Alternative Splicing Events

I. Elizabeth Cha[1], Katherine L. Hoblitzell[1],
and Eric C. Rouchka[1]
TR-ULBL-2007-03

December 12, 2007

[1]University of Louisville
Speed School of Engineering
Department of Computer Engineering and Computer Science
123 JB Speed Building
Louisville, Kentucky, USA 40292

icha@louisville.edu; Katie.hoblitzell@gmail.com; eric.rouchka@louisville.edu

UNIVERSITY of LOUISVILLE dare to be great
Bioinformatics Laboratory Technical Report Series

Bioinformatics Review

Alternative Splicing Events

I. Elizabeth Cha[1], Katherine L. Hoblitzell[1], and Eric C. Rouchka[1,*]

[1]Department of Computer Engineering and Computer Science, University of Louisville, 123 JB Speed Building, Louisville, KY, USA

UNIVERSITY OF LOUISVILLE BIOINFORMATICS LABORATORY TECHNICAL REPORT SERIES REPORT NUMBER **TR-ULBL-2007-03**

ABSTRACT

Motivation: Alternative splicing is an RNA splicing variation in which the coding regions for a eukaryotic gene, called ***exons***, are extracted and connected to produce alternative mRNA molecules, leading to a number of distinguishable proteins for a single transcriptional region. The resulting mRNA molecules known as ***isoforms*** or ***splice variants***, are becoming increasingly values as an important biological method for providing complexity and diversity. He we present a review of alternative splicing along with its potential role for a number of biological phenomenon.

1 INTRODUCTION

1.1 Intron/Exon Gene Structure

Eukaryotic gene transcription occurs in gene coding regions when a portion of genomic DNA is copied, or transcribed, into messenger RNA (mRNA). This pre-processed mRNA is maintained in the nucleus of the cell. It contains regions called ***exons*** that will eventually code for a protein, and other regions called ***introns*** that are noncoding. Processing of the mRNA removes the introns and splices exons together. A poly-A tail is also added to the end of the mRNA molecule that travels to the cytoplasm where it can be translated into a protein with the help of tRNA and ribosomes (Fig. 1).

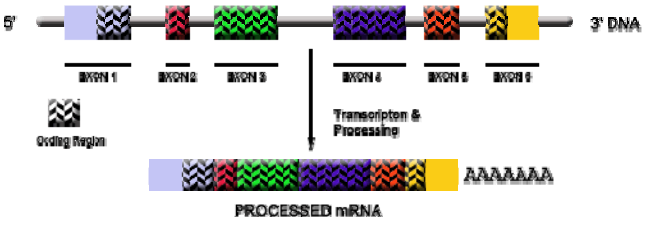

Fig. 1 Typical eukaryotic gene structure.

The exon/intron structure of genes was first described by Phillip A. Sharp of the Massachusetts Institute of Technology (MIT) and Richard J. Roberts of New England Bio-labs in 1977 who were working independently on understanding the structure of adenoviruses (Berget et al., 1977; Chow et al., 1977). In their work (for which they were awarded the Nobel Prize in Physiology or Medicine in 1993 "for their discoveries of split genes"), they found that in primary RNA transcripts, the non-coding regions, called ***introns***, are discarded for the final version of mRNA. In 1980, Randolph Wall discovered that the hypothesis that introns are always discarded and exons always included was not correct through his work with immunoglobulin (Moore et al., 1981; Wall et al., 1981; Early et al., 1980; Rogers et al., 1980; Calame et al., 1980; Rogers and Wall, 1980). At the time this was thought to be an anomaly. However, further research confirmed the findings of alternative splicing methods by which different patterns of exons were used to create alternative protein products. The discovery of alternative splicing was very crucial since it challenged the "one-gene-one-protein" hypothesis that molecular biologists had long believed to be true.

1.2 Alternative Splicing

The biological mechanism which allows for alternative splicing to occur was first described in 1984 as a splice machine now commonly known as a spliceosome (Krainer et al., 1984). The spliceosome is a complex of five small nuclear uridine rich (snRNA) molecules labeled U1, U2, U4, U5 and U6 along with protein factors. Spliceosomes recognize exon/intron boundaries through a 5' donor site, a 3' acceptor site, a branch point, and a polypyrimidine tract.

