Tel Aviv University

SACKLER SCHOOL OF MEDICINE

DEPARTMENT OF HUMAN GENETICS AND MOLECULAR MEDICINE

The characteristics, regulation and evolution of human alternatively spliced exons

THESIS SUBMITTED TO THE SENATE OF TEL-AVIV UNIVERSITY FOR THE DEGREE OF "DOCTOR OF PHILOSOPHY" BY

Rotem Sorek

January 2006

THIS WORK WAS CARRIED OUT UNDER THE SUPERVISION OF

Dr. Gil Ast

and

Prof. Ron Shamir

To my parents,

ACKNOWLEDGMENT

I would like to thank Dr. Gil Ast, my primary supervisor, for his great support, understanding, and advice; for setting up the best conditions possible that allowed my PhD studies to take place; and for optimistically pushing forward the research. I would also like to thank to Prof. Ron Shamir, my secondary supervisor, for his help in the computational and algorithmic aspects, and for equipping me with valuable advice at the right times. I have learned from both my supervisors the values of scientific integrity and persistence, and for this I am in deep gratitude.

Next, I would like to thank Galit Lev-Maor, for adding a great value of experimental work and ideas to my studies; and also to Noam Shomron, Amos Tanay, Adi Maron-Katz, and Irit Gat-Viks for their ideas and friendship throughout my PhD studies.

Also, I would like to thank the colleagues in Compugen with whom I worked, primarily Amit Novik, Dvir Dahary Pini Akiva, Amir Toporik, Ronen Shemesh, and many others, for their friendship and support.

Last but not least, I would like to thank my wonderful parents and my beloved family, my wife and best friend Zohar for being so supportive and encouraging in good and bad times, and my son Uri just for being.

PREFACE

This thesis is based on the following collection of seven articles that were published throughout the PhD period in leading scientific journals.

1. Sorek R., Ast G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. ***Genome Research*** 13(7):1631-1637.

2. Sorek R., Shamir R., Ast G. (2004) How prevalent is functional alternative splicing in the human genome? ***Trends in Genetics*** 20(2):68-71.

3. Sorek R., Shemesh R., Cohen Y., Basechess O., Ast G., Shamir R. (2004) A non-EST based method for exon-skipping prediction. ***Genome Research*** 14(8): 1617-1623.

4. Dror G., Sorek R., Shamir R. (2005) Accurate identification of alternatively spliced exons using support vector machine. ***Bioinformatics*** 21(7):897-901.

5. Lev-Maor G.*, Sorek R.*, Shomron N., Ast G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. ***Science*** 300(5623):1288-91 **(* equal contribution).**

6. Sorek R.*, Lev-Maor G.*, Reznik M.*, Dagan T., Belinky F., Graur D., Ast G. (2004) Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. ***Molecular Cell*** 14(2):221-231 **(* equal contribution)**.

7. Dagan T.*, Sorek R.*, Sharon E.*, Ast G., Graur D. (2004) AluGene: a database of Alu elements incorporated within protein-coding genes. ***Nucleic Acids Research*** 32: D489-D492. **(* equal contribution).**

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

CHAPTER 2: ARTICLES COLLECTION

CHAPTER 3: DISCUSSION

ABSTRACT

Alternative splicing increases protein diversity by allowing multiple, sometimes functionally distinct proteins to be coded from the same gene. It can be specific to tissues, stress conditions, and developmental and pathological states. Analyses of Expressed Sequence Tags (ESTs), which are currently the most widely used tool for predicting alternative splicing, revealed that this phenomenon is extremely abundant in metazoan genes. Despite the thousands of splicing events identified so far, regulatory mechanisms were determined only for a small subset of these events, leaving the regulation of most alternative splicing events uncharacterized.

Through the course of my PhD, I used computational tools to identify sequences that regulate alternative splicing and to study the evolution of new alternatively spliced exons. The study was conducted along two parallel lines: Comparative-genomics-based approach, and study of primate specific alternatively spliced exons derived from Alu retrotransposon elements. In the comparative genomics approach I compared the human genome and the recently sequenced mouse genome. As the human and mouse genomes have diverged significantly, sequences showing high human-mouse conservation are indicative of their functionality. I used this feature aiming at detecting novel sequences that regulate alternative splicing. In the second approach I used *Alu* retrotransposons, which are widespread mobile genetic elements that are unique to primates and are found in more than 1 million copies in the human genome. In a previous study I have shown that these elements can form novel alternatively spliced

exons in the human genome, and this was the basis for the second part of the research presented here.

Using comparative genomics, I revealed that intronic sequences near alternatively spliced exons are highly conserved between mammals, and contain motifs recognized by splicing regulatory proteins, indicating that such elements are involved in the regulation of alternative splicing. This discovery led to the development of a novel comparative genomics-based method to identify alternatively spliced exons without using ESTs. This also enabled the development of a method to distinguish between functional and non-functional alternative splicing, which resulted in the unexpected finding that many alternative splicing events observed in EST databases might be non-functional.

Through studying Alu elements that became exons, we have discovered the nucleotide positions that need to be changed in order to create an exon out of an intronic Alu element, and characterized the mechanisms underlying the regulation of their alternative splicing. Consequently, we were able to demonstrate how a small set of mutations in intronic Alu elements results in the creation of new exons. We further showed that such births of novel exons could either be the molecular basis for human genetic disorders, or increase the coding capacity of the human genome in the course of human evolution, depending on the type of mutation. This part of my research was conducted in collaboration with Galit Lev-Maor from the Ast's laboratory, who performed the laboratory (“wet”) experimentation.

Together, the results of these studies (described in the seven articles comprising this thesis), form a significant contribution to the current knowledge on alternative splicing regulation and evolution. Moreover, these results demonstrate that the combination of computational and experimental approaches is a powerful tool to study complex biological phenomena such as alternative splicing.

1. INTRODUCTION

1.1 RNA splicing

Most protein-coding genes from eukaryotes are interrupted by non-coding intervening sequences (introns). In the process of splicing, which occurs in the nucleus**,** the introns are cleaved out of the pre-mRNA and exons are joined together [1]. Splicing involves at least five small nuclear RNAs (snRNAs) and more than 140 different proteins [3]. The snRNAs and proteins are organized in a large ribonucleoprotein (RNP) particle called the spliceosome [4]. At least four conserved sites within the intron are recognized by the snRNAs and the spliceosomal proteins during splicing [5]. These are the 5' splice site (donor), the 3' splice site (acceptor), the poly pyrimidine tract (PPT) and the branching point [6, 7].

Two catalytic events occur during splicing. In the first, an adenosine residue in the branching point attacks the 5’ splice site. In the second, the free exon in the 5' attacks the 3' splice site. These events lead to the cutting of the intron and the ligation of the two flanking exons to each other [1].

1.2 Alternative splicing

The publication of the sequence of the human genome revealed that the gene count in humans is much lower than previously estimated [8]. While textbooks usually quoted the number 100,000 as the number of human genes, current estimations are that the human genome contains only up to 25,000 protein-coding genes [9]. How the great complexity of the human organism can be explained by this low number, which is less than twice the number of genes in the primitive worm C. elegans? The discrepancy is

now explained in part by the recent discoveries that at least half of all human genes undergo alternative splicing [8, 10-12].

Alternative splicing is a process by which several mRNA isoforms can be generated from a single gene. It occurs by using different 5'-3' pairs of splice sites while processing different molecules of the same pre-mRNA [5]. Alternative splicing involving coding exons results in the production of several proteins from the same mRNA [13]. Some alternative splicing events are obligatory, causing the production of two or more mRNA variants at constant ratios in all cells in which the gene is expressed. Other events are facultative and depend on such factors as sex, cell type, developmental stage, or physiological signals [14, 15]. Alternative splicing increases the diversity of the protein inventory within and between cells, and adds an additional regulatory dimension to the expression pattern of the organism [15, 16].

Alternative splicing events can be divided into different types, i.e., intron retention, exon skipping (also called cassette exon), alternative donor site, and alternative acceptor site (Figure 1.1). The most frequent type of alternative splicing is exon skipping, which accounts for approximately 38% of all alternative splicing events [17].

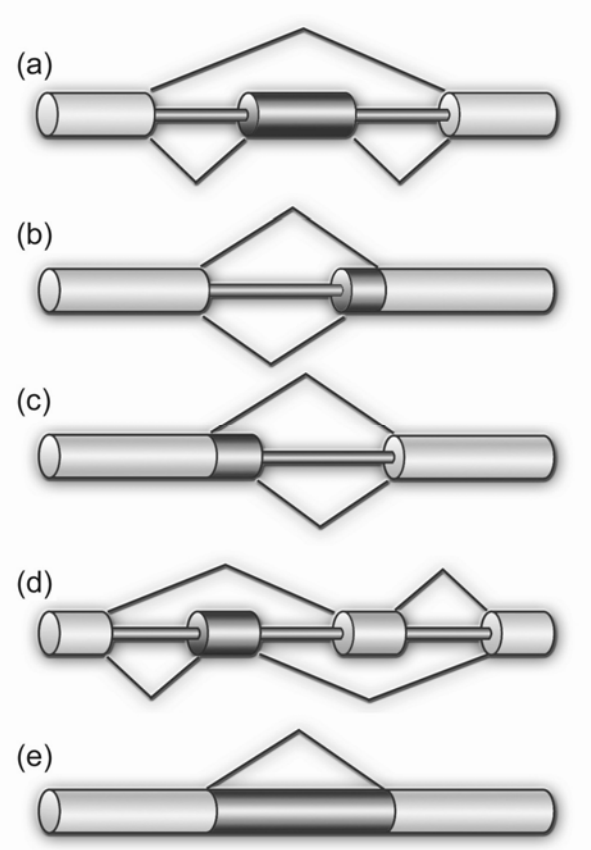


Figure 1.1: Types of alternative splicing. Exons are represented as wide cylinders, introns are thin cylinders. Alternative exons are dark. Alternative splicing pathways are shown by the diagonal lines above and below the gene. (a) Exon skipping. (b) Alternative 3' splice site (acceptor). (c) Alternative 5' splice site (donor). (d) Mutually exclusive exons. (e) Intron retention.

1.3 Prediction of alternative splicing using expressed sequence tags

The finding that alternative splicing could be deduced from alignments of expressed sequence tags (ESTs) and full-length mRNAs was a major breakthrough in alternative splicing research [14]. An EST represents a part of an mRNA, and is the

result of sequencing part of a cDNA clone that was generated from an mRNA [18]. The largest public collection of ESTs is dbEST [19], which is a division of GenBank that currently (as of December 2005) contains more than 31 million sequences, including more than 7 million ESTs from human.

