Published online 9 October 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D1027–D1030***
doi: 10.1093/nar/gkx892

TC3A: The Cancer 3′ UTR Atlas

Xin Feng[1,†], Lei Li[1,†], Eric J. Wagner[2] and Wei Li[1,*]

[1]Division of Biostatistics, Dan L. Duncan Cancer Center and Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA and [2]Department of Biochemistry & Molecular Biology, University of Texas Medical Branch at Galveston, Galveston, TX 77550, USA

Received August 13, 2017; Revised September 20, 2017; Editorial Decision September 21, 2017; Accepted September 22, 2017

ABSTRACT

Widespread alternative polyadenylation (APA) occurs during enhanced cellular proliferation and transformation. Recently, we demonstrated that CFIm25-mediated 3′ UTR shortening through APA promotes glioblastoma tumor growth *in vitro* and *in vivo*, further underscoring its significance to tumorigenesis. Here, we report The Cancer 3′ UTR Atlas (TC3A), a comprehensive resource of APA usage for 10,537 tumors across 32 cancer types. These APA events represent potentially novel prognostic biomarkers and may uncover novel mechanisms for the regulation of cancer driver genes. TC3A is built on top of the now *de facto* standard cBioPortal. Therefore, the large community of existing cBioPortal users and clinical researchers will find TC3A familiar and immediately usable. TC3A is currently fully functional and freely available at http://tc3a.org.

INTRODUCTION

Alternative polyadenylation (APA) is emerging as a new paradigm of post-transcriptional regulation (1) for >70% of human genes. By changing the position of polyadenylation, APA can generate transcripts with diverse 3′ UTRs that contain distinct *cis*-regulatory elements, such as miRNA binding sites, leading to altered function, stability, localization and translation efficiency of target RNAs. The role of APA in human cancers is only beginning to be appreciated. Both proliferating and transformed cells have been shown to favor shortened 3′ UTRs, potentially leading to the activation of proto-oncogenes through evasion of miRNA-mediated repression. In addition, our recent study identified CFIm25 (2), a master APA regulator, as a glioblastoma (GBM) tumor suppressor gene, further underscoring the importance of APA in cancer development.

Several APA databases exist with varying scopes. The first-generation APA databases, i.e. PolyA_DB2 (3) and PACdb (4), provide a limited amount of APA sites mainly based on the small expressed sequence tag (EST) database. With the rapid progress of next-generation sequencing (NGS), the second-generation APA databases were recently developed leveraging the various partitioned NGS data. For example, APADB (5) and APASdb (6) use specialized NGS protocols for APA detection, e.g. PolyA-seq, while effective are still currently limited to only a few tissue types and diseases. Furthermore, what was not clearly established from these databases is how pervasive and recurrent APA is in large clinical cohorts across distinct cancer types. The largest source of data that could mitigate this limitation, The Cancer Genome Atlas (TCGA), has devoted significant efforts to characterize numerous genomic, epigenomic, and transcriptomic features in thousands of tumors; however, they lack a PolyA-seq platform for APA analysis.

To at least partially fill this knowledge gap, we recently developed a powerful bioinformatics algorithm DaPars (7), one of the first of its kind, for the *de novo* identification of APA events based on localized changes in 3′ UTR RNA-seq read density between tumors and matched normal tissue (normals). DaPars has been shown to accurately detect the majority of APA events in our computational and experimental validations (7). We have applied DaPars to 358 TCGA tumor/normal pairs and revealed 1,346 recurrent APA target genes. Strikingly, the majority of APA events have shorter 3′ UTRs in cancers and tend to be more up-regulated in tumors, supporting the previous cell line model (1) that genes are up-regulated by shortening 3′ UTRs to escape miRNA-mediated repression. However, in other TCGA tumors beyond those 358 we previously reported, the critical target genes subject to APA, remain poorly understood.

