Published online 1 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D1055–D1059***
doi:10.1093/nar/gkr1171

SpliceDisease database: linking RNA splicing and disease

Juan Wang[1,2,*], Jie Zhang[3], Kaibo Li[1], Wei Zhao[2] and Qinghua Cui[1,2]

[1]Department of Biomedical Informatics, Peking University Health Science Center, [2]MOE Key Laboratory of Molecular Cardiology, Peking University, Beijing 100191 and [3]Beijing Fang Yuan Teng DA Technology Co. Ltd., Beijing 100083, China

Received August 20, 2011; Revised November 11, 2011; Accepted November 12, 2011

ABSTRACT

RNA splicing is an important aspect of gene regulation in many organisms. Splicing of RNA is regulated by complicated mechanisms involving numerous RNA-binding proteins and the intricate network of interactions among them. Mutations in *cis*-acting splicing elements or its regulatory proteins have been shown to be involved in human diseases. Defects in pre-mRNA splicing process have emerged as a common disease-causing mechanism. Therefore, a database integrating RNA splicing and disease associations would be helpful for understanding not only the RNA splicing but also its contribution to disease. In SpliceDisease database, we manually curated 2337 splicing mutation disease entries involving 303 genes and 370 diseases, which have been supported experimentally in 898 publications. The SpliceDisease database provides information including the change of the nucleotide in the sequence, the location of the mutation on the gene, the reference Pubmed ID and detailed description for the relationship among gene mutations, splicing defects and diseases. We standardized the names of the diseases and genes and provided links for these genes to NCBI and UCSC genome browser for further annotation and genomic sequences. For the location of the mutation, we give direct links of the entry to the respective position/region in the genome browser. The users can freely browse, search and download the data in SpliceDisease at http://cmbi.bjmu.edu.cn/sdisease.

INTRODUCTION

Cells need to regulate the expression of a gene in a specific level at specific time and space in order to fulfill specific task. Gene regulation is a ubiquitous phenomenon and is critical in every biological process (1). Mechanisms of gene regulation include the regulations of transcription, RNA processing and translation. In higher eukaryotes, pre-mRNA splicing plays an important role in gene regulation. The inclusion of different exons in mRNA—alternative splicing (AS)—enables one single gene to produce multiple different mRNAs, which can be further translated into different proteins called splice variants (2,3). New high-throughput sequencing technology has revealed that >90% of human genes undergo AS—a much higher percentage than anticipated (4). And recent genome-wide analyses have indicated that almost all primary transcripts from multi-exon human genes undergo alternative pre-mRNA splicing (5). Therefore, RNA splicing greatly increases the genomic complexity of higher eukaryotes (6). RNA splicing is tissue specific and studies highlight differences in the types of AS occurring commonly in different tissues. For example, the frequencies of alternative 3′ splice site and alternative 5′ splice site usage are ~50–100% higher in liver than in other investigated tissues (7). The importance of splicing is emphasized by its presence in species throughout the phylogenetic tree. Evolutionary studies, which have revealed the formation of *de novo* alternative exons and the evolution of exon–intron architecture, highlight the importance of AS in the diversification of the transcriptome, especially in humans (8).

As we stated earlier, RNA splicing is critical in many biological processes. Splicing of RNA is regulated by complicated mechanisms involving numerous RNA-binding proteins and the intricate network of interactions among them. Splicing in general, and AS in particular, if disrupted, can lead to disease. Therefore, mutations in *cis*-acting splicing elements or splicing machinery and the regulatory proteins which could compromise the accuracy of either constitutive or alternative splicing would have a profound impact on human pathogenesis. Defects in pre-mRNA splicing have been shown as a common disease-causing mechanism in several studies (9–11).

