Sun et al. BMC Plant Biology (2020) 20:119
https://doi.org/10.1186/s12870-020-2312-y

BMC Plant Biology

DATABASE Open Access

CuAS: a database of annotated transcripts generated by alternative splicing in cucumbers

Ying Sun[1†], Quanbao Zhang[1†], Bing Liu[1], Kui Lin[1], Zhonghua Zhang[2] and Erli Pang[1*]

Abstract

Background: Alternative splicing (AS) plays a critical regulatory role in modulating transcriptome and proteome diversity. In particular, it increases the functional diversity of proteins. Recent genome-wide analysis of AS using RNA-Seq has revealed that AS is highly pervasive in plants. Furthermore, it has been suggested that most AS events are subject to tissue-specific regulation.

Description: To reveal the functional characteristics induced by AS and tissue-specific splicing events, a database for exploring these characteristics is needed, especially in plants. To address these goals, we constructed a database of annotated transcripts generated by alternative splicing in cucumbers (CuAS: http://cmb.bnu.edu.cn/alt_iso/index.php) that integrates genomic annotations, isoform-level functions, isoform-level features, and tissue-specific AS events among multiple tissues. CuAS supports a retrieval system that identifies unique IDs (gene ID, isoform ID, UniProt ID, and gene name), chromosomal positions, and gene families, and a browser for visualization of each gene.

Conclusion: We believe that CuAS could be helpful for revealing the novel functional characteristics induced by AS and tissue-specific AS events in cucumbers. CuAS is freely available at http://cmb.bnu.edu.cn/alt_iso/index.php.

Keywords: Cucumber, Alternative splicing, Isoform-level function, Isoform-level features, Tissue-specific alternative splicing events

Background

Alternative splicing (AS) is an important post-transcriptional process by which multiple transcripts are generated from a single gene. It plays critical roles in adaption to the environment, development, and tissue specificity [1–4]. Additionally, it increases the functional diversity of proteins [2].

Since the first discovery of AS 40 years ago [5], an increasing number of alternatively spliced genes have been reported. With the development of sequencing technology, it has been found that AS is apparently highly pervasive in eukaryotes. Recently, based on RNA-Seq data, 95% of human genes [6] and 61% of *Arabidopsis* genes [7] were reported to undergo AS. In addition, the functions of AS have been investigated. Emerging experimental evidence indicates that AS can regulate the following properties of proteins: 1) binding to other proteins and nucleic acids [8], 2) the localization of proteins according to localization signals [9], 3) enzymatic properties [10], and 4) interactions with ligands [11]. Overall, AS can influence almost every aspect of protein functions [2].

Several AS databases such as ASpedia [12], VastDB [13], and DBATE [14] have been established, but these databases are for vertebrates, especially humans, and few of them address AS in plants. In plants exposed to environmental stress, many biological processes are regulated

* Correspondence: pangerli@bnu.edu.cn
†Ying Sun and Quanbao Zhang contributed equally to this work.
[1]MOE Key Laboratory for Biodiversity Science and Ecological Engineering and Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, No 19 Xinjiekouwai Street, Beijing 100875, China
Full list of author information is available at the end of the article

by alternative splicing [3]. With the development of sequencing technologies, the detection of AS in plants is coming of age [15]. Therefore, a database for the annotation of AS events and a retrieval system to query AS and explore the functions of alternatively spliced transcripts in plants is needed.

Here, we introduce a database of annotated transcripts generated by AS in cucumbers (CuAS) (*Cucumis sativus* L. var. *sativus* cv. 9930 and *Cucumis sativus* var. *hardwickii* PI 183967). The database provides five types of data: (1) genomic annotation, (2) AS events analysed from multiple tissues, (3) isoform features, (4) isoform functions, (5) and splicing events among tissues. The web application includes four components: an annotation database, a retrieval system, a browser, and tools. This user-friendly database will serve as a hub for revealing the functional characteristics induced by AS and tissue-specific AS events in cucumbers.

Construction and content

The CuAS database integrates genomic annotation, AS events from multiple tissues, isoform functions, isoform features, and tissue-specific splicing events. The integration steps are shown in Fig. 1.