Here we present The Cancer 3′ UTR Atlas (TC3A) as a new web resource of APA usage in human cancers. TC3A distinguishes itself in several ways: (i) TC3A, to the best of our knowledge, is the first repository to host a comprehensive compilation of APA events for more than ∼10,537 tumors across 32 cancer types (Supplementary Table S1). (ii) TC3A builds upon a proven user-friendly interface cBioPortal and provides visualizations and analytic functions in addition to data querying and download. TC3A users can easily explore the effects of APA usage using survival analysis and correlation analysis. (iii) TC3A uses standard RNA-

*To whom correspondence should be addressed. Tel: +1 713 798 7854; Fax: +1 713 798 2716; Email: wl1@bcm.edu
†These authors contributed equally to this work as first authors.

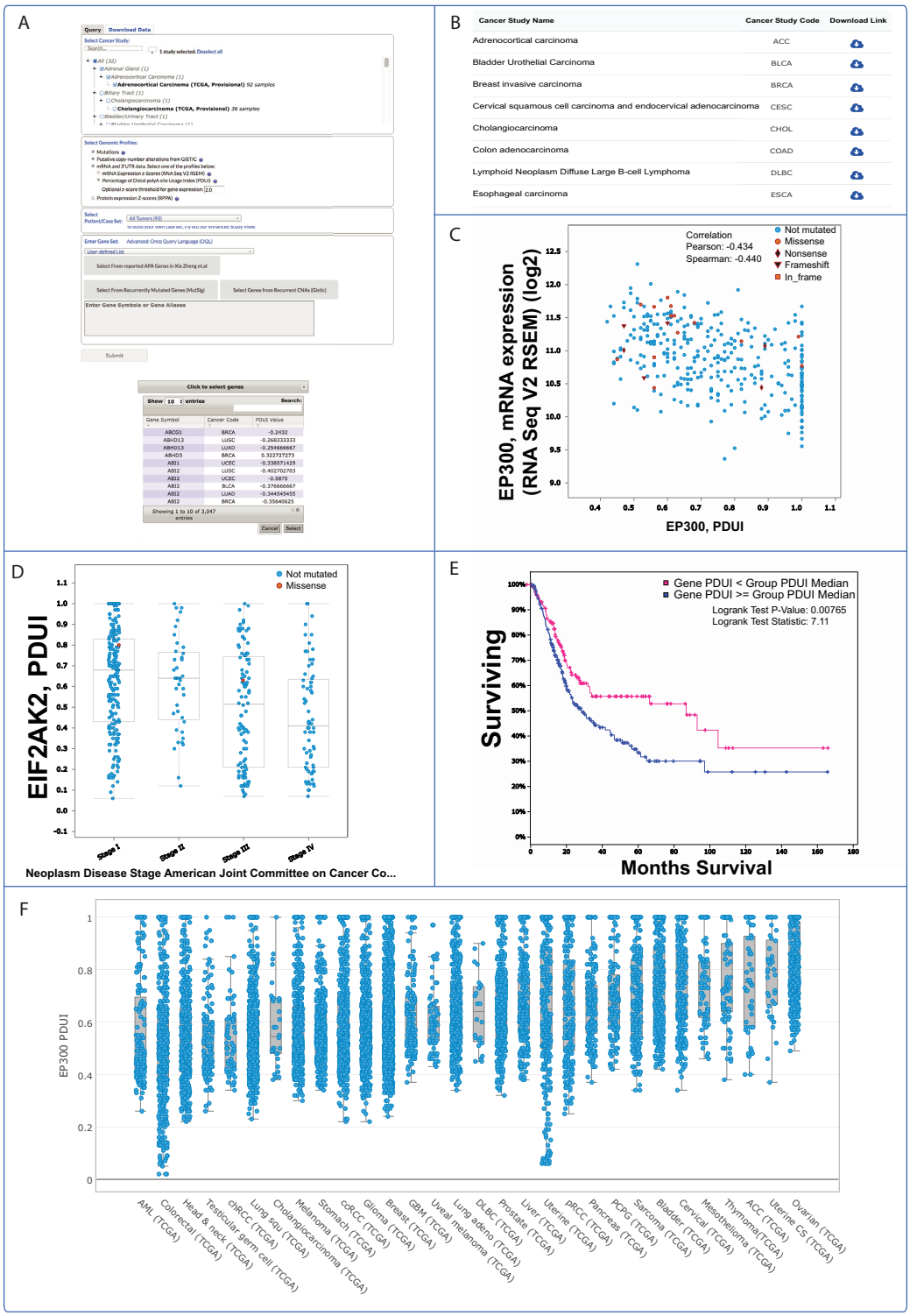

Figure 1. Summary of TC3A. (**A**) TC3A main query interface. An auxiliary table provides all reported APA genes. (**B**) TC3A provides download links to all PDUI values indexed by genes or cancer types. (**C**) TC3A Correlation analysis: the gene EP300's PDUI value negatively correlates with its gene expression value. To reproduce: Choose cancer type Bladder Urothelial Carcinoma (BLCA) and gene EP300. Select the 'Plots' tab. In the 'Horizontal Axis' panel, select 'mRNA' as 'Profile Type' and 'PDUI' as 'Profile Name'. (**D**) Clinical data analysis: plot EIF2AK2's PDUI values against the disease stages. To reproduce: Choose cancer type Kidney Renal Clear Cell Carcinoma (KIRC) and gene EIF2AK2. Select the 'Plots' tab. In the 'Horizontal Axis' panel, select 'Clinical Attribute', then pick 'Neoplasm Disease Stage American Joint Committee on Cancer Code'. In the 'Vertical Axis' panel, select 'mRNA' as 'Profile Type' and 'PDUI' as 'Profile Name'. (**E**) Survival analysis using PDUI values. To reproduce, Choose cancer type Bladder Urothelial Carcinoma (BLCA) and gene EP300. Then select the 'Survival' tab and scroll down to look for the figure titled as 'APA Overall Survival Kaplan-Meier Estimate'. (**F**) The cross-cancer PDUI value distribution plot shows the overview of APA events cross multiple cancers.