The canonical donor splice site recognized by the above spliceosome complex is marked by the nucleotides GT, while the canonical 3' acceptor site is AG. Other patterns confirmed within mammalian genomes include GC/AG, AT/AC, GT/GG, and TT/AG (Burset et al., 2001; Burset et al., 2000). The canonical pattern GT/AG is predicted to occur 99.24% of the time, with GC/AG occurring 0.69% and AT/AC 0.05% with all others accounting for 0.02% of all splice sites in mammals (Burset et al., 2001; Burset et al., 2000). The pattern AT/AC has been shown to result from a separate spliceosome complex involving U11 and U12 snRNAs.

At the beginning of canonical splicing, the U1 snRNA binds to the 5' end of the splice site which recruits several protein factors to form the commitment complex (Jamison et al., 1995; Jamison and Garcia-Blanco, 1992; Seraphin et al., 1988). This causes the branch site to attach to the donor's G nucleotide to form a phos-

*To whom correspondence should be addressed.

phodiester linkage. As part of this process, the U2 snRNA binds to the branch site. Afterwards, the U4-U5-U6 complex arrives and U6 replaces the U1 position. U1 and U4 disassociate from the complex, and U2 and U6 associate to for the lariat intron that gets removed. U5 aids in merging the two adjoining exons together. A diagram of the canonical spliceosome is given in Fig. 2.

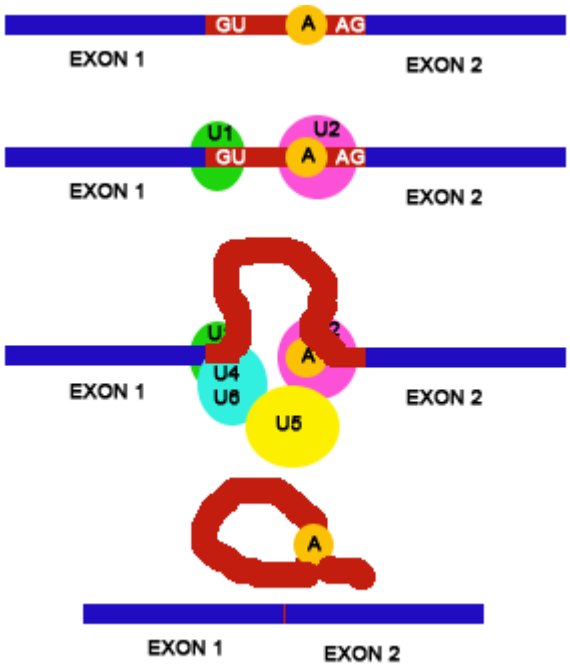

Fig. 2 Canonical spliceosome for intron splicing.

1.3 Alternative Splicing Events

Alternative splicing of gene coding regions works to produce alternative processed mRNA sequences. This may occur in such a way to only affect the 5' or 3' untranslated regions (UTRs), or it may alter the actual coding region, resulting in an altered translated protein. The possible alternative splicing events shown in Fig. 3 can be summarized as belonging to one of seven classes: alternative promoter selection, alternative polyadenylation sites, exon skipping, exon extension, exon truncation, exon retention, and intron retention.

Alternative promoter selection. In this method of splicing, alternative aminyl ends of the processed mRNA are produced by alternatively selecting promoter regions. These regions occur in the 5' untranslated region (UTR) of the mRNA. Such events do not alter the final translated protein product.

Alternative polyadenylation sites. In some cases, the 3' UTR of the processed mRNA is alternatively spliced producing alternate polyadenylation sites. Once again, these events will not alter the final translated protein product.

Exon skipping. This method of alternative splicing works by choosing certain exons to remove from the final processed mRNA product. As a result, each exon can be thought of as a "cassette" that can potentially be removed. Exon skipping results in an altered translated proteins (Sharp, 2005; Gupta et al., 2004a; Kan et al., 2002). This is the most commonly found version of alternative splicing in mammals.

