BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 14 2007, pages 1815–1823
doi:10.1093/bioinformatics/btm084

Databases and ontologies

EuSplice: a unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes

Ashwini Bhasi[1], Ram Vinay Pandey[2], Suriya Prabha Utharasamy[2] and Periannan Senapathy[1,*]

[1]Department of Human Genetics, Genome Technologies, Inc., 8000 Excelsior Drive, Madison, WI 53717, USA and [2]Department of Bioinformatics, International Center for Advanced Genomics and Proteomics, 383 1st Cross Street, Nehru Nagar, Chennai 600096, India

Received on November 17, 2006; revised on January 20, 2007; accepted on March 1, 2007

Advance Access publication March 7, 2007

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Despite increased availability of genome annotation data, a comprehensive resource for in-depth analysis of splice signal distributions and alternative splicing (AS) patterns in eukaryote genomes is still lacking. To meet this need, we have developed EuSplice—a unique splice-centric database which provides reliable splice signal and AS information for 23 eukaryotes.

Results: The EuSplice database contains 95 822 AS events and 2.1 million splice signals associated with over 270 000 protein-coding genes. The intuitive, user-friendly EuSplice web interface has powerful data mining and graphics capabilities for inter-genomic comparative analysis of splice signals, putative cryptic splice sites and AS events. Moreover, the seamless integration of splicing data to extensive gene-specific annotations, such as homolog annotations, functional information, mutations and sequence details makes EuSplice a powerful one-stop information resource for investigating the molecular mechanisms of complex splicing events, disease associations and the evolution of splicing in eukaryotes.

Availability: http://66.170.16.154/EuSplice

Contact: ps@genome.com

Supplementary information: Supplementary tables and figures at Bioinfo online.

1 INTRODUCTION

There is a growing need in biological research for data integration methodologies which efficiently exploit the wealth of information generated from whole genome sequencing projects. This need is particularly prominent in the rapidly advancing field of research examining the complex processes of mRNA splicing (Balvay *et al.*, 1993; Nilsen, 2003; Staley and Guthrie, 1998). The donor ((C|A) A G G T (A|G) A G T) and acceptor (Y_{10} N (C|T) A G (G|A)) splice signals at exon–intron junctions are key players in splicing events. A convergence of studies has established that these splice signals are uniformly observed in eukaryotes (Burset *et al.*, 2000; Mount, 1982; Senapathy *et al.*, 1990; Shapiro and Senapathy, 1987). The recognition of alternate splice signals at exon–intron junctions generates alternative splicing (AS) events, which ultimately produce multiple mRNA isoforms from the same gene template. The mechanism of exon–intron boundary detection by the spliceosome (Burge *et al.*, 1999), the effects of AS events on gene expression (Stetefeld and Ruegg, 2005), and the role of defective splicing in disease manifestations (Baralle and Baralle, 2005; Buratti *et al.*, 2006) are being intensely researched.

A unified and reliable database of splice signals and AS information that is integrally linked to publicly available gene annotations would be extremely useful to researchers. Currently, none of the available splice signal databases (Burset *et al.*, 2001; Chong *et al.*, 2004; Sheth *et al.*, 2006) provide information on the different AS event types (exon skipping, intron retention, cryptic exons, etc.). Although AS events in some well-studied eukaryotes have been deposited in various databases (Holste *et al.*, 2006; Huang *et al.*, 2005; Nagasaki *et al.*, 2006; Ratsch *et al.*, 2005; Stamm *et al.*, 2006), their annotations differ considerably between these databases, largely due to variations among the datasets and methodologies employed (Bonizzoni *et al.*, 2006). Moreover, the currently available splice signal and AS resources are limited to a few genomes and do not contain information on fungal genomes or newly sequenced mammalian genomes. Thus, investigators must perform numerous tedious maneuvers and browse multiple databases to gather information on a gene's AS events, splice signals associated with mRNA isoforms, functional annotations, disease associations, homolog annotations, polymorphisms and other related information. A great deal of time and effort could be saved if all of this information were available at one location in a comprehensive, user-friendly format.

We have developed EuSplice (http://66.170.16.154/EuSplice), a unique splice-centric information resource which provides reliable splicing information for protein-coding genes of the following 23 eukaryote genomes (12 higher eukaryotes and 11 fungi): *Homo sapiens, Pan troglodytes, Bos taurus, Canis familiaris, Mus musculus, Rattus norvegicus, Gallus gallus, Danio rerio, Drosophila melanogaster,*

[*]To whom correspondence should be addressed.

Apis mellifera, Caenorhabditis elegans, Arabidopsis thaliana, Aspergillus fumigatus, Candida albicans, Candida glabrata, Cryptococcus neoformans, Debaryomyces hansenii, Encephalitozoon cuniculi, Eremothecium gossypii, Kluyveromyces lactis, Schizosaccharomyces cerevisiae, Schizosacharomyces pombae and *Yarrowia lipolytica*. EuSplice provides complete splicing information for 277 399 protein-coding genes. It has a powerful web-based graphical user interface for efficient visualization and analysis of gene sequence and structure, splice signals, AS events, alternative transcription initiation (ATI), alternative transcription termination (ATT), functional annotations and disease associations as well as inter-genomic comparative analysis of these features. The user-friendly query options that support intelligent querying and information retrieval together with the extensive external database cross-linking make EuSplice an invaluable resource for researchers studying eukaryotic splice signals and AS events.

2 METHODS

2.1 Database design

EuSplice is a relational database which we developed in MySQL 4.1 (www.mysql.com) using a unique data extraction, validation and integration process (Fig. 1). It contains sequence and annotation data that were extracted from the following six source databases: RefSeq genome assemblies(ftp.ncbi.nih.gov/genomes), Entrez Gene (ftp.ncbi.nih.gov/gene), Homologene (ftp.ncbi.nih.gov/pub/HomoloGene), OMIM (ftp.ncbi.nih.gov/repository/OMIM), UTR db (ftp.ebi.ac.uk/pub/databases/UTR/) and IPI (ftp.ebi.ac.uk/pub/databases/IPI). The unique database design preserves the inherent independence of data extracted from various parent databases, while supporting query-based data integration. This design enables EuSplice to efficiently cope with periodic non-synchronized updates of the source databases. We will perform regular updates to the EuSplice database using automated programs.