Seq data to infer APA *de novo* events, which is a critical component because PolyA sequencing methods (8,9) have not been widely adopted by the cancer community. In contrast, RNA-seq has been extensively used in almost every cancer genomics project including TCGA.

DATA COLLECTION AND DATABASE CONTENT

We downloaded RNA-seq BAM files of 10,537 tumor samples across 32 TCGA cancer types from the UCSC Cancer Genomics Hub (CGHub). The original BAM files were then sorted and converted into bedgraph files using bedtools (10). Multiple bedgraph files within each cancer type were then jointly analyzed together to identify *de novo* APA sites with an updated version of DaPars that can handle multiple tumors samples without relying on normal tissues (https://github.com/3UTR/DaPars2).

For each transcript in tumors, we used a linear regression model to infer the exact location of APA site within 3′ UTR region from multiple tumor RNA-seq data by minimizing the deviation between the observed and the expected read density based on the two-polyA-site model, the most common model of APA regulation:

$$(W_L^{1,2...m*},\ W_S^{1,2,...m*},\ P^*)$$

$$= \arg \min_{W_L^{1,2...m},\ W_S^{1,2...m} \geq 0,\ 1<P<L} \sum_{i=1}^{m} ||C_i - (W_L^i I_L + W_S^i I_p)||_2^2$$

where W_L and W_S are the average abundances of transcripts with distal and proximal polyA sites for sample i, respectively. C_i is the normalized read coverage of sample i. L is the length of the annotated 3′ UTR region. m is the number of jointed samples. P is the estimated length of alternative proximal 3′ UTR. I_L and I_P are indicator. The optimal proximal site P* is the one with the minimal objective function value.

The degree of difference in APA usage in each tumor can be quantified as a change in Percentage of Distal polyA site Usage Index (ΔPDUI), which is capable of identifying 3′ UTR lengthening (positive index) or shortening (negative index).

