W98–W102 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 12 April 2017*
doi: 10.1093/nar/gkx247

GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses

Zefang Tang[1,†], Chenwei Li[1,2,†], Boxi Kang[1], Ge Gao[3], Cheng Li[2,3] and Zemin Zhang[1,2,4,5,*]

[1]BIOPIC, School of Life Sciences, Peking University, Beijing 100871, China, [2]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China, [3]School of Life Sciences, Peking University, Beijing 100871, China, [4]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA and [5]Beijing Advanced Innovation Center for Genomics, Peking University, Beijing 100871, China

Received February 12, 2017; Revised March 27, 2017; Editorial Decision March 31, 2017; Accepted April 05, 2017

ABSTRACT

Tremendous amount of RNA sequencing data have been produced by large consortium projects such as TCGA and GTEx, creating new opportunities for data mining and deeper understanding of gene functions. While certain existing web servers are valuable and widely used, many expression analysis functions needed by experimental biologists are still not adequately addressed by these tools. We introduce GEPIA (Gene Expression Profiling Interactive Analysis), a web-based tool to deliver fast and customizable functionalities based on TCGA and GTEx data. GEPIA provides key interactive and customizable functions including differential expression analysis, profiling plotting, correlation analysis, patient survival analysis, similar gene detection and dimensionality reduction analysis. The comprehensive expression analyses with simple clicking through GEPIA greatly facilitate data mining in wide research areas, scientific discussion and the therapeutic discovery process. GEPIA fills in the gap between cancer genomics big data and the delivery of integrated information to end users, thus helping unleash the value of the current data resources. GEPIA is available at http://gepia.cancer-pku.cn/.

INTRODUCTION

High-throughput RNA sequencing (RNA-Seq) has emerged as a powerful method for transcriptomic analysis (1), widely used for understanding gene functions and biological patterns, finding candidate drug targets and identifying biomarkers for disease classification and diagnosis (2). In recent years, the Cancer Genome Atlas (TCGA) (3) and Genotype-Tissue Expression (GTEx) (4,5) projects produced RNA-Seq data for tens of thousands of cancer and non-cancer samples, providing an unprecedented opportunity for many related fields including cancer biology. TCGA thus far has produced RNA-Seq data for 9736 tumor samples across 33 cancer types, in addition to data for 726 adjacent normal tissues. The imbalance between the tumor and normal data can cause inefficiency in various differential analyses. Fortunately, the GTEx project produced RNA-Seq data for over 8000 normal samples, albeit from unrelated donors. Such data cannot be directly combined for integrated analysis due to many differences in aspects like data processing pipelines and gene models. To make data from different sources more compatible, the UCSC Xena project (http://xena.ucsc.edu/) has recomputed all expression raw data based on a standard pipeline to minimize differences from distinct sources, thus allowing for the formation of the most comprehensive expression data up to date.

Methods for analyzing gene expression are numerous and diverse. Expression-based clustering, for example, can be divided into supervised and unsupervised methods. Gene expression differential analysis is a classical supervised method, leading to the finding tumor-specific genes by comparing tumor to normal groups. Those tumor-specific genes coding for 'targetable' proteins are often pursued as candidates for downstream analysis (6), such as those found as potential drug targets in prostate, colon and ovarian cancers (7–9). In addition, principal component analysis (PCA) is a common unsupervised method to reduce the dimensionality of high dimensional expression datasets while maintaining most of the variances. Li *et al*., for instance, discovered that T cell receptor (TCR) variable genes in brain cancer were different from other cancer types using PCA analysis based on the TCGA RNA-Seq data (10). Survival analysis based on gene expression levels is also widely used for evaluating the clinical importance of a given gene (11). Furthermore, since genes with similar expression patterns are likely to have related functions, it is often desirable to iden-

*To whom correspondence should be addressed. Tel: +86 10 6276 8190; Fax: +86 10 6276 8190; Email: zemin@pku.edu.cn
†These authors contributed equally to the paper as first authors.

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

tify genes with expression similarities to a known gene, such as a known cancer drug target, using an appropriate distance metric (e.g. Pearson's correlation coefficient) (6).