Exon extension. In exon extension, certain exons may be altered by adding in additional sequence to a transcript directly around a cassette exon. This will result in a slightly modified mRNA sequence (Sharp, 2005; Gupta et al., 2004a; Kan et al., 2002).

Exon truncation. Exon truncation occurs when a complete cassette exon is not fully added to a mRNA sequence. In effect, either the 5' or 3' end (or both) of the exon is removed, resulting a an altered translated protein (Modrek and Lee, 2002).

Exon retention. In exon retention, extra exons located in what is normally an intronic region, are added to the final mRNA sequence, resulting in a different translated protein (Modrek and Lee, 2002).

Intron retention. This method of alternative splicing occurs when an intron is retained in the final mRNA transcript (Sharp, 2005; Gupta et al., 2004a; Kan et al., 2002). As long as a stop codon is not found in the intron, the resulting translated protein will be al-

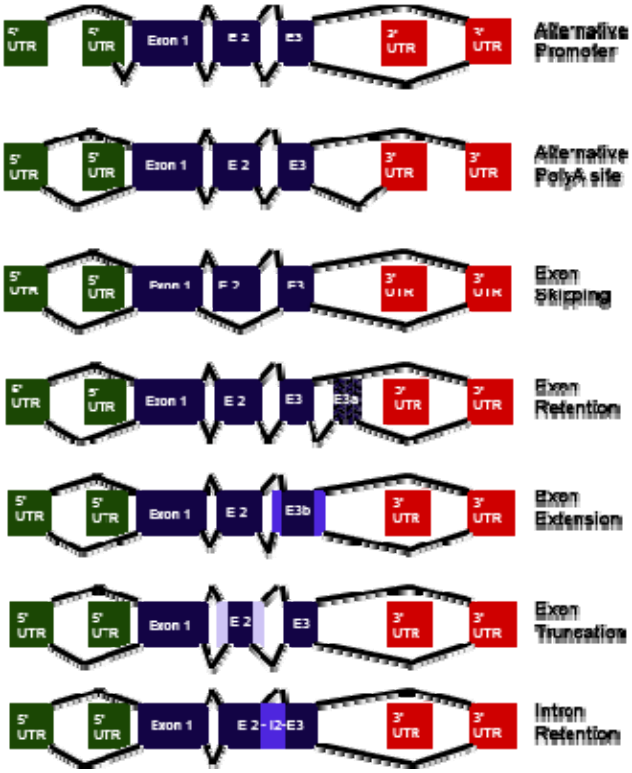

Fig. 3 Alternative Splicing Events.

tered. Intron retention is commonly found in plants and in lower multicellular organisms.

2 ROLES OF ALTERNATIVE SPLICING

Alternative splicing allows for a potential for multiple proteins to be produced by a single gene coding region, yielding a highly efficient genetic mechanism. This can also allow for eukaryotic organisms to adapt and evolve at a much faster rate due to the cassette nature. In this manner, a gene can evolve much faster by simply adding or removing exons at a faster rate than adding or removing whole genes. Alternative splicing has been implicated in protein structure evolution (Birzele et al., 2007).

Beyond just providing a method for variation amongst organisms, alternative splicing may play an important role in genetic diseases as well as cancer (Kim et al., 2008; Garcia-Blanco et al., 2004). There has been research to imply that diseases, such as growth deficiencies, Frasier Syndrome (Klamt et al., 1998), cystic fibrosis (Hull et al., 1994), frontotemporal dementia (Kar et al., 2005), Parkinsonism (Buee et al., 2000), endometriosis (Fujino et al., 2006), Ehlers-Danlos syndrome (Takahara et al., 2002) and certain cancers such as hereditary nonpolyposis colorectal (De et al., 2007) or prostate (Hayes et al., 2007) could be caused by alternative splicing.

