Wuhan 430071, China.

Correspondence:

Xing-Huan Wang, Department of Urology, Cancer Precision Diagnosis and Treatment and Translational Medicine Hubei Engineering Research Center, Department of Biological Repositories, Zhongnan hospital of Wuhan University, Wuhan, Hubei province, China. 430071. Email: wangxinghuan@whu.edu.cn

Sheng Li, Department of Urology, Cancer Precision Diagnosis and Treatment and Translational Medicine Hubei Engineering Research Center; Department of Biological Repositories, Zhongnan Hospital of Wuhan University, Wuhan 430071, China. Email: lisheng-znyy@whu.edu.cn

Abstract

Cancer is one of the main killers threatening human life and health. The exploration of biomarkers that can predict the long-term survival of cancer patients is of great significance for the risk stratification, treatment and prognosis of cancer patients. However, the screening and validation

process involving biomarkers is a complicated process. Currently, there are no unified tools for physicians and researchers to explore tumor biomarkers. Herein, we present CBioExplorer (**C**ancer **Bio**marker **Explorer**), a user-friendly web server and standalone application that includes 5 major modules (dimensionality reduction, benchmark experiment, prediction model, clinical annotation; and, biological annotation) and integrated a novel R package CuratedCancerPrognosisData that reviewed, curated and integrated the gene expression data and corresponding clinical data of 47,210 clinical samples from 268 gene expression studies of 43 common blood and solid tumors, for screening, validation, and annotation of cancer survival related biomarkers based on 6 machine learning survival learners from molecular level to clinical settings. The web server is available at http://cbioexplorer.znhospital.cn:8383/CBioExplorer/, and the standalone app and source code of CBioExplorer and CuratedCancerPrognosisData can be found at https://github.com/liuxiaoping2020/CBioExplorer or https://gitee.com/liuxiaoping2020/CBioExplorer.

Introduction

In recent years, with the advancement of high-throughput sequencing technologies including DNA sequencing, RNA sequencing, microarray, single-cell sequencing, etc., and their wide application in medical research and clinical practice, a large number of gene expression profiling studies of cancer patients has been published(1,2). The gene expression data of these gene expression profiling studies and the clinical data of the corresponding patients are mostly stored in public databases such as GEO (Gene Expression Omnibus) (https://www.ncbi.nlm.nih.gov/geo/), TCGA (The Cancer Genome Atlas) (https://portal.gdc.cancer.gov/), ArrayExpress (https://www.ebi.ac.uk/arrayexpress/), TARGET (Therapeutically Applicable Research To Generate Effective Treatments) (https://ocg.cancer.gov/programs/target), ICGC (https://dcc.icgc.org/), CGGA (Chinese Glioma Genome Atlas) (http://www.cgga.org.cn/), etc., and some of them are uploaded as research supplements on the official website of the journal or related research institutions. However, given that the data storage, preprocessing, and operation interface of each database are not completely the same, and the clinical data and gene expression data of each study have great differences in data collection, preprocessing, format, and documentation, people who want to make full and effective use of these high-throughput data to serve their research and guide clinical practice are facing great obstacles.

With the aging of the population and changes in people's lifestyles, malignant tumors have become one of the main killers threatening human health and life(3,4). Therefore, the development and validation of novel tumor markers that can be used for tumor diagnosis, risk stratification, and prognosis are of great significance for the early detection and early treatment of tumors. With the advancement of artificial intelligence, more and more machine learning strategies have been applied to the screening and validation of biomarkers for cancer patients(5,6). However, due to the lack of a unified, standardized and rigorous model and variable selection process, the reliability of relevant biomarkers is questionable.

Thus, in the present study, we developed and introduced CBioExplorer (**C**ancer **Bio**marker **Explorer**), a web server and standalone application that reviewed, curated and integrated the gene expression data and corresponding clinical data of 47,210 samples from 268 gene expression studies of 43 common blood and solid tumors, for screening, validation, and annotation of cancer survival related biomarkers based on 6 machine learning survival learners from molecular level to clinical settings (Figure 1).

Methods and materials

Data collection and curation

We searched and downloaded the gene expression profile studies of tumor patients in GEO, TCGA, ICGC, ArrayExpress, TARGET, CGGA and other public databases or websites. Data sets that meet the following criteria are included in our research: (1) The subject of the relevant research is tumor patients; (2) Dataset contains gene expression profiling data; (3) Dataset has reported at least one type of follow-up information, such as overall survival (OS), progression free survival (PFS), relapse free survival (RFS), disease free survival (DFS), distant metastasis free survival (DMFS), etc.); (4) The sample size of dataset is greater than 20. For data from GEO and ArrayExpress, R package 'GEOquery'(7) and 'ArrayExpress'(8) were used to download it, respectively. If the raw CEL data was available, robust multichip average (RMA) method(9) was used to normalize the raw data using R package 'affy'(10) or 'oligo'(11), otherwise, the normalized data was used. For data from TCGA and MMRF-CoMMpass(12,13), we downloaded the RNA-seq count data and transformed it to TPM (transcripts per kilobase million) value from GDC using R package 'TCGAbiolinks'(14). For data from TARGET, ICGC, CGGA, normalized data was downloaded and used indirectly. Annotation files provided by the submitters were used to annotate the gene expression profiling data, for multiple probes that match to the same gene, we choose the most variant probe, and when multiple genes correspond to the same probe, we dropped them since unspecific annotation. Clinical data was uniformly reformatted and curated using in-house R script per dataset, and independent double-check was conducted by investigators to ensure the accuracy of the curation. The workflow of the curation is summarized in supplementary figure 1.

Dimensionality reduction

CBioExplorer uses the 3 most commonly used bioinformatics methods to reduce the dimensionality of data: (1) weighted gene co-expression analysis (WGCNA)(14); (2) Survival related genes (SRGs) (3) Differentially expressed genes (DEGs)(15). WGCNA includes 3 steps: firstly, Euclidean distance-based sample network was used to filter outlying samples, and then a weighted gene co-expression network was constructed to identify gene modules whose expression profiles are similar based on adjacency matrix and appropriate soft threshold. Finally, correlations between gene modules and clinical features are calculated. Empirical Bayesian methods was used to perform DEG analysis using R package 'limma'(15). SRG was implemented based on univariate Cox proportional hazards regression (CoxPH) model using R package 'survival'.

Survival learners

For the benchmark experiment, CBioExplorer includes 6 embedded machine learning algorithms, namely, LASSO(16,17), Ridge(18,19), Elastic Net(20), Glmboost(21), Coxboost(22), RandomForestSRC(23) for survival analysis. LASSO (least absolute shrinkage and selection operator), proposed by Robert Tibshirani in 1996, obtains a more refined model by constructing a L-1 norm penalty function, which forces the sum of absolute values of coefficients to be less than a certain fixed value and set some regression coefficients to zero. Therefore, it is a regression analysis method that performs feature selection and regularization at the same time, and aims to enhance the prediction accuracy and interpretability of statistical models(16). Ridge regression is

similar to linear regression, both of which are to solve the over-fitting problem of standard linear regression. The difference is that ridge regression adds the L2 norm penalty(19). Elastic network integrates the L1 norm and the L2 norm, which makes it have both the variable selection and regularization advantages of lasso and ridge regression(20). Glmboost fits generalized linear model and conduct variable selection based on component-wise boosting(21). Unlike glmboost, CoxBoost fits a CoxPH model by component-wise likelihood based boosting(22). For feature selection, the above five survival learners reserve features with non-zero coefficients. Random forest is an ensemble model, and the original random forest (RF) is only available for regression and classification. randomforestSRC extends RF to survival analysis and conducts variable selection based on maximal subtree information(23). Parameter sets for the survival learners are summarized in supplementary table 1.

Benchmark experiment

Benchmark experiment is supported by the R package 'mlr'(24). To train and validate the survival learners, we apply cross validation (CV) and nested cross validation (nCV) to perform a benchmark experiment. During CV, the whole data set is randomly split training set and test set, then k-fold CV is applied to the training set: (1) divide the training set into equal K folds; (2) use the first fold as inner test set, and the rest as inner training set. (3) train the model and calculate the C-index of the model on the inner test set; (4) use a different fold as inner test set each time, and repeat steps (2) and (3) K times. (5) apply the best model to test set and external independent validation cohort. CV is designed for model selection, when the best model is selected, bootstrap resampling can be used to evaluate and compare the performances of different survival learners. The workflow of CV is summarized in supplementary figure 2.

With respect to nCV, the whole data set is divided into N outer folds, and then each outer fold is divided into training set and test set. Then, the main steps of nCV can be summarized as: (1) divide the training set into equal K folds; (2) use the first fold as inner test set, and the rest as inner training set. (3) train the model and calculate the C-index of the model on the inner test set; (4) use a different fold as inner test set each time, and repeat steps (2) and (3) K times. (5) apply the best model to outer fold test set. (6) select the best outer model features and parameters and train on the whole data set to get final model. (7) if the users have divided the whole data set into two parts, one is for training nCV, the other is for validation, then they can validate the final model on the validation part and external validation cohort, otherwise, they can validate the final model on external cohort. nCV utilizes multi-layer CV to implement model selection. The workflow is summarized in supplementary figure 3.