Currently, Xena, cBioPortal (http://www.cbioportal.org/), HPA (12) and Expression Atlas (13) have provided many useful visualization and analysis tools for gene expression analysis. Among numerous useful features provided by these tools, multiple types of genomic alteration data can be simultaneously displayed by cBioPortal. HPA is best at integrating protein information, while Expression Atlas distinguishes itself by providing multi-species expression data. While these web servers are exceptionally valuable and widely used, many additional expression analysis functions are often requested by experimental biologists but are not adequately addressed by those existing tools. For example, differential expression analyses are commonly performed, but this function is not available in cBioPortal, Xena or HPA, while Expression Atlas does not allow detailed tumor-normal comparison. In addition, while cBioPortal provides survival analysis based on mutation status, it lacks customizable selection of gene expression thresholds for patient cohort partitioning. Furthermore, none of the existing tools allow analyses based on the relationship of a pair of genes (e.g. sample partitioning based on the expression of one gene normalized by another). Moreover, none of the tools provide chromosomal distribution plots, similar gene detection, dimensionality reduction analysis or expression comparison among pathological stages. Based on the above needs that are not adequately addressed, we developed GEPIA (Gene Expression Profiling Interactive Analysis), a web-based tool to deliver fast and customizable functionalities to complement with the existing tools.

MATERIALS AND METHODS

Implementations

The GEPIA website is freely available to all users. It is built by the HTML5 and JavaScript libraries, including jQuery (http://jquery.com), Bootstrap (http://getbootstrap.com/) for the client-side user interface. The server-side and interactive data processing are carried out by PHP scripts (version 7.0.13). The web site automatically adjusts the look and feel according to different browsers and devices, ranging from desktop computers to tablets and smart phones. There is no login requirement for accessing any features in GEPIA.

To solve the imbalance between the tumor and normal data which can cause inefficiency in various differential analyses, we download the TCGA and GTEx gene expression data that are re-computed from raw RNA-Seq data by the UCSC Xena project based on a uniform pipeline (Figure 1). We consult with medical experts to determine the most appropriate sample grouping for tumor-normal comparisons. The datasets are stored in a MySQL relational database (version 5.7.17).

The GEPIA web server features are divided into seven major tabs: General, Differential Genes, Expression DIY, Survival, Similar Genes, Correlation and PCA, which provides key interactive functions corresponding to differential expression analysis, customizable profiling plotting, patient survival analysis, similar gene detection, correlation analysis and dimensionality reduction analysis (Figure 2).

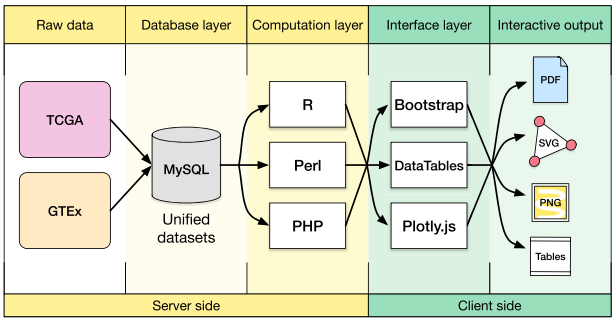

Figure 1. Schema describing data processing and data display for the GEPIA visualization tool.

All plotting features in GEPIA are developed using R (version 3.3.2) and Perl (version 5.22.1) programs. The GEPIA outputs consist of plots and tables. Static visualizations are rendered as Portable Document Format (PDF), Scalable Vector Graphics (SVG) and Portable Network Graphics (PNG) images. The rotatable 3D plots are built by the plotly.js library (https://plot.ly/). Tables are generated by the DataTables (https://www.datatables.net/) JavaScript library, allowing for data querying and selection.

Functionalities

Quick start. GEPIA provides a simple search interface. Users can enter a gene symbol (e.g. *ERBB2*) or an Ensembl ID (e.g. ENSG00000141736) in the 'Enter gene name' field to search for a gene of interest. Clicking the 'GoPIA!' button will lead to the expression profile of the input gene across all tumor and normal tissues in dot plots or body maps (Figure 2A). The dot plots and body maps are similar to those presented by the GE-mini mobile application (14). Users can also obtain basic gene annotation and a list of genes that are most similar to the input gene based on the entire datasets of tumor and normal samples.

Differential analysis. It is a common task to screen for candidate cancer drug targets, oncogenes or suppressor genes from differentially expressed genes based on the comparison of tumors and either matched normal or all normal samples (15,16). Meanwhile, genes with similar expression patterns that cluster along the chromosome often suggest the underlying genomic mechanisms leading to special expression characteristics (17,18). Accordingly, GEPIA allows users to input custom statistical methods and thresholds for a given dataset to dynamically obtain differentially expressed genes and their chromosomal distribution (Figure 2B). The statistical methods are presented in supplementary online information (the 'Help' page in GEPIA website).