Using multiple sequence alignment algorithms, the millions of publicly available ESTs and cDNAs can be utilized to identify alternative splicing events. Indeed, numerous independent studies conducted since 1999 have reported that 40%-75% of all human genes undergo alternative splicing, with an average of about three splice variants per gene [8, 10-12, 20]. However, as ESTs are only a sample of the transcriptome, they cannot report on all possible splice variants.

Recently, several splicing-specific microarrays were developed [20, 21]. By using probes from the exon-exon junction, these microarrays can measure differences in the expression of specific splice variants. Such microarrays have led to the discovery of numerous novel events of alternative splicing that were not indicated by ESTs [20]. However, even microarray experiments are not sufficient for the identification of all splice variants, as they do not sample all combinations of possible tissues, developmental stages and conditions. One of the aims of my thesis was to develop computational tools that would identify alternative splicing events without relying on experimental evidence such as ESTs or microarrays.

1.4 Regulation of alternative splicing

Despite the thousands of alternative splicing events identified to date, not much is known about the regulation of this process [13]. *Cis*-acting sequence elements found within exons, called exonic splicing enhancers (ESEs), show the ability to regulate alternative splicing [6, 22]. These sequences interact with splicing factors of the SR protein family, which recruit the splicing machinery to weak flanking splice sites, and enable the inclusion of the alternatively spliced exon in the mature mRNA [6]. Other *cis*-regulatory elements, such as exonic splicing silencers (ESS) and intronic splicing enhancers and silencers (ISE and ISS, respectively), were also shown to regulate individual cases of alternative splicing [13]. Sometimes multiple *cis*-acting elements cooperatively function to regulate the same alternatively spliced exon [23]. These *cis*-acting elements are relatively short, usually 4 to 10 nucleotides [6, 23, 24], and are generally found up to 150 bases from the splice site they regulate. Alternative splicing was also found to be affected by other factors, including the phosphorylation state of SR proteins, shuttling of SR and other proteins between the nucleus and the cytoplasm [5], and the promoter of the gene [25].

Although several *cis*-regualtory elements have been characterized for individual cases, it is still unclear how the majority of alternative splicing events are regulated [13, 26]. One of the aims of my Ph.D. thesis was to exploit the vast amount of sequence data currently existing in the public databases to discover more *cis*-acting elements that regulate alternative splicing.

1.5 The mouse genome

The draft sequence of the mouse genome [9] facilitates a great deal of advance in searching for sequence elements that regulate alternative splicing. The estimated 75 million years that passed since the divergence of the ancestor of human and mouse lineages have led to sequence conservation of functionally orthologous regions, while non-functional sequences in each genome had a faster rate of evolution [9]. Indeed, homologous human and mouse exons are, on the average, 85% identical in their sequence, while introns are much poorly conserved ranging between 35% and 69% identity [9, 27].

In spite of the general low similarity of non-coding sequences, numerous small regions that are conserved between human and mouse were recently found in introns [28-30]. The vast majority of them are not expressed and they, therefore, do not represent new protein-coding or RNA genes [29]. Some of these sequences are thought to have regulatory properties, because the high sequence conservation is probably an indication of biological importance. Their function, however, is still unknown.

Another aim of my Ph.D. thesis was to use human-mouse comparative genomics analysis to identify highly conserved intronic sequences that might be involved in the regulation of alternative splicing.

1.6 Alternatively spliced *Alu* elements

Alu elements are short interspersed elements (SINEs) of about 300 nucleotides, which amplify in primate genomes through a process of retroposition [31-33]. These elements have reached a copy number of about 1.4 million in the human genome, comprising more than 10% of it [8]. A typical Alu is built of two similar sequence elements (left and right arms) that are separated by a short A-rich linker. Most Alus have a long poly-A tail of about 20-100 bases [34].

It was shown that parts of Alu elements, predominantly on their antisense orientation, could be inserted into mature mRNAs by way of splicing (‘exonization’). This exonization process is facilitated by sequence motifs that resemble splice sites, which are found within the Alu sequence (Figure 1.2) [2, 35].

We have previously shown that most, if not all, Alu containing exons are alternatively spliced, and that more than 5% of human alternatively spliced exons are Alu-derived [35]. Since Alu elements are unique to primates, the Alu-exons might contribute to some of our human characteristic features. Exonized Alus are sequences sharing a similar ancestor. Each was inserted into a different intron and had undergone independent mutational changes that led to its exonization. This makes them a unique repository containing information on alternative splicing regulation. In the course of my Ph.D. thesis I have studied how Alu exons are born in the genome, and used exonized Alu elements to identify candidate regulatory sequences for alternative splicing.

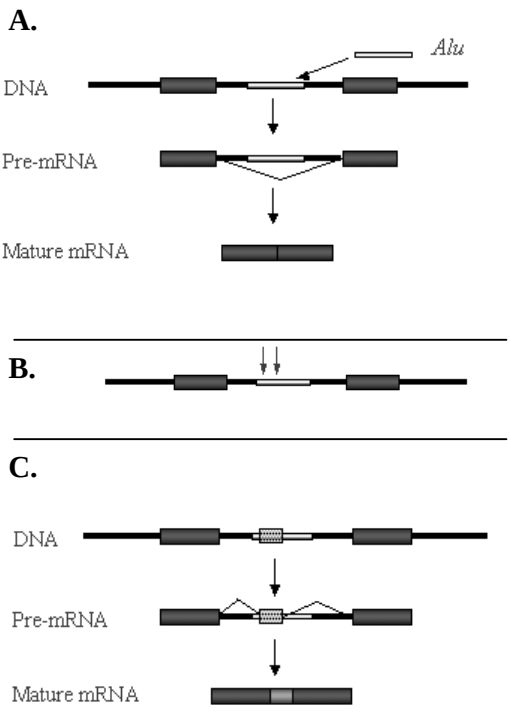


Figure 1.2: A model for the exonization of part of an Alu sequence, proposed by Makalowski et al. [2]. (a) An Alu element (thin light-shaded rectangle) is retroposed into the intron of a two-exon model gene. (b) Since it is genetically functionless, Alu has a high substitution rate. Mutations within the Alu sequence (marked as arrows) activate cryptic splice sites, i.e., segment of a sequence that resembles a splice site mutates into a functional splice site (c) The splicing mechanism recognizes the splice sites, and therefore part of the Alu sequence is exonized (the new exon is marked as a light-shaded box).

1.7. Summary of articles included in this thesis

This PhD research involved computational analyses to study the characteristics, regulation and evolution of alternative splicing in the human and mouse genomes. In general, this study streams along two lines of research: (A) Using comparative genomics methodologies to understand the differences between alternative and constitutive exons, and to reveal sequences involved in regulation of alternative splicing. (B) Characterization of the evolution and regulation of human specific Alu-derived alternative splicing, aiming at understanding how new exons are born in the human genome. The following articles contain the results of my PhD research.

A. Regulation and classification of alternative splicing using comparative genomics

(i). Sorek R., Ast G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. ***Genome Research*** **13(7):1631-1637.** In this article we discovered that exons undergoing alternative splicing are flanked by intronic sequences that are exceptionally conserved in mammals. These conserved intronic sequences extend, on average, 100bp from each side of the exon. Exons that do not undergo alternative splicing are not flanked by such conserved sequences. We have further hypothesized that these conserved flanking intronic sequences contain sequences that regulate alternative splicing. Indeed, the strongest motif we found in the conserved introns downstream to the alternatively spliced exons was previously shown to regulate alternative splicing by binding to the splicing regulatory protein FOX1. Until our finding it was generally believed that most of signals that regulate alternative splicing are found within the alternative exons themselves. Our finding shifted this paradigm, and localized much of these regulatory sequences to the flanking introns.

(ii). Sorek R., Shamir R., Ast G. (2004) How prevalent is functional alternative splicing in the human genome? *Trends in Genetics* 20(2):68-71. In this article we discovered additional features that distinguish alternatively spliced exons, and showed that these features could be used to classify splice variants as functional and non-functional. Our analysis implied that many, and perhaps most, of the splice variants present in EST databases are not functional, and provided efficient tests to asses the probability that a splice variant is functional.

(iii). Sorek R., Shemesh R., Cohen Y., Basechess O., Ast G., Shamir R. (2004) A non-EST based method for exon-skipping prediction. *Genome Research* 14(8): 1617-1623. In this article we used the features we discovered in the previous two papers and showed that these features could be utilized to detect alternative splicing without using ESTs. We developed a computational classifier that distinguished alternative from constitutive exons, and systematically identified exon-skipping events in the human genome. Using this classifier we computationally identified completely novel cases of exon-skipping that were not identified so far, and verified experimentally that these variants truly exist in human tissues. The method we presented was the first approach that successfully identifies alternative splicing without relying on any experimental data (such as expressed sequences or microarrays).

(iv). Dror G., Sorek R., Shamir R. (2005) Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics* 21(7):897-901. In this article we improved the exon skipping prediction method presented in our previous paper. By adding more features to the classification (228 features in this article

compared to 7 features in the previous analysis) and employing advanced machine-learning methodologies, we managed to increase the sensitivity of the classifier from 30% (previous article) to more than 50%, with a very low rate (<0.5%) of false positives.

B. Evolution and regulation of primate-specific alternatively spliced Alu exons.

(v). Lev-Maor G.*, Sorek R.*, Shomron N., Ast G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* 300(5623):1288-91 (* equal contribution). In this article we characterized the mutations that need to occur in Alu elements in order to create functional 3' splice sites within these elements. We further showed that such mutations can activate the exonization of Alu sequences and lead to the "birth" of new exons in the human genome. We demonstrated how such mutations can increase the coding capacity of the human genome by creating new alternatively spliced transcripts. Finally, we characterized a specific type of Alu-exonizing mutations that lead to genetic disorders in humans.