A number of studies suggest alternative splicing aids an organism's complexity by adding in more protein coding potential (Lander et al., 2001; Venter et al., 2001; Ewing and Green, 2000). However, separate analyses of eukaryotic transcriptomes using EST analysis (Brett et al., 2002) and Unigene and Homologene clusters (Kim et al., 2007) yield conflicting results. More complete EST sequencing of a number of genomes will only help to clarify this issue in the future.

3 RATES OF ALTERNATE SPLICING

Initially, alternative splicing was thought to be an anomaly (Wall et al., 1981). In the early 1990s, it was predicted that about 5% of all human genes provided alternatively spliced isoforms (Sharp, 1994). By the late 1990s, it was thought that alternative splicing is much more prevalent than previously thought, with the potential for 50% of human genes to undergo alternative splicing (Brett et al., 2002; Mironov et al., 1999). In recent studies it has been found that up to 74% of human genes are alternatively spliced (Blaustein et al., 2007).

The average mammalian gene has between eight and nine exons. While this may on the surface may seem small, this gives an average of 256 different cassette exon patterns. The largest number of known isoforms for a single coding region is the Drosophila DScam gene which has 38,016 distinct forms (Zipursky et al., 2006).

The possibility that human genes can produce a number of different proteins from a single gene leads to understanding of the complexity of such organisms that still have a relatively small percentage of variation amongst DNA. Within humans the similarity is measured to be around 99.5 -99.8 %. This leaves a small difference between .2-.5% of variation. Yet this difference is up to 15 million different nucleotides in humans (Kidd et al., 2004). Coupled with alternative splicing, variation is inflated, making it understandable that even though human DNA is only 1% different from a chimpanzee, the two species may have a wide gap in genetic expression.

4 METHODS FOR DETECTING ALTERNATIVE SPLICING

There are many factors that can be used to detect alternative splicing. The most popular methods for identifying alternative splicing sites are based around ESTs, microarrays, sequence length and AG-GT splicing site recognition.

4.1 EST Detection

ESTs were one of the first computational forms of identifying alternative splicing sites (Kan et al., 2001; Kan et al., 2000). Since ESTs represent segments of transcribed cDNAs, it is relatively straightforward to map them back onto genomic regions along with mRNA sequences to observe alternative patterns. ESTs, however, contain a lot of "extra" sequencing which can cause many alternative splicing hits to be artifacts of old sequence that is not expressed. In addition, EST sequencing does not have a uniform coverage among transcripts or genomes, so many isoforms go undetected (Gupta et al., 2004b).

4.2 Microarray Detection

Microarrays are one of the more definitive ways to study alternative splicing in the form of genome tiling arrays (Bertone et al., 2004) and exon junction arrays (Johnson et al., 2003). In model processes where characteristics are assumed there is a region for error and falsely classifying. With microarrays, mRNA can be compared to give a definite answer to the question whether a gene is alternatively spliced. The problem with microarrays is they may prove to be more time and resource consuming, particularly if tissue specificity is considered. If an alternative splice is created in a small percentage or only in a certain condition (which is often the case with tissue specific splicing) there may not be representatives of the alternative splicing in the data.

4.3 Classifying Features

There has been further research to indicate that in many cases of alternative splicing, alternatively spliced regions contain internal exons followed by conserved intronic sequences (Sorek et al., 2004). These sequences were not only found in the human genome but also in the mouse genome. It is believed that these intronic sequences could play a large role in facilitating the alternative splicing. Sorek et al. studying the correlating features between alternative spliced genes using the following features:

(1) Exon length. This plays a large role in identifying as most alternatively spliced exons are shorter than the conserved ones.

(2) Divisibility by three. This is fairly obvious, because obviously codons are arranged as three nucleotides. Therefore

by alternatively spliced exons being arranged as multiples of three it will not negatively affect transcription when certain sequences are coded or left uncoded.

(3) Conservation of alternatively spliced regions with the corresponding mouse sequences.

(4) Conservation of upstream and downstream intronic sequences.

