

# HOLLYWOOD: a comparative relational database of alternative splicing

Dirk Holste\*, George Huo<sup>1</sup>, Vivian Tung and Christopher B. Burge

Department of Biology and <sup>1</sup>Department of Computer Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02319, USA

Received August 15, 2005; Revised September 26, 2005; Accepted October 4, 2005

## ABSTRACT

**RNA splicing is an essential step in gene expression, and is often variable, giving rise to multiple alternatively spliced mRNA and protein isoforms from a single gene locus. The design of effective databases to support experimental and computational investigations of alternative splicing (AS) is a significant challenge. In an effort to integrate accurate exon and splice site annotation with current knowledge about splicing regulatory elements and predicted AS events, and to link information about the splicing of orthologous genes in different species, we have developed the HOLLYWOOD system. This database was built upon genomic annotation of splicing patterns of known genes derived from spliced alignment of complementary DNAs (cDNAs) and expressed sequence tags, and links features such as splice site sequence and strength, exonic splicing enhancers and silencers, conserved and non-conserved patterns of splicing, and cDNA library information for inferred alternative exons. HOLLYWOOD was implemented as a relational database and currently contains comprehensive information for human and mouse. It is accompanied by a web query tool that allows searches for sets of exons with specific splicing characteristics or splicing regulatory element composition, or gives a graphical or sequence-level summary of splicing patterns for a specific gene. A streamlined graphical representation of gene splicing patterns is provided, and these patterns can alternatively be layered onto existing information in the UCSC Genome Browser. The database is accessible at <http://hollywood.mit.edu>.**

## INTRODUCTION

Gene expression is controlled at several levels, and in meta-zoan genomes, where the majority of protein-coding genes contain introns, the splicing of precursors to mRNAs (pre-mRNAs) constitutes a critical step for regulation of gene expression (1–3). RNA splicing occurs in the nucleus and is catalyzed by a large ribonucleoprotein (RNP) complex known as the spliceosome, which is composed of several small nuclear RNPs and over one hundred proteins (4). The processing of pre-mRNAs is often variable, giving rise to multiple alternatively spliced mRNAs, which may serve to produce distinct protein isoforms (5–7). Typical mammalian gene loci span tens of thousands of nucleotides (nt), with an average of nine exons/eight introns and the coding region typically spanning ~1500 nt (8–10). In addition to the precise recognition of splice sites among many possible pseudo-sites, the removal of introns and the production of the correct message, the spliceosome must also produce tissue- and developmental stage-specific mRNA isoforms and integrate RNA splicing decisions with other steps in RNA processing, such as capping, cleavage and polyadenylation (11,12). Correct pre-mRNA splicing is generally required for cell viability. At least 15% of point mutations that cause genetic defects do so by altering splice site sequences (13), and the misregulation of alternative splicing (AS) is associated with a number of human diseases (14–17).

Alternative pre-mRNA splicing is estimated to affect more than half of actively transcribed human genes (18), and the systematic identification of AS events is important for the fundamental understanding of the regulation of gene expression in development, differentiation and human disease (19,20). A number of AS databases have been constructed, based on either searches of the scientific literature (21–23) or automated large-scale comparisons of transcript and genomic sequences (see below). The latter approach is made possible by the availability of large repositories of complementary

\*To whom correspondence should be addressed. Email: holste@mit.edu

Correspondence may also be addressed to Christopher B. Burge. Tel: +1 617 258 5997; Fax: +1 617 452 2936; Email: cburge@mit.edu

DNA (cDNA) sequences and expressed sequence tags (ESTs), derived from different tissues or cell lines. Available data enable large-scale computational analysis of AS in human and mouse, and a few other organisms, with an average of >200 transcripts available for each annotated human gene (24). Transcript-based AS databases include: the Alternative Splicing Database Project, ASAP (25), the Alternative Splicing Database, ASD (26), the Extended Alternatively Spliced EST Database, EASED (27), SPLICEINFO (28) and ECGENE (29), to name a few (18,30) (<http://hollywood.mit.edu/db/>). Bioinformatics studies relying on such databases have proven useful in revealing differences in AS patterns between tissues (31–33), in identifying conserved AS events in orthologous genes (34–38), and for describing disease-associated AS (39). However, the AS events recorded differ significantly between different databases, owing to differences in primary sequence data used, in the algorithms used to generate spliced alignments, and in the stringency of alignment quality filtering.