Expression DIY. GEPIA dynamically plots expression profiles of a given gene according to user-defined sample selections and methods (Figure 2C). The results can be presented in dot plots (in the 'profile' tab) or box plots (in

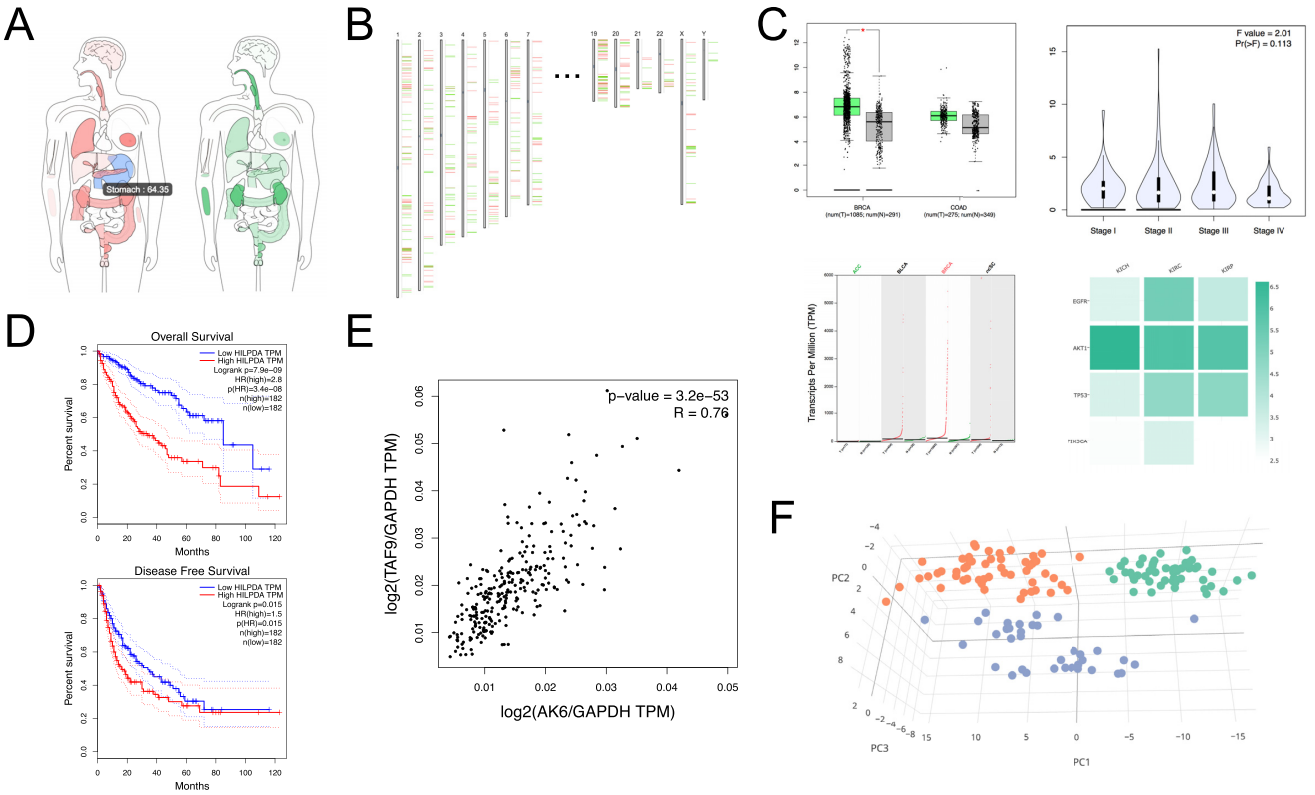

Figure 2. Examples of GEPIA outputs. (**A**) Users can check gene expression by body map in the 'General' tab. (**B**) The differential genes provided by GEPIA can be plotted as chromosomal distribution in the 'Differential Genes' tab. (**C**) GEPIA provides box plots, violin plots based on pathological stages, dot plots and matrix plots in the 'Expression DIY' tab. (**D**) The overall survival and disease-free survival analysis of gene of interest can be presented in the 'Survival' tab. Meanwhile, genes with the most significant association with patient survival can be identified. (**E**) GEPIA provides pairwise gene correlation analysis for given sets of TCGA and/or GTEx expression data in the 'Correlation' tab. Genes of interest can be normalized by other gene. (**F**) Principal component analysis results based on a set of input genes can be presented in the 'PCA' tab.

the 'Boxplot' tab). In addition, GEPIA plots gene expression by pathological stages based on the TCGA clinical annotation. Furthermore, for rapid comparison between different genes across multiple tissue types, GEPIA provides matrix plots based on input gene list and datasets of interest. The color density of each block represents the median expression value of a gene in a given tissue type, normalized by the maximum median expression value across all blocks. GEPIA provides plot modification parameters such as width, jitter size and group colors in various DIY Expression sub features.

