D1046–D1052 *Nucleic Acids Research, 2023, Vol. 51, Database issue* *Published online 31 August 2022*
https://doi.org/10.1093/nar/gkac736

ipaQTL-atlas: an atlas of intronic polyadenylation quantitative trait loci across human tissues

Xuelian Ma[1,†], Shumin Cheng[1,†], Ruofan Ding[1], Zhaozhao Zhao[2], XuDong Zou[1], Shouhong Guang[3], Qixuan Wang[1], Huan Jing[4], Chen Yu[5], Ting Ni[2] and Lei Li[1,*]

[1]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518055, China, [2]State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, Human Phenome Institute, School of Life Sciences and Huashan Hospital, Fudan University, Shanghai 200438, China, [3]Ministry of Education Key Laboratory for Membraneless Organelles and Cellular Dynamics, School of Life Sciences, Department of Obstetrics and Gynecology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, Biomedical Sciences and Health Laboratory of Anhui Province, University of Science and Technology of China, Hefei, Anhui 230027, P.R. China, [4]Department of Stomatology, Peking University Shenzhen Hospital, Shenzhen 518036, China and [5]Institute of Cancer Research, Shenzhen Bay Laboratory, Shenzhen 518055, China

Received July 14, 2022; Revised August 03, 2022; Editorial Decision August 12, 2022; Accepted August 18, 2022

ABSTRACT

Functional interpretation of disease-associated non-coding variants remains a significant challenge in the post-GWAS era. Our recent study has identified 3′UTR alternative polyadenylation (APA) quantitative trait loci (3′aQTLs) and connects APA events with QTLs as a major driver of human traits and diseases. Besides 3′UTR, APA events can also occur in intron regions, and increasing evidence has connected intronic polyadenylation with disease risk. However, systematic investigation of the roles of intronic polyadenylation in human diseases remained challenging due to the lack of a comprehensive database across a variety of human tissues. Here, we developed ipaQTL-atlas (http://bioinfo.szbl.ac.cn/ipaQTL) as the first comprehensive portal for intronic polyadenylation. The ipaQTL-atlas is based on the analysis of 15 170 RNA-seq data from 838 individuals across 49 Genotype-Tissue Expression (GTEx v8) tissues and contains ∼0.98 million SNPs associated with intronic APA events. It provides an interface for ipaQTLs search, genome browser, boxplots, and data download, as well as the visualization of GWAS and ipaQTL colocalization results. ipaQTL-atlas provides a one-stop portal to access intronic polyadenylation information and could significantly advance the discovery of APA-associated disease susceptibility genes.

INTRODUCTION

Genome-wide association studies (GWAS) have identified thousands of genomic non-coding variants statistically associated with many human complex traits and diseases. However, it remains a significant challenge to understand the molecular mechanism of how these non-coding genetic variants contribute to a physiological/pathological phenotype in humans. The characterization of molecular Quantitative Trait Loci, such as gene expression QTL (eQTL) (1), splicing QTL (sQTL) (1) or allele-specific QTL (aseQTL) (2), is a key step towards better understanding the effects of non-coding genetic variants on genes, pathways, and their function mechanism and serve as an essential link between genotype and disease phenotype. However, although these molecular QTLs have successfully explained numerous GWAS risk loci, a significant fraction of GWAS risk loci remains unexplained (1).

Alternative polyadenylation (APA), which occurs in >70% of human genes, plays an essential role in the post-transcriptional regulation (3,4). By employing different poly(A) sites (PASs), genes can either shorten or extend the 3′UTRs that contain multiple cis-regulatory elements such as miRNAs or RNA-binding protein (RBP) binding sites (5). APA can affect the stability, translation efficiency and cellular localization of mRNAs and proteins (3), which significantly impacts both normal development and progression of diseases such as cancer (6). Besides 3′UTR, APA also occurs in intronic regions in approximately 20% of human genes. In contrast to 3'UTR APAs, the use of intronic PASs generates mRNA isoforms that encode truncated proteins or generate noncoding RNAs (7). These truncated proteins could be dysfunctional, which result in a variety

*To whom correspondence should be addressed. Tel: + 0755 2684 9284; Email: lei.li@szbl.ac.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

of diseases such as leukemia, multiple myeloma and non-Alcoholic fatty liver (8,9). Regulators such as *CDK12* (10) can suppress intronic polyadenylation events and regulate full-length DNA repair gene. Meanwhile, the use of intronic APA is tightly regulated by variants near splicing sites. The somatic variants could interfere with splice site recognition and increase the use of intronic PASs of tumor suppressor genes (11).