DATABASE ORGANIZATION AND WEB INTERFACE

We organized APA usage data generated by DaPars into a series of MySQL tables. We adapted the original cBioPortal (11) web interface to our data and developed several analytical functions. cBioPortal is now the *de facto* standard among cancer biologists and clinical researchers who rely on TCGA datasets and has received nearly 2000 citations since 2013. With cBioPortal, researchers can easily access various molecular and clinical profiles from large-scale genomics datasets. These characteristics make cBioPortal a preferred foundation for TC3A.

Integrating APA data with cBioPortal

APA data is mathematically modelled as a two-dimensional matrix, or profile, where each cell represents a gene's PDUI value. The columns represent patient IDs and genes are indexed by rows. For each of the 32 cancer types, one such profile exists in the database. cBioPortal already provides all 32 cancer's genomic datasets. By aligning the patient ID and cancer type with existing cBioPortal datasets, APA data can be queried through various existing cBioPortal interfaces. To import the APA data, we used the dedicated importer module of cBioPortal. These APA data are then retrieved by Javascripts Ajax calls via Java Servlets end points. While doing on-demand analysis and plotting, Javascripts analyze data and call various plotting libraries to present the results.

The query interface

We adopted the original cBioPortal's main query interface (Figure 1A). Users can begin with typing the names of the genes in the box and the system will automatically detect any typos or alias and suggest corrections if necessary. An auxiliary table is provided to include all APA genes reported in the original DaPars paper (7). Genes can be searched and selected from this table. Cancer types are also specified in this interface and the available analysis functions will show up depending on the number of cancer types selected. When users choose multiple cancer types, the cross-cancer distribution analysis is available. If only one cancer type is selected, users can perform survival analysis and correlation analysis.

Data download

Users can freely download PDUI values for genes of interest in the 'Download Data' panel. The full dataset can also be downloaded individually for each cancer type, as shown in Figure 1B.

Correlation analysis

TC3A provides the function to examine the correlation of 3′ UTR usage with other molecular features. In Figure 1C, gene EP300's PDUI is reversely correlated with its mRNA gene expression. It is consistent with recent reports that binding by miR-132 to a site located after the annotated proximal polyA site can significantly reduce EP300's expression levels (12).

For any single gene, the portal also allows users to examine the association of a gene's PDUI values with its clinical data. In Figure 1D, we provide such an example where in kidney renal clear cell carcinoma (KIRC), the gene EIF2AK2's PDUI value is plotted against the neoplasm disease stage. The median PDUI values gradually reduced in later clinical stages.

Survival analysis

TC3A enables researchers to analyze which APA event impacts patient overall survival. Users can generate publication quality Kaplan-Meier plot with log-rank test *P*-value. In Figure 1E, we demonstrate an example using gene EP300, a known APA gene from PolyA_DB2 (3). Using only its PDUI values to separate patient groups, we observed a significant longer survival in bladder cancer patients with 3′ UTR shortening. In comparison, when using only its genomic data, the patients cannot be effectively separated. This is in accordance with our previous report that APA data may better predict patient survival (7).

Cross-cancer distribution analysis

When multiple cancer types are selected, in the 'Expression' tab, a box plot shows the overview of PDUI distribution in each cancer type for the gene of interest. Users can easily discover the cancer types that have dramatic APA changes. In Figure 1F, the cross-cancer distribution plot shows the PDUI values of gene EP300 across different cancer types.