*To whom correspondence should be addressed. Tel: +8610 82801585; Fax: +8610 82801001; Email: wjuan@hsc.pku.edu.cn

Table 1. Databases and tools of splicing mutation and alternative splicing

Resource	Description	URL
HGMD (15)	The Human Gene Mutation Database (HGMD) constitutes a comprehensive core collection of data on germ-line mutations in nuclear genes underlying or associated with human inherited disease	www.hgmd.org
DBASS5 (16,17)	A database of aberrant 5′ splice sites	http://www.dbass.org.uk/
DBASS3 (17,18)	A database of aberrant 3′ splice sites	http://www.dbass.org.uk/
ASDB (19)	Database of alternatively spliced genes	http://cbcg.nersc.gov/asdb
ssSNPTarget (20)	A genome-wide splice-site Single Nucleotide Polymorphism database	http://ssSNPTarget.org
EuSplice (21)	a splice-centric database which provides reliable splice signal and AS information for 23 eukaryotes	http://66.170.16.154/EuSplice
AsMamDB (22)	An alternative splice database of mammals	http://166.111.30.65/ASMAMDB.html
Alternative Splicing Database (23)	An alternative splicing database based on publications	http://cgsigma.cshl.org/new_alt_exon_db2/
TassDB2 (24)	A database of subtle alternative splicing events	http://www.tassdb.info
ISIS (25)	An intron information system	http://isis.bit.uq.edu.au/
ASPicDB (26)	A database of annotated transcript and protein variants generated by alternative splicing	http://www.caspur.it/ASPicDB/
STEPs (27)	A database of splice translational efficiency polymorphisms	http://dbstep.genes.org.uk/
Human Splicing Finder (14)	A tool to predict the effects of mutations on splicing signals or to identify splicing motifs in any human sequence	http://www.umd.be/HSF/
SpliceMiner (28)	A high-throughput database implementation of the NCBI Evidence Viewer for microarray splice variant analysis	http://discover.nci.nih.gov/spliceminer
Intronerator (29)	Exploring introns and alternative splicing in *Caenorhabditis elegans*	http://www.cse.ucsc.edu/~kent/intronerator
WebScipio (30)	Tool for Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology	http://www.webscipio.org
IsoEM (31)	Tool for the estimation of alternative splicing isoform frequencies from RNA-Seq data	http://dna.engr.uconn.edu/software/IsoEM/
MAISTAS (32)	A tool for automatic structural evaluation of alternative splicing products	http://maistas.bioinformatica.crs4.it/
HMMSplicer (33)	A tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data	http://derisilab.ucsf.edu/software/hmmsplicer
SFmap (34)	A web server for motif analysis and prediction of splicing factor binding sites	http://sfmap.technion.ac.il

As an example, a point mutation in exon 7 of SMN2 gene leads to exon 7 skipping and a truncated protein, which causes decreased effective rate of SMN protein production and motor neuron degenerative disease (12). Other studies indicate *trans*-acting mutations affect RNA-dependent functions and cause disease (9,13). A number of bioinformatics resources for RNA splicing have been developed during the past decade including databases and tools (Table 1). For example, Human Splicing Finder is a tool to predict the effects of mutations on splicing signals and can identify splicing motifs in human sequence (14). These resources have provided great help in the study and analysis of RNA splicing.

The above evidences have shown an increased importance of connecting the RNA splicing and diseases. For this reason, a high quality database linking RNA splicing and splicing mutations with disease will be of great help and be emergently needed in the study of both RNA splicing and disease. Although the human gene mutation database (HGMD, http://www.hgmd.org/) integrated this kind of data but there are big difference between HGMD and SpliceDisease database (15). Firstly, HGMD is not free and only provides 'search' function for registered users for limited days. Secondly, HGMD only provides information of point mutations of intronic sequence for splicing mutation. Thirdly, HGMD does not provide detailed descriptions for the relationship among gene mutations, splicing defects and diseases. On the other hand, SpliceDisease database is a free and comprehensive database containing *cis*-splicing sequence mutations and *trans*-acting splicing mutations that cause disease. SpliceDisease integrates detailed descriptions for the relationship among gene mutation, splicing defect and disease. And it provides direct links of EntreZ gene, genome browser, respective location of the mutation on the gene and PubMed for each literature. At present, the 'SpliceDisease' database is at its first step, it will be a valuable ongoing resource for the study of RNA splicing and disease.

