

User manual

Table of Contents

1. The databases

- 1.1. The Alternative splicing database (ASD)
- 1.2. The Alternative transcript diversity database (ATD)
- 1.3. The AEdb database

2. Derivation of the basic data

- 2.1. Start-up gene set and gene structure annotations
- 2.2. Cleaning
- 2.3. Identifying transcript-confirmed introns & exons
- 2.4. Derivation of Transcript classes
 - 2.4.1. Transcript structure
 - 2.4.2. Derivation of transcript classes
 - 2.4.3. Ambiguity in the assignment of transcripts
 - 2.4.4. Assignment of transcripts that contain ambiguities to classes
 - 2.4.4.1. Internal discontinuity in the transcript structure
 - 2.4.4.2. End ambiguities in the transcript structures
 - 2.4.4.3. Forming of classes and class ambiguity assignment
 - 2.4.4.4. Defining the quality of a class
 - 2.4.4.5. Defining the relationship between classes
 - 2.5. Delineation of events

3. ATD

- 3.1. Generating transcript patterns
- 3.2. Detecting transcription start sites
 - 3.2.1. Identification of TSS
 - 3.2.2. Mapping of TSS
- 3.3. Detecting poly(A) sites
 - 3.3.1. Detecting poly(A) tail and poly(A) cleavage (PAC) site
 - 3.3.2. Detecting poly(A) signal (PAS)
 - 3.3.3. AltPAS
 - 3.3.4. Choosing a representative cleavage site as poly(A) site
 - 3.3.5. Integrating poly(A) sites from the AltPAS and AltTrans pipelines

4. AEdb

5. Mapping and analysis

- 5.1. SNP mapping
 - 5.1.1. Identification of SNP positions in the genes
 - 5.1.2. Identification of exonic SNP positions from two or more
 - 5.1.3. Identification of alleles used at the exonic SNP positions

5.2. Expression states

6. Integration

6.1. AltSplice and AEdb-Sequence databases

7. Workbench

7.1 Intron analysis

7.2. Scoring ATG-context sequence

7.3. MZEF-SPC exon finder

7.4. Detection of short regulatory sequences

The work of the ASTD-EBI team, in collaboration with other institutes, is supported by grants from the EC: ATD consortium (LSHG-CT-2003-503329) and the Eurasnet consortium (LSHG-CT-2005-518238). It was supported by the ASD grant from the EC (QLRT-CT-2001-02062) until November 2005.

1. The databases

1.1. The Alternative splicing database (ASD)

<http://www.ebi.ac.uk/asd/>

Stamm S., Riethoven J-J.M., Le Texier V., Gopalakrishnan C., Kumanduri V., Tang Y., Barbosa-Morais N.L., and Thanaraj T.A.

ASD: a bioinformatics resource on alternative splicing.

Nucleic Acids Res 34: D46-D55 (2006)

Members of the European based ASD consortium have developed the Alternative Splicing Database (ASD) Project. The aims of the consortium were to analyse the mechanism of alternative splicing on a genome-wide scale by creating a database of alternative splice events and the resultant isoform splice patterns of genes from human and mouse.

A computational pipeline detects and characterises alternative introns and exons, alternative splice events, and isoform splice patterns and isoform peptide sequences. Value-added annotation includes expression states, human-mouse comparisons and allele-use at SNP positions.

Data is integrated from manually annotated databases: AEdb (described below) and UniProt peptide variants, with ASD splice patterns to add value of evidence to computationally predicted isoform splice events.

1.2. The Alternative transcript diversity (ATD) database

<http://www.ebi.ac.uk/atd/>

Le Texier V., Riethoven J-J., Kumanduri V., Gopalakrishnan C., Lopez F., Gautheret D. and Thanaraj T.A.

AltTrans: Transcript pattern variants annotated for both alternative splicing and alternative poly(A)denylation.

BMC Bioinformatics 7: 169 (2006)

Members of the European based ASD consortium have developed the Alternate Transcript Diversity (ATD) Project. The aim of the consortium is to understand the mechanisms that are responsible for the formation of transcript isoforms on a genome-wide scale by creating a value-added database of full-length alternate transcripts from human and mouse.

A computational pipeline detects and characterises transcription start sites, splice sites and poly(A) sites. These data are put in context of one another to generate full-length transcript isoforms. The isoform transcripts are annotated for various biological features, including expression states.

1.3. The AEdb database

<http://www.ebi.ac.uk/asd/aedb/index.html>

Stamm S., Zhu J., Nakai K., Stoilov P., Stoss O., and Zhang M.Q.
An alternative exon database and its statistical analysis.
DNA and Cell Biol., 19: 739-756 (2000)

AEdb is a manually curated database of exons captured from literature by Stefan Stamm (University of Erlangen, Germany). AEdb has several subsets of information: AEdb-Sequence (sequence and properties of alternatively splice exons), AEdb-Function (data on functional aspects of alternatively spliced exons), AEdb-motif (data and sequence of known splice regulatory motifs), AEdb-minigene (a collection of known minigene constructs for alternative splice events) and AEdb-disease (alternative exons that are causative to disease).

2. Derivation of the basic data

Clark F. and Thanaraj T.A.

Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human.

Hum. Mol. Genet. 11: 451-464 (2002)

Above citation provides very detailed information on the following stages.

2.1. Start-up gene set and gene structure annotations

This database of alternative splicing events is built on the set of genes from the Ensembl genome annotation project (<http://www.Ensembl.org/>). Ensembl defines a gene region along with multiple transcript structures. A gene in the pipeline represents a genomic region containing 5' flanking 3000 bases + Ensembl gene region + 3' flanking 3000 bases. We examine each gene region for the presence of alternatively spliced introns and exons and delineate the consequent alternative events. The events are annotated for various biological features including single nucleotide polymorphism (SNP) position, tissue expression state, human-mouse comparison of exons and events.

Transcript (Expressed sequence tag (EST) and messenger Ribonucleic Acid (mRNA)) sequences, as extracted from International Nucleotide sequence database collaboration (INSDC, <http://www.ebi.ac.uk/embl/>) database, are BLASTed (Basic Local Alignment Search Tool) against the Ensembl gene region to generate a high quality data set of gene-transcript alignments showing more than one high scoring match between gene and transcript sequences. Any transcript sequence that aligns exclusively to the flanking regions of a gene is discarded.

2.2. Cleaning

Before we start with the main process of trying to identify introns and exons we have several steps to clean our initial data set of known genes:

a. Removing immunoglobulin domain-containing genes.

All genes that have a high probability of belonging to the immunoglobulin family are removed by blasting against the latest release of Ig Germline genes (<http://www.ncbi.nlm.nih.gov/igblast/showGermline.cgi>) (pid=90%, coverage=90%).

b. Removing redundant genes.

