SNP2APA: a database for evaluating effects of genetic variants on alternative polyadenylation in human cancers

Yanbo Yang1,†, Qiong Zhang2,†, Ya-Ru Miao2, Jiajun Yang1, Wenqian Yang1, Fangda Yu1, Dongyang Wang1, An-Yuan Guo2,* and Jing Gong1,3,*

1Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, P. R. China, 2Hubei Bioinformatics and Molecular Imaging Key Laboratory, Key Laboratory of Molecular Biophysics of the Ministry of Education, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, P. R. China and 3College of Biomedicine and Health, Huazhong Agricultural University, Wuhan 430070, P. R. China

ABSTRACT

Alternative polyadenylation (APA) is an important post-transcriptional regulation that recognizes different polyadenylation signals (PASs), resulting in transcripts with different 3′ untranslated regions, thereby influencing a series of biological processes and functions. Recent studies have revealed that some single nucleotide polymorphisms (SNPs) could contribute to tumorigenesis and development through dysregulating APA. However, the associations between SNPs and APA in human cancers remain largely unknown. Here, using genotype and APA data of 9082 samples from The Cancer Genome Atlas (TCGA) and The Cancer 3′ UTR Atlas (TC3A), we systematically identified SNPs affecting APA events across 32 cancer types and defined them as APA quantitative trait loci (apaQTLs). As a result, a total of 467 942 cis-apaQTLs and 30 721 trans-apaQTLs were identified. By integrating apaQTLs with survival and genome-wide association studies (GWAS) data, we further identified 2154 apaQTLs associated with patient survival time and 151 342 apaQTLs located in GWAS loci. In addition, we designed an online tool to predict the effects of SNPs on PASs by utilizing PAS motif prediction tool. Finally, we developed SNP2APA, a user-friendly and intuitive database (http://gong.lab.hzau.edu.cn/SNP2APA/) for data browsing, searching, and downloading. SNP2APA will significantly improve our understanding of genetic variants and APA in human cancers.

INTRODUCTION

Alternative polyadenylation (APA) is a widespread phenomenon that generates transcript isoforms with different lengths of 3′ untranslated regions (3′UTR) by recognizing different polyadenylation signals (PASs) (1). More than 70% of human genes have multiple polyadenylation sites (2). As a common post-transcriptional modification mechanism, APA events may cause the alteration of important regulatory elements, such as miRNA binding sites and RNA protein binding sites, thus impacting the stability, localization and translation rate of mRNAs (3). APA modulation has been investigated in cells, tissues and different diseases. Previous studies have shown that APA often functions in a tissue- or cell-specific manner (4,5), and several APA dysregulations have been identified in human diseases (6–9), including cancers (10). A significant global 3′UTR shortening has been found in cancer cell lines and tumor samples, compared with normal samples (11). Another study pointed out that shortening or lengthening of 3′UTR might lead to a worse prognosis in some cancers. For example, kidney cancer samples with the shorter isoforms TMCO7 and PLXDC2 were found to have lower survival rates (12). However, research on the APA role and APA regulation in cancer is still at an early stage.

As the most common genetic variant, single nucleotide polymorphisms (SNPs) are major contributors to the differences in human disease susceptibility (13). Genome-wide association studies (GWAS) have identified thousands of SNPs associated with complex traits and diseases. Currently, most studies of the disease/trait-related SNPs remain...
at statistical level, and the biological mechanism underlying them is still largely unknown (14). Quantitative trait locus (QTL) mapping, such as eQTL and meQTL analysis, is a method used to evaluate the effects of genetic variants on intermediate molecular phenotypes, and has been demonstrated as a powerful tool to decipher the function of SNPs and prioritize genetic variants within GWAS loci (15–19). Recent studies have confirmed the associations between several APA quantitative trait loci (apaQTLs) and cancer. For example, the presence of a SNP in a canonical PAS within TP53 (AATAAA to AATACA) has been found to be highly associated with the processing of the impaired 3′ end of TP53 transcripts and increase the susceptibility to cancers including cutaneous basal cell carcinoma, prostate cancer, glioma and colorectal adenoma (20). However, large-scale genome-wide analyses of apaQTL have rarely been reported, and no database for apaQTLs in cancer is available. Recently, Feng et al. have used Percentage of Distal polyA site Usage Index (PDUI) to quantify APA events for 10,537 tumor samples across TCGA 32 cancer types (21). Therefore, it is feasible to add APA as an additional dimension to the existing cancer genomic analysis.