Recently, we constructed the first atlas of human 3′UTR APA Quantitative Trait Loci (3′aQTLs): i.e. ∼0.4 million genetic variants associated with the APA of target genes across 46 Genotype-Tissue Expression (GTEx) healthy tissues from 467 individuals (12). These 3′aQTLs represent the genetic basis of APA that can be used to interpret ∼16.1% of trait-associated non-coding variants, and are largely distinct from other molecular QTLs such as eQTLs or sQTLs. It has also been evidenced that genetic variations can influence intronic APA usage. For example, Mittleman *et al*. applied 3′seq to identify the genetic basis of APA in 52 HapMap Yoruba human lymphoblastoid cell lines and identified a subset of intronic APA QTLs (ipaQTLs) (13). However, these studies are limited to certain cell lines. To further explore to what extent these genetic variations affect intronic APA usage in a wide variety of human tissues, a comprehensive database dedicated to intronic polyadenylation is in urgent need.

To meet this goal, we developed ipaQTL-atlas, the first database for intronic polyadenylation based on 15 170 RNA-seq samples across 49 human Genotype-Tissue Expression (GTEx) project (version 8) tissues from 838 individuals. ipaQTL-atlas includes visual analysis tools of Gene/SNP searches, genome browsers, boxplots, colocalization and traits/diseases. This database aims to facilitate studies analyzing and validating the roles of intronic polyadenylation as important aspects to explain disease-associated risk SNPs.

MATERIAL AND METHODS

Data curation and processing

We downloaded the 17 382 RNA-sequencing samples across 52 tissues and two cell lines in 948 post-mortem donors from the Genotypic Tissue Expression (GTEx) project (dbGaP, phs000424.v8.p2) (1). After removing tissue types with the threshold of the sample size <70 and BAM files without genotype data, 15 170 sets of RNA-seq across 49 tissues were retained. Meanwhile, the low-quality samples from whole-genome sequencing data from the GTEx v8 release were filtered, and the variants were further imputed and phased using SHAPEIT v2 (14) and the subset VCF files using bcftools (15). The relevant sample description file is downloaded from the GTEx portal website (www.gtexportal.org).

Quantification of intronic polyadenylation usage using IPAFinder

The recently developed IPAFinder (7,11) can *de novo* identify and quantify dynamic intronic polyadenylation events using standard RNA-seq without any prior poly(A) site annotation. We first generated an APA annotation file based on the hg38 genome version using the script 'IPAFinder_GetAnno.py'. Then, we used 'IPAFinder_DetectIPA.py' to *de novo* identify the IPA sites of multiple RNA-seq samples and calculate the intronic poly(A) usage in each sample.

Covariate correction

We performed sample genotype correction to account for hidden batch effects and other unobserved covariates in each tissue. We used the R package PEER (16) to estimate potential covariates for IPUI values in each tissue with sex, genotyping platform, and top 5 genotype principal components as known covariates. The number of PEER factors was chosen according to the recommendations of the GTEx Consortium (17), with 15, 30 and 35 PEER factors selected for tissue sample sizes < 150, 150–250 and > 250, respectively.

ipaQTLs mapping

Based on the available IPUI values of genes in each tissue from above analyses, we removed genes with >50% entries missing and individuals with >80% missing data after quantification by IPAFinder, then used the R package impute tool (18) to impute missing IPUI values and normalize in each tissue. The normalized IPUI values were used as Matrix eQTL input. SNPs with a minor allele frequency of <0.01 were filtered. Furthermore, combined with covariates files, SNP and gene position files, we performed ipaQTLs mapping for each tissue separately using Matrix eQTL with linear regression model (19) to obtain the association between genetic variants and APA target genes in the 1-Mb interval of the intronic region.

Identification of GWAS-associated ipaQTLs

To identify trait-associated ipaQTLs, we obtained GWAS significant SNPs from the NHGRI GWAS catalog (20) and removed the SNPs without dbSNP addition. For each tissue, we extract the lead ipaQTLs per target gene. Based on dbSNP annotations, we extracted a list of SNPs in strong LD of CEU (Utah Residents with Northern and Western European Ancestry) with GWAS catalog labeling SNPs. Finally, for each tissue, we used a cut-off with the FDR < 0.05 and extract the lead ipaQTLs per target gene. ipaQTLs were considered to overlap with disease-associated loci if the lead ipaQTLs or its LD tag ($r^2 \geq 0.8$) mapped with a GWAS SNP.