Prediction model

When the benchmark experiments finished, CBioExplorer could calculate the fitted relative risk of patients in the training set, test set and external validation set. LASSO, Ridge, Elastic Net, Glmboost, and Coxboost calculate relative risk using the predict function, while randomforestSRC uses the sum of cumulative hazard function (CHF). Thus, based on the relative risk score, we build and validate prediction model for certain survival endpoints using time-dependent receiver operating characteristic curve (ROC) (implemented with R ‘survivalROC’(25)), Kaplan-Meier curve (implemented and visualized using R packages ‘survival’ and ‘survminer’) and CoxPH model. Moreover, CBioExplorer allows to construct nomogram that included the relative risk score and other clinical features, which helps researchers and physicians predict the survival

probability of cancer patients. The nomogram can be internally and externally validated and calibrated based on bootstrap resampling and calibration analysis.

Clinical annotation

To help users investigate the clinical relevance of the biomarkers they identified, CBioExporer allows users to analyze the correlation between a given biomarker and clinical features ("Correlation with clinical features" module), characterize the prognostication significance of given biomarkers (“Kaplan-Meier curve” module, “CoxPH model” module, and “Time-dependent ROC” module), identify genes correlated with given biomarkers (“Most correlated genes” module, “Correlation with specific gene” module), compare the expression levels of given biomarkers among different groups (“Gene expression in different groups” module), and investigate the relationship between given biomarkers and immune cell infiltration (“Correlation with Immune infiltration” module(26)) and cancer stemness score(27) (“Correlation with stemness score” module).

Biological annotation

Yu and colleagues developed clusterProfiler(28), a very outstanding R language package for gene functional annotation. CBioExplorer integrates some useful functions of clusterProfiler to annotate tumor markers, allowing users to perform functional annotation of their biomarkers regarding gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Molecular Signatures Database (MSigDb), and Reactome pathway based on over representation analysis (ORA)(29,30) and gene set enrichment analysis (GSEA)(31).

Results

Curated public gene expression studies

We have reviewed, curated, normalized and integrated the gene expression data and corresponding clinical data of 43 common blood and solid tumors from GEO, TCGA, ICGC, TARGET, ArrayExpress and other public databases. These public data from 47,210 clinical samples of 268 gene expression studies (Figure 2 and https://liuxiaoping2020.github.io/CBioExplorerDatasource/) was integrated to build an R package “CuratedCancerPrognosisData”, the associated source code was deposited at https://zenodo.org/record/5728447#.Ya9vhsj1dk4.

Case study: dimensionality reduction

CBioExplorer allows 3 methods (WGCNA, SRG, DEG) to conduct dimensionality reduction. WGCNA(32) is a biometric method that can cluster genes with similar expression levels or functions into the same module, while unassigned genes are categorized into grey modules. Bladder cancer represents one of the most common types of malignant cancers of human genitourinary system(33). Kim WJ et al. published a far-reaching bladder cancer gene expression study GSE13507 in 2010, which evaluated the predictive effect of bladder cancer prognosis-related genes for patients(34). In this example, we further used the WGCNA to perform dimensionality reduction on the GSE13507, and then screened genes that were closely related to the patient’s overall survival (OS) for subsequent studies. After Euclidean distance-based clustering, 1 sample was detected as outlier, the remaining 164 samples were used to construct co-expression network (Supplementary figure 4). Then, according to the soft-thresholding power 7 (Figure 3A), we constructed a co-expression network, which clustered all genes into 8 modules

(Figure 3B). Next, we analyzed the relationship between the gene modules and the clinical features of bladder cancer patients. As a result, 'red' module was most positively relevant to the OS, while brown module was most negatively correlated with the OS of bladder cancer patients (Figure 3C-D). Finally, we could output the genes of any module or non-gray module for subsequent research.

The CoxPH model is often used clinically to evaluate the impact of clinical phenotypes on patient survival. Thus, we use the breast cancer gene expression dataset METABRIC(35) to illustrate the use of univariate CoxPH model for dimensionality reduction analysis. The METABRIC project introduced a novel risk stratification system for patients with breast cancer based on multi-omics high-throughput data. We performed univariate CoxPH model to analyze the impact of a single gene on the OS of breast cancer patients. With adjusted P value < 0.0000005 as the screening criterion, we obtained a total of 185 genes that are closely related to the OS of patients (Supplementary table 2).

Exploring the DEGs between biological groups is of great significance to clarify the biological significance of the groups. Therefore, it is also very recommended to screen differentially expressed genes for dimensionality reduction analysis. Okayama et al. conducted transcription profiling of 226 stage I–II lung adenocarcinomas, which clustered these lung adenocarcinomas into two groups (group A: patients that are mainly males, ever-smokers, and advanced stages; group B: patients that are mainly never-smokers and early stages)(36). As shown in figure 3E-H, there were 3,542 significant DEGs detected at adjusted P value <0.005 between group A lung adenocarcinomas and group lung adenocarcinomas. The DEG were visualized using heatmap, volcano plot, MA plot, and adjusted P plot.

Case study: build and validate prediction model and construct associated nomogram.

To build a prediction model, METABRIC was treated as the discovery cohort, which was then randomly stratified into training set and test set according to a ratio of 0.85. The 185 genes selected by the CoxPH model were applied to train 3 machine learning models (lasso, elastic net, and glmboost) to prediction model based on 10-fold cross validation. The performances of these models were then evaluated and compared using 100 bootstraps. As shown in supplementary figure 5, the performance of elastic net outperformed both lasso and glmboost. Thus, elastic net was selected to build the prediction model. Results of time-dependent survival ROC suggested that the areas under the curves (AUCs) of the prediction model in the training set at 3-, 5-, 10-, and 15- years were 0.736, 0.727, 0.695, and 0.665 (Figure 4A), respectively, while the corresponding AUCs in the test set were 0.748, 0.707, 0.661 and 0.667, respectively (Figure 4B). KM plot suggested that patients in the low-risk group had significant better OS compared to those in the high-risk group in training set and test (Figure 4D-E). Moreover, the risk score remained an independent prognostic factor after adjusting for other clinical features of breast cancer patients (Supplementary figure 6).

TCGA-BRCA(37) is an independent multi-omics study of breast cancer, thus, we use it as an external validation cohort. Time-dependent survival ROC analysis suggested that the AUCs of the prediction model in the validation cohort were 0.610, 0.587, 0.567, and 0.556 at 3-, 5-, 7-, and 9-years respectively (Figure 4C) and the risk score could also stratify patients into different risk groups(Figure 4F, supplementary 7).

In order to facilitate physicians and researchers to predict the patient's long-term survival rate, CBioExplorer can draw a nomogram to predict the patient's long-term survival rate. In this

example, we included the patient's risk score, ER status, tumor size, lymph node metastasis status, age and grade to draw a nomogram predicting the OS probabilities of patients at 3-, 5-, and 10-years (Figure 4G). The user can estimate the survival probability of each patient based on the 'Total point' which is the sum of the 'Points' corresponding to each clinical feature. Then, we internally and externally validated and calibrated the performance of the nomogram (Supplementary figure 8,9).

Case study: clinical annotation

In order to further clarify the clinical significance of the molecular markers screened by CBioExplorer, we chose ABAT (4-aminobutyrate aminotransferase) for clinical annotation analysis. As shown in supplementary table 3, the expression level of ABAT was closely related to the patient's ER, tumor size, Grade, and OS. Survival analysis showed that the OS of breast cancer patients in the ABAT high expression group was significantly better than that in the ABAT low expression group (Figure 5A). After adjusting for other clinical factors, the expression level of ABAT still has independent prognostic significance for breast cancer patients (Figure 5B). At the same time, Spearman's correlation analysis showed that the expression levels of 30 genes were closely related to the expression of ABAT (at adjusted P value<0.000000005) (Figure 5C). The expression of ABAT in the ER-positive group was significantly lower than that in the ER-negative group (Figure 5D). ABAT expression was also correlated with stemness score of breast cancer (Figure 5E). Meanwhile, correlation analysis between the expression of ABAT and the enrichment score of ssGSEA analysis of the gene sets provided by Bindea et al.(38) suggested that the expression of ABAT was significantly correlated with enrichment score of activated dendritic cells, B cells, CD8 T cells, cytotoxic cells, dendritic cells, eosinophils, immature dendritic cells, macrophages, neutrophils, NK CD56bright cells, NK CD56dim cells, plasmacytoid dendritic cells, T cells, T helper cells, T central memory cells, T effector memory cells, T follicular helper cells, T gamma delta cells, Th1 cells, Th17 cells, Treg cells, angiogenesis, and antigen presentation machinery (Figure 5F).