In this study, by using the genotype and PDUI data, we developed a new computational pipeline to systematically perform apaQTL analyses across 32 cancer types. We further identified apaQTLs associated with patient overall survival time and apaQTLs located in GWAS linkage disequilibrium (LD) regions. The SNP2APA database (http://gong_lab.hzau.edu.cn/SNP2APA/) was constructed for browsing, searching and downloading the apaQTL data.

**MATERIALS AND METHODS**

**Collection and processing of genotype data**

We downloaded the genotype data across 32 cancer types from the TCGA data portal (https://portal.gdc.cancer.gov/) (22), which contained 898,620 SNPs called by Affymetrix SNP 6.0 array. We extracted 9082 samples with both genotype data and APA data available (Figure 1A). To increase the power for apaQTL discovery, IMPUTE2 was used to impute autosomal variants of all samples in each cancer type. The top five principal components in genotype data were included as covariates for apaQTL analysis (25). We first used the smartpca in the EIGENSTRAT program (27) to perform principal component analysis (PCA) of the genotype data for each cancer type. The top five principal components in genotype data were included as covariates for correcting the ethnicity differences. We additionally used PEER software (28) to analyse the APA data and obtained the first 15 PEER factors as covariates which were used for eliminating the possible batch effects and other confounders. Finally, other common confounders such as gender, age and tumor stage (25,29,30), were also included as covariates for apaQTL analysis.

**Identification of cis- and trans-apaQTL using MatrixEQTL**

For each cancer type, we evaluated pairwise associations between autosomal SNPs and APA events through linear regression by MatrixEQTL (31), a software for efficient QTL analysis. The SNP locations (hg19) were downloaded from dbSNP database (https://www.ncbi.nlm.nih.gov/projects/SNP) and distal PAS locations were extracted from the APA datasets. The SNPs with false discovery rates (FDRs) <0.05 calculated by MatrixEQTL and the absolute value of correlation coefficient (r) ≥0.3 were defined as apaQTLs (Figure 1C). Of them, we further defined the apaQTLs within 1 Mb from the distal PAS as the cis-apaQTLs (25), while defined the apaQTLs beyond that region or on another chromosome as the trans-apaQTLs.

**Identification of survival-associated apaQTLs**

To prioritize promising apaQTLs, we further examined the association between apaQTLs and patient survival time. The clinical data including survival time of patient were downloaded from TCGA data portal. For each apaQTL, the samples were divided into three groups by genotypes: homozygous genotype (AA), heterozygous genotype (Aa), and homozygous genotype (aa). Then the log-rank test was performed to examine the differences in survival time, and Kaplan–Meier (KM) curves were plotted for intuitive visualization of the survival time for each group. Finally, apaQTLs with FDR <0.05 were designated as survival-associated apaQTLs.