More recently, splicing-sensitive microarrays have been designed and used for the detailed analysis of tissue-specific and other types of AS (40–45), and a cross-linking/immunoprecipitation strategy has been introduced for the systematic identification of RNAs bound by a given splicing factor (46). These newly developed methods set a direction toward increasingly parallel experimental analysis of splicing regulation, and AS databases will become increasingly important in both experimental design and data analysis for these types of functional genomic approaches.

To aid in computational and large-scale experimental studies of AS, we developed HOLLYWOOD, a comparative relational database of AS. HOLLYWOOD integrates accurate exon and splice site annotation derived from spliced alignments of transcripts to genomic sequences with current knowledge about splicing regulatory elements and predicted AS events, and links information about the splicing of orthologous genes in different species to facilitate comparative analyses. A compact representation of the splicing pattern of any desired gene is provided, and sets of alternative or constitutive exons can be obtained using complex queries for features such as splice site strength, type of AS event, tissue expression, splicing regulatory element content or conservation of the AS event between human and mouse.

## HOLLYWOOD DATABASE

The design and implementation of HOLLYWOOD followed certain guiding principles: (i) all exon and isoform data should derive from high-quality spliced alignment of transcripts to genome sequences; (ii) AS events should be identified without requiring designation of an arbitrary ‘reference’ transcript; (iii) current knowledge about splice sites, splicing regulatory elements and predicted AS events should be incorporated into the database to allow efficient searches; (iv) the database should be integrated with other widely-used databases and genome browsers when possible; and (v) two main types of queries should be supported—queries for splicing information about a particular gene or genes and queries for sets of exons with particular properties. Examples of the output format for each of these types of queries are shown in Figure 1A for a gene query for the human fragile X mental retardation

syndrome-related (*FXR1*) gene, and in Figure 1B for an exon query yielding exon 16 of the *FXR1* gene.

HOLLYWOOD incorporates current knowledge about splice sites and splicing regulatory elements. It uses both a standard position-specific weight matrix model as well as a sophisticated maximum entropy-based model for the quantification of 3' and 5' splice site (3'ss and 5'ss) strength; the latter has been shown to more accurately distinguish authentic and pseudo splice sites (47,48). In addition to classical 3'ss and 5'ss motifs, it is now well established that other *cis*-regulatory elements including exonic splicing enhancers (ESEs) and silencers (ESSs) play common and important roles in exon and splice site choice (49). HOLLYWOOD annotates exons with sets of candidate ESE and ESS elements that have been identified in recent computational and experimental screens (50,51). The database also incorporates information about ‘alternative-conserved exons’ (ACEs)—orthologous exon pairs whose alternative splicing is conserved between human and mouse—from two sources: ~450 exons with transcript evidence of AS in both species are annotated, as well as ~2000 candidate ACEs predicted by the ACESCAN algorithm (35).

## Primary data

In building the HOLLYWOOD system, five major data sources were used, all of which are publicly available: (i) Ensembl gene chromosomal locations and gene identifiers (52), corresponding to genome assemblies from GoldenPath version hg16 of the human and version mm3 of the mouse genome (<http://genome.ucsc.edu/>); (ii) transcript sequences from GenBank release 139.0, including the repositories gbpri, gbrod and gbhtc; (iii) EST sequences from dbEST, release 01122004, totaling ~5.4 million human and ~4.5 million mouse ESTs [dbEST records were grouped into one of about 40 human or mouse primary tissue types according to their cDNA library information, as described previously (32)]; (iv) mammalian interspersed repeat sequences from the RepBase repository (53); and (v) sets of ESEs and ESSs from the RESCUE-ESE and FAS-hex2 datasets, respectively (50,51).

## Exon and feature annotation


Genomic sequences were extracted spanning an Ensembl gene from the start to the end of the annotation, plus an additional 5000 nt upstream and downstream of the start and end, respectively; these sequences are referred to as gene ‘slices’. The set of slices for all Ensembl genes was obtained from Ensembl (54). Use of the Ensembl gene annotation to define gene slices enables use of standard gene names and identifiers and enables linking to external databases. However, beyond the definition of slice boundaries, Ensembl annotation is not explicitly used in HOLLYWOOD: all exon/intron annotation and splicing information derives directly from transcript alignments. For convenience, HOLLYWOOD generally uses gene slice-based coordinates, which are converted to global chromosomal coordinates as needed.