Case study: biological annotation

CBioExplorer provides a variety of enrichment analysis and corresponding visualization methods so that researchers can clarify the biological function and significance of the biomarkers they screened. In this example, we performed GO enrichment analysis on 185 genes closely related to the OS of breast cancer patients selected by CBioExplorer. ORA and GSEA were implemented respectively. Figure 6A-D showed the top 10 GO terms (chromosome segregation, sister chromatid segregation, nuclear chromosome segregation, mitotic nuclear division, mitotic sister chromatid segregation, nuclear division, organelle fission, microtubule cytoskeleton organization involved in mitosis, regulation of chromosome segregation, and regulation of sister chromatid segregation) that the 185 genes were enriched in. While results of GSEA suggested that these genes were enriched in cell cycle, cellular process, cell cycle process, and cell division (Figure 6E-H)

Discussions

CBioExplorer, to our knowledge, is the first web and standalone app that integrates multiple popular machine learning algorithms and cross-validation strategies that integrates a variety of prefaces for tumor marker screening, model evaluation and validation, and clinical and biological

annotation of tumor markers. Mover, we also developed CuratedCancerPrognosisData, an R package that reviewed, integrated 47,210 clinical samples from 268 gene expression studies of 43 common blood and solid tumors. Compared with other similar online tools or standalone apps(39-50) (Supplementary table 4) based on public data from TCGA, GEO, etc., CBioExplorer has the following advantages: (1) CBioExplorer includes the largest disease types and samples, completely open to download the curated data for academic use, and provides a personalized data submission interface for researchers to analyze their own data. (2) CBioExplorer covers many analysis modules. These analysis modules can not only be connected in series as a whole pipeline to screen, evaluate, validate and annotate tumor molecular markers, but also can analyze certain modules individually to achieve specific research purposes. For example, users can perform DEG, WGCNA, and SRG analysis separately, or they can perform clinical annotations (survival analysis, immune infiltration analysis, etc.) for some molecular markers they are interested in. (3) CBioExplorer provides an online version and a standalone local app version. The online version can help users quickly conduct research on related molecular markers. At the same time, when users involve a large amount of computing, they can download the standalone app of CBioExploer for their analysis.

Taken together, we developed and introduced CBioExplorer, a web server and standalone application that reviewed, curated and integrated the gene expression data and corresponding clinical of 47,210 clinical samples from 268 gene expression studies of 43 common blood and solid tumors, for screening, validation, and annotation of cancer survival related biomarkers based on 6 machine learning survival learners from molecular level to clinical settings.

Conflicts of interests None

Data availability

CBioExplorer is publicly available as a web server at http://cbioexplorer.znhospital.cn:8383/CBioExplorer/, the associated source code (standalone app) can be downloaded from https://github.com/liuxiaoping2020/CBioExplorer or https://gitee.com/liuxiaoping2020/CBioExplorer. The Docker image for CBioExplorer can be found at https://hub.docker.com/r/liuxiaoping2020/cbioexplorer/tags. The source code for CuratedCancerPrognosisData is deposited at https://zenodo.org/record/5728447#.Ya9vhsj1dk4.

Funding None

Acknowledgement We thank Dr. Shixiang Wang, Sun Yat-sen University Cancer Center, for his important help in the deployment process of CBioExplorer.

References

1. Roth, S.C. (2019) What is genomic medicine? *J Med Libr Assoc*, **107**, 442-448.

2. Feero, W.G. and Guttmacher, A.E. (2014) Genomics, personalized medicine, and pediatrics. *Acad Pediatr*, **14**, 14-22.

3. Shammas, M.A. (2011) Telomeres, lifestyle, cancer, and aging. *Curr Opin Clin Nutr*

Metab Care, **14**, 28-34.

4. Niu, C., Eng, L., Qiu, X., Shen, X., Espin-Garcia, O., Song, Y., Pringle, D., Mahler, M., Halytskyy, O., Charow, R. *et al.* (2015) Lifestyle Behaviors in Elderly Cancer Survivors: A Comparison With Middle-Age Cancer Survivors. *J Oncol Pract*, **11**, e450-459.

5. Iqbal, M.J., Javed, Z., Sadia, H., Qureshi, I.A., Irshad, A., Ahmed, R., Malik, K., Raza, S., Abbas, A., Pezzani, R. *et al.* (2021) Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. *Cancer Cell Int*, **21**, 270.

6. Elemento, O., Leslie, C., Lundin, J. and Tourassi, G. (2021) Artificial intelligence in cancer research, diagnosis and therapy. *Nat Rev Cancer*.

7. Davis, S. and Meltzer, P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846-1847.

8. Kauffmann, A., Rayner, T.F., Parkinson, H., Kapushesky, M., Lukk, M., Brazma, A. and Huber, W. (2009) Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics*, **25**, 2092-2094.

9. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-264.

10. Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307-315.

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

12. Wang, Y., Sanchez, L., Siegel, D.S. and Wang, M.L. (2016) Elotuzumab for the

treatment of multiple myeloma. *J Hematol Oncol*, **9**, 55.

13. Settino, M. and Cannataro, M. (2021) MMRFBiolinks: an R-package for integrating and analyzing MMRF-CoMMpass data. *Brief Bioinform*, **22**.

14. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I. *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*, **44**, e71.

15. 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 Res*, **43**, e47.

16. Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met*, **58**, 267-288.

17. Tibshirani, R. (2011) Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc B*, **73**, 273-282.

18. Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression - Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-&.

19. Hoerl, A.E. and Kennard, R.W. (2000) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **42**, 80-86.

20. Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005). *J R Stat Soc B*, **67**, 768-768.

21. Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M. (2014) Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computation Stat*, **29**, 3-35.

22. De Bin, R. (2016) Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Computation Stat*, **31**, 513-531.

23. Nasejje, J.B., Mwambi, H., Dheda, K. and Lesosky, M. (2017) A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *Bmc Med Res Methodol*, **17**.

24. Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. and Jones, Z.M. (2016) mlr: Machine Learning in R. *J Mach Learn Res*, **17**.

25. Kamarudin, A.N., Cox, T. and Kolamunnage-Dona, R. (2017) Time-dependent ROC curve analysis in medical research: current methods and applications. *Bmc Med Res Methodol*, **17**, 53.

26. Jimenez-Sanchez, A., Cast, O. and Miller, M.L. (2019) Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Res*, **79**, 6238-6246.

27. Malta, T.M., Sokolov, A., Gentles, A.J., Burzykowski, T., Poisson, L., Weinstein, J.N., Kaminska, B., Huelsken, J., Omberg, L., Gevaert, O. *et al.* (2018) Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*, **173**, 338-354 e315.

28. Yu, G., Wang, L.G., Han, Y. and He, Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284-287.

29. Maleki, F., Ovens, K., Hogan, D.J. and Kusalik, A.J. (2020) Gene Set Analysis:

Challenges, Opportunities, and Future Research. *Front Genet*, **11**, 654.

30. Leong, H.S. and Kipling, D. (2009) Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic Acids Res*, **37**, e79.

31. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

32. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

33. (2017) Bladder cancer: diagnosis and management of bladder cancer: (c) NICE (2015) Bladder cancer: diagnosis and management of bladder cancer. *BJU Int*, **120**, 755-765.

34. Kim, W.J., Kim, E.J., Kim, S.K., Kim, Y.J., Ha, Y.S., Jeong, P., Kim, M.J., Yun, S.J., Lee, K.M., Moon, S.K. *et al.* (2010) Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. *Mol Cancer*, **9**, 3.

35. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346-352.

36. Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., Furuta, K., Tsuta, K., Shibata, T., Yamamoto, S. *et al.* (2012) Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res*, **72**, 100-111.

37. Cancer Genome Atlas, N. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61-70.

38. Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A.C., Angell, H., Fredriksen, T., Lafontaine, L., Berger, A. *et al.* (2013) Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, **39**, 782-795.

39. Tang, Z., Li, C., Kang, B., Gao, G., Li, C. and Zhang, Z. (2017) GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*, **45**, W98-W102.

40. Goswami, C.P. and Nakshatri, H. (2014) PROGgeneV2: enhancements on the existing database. *BMC Cancer*, **14**, 970.

41. Aguirre-Gamboa, R., Gomez-Rueda, H., Martinez-Ledesma, E., Martinez-Torteya, A., Chacolla-Huaringa, R., Rodriguez-Barrientos, A., Tamez-Pena, J.G. and Trevino, V. (2013) SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One*, **8**, e74250.

42. Gyorffy, B., Lanczky, A., Eklund, A.C., Denkert, C., Budczies, J., Li, Q. and Szallasi, Z. (2010) An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat*, **123**, 725-731.

43. Chandrashekar, D.S., Bashel, B., Balasubramanya, S.A.H., Creighton, C.J., Ponce-Rodriguez, I., Chakravarthi, B. and Varambally, S. (2017) UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia*,

19, 649-658.