Data sources

CuAS includes data from two varieties of cucumber: *Cucumis sativus* L. var. *sativus* cv. 9930 and *Cucumis sativus* var. *hardwickii* PI 183967. The genome sequences and genome annotations were collected from http://cmb.bnu.edu.cn/Cucumis_sativus_v20/. The RNA-Seq data of ten tissues from *Cucumis sativus* L. var. *sativus* cv. 9930 were downloaded from the SRA database (https://www.ncbi.nlm.nih.gov/sra/) (SRA: SRA046916), and the RNA-Seq data of seven tissues from *Cucumis sativus* var. *hardwickii* PI 183967 were obtained from the website http://cmb.bnu.edu.cn/Cucumis_sativus_v20/. The seven tissues included the roots, stems, leaves, male flowers, female flowers, fruit, and tendrils.

Identification of alternative splicing events and isoforms

In previous research based on RNA-Seq of ten tissues from *Cucumis sativus* L. var. *sativus* cv. 9930, we assembled transcripts by using TopHat and Cufflinks [16], respectively. These sets of transcripts were then compared with the reference genome annotation file using Cuffcompare. The transcripts were divided into 12 categories according to the output of Cuffcompare. Then, the following strategies were applied to obtain high-quality transcripts [17, 18]. First, all of the transcripts with three class codes (=, j, o) (http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/) were extracted from the output generated by Cuffcompare. The transcripts in the "j" and "o" classes were considered novel transcripts. Next, the novel transcripts with a single exon were removed, and we obtained an assembled cucumber transcriptome. To reduce potentially misassembled transcripts, each novel splice junction was required to be supported by at least ten reads, and each known splice junction was required to be supported by at least one read. According to these criteria, transcripts supported by certain splice junction reads were obtained. Finally, transcripts per million