Large-scale spliced alignments of transcript sequences to genomic DNA are conducted using the genome annotation system GENOA (<http://genes.mit.edu/genoa>), which will be described in greater detail elsewhere. Briefly, GENOA detects statistically significant blocks of identity between repeat-filtered

cDNA sequences and gene slices, then conducts spliced alignments of best-matched cDNAs to corresponding gene loci using the algorithm mRNAvsGEN. To avoid problems attendant to automated annotation of genomic regions, which are subject to frequent rearrangement such as immunoglobulin loci, cDNAs from certain classes of immune-related genes

are optionally excluded by GENOA. Statistically significant matches are then identified between EST sequences and aligned repeat-filtered cDNAs, and best-matched ESTs are aligned to the corresponding gene slices using the SIM4 algorithm (55). GENOA was applied with stringent alignment criteria, requiring a sequence identity above 93% for cDNA

**A**



## Alternatively spliced mRNA

# Hollywood.mit.edu

Genomic coordinates of the gene locus, ranging from first to last exon regions

**Hollywood search result**

Genus: **Homo sapiens**  
 Species: **314181.315:187215.8561**  
 Chromosome: **forebrain**  
 Strand: **forward**  
 Ensembl gene identifier: **ENSG00000114416**  
 Ensembl gene name: **FXR1**  
 Gene description: **fragile x mental retardation syndrome related protein 1**  
 EST-derived tissue types: **lyc cell (2), blood (1), brain (20), breast (2), cervix (3), cochlea (1), colon (1), endometrium (1), eye retina (3), fetal brain (2), genitourinary (1), germ cell (3), head and neck (4), heart (4), kidney (2), lung (6), muscle (5), nervous system (4), ovary (1), pancreas (5), placenta (6), pooled tissue (9), prostate (1), skin (2), stomach (5), testis (11), thyroid (9), uterus (2), whole body (1)**

Legend with color-encoded categories of constitutive and alternative exons

- First/Last exon regions
- Ensembl 100% annotated exon
- Alternative exon
- Retained intron
- Constitutive exon
- Internal exon
- ACEScan(+) exon
- ACEScan(-) exon

Link to Ensembl database with further information about gene and gene expression

**Reference exon sets**

Hollywood annotation  
 ACEScan(+/-) exons  
 Ensembl annotation

Alternative splice form  
 Transcript/Locus  
 UCSC genome browser view

5' exon cont: [cDNA:AY141428](#)  
 5' exon extension: [EST:BX143997](#)

5' exon cont: [cDNA:AY141428](#)  
 5' exon extension: [EST:BU195111](#)

Retained intron: [EST:BU167769](#)  
 Spliced intron: [cDNA:AY141428](#)

Exon inclusion: [cDNA:AY141428](#)  
 Exon skipping: [cDNA:BC029551](#)

Click on either one of the transcript images to view larger version (if compressed due to the presence of a larger number of exons)

Tissue categories and corresponding number of occurrences, derived from cDNA libraries for ESTs

Graphical representation of internal exons, and first and last exon regions

Predicted alternative conserved exons (ACEs) and link to ACEScan webserver for additional information

Link to GenBank database with pertinent information about the transcript record and sequence retrieval

Representative transcripts, with exon and intron sizes, for supporting splicing classification for alternative exons

Link to UCSC Genome Browser layered with representative Hollywood transcripts

**B**



5' 3'

9.03 bits 182,014,014..182,014,105 9.33 bits

```

tctttaacag|TCACAGTTGCAGATTATATTTCTAGAGCTGAGTCTCAGAGCAGACAAAGAAACCTCCCAAGGGAACCTTTGGCTAAAAACAAGAAAGAAATG|gtaaggagaa
|||||
tctttaacag|TCACAGTTGCAGATTATATTTCTAGAGCTGAGTCTCAGAGCAGACAAAGAAACCTCCCAAGGGAACCTTTGGCTAAAAACAAGAAAGAAATG|gtaaggagaa
    
```