2.2 Data extraction and database population

The sequences and annotations of the 23 eukaryote genomes were downloaded from ftp.ncbi.nih.gov/genomes on 21 July 2006. The following information was extracted for all genes: (1) co-ordinates of the gene, mRNA and coding sequences (CDSs), (2) Entrez Gene ID, (3) RefSeq mRNA ID, (4) RefSeq protein ID, (5) RefSeq contig ID, (6) chromosome number and (7) strand. The relationship of each gene to its multiple mRNA elements was represented as a structured set of relationships in XML (Extensible Markup Language) format. The creation of these XML gene indices involved extensive filtering of genome assembly annotations and multiple quality checks as described elsewhere (A.Bhasi, P.Philip, R.V.Pandey, S.Sahu and P.Senapathy, in preparation).

Gene, exon, intron, untranslated regions (UTRs) and splice signal sequences were extracted from the Refseq genome sequence files using the XML gene indices created for each genome. Parsers written in Perl, which utilize the Bioperl modules (www.bioperl.org) were developed for efficient data extraction. The XML gene index files were also used to extract gene-specific information from Entrez Gene, IPI, Homologene, OMIM and UTRdb. The multi-faceted data extracted from each of these resources were independently populated into the EuSplice database. A detailed schema for the data-extraction procedure is given in Supplementary Figure 1.

2.3 Web interface Development

The EuSplice web interface was developed in CGI/Perl (www.perl.org) and runs on an Apache 2.0.53 (www.apache.org) web server. It performs intelligent queries to the back-end mySQL database using the Perl:DBI module and provides the user with integrated information on the gene of interest. The graphical display in the web interface was implemented using GD.pm (http://search.cpan.org/dist/GD/GD.pm).

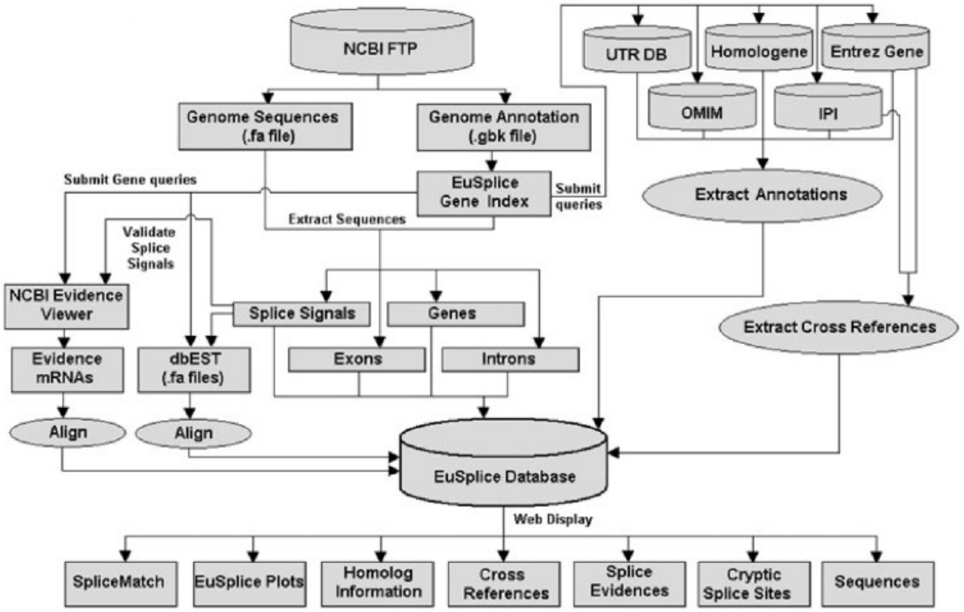


Fig. 1. Data flow diagram for EuSplice database creation.

2.4 Position weight matrices and Shapiro–Senapathy splice scores

Position weight matrices (PWMs), which are essentially the percentages of each nucleotide within the 15 acceptor splice signal positions and the 9 donor splice junction positions, were calculated for each of the 12 higher eukaryotes. A single PWM for fungi was calculated using the same methodology with a comprehensive set of splice sites compiled from all 11 fungi. Splice scores, which provide a reliable index of the conformation of each splice sequence to the expected splice consensus sequence, were calculated using the PWM as described in detail previously (Shapiro and Senapathy, 1987). A splice signal that fully conforms to the splice consensus sequence is given a score of 100, and the score decreases relative to the extent by which a splice signal varies from the consensus sequence. Putative cryptic splice sites around splice junctions were also detected using the splice scoring technique.

2.5 Classification of alternative transcription events

The exon-specific position variation data associated with transcript variant annotations in RefSeq were identified and categorized using a rule-based classification procedure (A.Bhasi, R.V.Pandey, P.Philip and P.Senapathy, in preparation). First, the exons from the various isoforms of a gene were categorized into specific alternative exon clusters based on exon-specific position co-ordinates (start and end positions) within the gene. Next, the various classes of exon-specific position variations were identified within each exon cluster. This classification procedure enabled us to identify six types of AS events and two types of alternative transcription events (ATI and ATT) in the Refseq mRNA transcripts of the 10 higher eukaryotes and 4 fungi.

2.6 Validation of splice junctions in EuSplice

Two independent methodologies were applied in order to confirm the authenticity of splice junctions and associated splice signals in the EuSplice database. The first methodology aligned splice junctions to expressed sequence tag (EST) sequences (Burset *et al*., 2000; Thanaraj, 1999); and the second aligned splice junctions to NCBI evidence mRNA sequences (Chong *et al*., 2004). Many splice junctions were doubly validated, which resulted in their increased reliability.

Rather than the traditional methodology of EST-based splice junction validation (Burset *et al*., 2000; Thanaraj, 1999) which employs a general eukaryote EST database, here we used genome-specific EST databases. The EST sequences for each of the 12 higher eukaryote genomes were extracted separately from dbEST (Boguski *et al*., 1993) to create genome-specific EST databases. For the 11 fungi, a single fungal EST database was created. This tactic of using specific EST databases increased the sensitivity and reliability of splice junction validation and thus helped to decrease false positives.

Exon junction sequences 80 bp in length were created for each splice junction present in EuSplice; each of these sequences consisted of the last 40 bases from the 3′ end of the donor exon (ExonD) and the first 40 bases from the 5′end of the acceptor exon (ExonA). These sequences were submitted for BLAST (blastn) searches against the corresponding genome-specific EST databases. In the BLAST hits, if a minimum of 10 continuous bases comprising of the splice junction (5|5: *last 5 bases of ExonD and first 5 bases of ExonA*), continuously aligned to an EST with no gaps, then the splice junction was identified as a ‘Confirmed’ splice junction. If this criterion was not met, then the splice junction was tagged as ‘Unconfirmed’. Splice junctions with continuous alignment of at least 30 (15|15) bases at the splice junction were tagged as ‘Reliable’. Splice junctions with the full 80 bases continuously aligned to the EST (40|40) were considered well supported with evidence and tagged as ‘Valid’.