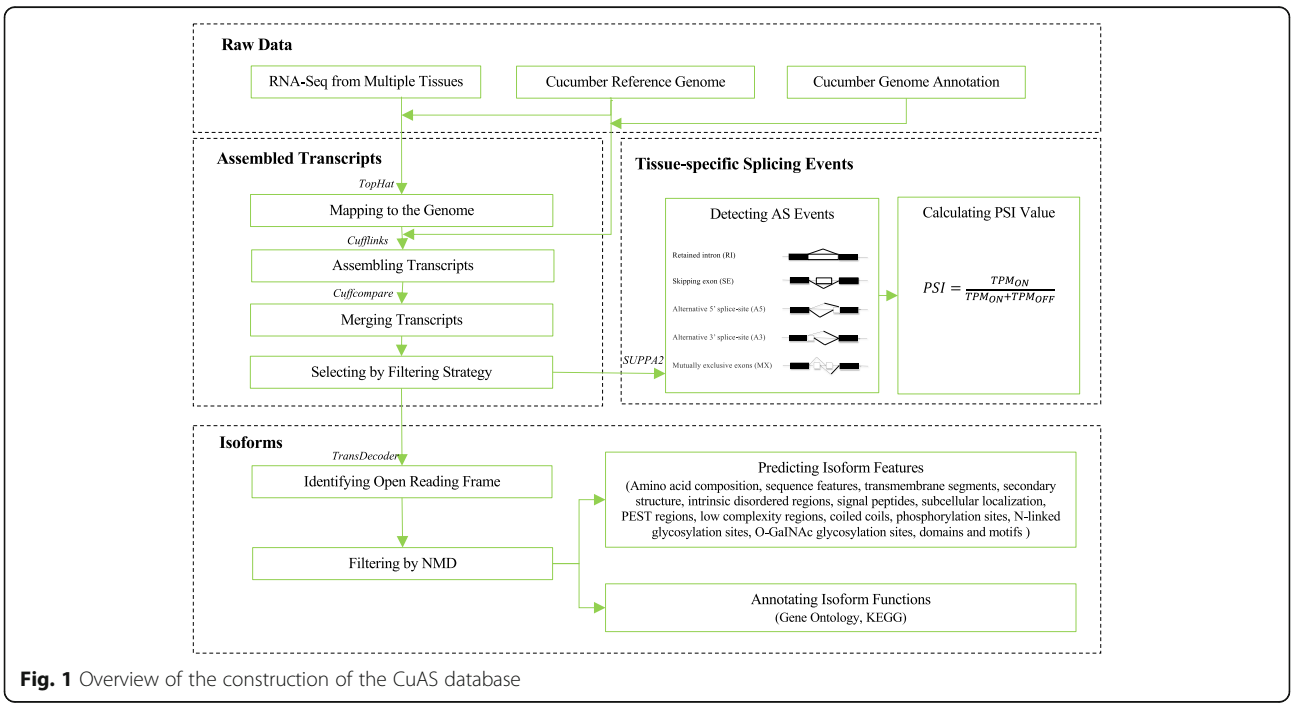

Fig. 1 Overview of the construction of the CuAS database

reads (TPM) values were calculated by using Salmon (version 0.13.0) [19], and the transcripts with TPM values of greater than or equal to one in at least one sample were used for the analysis [20]. With the implementation of a series of filters, a high-quality putative transcriptome was obtained. Based on the obtained transcripts, AS events were identified by using SUPPA2 (version 2.3) [21]. The AS events were classified into five types: retained intron (RI), skipped exon (SE), alternative 3′ splice-sites (A3), alternative 5′ splice-sites (A5), and mutually exclusive exons (MX).

To better understand the impact of differentially spliced isoforms encoded by a single gene, we used TransDecoder (https://github.com/TransDecoder/TransDecoder, version 3.0.1) to identify the candidate coding regions in the assembled transcripts. TransDecoder performs homology searches against Pfam 30.0 [22] and the UniProt database (version 2016_11) [23] to obtain supporting evidence for the open reading frames (ORFs). We selected the single best ORF for each transcript using the parameter "-single_best_orf". If a premature termination codon was located more than 55 nucleotides from the last splice junction, the transcript was considered to be a result of nonsense-mediated mRNA decay (NMD) [24–26]. Any transcript with an ORF that is greater than or equal to 300 bp in length that did not show NMD was retained for further analysis. The same software and parameters were used for *Cucumis sativus* var. *hardwickii* PI 183967.

Functional annotation at the isoform level

First, we performed a Blast2GO [27] analysis that assigned gene ontology terms to each isoform. Blast2GO performed a BLASTP search (E-value 1e-05) against the UniProt (release 2017_06) database. Then, the identified isoforms were mapped to reference canonical pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/, version 90.1) [28]. KAAS (KEGG Automatic Annotation Server, https://www.genome.jp/tools/kaas/) was used to assign KEGG pathways.

Prediction of features at the isoform level

The software used for the prediction of isoform features is listed in Table 1. In total, 15 types of features were predicted, including the amino acid composition, sequence features, transmembrane segments, secondary structure, regions of intrinsic disorder, signal peptides, subcellular localization, PEST regions, low-complexity regions, coiled coils, phosphorylation sites, N-linked glycosylation sites, O-GaINAc glycosylation sites, domains, and motifs.

The transmembrane segments, secondary structure, and regions of intrinsic disorder were searched against the UniRef90 dataset (release 2016_01). Domains and motifs were assigned using InterProScan 5.24 [29].

Table 1 Software used for isoform feature prediction

Feature Group	Software	Reference
Amino acid composition	EMBOSS-6.6.0	[39]
Sequence features	EMBOSS-6.6.0	[39]
Gravy	GRAVY calculator	(no warranty)
Transmembrane segments	MEMSAT 3.0	[40]
Secondary structure	PSIPRED 4.0	[41]
Intrinsically disordered regions	DISOPRED 3.16	[42]
Signal peptides	SingIP 4.0	[43]
Subcellular localization	YLoc	[44]
PEST regions	EMBOSS-6.6.0	[39]
Low complexity regions	EMBOSS-6.6.0	[39]
Coiled coils	EMBOSS-6.6.0	[39]
Phosphorylation sites	NetPhos-3.1	[45]
N-linked glycosylation sites	NetNGlyc-1.0c	[45]
O-GalNAc-glycosylation sites	NetOglyc-3.1d	[45]
Domains (Pfam)	InterProScan 5.24	[29]
Motifs (Prosite)	InterProScan 5.24	[29]

Tissue-specific splicing events

To investigate tissue-specific splicing events, the percent spliced-in index (PSI), which is a representative AS event measurement, was quantified for all AS events. The PSI measures the fraction of the mRNAs expressed from a gene that contains a specific form resulting from an AS event [30]. The reads were used to quantify transcript abundances with Salmon [19], and the PSI values [31] among tissues were calculated by SUPPA2 for all AS events.