Construction of the ipaQTL-atlas website

The ipaQTL-atlas website was constructed on a standard LAMP (Linux + Apache + MySQL + PHP) system. All processed and annotated data in the ipaQTL-atlas were stored in MySQL (www.mysql.com). The R package LocusComparer (21) and in-house R scripts were used to perform data analyses and visualization. The interactive web pages were implemented using HTML, CSS, JavaScript, and PHP languages (www.php.net), with several JavaScript

libraries (JQuery.js, DataTable.js, and IGV.js) and Bootstrap framework (a popular framework for developing interactive websites) on Red Hat Linux powered by an Apache server (www.apache.org). The ipaQTL-atlas is freely available online without registration or login requirements for access, and optimized for Chrome (recommended).

RESULTS

Atlas of ipaQTLs across 49 human tissues

We analyzed 15 170 RNA-seq samples across 49 human tissues from GTEx version 8 (Figure 1A, B). After quantifying by IPAFinder, we obtained 4826 IpA events for each tissue type, ranging from 1678 for Brain Amygdala to 8514 for Breast Mammary Tissue. Through ipaQTL mapping, a total of 0.98 million common genetic variants associated with ipaQTLs were identified; the median was 58 317 per tissue type (Figure 1C). The number of IpA events was significantly correlated with the sample size in each tissue (Spearman correlation, $R = 0.89$, P-value = 2.2e-16, Supplementary Figure S1). The strong correlation between the number of IpA events and sample size suggests that more ipaQTLs will continue to be identified as additional RNA-seq datasets become available. In addition, in this version of the ipaQTLs atlas, we further divide IpA into compound terminal exon IpA (or composite IpA) and skip terminal exon IpA (or skipped IpA), and provide ipaQTL name, allele (Ref/Alt), IPA site, sample size and other helpful information.

To investigate the global distribution of ipaQTLs in the human genome, we used Manhattan plots to visualize the locations of ipaQTLs and their associated P-values. Mittleman *et al*. (13) previously quantified the APA usage across 52 Yoruba HapMap lymphoblastoid cell lines (LCLs) samples and mapped 225 ipaQTLs at 10% false discovery rate (FDR). By comparing with our ipaQTLs from Cells EBV-transformed lymphocyte cell lines, we found that previously reported IpA genes were successfully detected such as *PMS2P1*, *MRPL45P2*, *HVCN1*, etc. The strong association between SNP rs3873746 and the IpA of *PMS2P1* was replicated in our analysis. *PMS2P1* is a pseudogene involved in mismatch repair and plays a critical regulatory role in hemoglobin measurement and late-onset Alzheimer's disease (22–25). Our further analysis revealed that this genetic effect is shared in 28 other tissues (Supplementary Figure S2), suggesting that multi-tissue contextual analysis of this locus could aid further studies on how *PMS2P1* variants contribute to complex diseases. Indeed, most of the detected ipaQTLs genes represent novel events. Several new ipaQTLs genes are of particular note, including *ZNF880*, a Zinc Finger Protein involved in the transcriptional regulation of RNA polymerase II associated with childhood acute lymphoblastic leukemia and multiple cancer associations (26–28).

Web design and interface

We developed a user-friendly database ipaQTL-atlas (http://bioinfo.szbl.ac.cn/ipaQTL) for searching, browsing, visualizing and downloading 0.9 million common genetic variants associated with ipaQTLs across 49 human tissues. The homepage provides the summary and description of the database. The ipaQTL-atlas can annotate ipaQTLs for two types: composite and skipped. For two types of ipaQTLs, the database provides 'Gene/SNP Search' to search the gene name or SNP rs ID across one specific selected tissue type or all 49 human tissues (Figure 2A). Besides, the ipaQTL genome browser (Figure 2B), ipaQTL boxplot (Figure 2C), and GWAS-ipaQTL colocalization event visualization (Figure 2D) are also available in our database. In addition, a list of GWAS-associated ipaQTLs is also provided for users to further investigate the mechanisms of ipaQTLs in human traits and diseases (Figure 2E).