Our second methodology for splice junction validation involved alignment of splice junctions to NCBI evidence mRNAs—mRNA sequences utilized by Refseq to identify and create the RefSeq transcript model. Splice junctions from the 10 EuSplice genomes (*H.sapiens, P.troglodytes, B.taurus, M.musculus, R.norvegicus, C.familiaris, G.gallus, A.mellifera, D.rerio, D.melanogaster*) for which NCBI evidences are available, were successfully validated using this methodology. The NCBI Evidence viewer (Wheeler *et al*., 2006) web pages for genes from each of these genomes were downloaded and parsed, and the evidence mRNA IDs for specific RefSeq transcripts were extracted. Next, the evidence mRNA sequences were retrieved from the genome-specific mRNA FASTA files downloaded from UCSC Genome Browser (Karolchik *et al*., 2003) and each exon was aligned to the corresponding evidence mRNA using Emboss’ needle program (http://emboss.sourceforge.net/) with default gap opening and extension options. A splice junction was considered as well supported with mRNA evidence and tagged as ‘Reliable’ if a continuous alignment with no gaps in the last 15 bases of ExonD and the first 15 bases of ExonA (15|15) was observed. If a continuous alignment of at least five bases was found in both ExonD and in ExonA (5|5), then the splice junction was identified as ‘Confirmed’; otherwise it was designated as ‘Unconfirmed’.

2.7 Alignment of splice signals by *SpliceBlast*

The *SpliceBlast* tool utilizes blastn to align user-specified query sequences to the splice signals present in the EuSplice database. Since splice signal sequences were too short (9 bases for the donor and 15 bases for the acceptor), to retrieve alignment using the default blast options, we used the options: ‘***blastall -p blastn -d blast database -F F -e 1000 -W 7 -G 1 -E 1 -q -1 -i my_query -o blast output***’. This ensured efficient sequence alignment of the short splice signals to a given query sequence.

3 RESULTS

The EuSplice database contents are summarized in Table 1. Of the 277 399 genes in EuSplice, there are 23 678 human genes which contain 273 426 splice junctions. The authenticity of 92.8% of these human splice junctions was confirmed; 17.3% of the human genes in EuSplice are alternatively spliced, 80.9% have at least one homologous gene and 44.7% have disease associations recorded in OMIM. There are 1.38 million mammalian splice sites annotated in EuSplice. Although, splicing is less frequent in fungi than in mammals, the 11 fungal genomes in EuSplice contain a total of 59 857 splice junctions, and four of the fungal genomes have AS events. The principal features of EuSplice are summarized below.

3.1 Donor and acceptor splice signals from 23 genomes

EuSplice is the largest repository of eukaryotic splice signals and provides access to 2.1 million donor–acceptor splice signal pairs for querying, comparative analysis, visualization and download through its user-friendly web interface. The authenticity of the splice junction data presented in EuSplice was confirmed by alignment to mRNA and EST sequences (Materials & Methods). The splice signals were ranked using Shapiro–Senapathy splice consensus scoring (Shapiro and Senapathy, 1987), which provides a reliable index of the variation of each splice signal from the expected consensus sequences.

Table 2 shows the tendency of splice signals in the EuSplice database to conform to the known splice consensus for the donor and acceptor splice signals. The 9-nt donor splice consensus of (C|A) A G G T (A|G) A G T was uniformly observed in the 12 higher eukaryotes and in the comprehensive fungal splice consensus. Interestingly, we found an increased representation of A at the +3 position with a corresponding lower frequency of C at this position, in *A.thaliana* and *A.mellifera*. Similarly, we found that the frequency of A was significantly higher than G at the −3 position in *C.elegans* and fungi. In the case of the 15-nt acceptor consensus (Y_{10} N (C|T) A G (G|A)), some variations were observed in the pyramidine tract region (−14 to −5). In *H.sapiens, P.troglodytes, B.taurus, C.familiaris, M.musculus, R.norvegicus, G.gallus, D.rerio*, and *D.melanogaster*, the nucleotide T was found to be more strongly represented than C in the pyramidine tract. In *A.mellifera, C.elegans*, *A.thaliana* and fungi, T was again the most frequent nucleotide in the −14 to −5 region; however interestingly, A, rather than C, had the second highest nucleotide frequency. Only 1.9% of the splice signals in EuSplice are non-canonical, and hence, do not conform to the expected canonical GT-AG (Supplementary Table 1). The most commonly observed non-canonical splice signals are GC-AG and AT-AC.

Table 1. Summary of EuSplice contents

Organism	Total genes	Genes with AS (%)	Genes with homologs (%)	Total splice junctions	Validated* splice junctions (%)
H.sapiens	23678	17.33	80.92	273426	92.77
P.troglodytes	20972	–	59.72	167605	57.7
B.taurus	14538	19.2	–	215949	40.41
C.familiaris	19347	22.74	84.60	338993	72.46
M.musculus	24109	6.86	83.40	212090	94.90
R.norvegicus	21354	1.88	74.85	169778	74.84
G.gallus	13518	–	75.72	154241	74.24
D.rerio	18498	5.53	–	175840	74.27
D.melanogaster	13629	21.74	43.72	74217	89.44
A.mellifera	3652	4.22	–	29658	59.32
C.elegans	20051	9.72	25.19	125971	61.69
A.thaliana	26536	11.89	17.44	141159	62.92
A.fumigatus	9923	–	–	18199	42.6
C.albicans	403	–	–	12	41.67
C.glabrata	5180	0.02	–	72	38.89
C.neoformans	6273	2.97	–	35349	74.57
D.hansenii	6309	0.1	–	330	35.76
E.cuniculi	1996	–	–	6	83.33
E.gossypii	4718	–	79.69	200	40.5
K.lactis	5319	–	–	117	43.49
S.cerevisiae	5850	–	73.11	243	41.98
S.pombe	5027	–	56.26	4653	35.05
Y.lipolytica	6519	0.02	–	676	45.71

*Splice junctions were validated by alignment to EST and mRNA sequences.