Prediction of gene descriptions and gene families

The functional description of the genes was provided by the AHRD tool (https://github.com/groupschoof/AHRD) based on the results of BLASTP searches against UniProt and TAIR. In regard to gene families, transcription factors (TFs), transcriptional regulators (TRs), and protein kinases (PKs) were identified by iTAK (version 1.7) [32]. While splicing-related genes were identified by OrthoFinder (version: 2.3.1) [33] against the sequences of *Arabidopsis* [34], including small nuclear ribonucleoproteins, splicing factors, splicing regulation-related proteins, novel spliceosome proteins, and possible splicing-related proteins.

Web implementation

The web interface is implemented with PHP programming, HTML, and JavaScript. All the graphs are generated through the plug-in ECharts [35]. All the tables are in the style of Layui (https://www.layui.com/). Poshy Tip

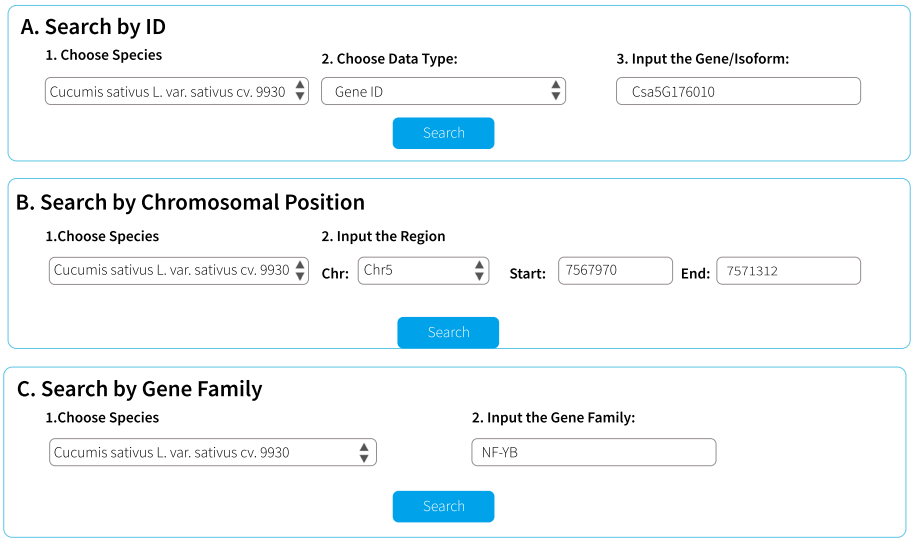

Fig. 2 Search tools of CuAS

(https://github.com/vadikom/poshytip) is applied to show the position of amino acids.

Utility and discussion

The CuAS system contains four components: an annotation database, a retrieval system, a browser, and tools (BLAST and JBrowse).

Database overview

In total, a set of 60,643 transcripts (36,274 from *Cucumis sativus* L. var. *sativus* cv. 9930 and 24,369 from *Cucumis sativus* var. *hardwickii* PI 183967) was obtained. Based on these transcripts, 10,748 AS events (6673 from *Cucumis sativus* L. var. *sativus* cv. 9930 and 4075 from *Cucumis sativus* var. *hardwickii* PI 183967) were predicted, and 49,018 isoforms (28,588 from *Cucumis sativus* L.