**Identification of GWAS-associated apaQTLs**

GWAS has been successfully used for identifying thousands of disease susceptibility loci, but it remains a challenge to...
pinpoint the causal variants and decipher their underlying mechanisms. To facilitate the interpretation of GWAS results, we integrated apaQTLs with the existing GWAS risk loci to explore trait/disease-associated apaQTLs. We downloaded all the risk tag SNPs identified in GWAS studies from GWAS catalog (http://www.ebi.ac.uk/gwas, accessed September 2018) (32). Then the SNPs in linkage disequilibrium (LD) regions with GWAS tag SNPs were extracted from SNAP (https://personal.broadinstitute.org/plin/snap/ldsearch.php) (33). The parameters were set as follows: (i) SNP dataset: 1000 Genomes, (ii) $r^2$ (the square of the Pearson correlation coefficient of LD) threshold: 0.5, (iii) population panel: CEU (Utah residents with northern and western European ancestry), (iv) distance limit: 500 kb. Finally, we defined apaQTLs that overlapped with these GWAS tag SNPs and LD SNPs as GWAS-associated apaQTLs.

**DATABASE CONSTRUCTION AND CONTENT**

All results mentioned above were stored into MongoDB database (version 3.4.2) in the form of relation tables. A user-friendly web interface, SNP2APA (http://gonglab.hzau.edu.cn/SNP2APA/), was constructed to support data browsing, searching, downloading and PAS online prediction (Figure 1D and E), based on Flask (version 1.0.3) framework with Angularjs (version 1.6.1) as the JavaScript library. It was running on Apache2 web server (version 2.4.18). We have tested SNP2APA on various web browsers, including Chrome (recommended), Firefox, Opera, Internet Explorer, Windows Edge and Safari of macOS.

**Data summary of SNP2APA**

In total, SNP2APA included 9082 tumor samples across 32 cancer types with both genotype data and APA data available for apaQTL analysis. The sample sizes for each cancer type ranged from 36 in cholangiocarcinoma (CHOL) to 1,091 in invasive breast carcinoma (BRCA) with a median of 221 (Table 1). After genotype imputation and quality control, 4 390 660 SNPs on average per each cancer type were included for further analysis, ranging from 2 746 335 for BRCA to 5 143 663 for acute myeloid leukemia (LAML). After filtering APA events by both the rate of missing PDUI value >0.1 and PDUI standard deviation >0.05, we obtained an average of 4143 APA events per can-
SNP2APA primarily provided four kinds of datasets: cis- and trans-apaQTLs, survival apaQTLs and GWAS-associated apaQTLs (Figure 2A and B). In the cis-apaQTL analysis, a total of 467,942 cis-apaQTLs across 32 cancer types were identified at the level of FDR < 0.05 and |r| ≥ 0.3, with a median of 14,811 apaQTLs per cancer type, minimum of 1,580 in lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), and maximum of 34,381 in glioblastoma multiforme (GBM). In the trans-apaQTL analysis, a total of 30,721 trans-apaQTLs across 32 cancer types were identified at P-value < 1 × 10^{-8} and |r| ≥ 0.3, with a median of 936 apaQTLs per cancer type, minimum of nine in thyroid carcinoma (THCA), and maximum of 2,171 in DLBC.

Survival and GWAS associated apaQTLs

To prioritize promising apaQTLs, we associated apaQTLs with the survival data of patients downloaded from the TCGA portal. A total of 2,154 apaQTLs associated with overall survival time across 32 cancer types at FDR < 0.05, were identified and included in SNP2APA. For example, we found that rs10247994 was highly associated with patient overall survival time in kidney renal clear cell carcinoma (KIRC) (Figure 2C). The significant differences in PDUI values among corresponding genotypes of rs10247994 were observed, indicating that this SNP might play an important role in regulating the APA event of *PUSH* gene in KIRC (Figure 2C).

We further mapped apaQTL results to SNPs in GWAS regions and identified a total of 151,342 apaQTLs overlapping with GWAS LD regions with one or multiple traits. For example, rs2303282, as a risk SNP, was reported to be associated with BRCA (34). In our study, we found that rs370151 was in LD with the rs2303282 (LD r^2 = 0.87) and was highly associated with APA event of *AMFR* gene. *AMFR* was reported to encode a tumor motor stimulating protein receptor (35). Thus, it could be inferred that rs370151 might play an important role in breast cancer by affecting APA events (Figure 2D).