These characteristics were combined to form a rule where there was a 95% identity with the mouse exon, the exon was a multiple of three, the local alignment was at least 15 intronic nucleotides upstream of the exon with 85% identity and matched perfectly with twelve consecutive intronic nucleotides downstream. This rule only matched 31% of the alternatively spliced sequence in the test data, but did not match any of the conservatively spliced sequences. This rule was validated and proved to have a 99.72% specificity and a 32.3% sensitivity. When validating the rule by applying it to a much wider range of sequences, there were many cases where alternatively spliced exons were named that have no EST backing. 60% of the classifications did seem to have further evidence with most of the other 40% still exhibiting some type of splicing abnormality.

5 TISSUE SPECIFIC ISOFORMS

In more recent studies, the relationship between tissue type and alternative splicing has been explored. Tissue type is very important when analyzing the expression of gene and proteins in an organism. An organism has only one genome, therefore each cell contains the same DNA, the same mapping for proteins. What makes tissue cells different are their functions and their ability to express different parts of the DNA. Therefore there needs to be methods to control the expression of genes. Alternative splicing is now being investigated for its ability to control gene expression which could translate to tissue type control of gene expression. Problems arise when this alternative splicing does not happen correctly and tissue types begin to dedifferentiate which is what creates cancer. Though alternative splicing is not the foremost or only cause of cancer it is now being used to target with drugs, serve as a marker for diagnosis and study many cancers.

Tissue specific genes are genes that create proteins only for specified tissue types. About 58% of genes tested to be specific to one tissue type according to the following equations (Noh et al., 2006):

$$p(y|x) = \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x!y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}} \quad (1)$$

In equation 1, N_2 is the number of ESTs from specific tissues, N_1 is ESTs from other tissues, making $N = N_1 + N_2$ equal to the total

$$Pvalue = \sum_y^n p(y|x) = \sum_y^n \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x!y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}} \quad (2)$$

amount of ESTs. x + y = n with y being the number of ESTs in a specific tissue type from a certain gene and x being the other ESTs from that gene. With these two variables the equations 1 and 2 are

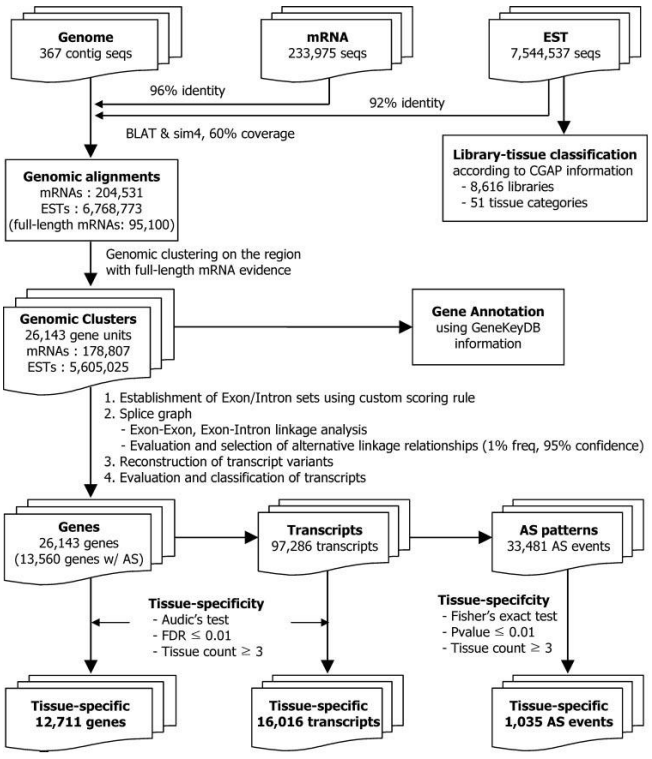

Fig. 4 Pipeline for detecting tissue specific isoforms (Noh et al., 2006).

formed along with a null hypothesis. In order for the null hypothesis to be disproven the P-value must be lower than .01. The null hypothesis states that the ratio of tissue specific ESTs in a gene is proportional to the amount of tissue specific ESTs. Disproving this hypothesis basically means that a gene has been found with a high confidence that seems to code for one tissue type and does not follow the averages.