The gene set is BLATed (The BLAST-like Alignment Tool) (ungapped) against itself to find those genes that are similar to a very high degree (pid=99%, coverage=90%). Transcript sequences that align to more than one gene or ambiguously with more than one region on a single gene are identified and the longest copy of the duplicate genes is kept.

c. Repeat masking.

While repeat masking is set as one of the parameters when aligning the ESTs and mRNAs to the Ensembl gene regions, it is run on the input sequence for BLAT through the program RepeatMasker (<http://www.repeatmasker.org/>).

2.3. Identifying transcript-confirmed introns & exons

Alignment gaps on gene sequences are considered as potential introns. Their validation as transcript-confirmed introns is a crucial step in AltSplice pipeline. Alignment matches on the gene sequence are accepted as a confirmed exon if flanked on either side by a confirmed intron. Each transcript sequence that aligns with a gene is described by its exon-intron structure. Transcript sequences that map to a gene are then grouped into classes in a manner that the member transcripts from a class show same exon-intron structure. The longest representative from each such class denotes a unique splice pattern for the gene. Overlapping exons and introns from the isoform splice patterns are then examined to delineate alternative splice events.

Defining introns:

Introns are determined by blasting a recent database of EST's and mRNA's against the gene set, and then analyse the resulting high scoring pairs (HSPs).

Various checks ensure the pipeline uses HSPs of a good quality: only extract HSPs with pid larger than 95% and a minimum length of 25 bases (therefore small exons are excluded from the pipeline), remove all transcripts (and HSPs) that map to more than one gene, choose only one gene/transcript strand combination (based on highest coverage of HSPs, highest pid), remove HSPs that map to the untranslated flanking region, remove HSPs that have more than 20% overlap within a gene on the EST side, and remove all HSPs that are the only HSP for a particular transcript.

A pair of successive HSPs is used to define an intron. These 'initial introns' go through various stages of validation to become a 'candidate intron'. The pipeline only allows the canonical *GT-AG*, *GC-AG*, or *AT-AC* type introns in our data set.

- a. If there is no gap or overlap on the EST between two successive HSPs, the intron is kept.
- b. If there is an overlap on the transcript sequence, firstly we check on the length of overlap; if the overlap length is > 10 bases, the 'initial intron' is removed from the data set.
- c. If the overlap is ≤ 10 bases, we try to shift the HSPs around (in such a manner that the order of coding nucleotides is not changed) to the effect that we get no overlap/gap on the EST side, and a 'candidate intron' is made. Multiple possible 'candidate introns' can be formed this way.

- d. If there is a gap on the transcript sequence of more than one base, they go through a 'patching' routine (described below).
- e. If there is a gap on the transcript sequence of 1 base, the pipeline tries to fix this where the sequence data is available making a 'candidate intron'. If the gap could not be sorted out the 'initial intron' is removed.
- f. If there is no gap or overlap on the EST between two successive HSPs and the 'initial intron' shows up as a non-canonical intron type, then we do 'complementary shift' over a 4-base window at both the ends of the intron to see whether a canonical intron can be obtained. The restrictive criteria are that such a process should not lead to a gap or overlap on the EST or a mismatch in a 4-base exonic region around the gap. If this is not possible, then no intron is delineated from this transcript match, otherwise the intron is kept.

Consistency checks of 'candidate introns'

- 1) 10 bases upstream and downstream from the intron, the transcript and gene sequence must not have more than 1 mismatch on either side.
- 2) An intron that passed the above and has one mismatch in either or both the up/downstream regions must be confirmed by the same intron in a different transcript that had a perfect 100% match in both of the regions.

Introns passing this check are called 'confirmed valid introns' and these are the basis for determining our exons.

At this stage we do 'patching' to fill large gaps that may be present on the transcript sequence between HSPs. The gap must be present in both the transcript and in the gene. If a transcript has got a internal gap of >15 bases (or a terminal gap of >30 bases), we do a localised blast via bl2seq between the gap region (with an extra 10 bases on either side of the gap) on the transcript and that on the gene with reduced pid conditions to see if we can find another useful HSP that fills the gap on the transcript. Any intron determined via this method will follow the above procedures and checks.

Our list of 'confirmed valid introns' is used to delineate 'confirmed valid exons'. Between each successive pair of 'confirmed valid introns' is a 'confirmed valid exon' on the additional condition that there may not be a gap on the transcript alignment in between those introns.

The combined list of introns and exons (sorted by transcript, then start location) is used in the next steps to group the transcripts and delineate the events.

2.4. Derivation of transcript classes

2.4.1. Transcript structure

A transcript sequence is retained in the data set only if it has led to delineation of at least one 'confirmed valid intron'. Each of the retained transcript sequence represents a splice pattern of the gene; a splice pattern is represented as a string of exons as encoded in the transcript sequence.