(vi). Sorek R.*, Lev-Maor G.*, Reznik M.*, Dagan T., Belinky F., Graur D., Ast G. (2004) Minimal conditions for exonization of intronic sequences: 5' splice site formation in Alu exons. *Molecular Cell* 14(2):221-231 (* equal contribution). In this article, which is a direct extension of our previous article, we computationally characterized the mutations that need to occur in order to create functional 5' splice sites within Alu sequences. This was done by calculation of the nucleotide frequency in each position along the Alu consensus, followed by statistical comparison of these

frequencies between Alus that became exons and more than a million un-exonized genomic Alus. Using the findings of this analysis, along with the results from our previous paper, we located several thousands of Alus in the genome that are susceptible of becoming exons and thus might be the cause for genetic disorders.

(vii). Dagan T.*, Sorek R.*, Sharon E.*, Ast G., Graur D. (2004) AluGene: a database of Alu elements incorporated within protein-coding genes. ***Nucleic Acids Research*** **32: D489-D492. (* equal contribution).** In this article we devised a database that contains all the Alu elements in the human genome, and enables selective extraction of Alus according to their position relative to specific genes. The database also enables the extraction of Alus that might become new exons according to specific motifs.

Intronic Sequences Flanking Alternatively Spliced Exons Are Conserved Between Human and Mouse

Rotem Sorek[1,2] and Gil Ast[1,3]

[1]*Department of Human Genetics, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel;* [2]*Compugen, Ltd., Tel Aviv 69512, Israel*

Comparison of the sequences of mouse and human genomes revealed a surprising number of nonexonic, nonexpressed conserved sequences, for which no function could be assigned. To study the possible correlation between these conserved intronic sequences and alternative splicing regulation, we developed a method to identify exons that are alternatively spliced in both human and mouse. We compiled two exon sets: one of alternatively spliced conserved exons and another of constitutively spliced conserved exons. We found that 77% of the conserved alternatively spliced exons were flanked on both sides by long conserved intronic sequences. In comparison, only 17% of the conserved constitutively spliced exons were flanked by such conserved intronic sequences. The average length of the conserved intronic sequences was 103 bases in the upstream intron and 94 bases in the downstream intron. The average identity levels in the immediately flanking intronic sequences were 88% and 80% for the upstream and downstream introns, respectively, higher than the conservation levels of 77% that were measured in promoter regions. Our results suggest that the function of many of the intronic sequence blocks that are conserved between human and mouse is the regulation of alternative splicing.

[Supplemental material is available online at www.genome.org.]

The recently published draft sequence of the mouse genome (Waterston et al. 2002) facilitates a great advance in searching for *cis*-regulatory sequence elements. The 75 million years that have passed since the divergence of the ancestor of the human and mouse lineages allowed a substantial divergence in neutral DNA; the constraint on functional elements has kept them conserved. Indeed, homologous human and mouse exons are, on the average, 85% identical in their sequences, but introns are more poorly conserved: 60% of the nonexonic sequences are nonalignable, and in the alignable regions the average identity level is 69% (Waterston et al. 2002).

However, numerous regions that are conserved between human and mouse are also found in introns (Hardison et al. 1997). Comparison between the human chromosome 21 and the corresponding genomic sequences in mouse revealed that only one-third of the conserved blocks are exons (Dermitzakis et al. 2002). The other two-thirds of highly conserved sequences are intronic and intergenic. These conserved elements were found to be unexpressed in microarray experiments. Thus, the conclusion was that they are probably *cis*-regulatory sequence elements, but no function could be assigned to most of them (Dermitzakis et al. 2002). We decided to check the possible correlation between the conserved intronic sequences and alternative splicing regulation.

Alternative splicing, a process by which several mRNA isoforms can be generated from a single gene, has received a great deal of attention recently. The current estimates are that 35%–59% of all human genes undergo alternative splicing (Mironov et al. 1999; Brett et al. 2000; International Human Genome Sequencing Consortium 2001). Despite the thousands of alternative splicing events identified to date, very little is known about the regulation of this process (Maniatis and Tasic 2002). *Cis*-acting sequence elements found within exons, called exonic splicing enhancers (ESEs), show the ability to regulate alternative splicing (Tacke and Manley 1999; Blencowe 2000). These sequences interact with proteins of the SR family, which recruit the splicing machinery toward weak, flanking splice sites and enable the inclusion of the alternatively spliced exon in the mature mRNA (Blencowe 2000). Other *cis*-regulatory elements for splicing, such as exonic splicing silencers (ESSs) and intronic splicing enhancers and silencers (ISEs and ISSs, respectively), were also shown to regulate individual cases of alternative splicing (Maniatis and Tasic 2002). Sometimes, multiple *cis*-acting elements cooperatively function to regulate the same alternatively spliced exon (Cartegni et al. 2002). Alternative splicing was also found to be affected by other factors, including the phosphorylation state of SR proteins, the migration of SR and other proteins between the nucleus and the cytoplasm (Chabot 1996), and the promoter of the gene (Cramer et al. 1999).

Although several *cis*-regulatory elements have been characterized for individual cases, it is still unclear how the majority of alternative splicing events are regulated (Maniatis and Tasic 2002; Modrek and Lee 2002). In addition, intronic elements that regulate splicing have received little attention compared to exonic regulatory elements. We used the sequence of the mouse genome to locate homologous exons that are alternatively spliced in both human and mouse. We found that most of these exons are flanked by long conserved intronic sequences, whereas constitutively spliced exons usually are not flanked by such sequences. Our results suggest that many of the previously uncharacterized intronic sequences conserved between human and mouse are involved in the regulation of alternative splicing.

[3]Corresponding author.
E-MAIL gilast@post.tau.ac.il; FAX 972-3-640-9900.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1208803.

RESULTS

To investigate the correlation between conserved intronic sequences and alternatively spliced exons, we began by obtaining the intron–exon structures of human genes by using the output of the LEADS software platform (Shoshan et al. 2001; Sorek et al. 2002), run on the draft human genome build 30 and the cDNAs and ESTs from GenBank version 131. This software aligns expressed sequences to the genome, taking alternative splicing into account (the LEADS process is described in detail in [Sorek et al. 2002] and [Sorek and Safer 2003]). Using the same methods described in Sorek et al. (2002), we utilized the LEADS output to collect reliable data sets of 3583 human alternatively spliced internal exons and 7557 human constitutively spliced internal exons.

To identify exons that are conserved in mouse, we first aligned the mouse ESTs (from GenBank version 131) to the human genome, as described in the Methods section. A constitutively spliced exon was deemed "conserved" if there were mouse ESTs aligned to the constitutively spliced exon, as well as to the exons immediately adjacent to it. We were able to identify a conserved mouse counterpart for 1966 human constitutively spliced internal exons. This is, of course, fewer than the actual fraction of constitutively spliced exons conserved between human and mouse. There are several reasons for not identifying all of the conserved exons. First, our method depends on the existence of mouse ESTs covering the actual exon and the two flanking exons by at least 25 bp on each side (see Methods). Second, the method requires the borders of the exons to be kept conserved between human and mouse. Third, the current draft sequence of the mouse genome is still incomplete.

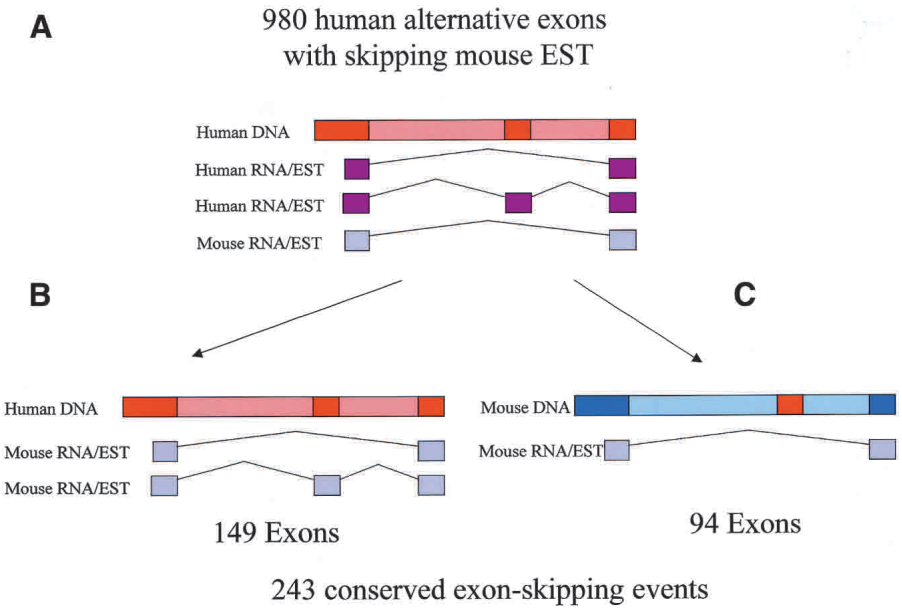


Figure 1 Finding exon-skipping events that are conserved between human and mouse: 3583 exon-skipping events were found in the human genome, using the methods described in Sorek et al. (2002). (*A*) For 980 of these human exons, a mouse EST spanning the intron that represents the exon-skipping variant was found. Human ESTs appear in purple; mouse ESTs are in light blue. (*B*,*C*) The two possible ways to identify an exon as conserved in mouse. (*B*) Identification of mouse ESTs that contain the exon, as well as the two flanking exons. (*C*) If the exon was not represented in mouse ESTs, the sequence of the human exon was searched against the intron spanned by the skipping mouse EST on the mouse genome. If a significant conservation (above 80%) was found, and the alignment spanned the full length of the human exon, the exon was declared as conserved.

A human exon-skipping was deemed "conserved" in mouse if both splice variants (the variant that skips the exon and the variant that contains the exon) were supported by mouse ESTs. For 149 exon-skipping events, both variants were found in mouse ESTs (Fig. 1B). However, when the variant that contains the alternatively spliced exon is a rare variant, or a variant unique to a tissue that is not represented in mouse EST libraries, there might be no mouse EST covering it. Nevertheless, if the human exon is truly conserved in the mouse transcriptome, we would expect its DNA sequence to be conserved between human and mouse. Here, we rely on the fact that although exons are conserved between the human and mouse genomes at an average level of 85%, introns are conserved at much lower levels (Waterston et al. 2002). Therefore, in cases where there was a skipping variant evident in mouse ESTs, but there was no mouse EST showing the variant that contains the exon, we aligned the sequence of the human exon to the relevant intron in the mouse genome (Fig. 1C). The exon was declared conserved if a significant conservation above 80% identity was found, if the alignment spanned the full length of the human exon, and if the exon was flanked by the canonical AG/GT acceptor and donor sites in the mouse genome. Using this approach we identified 94 additional exon-skipping events conserved between human and mouse. Complete data for the 243 conserved alternatively spliced exons, along with the flanking intronic sequences, are provided as Supplementary Material (available online at www.genome.org).