3.2 Alternative transcription events

EuSplice allows extensive analysis and visualization of 95 822 AS events, 23 556 ATI events and 16 262 ATT events (Fig. 2). Exon skipping accounted for 34.63% of the AS events, while alternative acceptor (17.74%), cryptic exons (16.85%), alternative donor (16.06%) and intron retention (12.87%) accounted for progressively lesser proportions. Exons with both alternative acceptor and donor splice sites accounted for only 1.84% of the total AS events.

3.3 Comparative analysis of homologs

Exploiting the ready availability of splice and gene information from the 23 eukaryote genomes, we inter-linked individual EuSplice genes to their homologs using information extracted from the Homologene database (Wheeler *et al*., 2006). Thus far, 120 884 EuSplice genes have been linked to their respective homologous genes in this manner and this facility allows users to efficiently perform comparative analyses of splicing patterns of known homologs.

Table 2. Splice consensus summary for the 12 higher eukaryotes and comprehensive splice consensus summary for the 11 fungi calculated from EuSplice

Organism	Donor	Acceptor
H.sapiens	$M_{70}A_{63}G_{80}\mid G_{100}T_{99}R_{94}A_{69}G_{78}T_{48}$	$Y_{76}Y_{78}Y_{80}Y_{81}Y_{80}Y_{78}Y_{78}Y_{79}Y_{85}Y_{85}N_{55}C_{65}A_{100}G_{100}\mid R_{74}$
P.troglodytes	$M_{66}A_{57}G_{73}\mid G_{98}T_{97}R_{90}A_{62}G_{71}T_{43}$	$Y_{73}Y_{75}Y_{77}Y_{78}Y_{77}Y_{75}Y_{75}Y_{76}Y_{81}Y_{81}N_{56}Y_{90}A_{99}G_{98}\mid R_{71}$
B.taurus	$M_{69}A_{62}G_{78}\mid G_{100}T_{98}R_{92}A_{66}G_{76}T_{46}$	$Y_{76}Y_{77}Y_{79}Y_{81}Y_{79}Y_{78}Y_{77}Y_{79}Y_{83}Y_{83}N_{55}Y_{92}A_{100}G_{100}\mid G_{48}$
C.familiaris	$M_{69}A_{63}G_{78}\mid G_{100}T_{98}R_{92}A_{66}G_{75}T_{47}$	$Y_{75}Y_{77}Y_{78}Y_{80}Y_{79}Y_{77}Y_{77}Y_{78}Y_{83}Y_{83}N_{55}Y_{92}A_{100}G_{100}\mid R_{74}$
M.musculus	$M_{70}A_{64}G_{80}\mid G_{100}\ T_{99}R_{94}A_{70}G_{79}T_{48}$	$Y_{76}Y_{78}Y_{79}Y_{81}Y_{80}Y_{79}Y_{79}Y_{80}Y_{85}Y_{85}N_{80}Y_{94}A_{100}G_{100}\mid R_{75}$
R.norvegicus	$M_{69}A_{63}G_{79}\mid G_{100}T_{99}R_{93}A_{69}G_{78}T_{47}$	$Y_{76}Y_{77}Y_{79}Y_{81}Y_{80}Y_{79}Y_{78}Y_{79}Y_{84}Y_{85}N_{55}Y_{93}A_{100}G_{99}\mid R_{74}$
G.gallus	$M_{67}A_{60}G_{74}\mid G_{100}T_{99}R_{92}A_{64}G_{72}T_{45}$	$T_{50}T_{52}T_{53}T_{56}Y_{79}Y_{77}Y_{78}Y_{78}Y_{82}T_{59}N_{54}Y_{91}A_{100}G_{100}\mid R_{70}$
D.rerio	$M_{69}A_{61}G_{74}\mid G_{99}T_{98}R_{85}A_{64}G_{66}W_{72}$	$T_{53}T_{55}T_{55}T_{56}T_{55}T_{52}T_{52}T_{52}Y_{78}T_{60}N_{63}Y_{92}A_{100}G_{99}\mid R_{72}$
D.melanogaster	$M_{64}A_{52}G_{65}\mid G_{100}T_{99}R_{94}A_{75}G_{84}T_{69}$	$T_{45}T_{46}T_{46}T_{48}T_{47}Y_{70}Y_{67}Y_{74}T_{58}T_{62}N_{84}Y_{94}A_{100}G_{100}\mid R_{63}$
A.mellifera	$M_{66}A_{59}G_{62}\mid G_{100}T_{98}A_{74}A_{63}R_{79}T_{60}$	$W_{80}W_{80}W_{81}W_{81}W_{80}W_{79}W_{78}T_{52}T_{67}T_{70}N_{73}Y_{91}A_{100}G_{100}\mid R_{74}$
C.elegans	$A_{40}A_{54}G_{58}\mid G_{99}T_{99}R_{82}A_{65}G_{73}T_{61}$	$W_{76}W_{76}W_{75}W_{75}W_{77}W_{80}W_{85}W_{85}T_{88}T_{97}N_{66}C_{82}A_{100}G_{99}\mid R_{72}$
A.thaliana	$M_{69}A_{64}G_{77}\mid G_{100}T_{99}A_{66}W_{81}G_{50}T_{51}$	$T_{51}T_{52}T_{52}T_{51}T_{47}T_{49}T_{50}T_{50}T_{52}T_{62}N_{91}Y_{92}A_{100}G_{100}\mid R_{77}$
Fungi	$A_{35}A_{39}G_{47}\mid G_{100}T_{99}R_{90}A_{56}G_{90}T_{72}$	$W_{60}W_{62}W_{61}W_{60}W_{59}T_{33}T_{35}T_{35}T_{35}T_{41}N_{36}Y_{89}A_{100}G_{100}\mid R_{58}$

N stands for any of the 4 nt (A, T, G or C) and R: A or G; Y: T or C; M: A or C; W: A or T.

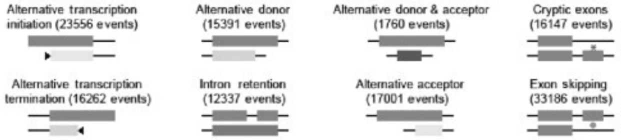

Alternative transcription initiation (23556 events)
Alternative donor (15391 events)
Alternative donor & acceptor (1760 events)
Cryptic exons (16147 events)
Alternative transcription termination (16262 events)
Intron retention (12337 events)
Alternative acceptor (17001 events)
Exon skipping (33186 events)

Fig. 2. Alternative splicing, transcription initiation and transcription termination events in EuSplice.