AGATTA TTTCTA
AGCAGA
CCAAGGGAAC
AACAAG

TTCTAG
AGACAA
GAAACT
ACAAG

GACAAA
GAACT
ACAAGA

ACAAG
GAACT
CAAGAA

CAANGA
GAACT
AGAAG

AAGAAA
GAACT
GAAAG

AAGAAA
GAACT
AAGAAA

AAGAAA
GAACT
AAGAAA

AAGAAA
GAACT
AAGAAA

AAGAAA
GAACT
AAGAAA

AAGAAA
GAACT
AAGAAA

AAGAAA
GAACT
AAGAAA

```

>Exon: ENSG00000114416:182014014..182014105
Genus: Homo
Species: sapiens
Gene name: FRAGILE X MENTAL RETARDATION SYNDROME RELATED PROTEIN 1
Chromosome: 3
Strand: 1
Chromosome coordinates: 182014014..182014105
Exon position in transcript: I
Exon length (bp): 92
3' splice site: tttttttcattctttaacag|TCACA
5' splice site: AAATG|gtaaggagaa
Splicing characterization: SE
MaxENT 3' splice site score: 9.03
MaxENT 5' splice site score: 9.33
PWM 3' splice site score: 8.6
PWM 5' splice site score: 8.79
ACEScan score: 0.170779
RESCUE-ESE motif(s): agcaga, agacaa, gacaaa, acaaaag, caaaga, aaagaa, aagaaa, agaaac, gaaacc, aaacct, ggaaac, gaaact, aacaag, acaaga, caagaa, aagaaa, agaaag, gaaaga, aaagaa, aagaaa
RESCUE-ESE initial coordinate(s): 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 61, 62, 77, 78, 79, 80, 81, 82, 83, 84
ESS motif(s): agatta, tttcta, ttctag, coaagg
ESS initial coordinate(s): 10, 18, 19, 56
Sequence: tcacagttgc agattatatt tctagagctg agtctcagag cagacaaaga aaacctcccaa
gggaaacttt ggctaaaaa aagaagaaa tg
    
```

alignments. For ESTs, the first and last aligned segments were required to be at least 30 nt long, with a sequence identity of at least 90%, and the entire alignment was required to have a sequence identity of at least 90%, over at least 90% of EST nucleotides. Using ~22 200 human and 25 000 mouse gene slices, these alignments criteria were passed by ~79 000 out of 115 000 human cDNAs for ~19 300 gene slices, and by roughly one-fifth out of 5.4 million human ESTs, highlighting the stringency of the applied quality filter. The same alignment criteria were passed by only ~27 000 out of 102 000 mouse cDNAs for ~13 500 gene slices, while roughly one-fifth out of 4.1 million mouse ESTs met these criteria. GENOA aligned 2–4% of ESTs and ~1% of cDNAs to multiple loci on different chromosomes.

The annotation of exons as constitutive or alternative is made by the program runHOLLYWOOD, which implements a set of computational rules to identify splice types of alternative exons (to be described elsewhere). HOLLYWOOD annotates constitutive exons, skipped exons, mutually exclusive exons, alternative 3'ss and/or 5'ss exons and retained introns. By default, every exon is labeled as 'constitutive' and this label remains unless specific criteria are met for annotation as another alternative. Figure 2 shows for the human and mouse genome the numbers of annotated constitutive and alternative exons, together with a pictorial representation of the criteria required for identifying each of these alternative exon types. This annotation is not restricted to one splice type per exon, but allows for exons to be included in multiple categories, e.g. an exon may exhibit both skipping and alternative 5'ss usage.

### Data model and implementation as a relational database

The HOLLYWOOD system consists of a generalized alignment parser framework, a relational database and a web interface. HOLLYWOOD defines a relational data model that distinguishes three primary tables—'exon', 'gene' and 'transcript'—such that updated or new information can be represented in a structure consistent with existing data. A generalized parsing module simplifies the process of incorporating information, which is provided in flat file format. The parser currently inserts data into the PostgreSQL relational database management system, but could in principle support various database back ends. The structured query language (SQL) provides the ability to perform a wide range of powerful queries. HOLLYWOOD can also be queried through a web interface (Figure 3), with which users can build queries without knowledge of SQL. This interface allows users to retrieve sets of exons or transcripts that satisfy constraints defined

on any number of supported features. Data are output in flat file or XML-based formats for downstream bioinformatic analysis.

A primary design goal was to ease the process of importing new data, and to this end HOLLYWOOD utilizes Perl packages that hide database implementation details. The parser interacts with the relational database backend, and the data model is optimized for efficient storage and data retrieval. Proper normalization techniques are employed in HOLLYWOOD, which facilitate the removal of redundant information and contribute to the logical organization of the data model, and the primary exon, gene and transcript data were decoupled in order to minimize the number of tables that have to be reloaded when a single input file is updated. For instance, for an update of gene names/descriptions one only needs to reload the corresponding table, and without further dependencies, the rest of the database remains unaffected.