ENSG00000133678

CLASS 1

~3013..3117,6172..6275,6476..6535,14711..14762,15128..~15622

CLASS 2 ~3364..3515, 6172..6275, 6476..6535, 14711..14762, 15128..~15583

CLASS 3 ~2996..3117, 6172..6275, 6476..6535, 15128..~15463

CLASS 4 ~3367..3515, 6172..6275, 6476..6535, 15128..~16882

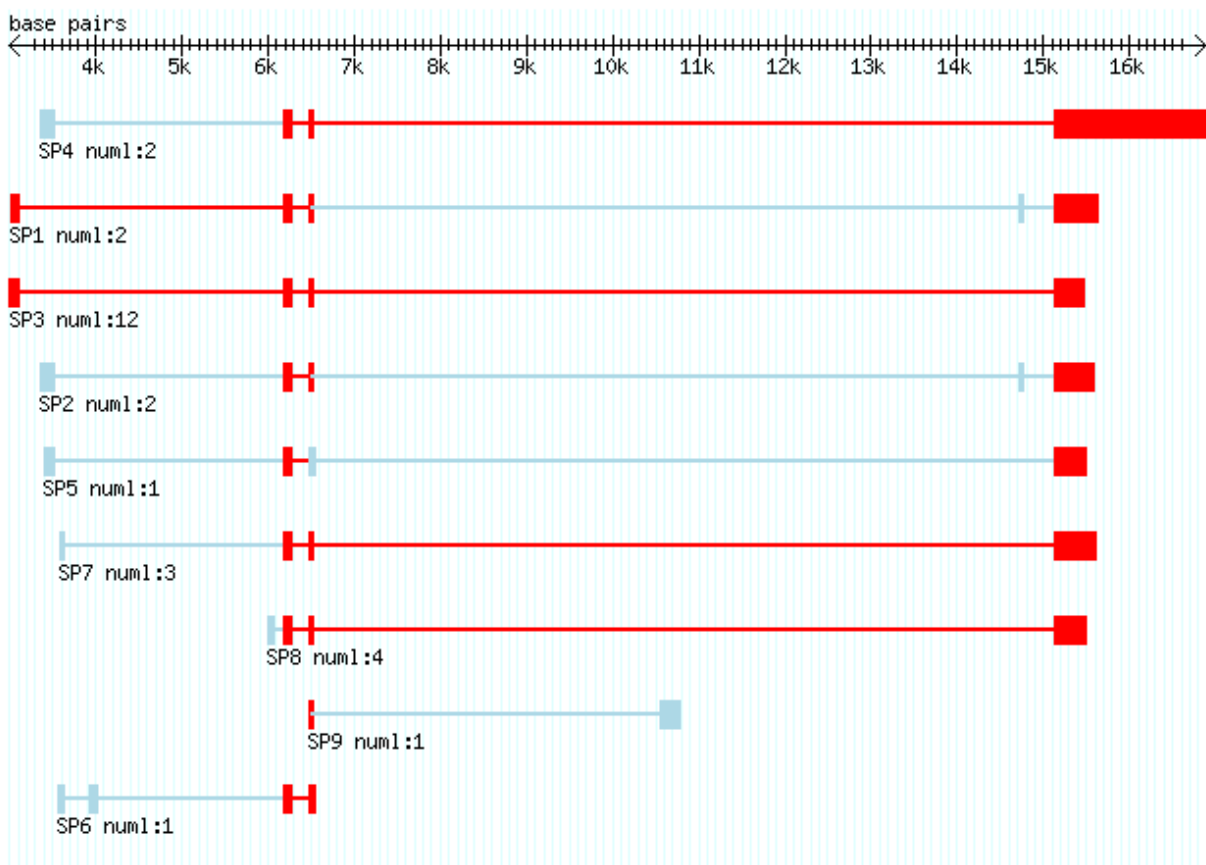
CLASS 5 ~3415..3515, 6172..6275, 6476..6548, 15128..~15486

CLASS 6 ~3576..3638, 3925..4023, 6172..6275, 6476..~6536

CLASS 7 ~3578..3638, 6172..6275, 6476..6535, 15128..~15618

CLASS 8 ~6007..6065, 6172..6275, 6476..6535, 15128..~15489

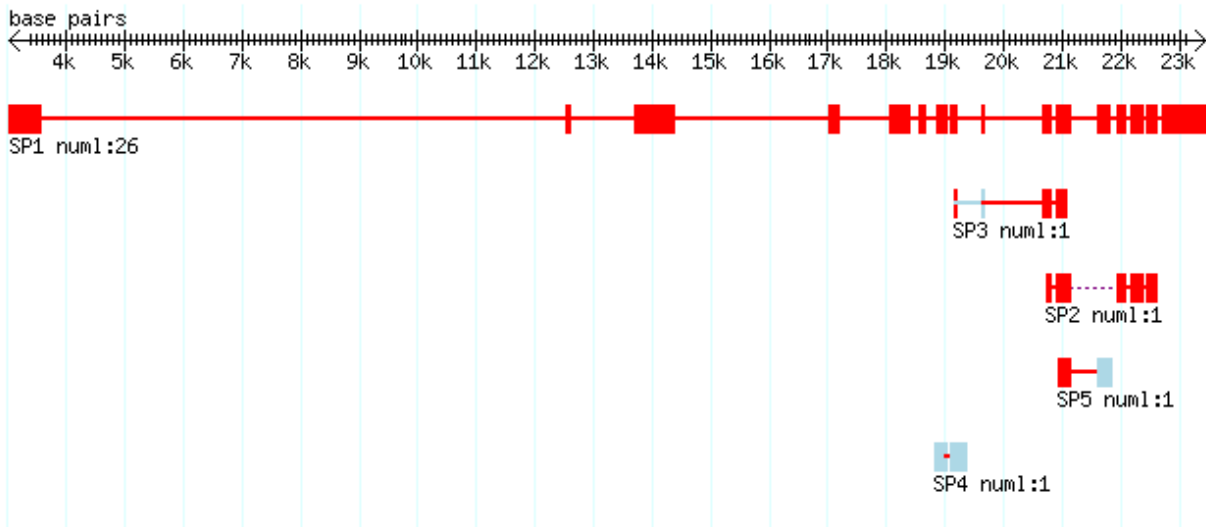
CLASS 9 ~6475..6535, 10562..~10772



Unambiguous definition of 5' boundary of an exon requires that the exon is flanked at its 5' end by a 'confirmed valid intron'. Unambiguous definition of 3' boundary of an exon requires that the exon is flanked at its 3' end by a 'confirmed valid intron'.

The extent of the region defining the terminal exons depends on the length of the EST/mRNA sequence and thus the terminal boundaries of the first and last exon as encoded in the EST/mRNA sequence are thus always ambiguous and are tagged with '~'

(terminal ambiguity). The presence of an initial or candidate intron in the transcript structure requires that the immediate boundary of each of two the flanking exons is tagged with '~' (internal ambiguity) - see the example below:



SP2: ~20729..20801, 20889..~21136, ~21952..22080, 22174..22367, 22455..~22604

If a transcript sequence shows a structure with more than one 'internal ambiguity', then the transcript is removed from the data set.

2.4.2. Derivation of transcript classes

The different transcript sequences that map to a gene are grouped into classes, each class shows a specific splice pattern across a region along the length of the gene. In principle, each class represents a distinct splice pattern.

For each class, the longest of the member transcript sequences (and its splice pattern) is chosen as a representative. Using such representatives, we carry out delineation of alternative splice events.

2.4.3. Ambiguity in the assignment of transcripts

It is possible that certain transcripts, which represent parts of a splice pattern, cannot be assigned unambiguously to a single class. Thus, every transcript is tagged with identifiers of the classes to which the transcript sequence can belong. See example below.