In order to construct a model of tissue specific alternative splicing a p value was also require to measure whether an alternative splice pattern met a given threshold of support.

$$P(|Y| \geq k) = \sum_{i=k}^{n} P(i); P(i) = \frac{n!}{i!(n-i)!} f^i (1-f)^{n-i} \quad (3)$$

In equation three, k is the number of ESTs that support the alternative splice, and n is the total amount of ESTs supporting all splice patterns. The variable f represents the frequency of the splicing. If P(|Y| >=k|n,f) <.05 then a 95% confidence interval has been reached and the criteria for alternative splicing has been met.

These equations along with a few other methods have contributed to making the model used by Noh et al. to detect tissue specific alternative splicing is seen in Fig. 3.

6 COMPARATIVE GENOMICS

Comparative Genomics has been a very useful tool in studying alternative splicing. Many studies have used the comparison between human DNA and mouse DNA to get different models of alternative splicing. This gives more support to the alternative splicing sites. This was used in the model that was discussed in

section 5. It demonstrates that tissue specific alternative splicing is not exclusively held to humans but is actually quite prevalent in mice as well. Mice demonstrate a higher percentage of tissue specific alternative splicing events. What was most interesting from this data is how cassette exons in both were found to be the main alternative splicing event Nothing else came close to the number of total alternative splicing events (in humans this was 13,191). The percentage of tissue-specific events was also slightly higher at 4.2 for humans. The second most prevalent alternative splicing event was intron retention though it was not highly tissue specific with only 2.1 % in humans. Overlapping exons were by far the smallest percentage amongst all of the alternative splicing events. Cassette exons prove to be the most important factor in alternative splicing.

7 DISCUSSION

Alternative splicing of mRNA transcripts is critical to providing gene diversity within eukaryotic organisms. As more high-throughput data becomes available in the form of EST sequences and tiling arrays, the role, frequency, and tissue specificity of isoforms will become more apparent.

ACKNOWLEDGEMENTS

ER is supported by NIH-NCRR grant P20RR16481 and NIH-NIEHS grant P30ES014443. The contents of this manuscript are solely the responsibility of the authors and do not represent the official views of NCRR, NIEHS, or NIH.

REFERENCES

Berget,S.M., Moore,C. and Sharp,P.A. (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A*, 74, 3171-3175.

Bertone,P. et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306, 2242-2246.

Birzele,F., Csaba,G. and Zimmer,R. (2007) Alternative splicing and protein structure evolution. *Nucleic Acids Res.*

Blaustein,M., Pelisch,F. and Srebrow,A. (2007) Signals, pathways and splicing regulation. *Int. J. Biochem. Cell Biol.*, 39, 2031-2048.

Brett,D. et al. (2002) Alternative splicing and genome complexity. *Nat. Genet.*, 30, 29-30.

Buee,L. et al. (2000) Tau protein isoforms, phosphorylation and role in neurodegenerative disorders. *Brain Res. Brain Res. Rev.*, 33, 95-130.

Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, 28, 4364-4375.

Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2001) SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res.*, 29, 255-259.

Calame,K. et al. (1980) Mouse Cmu heavy chain immunoglobulin gene segment contains three intervening sequences separating domains. *Nature*, 284, 452-455.

Chow,L.T. et al. (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12, 1-8.

De,R.M. et al. (2007) Alternative splicing and nonsense-mediated mRNA decay in the regulation of a new adenomatous polyposis coli transcript. *Gene*, 395, 8-14.

Early,P. et al. (1980) Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell*, 20, 313-319.

Ewing,B. and Green,P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.*, 25, 232-234.

Fujino,K. et al. (2006) Transcriptional expression of survivin and its splice variants in endometriosis. *Mol. Hum. Reprod.*, 12, 383-388.

Garcia-Blanco,M.A., Baraniak,A.P. and Lasda,E.L. (2004) Alternative splicing in disease and therapy. *Nat. Biotechnol.*, 22, 535-546.