DATA SOURCES AND IMPLEMENTATION

RNA splicing and disease-related literature was acquired by PubMed search using the keywords 'splice', 'splicing' and 'spliced'. Literatures with titles including 'mutation spectrum', 'mutational spectrum', 'mutation analysis' and 'mutation screening' were also obtained. We then curated the data manually and retrieved the association between RNA splicing and splicing mutations in the gene and disease of interest. The data were double checked by different people. We standardized the disease names and gene names based on NLM Mesh Browser and

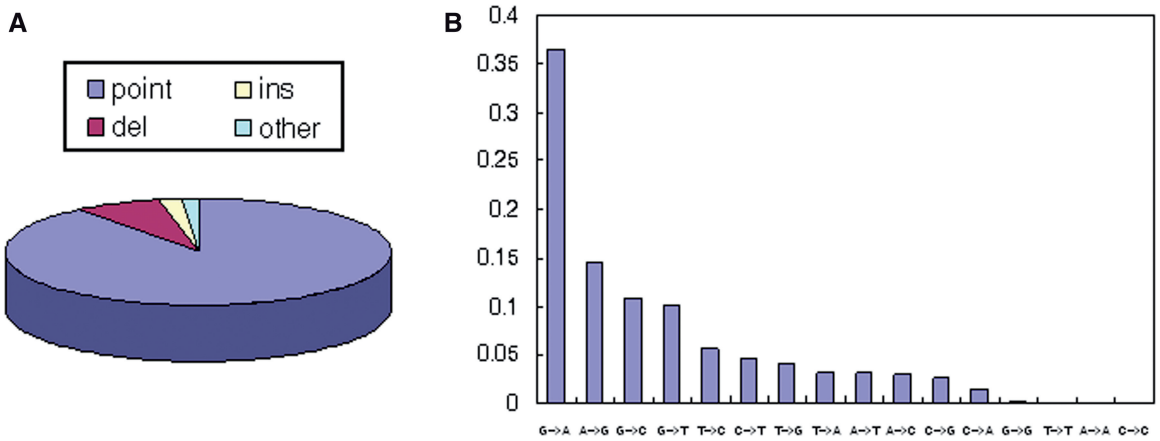

Figure 1. Distribution of mutation type and distribution of point mutation type in the SpliceDisease database. (**A**) Splicing mutation type: point, point mutation; ins, insertion mutation; del, deletion mutation; other, other types. (**B**) Axes of the histogram represent the proportions of different nucleotide substitutions in whole point mutations.

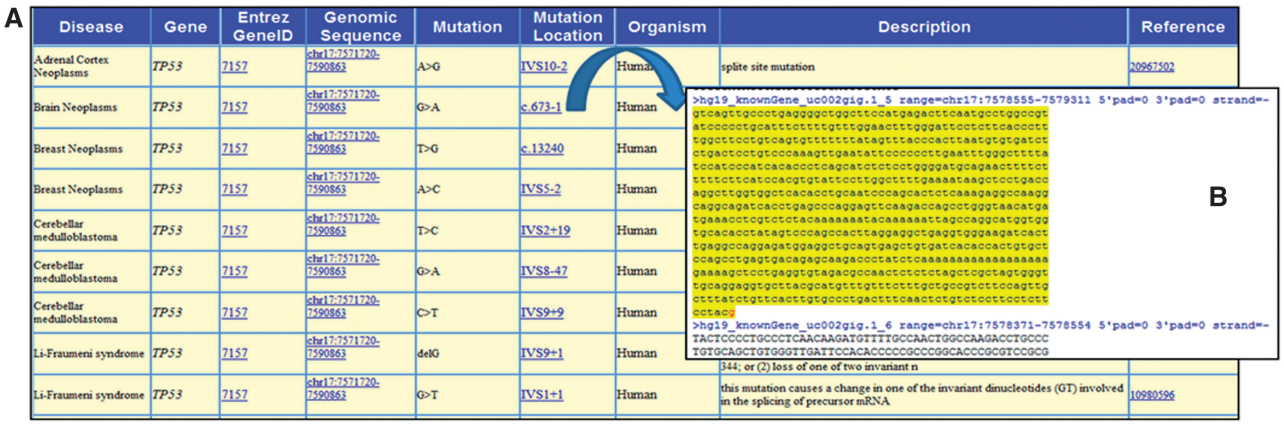

Figure 2. SpliceDisease results page. (**A**) Once a user runs a search, there comes the result summary page that includes nine items. (**B**) The direct link for entry to the respective position of mutation. The sequence of exon shows in upper case and intron shows in lower case. And one FASTA record per region (exon, intron) is used in the sequence file. The inton/exon of the location of mutation is highlighted in yellow color and specific nucleotide is marked in red color.