3.4 Extensive data integration

Our aim in creating EuSplice was not only to provide reliable splicing information for a wide array of genomes, but also to minimize the encumbrance of cross-navigation or 'web surfing' involved in accessing all relevant information for genes of interest. Mindful of the value of data integration in biological research (Stein, 2003), we have integrated information extracted from various resources, such as gene nomenclature annotations, location, gene function, disease association and UTR features. Sequences for genes, introns, exons, UTRs and CDSs are readily available. Functional annotations from the Gene Ontology (GO) database (Ashburner *et al*., 2000) and GeneRIF (Gene Reference Into Function) annotations from the Entrez Gene database (Wheeler *et al*., 2005) have also been integrated into EuSplice. The user-friendly web interface in EuSplice also incorporates 'link integration', whereby access to a variety of external resources is provided in the form of hyperlinks, allowing the user to seamlessly navigate across various databases without getting lost in a labyrinth of independent database queries. The EuSplice gene records are linked to 23 database resources (Galperin, 2006; see Supplementary Table 2), which ultimately provide over 1.5 million cross-references.

3.5 Intelligent querying

The search options available in EuSplice have been carefully designed to ensure that the user can perform intelligent querying to extract and analyze relevant information.

3.5.1 General keyword searches Performing a general keyword search results in a Perl regular expression-based query which retrieves relevant genes from the numerous tables within the database. For example, a general search for the term '*cancer*' retrieves 2381 genes from 12 genomes. While some of these genes are retrieved due to matches in the gene nomenclature, others are retrieved due to matches in gene function or disease association.

3.5.2 Specific searches The user can narrow their search by selecting specific search options. A key word search can be restricted by species or even to matches that include the term specifically within the gene name or gene symbol. Additionally ID-specific queries, such as those based on Entrez Gene ID, Refseq Transcript ID, Refseq Protein ID, OMIM ID, Gene Symbol and GO ID, can also be performed. For example, limiting a search for the term '*angiotensin*' to the *Human Genome* and to *Gene Name/Gene Synonyms*, results in retrieval of 9 specific hits, while a general keyword search would have retrieved 274 hits from 10 different species.

3.5.3 Function-based and disease-based queries Genes can be explored based on their functional role as well as their disease associations. For example, entering the query '*carboxypeptidase activity*' after selecting '*Human genome*' from the Organism Menu and the '*Function Search*' from the ID Menu will retrieve 28 human genes with this functional role. Meanwhile, the disease-based query '*breast cancer*' retrieves 19 genes with breast cancer associations specified in OMIM.

3.5.4 Genome locus-specific searches Genome locus-targeted searches are especially useful for researchers who are interested in analyzing splice signals and AS events within a priori defined genomic regions. For example, a query to retrieve all genes at map location *1q21.1* returns 76 genes, whereas a query to retrieve all genes in a *1 MB* region starting from base position *10 000* on *human chromosome 1* retrieves 26 genes.

3.5.5 SpliceBlast and SpliceMatch These unique data-mining tools allow users to efficiently query EuSplice's extensive splice signal dataset. *SpliceBlast* aligns splice signals from EuSplice database to user-specified query sequences and is a useful tool for the identification of potential intron/exon structures within sequences of interest. Since the alignments are made against known splice signals of annotated genes, they can be used as an additional facility for the analysis of the reliability of splice junctions predicted by various splice prediction techniques. The results of a sample *SpliceBlast* query is given in Supplementary Table 3. *SpliceMatch* is particularly useful when a user wants to identify the prevalence of splice signals in an organism or would like to analyze the distribution pattern of splice signals in various eukaryotes. *SpliceMatch* can also be used to retrieve exons and introns and to examine the distribution of putative 'cryptic splice signals' (Nelson and Green, 1990) around splice junctions of a gene of interest.

3.6 Graphical visualization of the gene, splice junctions and AS events

The graphics display options in EuSplice allow in-depth exploration of features of the gene and provide the user with a clear view of the gene structure, splice junctions and AS events. They can be divided into the following three main categories:

Gene Plots: These plots provide graphical representation of a gene's 5′UTRs, CDSs and 3′UTRs. Detailed information about particular exons and introns, such as gene-specific position co-ordinates and exon classification based on the coding sequence content, can also be viewed in this plot using the mouse-over pop-up features.

Splice Plots: Splice junctions and flanking sequences can be visualized in Splice Plots. They provide a user-friendly platform for analyzing the distribution of splice signal sequences and putative cryptic splice sites.

AS Plots: These plots not only depict AS, ATI and ATT events (Fig. 2), they also reveal alternative donor and acceptor splice sites of alternatively spliced exons relative to those of the constitutive exons. The splice score associated with each alternatively spliced site provides a reliable index of splice signal strength.