CONCLUSION AND FUTURE DIRECTIONS

We developed SNP2APA as a resource providing comprehensive apaQTLs across 32 cancer types. To the best of our knowledge, this is the first database systematically evaluating the effects of the genetic variants on APA, especially in multiple cancer types with a large sample size. In recent years, increasing studies have suggested that APA is likely to play important roles in cancer. Therefore, it is urgent to add APA as an additional dimension to existing cancer genomic analysis. In this version of TC3A, by using genotype and APA data of 9,082 tumor samples, we provided numerous apaQTLs among multiple cancer types and identified abundant apaQTLs associated with patient survival time or located in known GWAS loci. To explore the impact of SNPs on PAS, we also designed an online tool for users to predict functional apaQTLs. The SNP2APA database will greatly facilitate the interpretation of risk SNPs identified in genetic studies. In the future, with the increasing number of RNA-Seq datasets and genotype data from large consortium projects, we will continue to update the SNP2APA database. We believe that our database will be of particular interest to researchers in the field of genetic variants and APA in cancer.
Figure 2. The interface of SNP2APA database. (A) Browser bar in SNP2APA. (B) Modules of cis- and trans-apaQTL, survival apaQTL and GWAS apaQTL. (C) An example of survival apaQTL. KM plot indicated that rs10247994 in KIRC was highly associated with patient survival time, and box plot indicated that rs10247994 in KIRC was highly associated with PDUI values of the APA event in PUSH gene. (D) An example of GWAS apaQTL. Box plot indicated that GWAS associated apaQTL rs370151 in BRCA was highly associated with PDUI values of the APA event in AMFR. (E) Search results of cis-apaQTL dataset. (F) The heatmap displaying the correlation coefficient (r) of apaQTLs in the ‘Pancan-apaQTL’ page. The label for y-axis contains SNP ID, gene symbol of APA and APA event. (G) The input of online PAS prediction tool.
FUNDING

National Natural Science Foundation of China (NSFC) [31970644 to J.G., 31822030 and 31771458 to A.Y.G.].

Conflict of interest statement. None declared.

REFERENCES

1. Gruber, A.J., Schmidt, R., Gruber, A.R., Martin, G., Ghosh, S., Belmadani, M., Keller, W. and Zavolan, M. (2016) A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.*, 26, 1145–1159.

2. Tian, B. and Manley, J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, 18, 18–30.

3. Elkon, R., Ugalde, A.P. and Agami, R. (2013) Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.*, 14, 496–506.

4. MacDonald, C.C. (2019) Tissue-specific mechanisms of alternative polyadenylation: Testis, brain, and beyond (2018 update). *Wires RNA*, 10, e1526.

5. Di Giammartino, D.C., Nishida, K. and Manley, J.L. (2011) Mechanisms and consequences of alternative polyadenylation. *Mol. Cell.*, 43, 853–866.

6. Chang, J.W., Yeh, H.S. and Yong, J. (2017) Alternative polyadenylation in human diseases. *Endocrinol. Metab. (Seoul)*, 32, 413–421.

7. Bacchetta, R., Barzaghi, F. and Roncarolo, M.G. (2018) From IPEX syndrome to FOXP1 mutation: a lesson on immune dysregulation. *Am. N. Y. Acad. Sci.*, 1417, 5–22.

8. Bennett, C.L., Brunkow, M.E., Ramsdell, F., O'Briant, K.C., Zhu, Q., Fuleihan, R.L., Shigeoka, A.O., Ochs, H.D. and Chance, P.F. (2001) A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA → AAUGAA) leads to the IPEX syndrome. *Immunogenetics*, 53, 435–439.

9. Garin, I., Edghill, E.L., Akerman, I., Rubio-Cabezas, O., Rica, I., Locke, J.M., Maestro, M.A., Alshaikh, A., Bundak, R., del Castillo, G. et al. (2010) Recessive mutations in the INS gene result in neonatal diabetes through reduced insulin biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.*, 107, 3105–3110.