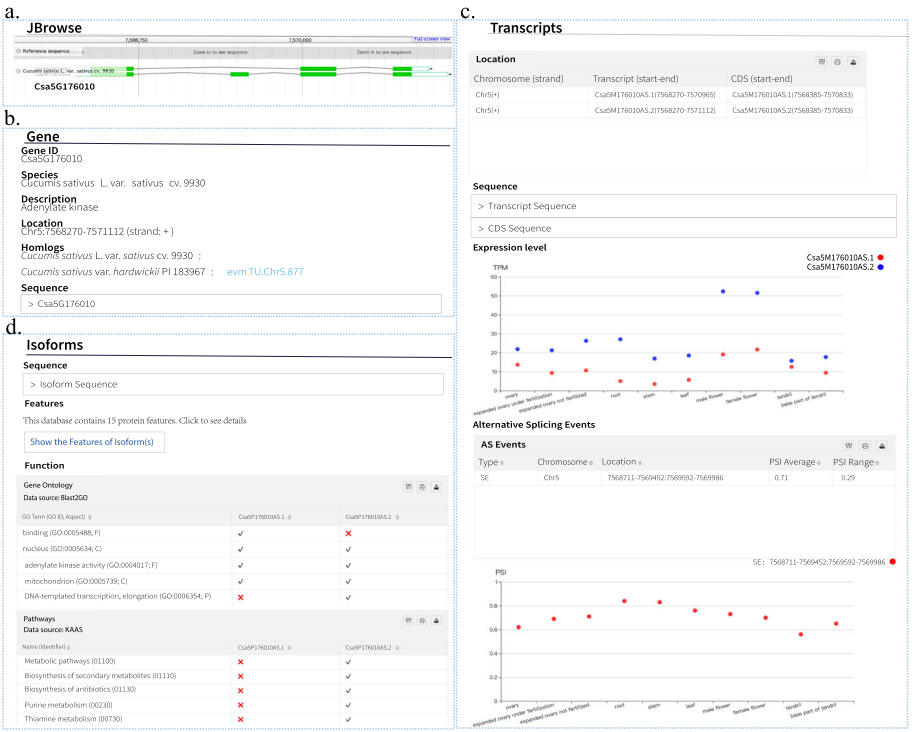

Fig. 3 Web interface of CuAS. (**a**) JBrowse. (**b**) At the gene level. (**c**) At the transcript level. (**d**) At the isoform level

var. *sativus* cv. 9930 and 20,430 from *Cucumis sativus* var. *hardwickii* PI 183967) were retained for the analysis of features and functions. Isoform functions were annotated with Gene Ontology [36] and KEGG [28] terms. Regarding isoform features, 15 types of features were predicted. In addition, the PSI values were quantified for all AS events (see Construction and Content).

Web interface

The CuAS web-interface provides access to genomic annotation, functional annotation at the isoform level, features at the isoform level, and tissue-specific AS events. The data can be queried using three input formats: ID (gene ID/isoform ID/UniProt ID/gene name), chromosomal position, and gene family (Fig. 2, e.g., *Csa5G176010*). These input data can be used to search AS events among tissues and their relevant annotations.

Search results are categorized and visualized on the results page, as illustrated in Fig. 3 by using the example of *Csa5G176010*. The structures of the two transcripts encoded by *Csa5G176010* are displayed by JBrowse (Fig. 3a). The results are organized at three levels, the gene, transcript, and isoform levels. At the gene level, we list the basic information of the gene and its homologs in the two cucumbers (Fig. 3b). At the transcript level, the transcript expression abundance, predicted AS events, and PSI values of these events are reported for each query gene among tissues. This is also illustrated in Fig. 3c, in which a SE event is detected for *Csa5G176010*. The two transcripts are expressed in all the tissues. At the isoform level, the isoform functional annotations (GO annotation and KEGG pathway annotation) and