44. Liu, C.J., Hu, F.F., Xia, M.X., Han, L., Zhang, Q. and Guo, A.Y. (2018) GSCALite: a web server for gene set cancer analysis. *Bioinformatics*, **34**, 3771-3772.

45. Han, S., Kim, D., Kim, Y., Choi, K., Miller, J.E., Kim, D. and Lee, Y. (2018) CAS-viewer: web-based tool for splicing-guided integrative analysis of multi-omics cancer data. *BMC Med Genomics*, **11**, 25.

46. Anaya, J. (2016) OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *Peerj Comput Sci*.

47. Jang, Y., Seo, J., Jang, I., Lee, B., Kim, S. and Lee, S. (2019) CaPSSA: visual evaluation of cancer biomarker genes for patient stratification and survival analysis using mutation and expression data. *Bioinformatics*, **35**, 5341-5343.

48. Dong, H., Wang, Q., Li, N., Lv, J., Ge, L., Yang, M., Zhang, G., An, Y., Wang, F., Xie, L. *et al.* (2019) OSgbm: An Online Consensus Survival Analysis Web Server for Glioblastoma. *Front Genet*, **10**, 1378.

49. Gentles, A.J., Newman, A.M., Liu, C.L., Bratman, S.V., Feng, W., Kim, D., Nair, V.S., Xu, Y., Khuong, A., Hoang, C.D. *et al.* (2015) The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med*, **21**, 938-945.

50. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, **2**, 401-404.

References with PubMed links

1. Roth, S.C. (2019) What is genomic medicine? *J Med Libr Assoc*, **107**, 442-448.

http://www.ncbi.nlm.nih.gov/pubmed/31258451

http://dx.doi.org/10.5195/jmla.2019.604

2. Feero, W.G. and Guttmacher, A.E. (2014) Genomics, personalized medicine, and pediatrics. *Acad Pediatr*, **14**, 14-22.

http://www.ncbi.nlm.nih.gov/pubmed/24369865

http://dx.doi.org/10.1016/j.acap.2013.06.008

3. Shammas, M.A. (2011) Telomeres, lifestyle, cancer, and aging. *Curr Opin Clin Nutr Metab Care*, **14**, 28-34.

http://www.ncbi.nlm.nih.gov/pubmed/21102320

http://dx.doi.org/10.1097/MCO.0b013e32834121b1

4. Niu, C., Eng, L., Qiu, X., Shen, X., Espin-Garcia, O., Song, Y., Pringle, D., Mahler, M., Halytskyy, O., Charow, R. *et al.* (2015) Lifestyle Behaviors in Elderly Cancer Survivors: A Comparison With Middle-Age Cancer Survivors. *J Oncol Pract*, **11**, e450-459.

http://www.ncbi.nlm.nih.gov/pubmed/26060227

http://dx.doi.org/10.1200/JOP.2014.002287

5. Iqbal, M.J., Javed, Z., Sadia, H., Qureshi, I.A., Irshad, A., Ahmed, R., Malik, K., Raza,

S., Abbas, A., Pezzani, R. *et al.* (2021) Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. *Cancer Cell Int*, **21**, 270.

http://www.ncbi.nlm.nih.gov/pubmed/34020642

http://dx.doi.org/10.1186/s12935-021-01981-1

6. Elemento, O., Leslie, C., Lundin, J. and Tourassi, G. (2021) Artificial intelligence in cancer research, diagnosis and therapy. *Nat Rev Cancer*.

http://www.ncbi.nlm.nih.gov/pubmed/34535775

http://dx.doi.org/10.1038/s41568-021-00399-1

7. Davis, S. and Meltzer, P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846-1847.

http://www.ncbi.nlm.nih.gov/pubmed/17496320

http://dx.doi.org/10.1093/bioinformatics/btm254

8. Kauffmann, A., Rayner, T.F., Parkinson, H., Kapushesky, M., Lukk, M., Brazma, A. and Huber, W. (2009) Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics*, **25**, 2092-2094.

http://www.ncbi.nlm.nih.gov/pubmed/19505942

http://dx.doi.org/10.1093/bioinformatics/btp354

9. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U.

and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-264.

http://www.ncbi.nlm.nih.gov/pubmed/12925520

http://dx.doi.org/10.1093/biostatistics/4.2.249

10. Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307-315.

http://www.ncbi.nlm.nih.gov/pubmed/14960456

http://dx.doi.org/10.1093/bioinformatics/btg405

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

http://www.ncbi.nlm.nih.gov/pubmed/20688976

http://dx.doi.org/10.1093/bioinformatics/btq431

12. Wang, Y., Sanchez, L., Siegel, D.S. and Wang, M.L. (2016) Elotuzumab for the treatment of multiple myeloma. *J Hematol Oncol*, **9**, 55.

http://www.ncbi.nlm.nih.gov/pubmed/27417553

http://dx.doi.org/10.1186/s13045-016-0284-z

13. Settino, M. and Cannataro, M. (2021) MMRFBiolinks: an R-package for integrating and analyzing MMRF-CoMMpass data. *Brief Bioinform*, **22**.

http://www.ncbi.nlm.nih.gov/pubmed/33821961

http://dx.doi.org/10.1093/bib/bbab050

14. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I. *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*, **44**, e71.

http://www.ncbi.nlm.nih.gov/pubmed/26704973

http://dx.doi.org/10.1093/nar/gkv1507

15. 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 Res*, **43**, e47.

http://www.ncbi.nlm.nih.gov/pubmed/25605792

http://dx.doi.org/10.1093/nar/gkv007

16. Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met*, **58**, 267-288.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:A1996TU31400017

http://dx.doi.org/DOI 10.1111/j.2517-6161.1996.tb02080.x

17. Tibshirani, R. (2011) Regression shrinkage and selection via the lasso: a retrospective.

J R Stat Soc B, **73**, 273-282.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:000290575300001

http://dx.doi.org/DOI 10.1111/j.1467-9868.2011.00771.x

18. Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression - Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-&.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:A1970F898700005

http://dx.doi.org/Doi 10.1080/00401706.1970.10488634

19. Hoerl, A.E. and Kennard, R.W. (2000) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **42**, 80-86.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:000084790800013

http://dx.doi.org/Doi 10.2307/1271436

20. Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005). *J R Stat Soc B*, **67**, 768-768.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:000233203400009

http://dx.doi.org/DOI 10.1111/j.1467-9868.2005.00527.x

21. Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M. (2014) Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computation Stat*, **29**, 3-35.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:000331741200002

http://dx.doi.org/10.1007/s00180-012-0382-5

22. De Bin, R. (2016) Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Computation Stat*, **31**, 513-531.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:000374375800006

http://dx.doi.org/10.1007/s00180-015-0642-2

23. Nasejje, J.B., Mwambi, H., Dheda, K. and Lesosky, M. (2017) A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *Bmc Med Res Methodol*, **17**.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:000406576700001

http://dx.doi.org/ARTN 115

10.1186/s12874-017-0383-8

24. Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. and Jones, Z.M. (2016) mlr: Machine Learning in R. *J Mach Learn Res*, **17**.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:000391675400001

25. Kamarudin, A.N., Cox, T. and Kolamunnage-Dona, R. (2017) Time-dependent ROC curve analysis in medical research: current methods and applications. *Bmc Med Res*

Methodol, **17**, 53.

http://www.ncbi.nlm.nih.gov/pubmed/28388943

http://dx.doi.org/10.1186/s12874-017-0332-6

26. Jimenez-Sanchez, A., Cast, O. and Miller, M.L. (2019) Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Res*, **79**, 6238-6246.

http://www.ncbi.nlm.nih.gov/pubmed/31641033

http://dx.doi.org/10.1158/0008-5472.CAN-18-3560

27. Malta, T.M., Sokolov, A., Gentles, A.J., Burzykowski, T., Poisson, L., Weinstein, J.N., Kaminska, B., Huelsken, J., Omberg, L., Gevaert, O. *et al.* (2018) Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*, **173**, 338-354 e315.

http://www.ncbi.nlm.nih.gov/pubmed/29625051

http://dx.doi.org/10.1016/j.cell.2018.03.034

28. Yu, G., Wang, L.G., Han, Y. and He, Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284-287.

http://www.ncbi.nlm.nih.gov/pubmed/22455463

http://dx.doi.org/10.1089/omi.2011.0118

29. Maleki, F., Ovens, K., Hogan, D.J. and Kusalik, A.J. (2020) Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front Genet*, **11**, 654.