Survival analysis. GEPIA performs survival analysis based on gene expression levels (Figure 2D). This function allows users to select their custom cancer types for overall or disease-free survival analysis. For example, to examine the survival curves of an input gene in lung cancer, a user can select lung squamous cell carcinoma (LUSC) only or choose both LUSC and lung adenocarcinoma for the survival analyses. GEPIA uses log-rank test, sometimes called the Mantel–Cox test, for the hypothesis evaluation. The cox proportional hazard ratio and the 95% confidence interval information can also be included in the survival plot. The thresholds for high/low expression level cohorts can be adjusted.

For survival analysis, GEPIA also provides a gene normalization feature that allows the relative expression of two different genes as input. For example, when investigating gene *FOXP3* in cancer survival analysis, users can also input another gene such as *CD3G* to normalize the expression of *FOXP3*. In such case, GEPIA will perform the survival analysis based on the *FOXP3*/*CD3G* relative expression levels. Furthermore, GEPIA can also present top genes that are most associated with cancer patient survival. The gene list is ranked by *P*-values of survival analysis based on any input cancer types.

Similar genes. With this function, users can rapidly identify additional genes with expression features similar to an input gene of interest, such as a known drug target. Datasets can be selected to represent one cancer type or multiple cancer and normal tissue combinations, and this function reports a list of genes with similar expression pattern to the input gene across selected datasets.

Correlation analysis. This function performs pairwise gene correlation analysis for any given sets of TCGA and/or GTEx expression data (Figure 2E), using methods such as the Pearson, Spearman and Kendall correlation statistics. For this feature, one gene can also be normalized by other

gene. For example, users can examine the simple correlation coefficient between the *AK6-TAF9* gene pair, or check the correlation analysis result between *AK6/GAPDH* and *TAF9/GAPDH* relative ratios.

Dimensionality reduction. For a given gene list and sample dataset, GEPIA provides PCA, yielding the rotatable 3D plots (Figure 2F). This feature could reveal subsets of certain cancer type as stratified by input gene list, or confirm whether a gene set could be further used as effective biomarkers. GEPIA presents a 3D plot of top three principal components (PC) and generates a bar plot for variances interpreted by each PC. GEPIA also presents 2D plot or 3D plot based on user-specified PCs.

Results availability

After submission of an analysis request, GEPIA will provide the vector image result for users. All the results provided by GEPIA are publication-ready. The PDF and the SVG download is available by clicking the button next to the results. A tutorial and an example video is also available in the 'Help' page in GEPIA. These vector statistical plots can be downloaded for modification using Adobe Illustrator.

Documentation

GEPIA documentation is available and can be accessed by clicking the 'Help' link in the top right navigation bar. The documentation contains the description of each feature function and the introduction of parameters in each feature as well as the results of each analysis. Meanwhile, GEPIA also provides an 'Example' link for quick view of all GEPIA features in the top right navigation bar. In addition to these links, users can click the 'Help' button in each feature tab to open the collapsed tooltips that give concise explanations and detail of each parameter.

DISCUSSION

GEPIA is an interactive web application for gene expression analysis based on 9736 tumors and 8587 normal samples from the TCGA and the GTEx databases, using the output of a standard processing pipeline for RNA sequencing data. Analysis results cover ∼20 000 coding and ∼25 000 non-coding genes, as well as ∼14 000 pseudogenes and ∼400 T-cell receptor segments.

GEPIA enables experimental biologists without any computational programming skills to perform a diverse range of gene expression analyses. By using GEPIA, experimental biologists can easily explore the large TCGA and GTEx datasets, ask specific questions and test their hypotheses. For example, one can easily find the *MPO* gene specifically expressed in leukemia and *UPK2* specifically expressed in bladder cancer. Genes with the most significant association with patient survival can be identified, including the *MCTS1* gene in breast cancer and the *HILPDA* gene in liver cancer. Genes like *PGAP3* and *GRB7* can also be quickly identified to have similar expression pattern with cancer drug target *ERBB2*. Meanwhile, the flexible customization parameters of GEPIA also enable users to extensively customize the visualization, for example, by changing the colors or modifying the width of expression plots.

