

# TassDB2 - A user guide

## Introduction

TassDB2 (TAndem Splice Site DataBase, version 2), is a database of alternative splice sites in mouse and human which are separated by distances of 2-12 nt.

In the following, the notation  $\Delta x$  is used to denote a subtle splice event involving sites separated by  $x$  nucleotides, so for example, the class  $\Delta 3$  means all GYNGYN and NAGNAG AS events (Y stands for C or T; N for A, C, G, or T), and so on.

TassDB2 includes data on the conservation of the tandem motifs in five vertebrates (human, mouse, dog, chicken and zebrafish). Thus, TassDB2 provides comprehensive information on 22 event types. A user-friendly search interface features both a “quick search” mode, in which a user can search using gene symbol, as well as an “advanced search” mode, in which several different criteria can be specified by the user, and the possibility to download result datasets.

In the following we describe the search interfaces and explain the results page.

### *Nomenclature for the splice sites and transcripts*

Tandem alternative splicing using the intron-proximal site (E acceptor / e donor - the distal part of the tandem becomes EXONIC) results in the E/e transcript, whereas splicing using the intron-distal site (I acceptor / i donor - as the entire tandem becomes INTRONIC) in the I/i transcript.

## **Quick search interface**

The quick search interface is meant for cases where a user is looking for a single gene or possibly genes with similar names (like gene families). One only needs to provide a gene symbol, alias or transcript ID as a search string. The interface reports the gene(s) with tandem splice site(s) in a table, and the user can click on the gene symbol(s) to view the details (see Result page below).

## **Search query**

The user can enter a gene symbol (like BRCA1) or a transcript ID. A transcript ID can either be a RefSeq ID (like NM\_000014) or a UCSC Known Gene ID (like uc001lsx). It is recommended to use gene symbols approved by the nomenclature committees for human (<http://www.genenames.org>) and mouse (<http://www.informatics.jax.org/mgihome/nomen>). You may also enter a gene alias (like RNF53) although it is not guaranteed that the lists of aliases are exhaustive. The search is an exact one and is case insensitive. However, an asterisk (\*) may be used as a wildcard on the right side of your search term. For example, "STAT\*" will retrieve all hits for genes as STAT1, STAT2, STAT3, STAT4, STAT5A, STAT5B, and STAT6 but also for STATH, the gene for statherin.

## Advanced search interface

TassDB2 provides an “advanced search interface” for situations where users are interested in tandem splice sites with specific features.

In the following the various options provided in the advanced search interface are explained in more detail.

### Delta

Delta is the distance between the splice sites in a tandem. E.g., GYNGYN and NAGNAG represent delta 3 tandem sites (Y stands for C or T; N for A, C, G, or T).

### **EST/mRNA confirmed**

A tandem splice site with at least one EST/mRNA for each of the e/E and i/I transcripts is called confirmed. You can restrict the search to those by clicking the “confirmed” button. You can also search for sites whose e/i and/or E/I transcript is represented by other numbers of ESTs/mRNAs using " $\geq$ ", "=", or " $\leq$ ".

Additionally, the minor isoform ratio can be used to search for tandem sites that generate both isoforms to a certain degree. This ratio is computed by the numbers  $n$  of ESTs/mRNAs representing e/i or E/I transcript as  $\min(n_e, n_i)/(n_e + n_i)$  and  $\min(n_E, n_I)/(n_E + n_I)$ , respectively. The minor isoform ratio is thus given as a fraction, with values between 0 and 0.5.

### **Splice site scores**

The splice site scores are computed with MaxEntScan ([http://genes.mit.edu/burgelab/maxent/Xmaxentscan\\_scoreseq.html](http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html)). A comparison of the scores can give a hint which splice site is preferentially used. The splice site with the higher score is often covered by more ESTs.

### **Tandem site conservation**

Conservation was computed based purely on the two splice sites. Only full matches are considered as conserved. A dash indicates that either the tandem splice site was not conserved, or that the sequence could not be found in the genome. Here, conservation simply means that both species contain tandem splice sites, the neighbouring nucleotides need not be conserved.

### **UTR/CDS**

The location of the tandem splice sites in the transcript which can be restricted either to untranslated regions (UTR) and/or the coding sequence (CDS). If the tandem splice site is located in the CDS the effect of the alternative splice event on the protein will be calculated.

## Results page

The result of the search consists of two parts: (i) a summary table listing the affected gene(s) and their number of tandem splice sites of each type, and (ii) detailed gene specific tables containing information regarding the individual tandem splice sites. These detailed result tables also provide links to the ESTs/mRNAs for both splice forms as well as links to the UCSC genome browser.