10. Erson-Bensan, A.E. and Can, T. (2016) Alternative polyadenylation: another foe in cancer. *Mol. Cancer Res.*, 14, 507–517.

11. Xiang, Y., Ye, Y., Lou, Y., Yang, Y., Cai, Z., Zhang, Z., Mills, T., Chen, N.Y. and Cheng, N. (2018) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, 50, 384–394.

12. 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, 5274.

13. Shastry, B.S. (2002) SNP alleles in human disease and evolution. *J. Hum. Genet.*, 47, 561–566.

14. Do, C., Shearer, A., Suzuki, M., Terry, M.B., Gelernter, J., Greally, J.M. and Tycko, B. (2017) Genetic-epigenetic interactions in cis: a major focus in the post-GWAS era. *Genome Biol.*, 18, 120.

15. Westra, H.J., Peters, M.J., Esko, T., Yaghooikar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E. et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, 45, 1238–1243.
16. Xiong,H.Y., Alipanahi,B., Lec,L.I., Bretschneider,H., Merico,D., Yuen,R.K.C., Hua,Y.M., Gueroussov,S., Najafabadi,H.S., Hughes,T.R. et al. (2015) The human splicing code reveals new insights into the genetic determinants of disease. Science, 347, 1254806.
17. Takata,A., Matsumoto,N. and Kato,T. (2017) Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. Nat. Commun., 8, 14519.
18. Gong,J., Mei,S.F., Liu,C.J., Xiang,Y., Ye,Y.Q., Zhang,Z., Feng,J., Liu,R.Y., Diao,L.X., Guo,A.Y. et al. (2018) PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res., 2020, Vol.48, Database issue.
20. Stacey,S.N., Sulem,P., Jonasdottir,A., Masson,G., Gudmundsson,J., Feng,X., Li,L., Wagner,E.J. and Li,W. (2018) TC3A: the cancer 3 Nucleic Acids Research, 2020, Vol.48, Database issue.
21. Cancer Genome Atlas, N. (2012) Comprehensive molecular portraits of human breast tumours. Nature, 490, 61–70.
22. Price,A.L., Patterson,N.J., Plenge,R.M., Weinblatt,M.E., Shadick,N.A. and Reich,D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet., 38, 904–909.
23. Stegle,O., Parts,L., Piipari,M., Winn,J. and Durbin,R. (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc., 7, 500–507.
24. Schulz,H., Ruppert,A.K., Herms,S., Wolf,C., Mirza-Schreiber,N., Stegle,O., Czamara,D., Forstner,A.J., Sivalingam,S., Schoch,S. et al. (2017) Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. Nat. Commun., 8, 1511.
25. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A., Morales,J. et al. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res., 45, D896–D901.
26. Johnson,A.D., Handsaker,R.E., Pulit,S.L., Nizzari,M.M., O’Donnell,C.J. and de Bakker,P.I.W. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics, 24, 2938–2939.
27. Michailidis,K., Lindstrom,S., Dennis,J., Beesley,J., Hui,S., Kar,S., Lemacon,A., Soucy,P., Glubb,D., Rostamianfar,A. et al. (2017) Association analysis identifies 65 new breast cancer risk loci. Nature, 551, 92–94.
28. Jiang,W.G., Raz,A., Douglas-Jones,A. and Mansel,R.E. (2006) Expression of autocrine motility factor (AMF) and its receptor, AMFR, in human breast cancer. J. Histochem. Cytochem., 54, 231–241.
29. Kalkatawi,M., Rangkuti,F., Schramm,M., Jankovic,B.R., Kamau,A., Chowdhary,R., Archer,J.A.C. and Bajic,V.B. (2012) Dragon PolyA Spotter: predictor of poly(A) motifs within human genomic DNA sequences. Bioinformatics, 28, 127–129.