### HOLLYWOOD web interface and example applications

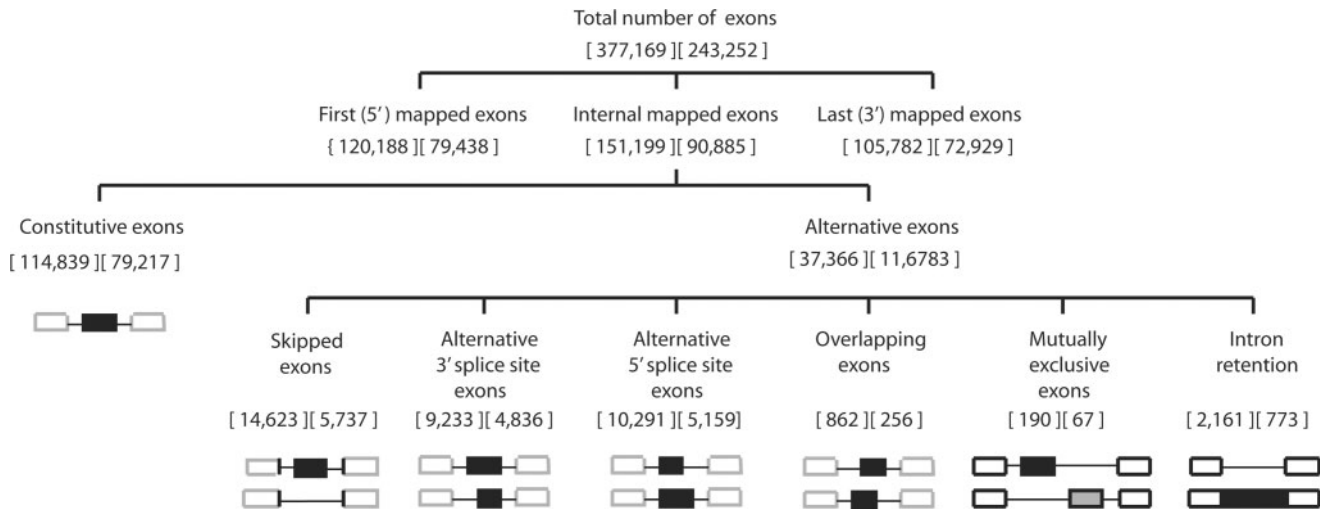
The HOLLYWOOD system is accessible at <http://hollywood.mit.edu>. Human and mouse gene slices, comprising roughly one-third of the human or mouse genome, are available for download and organized by chromosomes. Gene locus-based (local) coordinates are provided in 5'–3' direction corresponding to the transcriptional orientation of the gene in each slice; corresponding chromosome-based (global) coordinates are also provided.

### Retrieval of exon sequences and features

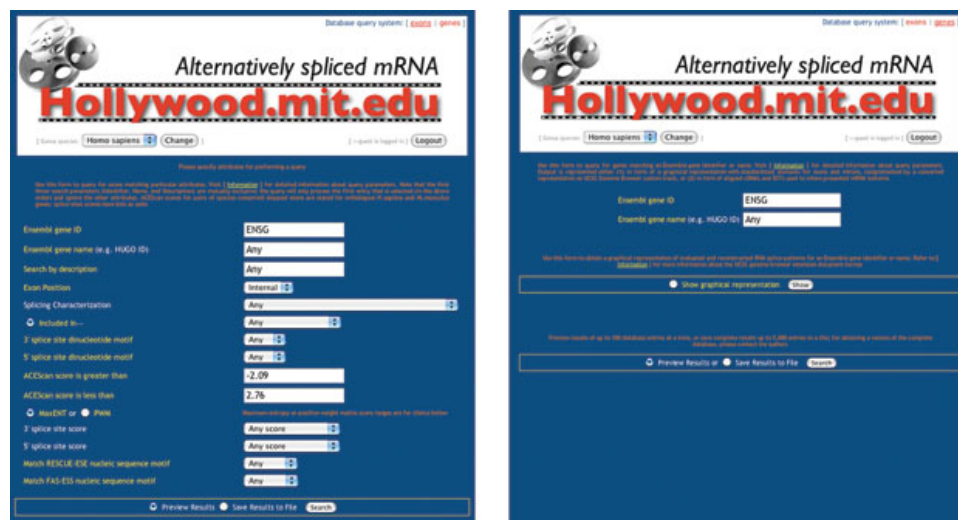
A feature-selection form allows the user of HOLLYWOOD to extract sets of constitutive and/or alternative exons, either for a single gene, specified by its Ensembl gene identifier or name, or for genes that share similar descriptions (e.g. kinases). The user can select exon features such as internal, constitutive or alternative (e.g. skipped), inclusion in transcripts derived from a particular tissue, conservation or presence of a particular ESE hexamer (e.g. GAAGAA). After the user has selected the features and submitted the form, exon records that meet the criteria are retrieved from the database. Figure 1A shows the data structure of such a record for exon E16 of the *FXR1* gene, which undergoes tissue-specific AS (32,56).