A tandem splice site can affect more than one transcript. If the transcript specific data differ between transcripts, TassDB2 shows detailed result tables with more than two columns. Features that differ between transcripts are shown in black while those that are identical in all transcripts are shown in grey color.

The following details for identified tandem splice sites are given in table:

locus	The chromosome number and the position of the tandem splice site on the chromosome. The link refers to the UCSC genome browser and shows the tandem splice site in the context of other genome annotations.
sequence context	The tandem splice site (in red) as well as its exonic or intron flank is shown. Lower case letters represent intronic nucleotides, upper case letters exonic nucleotides.
splice site scores Ee/Ii	The splice site scores are computed with MaxEntScan <a href="http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html">http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html</a> (G Yeo & C Burge (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11 (2-3) 377-394). A comparison of the scores can give a hint which splice site is preferentially used. The splice site with the higher score is often covered by more ESTs. The score 'NaN' (not-a-number) means that the splice site context contains characters other than A, C, G, or T. NOTE: MaxEntScan was trained on human data. Thus, the splice site scores for species other than human might have less meaning. However, since the splice site nucleotide preferences are quite similar between species (at least in vertebrates), we believe that scores for other species are useful as well.
conservation	If the tandem splice site is conserved in mouse/human, dog, chicken, or zebrafish, the respective tandem site is shown together with a link to the USCS genome browser. If the site is not conserved or the sequence could not be found in the genome, only a dash is given.

transcript	Accession number(s) of the transcript model(s) harboring the tandem splice site. Clicking on the accession number opens a window displaying the RefSeq entry at NCBI or the UCSC gene model. Features that differ between transcripts are shown in black while those that are identical in all transcripts are shown in grey color.
exon number	The number of the exon that is upstream of a tandem donor and downstream of a tandem acceptor, respectively.
annotated splice site	The tandem splice site (e or i for a GY...GY donor, E or I for a AG...AG acceptor) that is used in the respective transcript.
number of Ee/Ii transcripts	The number of ESTs/mRNAs that represent e/I transcripts for tandem donors and E/I transcripts for tandem acceptors, respectively. Clicking on the link will open a window containing a list of all these ESTs/mRNAs and their description linked to the respective database entries.
protein impact	If the tandem splice site is located in the CDS the effect of the alternative splice event on the protein is given. Single amino acid events are shown as “indel G”, the exchange of a dipeptide for an unrelated amino acid as “GN vs. D”. Stop codons are shown as an asterisk (*). A frameshift (induced by a delta that is not 3, 6, 9 or 12) has a much larger effect than these local amino acid substitutions. This case, or a tandem splice site location in the UTR, is indicated by a dash. If the mRNA sequence of a codon contains a character other than A, C, G, T, the translated codon is shown as X.
position in protein	The position of the amino acid that is exactly upstream of the putative amino acid variation event is given. If the alternative splicing event at the tandem splice site is located in the UTR, a dash is displayed.
NMD	A transcript is considered as an nonsense-mediated decay (NMD) candidate if the stop codon is 50 nt or more upstream of the last exon-exon junction.

Result details can be downloaded in csv table format (text, tab separated). First line of the downloaded file contains the number of identified tandem sites. The third line contains column names; tab separated. Each following line refers to an affected transcript.

**Columns:**

GeneSymbol	gene symbol approved by the HUGO Gene Nomenclature Committee (HGNC) <a href="http://www.genenames.org">www.genenames.org</a>
Aliases	incomplete list of known aliases of the approved gene symbol
GeneName	gene name approved by HGNC <a href="http://www.genenames.org">www.genenames.org</a>

SpliceSite	donor/acceptor
Species	human/mouse
Locus	chromosome and the position of the tandem splice site
Sequence	UPPERCASE: exonic (23nt), tandem (Delta+3nt); lowercase: intronic (17nt)
ScoreE/e	MaxEntScan score of the E/e splice site. The score "NaN" (not-a-number) means that the splice site context contains characters other than A, C, G, T
ScoreI/i	MaxEntScan score of the I/i splice site. The score "NaN" (not-a-number) means that the splice site context contains characters other than A, C, G, T
Transcript	GenBank/EMBLsdb accession number(s) harbouring the tandem splice site
ExonNum	number of the exon affected by the tandem splice site
Annotation	symbol of the splice site (E, e, I or i) utilized in the transcript
NumEtrans	number of ESTs/mRNAs confirming the E/e site
NumItrans	number of ESTs/mRNAs confirming the I/i site
ProteinImpact	CDS: "indel", "vs." (substitution), single aa code, "*" stop codon; UTR/Frameshift: "-", mRNA contains character other than A,C,G,T: "X".
PosProtein	CDS: position of the amino acid upstream of the variation; UTR: "-"