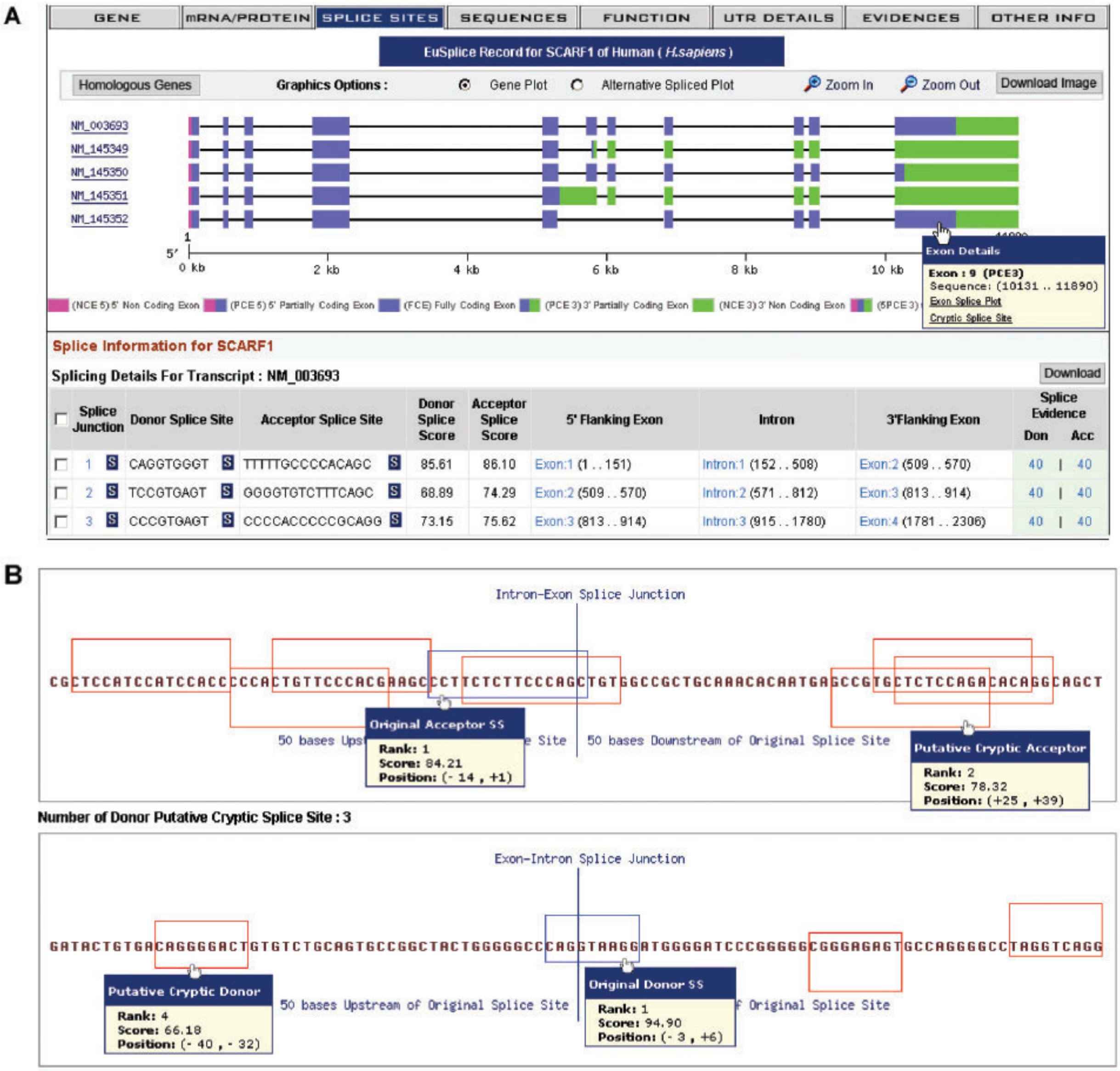
3.7 Exploring EuSplice with a sample query

The features of EuSplice are demonstrated here with a specific example: the scavenger receptor class F, member 1 (*SCARF1*) gene; an endothelial receptor for acetylated low-density lipoprotein (Ace-LDH). In the EuSplice query page, a general keyword search for 'scavenger receptor class F, member 1' retrieves five genes from the database. These include the human *SCARF1* gene, the mouse *Scarf1* gene, a predicted rat gene (*Scarf1_predicted)* and genes in cow (LOC616566) and dog (*LOC491194*) that are annotated as *SCARF1* precursors.

The EuSplice result page (Fig. 3A) for human *SCARF1* gene consists of two sections: the upper section contains the EuSplice graphical display options and the lower section contains the splice signal data for *SCARF1*. The five mRNA isoforms of *SCARF1* can be explored using the display option available in the Gene Plot. The exon color coding reveals that while all isoforms have partially coding 5′ and 3′ exons, only 2 isoforms have 3′ fully non-coding exons.

There are a total of 47 splice junctions associated with the 5 isoforms of the *SCARF1* gene (Fig. 3A). Splice signal sequences, splice consensus scores and the details of the specific intron and flanking exons are provided. All splice junctions in *SCARF1* have been confirmed and 34 have a full 80 bases of continuous alignment (40|40) indicative of a fully validated entry. The details of the evidence source and match alignment are available in the 'EVIDENCES' page (not shown).

A sample putative cryptic splice plot of a *SCARF1* exon is shown in Figure 3B. There are seven putative cryptic acceptor splice sites around the start of the exon 5 of the NM_145351 isoform, 4 upstream of the actual acceptor splice signal and



Exon : 9 (PCE3)
Sequence: (10131 .. 11890)

Splice Information for SCARF1

Splicing Details For Transcript : NM_003693

□	Splice Junction		Donor Splice Site		Acceptor Splice Site		Donor Splice Score	Acceptor Splice Score	5' Flanking Exon	Intron	3'Flanking Exon	Splice Evidence		
												Don		Acc
□	1	S	CAGGTGGGT	S	TTTTTGCCCCACAGC	S	85.61	86.10	Exon:1 (1 . . 151)	Intron:1 (152 . . 508)	Exon:2 (509 . . 570)	40	|	40
□	2	S	TCCGTGAGT	S	GGGGTGTCTTTCAGC	S	68.89	74.29	Exon:2 (509 . . 570)	Intron:2 (571 . . 812)	Exon:3 (813 . . 914)	40	|	40
□	3	S	CCCGTGAGT	S	CCCCACCCCCGCAGG	S	73.15	75.62	Exon:3 (813 . . 914)	Intron:3 (915 . . 1780)	Exon:4 (1781 . . 2306)	40	|	40

```
CGCTCCATCCATCCACCCCCACTGTTCCCACGAAGCCCTTCTCTTCCCAGCTGTGGCCGCTGCAAACACAATGAGCCGTGCTCTCCAGACACAGGCAGCT
```

Original Acceptor SS
Rank: 1
Score: 84.21
Position: (- 14 , +1)

Putative Cryptic Acceptor
Rank: 2
Score: 78.32
Position: (+25 , +39)

Number of Donor Putative Cryptic Splice Site : 3

```
GATACTGTGACAGGGGACTGTGTCTGCAGTGCCGGCTACTGGGGGCCCAGGTAAGGATGGGGATCCCGGGGGCGGGAGAGTGCCAGGGGCCTAGGTCAGG
```

Putative Cryptic Donor
Rank: 4
Score: 66.18
Position: (- 40 , - 32)