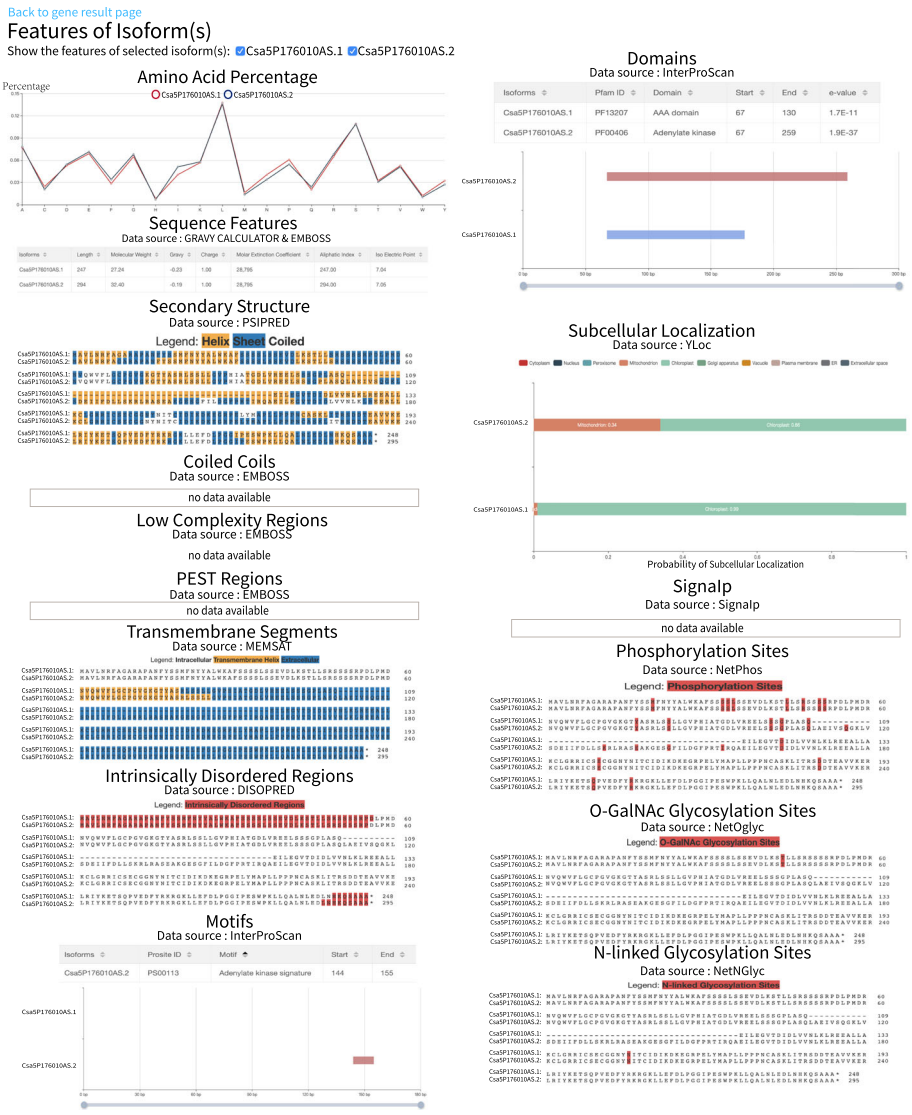

Fig. 4 Isoform features of CuAS

features of the gene isoforms are provided. As shown in Fig. 3d, the two isoforms of *Csa5G176010* present some different functions, such as "binding" and "AMP salvage".

The features of the alternative isoforms can be retained by clicking "Show the Features of Isoform(s)" (Fig. 3d). The list of features is shown on the isoform feature page (Fig. 4), including the amino acid composition, sequence features, transmembrane segments, secondary structure, regions of intrinsic disorder, signal peptides, subcellular localization, PEST regions, low-complexity regions, coiled coils, phosphorylation sites, N-linked glycosylation sites, O-GaINAc glycosylation sites, domains, and motifs (see Construction and Content). As shown in Fig. 4 using *Csa5G176010* as an example, Csa5P176010AS.1 includes the "Adenylate kinase signature" motif, but Csa5P176010AS.2 does not include the motif. In addition, there are different functional characteristics between the two transcripts. These results suggest that the SE event detected in *Csa5G176010* has an influence on the function of isoforms.

In addition, two tools are provided: BLAST and JBrowse. BLAST is used to find the homologous sequences of cucumbers. Users can paste their DNA or protein query sequences in the "Query Sequence" box. Users can set search parameters such as the search databases, search programs, maximum number of hits, and E-values. Users can choose the search database by selecting "Searching Against". Eight BLAST databases including genes, transcripts, CDSs, and isoforms from the two cucumbers were generated for BLAST searches. The search program (BLASTN, TBLASTX, BLASTX, TBLASTN, or BLASTP) can be chosen by selecting "Program", according to the query sequence and the search database. "Advanced options" can be used to set the maximum number of hits and E-values. JBrowse was applied to visualize the genomic features of cucumbers, including transcripts from multiple cucumber tissues.