## **Database construction and content**

### **Data**

The annotation pipeline of TassDB2 is based on the following:

- Transcript-to-genome mappings: As per the UCSC genome browser [1]
- Gene annotation: RefSeq annotation as well as the UCSC 'knownGene' set
- Genome builds: hg18 (human) and mm9 (mouse)
- Exon–intron structure and protein-coding sequence (CDS): as per the UCSC annotation

Subtle AS events were identified using BLAST against all ESTs and mRNAs from the respective species as described in [2, 3].

For each tandem splice site and the confirmed or putative AS event, TassDB2 contains the following data:

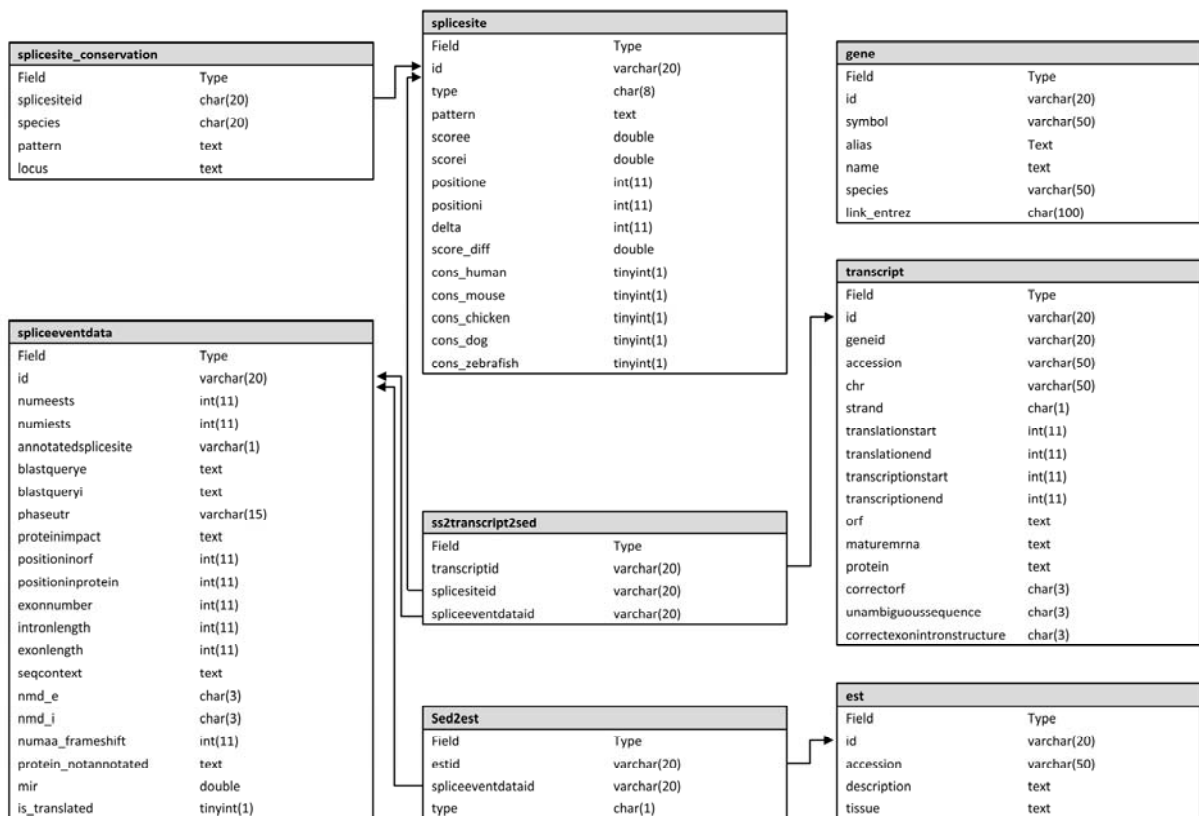
- Splice site motif
- Genomic locus
- Location in the transcript (5'/3'-UTR or CDS with intron phase 0/1/2)
- The (predicted) impact of the splice event on the protein
- The sequences and length of the up-/downstream exon and the intron
- Information about the ESTs/mRNAs that indicate usage (if any) of the splice sites

### **Database Design**

The web-frontend to TassDB2 is created in HTML with PHP and JavaScript. The data is stored in a relational database, running under the MySQL database system. The data is primarily organized in the database tables `splicesite`, `spliceeventdata`, and `transcript`.