http://www.ncbi.nlm.nih.gov/pubmed/32695141

http://dx.doi.org/10.3389/fgene.2020.00654

30. Leong, H.S. and Kipling, D. (2009) Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic Acids Res*, **37**, e79.

http://www.ncbi.nlm.nih.gov/pubmed/19429895

http://dx.doi.org/10.1093/nar/gkp310

31. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

http://www.ncbi.nlm.nih.gov/pubmed/16199517

http://dx.doi.org/10.1073/pnas.0506580102

32. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

http://www.ncbi.nlm.nih.gov/pubmed/19114008

http://dx.doi.org/10.1186/1471-2105-9-559

33. (2017) Bladder cancer: diagnosis and management of bladder cancer: (c) NICE (2015) Bladder cancer: diagnosis and management of bladder cancer. *BJU Int*, **120**, 755-765.

http://www.ncbi.nlm.nih.gov/pubmed/29168333

http://dx.doi.org/10.1111/bju.14045

34. Kim, W.J., Kim, E.J., Kim, S.K., Kim, Y.J., Ha, Y.S., Jeong, P., Kim, M.J., Yun, S.J., Lee, K.M., Moon, S.K. *et al.* (2010) Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. *Mol Cancer*, **9**, 3.

http://www.ncbi.nlm.nih.gov/pubmed/20059769

http://dx.doi.org/10.1186/1476-4598-9-3

35. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346-352.

http://www.ncbi.nlm.nih.gov/pubmed/22522925

http://dx.doi.org/10.1038/nature10983

36. Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., Furuta, K., Tsuta, K., Shibata, T., Yamamoto, S. *et al.* (2012) Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res*, **72**, 100-111.

http://www.ncbi.nlm.nih.gov/pubmed/22080568

http://dx.doi.org/10.1158/0008-5472.CAN-11-1403

37. Cancer Genome Atlas, N. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61-70.

http://www.ncbi.nlm.nih.gov/pubmed/23000897

http://dx.doi.org/10.1038/nature11412

38. Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A.C., Angell, H., Fredriksen, T., Lafontaine, L., Berger, A. *et al.* (2013) Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, **39**, 782-795.

http://www.ncbi.nlm.nih.gov/pubmed/24138885

http://dx.doi.org/10.1016/j.immuni.2013.10.003

39. Tang, Z., Li, C., Kang, B., Gao, G., Li, C. and Zhang, Z. (2017) GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*, **45**, W98-W102.

http://www.ncbi.nlm.nih.gov/pubmed/28407145

http://dx.doi.org/10.1093/nar/gkx247

40. Goswami, C.P. and Nakshatri, H. (2014) PROGgeneV2: enhancements on the

existing database. *BMC Cancer*, **14**, 970.

http://www.ncbi.nlm.nih.gov/pubmed/25518851

http://dx.doi.org/10.1186/1471-2407-14-970

41. Aguirre-Gamboa, R., Gomez-Rueda, H., Martinez-Ledesma, E., Martinez-Torteya, A., Chacolla-Huaringa, R., Rodriguez-Barrientos, A., Tamez-Pena, J.G. and Trevino, V. (2013) SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One*, **8**, e74250.

http://www.ncbi.nlm.nih.gov/pubmed/24066126

http://dx.doi.org/10.1371/journal.pone.0074250

42. Gyorffy, B., Lanczky, A., Eklund, A.C., Denkert, C., Budczies, J., Li, Q. and Szallasi, Z. (2010) An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat*, **123**, 725-731.

http://www.ncbi.nlm.nih.gov/pubmed/20020197

http://dx.doi.org/10.1007/s10549-009-0674-9

43. Chandrashekar, D.S., Bashel, B., Balasubramanya, S.A.H., Creighton, C.J., Ponce-Rodriguez, I., Chakravarthi, B. and Varambally, S. (2017) UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia*, **19**, 649-658.

http://www.ncbi.nlm.nih.gov/pubmed/28732212

http://dx.doi.org/10.1016/j.neo.2017.05.002

44. Liu, C.J., Hu, F.F., Xia, M.X., Han, L., Zhang, Q. and Guo, A.Y. (2018) GSCALite: a web server for gene set cancer analysis. *Bioinformatics*, **34**, 3771-3772.

http://www.ncbi.nlm.nih.gov/pubmed/29790900

http://dx.doi.org/10.1093/bioinformatics/bty411

45. Han, S., Kim, D., Kim, Y., Choi, K., Miller, J.E., Kim, D. and Lee, Y. (2018) CAS-viewer: web-based tool for splicing-guided integrative analysis of multi-omics cancer data. *BMC Med Genomics*, **11**, 25.

http://www.ncbi.nlm.nih.gov/pubmed/29697367

http://dx.doi.org/10.1186/s12920-018-0348-8

46. Anaya, J. (2016) OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *Peerj Comput Sci*.

http://www.ncbi.nlm.nih.gov/pubmed/WOS:000437457800001

http://dx.doi.org/ARTN e67

10.7717/peerj-cs.67

47. Jang, Y., Seo, J., Jang, I., Lee, B., Kim, S. and Lee, S. (2019) CaPSSA: visual evaluation of cancer biomarker genes for patient stratification and survival analysis

using mutation and expression data. *Bioinformatics*, **35**, 5341-5343.

http://www.ncbi.nlm.nih.gov/pubmed/31228188

http://dx.doi.org/10.1093/bioinformatics/btz516

48. Dong, H., Wang, Q., Li, N., Lv, J., Ge, L., Yang, M., Zhang, G., An, Y., Wang, F., Xie, L. *et al.* (2019) OSgbm: An Online Consensus Survival Analysis Web Server for Glioblastoma. *Front Genet*, **10**, 1378.

http://www.ncbi.nlm.nih.gov/pubmed/32153627

http://dx.doi.org/10.3389/fgene.2019.01378

49. Gentles, A.J., Newman, A.M., Liu, C.L., Bratman, S.V., Feng, W., Kim, D., Nair, V.S., Xu, Y., Khuong, A., Hoang, C.D. *et al.* (2015) The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med*, **21**, 938-945.

http://www.ncbi.nlm.nih.gov/pubmed/26193342

http://dx.doi.org/10.1038/nm.3909

50. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, **2**, 401-404.

http://www.ncbi.nlm.nih.gov/pubmed/22588877

http://dx.doi.org/10.1158/2159-8290.CD-12-0095

Figure 1. Overview of CBioExplorer. The whole pipeline of CBioExplorer includes data collection, data curation, dimensionality reduction using three methods (WGCNA, univariate Cox proportional hazards regression model, differentially expressed gene analysis), benchmark experiment with 6 machine learning learners (Lasso, Ridge, Elastic net, Glmboost, Coxboost, Randomforest) using cross validation and nested cross validation based on R package, prediction model construction using Cox proportional hazards regression model and nomogram, clinical annotation using a variety of clinical approaches, and biological annotation using over-representation analysis (ORA) and gene set enrichment analysis (GSEA). Abbreviations: TCGA, The Cancer Genome Atlas; GEO, gene expression omnibus; TARGET, Therapeutically Applicable Research To Generate Effective Treatments; GEP, expression profile; ICGC, International Cancer Genome Consortium; WGCNA, weighted gene co-expression network analysis; SRG, survival- related gene; DEG, differentially expressed gene; CV, cross validation; nCV, nested cross validation;