GEPIA is a time-saving and intuitive tool for unleashing the value of the big genomic data in TCGA and GTEx. It complements well with other available tools such as cBioPortal and Expression Atlas. With the continuous user feedback and further enhancement, GEPIA has the potential to become an integral part of routine data analyses for experimental biologists.

ACKNOWLEDGEMENTS

C.L. and B.K. built the server base system; Z.T., C.L. and B.K. designed the user interface; Z.T. constructed the integrated datasets, and constructed the backend computation pipeline; Z.T. and C.L. developed the interactive analysis tools; Z.T. and B.K. wrote the server documents; C.L and G.G. provided critical input on analysis methods. Z.Z. obtained funding and supervised the project, and oversaw the manuscript preparation. We thank Lei Zhang, Chunhong Zheng and Xinyi Guo among many others for providing critical comments.

FUNDING

Beijing Advanced Innovation Center for Genomics; National Natural Science Foundation of China [Program 31530036, 81573022]; Key Technologies R&D Program [2016YFC0900100]. Funding for open access charge: Beijing Advanced Innovation Center for Genomics.
Conflict of interest statement. None declared.

REFERENCES

1. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
2. Loven,J., Orlando,D.A., Sigova,A.A., Lin,C.Y., Rahl,P.B., Burge,C.B., Levens,D.L., Lee,T.I. and Young,R.A. (2012) Revisiting global gene expression analysis. *Cell*, **151**, 476–482.
3. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
4. Ardlie,K.G., Deluca,D.S., Segrè,A.V., Sullivan,T.J., Young,T.R., Gelfand,E.T., Trowbridge,C.A., Maller,J.B., Tukiainen,T., Lek,M. *et al.* (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
5. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
6. Desany,B. and Zhang,Z. (2004) Bioinformatics and cancer target discovery. *Drug Discov. Today*, **9**, 795–802.
7. Gray,D., Jubb,A.M., Hogue,D., Dowd,P., Kljavin,N., Yi,S., Bai,W., Frantz,G., Zhang,Z., Koeppen,H. *et al.* (2005) Maternal embryonic leucine zipper kinase/murine protein serine-threonine kinase 38 is a promising therapeutic target for multiple cancers. *Cancer Res.*, **65**, 9751–9761.
8. Welsh,J.B., Sapinoso,L.M., Su,A.I., Kern,S.G., Wang-Rodriguez,J., Moskaluk,C.A., Frierson,H.F. Jr and Hampton,G.M. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.*, **61**, 5974–5978.
9. Welsh,J.B., Zarrinkar,P.P., Sapinoso,L.M., Kern,S.G., Behling,C.A., Monk,B.J., Lockhart,D.J., Burger,R.A. and Hampton,G.M. (2001) Analysis of gene expression profiles in normal and neoplastic ovarian

tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 1176–1181.
10. Li,B., Li,T., Pignon,J.C., Wang,B., Wang,J., Shukla,S.A., Dou,R., Chen,Q., Hodi,F.S., Choueiri,T.K. *et al.* (2016) Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.*, **48**, 725–732.
11. Plitas,G., Konopacki,C., Wu,K., Bos,P.D., Morrow,M., Putintseva,E.V., Chudakov,D.M. and Rudensky,A.Y. (2016) Regulatory T cells exhibit distinct features in human breast cancer. *Immunity*, **45**, 1122–1134.
12. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
13. Petryszak,R., Keays,M., Tang,Y.A., Fonseca,N.A., Barrera,E., Burdett,T., Fullgrabe,A., Fuentes,A.M., Jupp,S., Koskinen,S. *et al.* (2016) Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.
14. Tang,Z., Li,C., Zhang,K., Yang,M. and Hu,X. (2017) GE-mini: a mobile APP for large-scale gene expression visualization. *Bioinformatics*, **33**, 941–943.
15. Li,L., Chaudhuri,A., Chant,J. and Tang,Z. (2007) PADGE: analysis of heterogeneous patterns of differential gene expression. *Physiol. Genomics*, **32**, 154–159.
16. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.W., Varambally,S., Cao,X., Tchinda,J., Kuefer,R. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.
17. Stroud,H., Feng,S., Morey Kinney,S., Pradhan,S. and Jacobsen,S.E. (2011) 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.*, **12**, R54.
18. Zhou,Y., Luoh,S.M., Zhang,Y., Watanabe,C., Wu,T.D., Ostland,M., Wood,W.I. and Zhang,Z. (2003) Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. *Cancer Res.*, **63**, 5781–5784.