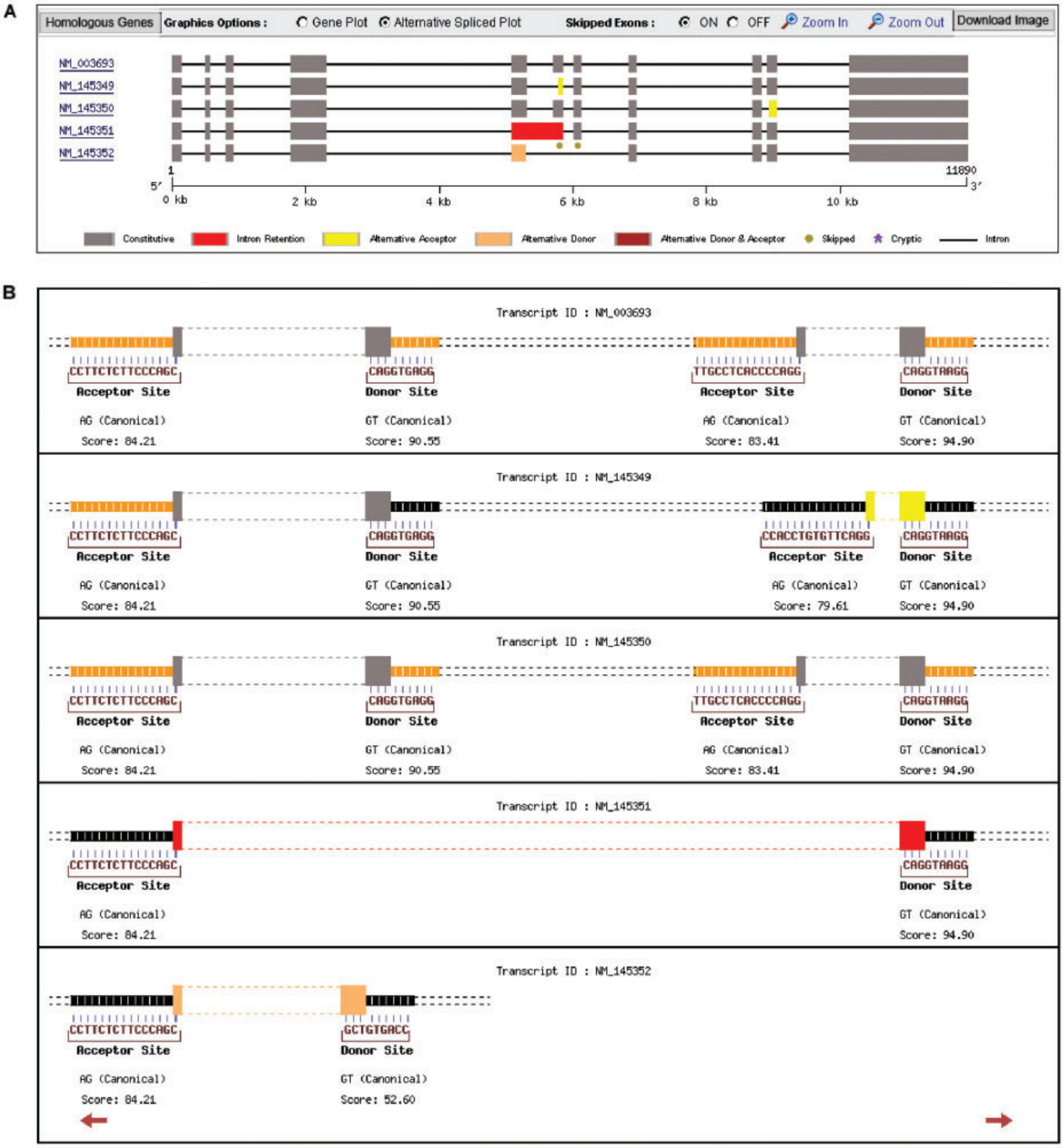
Original Donor SS
Rank: 1
Score: 94.90
Position: (- 3 , +6)

Fig. 3. (**A**) EuSplice result page with Gene Plot and splice signal information for human *SCARF1* and (**B**) Putative cryptic splice signals around the donor and acceptor splice sites of Exon 5 of NM_14351.

3 downstream of it. At the exon end, there are three putative cryptic donor splice sites, one of which is upstream of the exon end, and two of which are downstream. Note that the splice scores of the actual splice signals are much higher than that of the putative ones.

Using the '*SpliceMatch*' button of splice junction 10 of the NM_003693 isoform, we identified which genes share the donor–acceptor splice signal combination with this intron. *SpliceMatch* retrieved matches in three genomes: human, dog and cow. All five isoforms of *SCARF1* in the human genome have the same intron in this region (intron position 9040–10130). The dog gene *LOC491194* has a single intron with the same donor-acceptor splice signal. This gene is a known homolog of human *SCARF1*, and its details are available in the *SCARF1* 'Homologous Genes' page. The last intron of two isoforms (XM_870136 and XM_883180) in the cow gene *LOC616566* has this specific splice signal combination. Although this gene has not been annotated as a known homolog of human *SCARF1* in the Homologene database (and hence not presented in the *SCARF1* Homologous Genes page), the 'Gene Name'-specific query did retrieve this gene as a hit.

SCARF1 has four types of AS events (Fig. 4A). The color-coded exon display clearly indicates that alternative acceptor splice sites are present in two mRNA



Graphics Options : ○ Gene Plot ● Alternative Spliced Plot
Skipped Exons : ● ON ○ OFF

Transcript ID : NM_003693

CCTTCTCTTCCCAGC
Acceptor Site
AG (Canonical)
Score: 84.21

CAGGTGAGG
Donor Site
GT (Canonical)
Score: 90.55

TTGCCTCACCCCAGG
Acceptor Site
AG (Canonical)
Score: 83.41

CAGGTAAGG
Donor Site
GT (Canonical)
Score: 94.90

Transcript ID : NM_145349

CCTTCTCTTCCCAGC
Acceptor Site
AG (Canonical)
Score: 84.21

CAGGTGAGG
Donor Site
GT (Canonical)
Score: 90.55

CCACCTGTGTTCAGG
Acceptor Site
AG (Canonical)
Score: 79.61

CAGGTAAGG
Donor Site
GT (Canonical)
Score: 94.90

Transcript ID : NM_145350

CCTTCTCTTCCCAGC
Acceptor Site
AG (Canonical)
Score: 84.21

CAGGTGAGG
Donor Site
GT (Canonical)
Score: 90.55

TTGCCTCACCCCAGG
Acceptor Site
AG (Canonical)
Score: 83.41

CAGGTAAGG
Donor Site
GT (Canonical)
Score: 94.90

Transcript ID : NM_145351

CCTTCTCTTCCCAGC
Acceptor Site
AG (Canonical)
Score: 84.21

CAGGTAAGG
Donor Site
GT (Canonical)
Score: 94.90

Transcript ID : NM_145352

CCTTCTCTTCCCAGC
Acceptor Site
AG (Canonical)
Score: 84.21

GCTGTGACC
Donor Site
GT (Canonical)
Score: 52.60

Fig. 4. (**A**) AS plot for *SCARF1* and (**B**) Exon isoform plot for *SCARF1*.

isoforms: one in exon 6 of NM_145349 and the other in exon 10 of NM_145350. An alternative donor splice site occurs in exon 5 of NM_145352, and an intron retention occurs in exon 5 of NM_145351. Finally, there are two exon skipping events, both in NM_145352.