Our database offers HTTP links to download the genome sequence, transcript sequences, putative CDSs, and protein sequences in FASTA format. The gene structure annotations can be obtained in the GFF3 format. The list of IDs mapping to UniProt can be obtained. AS events and PSI values can also be downloaded. The list of data files, including isoform features as well as isoform functions, is also accessible in text format. The detailed user manual is available on the CuAS website.

Conclusions

The advent of RNA-Seq has driven the rapid expansion of transcriptomics. This adds the gap between functional characteristics and transcripts, which is a critical step when trying to understand how diversity may arise from AS. CuAS provides a resource for exploring the relationships between functional features and AS transcripts predicted from multiple tissues in cucumbers, and tissue-specific AS events can be obtained from PSI values. CuAS will help reveal the novel functional features induced by AS and tissue-specific AS events in plants.

CuAS is an ongoing project, and we plan to further develop it in the next release. In particular, we are going to add variation annotations for AS sites and explore the relationship between variation and AS. We also plan to include data related to other organisms, such as *Cucumis melo* L. [37] and *Citrullus lanatus* [38], which will be helpful for achieving a better understanding of AS through comparative analyses in Cucurbitaceae.

Abbreviations

A3: alternative 3′ splice-sites; A5: alternative 5′ splice-sites; AS: alternative splicing; CuAS: a database of annotated transcripts generated by alternative splicing in cucumbers; MX: mutually exclusive exons; NMD: nonsense-mediated mRNA decay; ORF: open reading frame; PKs: protein kinases; PSI: percent spliced-in index; RI: retained intron; SE: skipped exon; TFs: transcription factors; TPM: transcripts per million reads; TRs: transcriptional regulators

Acknowledgements

We are thankful to the editors and anonymous reviewers for insightful feedback on the manuscript.

Authors' contributions

ELP and YS conceived and designed the analyses. YS, QBZ and BL participated in the database and implemented the user interface. ELP, YS and KL drafted the manuscript. ZHZ provided essential suggestions for this work. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 31571361). The funding agency had no role in the design of the study and collection, analysis and interpretation of data or in writing the manuscript.

Availability of data and materials

CuAS is freely available at http://cmb.bnu.edu.cn/alt_iso/index.php. The dataset can be downloaded from http://cmb.bnu.edu.cn/alt_iso/index.php/download. The detailed user manual is available at http://cmb.bnu.edu.cn/alt_iso/index.php/help. The website is optimized for Internet Explorer, Mozilla Firefox, Google Chrome, and Safari.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]MOE Key Laboratory for Biodiversity Science and Ecological Engineering and Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, No 19 Xinjiekouwai Street, Beijing 100875, China. [2]Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Ministry of Agriculture, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China.

Received: 11 June 2019 Accepted: 26 February 2020
Published online: 18 March 2020