>ENSG00000133678

CLASS 1

BI544962-1 ~3013..3117,6172..6275,6476..6535,

14711..14762,15128..~15622
BI752543-1 ~6475..6535,14711..14762,15128..~15683
BE737009-1,2,3,4,5 ~6172..6275,6476..~6536
BM925342-1,2,7 ~3015..3117,6172..~6278
AK023325-1,2,7 ~3001..3117,6172..~6278
BI524079-1,2,7 ~3035..3117,6172..~6278
BE082643-1,2,3,4,5 ~6186..6275,6476..~6536
BI753011-1,2,7 ~3028..3117,6172..~6278
BI765313-1,2,7 ~3012..3117,6172..~6278

CLASS 2

BG499551-2 ~3001..3117,6172..6275,6476..6535,15128..
~15644
BC022252-2 ~3001..3117,6172..6275,6476..6535,15128..
~16882
BM722685-2 ~3014..3117,6172..6275,6476..6535,15128..
~15515
BM829225-2 ~3022..3117,6172..6275,6476..6535,15128..
~15548
BG568347-2,3,4 ~6172..6275,6476..6535,15128..~15597

CLASS 3

BM786580-3 ~6007..6065,6172..6275,6476..6535,15128..
~15489
BE883009-3 ~6007..6065,6172..~6278

CLASS 4

AL708446-4 ~3578..3638,6172..6275,6476..6535,15128..
~15618

CLASS 5

BG674710-5 ~3368..3515,6172..6275,6476..~6536

CLASS 6

AV687145-6 ~3588..3638,6172..~6278,~6475..6535,15128..
~15236

CLASS 7

AV655687-7 ~3022..3117,6172..~6278,~6475..6535,15128..
~15236

CLASS 8

AA190535-8 ~6475..6535,10562..~10772

2.4.4. Assignment of transcripts that contain ambiguities to classes

2.4.4.1. Internal discontinuity in the transcript structure

These arise because:

- a. there exists an unsorted gap in the transcript sequence between successive HSPs. In these cases, the transcript is assigned to a class if the gap on the gene is comparable in length to that of the definite intron in the representative class with a tolerance of 10 bases.
- b. there exists an unsorted overlap on the transcript sequence between successive HSPs. In these cases, the transcript is assigned to a class if the overlap on the gene is comparable in length to that of the definite intron in the representative class with a tolerance of 2x overlap length.
- c. there exists a non-canonical intron on the transcript sequence between successive HSPs. Treat the same as b. above for unsorted overlap.

If a transcript has got more than one internal ambiguity, it is removed from the data set.

2.4.4.2. End Ambiguities in the transcript structures

These arise when the exterior boundaries of the terminal exons cannot be defined unambiguously due the ending of transcript sequences.

Comparing two transcript structures forms:

- a. If the ends of both the exons under comparison are both ambiguous, the ends are considered to be similar.
- b. If the end of the exon from one transcript is definite and from the other transcript is ambiguous, then the following criteria are used
 - (i) If the ambiguous end is extending the other exon by ≥ 25 bases, then the ends are considered to be dissimilar.

$(5000..6000) = (\sim 4981..6000)$

$(5000..6000) \neq (\sim 4900..6000)$

$(5000..6000) \neq (5000..\sim 6034)$

- (ii) If the ambiguous end is shorter than the definite end, then they are considered to be similar.

Examples:

(5000..6000) = (~5104..6000)
(5000..6000) = (5000..~5900)

2.4.4.3. Summary: forming of classes and class ambiguity assignment

The number of introns sorts the transcript structures. The transcript with the highest number of introns is used to seed the first class. Any transcript structure that follows is shorter or equal in length with regard to number of introns. This is important since the exact intron positions are used as the main indicator for whether or not to assign a transcript structure to a class. If all introns from a structure match the introns from a class representative a transcript class could be formed taking into account the following rules.

Rule 1. Flanking ambiguous 'exon' extension

The flanking ambiguous 'exons' should not extend past a well-defined structure in the other class by more than 25 bases. If they do the current transcript structure cannot fit in this class.

Rule 2. No full intron inclusion in exon

None of the exon in the representative transcript should include a full intron from any other transcript of the class. If it does the current transcript structure cannot fit in this class.

Rule 3. Internal ambiguities extension check

a. Overlap case:

The transcript is assigned to a class if the alignment gap on the gene is comparable in length to that of a confirmed intron in the representative transcript structure, with a tolerance of 2 x overlap length.

b. Gap cases:

The transcript is assigned to a class if the alignment gap on the gene is comparable in length to that of a confirmed intron in the representative transcript structure, with a tolerance of 10 bases.

c. Non-canonical intron; same as b above.

Rule 4. Internal ambiguity occupancy check

Whenever there is an internal ambiguity we need to make sure that 1. at the place of the internal ambiguity the other transcript doesn't have more than one intron

2. no full 'confirmed valid exon' is covering the whole of the ambiguity.

If both sub rules pass, then we can add the transcript to the class.

Note: Rule 3 and Rule 4 are only performed in case there is an internal ambiguity in either (or both) the transcript structure to be classified or the representative of the current class.

2.4.4.4. Defining the quality of a class

If a class has all its members having an internal discontinuity add an 'amb' tag to the class number.

2.4.4.5. Defining the relationship between classes

An upper triangular matrix can be created for this purpose.

Class a with class b =

D - represents distinct regions on the gene

S - staggering overlap (overlaps over a region with same structure)

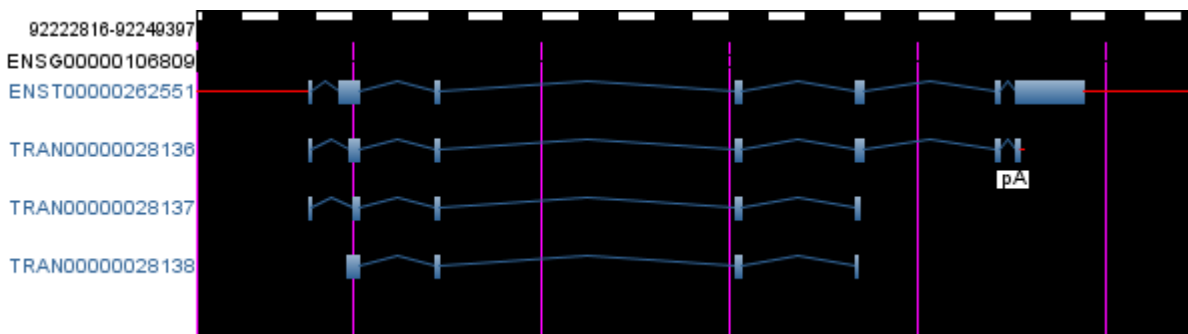
A - overlapping regions but with different structures

2.5. Delineation of events

We delineate the following types of alternative events.

a. Intron Isoforms: These are delineated by examining a pair of splice patterns wherein an intron from one has undergone modifications at either or both of its boundaries in the other splice pattern. The outer boundaries of the flanking exons can have ambiguous ends.

Notations: II-5p, II-3p, II-5p3p. 5p indicates modification at the 5' end of the intron; 3p indicates modification at the 3' end.