Exploring the AS events in exon five in NM_145351 via the 'View Exon Isoforms' link reveals the differences in the splice junction signals among the five isoforms (Fig. 4B). Exon five donor and acceptor splice sites for three isoforms (NM_003693, NM_145349 and NM_145350) are the same. However, although this constitutive donor splice signal has a high splice score (90.55), it is ignored in NM_145351, resulting in an intron retention event. The skipping of an exon at 5082–5836 position in the NM_145352 isoform is also clearly depicted. The plot also shows that the NM_145352 has a relatively short exon five because an alternative donor splice signal with a low splice score (52.60) is selected. Similarly, the shorter exon six in the NM_145349 is due to the selection of an alternative acceptor splice signal with a lower splice score (79.61) instead of the constitutive acceptor splice signal which has a splice score of 83.41.

Clicking the 'Homologous Genes' button in the result page reveals that human *SCARF1* has three homologs, one each in mouse, rat and dog. The user can compare the AS events, splice signals and gene annotations among these homologs using the various options available in this page (Supplementary Fig. 2). The homolog AS plot clearly indicates that while the rat homolog is alternatively spliced, the mouse and dog homologs are not.

Integrated gene information can be accessed by selecting the 'GENE', 'mRNA/PROTEIN', 'SEQUENCES', 'FUNCTION', 'UTR DETAILS' and 'OTHER INFO' buttons at the top of the EuSplice result page, which opens specific information pages (Supplementary Fig. 3) below the graphical gene display. Accordingly, EuSplice revealed that *SCARF1* has 2 GeneRIFs and 10 known GO functions, including transmembrane receptor activity, low-density lipoprotein binding and low-density lipoprotein binding catabolism, and that there are SINE and Alu repeat sequences present in the 3′UTRs of 4 mRNA isoforms of *SCARF1*. Database cross-references for *SCARF1* can be accessed through the 'OTHER INFO' drop-down menu and include one Unigene cluster (Unigene ID: 534497), six STS entries, five literature citations, one disease association in OMIM and references to *SCARF1* in HGNC, HPRD, CCDS and H-INV databases.

3.8 Future directions

EuSplice is an ongoing project, and we will continue to enhance the database content with the additional curated and computed data on the splicing process as well as expand the analytical capabilities of the web interface with further novel data mining and visualization tools. We plan to include splicing annotations for several additional genomes and add extensive curated and computed information on splice mutations, branch point sequences, experimentally verified cryptic splice sites, splicing enhancers and other splicing regulatory sequences. We will also integrate information on splicing pathways and provide tools for the analysis of gene expression variations due to AS. Furthermore, we are developing several software modules for the analysis of mRNA secondary structure at splice junctions, splice-related polymorphisms, role of AS in disease manifestation and the phylogenetic variation of splicing patterns in eukaryotes.

ACKNOWLEDGEMENTS

We thank Philge Philip, Sanjog Sahu, Ramkumar Sundaram, Vinoth Sankaran, Prabu Raja, Arasi Nagarajan and Anup Viswanathan for technical support and Ann E. Power for valuable comments on the manuscript.

Conflict of Interest: None declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Balvay,L. *et al.* (1993) Pre-mRNA secondary structure and the regulation of splicing. *Bioessays*, **15**, 165–169.

Baralle,D. and Baralle,M. (2005) Splicing in action: assessing disease causing sequence changes. *J. Med. Genet.*, **42**, 737–748.

Boguski,M.S. *et al.* (1993) dbEST – database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.

Bonizzoni,P. *et al.* (2006) Computational methods for alternative splicing prediction. *Brief Funct. Genomic. Proteomic.*, **5**, 46–51.

Buratti,E. *et al.* (2006) Defective splicing, disease and therapy: searching for master checkpoints in exon definition. *Nucleic Acids Res.*, **34**, 3494–3510.

Burge,C.B. *et al.* (1999) Splicing of precursors to mRNAs by the spliceosomes. In: Gesteland,R.F., Cech,T. and Atkins,J.F. (eds), *The RNA World. 2nd edn.* Cold Spring Harbor Laboratory Press, Plainview, NY, pp. 525–560.

Burset,M. *et al.* (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.

Burset,M. *et al.* (2001) SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res.*, **29**, 255–259.

Chong,A. *et al.* (2004) Information for the coordinates of exons (ICE): a human splice sites database. *Genomics*, **84**, 762–766.

Galperin,M.Y. (2006) The molecular biology database collection: 2006 update. *Nucleic Acids Res.*, **34**, D3–D5.

Holste,D. *et al.* (2006) HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res.*, **34**, D56–D62.

Huang,H.D. *et al.* (2005) SpliceInfo: an information repository for mRNA alternative splicing in human genome. *Nucleic Acids Res.*, **33**, D80–D85.

Karolchik,D. *et al.* (2003) The UCSC Genome Browser database. *Nucleic Acids Res.*, **31**, 51–54.

Mount,S.M. (1982) A catalogue of splice junction sequences. *Nucleic Acids Res.*, **10**, 459–472.

Nagasaki,H. *et al.* (2006) Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. *Bioinformatics*, **22**, 1211–1216.

Nelson,K.K. and Green,M.R. (1990) Mechanism for cryptic splice site activation during pre-mRNA splicing. *Proc. Natl. Acad. Sci. USA*, **87**, 6253–6257.

Nilsen,T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell. *Bioessays*, **25**, 1147–1149.

Ratsch,G. *et al.* (2005) RASE: recognition of alternatively spliced exons in C.elegans. *Bioinformatics*, **21**, i369–i377.

Senapathy,P. *et al.* (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Meth. Enzymol.*, **183**, 252–278.

Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.

Sheth,N. *et al.* (2006) Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res., **34**, 3955–3967.

Staley,J.P. and Guthrie,C. (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell.*, **92**, 315–326.

Stamm,S. *et al*. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res*., **34**, D46–D55.

Stein,L.D. (2003) Integrating biological databases. *Nat. Rev. Genet*., **4**, 337–345.

Stetefeld,J. and Ruegg,M.A. (2005) Structural and functional diversity generated by alternative mRNA splicing. *Trends Biochem. Sci*., **30**, 515–521.

Thanaraj,T.A. (1999) A clean data set of EST-confirmed splice sites from Homo sapiens and standards for clean-up procedures. *Nucleic Acids Res*., **27**, 2627–2637.

Wheeler,D.L. *et al*. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **34**, D173–D180.