Gupta,S. et al. (2004a) Genome wide identification and classification of alternative splicing based on EST data. *Bioinformatics*, 20, 2579-2585.

Gupta,S. et al. (2004b) Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC. Genomics*, 5, 72.

Hayes,N.V. et al. (2007) Identification and characterization of novel spliced variants of neuregulin 4 in prostate cancer. *Clin. Cancer Res.*, 13, 3147-3155.

Hull,J., Shackleton,S. and Harris,A. (1994) Analysis of mutations and alternative splicing patterns in the CFTR gene using mRNA derived from nasal epithelial cells. *Hum. Mol. Genet.*, 3, 1141-1146.

Jamison,S.F. and Garcia-Blanco,M.A. (1992) An ATP-independent U2 small nuclear ribonucleoprotein particle/precursor mRNA complex requires both splice sites and the polypyrimidine tract. *Proc. Natl. Acad. Sci. U. S. A*, 89, 5482-5486.

Jamison,S.F. et al. (1995) U1 snRNP-ASF/SF2 interaction and 5' splice site recognition: characterization of required elements. *Nucleic Acids Res.*, 23, 3260-3267.

Johnson,J.M. et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 302, 2141-2144.

Kan,Z. et al. (2000) UTR reconstruction and analysis using genomically aligned EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8, 218-227.

Kan,Z. et al. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, 11, 889-900.

Kan,Z., States,D. and Gish,W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res.*, 12, 1837-1845.

Kar,A. et al. (2005) Tau alternative splicing and frontotemporal dementia. *Alzheimer Dis. Assoc. Disord.*, 19 Suppl 1, S29-S36.

Kidd,K.K. et al. (2004) Understanding human DNA sequence variation. *J. Hered.*, 95, 406-420.

Kim,E., Goren,A. and Ast,G. (2008) Insights into the connection between cancer and alternative splicing. *Trends Genet.*, 24, 7-10.

Kim,E., Magen,A. and Ast,G. (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.*, 35, 125-131.

Klamt,B. et al. (1998) Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of

WT1 +/-KTS splice isoforms. *Hum. Mol. Genet.*, 7, 709-714.

Krainer,A.R. et al. (1984) Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. *Cell*, 36, 993-1005.

Lander,E.S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921.

Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, 9, 1288-1293.

Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, 30, 13-19.

Moore,K.W. et al. (1981) Expression of IgD may use both DNA rearrangement and RNA splicing mechanisms. *Proc. Natl. Acad. Sci. U. S. A*, 78, 1800-1804.

Noh,S.J. et al. (2006) TISA: tissue-specific alternative splicing in human and mouse genes. *DNA Res.*, 13, 229-243.

Rogers,J. et al. (1980) Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell*, 20, 303-312.

Rogers,J. and Wall,R. (1980) A mechanism for RNA splicing. *Proc. Natl. Acad. Sci. U. S. A*, 77, 1877-1879.

Seraphin,B., Kretzner,L. and Rosbash,M. (1988) A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. *EMBO J.*, 7, 2533-2538.

Sharp,P.A. (1994) Split genes and RNA splicing. *Cell*, 77, 805-815.

Sharp,P.A. (2005) The discovery of split genes and RNA splicing. *Trends Biochem. Sci.*, 30, 279-281.

Sorek,R. et al. (2004) A non-EST-based method for exon-skipping prediction. *Genome Res.*, 14, 1617-1623.

Takahara,K. et al. (2002) Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-alpha1(V) N-propeptides and Ehlers-Danlos syndrome type I. *Am. J. Hum. Genet.*, 71, 451-465.

Venter,J.C. et al. (2001) The sequence of the human genome. *Science*, 291, 1304-1351.

Wall,R. et al. (1981) RNA processing in immunoglobulin gene expression. *Cold Spring Harb. Symp. Quant. Biol.*, 45 Pt 2, 879-885.

Zipursky,S.L., Wojtowicz,W.M. and Hattori,D. (2006) Got diversity? Wiring the fly brain with Dscam. *Trends Biochem. Sci.*, 31, 581-588.