EntreZ gene. Each gene was linked to NCBI for comprehensive annotations and to UCSC genome browser for genomic sequence. The mutations of genes were annotated as well including nucleotide change and location on the sequence. We used the nomenclature for description of sequence variants and exon/intron numbering according to den Dunnen and Antonarakis (35). For example: 'c.' for a *cDNA* sequence; IVS for intron sequence; substitutions are designated by a '>' character. We also gave direct link for entry to the respective position of mutation. In the sequence file, the intron/exon of the mutation location is highlighted in yellow color and the specific nucleotide is marked in red color. PubMed IDs and hyperlinks to PubMed were also provided for each literature. More importantly, we curated the detailed description for the relationship among gene mutation, splicing defect and disease.

As a result, we manually curated 2337 splicing mutation-disease entries including 303 genes and 370 diseases from 898 publications. In the 2337 entries, ~89% of them are point mutations (Figure 1A) among which >50% are mutations between G and A (36.5% G > A and 14.6% A > G) (Figure 1B).

All data were organized in the 'SpliceDisease' database using PostgreSQL 9.0, a lightweight database management system. The website is presented using Apache Tomcat 7.0, a JSP&Java web framework which is available at http://cmbi.bjmu.edu.cn/sdisease/.

USING SpliceDisease

SpliceDisease is a user-friendly designed database. The homepage has been designed to provide an organized venue to access all data. The top banner section of the homepage has tabs for 'Browser', 'Search', 'Submit', 'Download' and 'Help', respectively. When a user performs a search in SpliceDisease, he can use the 'Browser' to select the disease or gene of interest or use the 'Search' which supports fuzzy queries to find it. The page of result contains nine items disease name and gene symbol, gene EntreZ ID (link to NCBI gene

database), chromosome location of genomic sequence (link to UCSC genome browser), mutation, mutation location (direct link to respective position of mutation in the genome browser automatically), organism, description and reference (link to PubMed database) (Figure 2).

All data in SpliceDisease can be downloaded in a file of csv format. These data will facilitate study of exploitation of splicing mutational mechanisms, understanding of RNA biology and helping to discover new therapeutic targets.

FUTURE EXTENSIONS

The SpliceDisease database is in the first step of the project and further extensions will be developed. As we described earlier, a number of bioinformatics resources for RNA splicing have been developed. Therefore, we plan to integrate some related bioinformatics resources in the near future. We will also incorporate the expression data of different mRNA isoforms. As the data accumulation, we will add more *trans*-acting splicing mutations that cause disease. Finally, SpliceDisease will be continuously updated.

FUNDING

Funding for open access charge: National Natural Science Foundation of China (Grant No. 81001481).

Conflict of interest statement. None declared.