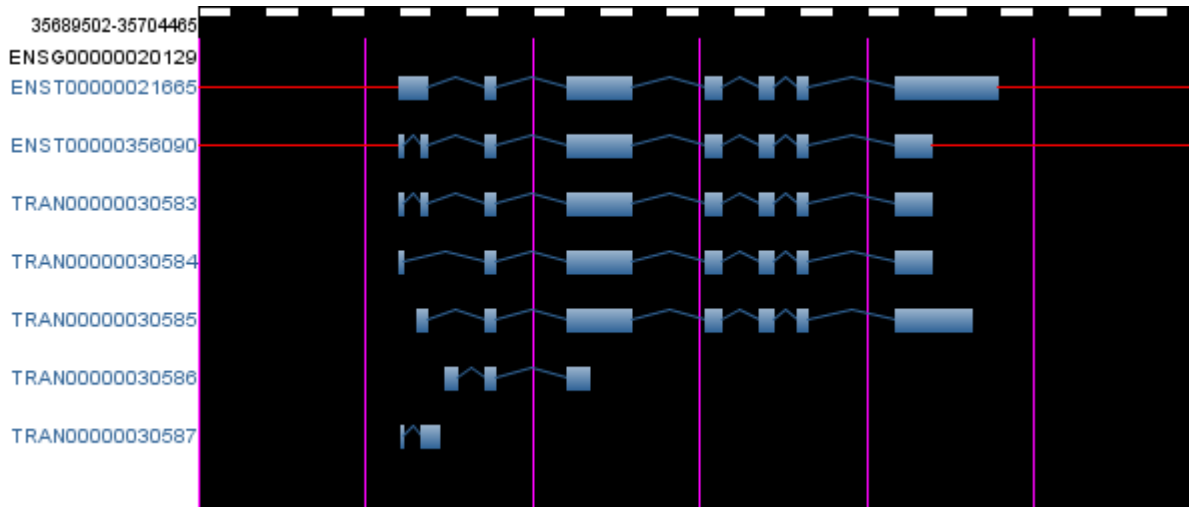
For above splice graph:

II-3p TRAN00000028136/TRAN00000028137

EI-5p TRAN00000028136/TRAN00000028137

b. Exon Isoforms: These are delineated by examining a pair of splice patterns wherein an exon from one has undergone modifications at either or both of its boundaries in the other splice pattern. The requirement is that both the boundaries of the exons are unambiguously defined in the representative splice patterns.

Notations: EI-5p, EI-3p, EI-5p3p. 5p indicates modification at the 5' end of the exon; 3p indicates modification at the 3' end.



For above splice graph:

II-5P TRAN00000030584/TRAN00000030585

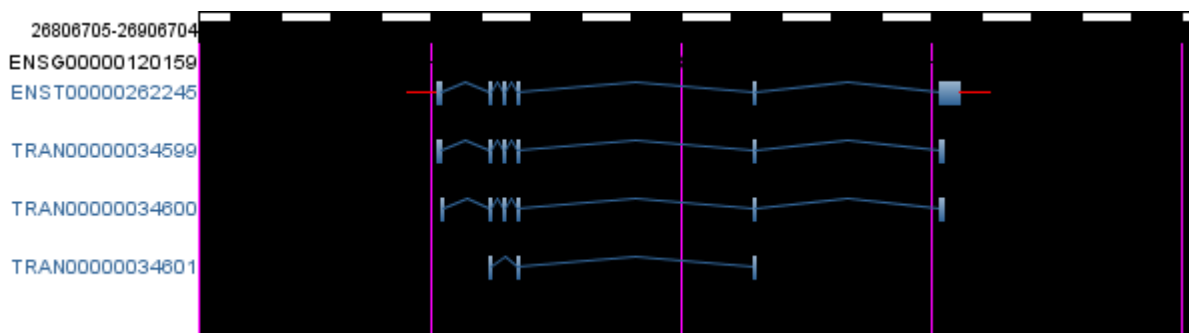
SCE TRAN00000030583/TRAN00000030584

c. Cassette Exons: These are delineated by examining a pair of splice patterns wherein an exon (or more than one consecutive exon) is present in one splice pattern and absent in the other. The cassette exon event can be termed as either a 'skipped exon event' or 'cryptic exon event' depending on which of the two isoform splice pattern is considered as constitutive. An exon that is present in a constitutive splice pattern but absent in the alternative pattern is called a skipped exon event; and an exon that is present in the alternative splice pattern but absent in the constitutive splice pattern is called a cryptic exon event.

Notation:

SCE - simple cassette exon event. The flanking exons do not undergo modifications.

CCE-IEB-5p3p - complex cassette exon event accompanied by modifications at the 5' and 3' boundaries of the 5' and 3' flanking exons.



For above splice graph:

SCE TRAN00000034599/TRAN00000034601

SCE TRAN00000034600/TRAN00000034601

II-5P TRAN00000034599/TRAN00000034600

The second component in the notation can take the form of IB or EB or IEB depending on which boundary of a flanking exon undergo modification - IB indicating the interior boundary (3' end of 5' flanking exon; 5' end of 3' flanking exon); EB indicating the exterior boundary (5' end of 5' flanking exon; 3' end of 3' flanking exon).

The third component in the notation can take the form 5p, 3p, or 5p depending on which of the flanking exons undergo modification.

d. Mutually Exclusive Exons: These are delineated by examining a pair of splice patterns where an exon present in one splice pattern is not present in the other, and vice versa.

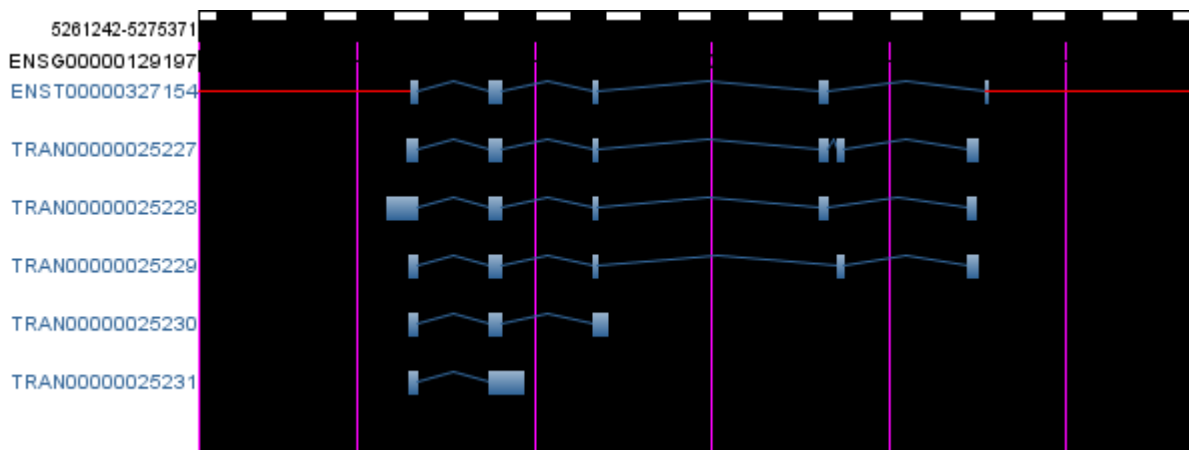
Notation:

SME - simple mutually exclusive exon event. The flanking exons do not undergo modifications.