ipaQTLs searching and querying

Our database provides a convenient search interface for querying gene(s) and SNP rs ID by commonly used keywords such as ipaQTL ID, gene symbol/alias, and ipaQTL type. Suppose users did not choose the particular tissue, the default tissue is 'Adipose Subcutaneous'. If they want to query in all tissues, the option 'All Tissues' need to be selected. Notably, the sample size of the selected tissue has been given in parentheses, together with the tissue type. Similarly, the number of all samples, regardless of tissue type, is written with 'All Tissues'. In that case, you will get a table with the tissue types, ipaQTLs ID, gene symbol, SNP rs ID, Ref allele (the allele in the reference genome), Alt allele (any other allele found at that locus), and P-value of each ipaQTL item for the queried gene or SNP with the default tissue of adipose subcutaneous (Figure 2A). For example, the query of *SPG7* will obtain 2811 significant ipaQTL items in default tissue adipose subcutaneous. Users can further filter these ipaQTLs by inputting custom filter relevant keywords (e.g. rs4785691) into the 'Search' field at the top-right corner of the table. For each ipaQTL item, we also provide the 'Browser' button with the genome browser hyperlink based on the related gene ID. On the other hand, the 'Boxplot' button is paralleled to allow users to visualize the ipaQTLs in high-resolution boxplot figures. In the end, it is possible for users to get no data available in the table with the query of *SPG7* in the tissue of whole blood because there is no predicted ipaQTLs of *SPG7* in the tissue of whole blood.

Data browsing by genome browser

In the 'Genome Browser' section, users can explore the ipaQTLs across human tissues in an interactive genome browser using the gene symbol (e.g. *NFKB1*), SNP rs ID (e.g. rs4785690) or genome position (e.g. chr15:40155795–41028618). Taking 'IVD' as an example in the genome browser, we can search this locus in the search box and find all significant ipaQTLs for *IVD* (red points) in adipose subcutaneous tissue (Figure 2B). Clicking the dot of interest will show details of the SNP, including rs ID and P-value. Only ipaQTLs of the queried gene are labeled in red in the genome browser, whereas ipaQTLs of other genes are marked in grey. The genome browser also exhibits gene structure annotation, GWAS catalog risk SNPs (20), and PolyA_DB3 polyA sites (29) tracks, which allow users to integrate these data with ipaQTLs. In addition, users can