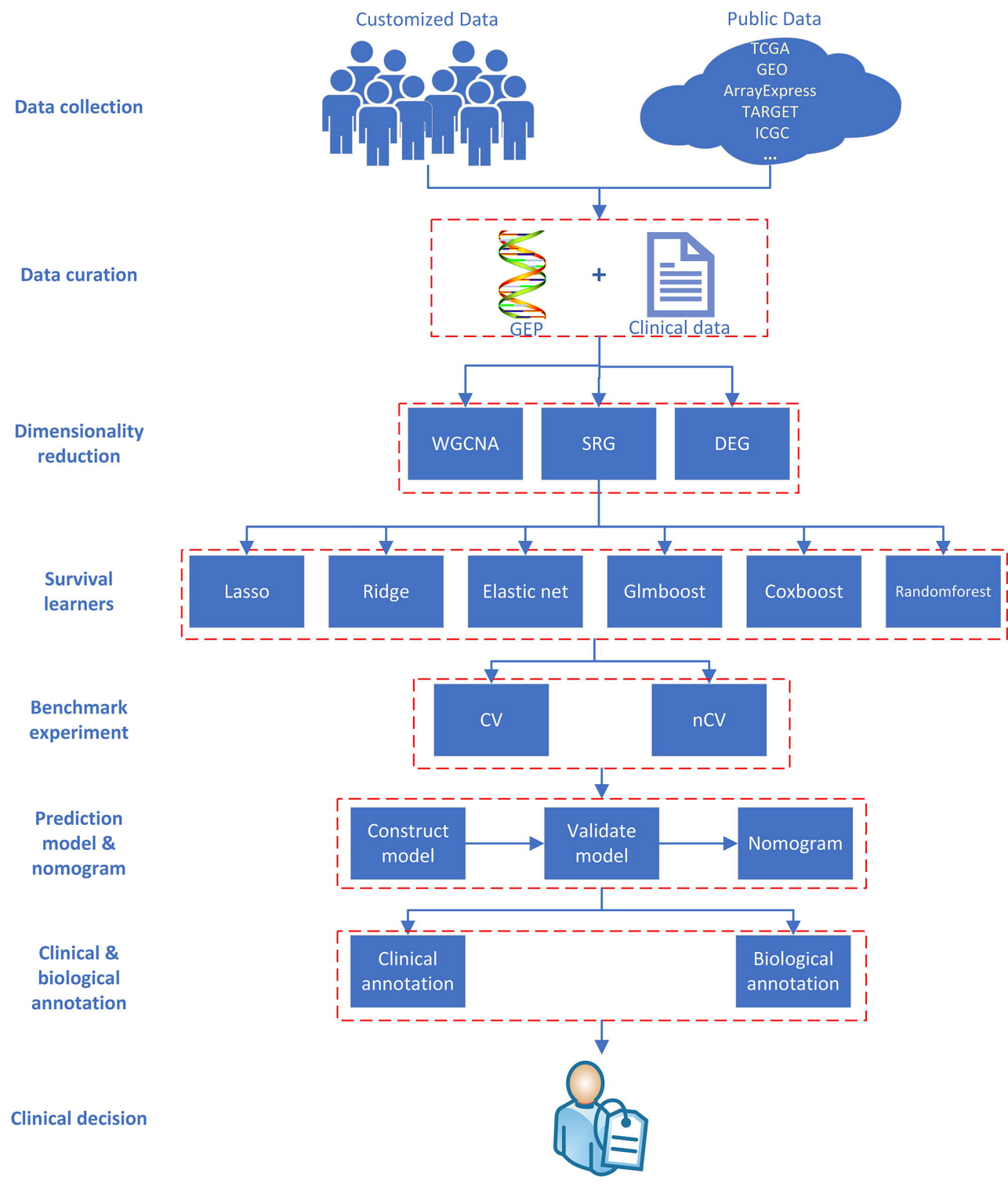
Figure 2. Total number of samples and datasets of included in CuratedCancerPrognosisData.

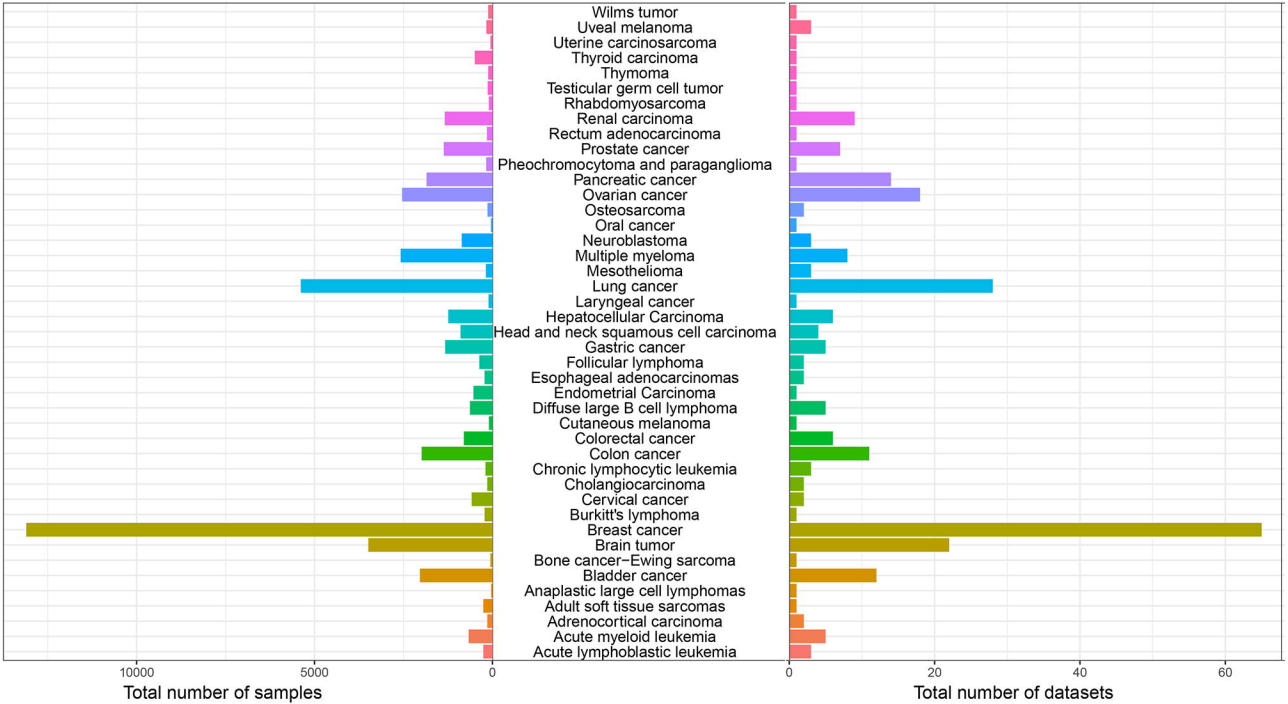
Figure 3. Main result output of WGCNA and DEG analysis. (A) Selection of soft threshold power; (B). Modules detected by network analysis. (C) Module -trait relationships. (D) Gene Significance (GS) for grade vs. Module Membership (MM) in the turquoise module. (E) Heatmap showing differentially expressed genes in group A and group B. (F) Volcano plot showing differentially expressed genes in group A and group B. (G) MA plot showing differentially expressed genes in group A and group B. (H) P adjusting plot of differentially expressed gene analysis.

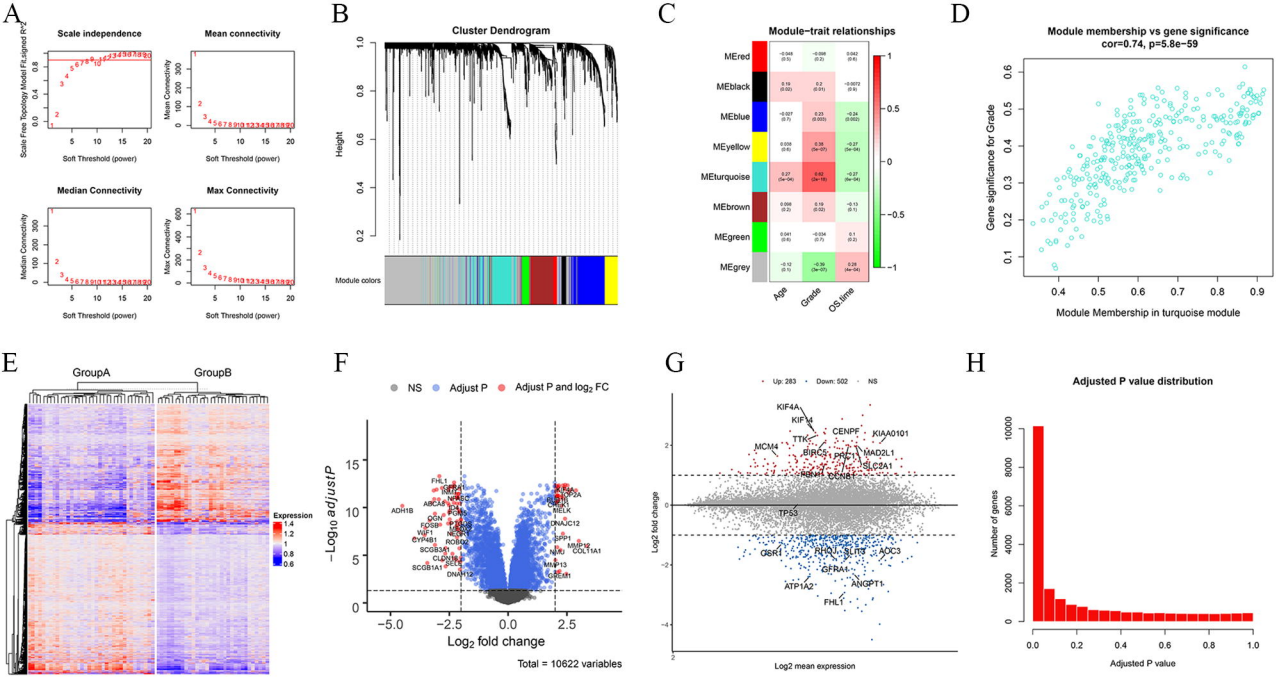
Figure 4. Construction and evaluation and translation of the prediction model. (A) Areas under the curve (AUCs) of time dependent ROC analysis at 3-, 5-. 10-, 15- year in the training set. (B) AUCs of time dependent ROC analysis at 3-, 5-. 10-, 15- year in the test set. (C) AUCs of time dependent ROC analysis at 3-, 5-. 10-, 15- year in the validation set. (D) Survival differences between the low-risk group and high-risk group in the training set. (E) Survival differences between the low-risk group and high-risk group in the test set. (F) Survival differences between the low-risk group and high-risk group in the validation set. (G) Nomogram prediction the 3-, 5-, 7- year survival probabilities based on CoxPH model that integrates the risk, ER status, tumor size, node status, age, and grade of breast cancer patients.