REFERENCES

1. Holste,D. and Ohler,U. (2008) Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events. *PLoS Comput. Biol.*, **4**, e21.
2. Xing,Y., Resch,A. and Lee,C. (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.*, **14**, 426–441.
3. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.
4. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
5. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
6. Brett,D., Pospisil,H., Valcarcel,J., Reich,J. and Bork,P. (2002) Alternative splicing and genome complexity. *Nat. Genet.*, **30**, 29–30.
7. Yeo,G., Holste,D., Kreiman,G. and Burge,C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol.*, **5**, R74.
8. Keren,H., Lev-Maor,G. and Ast,G. (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, **11**, 345–355.
9. Cooper,T.A., Wan,L. and Dreyfuss,G. (2009) RNA and disease. *Cell*, **136**, 777–793.
10. Faustino,N.A. and Cooper,T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev.*, **17**, 419–437.
11. Tazi,J., Bakkour,N. and Stamm,S. (2009) Alternative splicing and disease. *Biochim. Biophys. Acta*, **1792**, 14–26.
12. Wirth,B., Brichta,L. and Hahnen,E. (2006) Spinal muscular atrophy: from gene to therapy. *Semin. Pediatr. Neurol.*, **13**, 121–131.
13. Liu,F. and Gong,C.X. (2008) Tau exon 10 alternative splicing and tauopathies. *Mol. Neurodegener.*, **3**, 8.
14. Desmet,F.O., Hamroun,D., Lalande,M., Collod-Beroud,G., Claustres,M. and Beroud,C. (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.*, **37**, e67.
15. Cooper,D.N., Stenson,P.D. and Chuzhanova,N.A. (2006) The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr. Protoc. Bioinform.*, **Chapter 1**, Unit 1 13.
16. Buratti,E., Chivers,M., Kralovicova,J., Romano,M., Baralle,M., Krainer,A.R. and Vorechovsky,I. (2007) Aberrant 5′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.*, **35**, 4250–4263.
17. Buratti,E., Chivers,M., Hwang,G. and Vorechovsky,I. (2011) DBASS3 and DBASS5: databases of aberrant 3′- and 5′-splice sites. *Nucleic Acids Res.*, **39**, D86–D91.
18. Vorechovsky,I. (2006) Aberrant 3′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.*, **34**, 4630–4641.
19. Dralyuk,I., Brudno,M., Gelfand,M.S., Zorn,M. and Dubchak,I. (2000) ASDB: database of alternatively spliced genes. *Nucleic Acids Res.*, **28**, 296–297.
20. Yang,J.O., Kim,W.Y. and Bhak,J. (2009) ssSNPTarget: genome-wide splice-site Single Nucleotide Polymorphism database. *Hum. Mutat.*, **30**, E1010–E1020.
21. Bhasi,A., Pandey,R.V., Utharasamy,S.P. and Senapathy,P. (2007) EuSplice: a unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes. *Bioinformatics*, **23**, 1815–1823.
22. Ji,H., Zhou,Q., Wen,F., Xia,H., Lu,X. and Li,Y. (2001) AsMamDB: an alternative splice database of mammals. *Nucleic Acids Res.*, **29**, 260–263.
23. Stamm,S., Zhu,J., Nakai,K., Stoilov,P., Stoss,O. and Zhang,M.Q. (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol.*, **19**, 739–756.
24. Sinha,R., Lenser,T., Jahn,N., Gausmann,U., Friedel,S., Szafranski,K., Huse,K., Rosenstiel,P., Hampe,J., Schuster,S. *et al.* (2010) TassDB2 - A comprehensive database of subtle alternative splicing events. *BMC Bioinformatics*, **11**, 216.
25. Croft,L., Schandorff,S., Clark,F., Burrage,K., Arctander,P. and Mattick,J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat. Genet.*, **24**, 340–341.
26. Castrignano,T., D'Antonio,M., Anselmo,A., Carrabino,D., D'Onorio De Meo,A., D'Erchia,A.M., Licciulli,F., Mangiulli,M., Mignone,F., Pavesi,G. *et al.* (2008) ASPicDB: a database resource for alternative splicing analysis. *Bioinformatics*, **24**, 1300–1304.
27. Raistrick,C.A., Day,I.N. and Gaunt,T.R. (2010) Genome-wide data-mining of candidate human splice translational efficiency polymorphisms (STEPs) and an online database. *PLoS One*, **5**, e13340.
28. Kahn,A.B., Ryan,M.C., Liu,H., Zeeberg,B.R., Jamison,D.C. and Weinstein,J.N. (2007) SpliceMiner: a high-throughput database implementation of the NCBI Evidence Viewer for microarray splice variant analysis. *BMC Bioinformatics*, **8**, 75.
29. Kent,W.J. and Zahler,A.M. (2000) The intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans*. *Nucleic Acids Res.*, **28**, 91–93.
30. Pillmann,H., Hatje,K., Odronitz,F., Hammesfahr,B. and Kollmar,M. (2011) Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology. *BMC Bioinformatics*, **12**, 270.
31. Nicolae,M., Mangul,S., Mandoiu,I.I. and Zelikovsky,A. (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.*, **6**, 9.
32. Floris,M., Raimondo,D., Leoni,G., Orsini,M., Marcatili,P. and Tramontano,A. (2011) MAISTAS: a tool for automatic structural

evaluation of alternative splicing products. *Bioinformatics*, **27**, 1625–1629.
33. Dimon,M.T., Sorber,K. and DeRisi,J.L. (2010) HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS One*, **5**, e13875.
34. Paz,I., Akerman,M., Dror,I., Kosti,I. and Mandel-Gutfreund,Y. (2010) SFmap: a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Res.*, **38**, W281–W285.
35. den Dunnen,J.T. and Antonarakis,S.E. (2001) Nomenclature for the description of human sequence variations. *Hum. Genet.*, **109**, 121–124.