CME-IEB-5p3p - complex mutually exclusive exon event accompanied by modifications at the 5' and 3' boundaries of the 5' and 3' flanking exons.

The second component in the notation can take the form of IB or EB or IEB depending on which boundary of a flanking exon undergo modification - IB indicating the interior boundary (3' end of 5' flanking exon; 5' end of 3' flanking exon); EB indicating the exterior boundary (5' end of 5' flanking exon; 3' end of 3' flanking exon).

The third component in the notation can take the form 5p, 3p, or 5p depending on which of the flanking exons undergo modification.



For above splice graph:

SME TRAN00000025228/TRAN00000025229

SCE TRAN00000025227/TRAN00000025229

SCE TRAN00000025227/TRAN00000025228

e. Intron Retention: These are delineated by examining a pair of splice patterns wherein an intron from one splice pattern is completely contained within an exon of the other.

Notation:

SIR - simple intron retention event. The exons that flank the retained intron do not undergo modifications.

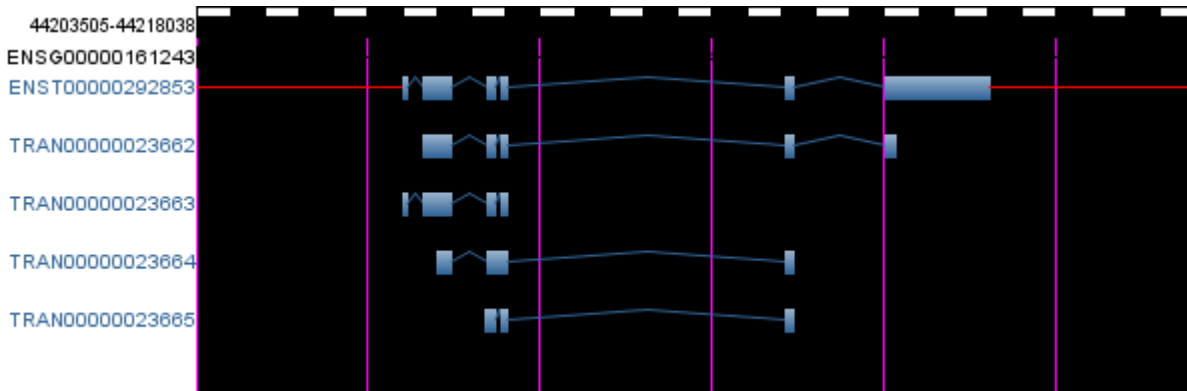
CIR-EB-5p3p - complex intron retention event accompanied by modifications at the exterior boundaries of the 5' and 3' flanking exons.

The second component in the notation can take the only form of EB.

The third component in the notation can take the form 5p, 3p, or 5p depending on which of the flanking exons undergo modification.

CIR-CE-5p or CIR-CE-3p - complex intron retention event accompanied by skipping of 5p or 3p (as the case may be) exon that flanks the retained intron.

CIR-ME-5p - complex intron retention event accompanied by the skipping of the 5p flanking (the retained intron) exon and the appearance of a cryptic exon at 5p or 3p



For above splice graph:

SCE TRAN00000023662/TRAN00000023664

SCE TRAN00000023663/TRAN00000023664

CIR-EB-5P TRAN00000023664/TRAN00000023665

f. Simple versus Complex events: The above events 3-5 are further characterized as either 'Complex' or 'Simple' depending on whether the event accompanies modifications to either of (or both) the boundaries of either of (or both) the flanking exons or not. An ambiguously defined boundary for the flanking exon is also treated as modification for the above classification purpose if the modification exceeds 25 bases.

In all the above cases, a boundary of an exon is said to undergo modification if it is defined unambiguously in both the transcripts and if it shows difference. In situations where the boundary is unambiguously defined in one transcript but not in the other, then we follow the following criteria. The unambiguously defined position should extend the exon by at least 25 bases. The choice of 25 was made by taking into account the allowed

pid of 95% for alignments and that mismatches usually occur at the end of the HSPs.

e. Incomplete Events: The above events 3-5 are classified as Incomplete, when the nucleotide region encompassing all the component introns defining the event in a splice pattern is not completely defined in the other splice pattern. Thus, it is to be understood that the basic event is not completely defined.

3. ATD

3.1. Generating transcript patterns

AltSplice splice patterns and the gene-transcript alignments form the basis of AltTrans pipeline that delineates alternate transcript patterns. Each of the gene-transcript alignments confirming a splice pattern from AltSplice is examined for the presence of a transcription start site and poly(A) site that initiate and terminate the transcript sequence, respectively.

3.2. Detecting transcription start sites

Determination of the mRNA start site is of key importance for the identification of Transcription Start Site [TSS] and transcriptional regulation of gene expression. Experimental knowledge of the precise 5' ends of cDNAs should facilitate the identification and characterization of TSSs.

As a first experimental step in this direction, Suzuki et al. used the oligo-capping method to identify TSSs from the enriched full length human cDNA libraries which they have made available to the public. The 'oligo-capping' method enabled the introduction of a sequence tag to the first base of an mRNA by replacing the cap structure of the mRNA. Using cDNA libraries made from oligo-capped mRNAs, we could identify the transcriptional start site of an individual transcript.

For the mouse genome, the findings of the FANTOM3/Genome Network project have redefined the mouse transcriptome by introducing an extensive collection of novel cDNAs and millions of sequenced tags corresponding to the 5' end mRNAs. For mouse, we used RIKEN 5' ESTs rather than RIKEN full-length cDNAs since 5' ESTs have more libraries sequenced and have a better coverage. Although many TSSs from mouse can be inferred from the 5' ends of full-length cDNAs, the depth of coverage is limited. To increase the depth of coverage, we will use transcripts from the cap analysis of gene expression (CAGE) in our next release.

3.2.1. Identification of TSS

For human, the oligo-capping method was used to generate 1.4 million 500 nucleotide sequences of the 5' end of full-length cDNAs. The cDNAs were collected from libraries constructed from 164 kinds of human tissues and cultured cells using the cap selection method. These sequences were generated focusing on TSS and splicing variations.

For Mouse, we used 195461 unique 5' -end one-pass sequences from RIKEN full length cDNA libraries.