To check the conservation between the intronic sequences immediately flanking alternatively spliced exons, we used Sim4 (Florea et al. 1998) to align the last 100 bases of the intron that is upstream of the human alternatively spliced exon to the last 100 bases of the intron upstream of the respective exon in the mouse genome. We repeated this analysis for the first 100 bases of the intron that is downstream from the alternatively spliced exon. A significant alignment was found for 223/243 (92%) human and mouse 100 bases of upstream introns, and for 199/243 (82%) human and mouse 100 bases of downstream introns. For 188/243 (77%) of the exons, conserved sequences were found in both the upstream and downstream introns. These percentages were similar in both the subset of 149 EST confirmed exons and the subset of the 94 exons supported by the mouse genomic sequence only (76.5% and 78.5%, respectively, P=0.92). This demonstrates the homogeneity of our group of conserved alternatively spliced exons, and further indicates that the 94 genome-supported exons are real events of conserved alternatively spliced exons.

We repeated the same analysis on constitutively spliced exons conserved between human and mouse. For these, a significant

alignment was found for 886/1966 (45%) human and mouse last 100 bases of upstream introns, and for 691/1966 (35%) human and mouse first 100 bases of downstream introns. Only 343/1966 (17%) of the constitutively spliced exons had conserved sequences in both the upstream and downstream introns.

We further characterized the alignment between the intronic regions close to conserved alternatively spliced exons. Whenever the conserved intronic region exceeded the 100 bases aligned, we extended the alignment to the end of the conserved region. For alternatively spliced exons, the average length of the conserved intronic sequences immediately flanking the exon was 103 bases in the upstream intron and 94 bases in the downstream intron (medians 72 and 77, respectively). In the 17% of constitutively spliced exons for which a significant conservation was identified, the average length of the conserved intronic sequences was 41 and 38 bases for upstream and downstream introns, respectively (medians 34 and 30).

Figure 2 presents the per-position conservation near the splice sites. For alternatively spliced exons, the average identity level in the last 30 bases of the upstream intron was 88%, and 80% for the first 30 bases of the downstream intron. For comparison, promoter regions are conserved between human and mouse at an average level of about 77% (Waterston et al. 2002). The average identity levels gradually decrease with the distance from the splice site, but remain significantly higher than those of constitutively spliced exons (Table 1).

A typical example of conserved intronic elements flanking an alternatively spliced exon is presented in Figure 3. This

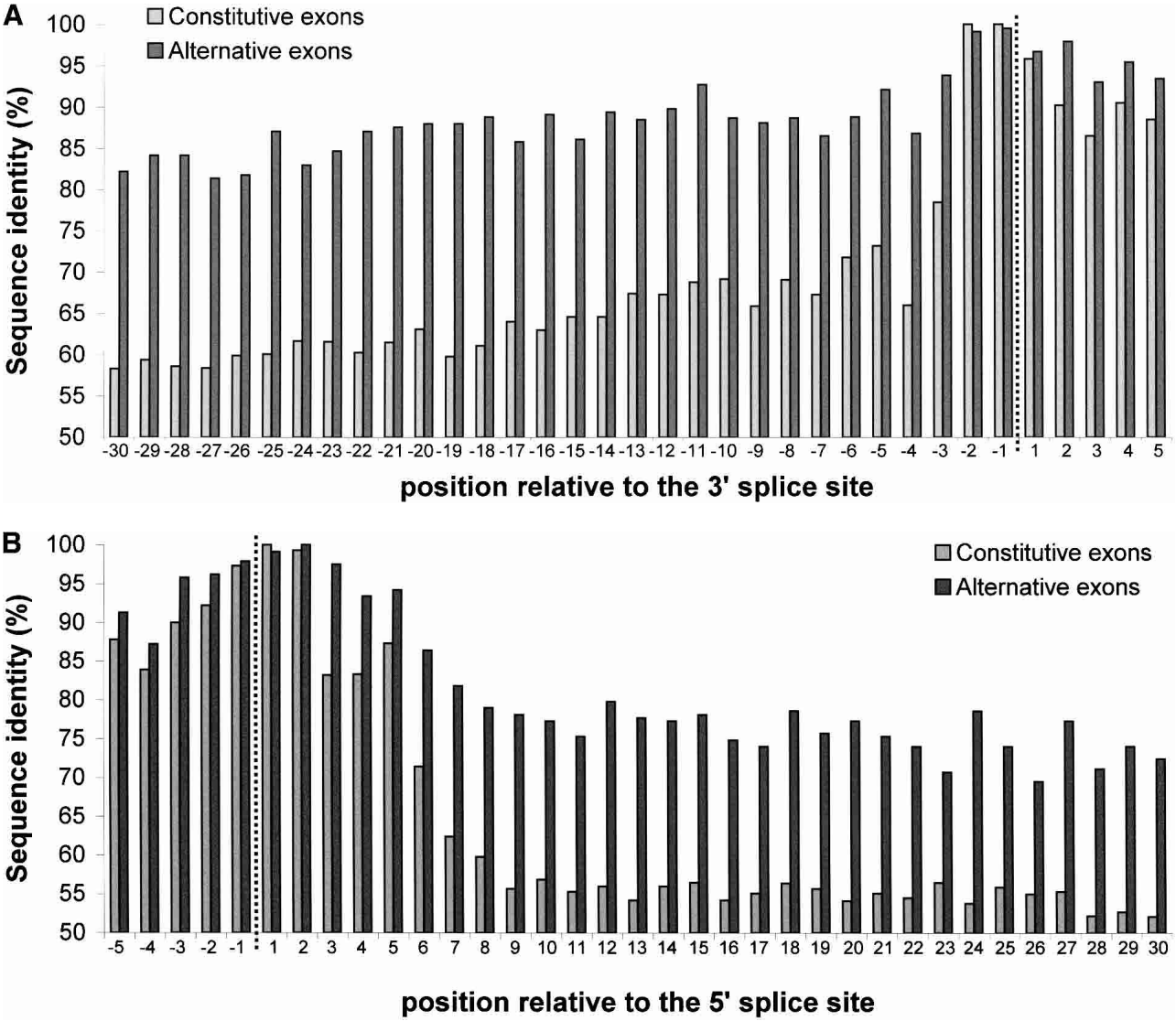


Figure 2 Per-position conservation near alternatively and constitutively spliced exons. Intronic regions near the splice site were aligned, using GAP (global alignment program) of the GCG package, and identity levels were calculated for each position as described in Methods. All 243 alternative exons and 1966 constitutive exons were used for this analysis. (*A*) Conservation near the 3′ splice site. Data for the last 30 nt of the intron and the first 5 nt of the exon are shown. Dashed line marks the border between the intron and the exon. (*B*) Conservation near the 5′ splice site. Data for the last 5 nt of the exon and the first 30 nt of the intron are shown.

Table 1. Percent Intronic Conservation as a Function of the Distance From the Splice Site[a]

	Upstream introns		Downstream introns	
Distance from splice site (bp)	Constitutively spliced	Alternatively spliced	Constitutively spliced	Alternatively spliced
1–6[b]	81.6	93.3	87.4	95.1
7–30	63.1	86.7	55.1	75.9
31–60	54.9	76.0	52.0	70.1
61–90	50.5	65.8	49.0	65.6

[a] Shown are the average identity levels in windows of 30 bases.
[b] Conservation for the nucleotides that define the splice site (intronic bases 1–6) was calculated separately.

figure shows a VISTA conservation graph (Mayor et al. 2000) for exons 4–8 of the gene *KCND3,* corresponding to RefSeq NM_004980. Exon 6 (marked by an arrow) was found, in our analysis, to be an alternatively spliced exon, conserved between human and mouse. The other four exons are constitutively spliced. Long conserved intronic regions (colored orange) are clearly seen flanking the alternatively spliced exon; such long conserved regions are not visible near any of the four constitutively spliced exons. These flanking intronic regions are also highly conserved in rats, a fact that further supports their functionality (Fig. 3C). Another conserved region is visible near (~400 bp upstream), but not adjacent to, the alternatively spliced exon. This region might be an alternatively spliced exon not yet identified or might serve as another *cis*-regulatory element.

The sequences we identified in this study could serve as templates for the identification of regulatory sequences for alternative splicing. To demonstrate this, we conducted a hexamer count in the first and last 100 bases of the downstream and upstream introns, respectively, that flank the human alternatively spliced exons. For this, we took only the 132 cases where the length of the conserved stretch was more than 50 bases. The most abundant hexamer in the conserved intronic sequences downstream of the alternatively spliced exons was TGCATG (excluding TTTTTT, which was also the most abundant in nonconserved introns). This hexamer appeared in 24 (18%) of the examined sequences (appearing twice in five of the sequences), ninefold over the expected frequency. In 93% of the cases, the TGCATG hexamer was conserved in mouse. This hexamer was over-abundant neither in the 100 bases downstream of constitutively spliced exons, nor in intronic sequences located upstream from the alternatively spliced ex-