As an example, a query with feature selection '*hnRNP*' for gene description, 'Skipped exon' for splice type, and 'Internal' for exon position retrieved ~40 human exons, including skipped exons in genes encoding hnRNP A1, hnRNP C and hnRNP R. A second query for exons with features 'Internal', 'Skipped exon', 'Testis' for tissue type and 'TTCCTT' for ESS

**Figure 1.** (A) Screen shot of the HOLLYWOOD graphical interface summarizing splicing patterns for a single gene. The top of the interface summarizes species, locus, Ensembl-linked gene number and name, gene description, and EST-derived tissue types with corresponding number of occurrences. The annotation performed by HOLLYWOOD is complemented by ACESCAN-predicted splice-conservation and exons that were annotated by Ensembl. Color-coded boxes are used to link display features with explanatory information. For each alternative exon splice type, HOLLYWOOD displays GenBank-linked accession numbers and primary transcript structures of representative pairs of transcripts, which can be used to identify the alternative exons, and layers each structure onto the UCSC Genome Browser. (B) HOLLYWOOD exon record for skipped exon E16 identified in the human fragile X mental retardation syndrome-related (*FXR1*) gene. *FXR1* is an autosomal homolog of the *FXR* gene and encodes an RNA-binding protein. Figure 1 shows data for exon E16, by searching HOLLYWOOD with '*FXR1*' for gene name and 'Skipping' for exon type, with a schematic representation of features at top and the standard text output below. *FXR1* is shown with two isoforms that alternatively skip/include E16. Skipping of E16 results in a shift of the reading-frame that is predicted to alter and shorten the C-terminus of the *FXR1* protein. E16 is an ACESCAN[+] exon with a score of ~0.2, and hence the orthologous exon of the mouse *FXR1* is predicted to undergo exon skipping. This exon is perfectly conserved in sequence and contains two clusters of RESCUE-ESE hexamers. Transcripts aligned to the locus of *FXR1* show E16 included in more than a dozen transcripts (two cDNAs AY341428 and HSU25165, and >10 ESTs) and excluded in ~30 other transcripts (cDNA BC028983 and other ESTs), and ESTs suggest that *FXR1* is expressed in many tissues.



**Figure 2.** Tree representation of numbers of human (left) and mouse (right) alternative and constitutive exons in HOLLYWOOD. All exons are supported by spliced alignments of transcript sequences with minimal acceptor and donor splice sites AG/ and /GT or /GC, respectively. Splice sites required to identify constitutive and alternative exons are marked bolded in black. HOLLYWOOD branches its annotation as follows: (i) on the first level, it distinguishes between first, internal and last exons; (ii) on the next level, it distinguishes between constitutive exons, with constant 3' and 5' splice sites, and alternative exons, with varying 3' and/or 5' splice sites; (iii) on the last level, alternative exons are annotated as skipped, alternative 3' splice sites, alternative 5' splice sites, overlapping, or mutually exclusive exons. In addition, introns that are retained in mature mRNA are annotated as intron retention events. Alternative exons may undergo multiple splice variations and can belong to multiple branches.



**Figure 3.** Web interface offering two feature-selection forms for searches for sets of exons with specific splicing characteristics or splicing regulatory element composition (left), or for the splicing picture for a specific gene (right). Each feature is linked to the online documentation with explanatory information about feature utilization, values or nomenclature. Ensembl gene identifiers are most reliable for querying, as gene names are often not standardized, and response time is typically within seconds, depending on the complexity of the query.

sequence element retrieved ~160 skipped human exons, which included the sperm-specific thioredoxin 2 (*Sptrx*) gene and the gene encoding the spermatogenesis cell apoptosis-related protein 1. A third query for exons with features 'Internal', 'Constitutive exon' and positive ACEScan scores (orthologous exon pairs predicted to be alternatively spliced in both human and mouse). The HOLLYWOOD system retrieved ~1400 ACEs, including exon 5 of the tissue-specific RNA-binding protein NOVA-1, which has recently been identified as a skipped exon that is auto-regulated by NOVA-1 (57). As a final example, HOLLYWOOD was queried first with the feature selection 'Skipped exon' for splice type, 'Brain' and 'GAA-GAA' for ESE sequence element, and was then queried with

'GGTAAG' for ESS sequence element, keeping the remaining features as selected previously. It retrieved ~470 exon records for the first query, and ~50 exon records for the latter query. These examples anticipate some of the types of analysis of alternative exons and functional sequence elements that can be conducted using HOLLYWOOD.