References

1. Lopez AJ. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Annu Rev Genet. 1998;32:279–305.
2. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. Function of alternative splicing. Gene. 2013;514(1):1–30.
3. Staiger D, Brown JW. Alternative splicing at the intersection of biological timing, development, and stress responses. Plant Cell. 2013;25(10):3640–56.
4. Lee Y, Rio DC. Mechanisms and regulation of alternative pre-mRNA splicing. Annu Rev Biochem. 2015;84:291–323.
5. Berget SM, Moore C, Sharp PA. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A. 1977;74(8):3171–5.
6. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40(12):1413–5.
7. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome Res. 2012;22(6):1184–95.
8. Belfiore A, Frasca F, Pandini G, Sciacca L, Vigneri R. Insulin receptor isoforms and insulin receptor/insulin-like growth factor receptor hybrids in physiology and disease. Endocr Rev. 2009;30(6):586–623.
9. Chu HY, Ohtoshi A. Cloning and functional analysis of hypothalamic homeobox gene Bsx1a and its isoform, Bsx1b. Mol Cell Biol. 2007;27(10):3743–9.
10. Bertolesi GE, Michaiel G, McFarlane S. Two heparanase splicing variants with distinct properties are necessary in early *Xenopus* development. J Biol Chem. 2008;283(23):16004–16.
11. Ko J, Fuccillo MV, Malenka RC, Sudhof TC. LRRTM2 functions as a neurexin ligand in promoting excitatory synapse formation. Neuron. 2009;64(6):791–8.
12. Hyung D, Kim J, Cho SY, Park C. ASpedia: a comprehensive encyclopedia of human alternative splicing. Nucleic Acids Res. 2018;46(Database issue):D58–63.
13. Tapial J, Kch H, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, Quesnel-Vallières M, Permanyer J, Sodaei R, Marquez Y. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. 2017;27(10):1759–68.
14. Bianchi V, Colantoni A, Calderone A, Ausiello G, Ferrè F, Helmercitterich M. DBATE: database of alternative transcripts expression. Database. 2013;2013(6):bat050.
15. Syed NH, Kalyna M, Marquez Y, Barta A, Brown JW. Alternative splicing in plants--coming of age. Trends Plant Sci. 2012;17(10):616–23.
16. Sun Y, Hou H, Song H, Lin K, Zhang Z, Hu J, Pang E. The comparison of alternative splicing among the multiple tissues in cucumber. BMC Plant Biol. 2018;18(1):5.
17. Dong C, He F, Berkowitz O, Liu J, Cao P, Tang M, Shi H, Wang W, Li Q, Shen Z, et al. Alternative splicing plays a critical role in maintaining mineral nutrient homeostasis in rice (*Oryza sativa*). Plant Cell. 2018;30(10):2267–85.
18. Thatcher SR, Danilevskaya ON, Meng X, Beatty M, Zastrow-Hayes G, Harris C, Van Allen B, Habben J, Li B. Genome-wide analysis of alternative splicing during development and drought stress in maize. Plant Physiol. 2016;170(1):586–99.
19. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9.
20. Wagner GP, Kin K, Lynch VJ. A model based criterion for gene expression calls using RNA-seq data. Theory Biosci. 2013;132(3):159–64.
21. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyras E. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19(1):40.
22. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–85.
23. UniProt CT. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46(5):2699.
24. Nagy E, Maquat LE. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci. 1998;23(6):198–9.
25. Ohtani M, Wachter A. NMD-based gene regulation-a strategy for fitness enhancement in plants? Plant Cell Physiol. 2019;60(9):1953–60.
26. Kalyna M, Simpson CG, Syed NH, Lewandowska D, Marquez Y, Kusenda B, Marshall J, Fuller J, Cardle L, McNicol J, et al. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. Nucleic Acids Res. 2012;40(6):2454–69.
27. Conesa A, Gotz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008;2008:619832.
28. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32(Database issue):D277–80.
29. Zdobnov EM, Apweiler R. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17(9):847–8.
30. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470–6.
31. Alamancos GP, Pagès A, Trincado JL, Bellora N, Eyras E. Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA. 2015;21:1521–31.
32. Zheng Y, Jiao C, Sun H, Rosli HG, Pombo MA, Zhang P, Banf M, Dai X, Martin GB, Giovannoni JJ, et al. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol Plant. 2016;9(12):1667–70.
33. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.
34. Wang BB, Brendel V. The ASRG database: identification and survey of *Arabidopsis thaliana* genes involved in pre-mRNA splicing. Genome Biol. 2004;5(12):R102.
35. Li D, Mei H, Shen Y, Su S, Zhang W, Wang J, Zu M, Chen W. ECharts: a declarative framework for rapid construction of web-based visualization. Visual Informatics. 2018;2(2):136–46.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000;25(1):25–9.
37. Garciamas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González VM, Hénaff E, Câmara F, Cozzuto L, Lowy E. The genome of melon (*Cucumis melo* L.). Proc Natl Acad Sci U S A. 2012;109(29):11872.
38. Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet. 2013;45(1):51–8.
39. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16(6):276–7.
40. Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry. 1994;33(10):3038–49.
41. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292(2):195–202.
42. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 2015;31(6):857–63.
43. Bendtsen JD, Nielsen H, Von HG, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol. 2004;340(4):783–95.
44. Briesemeister S, Rahnenführer J, Kohlbacher O. YLoc - an interpretable web server for predicting subcellular localization. Nucleic Acids Res. 2010;38(Web Server issue):497–502.
45. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics. 2004;4(6):1633–49.

Publisher's Note