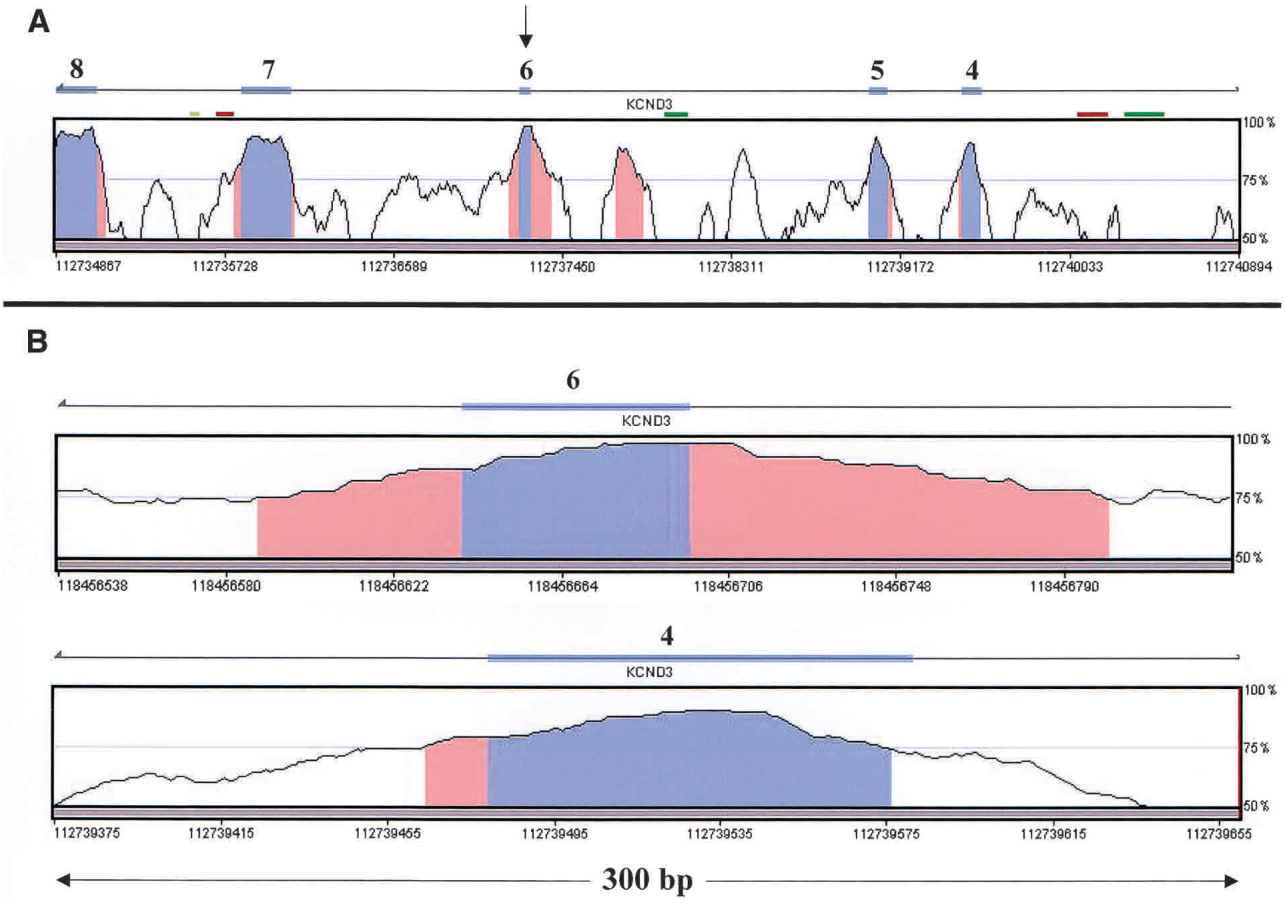


Figure 3 (Continued on facing page)

ons. The TGCATG hexamer was previously shown to regulate alternative splicing of several exons, specifically when found in the downstream intron (Lim and Sharp 1998; Deguillien et al. 2001).

DISCUSSION

Our finding that long conserved intronic elements are found near alternatively spliced exons is intriguing, especially because the average level of conservation between human and mouse intronic sequences is relatively low (Waterston et al. 2002). These findings suggest that there might be a regulatory mechanism common to many alternatively spliced exons that involves the intronic sequences immediately flanking these exons. Because typical binding sites for RNA-splicing regulatory proteins are relatively short (4–10 nucleotides; Tacke and Manley 1999; Blencowe 2000; Cartegni et al. 2002; Fairbrother et al. 2002; Maniatis and Tasic 2002), the fact that the length of the conserved regions usually exceeds 50 bases implies the involvement of multiple factors in this regulation. It is also possible that some of the intronic sequences create a secondary structure, or are involved in interactions with the transcriptional machinery, or in chromatin remodeling. All of these processes were shown to regulate alternative splicing (Caceres and Kornblihtt 2002; Damgaard et al. 2002; Manley 2002).

It was recently reported that the frequency of SINEs (short interspersed repetitive elements) insertion into the 150 intronic bases that flank exons tends to be lower than the frequency of their insertion into other parts of introns (Majewski and Ott 2002). Those authors, therefore, concluded that the first and last 150 bp of introns are likely to contain elements required for the splicing process. Our findings fit well with the distance reported by Majewski and Ott (2002) and indicate that most of this tendency might be contributed by introns flanking alternatively spliced, rather than constitutively spliced exons.

We found that 17% of the constitutively spliced exons conserved between human and mouse were flanked on both sides by conserved intronic sequences. These exons were declared constitutively spliced, because many expressed sequences showed their existence, but no expressed sequence skipped them. However, it is possible that these exons are actually alternatively spliced, and the splice isoform skipping the exon is rare, or condition-specific, and therefore not represented in the EST database.

A recent study found that 41% of all mouse genes undergo alternative splicing, with about two alternatively spliced exons observed per alternatively spliced gene (Okazaki et al. 2002). Our results suggest that most of the alternatively spliced exons are flanked on both sides by conserved intronic sequences, averaging about 100 bases in length; these conserved intronic elements possibly function in alternative splicing regulation. This implies that a few million intronic nucleotides in the sequence of the mouse and human genomes are involved in the regulation of alternative splicing. How this regulation occurs remains to be determined.

C

```
mouse    GAATGCTGGGATTGAGAAATTGAATGTCAAAACTGG------AATGCTGGAATGTACAAT
rat      GAATGCTGGGATTGAGAAATTGAATGTCAAGACTGG------AATGCTGGAATGTACAAC
human    GAATGCCGGGATCAAGAGATTGAATGTCAAAACTGGCTTGGGAATGCTGGAATGTACAAT
         ****** *****  *** ************ *****      ****************

mouse    CAATGGTGTTTCTATCTTCTGTTGGCATGTTGTCCTGTAG GGGTTATCCTATCTTGTGGA
rat      CAATGGTGTTTCTATCTTCTGTTGGCATGTTGTCCTGTAG GGGTTATCCTATCTTGTGGA
human    CAATGGTGTTTTTATCTTCTGTTGGCATGTTGTCCTGTAG GGGTTGTCCTATCTTGTGGA
         *********** ***************************** ***** **************

mouse    TGATCCCCTGTTATCTGTACGAACCTCCACCATCAAG GTATAACTTTATAAA--TTCAAT
rat      TGATCCCCTGTTGTCTGTACGAACCTCCACCATCAAG GTATAACTTTATAAA--TTCACT
human    TGATCCCCTGTTATCTGTACGAACCTCCACCATCAAG GTATAACTTTTTAAAAATTCAAT
         ************ ********************** ********** ****  **** *

mouse    TG---TTCTTCTCTCTGAG--TCTTGAGTCTGTGGAT-CTACTCTGCTTACCCCT-AGCC
rat      TG---TTCTTCTCT--GAC--TCTTGAGTCTGTGGAT-CTACTCTGCTTACCCCT-AGCC
human    TGATGTTTTTCTCTCTGATAATTCTGAGTCTGTGGATTCTCCCCTTTTTACCCCCCGGCC
         **   ** ******  **   *  ************* ** * **  *******   ***

mouse    TGGTTTTCTGCATGTGC
rat      TGGTTTTCTGCATGTGC
human    TGGTTCTGTGCATGTGC
         ***** * *********
```

Figure 3 Human–mouse alignment of the *KCND3* gene, corresponding to RefSeq NM_004980 (from VISTA browser, http://pipeline.lbl.gov/vistabrowser/). *x*-axis: The nucleotide coordinates on human chromosome 1, according to the assembly version of the human genome from June 2002. *y*-axis: The level of conservation between the human genome and the corresponding mouse genome, according to the MGSCv3 assembly version of the mouse genome. (*A*) Blue bars above the conservation area correspond to annotated exons 4–8 of *KCND3*. Blue areas within the conservation graph mark exons; orange areas mark conserved nonexonic sequences. The exon marked with an arrow (exon 6) is an alternatively spliced one; the others are constitutively spliced exons. (*B*) Enlarged view of the conservation graphs of the alternatively spliced exon (exon 6), and one of the constitutively spliced exons (exon 4) is presented to show the relative lengths of the conserved areas near the exons. (*C*) Human, mouse, and rat alignment of exon 6, as well as the 100 bases upstream and downstream of the exon. Exon sequence is bold; asterisks mark identity in all three organisms. Bold and underline mark the hexamer TGCATG, which previously showed the ability to regulate alternative splicing when found in introns downstream to alternatively spliced exons (Lim and Sharp 1998; Deguillien et al. 2001).

METHODS

Human ESTs and cDNAs were obtained from NCBI GenBank version 131 (August 2002; www.ncbi.nlm.nih.gov/dbEST) and aligned to the human genome build 30 (August 2002; www.ncbi.nlm.nih.gov/genome/guide/human) using the LEADS clustering and assembly system as described in Sorek et al. (2002). Briefly, the software cleans the expressed sequences from vectors and immunoglobulins, masking them for repeats and low-complexity sequences. The software then aligns the expressed sequences to the genome, taking alternative splicing into account, and clusters overlapping expressed sequences into "clusters" that represent genes or partial genes.

Alternatively spliced internal exons and constitutively spliced internal exons were identified using the same methods described in Sorek et al. (2002). In short, these methods screen for reliable exons requiring canonical splice sites and discarding possible genomic contamination events. "Constitutively

spliced internal exon" was defined as an internal exon supported by at least four sequences, for which no alternative splicing was observed. "Alternatively spliced internal exon" was defined as such if there was at least one sequence that contained both the internal exon and the two flanking exons (exon inclusion), and one sequence that contained the two flanking exons, but skipped the middle one (exon skipping).

Mouse ESTs and cDNAs from GenBank version 131 were aligned to the UCSC mouse genome, assembly version 3 (ftp.ensembl.org/pub/assembly/mouse/mgsc_assembly_3/) as follows. Mouse ESTs and cDNAs were cleaned from terminal vector sequences, and low-complexity stretches and repeats in the expressed sequences were masked. Sequences with internal vector contamination were discarded, as were sequences identified as immunoglobulins or T-cell receptors. In the next stage, expressed sequences were heuristically compared with the genome to find likely high-quality hits. They were then aligned to the genome using a spliced alignment model that allows long gaps. Single hits of mouse expressed sequences to the human genome shorter than 20 bases, or having less than 75% identity to the human genome, were discarded. Using these parameters, 1,341,274 mouse ESTs were mapped to the human genome, 511,381 of them having all their introns obeying the GT/AG or GC/AG rules.

To determine whether the borders of a human intron (which define the borders of the flanking exons) were conserved in mouse, a mouse EST spanning the same intron-borders, while aligned to the human genome, was required (with alignment of at least 25 bp on each side of the exon–exon junction). In addition, this mouse EST was required to span an intron (i.e., open a long gap) at the same position along the EST, when aligned to the mouse genome.