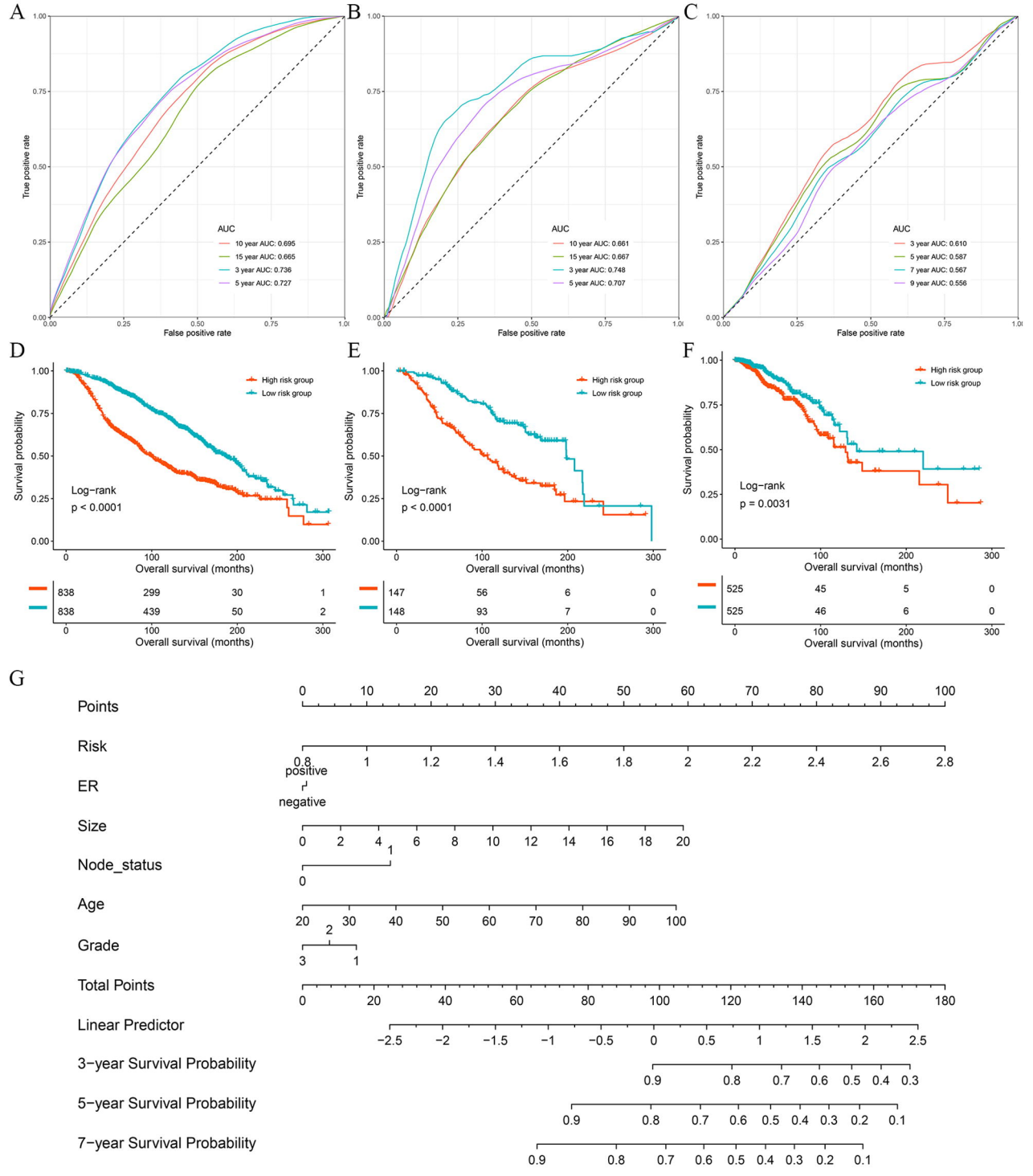
Figure 5. Clinical annotation ABAT in the METABRIC cohort. (A) Survival differences between ABAT low and high expression groups. (B) CoxPH model identifying the prediction ability of ABAT. (C) Most correlated genes of ABAT based on Spearman's correlation analysis. (D) Relative expression of ABAT in the ER- negative and ER- positive groups. (E) Correlation between the relative expression ABAT and the enrichment score of stemness. (F) Correlation between the relative expression ABAT and the enrichment scores of the immune cells infiltrated in the microenvironment of breast cancer cells.

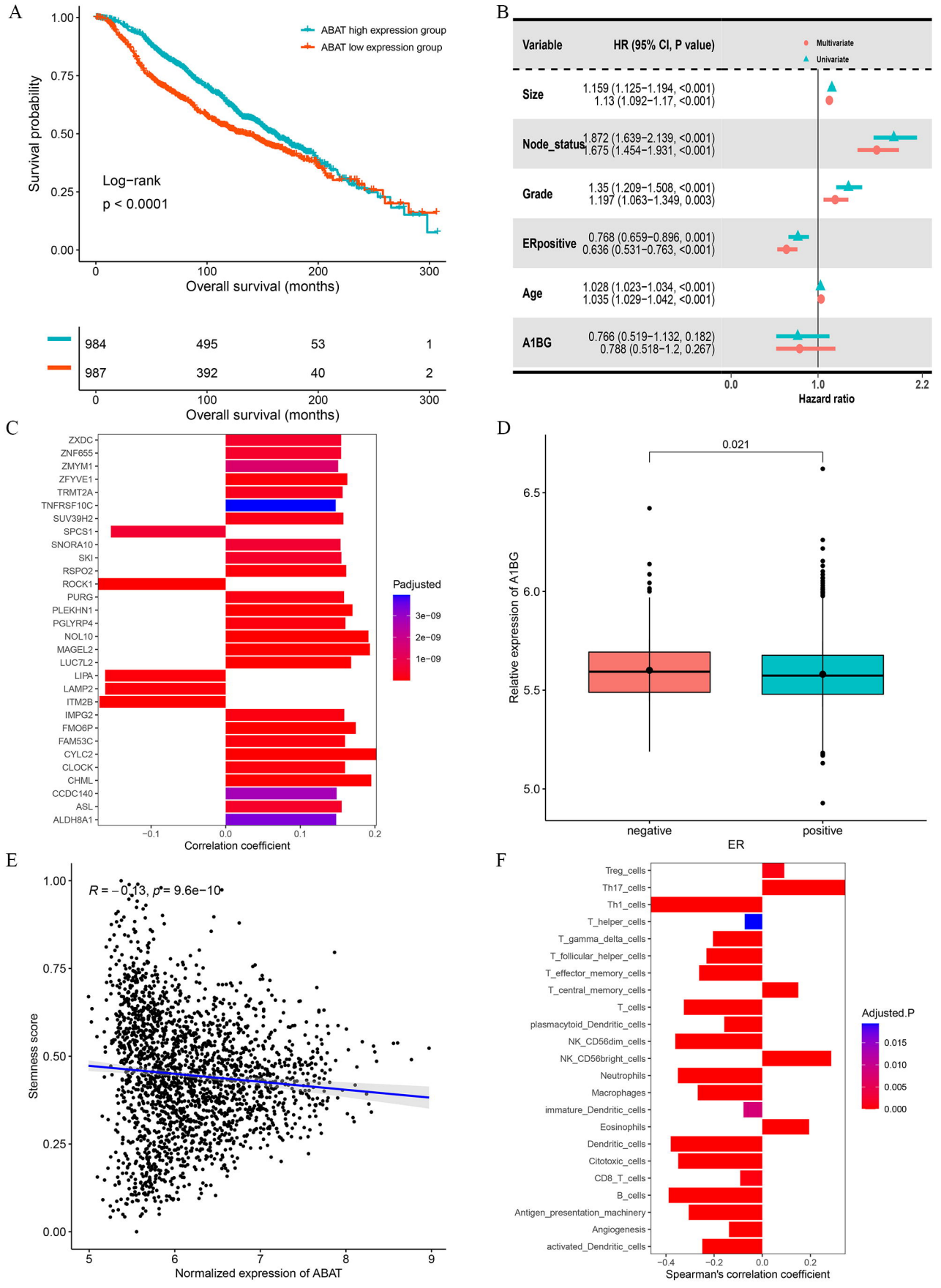
Figure 6. Functional enrichment analysis of the biomarkers identified by CBioExplorer. (A-D) Top 10 gene ontology biological process terms that enriched in the biomarkers. (E-H) Top gene ontology biological process terms identified by gene set enrichment analysis.