### Viewing AS patterns

In addition to specific sets of exons, queried for by employing the feature selection form, HOLLYWOOD allows the user to display a summary of the splicing information contained in the database for one gene at a time. The user simply queries for

a specific gene by name or by Ensembl identifier, and selects the display option. Alternatively, one may obtain a sequence-based representation of the display, and download all transcripts mapped to the gene locus.

After a gene is selected, a diagram illustrating splicing patterns is computed in real-time and displayed. Such a display for the human *FXR1* gene is shown in Figure 1A, supported by legends. The HOLLYWOOD annotation is presented as a set of evaluated reference exons, each of unit size and color-coded as first, internal or last exon, with internal exons further annotated according to their splicing properties as constitutive or alternative. The HOLLYWOOD annotation shown derives from exons supported by spliced alignments with minimal consensus 3' splice and 5' splice sequences (AG/ and /GT or /GC, respectively). For clarity, the first and last segments of EST alignments (which generally correspond to incomplete portions of exons) are not displayed. Similarly, first and last exons are clustered and represented as first/last 'exon regions' to simplify the diagrams and focus attention on splicing-related (rather than transcription- or polyadenylation-related) information. HOLLYWOOD does not display all obtained spliced alignments (which often number in the hundreds or more), but provides a streamlined representation, displaying pairs of 'representative transcripts', which support the inference of each AS event.

Finally, the HOLLYWOOD display can conveniently be layered onto the UCSC Genome Browser, by using custom tracks that are provided as links on the display, for detailed comparisons with other existing annotations and inspection of multi-species sequence conservation.

## HOLLYWOOD MAINTENANCE AND DEVELOPMENT

The first-phase development of HOLLYWOOD has matured such that the public release 1.0 can serve as a central resource for data and graphical display of patterns of alternative pre-mRNA splicing. Updates, improvements and further developments will be ongoing, and are currently envisioned as extensions for comparative genomics, perhaps with layering onto the Ensembl Genome Browser, integration of external annotations and incorporation of splicing-specific microarray data with links to exon records.

## ACKNOWLEDGEMENTS

The authors thank L. P. Lim and R.-F. Yeh for previous work that contributed to HOLLYWOOD, W.G. Fairbrother and Z. Wang for contributing the datasets of RESCUE-ESE and FAS-ESS sequence elements, respectively, G. Yeo for contributing the maximum entropy model of splice sites and ACESCAN predictions, and U. Ohler for stimulating discussions. This work was supported by grants from the NSF and the NIH (C.B.B.). Funding to pay the Open Access publication charges for this article was provided by NSF and NIH funds.

*Conflict of interest statement.* None declared.

## REFERENCES

- Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.

- Lopez,A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.*, **32**, 279–305.
- Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
- Jurica,M.S. and Moore,M.J. (2003) Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell*, **12**, 5–14.
- Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.*, **3**, 285–298.
- Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
- Grabowski,P.J. and Black,D.L. (2001) Alternative RNA splicing in the nervous system. *Prog. Neurobiol.*, **65**, 289–308.
- Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,J., Baldwin,K., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
- International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
- Kornblihtt,A.R. (2005) Promoter usage and alternative splicing. *Curr. Opin. Cell Biol.*, **17**, 262–268.
- Proudfoot,N. (1996) Ending the message is not so simple. *Cell*, **87**, 779–781.
- Krawczak,M., Reiss,J. and Cooper,D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.*, **90**, 41–54.
- Faustino,N.A. and Cooper,T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev.*, **17**, 419–437.
- Dredge,B.K., Polydorides,A.D. and Darnell,R.B. (2001) The splice of life: alternative splicing and neurological disease. *Nature Rev. Neurosci.*, **2**, 43–50.
- Garcia-Blanco,M.A., Baraniak,A.P. and Lasda,E.L. (2004) Alternative splicing in disease and therapy. *Nat. Biotechnol.*, **22**, 535–546.
- Caceres,J.F. and Kornblihtt,A.R. (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.*, **18**, 186–193.
- Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
- Black,D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.
- Pagani,F. and Baralle,F.E. (2004) Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Rev. Genet.*, **5**, 389–396.
- Stamm,S., Zhu,J., Nakai,K., Stoilov,P., Stoss,O. and Zhang,M.Q. (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol.*, **19**, 739–756.
- Shah,P., Jensen,L.J., Boue,S. and Bork,P. (2005) Extraction of transcript diversity from scientific literature. *PLoS Comput. Biol.*, **1**, 67–72.
- Zheng,C.L., Nair,T.M., Gribskov,M., Kwon,Y.S., Li,H.R. and Fu,X.D. (2004) A database designed to computationally aid an experimental approach to alternative splicing. *Pac. Symp. Biocomput.*, 78–88.
- Boguski,M.S. (1995) The turning point in genome research. *Trends Biochem. Sci.*, **20**, 295–296.
- Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res.*, **31**, 101–105.
- Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and MuiLu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res.*, **32**, D64–D69.
- Pospisil,H., Herrmann,A., Bortfeldt,R.H. and Reich,J.G. (2004) EASED: Extended Alternatively Spliced EST Database. *Nucleic Acids Res.*, **32**, D70–D74.
- Huang,H.D., Hornig,J.T., Lin,F.M., Chang,Y.C. and Huang,C.C. (2005) SpliceInfo: an information repository for mRNA alternative splicing in human genome. *Nucleic Acids Res.*, **33**, D80–D85.
- Kim,P., Kim,N., Lee,Y., Kim,B., Shin,Y. and Lee,S. (2005) ECgene: genome annotation for alternative splicing. *Nucleic Acids Res.*, **33**, D75–D79.
- Lareau,L.F., Green,R.E., Bhatnagar,R.S. and Brenner,S.E. (2004) The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.*, **14**, 273–282.

31. Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754–3766.
32. Yeo,G., Holste,D., Kreiman,G. and Burge,C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol.*, **5**, R74.
33. Taneri,B., Snyder,B., Novoradovsky,A. and Gaasterland,T. (2004) Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. *Genome Biol.*, **5**, R75.
34. Sorek,R., Shemesh,R., Cohen,Y., Basechess,O., Ast,G. and Shamir,R. (2004) A non-EST-based method for exon-skipping prediction. *Genome Res.*, **14**, 1617–1623.
35. Yeo,G.W., Van Nostrand,E., Holste,D., Poggio,T. and Burge,C.B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl Acad. Sci. USA*, **102**, 2850–2855.
36. Ohler,U., Shomron,N. and Burge,C.B. (2005) Recognition of unknown conserved alternatively spliced exons. *PLoS Comput. Biol.*, **1**, e15.
37. Modrek,B. and Lee,C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genet.*, **34**, 177–180.
38. Thanaraj,T.A., Clark,F. and Muilu,J. (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Res.*, **31**, 2544–2552.
39. Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.*, **31**, 5635–5643.
40. Johnson,J.M., Castle,J., Garrett-Engle,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
41. Hu,G.K., Madore,S.J., Moldover,B., Jatkoa,T., Balaban,D., Thomas,J. and Wang,Y. (2001) Predicting splice variant from DNA chip expression data. *Genome Res.*, **11**, 1237–1245.
42. Clark,T.A., Sugnet,C.W. and Ares,M.Jr (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**, 907–910.
43. Neves,G., Zucker,J., Daly,M. and Chess,A. (2004) Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nature Genet.*, **36**, 240–246.
44. Pan,Q., Shai,O., Misquitta,C., Zhang,W., Saltzman,A.L., Mohammad,N., Babak,T., Siu,H., Hughes,T.R., Morris,Q.D. *et al.* (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell*, **16**, 929–941.
45. Blanchette,M., Labourier,E., Green,R.E., Brenner,S.E. and Rio,D.C. (2004) Genome-wide analysis reveals an unexpected function for the *Drosophila* splicing factor U2AF50 in the nuclear export of intronless mRNAs. *Mol. Cell*, **14**, 775–786.
46. Ule,J., Jensen,K.B., Ruggiu,M., Mele,A., Ule,A. and Darnell,R.B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science*, **302**, 1212–1215.
47. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
48. Eng,L., Coutinho,G., Nahas,S., Yeo,G., Tanouye,R., Babaei,M., Dork,T., Burge,C. and Gatti,R.A. (2004) Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: maximum entropy estimates of splice junction strengths. *Hum. Mutat.*, **23**, 67–76.
49. Smith,C.W. and Valcarcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381–388.
50. Fairbrother,W.G., Yeh,R.F., Sharp,P.A. and Burge,C.B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.
51. Wang,Z., Rolish,M.E., Yeo,G., Tung,V., Mawson,M. and Burge,C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.
52. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.
53. Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **16**, 418–420.
54. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
55. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
56. Kirkpatrick,L.L., McIlwain,K.A. and Nelson,D.L. (1999) Alternative splicing in the murine and human FXR1 genes. *Genomics*, **59**, 193–202.
57. Ule,J., Ule,A., Spencer,J., Williams,A., Hu,J.S., Cline,M., Wang,H., Clark,T., Fraser,C., Ruggiu,M. *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse. *Nature Genet.*, **37**, 844–852.