Alignment of intronic regions was performed using the local alignment program Sim4 (Florea et al. 1998). An alignment was considered significant according to Sim4 default parameters. This program detects exact matches of length 12 and extends them in both directions with a score of 1 for a match and -5 for a mismatch, stopping when extensions no longer increase the score (Florea et al. 1998). The end of the Sim4 alignment was considered the end of the conserved region. In cases in which the alignment spanned the entire 100 bases, the next 100 intronic bases were aligned, and so forth, until the alignment stopped. A minimal significant alignment was, therefore, of a length of at least 12 exactly matching bases. Lengths of alignments and identity levels were parsed from Sim4 standard output.

For per-position conservation calculation, the first and last 100 bases of the downstream and upstream introns, respectively, that flank the alternatively spliced exons were aligned to their mouse counterparts using the GCG global alignment program GAP (default parameters). For each position, a parameter 'N' was assigned, such as N= the number of cases (out of 243 exons) in which this position was conserved. A per-position conservation value 'C' was calculated such as $C = (N/243) \times 100$, that is, the percentage of cases in which this position was conserved. This procedure was repeated for the 1966 constitutively spliced exons.

Overrepresentation of TGCATG hexamer was calculated as follows. We analyzed 132 downstream intronic sequences; each of these sequences was 100 bases long. In each such sequence, 95 hexamers are possible. As expected by chance, the probability for a specific hexamer is 1/4096, so that in each 100-base sequence the probability of appearance of a specific hexamer was $95/4096 = 0.023$. The mean expected frequency for a specific hexamer in 132 sequences was, therefore, $0.023 \times 132 = 3.06$. Hexamer counts for alternatively spliced and constitutively spliced exons are provided as Supplementary Material.

ACKNOWLEDGMENTS

We thank Eyal Fink and Guy Kol for enabling the study with the mouse LEADS productions; Erez Levanon and Dr. Zurit Levine for fruitful discussions; Han Xie, Pini Akiva, Yossef Kliger, Gad Cojocaru, Dvir Dahary, Kinneret Savitsky, Erez Levanon, and Galit Rotman for helpful comments. This work was partially supported by a grant from the Israel Science Foundation to G.A.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Blencowe, B.J. 2000. Exonic splicing enhancers: Mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci.* **25:** 106–110.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474:** 83–86.

Caceres, J.F. and Kornblihtt, A.R. 2002. Alternative splicing: Multiple control mechanisms and involvement in human disease. *Trends Genet.* **18:** 186–193.

Cartegni, L., Chew, S.L., and Krainer, A.R. 2002. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nat. Rev. Genet.* **3:** 285–298.

Chabot, B. 1996. Directing alternative splicing: Cast and scenarios. *Trends Genet.* **12:** 472–478.

Cramer, P., Caceres, J.F., Cazalla, D., Kadener, S., Muro, A.F., Baralle, F.E., and Kornblihtt, A.R. 1999. Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol. Cell* **4:** 251–258.

Damgaard, C.K., Tange, T.O., and Kjems, J. 2002. hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure. *RNA* **8:** 1401–1415.

Deguillien, M., Huang, S.C., Moriniere, M., Dreumont, N., Benz Jr., E.J., and Baklouti, F. 2001. Multiple *cis* elements regulate an alternative splicing event at 4.1R pre-mRNA during erythroid differentiation. *Blood* **98:** 3809–3816.

Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., Flegel, V., Bucher, P., Jongeneel, C.V., et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420:** 578–582.

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.

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.

Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lim, L.P. and Sharp, P.A. 1998. Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats. *Mol. Cell. Biol.* **18:** 3900–3906.

Majewski, J. and Ott, J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12:** 1827–1836.

Maniatis, T. and Tasic, B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418:** 236–243.

Manley, J.L. 2002. Nuclear coupling: RNA processing reaches back to transcription. *Nat. Struct. Biol.* **9:** 790–791.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. 2000. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046–1047.

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.
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.
Shoshan, A., Grebinskiy, V., Magen, A., Scolnicov, A., Fink, E., Lehavi, D., and Wasserman, A. 2001. Designing oligo libraries taking alternative splicing into account. In *Proceedings of SPIE: Microarrays: Optical technologies and informatics* (eds. M.L. Bittner, Y. Chen, A.N. Dorsel, and E.R. Dougherty), pp. 86–95. E.R. Vol 4266. SPIE, Bellingham, WA.
Sorek, R. and Safer, H.M. 2003. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.* **31:** 1067–1074.
Sorek, R., Ast, G., and Graur, D. 2002. Alu-containing exons are alternatively spliced. *Genome Res.* **12:** 1060–1067.
Tacke, R. and Manley, J.L. 1999. Determinants of SR protein specificity. *Curr. Opin. Cell. Biol.* **11:** 358–362.
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.

WEB SITE REFERENCES

www.ncbi.nlm.nih.gov/dbEST; Database of expressed sequence tags.
www.ncbi.nlm.nih.gov/genome/guide/human; Human genomic sequence.
ftp.ensembl.org/pub/assembly/mouse/mgsc_assembly_3/; Mouse genomic sequence.
http://pipeline.lbl.gov/vistabrowser/; VISTA Genome Browser.

Received January 22, 2003; accepted in revised form March 26, 2003.

How prevalent is functional alternative splicing in the human genome?☆

Rotem Sorek[1,2], Ron Shamir[3] and Gil Ast[1]

[1]Department of Human Genetics, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel
[2]Compugen, 72 Pinchas Rosen Street, Tel Aviv 69512, Israel
[3]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Comparative analyses of ESTs and cDNAs with genomic DNA predict a high frequency of alternative splicing in human genes. However, there is an ongoing debate as to how many of these predicted splice variants are functional and how many are the result of aberrant splicing (or 'noise'). To address this question, we compared alternatively spliced cassette exons that are conserved between human and mouse with EST-predicted cassette exons that are not conserved in the mouse genome. Presumably, conserved exon-skipping events represent functional alternative splicing. We show that conserved (functional) cassette exons possess unique characteristics in size, repeat content and in their influence on the protein. By contrast, most non-conserved cassette exons do not share these characteristics. We conclude that a significant portion of cassette exons evident in EST databases is not functional, and might result from aberrant rather than regulated splicing.

Numerous studies have shown that alternative splicing is prevalent in mammalian genomes. Using ESTs and cDNAs aligned to the genomic sequence, these studies estimate that between 35% and 59% of all human genes undergo alternative splicing [1,2]. However, it is not clear how many of the splice variants predicted from ESTs are functional and how many represent aberrant splicing ('noise') or EST artefacts (such as genomic contamination) [3–5]. An mRNA variant can be defined as being 'functional' if it is required during the life-cycle of the organism and activated in a regulated manner.

Aberrant alternative splicing

Somatic mutations within splice sites or introns could result in aberrant splicing, leading to non-functional mRNAs; ESTs derived from these mRNAs would be indistinguishable from normal splice variants. Because somatic mutations are prevalent in cancer related tissues, and >50% of the ESTs in dbEST come from cancer, cell-lines or tumor tissues [6], such spurious variants can be ubiquitous in dbEST.

Splicosomal mistakes have also been proposed as a mechanism that can result in non-functional transcripts [3]. In common with any complex biological machine, the splicosome can 'slip' and identify cryptic splice sites in introns as normal splice sites, thus, inserting part of an intron into the mature mRNA. Obviously such mistakes would not represent functional, regulated alternative splicing.

What portion of the observed splice variants represents functional alternative splicing? We employed a comparative genomics approach to address this question, by compiling a dataset of exon-skipping events (cassette exons) that are conserved between human and mouse. The conservation of such events in both human and mouse species, which diverged from their common ancestor 75–110 million years ago, suggests the functional importance of these exons.

Detecting cassette exons conserved between human and mouse

We have recently collected a dataset of 980 EST-predicted human alternatively spliced cassette exons [7]. From these 980 exons, 243 (25%) were also found to be alternatively spliced in mouse ['conserved alternatively spliced exons' (CAS exons)]. The remaining 737 (75%) are 'non-conserved alternatively spliced exons' (non-CAS exons) (Box 1). Low levels of alternative splicing conservation between human and mouse were also observed in other studies [8,9]. The method that was used to locate cassette exons and human–mouse conservation is described in detail in Ref. [7]. Calculated features of the 980 exons used for this study appear as supplementary material online.

Box 1. Finding exon-skipping events that are conserved between humans and mouse

The initial set of exons comprised 980 apparently alternatively spliced human exons, for which a mouse EST spanning the intron that represents the exon-skipping variant was found. Two strategies were used to identify an exon as being conserved in mouse: (i) identification of mouse ESTs that contain the exon and the two flanking exons; and (ii) if the exon was not represented in mouse ESTs, the sequence of the human exon was searched against the intron spanned by the skipping mouse EST on the mouse genome. If a significant conservation (>80%) was found, the alignment spanned the full length of the human exon, and the exon was flanked by the canonical AG acceptor and GT donor sites in the mouse genome, then the exon was declared as conserved. For 243 exons (25% of 980), conserved alternative splicing was detected in mouse. Detailed description of the methods can be found in Ref. [7].

☆ Supplementary data associated with this article can be found at doi: 10.1016/j.tig.2003.12.004

Corresponding author: Gil Ast (gilast@post.tau.ac.il).

Comparing conserved with non-conserved cassette exons

Presumably, orthologous exons that are alternatively spliced both in human and mouse have functional importance. We can therefore regard the group of CAS exons as a representative group of functional, alternatively spliced exons.

Non-CAS exons, on the other hand, can also be functional, representing exons created after the divergence of the human and the mouse lineages. However, if these exons, as a group, were indeed functional, we might expect them to have the same general characteristics as CAS exons. Therefore, we compared several features between the two groups of exons predicted by ESTs.

We first sought to understand the influence of the alternatively spliced exons on the proteins in which they are inserted. From the 243 CAS exons identified, 195 had an unambiguous coding-region annotation in the GenBank cDNAs. Of these, 191 (98%) were located within the protein-coding region and four were located within the untranslated region (UTR). This finding is not surprising because UTRs are mostly found in the terminal exons [10], whereas the exons in our study were internal exons. A similar percentage of the non-CAS exons were located in the coding sequence and UTR.