The table `splicesite` contains sequence-dependent information such as the genomic locus, the splice site pattern with its sequence context, the splice site scores, and conserved tandem sequences (if available) in human/mouse, chicken, dog, and zebrafish. All transcript-dependent data is stored in table `spliceeventdata`: the transcripts which have the tandem site in their exon-intron structure, the annotated splice site, the number of ESTs for each (potential) tandem splice variant along with the two BLAST queries used to find the ESTs, the predicted protein impact, and the NMD prediction. The table `transcript` contains the information on the transcripts that is independent from the splice sites. The three main tables are linked through the `ss2transcript2sed` table.





Additionally, each splice site is linked to information on its gene (table gene), and its conservation in other species (table splicesite\_conservation; species are human, mouse, dog, chicken, zebrafish, representing the major vertebrate clades). Here, conservation simply means that both species contain tandem splice sites – the neighbouring nucleotides need not be conserved. The splicing events are linked to their supporting ESTs in the table est. The user interface contains links giving a detailed description of each data field.

## Specifications of the software and hardware running TassDB2

- Database language: mysql 5.0.67
- Scripting language: php 5.2.11
- Webservice software: apache 2.2.10
- Operating system: Linux (openSUSE 11.1, kernel 2.6.27)
- CPU: 2 quad-core intel @ 2.33 GHz, 16 GB ram

## Examples

(1) Searching for all confirmed tandem splice sites in the gene HHIP (hedgehog interacting protein) in human leads to the result page shown in Fig. 3: HHIP has one confirmed  $\Delta 4$  tandem acceptor event, with the upstream and downstream acceptor supported by 30 and 34 ESTs/mRNAs, respectively. The event is predicted to lead to targeting by NMD according to one of the representative transcripts (uc003ijs.1, NM\_022475), but not according to the other (uc003ijr.1).

(2) Searching for all confirmed tandem splicing events with a minor isoform ratio of  $\geq 0.45$  yields 300 results, and increasing the threshold of supporting ESTs/mRNAs to  $\geq 10$  for each variant yields 170 results.

**TassDB2 - Tandem Splice Site DataBase**

Search Results: 1


Species	Gene Symbol	Gene Name	Delta	# Donor	# Acceptor
Human	<a href="#">HHIP</a> (HIP, FLJ20992)	hedgehog interacting protein	4		1

Details:

Clicking on the yellow line headers opens an information window

**HHIP** (HIP, FLJ20992) **hedgehog interacting protein**  
 acceptor, Human, E/I delta 4 nt

locus	<a href="#">chr4:145800227-145800251</a>	
sequence context	gttttctttaattgttTAGAAAGCACAAACACAACTGCTTCTGTAT	
splice site scores Ee/Ii	6.16 / -1.72	
conservation	-	
transcript	<a href="#">uc003ijs.1, NM_022475</a>	<a href="#">uc003ijr.1</a>
exon number	4	4
annotated splice site	E	E
number of Ee/Ii transcripts	30 / 34	30 / 34
protein impact	-	-
position in protein	aa: 209	aa: 209
NMD Ee/Ii	no / yes	no / no



## BayNAGNAG webserver

The TassDB2 resource also includes the BayNAGNAG webserver (available at <http://www.tassdb.info/baynagnag/>), which uses Bayesian networks to predict the splicing outcome at NAGNAG tandem splice sites in an EST/mRNA independent way based on splice site features [4].

## Availability and requirements

TassDB2 is freely available for online use at <http://www.tassdb.info>.

TassDB2 can be used via any standard internet browser.

## Acknowledgements

This work was supported by grants from the German Ministry of Education and Research (01GS08182, 01GS0809, 0313652), the Deutsche Forschungsgemeinschaft (SFB604-02, Ha3091/2-1, Hi1423/2-1, Hu498/3-1) and the Human Frontier Science Program (Fellowship LT000896/2009-L).

## References

1. Karolchik D, Baertsch R, Diekhans R, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31**:51 - 54.
2. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity.** *Nat Genet* 2004, **36**(12):1255-1257.
3. Hiller M, Huse K, Szafranski K, Rosenstiel P, Schreiber S, Backofen R, Platzer M: **Phylogenetically widespread alternative splicing at unusual GYNGYN donors.** *Genome Biology* 2006, **7**(7):R65.
4. Sinha R, Nikolajewa S, Szafranski K, Hiller M, Jahn N, Huse K, Platzer M, Backofen R: **Accurate prediction of NAGNAG alternative splicing.** *Nucl Acids Res* 2009, **37**(11):3569-3579.