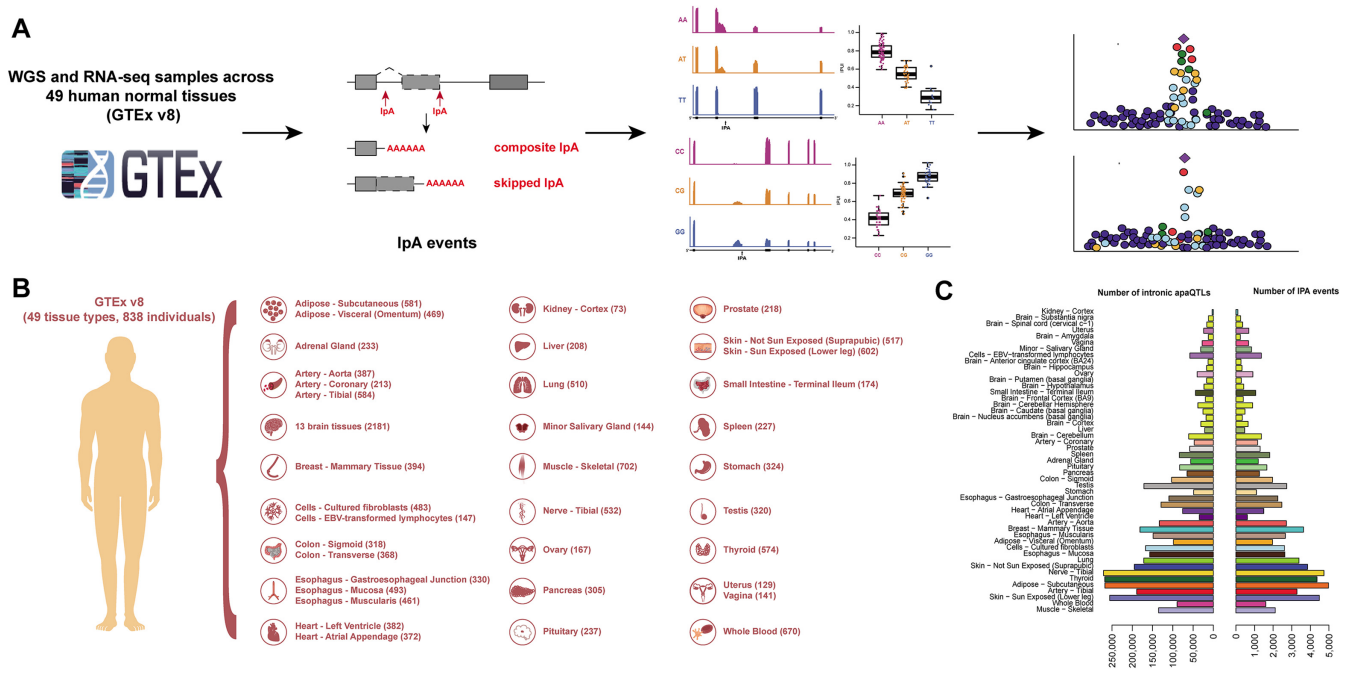

Figure 1. Data processing and data statistics in ipaQTL-atlas. (**A**) Workflows of ipaQTL-atlas. IPA with the arrow denotes the position of the corresponding intronic polyadenylation site. WGS, whole-genome sequencing; ipaQTLs, intronic polyadenylation quantitative trait loci. (**B**) The number of RNA-seq samples for each tissue used in the ipaQTL-atlas was distributed. (**C**) The distribution of the number of IpA events and significant ipaQTLs SNPs (FDR ≤ 0.05) for each tissue was sorted by the tissue sample sizes. Each color indicates a tissue of origin.

download the figures of the browser tracks in SVG format by clicking on the 'Save SVG' button at the top-right corner of the genome browser.

Visualizing tools

The 'ipaQTLs Boxplots' section was allowed to draw plots through the button from the result page of 'Genome Browser' and 'Traits/Diseases'. Of course, it also can be accessed directly through the navigation bar at the top of each page to customize boxplots for every ipaQTL as a flexible online tool. Instruction step by step is also described on the 'Help' page. In the case of the ipaQTL ID (e.g. ARPC2|intron_12|chr2:218238850–218238948), users can draw the boxplot by providing rs ID (e.g. rs13392177) and tissue name (e.g. Brain Hippocampus) with the default or user-defined colors (Figure 2C), which can be visualized and downloaded concurrently as a publishable PDF document. This tool could facilitate a deep understanding and direct sight for the different normalized IPUI values in the different genotypes caused by ipaQTLs.

Based on the widely used R package LocusCompare (21), we provide an online server in the 'GWAS-ipaQTLs colocalization event visualization' part, which helps users conveniently visualize GWAS-ipaQTLs colocalization events using their GWAS data. For instance, by inputting the ipaQTL ID (e.g. PMS2P3|intron_3|chr7:75510862–75510930), tissue name (e.g. Colon Sigmoid), and a two-column text with the rs ID and corresponding *P*-value, the plot of the GWAS-ipaQTLs colocalization event is shown at the intron region of the PMS2P3 locus, which can be downloaded as a PDF file (Figure 2D).

Traits/diseases correlation

To link ipaQTL variants to human genetic traits and diseases, we also supply a list of GWAS-associated ipaQTLs, defined when the lead ipaQTL variants are overlapped with the GWAS catalog (20) tag SNPs. Like the 'Gene/SNP' search page, we also show a table on the 'Traits/diseases' page with the optional tissue type. By default, users can view the full information of 2177 traits/diseases from GWAS Catalog across all 49 tissues. We also provide the 'Boxplot' button for each ipaQTL item to allow users to visualize the ipaQTL item in the boxplot figures. Here, we displayed an example of ipaQTL (e.g. PEX6|intron_21|chr6:42969477–42969667) with Alzheimer's disease in the brain frontal cortex BA9 tissue (Figure 2E). First, you need to select the tissue brain frontal cortex BA9. Second, the keyword 'Alzheimer' need to be input into the filter box at the top-right corner of the table. That example allows researchers to investigate the mechanisms of ipaQTLs in human traits and diseases.