The influence of the CAS exons on the protein coding sequence (CDS) was significantly different from the influence of the non-CAS exons (Figure 1). In 147 of 191 (77%) CAS exons that were located within the protein-coding region, the insertion of the alternatively spliced exon did not alter the reading frame and only eight of these 147 exons (5%) contained an in-frame stop codon. This means that 73% of the conserved alternative exons are 'peptide cassettes' (i.e. they insert a short amino-acid sequence into the translated protein without changing the coding frame) (Figure 1). These results indicate a strong tendency of functional cassette exons to add or remove amino acids within the protein sequence, rather than inflict a dramatic change on it.

By comparison, only 206 of the 510 (40%) non-CAS exons preserved the reading frame; when an alternatively spliced exon was inserted in frame, almost half (47%) of these contained a stop codon. Thus, only 109 of the 510 (21%) non-CAS exons were indeed 'peptide cassette' exons (Figure 1).

We examined in detail the cases in which the alternatively spliced exons caused a frame shift in the CDS. Of the 44 CAS exons that caused a frame shift, 27 (61%) actually made the protein longer by suppressing a nearby stop codon and only four exon insertions (9%) resulted in a protein shorter than 100 amino acids. This indicates that, in a substantial fraction of the conserved alternatively spliced exons that cause a frame shift, the exon insertion changes only the C-terminus of the protein.

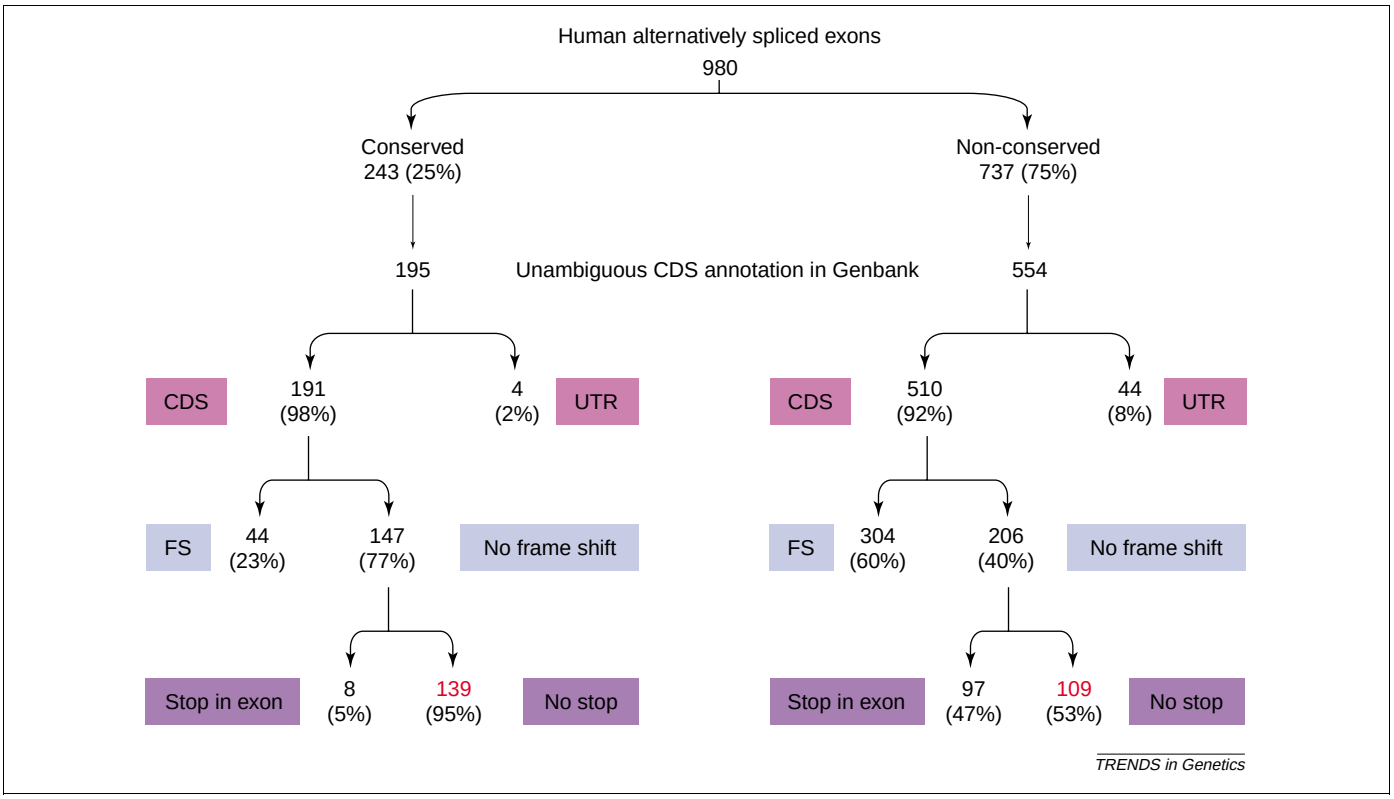


Figure 1. The influence of alternatively spliced exons on the protein-coding sequence. Our dataset of human alternatively spliced exons contained 980 exons. These were divided into two subsets: (i) exons whose existence was also confirmed by mouse ESTs or mouse genomic sequence (243 exons, termed 'conserved alternatively spliced exons'); and (ii) exons for which no such evidence was detected (737 exons, termed 'non-conserved alternatively spliced exons'). The influence of these exons on the protein-coding sequence is indicated by the tree-like diagram. Those exons with 'unambiguous CDS annotation' [i.e. all annotated GenBank cDNAs in which the exon appeared had the same annotation (either UTR or CDS)] were analyzed further. Peptide cassettes were found in 73% (139 of 191) and 21% (109 of 510) of the conserved alternatively spliced and non-conserved alternatively spliced exons, respectively (shown in red). Definitions: frame shift, the length of the exon is not a multiple of three; stop in exon, the exon contains an in-frame stop codon; peptide cassette, contains exons that neither cause a frame shift nor contain a stop codon. Abbreviations: CDS, coding sequence; FS, frame shift; UTR, untranslated region.

By contrast, of the 304 non-CAS exons that caused a frame shift, only 25 (8%) resulted in a longer protein, whereas 91 (30%) resulted in a protein shorter than 100 amino acids. This implies that insertion of non-CAS exons into the mature mRNA frequently has a major effect on the protein sequence.

Frequencies of ESTs, repeats and size as measures of function

It was recently suggested that a higher number of ESTs/mRNAs supporting a splice variant correlates with its functionality [11]. Our results agree with this observation: although CAS exons were on average supported by nine sequences (median 3), the average EST and mRNA support for non-CAS exons was 2.2 sequences (median 1). However, sequence support by itself is not sufficient for detecting functional alternative splicing: in our study, 30% of the CAS exons were supported by a single human expressed sequence.

We have previously shown that point mutations within silent intronic *Alu* elements can result in the creation of new alternatively spliced exons [12]. Such *Alu*-derived exons represent at least 5% of the alternatively spliced cassette exons found in dbEST [13]; however, it is unclear whether these *Alu* exons are functional [14]. Other repetitive elements, such as the RTE-1 retrotransposon in cattle, have also shown the ability to be exonized (i.e. become exons via a splicing-mediated process) [15].

To check how many alternatively spliced exons are the result of such exonization, we performed a BLAST search of the exons against a database of mammalian repeats. Only one of the 243 CAS exons had a significant 'hit' (e-value $<10^{-10}$) to a mammalian interspersed repeat (MIR). By contrast, 191 of the 737 non-CAS exons (26%) had significant 'hits' to a repeat. In 147 (77%) of these cases, the repeat was an *Alu* retrotransposon, which is unique to primates. The repeat content within exons, therefore, is another feature in which CAS exons differ from non-CAS exons.

Conserved and non-conserved cassette exons also differ in their exon length distribution. The average length of a CAS exon was 87 bases (median 76). By contrast, the average length of non-CAS exons was 116 (median 104). The difference between these two distributions is statistically significant ($P < 10^{-6}$). (For comparison, the average length of constitutively spliced exons is 129 bases [16].) Thus, non-CAS exons are significantly longer than CAS exons.

Which exons are functional?

We detected a conserved mouse exon for 25% of the human cassette exons in our set. These CAS exons most probably have functional importance. In principle, some of the 75% human non-CAS exons could also be functional; however, these might be expected to have similar characteristics to those of the CAS exons. This is not the case. We have shown that the group of non-CAS exons significantly differs from the group of CAS exons in many important parameters (discussed previously) (these differences are summarized in Table 1). This suggests that many of the apparently non-conserved splice variants in the human genome are non-functional. It is noteworthy that this claim is based on the assumption that functional non-CAS exons ought to have properties similar to those of CAS exons – this has yet to be verified.

To further test this assumption, we examined non-CAS exons that are supported by multiple ESTs. Of the non-CAS exons, 21% have no frame shift and no stop codon. However, in the subset of non-CAS exons that are supported by five sequences or more and are found in the CDS (37 exons), 54% have no frame shift and no stop codon. Therefore, some of the non-CAS exons indeed represent new functional exons that are specific to the human lineage. In addition, this supports our claim that functional non-CAS exons have properties similar to those of CAS exons.

Non-CAS exons might have an important evolutionary role even if they are not functional. These rarely expressed splice isoforms 'suggest' new variants while keeping the genomic repertoire intact. Such new variants can have a positive effect on the organism and become fixed during evolution. Indeed there are several examples, such as the *Alu*-derived exon in the ADARB1 gene [17], in which repetitive elements were found to be recruited from an intron during evolution and fixed as new alternatively

Table 1. Features differentiating between conserved alternatively spliced exons and non-conserved alternatively spliced exons

Features	Conserved alternatively spliced exons	Non-conserved alternatively spliced exons[a]	*P* value[b]
Average size	87	116	$P < 10^{-6}$
Percentage of exons that are a multiple of three	77% (147/191)	40% (206/510)	$P < 10^{-5}$
Percentage of exons that are 'peptide cassettes' [c]	73% (139/191)	21% (109/510)	$P < 10^{-15}$
Percentage of exon insertions that result in a longer protein (from a total of exons that are not a multiple of three)	61% (27/44)	8% (25/304)	$P < 10^{-9}$
Percentage of exon insertions that result in a protein <100 amino acids (from a total of exons that are not a multiple of three)	9% (4/44)	30% (91/304)	$P < 0.02$
Average supporting expressed sequences	9	2.2	$P < 10^{-6}$
Percentage of exons that contain repetitive elements[d]	<0.5% (1/243)	26% (191/737)	$P < 10^{-20}$

[a]Non-conserved alternatively spliced exons are human exons that were found to be alternatively spliced in human ESTs but were not found in the mouse genome.
[b]The *P* value was calculated using Fisher's exact test, except for the 'average size' and 'average support', for which *P* values were calculated using student's *t* test.
[c]A 'peptide cassette' exon is defined such as it neither causes a frame shift nor contains a stop codon, so that the effect of its insertion or deletion on the translated protein is a local insertion or deletion of a peptide.
[d]Exons were aligned to a database of repetitive elements, and 'hits' with e-value $<10^{-10}$ were considered positive.