From the transcript patterns, we generated the sequence of the first exon [FE] and upstream UTR for identification of TSSs. These were BLASTed against the putative TSS dataset: 5' -ends of the oligo-capped human cDNAs and 5' -end one-pass mouse sequences.

Criteria for determining a TSS from the unique and exact mapping of the UTR-FE against

the putative TSS dataset are:

- (i) to rule out vector contamination the first 6 bp of alignment are discarded
- (ii) it has a 95% pid match from the alignment;
- (iii) the 95% pid has to include the first 100 base pairs of the alignment.

3.2.2. Mapping of TSS

For every UTR-FE alignment, all such putative TSS positions are identified. Where there is more than one match, the first identified 5' end will only be considered, thereby defining the longest transcript. The TSS position is converted to the gene position using the coordinates of the first exon of the splice pattern.

Before generating the final TSS data set, we excluded those cDNA sequences whose start has been mapped inside an exon. This was intended to minimize dubious identification of TSSs due to erroneously cloned truncated cDNAs. Erroneous oligo-capping of an immaturely spliced transcript generates these truncated cDNAs. We presumed that those cDNAs whose 5'-ends were located outside of the exonic regions were not truncated forms of any known types of transcripts. Although it is relatively minor, "oligo-cap" cDNA collection does include a certain population of those erroneously cloned species. Using the above criteria, we consider that most of the selected 5'-ends of the cDNAs should correspond to actual TSSs, reducing the false-positive identification of TSSs.

3.3. Detecting poly(A) sites

3.3.1. Detecting poly(A) tail and poly(A) cleavage (PAC) site

Each of the gene-transcript alignments is examined for the presence of a 3' dangling end on the transcript sequence. Only those alignments that show 3' dangling ends of length at least 8 bases are considered further. The transcript region -5 to +5 from the end of alignment is examined for the start of a poly(A) tail: a string of 8 or more adenosines. We do want to include transcript patterns involving those gene-EST alignments with long dangling ends (artefacts in EST sequences or of genomic 'contaminations'). Consequently, the ratios between the length of the dangling ends and minimum length of the poly(A) tail are as follows:

- (a) If dangling end is less than or equal to 50 bp = poly(A) is 8bp
- (b) More than 50 and less than or equal to 100 = poly(A) is 10bp
- (c) More than 100 and less than or equal to 150 = poly(A) is 15bp
- (d) More than 150 bp = poly(A) is 20 bp

Since it is often the case that non-adenosine bases interrupt a run of adenosines, we allow mismatches to a maximum of 10% of the positions in the identified string. If more than one poly(A) tail is identified starting in the -5 to +5 region, the one with the highest

composition of adenosines is chosen as the authentic poly(A) tail. The gene position corresponding to the start of poly(A) tail is considered as the cleavage site.

3.3.2. Detecting poly(A) signal (PAS)

The gene sequence 40 nucleotides 5' to the identified cleavage site is scanned for the presence of one of the 13 variant poly(A) signals (namely, AAUAAA, AUUAAA, UAUAAA, AGUAAA, AAGAAA, AAUAUA, AAUACA, CAUAAA, GAUAAA, AAUGAA, UUUAAA, ACUAAA and AAUAGA) with no mismatch allowed. For every gene-transcript alignment, all such motifs are identified. If multiple matches exist a representative motif is chosen by a combination of rank and position.

3.3.3. AltPAS

3' EST sequences and full-length cDNA sequences are obtained from transcript resources such as dbEST, H-Inv, and FANTOM and are aligned to the repeat-masked genome using the MegaBlast program. High scoring matches clustered in a manner that transcript members from a cluster have their end positions located within a range of 10 nucleotides from each other. Each cluster is then analyzed using a sliding 10-nt window to locate the most likely cleavage site, defined as the position where the window contains the ends of most transcripts. Alignment hits with more than 5 unmatched positions at cleavage site are discarded. Cleavage sites that are flanked by A-rich region (at least 9 out of 10 nt positions are adenosines) in the 50 nt downstream genomic sequence, and those that do not contain one of the known poly(A) signals in the 30 nt upstream region are discarded. Of the remaining cleavage sites, only those that are supported by at least two transcript sequences are retained as potential poly(A) sites. The poly(A) sites, thus identified, are denoted using genome coordinates. Assignment of the detected poly(A) sites to Ensembl genes is carried out as below: A poly(A) site is assigned to the Ensembl gene to which the site's genome location can be mapped; if the genome location of a poly(A) site does not map to any annotated gene, it is assigned to the nearest 5' gene, provided that the distance to gene is less than 3000 bases.

3.3.4 Choosing a representative cleavage site as poly(A) site

It is often the case that some of the identified cleavage sites are close to one another. Given that a poly(A) site can harbour multiple cleavage sites and also that errors in sequences can lead to small differences in the locations of identified cleavage sites, it is possible that the adjacent sites are not distinct poly(A) sites. The 5' most site from a cluster of sites is chosen as the representative poly(A) site for that group of transcripts.

3.3.5. Integrating poly(A) sites from the AltPAS and AltTrans pipelines

Adjacent poly(A) sites are grouped and a representative poly(A) site is chosen from each group (as mentioned above) with the following variation. If a group contains sites from both AltTrans and AltPAS, the 5' most AltTrans site is chosen as the representative. Such a set of representative poly(A) sites is subsequently used to annotate a gene with all

potential poly(A) sites, and to annotate a transcript pattern for potentially "skipped" poly(A) sites.

4. AEdb

AEdb is a comprehensive database of alternative exons from the literature. Most alternative exons are cassette exons and are expressed in more than two tissues. Of all exons whose expression was reported to be specific for a certain tissue, the majority were expressed in the brain.

Statistical analysis of the dataset shows the length of constitutive exons follows a normal distribution; the distribution of alternative exons is skewed toward smaller ones.

Furthermore, alternative-exon splice sites deviate more from the consensus: their 3' splice sites are characterized by a higher purine content in the polypyrimidine stretch, and their 5' splice sites deviate from the consensus sequence mostly at the +4 and +5 positions. For exons expressed in a single tissue, adenosine is more frequently used at the -3 position of the 3' splice site. In addition to the known AC-rich and purine-rich exonic sequence elements, sequence comparison using a Gibbs algorithm identified several motifs in exons surrounded by weak splice sites and in tissue-specific exons. Together, these data indicate a combinatorial effect of weak splice sites, atypical nucleotide usage at certain positions, and functional enhancers as an important contribution to alternative-exon regulation.

There are five datasets that can be queried or downloaded.