Downloading data

The download page enables users to download all the results of ipaQTLs across 49 human tissues and a table of all human trait- and disease-associated ipaQTLs for further custom analysis. On the download page, the default drop-down box on the top table header is usually the setting of all tissues shown, so that most users would probably look through the descriptive information, including tissue type, data source, sample size and file name of the download file by single tissue. The options of tissue types have been provided for users to select the spe-

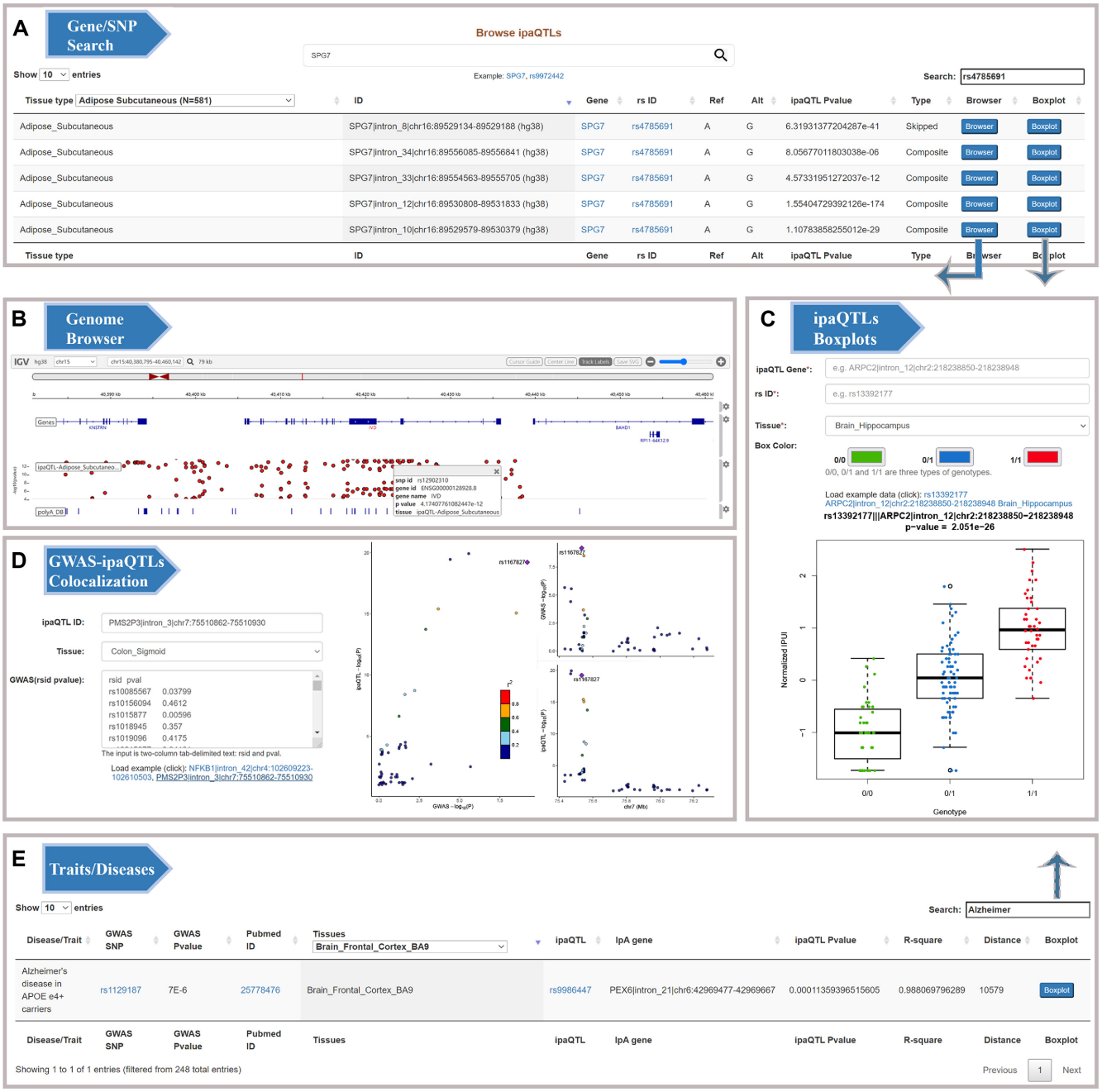

Figure 2. The web interface of ipaQTL-atlas. (**A**) ipaQTLs query interface and result visualization for 'Gene/SNP Search'. (**B**) An example of the genome browser view shows the ipaQTLs of adipose subcutaneous tissue at the *IVD* locus. (**C**) The interface of 'ipaQTLs Boxplots' and an example of the ipaQTLs boxplot for *ARPC2* and rs13392177 in brain hippocampus. (**D**) The interface of the 'GWAS-ipaQTLs Colocalization' and an example of the LocusCompare plot at the intronic region of the *PMS2P3* locus with T2D GWAS *P*-values and ipaQTLs *P*-values in colon sigmoid tissue. (**E**) The interface of the 'Traits/Diseases' and an example of the ipaQTLs with Alzheimer's disease in brain frontal cortex BA9 tissue.