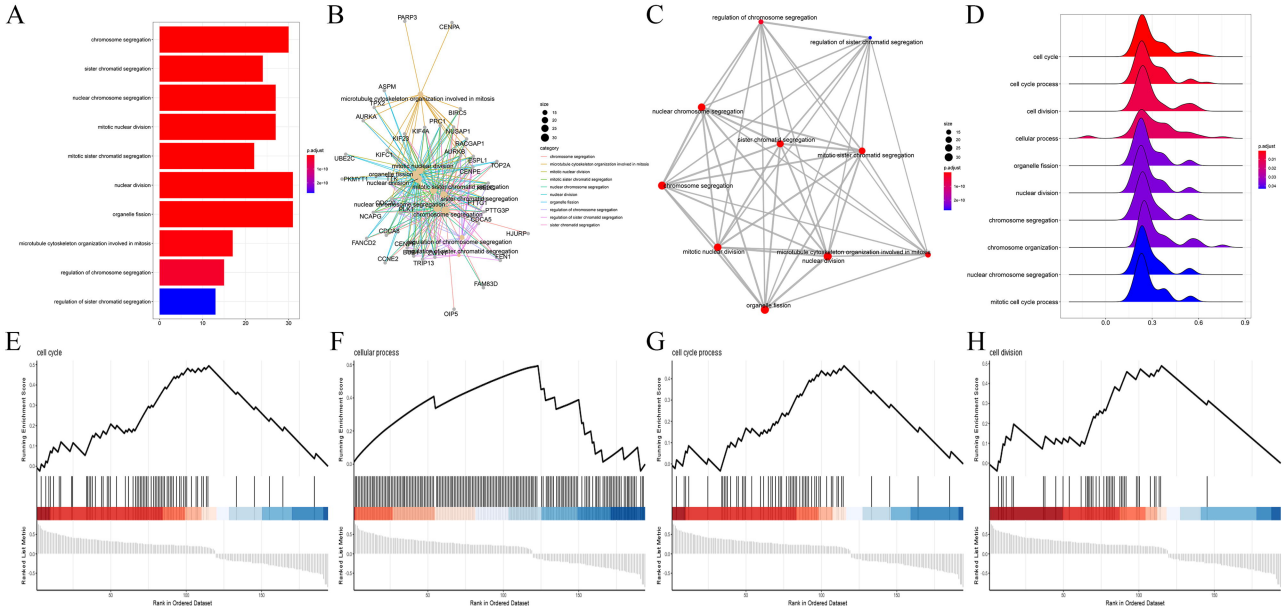

Customized Data
Public Data
TCGA
GEO
ArrayExpress
TARGET
ICGC
...
Data collection
Data curation
GEP
+
Clinical data
Dimensionality reduction
WGCNA
SRG
DEG
Survival learners
Lasso
Ridge
Elastic net
Glmboost
Coxboost
Randomforest
Benchmark experiment
CV
nCV
Prediction model & nomogram
Construct model
Validate model
Nomogram
Clinical & biological annotation
Clinical annotation
Biological annotation
Clinical decision


Wilms tumor
Uveal melanoma
Uterine carcinosarcoma
Thyroid carcinoma
Thymoma
Testicular germ cell tumor
Rhabdomyosarcoma
Renal carcinoma
Rectum adenocarcinoma
Prostate cancer
Pheochromocytoma and paraganglioma
Pancreatic cancer
Ovarian cancer
Osteosarcoma
Oral cancer
Neuroblastoma
Multiple myeloma
Mesothelioma
Lung cancer
Laryngeal cancer
Hepatocellular Carcinoma
Head and neck squamous cell carcinoma
Gastric cancer
Follicular lymphoma
Esophageal adenocarcinomas
Endometrial Carcinoma
Diffuse large B cell lymphoma
Cutaneous melanoma
Colorectal cancer
Colon cancer
Chronic lymphocytic leukemia
Cholangiocarcinoma
Cervical cancer
Burkitt's lymphoma
Breast cancer
Brain tumor
Bone cancer−Ewing sarcoma
Bladder cancer
Anaplastic large cell lymphomas
Adult soft tissue sarcomas
Adrenocortical carcinoma
Acute myeloid leukemia
Acute lymphoblastic leukemia
10000
5000
0
Total number of samples
0
20
40
60
Total number of datasets


A
Scale independence
Scale Free Topology Model Fit,signed R²
Soft Threshold (power)
Mean connectivity
Mean Connectivity
Median Connectivity
Max Connectivity
B
Cluster Dendrogram
Height
Module colors
C
Module−trait relationships
MEred
MEblack
MEblue
MEyellow
MEturquoise
MEbrown
MEgreen
MEgrey
Age
Grade
OS_time
D
Module membership vs gene significance
cor=0.74, p=5.8e−59
Gene significance for Grade
Module Membership in turquoise module
E
GroupA
GroupB
Expression
F
NS
Adjust P
Adjust P and log2 FC
−Log10 adjustP
Log2 fold change
Total = 10622 variables
G
Up: 283
Down: 502
NS
Log2 fold change
Log2 mean expression
H
Adjusted P value distribution
Number of genes
Adjusted P value


A
True positive rate
False positive rate
AUC
10 year AUC: 0.695
15 year AUC: 0.665
3 year AUC: 0.736
5 year AUC: 0.727
B
True positive rate
False positive rate
AUC
10 year AUC: 0.661
15 year AUC: 0.667
3 year AUC: 0.748
5 year AUC: 0.707
C
True positive rate
False positive rate
AUC
3 year AUC: 0.610
5 year AUC: 0.587
7 year AUC: 0.567
9 year AUC: 0.556
D
Survival probability
High risk group
Low risk group
Log−rank
p < 0.0001
Overall survival (months)
838 299 30 1
838 439 50 2
Overall survival (months)
E
Survival probability
High risk group
Low risk group
Log−rank
p < 0.0001
Overall survival (months)
147 56 6 0
148 93 7 0
Overall survival (months)
F
Survival probability
High risk group
Low risk group
Log−rank
p = 0.0031
Overall survival (months)
525 45 5 0
525 46 6 0
Overall survival (months)
G
Points
Risk
ER
positive
negative
Size
Node_status
Age
Grade
Total Points
Linear Predictor
3−year Survival Probability
5−year Survival Probability
7−year Survival Probability


A
ABAT high expression group
ABAT low expression group
Survival probability
Log−rank
p < 0.0001
Overall survival (months)
984 495 53 1
987 392 40 2
Overall survival (months)
B
Variable
HR (95% CI, P value)
Multivariate
Univariate
Size
1.159 (1.125−1.194, <0.001)
1.13 (1.092−1.17, <0.001)
Node_status
1.872 (1.639−2.139, <0.001)
1.675 (1.454−1.931, <0.001)
Grade
1.35 (1.209−1.508, <0.001)
1.197 (1.063−1.349, 0.003)
ERpositive
0.768 (0.659−0.896, 0.001)
0.636 (0.531−0.763, <0.001)
Age
1.028 (1.023−1.034, <0.001)
1.035 (1.029−1.042, <0.001)
A1BG
0.766 (0.519−1.132, 0.182)
0.788 (0.518−1.2, 0.267)
Hazard ratio
C
Padjusted
Correlation coefficient
D
0.021
Relative expression of A1BG
negative
positive
ER
E
R = −0.13, p = 9.6e−10
Stemness score
Normalized expression of ABAT
F
Adjusted.P
Spearman's correlation coefficient


A
chromosome segregation
sister chromatid segregation
nuclear chromosome segregation
mitotic nuclear division
mitotic sister chromatid segregation
nuclear division
organelle fission
microtubule cytoskeleton organization involved in mitosis
regulation of chromosome segregation
regulation of sister chromatid segregation
p.adjust
1e-10
2e-10
B
PARP3
CENPA
ASPM
microtubule cytoskeleton organization involved in mitosis
TEX2
AURKA
BIRC5
PRC1
KIF4A
NUSAP1
KIF23
RACGAP1
KIFC1
AURKB
UBE2C
ESPL1
TOP2A
mitotic nuclear division
CENPE
organelle fission
nuclear division
mitotic sister chromatid segregation
sister chromatid segregation
nuclear chromosome segregation
PTTG1
PTTG3P
NCAPG
chromosome segregation
CDCA5
CDCA8
HJURP
FANCD2
regulation of chromosome segregation
regulation of sister chromatid segregation
CCNE2
TRIP13
KEN1
FAM83D
CIP5
size
category
C
regulation of chromosome segregation
regulation of sister chromatid segregation
nuclear chromosome segregation
sister chromatid segregation
mitotic sister chromatid segregation
chromosome segregation
mitotic nuclear division
microtubule cytoskeleton organization involved in mitosis
nuclear division
organelle fission
D
cell cycle
cell cycle process
cell division
cellular process
organelle fission
nuclear division
chromosome segregation
chromosome organization
nuclear chromosome segregation
mitotic cell cycle process
p.adjust
E
cell cycle
F
cellular process
G
cell cycle process
H
cell division
Running Enrichment Score
Ranked List Metric
Rank in Ordered Dataset