DISCUSSION

We present TC3A as a resource of APA usage for 10,537 tumors across 32 cancer types. This unique resource focuses on human cancers and utilizes routinely available large-scale RNA-Seq datasets from TCGA. Recent reports suggest that APA events are likely to be novel prognostic biomarkers for survival (7,13). It is therefore urgent to add APA usage as an additional dimension to existing cancer genomic analysis. In TC3A, we compiled APA events detected by our updated method DaPars (7). We then systematically integrated APA with existing cancer data from TCGA within the framework of cBioPortal, a preferred database interface by cancer researchers. In this version of TC3A data release, we provide APA data for 10,537 tumors across 32 cancer types. Due to the limitations of DaPars, we used the annotated 3′UTR region as distal polyA site without including the lengthening events, and other APA events occurred outside of 3′UTR regions were not considered such as coding region APA (14). The 3′UTR usage calculated for each transcript was based on two-polyA site model, thus a few genes containing more than two polyA sites may be under-represented. It is expected with the increasing number of RNA-Seq datasets from Genomic Data Common (GDC, https://portal.gdc.cancer.gov/) and other similar consortium projects such as International Cancer Genome Consortium Data (ICGC) (15), TC3A will grow by incorporating newly generated RNA-seq data in the future. In summary, TC3A will enable biologists to explore functional consequences of APA events in human cancers, together with other genomics profiles in cBioPortal.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Jianzhong Su, Hyun Jung Park, Fanglue Peng and Jie Lyu in Li Lab for their helpful discussions.

FUNDING

US National Institutes of Health (NIH) [R01HG007538, R01CA193466]; Cancer Prevention Research Institute of Texas (CPRIT) [RP150292 to W.L., RP140800 to E.J.W.]. Funding for open access charge: NIH [R01CA193466].
Conflict of interest statement. None declared.

REFERENCES

1. Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
2. Masamha,C.P., Xia,Z., Yang,J., Albrecht,T.R., Li,M., Shyu,A.-B., Li,W. and Wagner,E.J. (2014) CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature*, **510**, 412–416.
3. Lee,J.Y., Yeh,I., Park,J.Y. and Tian,B. (2007) PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.*, **35**, D165–D168.
4. Brockman,J.M., Singh,P., Liu,D., Quinlan,S., Salisbury,J. and Graber,J.H. (2005) PACdb: PolyA Cleavage Site and 3′-UTR Database. *Bioinformatics*, **21**, 3691–3693.
5. Müller,S., Rycak,L., Afonso-Grunz,F., Winter,P., Zawada,A.M., Damrath,E., Scheider,J., Schmäh,J., Koch,I., Kahl,G. *et al.* (2014) APADB: a database for alternative polyadenylation and microRNA regulation events. *Database (Oxford)*, **2014**, bau076.
6. You,L., Wu,J., Feng,Y., Fu,Y., Guo,Y., Long,L., Zhang,H., Luan,Y., Tian,P., Chen,L. *et al.* (2015) APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.*, **43**, D59–D67.
7. Xia,Z., Donehower,L.A., Cooper,T.A., Neilson,J.R., Wheeler,D.A., Wagner,E.J. and Li,W. (2014) Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′UTR landscape across seven tumour types. *Nat. Commun.*, **5**, 1–13.
8. Lianoglou,S., Garg,V., Yang,J.L., Leslie,C.S. and Mayr,C. (2013) Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, **27**, 2380–2396.
9. Chang,H., Lim,J., Ha,M. and Kim,V.N. (2014) TAIL-seq: genome-wide determination of poly(A) tail length and 3′ end modifications. *Mol. Cell*, **53**, 1044–1052.
10. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
11. 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.
12. Lagos,D., Pollara,G., Henderson,S., Gratrix,F., Fabani,M., Milne,R.S.B., Gotch,F. and Boshoff,C. (2010) miR-132 regulates antiviral innate immunity through suppression of the p300 transcriptional co-activator. *Nat. Cell Biol.*, **12**, 513–519.
13. Wang,L., Hu,X., Wang,P. and Shao,Z.-M. (2016) The 3′UTR signature defines a highly metastatic subgroup of triple-negative breast cancer. *Oncotarget*, **7**, 59834–59844.
14. Di Giammartino,D.C., Nishida,K. and Manley,J.L. (2011) Mechanisms and consequences of alternative polyadenylation. *Mol. Cell*, **43**, 853–866.
15. Zhang,J., Baran,J., Cros,A., Guberman,J.M., Haider,S., Hsu,J., Liang,Y., Rivkin,E., Wang,J., Whitty,B. *et al.* (2011) International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database (Oxford)*, **2011**, bar026.