cific tissue to download. Furthermore, users can gain more details about how to download through the link of the 'Guide' button to the help page. In the 'Gene/SNP' search page, users can also download a formatted text file for the queried gene name or SNP rs ID across all tissues, a table with the same columns as the search results on the web page.

SUMMARY AND FUTURE DIRECTIONS

Although several molecular QTL resources derived from GTEx human normal tissues (30) have been developed to explore genetic effects, there is no resource for intronic polyadenylation. As a result, we developed a powerful and user-friendly database, ipaQTL-atlas, providing comprehensive ipaQTLs across 49 human tissue types to

explain a significant fraction of GWAS risk SNPs. As the first database systematically identifying ipaQTLs and disease-related ipaQTLs, ipaQTL-atlas provides millions of ipaQTLs information for users to query, browse and download. Furthermore, we designed high-efficiency online tools, including boxplot drawer, GWAS-ipaQTLs colocalization visualization and ipaQTLs genome browser, to significantly facilitate the functional interpretation of SNPs related to intronic polyadenylation. In the future, updating the ipaQTL-atlas database is extremely necessary because of the increasing number of RNA-seq datasets and genotype data from large consortium projects. Accordingly, we will continue to maintain and update our database with the release of GTEx and other projects. The more various ipaQTLs will be identified in different tissue and cell types. We believe our database will be of particular interest to researchers in genetic variants and APA. The ipaQTLs from more normal tissue types of more individuals, and even from cancer tissue types will be added to our database for follow-up updates. In conclusion, ipaQTL-atlas will become an important resource for the genetic research community and provides essential supplements to interpret the function of non-coding disease risk variants in shaping human phenotypic diversity.

DATA AVAILABILITY

The data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.v8.p2.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank members of the Li laboratory for helpful discussions. We also thank Shenzhen Bay Laboratory supercomputing center for high-computing support.

FUNDING

National Natural Science Foundation of China [32100533 to L.L.], and Shenzhen Bay Laboratory Open Fund (SZBL2021080601001 to L.L.). Funding for open access charge: National Natural Science Foundation of China [32100533].
Conflict of interest statement. None declared.