AEdb-Sequence - the sequence and properties of alternatively spliced exons

AEdb-Function - data on functional aspects of alternatively spliced exons

AEdb-Motif - data and sequence of known splice regulatory motifs, ESE, ESS, ISE, ISS

AEdb-Minigene - a collection of known minigene constructs for alternative splice events

AEdb-Disease - diseases associated with splicing

5. Mapping and analysis

5.1. SNP-mediated alternative splicing

There are at least two factors that modulate alternative splice site selection

- (i) relative concentration of splicing-associated proteins
- (ii) differences in the nucleotide compositions at the splice signals (such as donor or acceptor sites, BPS, PPT or splice enhancer/silencer sequences) associated with alternative splice sites.

It is well documented in the literature that nucleotide changes in sequences associated with splice signals can cause abolition of the use of a site and can instead promote use of an alternative splice site. Sequence variation, either through heredity or through acquired mutations, among individuals of a species can regulate selection of alternative splice sites. This leads to formation of isoform transcripts between two individuals carrying different polymorphic version of the gene. Resources, such as EST sequences, can indeed consist of isoform transcripts that are generated from more than one polymorphic versions of genes.

EST/mRNA data resources are often used to identify isoform splice events that a gene can undergo. It is of interest to see whether the delineated transcript isoforms are resultant of the 'cellular code' or whether they are arising from multiple polymorphic versions of the gene.

5.1.1. Identification of SNP positions in the genes

Human HGVbase, which is a collection of predominantly SNP's, was used to identify SNP positions in each of the genes that showed two or more distinct splice pattern isoforms. SNP records include polymorphisms (sequence variations in which the most abundant allele has a frequency of $< 99\%$) as well as variations with rare or single occurrence alleles, plus disease related and disease-causing clinical mutations. HGVbase lists 25-base regions upstream and downstream of the SNP position. dbSNP lists much longer sequences. The first part of the method is to identify hits from the HGVbase entries and the genes from the data set (criteria are: pid for each of the 25-base region matches should be more than 96%; a total of at least 48-base match region). Once the HGVbase hits have been identified, a further blast is done with the corresponding dbSNP entries (with criteria of 95% pid and match length of at least 50% of the dbSNP length or >200 bases).

5.1.2. Identification of exonic SNP positions from two or more isoform splice patterns

Considered further are only those SNP positions that occurred in the exonic regions from two or more distinct splice patterns corresponding to a gene. Two distinct splice patterns that contain a SNP position are isoforms to one another.

5.1.3. Identification of alleles used at the exonic SNP positions

At each of the SNP positions, the alleles used in the supporting transcript sequences from each of the splice patterns (of a gene) were identified. In this manner, we can associate the isoform splice patterns with the polymorphicity of the gene. If the transcript sequence showed any mismatch with the gene in the 5-base regions on either side of the SNP position, the transcript sequence was not considered further unless the mismatch position is a nearby SNP position. The above work has been carried out also at the events level, where two sets of transcript sequences are generated for every observed alternative events - each set demonstrating one of the two isoforms of an event.

5.2. Expression states

The transcripts that confirm the splice patterns are mostly EST sequences; it is the case that around only 10% of the transcript sequences are mRNA sequences. EST libraries are fairly well annotated towards the three expression states (tissue type, development stage, and disease state). eVOC provides controlled vocabularies for defining the expression states and such vocabularies have been mapped to EST libraries (such a controlled vocabulary and mapping is available publicly for human). Some of the mRNA sequences are annotated in EMBL data distribution for clone_library name; such mRNA sequences can be associated with the annotation of expression state provided their clone_library names are also seen associated with an EST library.

A look-up table was created giving association of EST/mRNA accession number, EST library Id, and Clone_library name. For each of the EST library, we have short-listed specific terms (at three hierarchical levels) using the eVOC mapping. Such specific terms are also added to the look-up table.

Transcript sequences that confirm each of the observed splice patterns in AltSplice are grouped as per clone_library names and these groups qualify every splice pattern. The look-up table mentioned in the previous paragraph was used to derive the expression states (for each of tissue type, development stage, disease state) at three hierarchical levels and these derived states qualify each of the observed splice patterns.

6. Integration of AltSplice and AEdb-Sequence databases.

While AltSplice contains computationally derived data on alternative splice events, AEdb contains experimentally known data on alternatively spliced exons. Thus AEdb can provide experimental validation for the events in AltSplice.

These two databases have different content and different format. AEdb contains the nucleotide sequences of the alternatively spliced exon and of the flanking constitutive exons. These sequences are blasted against the exon sequences in AltSplice. An AEdb entry is associated with an AltSplice entry if the sequences of the three exons (alternatively spliced exon and the two flanking constitutive exons) match against the sequences of the exons from the AltSplice entry. By this process, the entries from AltSplice and AEdb, the corresponding exons, and the corresponding splice events are identified and are annotated so & hyperlinked in both the databases. Web query interfaces allow retrieving this different information.

7. Workbench

The workbench provides a set of online tools that enable users to carry out analysis of pre-mRNA sequences. It includes tools for intron analysis, scoring ATG-context sequence, finding exons and identifying splicing regulatory sequences. These tools are accessed either through a single wrapper interface or through interfaces that are specialised for individual tools.

7.1. Intron analysis

The tool examines intron sequences (as provided by the user) for putative branch point (BP) sites and polypyrimidine tracts (PPT). It further calculates the strength of the donor and acceptor sites. The user has a choice of weight matrices for donor and acceptor sites tailored for different intron types, such as U2-type *GT-AG* and *GC-AG* and U12-type *GT-AG* and *AT-AC*.

7.2. Scoring ATG-context sequence

This tool examines each occurrence of *ATG* in a given transcript sequence for its ability to act as translation start codon. Each *ATG* is scored for Kozak's *ATG*-context sequence using a weight matrix that we built from experimentally confirmed translation initiation sites. The sequence of translated peptide from each occurrence of *ATG* is presented along with the *ATG*-context score. FASTA/BLAST searches against UniProt sequence data can be launched for each of the translated peptide sequence.

7.3. MZEF-SPC exon finder

This tool identifies potential exons in a given nucleotide sequence. It integrates Michael Zhang's Exon Finder and Thanaraj's SpliceProximalCheck. MZEF identifies putative exons using quadratic discriminant analysis. SPC is a decision tree implementation of splicing signals that differentiate genuine human splicing sites from the proximal false sites and thus specialises in validating the predicted exon boundary for exactness.

7.4. Detection of short regulatory sequences

Exons are regulated by short, degenerative sequences that bind to interacting splicing factors and proteins. These sequences are collected in the AEdb-Motif database. The Regulatory Sequence tool uses these motifs to examine a given nucleotide sequence for their presence. Users have a choice to specify the extent of allowed mismatches. The identified motifs are hyperlinked to the corresponding entries in AEdb-Motif database. The splicing rainbow is a visualization tool that colour-codes presence of different regulatory motifs in a user-supplied sequence.