REFERENCES

1. Aguet,F., Anand,S., Ardlie,K.G., Gabriel,S., Getz,G.A., Graubert,A., Hadley,K., Handsaker,R.E., Huang,K.H., Kashin,S. *et al.* (2020) The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science (New York, N.Y.)*, **369**, 1318–1330.
2. He,L., Loika,Y. and Kulminski,A.M. (2022) Allele-specific analysis reveals exon- and cell-type-specific regulatory effects of alzheimer's disease-associated genetic variants. *Transl. Psychiatry*, **12**, 163.
3. Tian,B. and Manley,J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, **18**, 18–30.
4. Hong,W., Ruan,H., Zhang,Z., Ye,Y., Liu,Y., Li,S., Jing,Y., Zhang,H., Diao,L., Liang,H. *et al.* (2020) APAatlas: decoding alternative polyadenylation across human tissues. *Nucleic Acids Res.*, **48**, D34–D39.
5. Mayr,C. (2017) Regulation by 3′-Untranslated regions. *Ann. Rev. Genet.*, **51**, 171–194.
6. Yang,S.W., Li,L., Connelly,J.P., Porter,S.N., Kodali,K., Gan,H., Park,J.M., Tacer,K.F., Tillman,H., Peng,J. *et al.* (2020) A cancer-specific ubiquitin ligase drives mRNA alternative polyadenylation by ubiquitinating the mRNA 3′ end processing complex. *Mol. Cell*, **77**, 1206–1221.
7. Zhao,Z., Xu,Q., Wei,R., Wang,W., Ding,D., Yang,Y., Yao,J., Zhang,L., Hu,Y.Q., Wei,G. *et al.* (2021) Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. *Genome Res.*, **31**, 2095–2106.
8. Tian,B., Pan,Z. and Lee,J.Y. (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.*, **17**, 156–165.
9. Singh,I., Lee,S.H., Sperling,A.S., Samur,M.K., Tai,Y.T., Fulciniti,M., Munshi,N.C., Mayr,C. and Leslie,C.S. (2018) Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat. Commun.*, **9**, 1716.
10. Dubbury,S.J., Boutz,P.L. and Sharp,P.A. (2018) CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature*, **564**, 141–145.
11. Zhao,Z., Xu,Q., Wei,R., Huang,L., Wang,W., Wei,G. and Ni,T. (2021) Comprehensive characterization of somatic variants associated with intronic polyadenylation in human cancers. *Nucleic Acids Res.*, **49**, 10369–10381.
12. Mazin,P.V., Khaitovich,P., Cardoso-Moreira,M. and Kaessmann,H. (2021) Alternative splicing during mammalian organ development. *Nat. Genet.*, **53**, 925–934.
13. Mittleman,B.E., Pott,S., Warland,S., Zeng,T., Mu,Z., Kaur,M., Gilad,Y. and Li,Y. (2020) Alternative polyadenylation mediates genetic regulation of gene expression. *Elife*, **9**, e57492.
14. Delaneau,O., Zagury,J.-F. and Marchini,J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods*, **10**, 5–6.
15. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A. and Davies,R.M. (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.
16. 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.
17. Consortium,GTEx (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
18. Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
19. Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.
20. 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.
21. Liu,B., Gloudemans,M.J., Rao,A.S., Ingelsson,E. and Montgomery,S.B. (2019) Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.*, **51**, 768–769.
22. Jansen,I.E., Savage,J.E., Watanabe,K., Bryois,J., Williams,D.M., Steinberg,S., Sealock,J., Karlsson,I.K., Hägg,S., Athanasiu,L. *et al.* (2019) Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nat. Genet.*, **51**, 404–413.
23. Wightman,D.P., Jansen,I.E., Savage,J.E., Shadrin,A.A., Bahrami,S., Holland,D., Rongve,A., Børte,S., Winsvold,B.S., Drange,O.K. *et al.* (2021) A genome-wide association study with 1,126,563 individuals identifies new risk loci for alzheimer's disease. *Nat. Genet.*, **53**, 1276–1282.
24. Oskarsson,G.R., Oddsson,A., Magnusson,M.K., Kristjansson,R.P., Halldorsson,G.H., Ferkingstad,E., Zink,F., Helgadottir,A.,

Ivarsdottir,E.V., Arnadottir,G.A. *et al.* (2020) Predicted loss and gain of function mutations in ACO1 are associated with erythropoiesis. *Commun. Biol.*, **3**, 189.
25. Astle,W.J., Elding,H., Jiang,T., Allen,D., Ruklisa,D., Mann,A.L., Mead,D., Bouman,H., Riveros-Mckay,F., Kostadima,M.A. *et al.* (2016) The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, **167**, 1415–1429.
26. Dong,C., Cesarano,A., Bombaci,G., Reiter,J.L., Yu,C.Y., Wang,Y., Jiang,Z., Zaid,M.A., Huang,K., Lu,X. *et al.* (2021) Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. *Oncogene*, **40**, 6130–6138.
27. Yue,L., Wentao,L., Xin,Z., Jingjing,H., Xiaoyan,Z., Na,F., Tonghui,M. and Dalin,L. (2020) Human epidermal growth factor receptor 2-positive metastatic breast cancer with novel epidermal growth factor receptor -ZNF880 fusion and epidermal growth factor receptor E114K mutations effectively treated with pyrotinib: a case report. *Medicine*, **99**, e23406.
28. Gawad,C., Koh,W. and Quake,S.R. (2014) Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl Acad. Sci. USA*, **111**, 17947–17952.
29. Wang,R., Nambiar,R., Zheng,D. and Tian,B. (2018) PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res.*, **46**, D315–D319.
30. Cui,Y., Peng,F., Wang,D., Li,Y., Li,J.S., Li,L. and Li,W. (2022) 3′aQTL-atlas: an atlas of 3′UTR alternative polyadenylation quantitative trait loci across human normal tissues. *Nucleic Acids Res.*, **50**, D39–D45.