Box 2. Methods

Detailed methods for compilation of exon sets are found in Refs [7] and [13]. Fisher's exact statistical test was used for calculating *P* values of the parameters differentiating between the two exons sets, except for the 'average size' and 'average support', for which *P* value was calculated using student's *t* test. Human–mouse exon skipping data from ASAP were downloaded from http://www.bioinformatics.ucla.edu/ASAP/data/Comparative_Genomics/Hs_Mm_exon_skip_status_table. Lengths of these exons were extracted from: http://www.bioinformatics.ucla.edu/ASAP/data/Alt_Splice/Human/January_2002/ exon_obs_table.

spliced exons. Such a recruitment of intronic sequences as new alternatively spliced exons was previously proposed to be major evolutionary driving force towards the emergence of new protein sequences in eukaryotes [8,9,13,15,18].

Comparison with other results

We examined the results of Modrek and Lee [9], who compared exon-skipping events between human and mouse using ASAP, an EST-based alternative splicing database. 127 ASAP exons were identified in which both variants (exon inclusion and exon skipping) were observed in human and in mouse (equivalent to our CAS exons); in 78 (61%) the length of the exon was a multiple of three (i.e. did not cause a frame shift). By contrast, of the 427 ASAP human exons that were predicted to be skipped in human but were not conserved in mouse (non-CAS exons), only 164 (38%) had an exon length that was a multiple of three. These numbers are similar to the numbers calculated from our set of exons, showing that our results are not database dependent (Box 2).

Our results suggest that 73% of functional cassette exons neither change the coding frame nor introduce a premature stop codon. However, Zavolan *et al.* reported that only 178 of 423 (42%) of the alternative splicing detected in mouse full-length mRNAs preserved the reading frame [19]. Another study that detected alternative splicing using ESTs and cDNAs also reported that only 40% of the deletions or insertions of a new sequence in the middle of the protein were in frame [20]. This contradiction probably stems from the fact that these two studies used alternative splicing predicted from alignments of expressed sequences (ESTs and cDNAs), which contained both conserved and non-conserved splice variants. Indeed, combining the numbers from both our sets (CAS and non-CAS exons) results in 248 out of 701 (35%) of the exons preserving the reading frame, similar to the results in these studies. [19,20] In a study of a sample of 1000 alternatively spliced exons compiled from the literature, ~78% were identified as 'peptide cassette' exons (neither contained a stop codon nor introduced a frame-shift) [21] – similar to the 73% we detected in CAS exons. These data strongly supports our results because experimentally confirmed splice variants are more likely to be functional.

We have shown that a comparative genomics approach can be useful for assessing the functionality of splice variants. In the future, the parameters defining functional cassette exons could be used for *de novo* identification of alternative splicing in organisms for which no EST data exist.

Acknowledgements

We thank Dan Graur for inspiration; Han Xie, Amit Novik, Dvir Dahary and Ami Haviv for insightful discussions; Zurit Levine for technical help; and Eli Eisenberg for critical reading of the manuscript.

References

1 Mironov, A.A. *et al.* (1999) Frequent alternative splicing of human genes. *Genome Res.* 9, 1288–1293
2 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
3 Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107
4 Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* 30, 13–19
5 Maniatis, T. and Tasic, B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418, 236–243
6 Baranova, A.V. *et al.* (2001) *In silico* screening for tumour-specific expressed sequences in human genome. *FEBS Lett.* 508, 143–148
7 Sorek, R. and Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* 13, 1631–1637
8 Nurtdinov, R.N. *et al.* (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.* 12, 1313–1320
9 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. *Nat. Genet.* 34, 177–180
10 Deutsch, M. and Long, M. (1999) Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.* 27, 3219–3228
11 Kan, Z. *et al.* (2002) Selecting for functional alternative splices in ESTs. *Genome Res.* 12, 1837–1845
12 Lev-Maor, G. *et al.* (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science* 300, 1288–1291
13 Sorek, R. *et al.* (2002) Alu-containing exons are alternatively spliced. *Genome Res.* 12, 1060–1067
14 Pavlicek, A. *et al.* (2002) Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? *FEBS Lett.* 523, 252–253
15 Makalowski, W. (2003) Genomics. Not junk after all. *Science* 300, 1246–1247
16 Zavolan, M. *et al.* (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* 13, 1290–1300
17 Lai, F. *et al.* (1997) Editing of glutamate receptor B subunit ion channel RNAs by four alternatively spliced DRADA2 double-stranded RNA adenosine deaminases. *Mol. Cell. Biol.* 17, 2413–2424
18 Kondrashov, F.A. and Koonin, E.V. (2003) Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet.* 19, 115–119
19 Zavolan, M. *et al.* (2002) Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Res.* 12, 1377–1385
20 Modrek, B. *et al.* (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29, 2850–2859
21 Thanaraj, T.A. and Stamm, S. (2003) Prediction and statistical analysis of alternatively spliced exons. *Prog. Mol. Subcell. Biol.* 31, 1–31

0168-9525/$ - see front matter
doi:10.1016/j.tig.2003.12.004

A Non-EST-Based Method for Exon-Skipping Prediction

Rotem Sorek,[1,2,4] Ronen Shemesh,[2] Yuval Cohen,[2] Ortal Basechess,[2] Gil Ast,[1] and Ron Shamir[3]

[1]*Department of Human Genetics, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel;* [2]*Compugen, Tel Aviv 69512, Israel;* [3]*School of Computer Science, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel*

It is estimated that between 35% and 74% of all human genes can undergo alternative splicing. Currently, the most efficient methods for large-scale detection of alternative splicing use expressed sequence tags (ESTs) or microarray analysis. As these methods merely sample the transcriptome, splice variants that do not appear in deeply sampled tissues have a low probability of being detected. We present a new method by which we can predict that an internal exon is skipped (namely whether it is a cassette-exon) merely based on its naked genomic sequence and on the sequence of its mouse ortholog. No other data, such as ESTs, are required for the prediction. Using our method, which was experimentally validated, we detected hundreds of novel splice variants that were not detectable using ESTs. We show that a substantial fraction of the splice variants in the human genome could not be identified through current human EST or cDNA data.

[Supplemental material is available online at www.genome.org.]

Alternative splicing is a mechanism allowing one gene to produce multiple, sometimes functionally distinct, protein variants (Maniatis and Tasic 2002). In recent years, numerous studies have shown that the phenomenon of alternative splicing is very prevalent in mammalian genes (Mironov et al. 1999; Brett et al. 2000; Kan et al. 2001, 2002; Lander et al. 2001; Modrek et al. 2001). All of these studies used expressed sequence tags (ESTs) or cDNAs for detection of alternative splicing. Other studies used microarrays specifically designed for detection of splice variants (Johnson et al. 2003).

Although much progress has been made in the field of computational detection of alternative splicing in recent years (for review, see Graveley 2001; Modrek and Lee 2002), the full extent of splice variants in the human genome is far from being completely known. ESTs, which are the main source of information for alternative splicing prediction, are a problematic source of information, as they are merely a sample of the transcriptome. The detection of a particular splice variant by ESTs is possible only if its transcription level is sufficiently high in a tissue type for which an EST library has been prepared. Moreover, as most ESTs are generated from the 5′ or the 3′ termini of the transcript, dbEST is biased towards underrepresentation of splice variants involving exons that are in the middle of long transcripts (Johnson et al. 2003). In addition, ESTs are very noisy and contain numerous erroneous sequences (Sorek and Safer 2003; Sorek et al. 2004), so that some of the EST-predicted splice variants may be artifacts (Modrek and Lee 2002).

Indeed, Johnson and colleagues, who recently investigated the extent of human alternative splicing using large-scale microarray experiments, reported on numerous events of alternative splicing that were not represented in ESTs (Johnson et al. 2003). However, even microarray experiments are not sufficient for the identification of all splice variants, as they do not sample all combinations of possible tissues, developmental stages, and conditions.

Comparative genomics has recently proven a useful approach for alternative splicing research (Modrek and Lee 2003; Nurtdinov et al. 2003; Sorek and Ast 2003; Sorek et al. 2004; Resch et al. 2004). Specifically, we have found that conserved alternatively spliced internal exons (of the "cassette-exons" type) are usually flanked by intronic sequences that are conserved between human and mouse, a feature only rarely seen in constitutively spliced exons (Sorek and Ast 2003). These conserved intronic sequences are probably involved in the regulation of alternative splicing. We have also recently found that alternative exons that are conserved between human and mouse possess characteristics, such as smaller size and divisibility by three, which distinguish them from nonconserved alternatively spliced exons (Sorek et al. 2004). In the present study we show (and experimentally verify) how the combination of these and additional features, which distinguish alternative from constitutive exons, can be used for the accurate prediction of whether an exon is an alternative cassette exon, even when there are no ESTs that indicate its skipping.

RESULTS AND DISCUSSION

To identify and characterize features that distinguish between alternative and constitutive exons, we used the training exons sets from Sorek and Ast (2003), which contained 243 alternative internal exons (cassettes) and 1753 constitutive internal exons that are conserved between human and mouse (see Methods). These sets were based on EST analysis of GenBank (release 131), where exons were defined as constitutive if there were at least four expressed sequences supporting them, and no EST skipping them, both in human and in mouse.

Table 1 summarizes the major classifying features that we characterized. In short, alternatively spliced exons are flanked by intronic sequences that are more conserved between human and mouse; they are shorter than constitutively spliced exons; their size tends to be a multiple of three; and they have higher identity

[4]Corresponding author.
E-MAIL rotem@compugen.co.il; FAX 972 3-7658555.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2572604.

Table 1. Features Differentiating Between Alternatively Spliced and Constitutively Spliced Exons

	Alternatively spliced exons	Constitutively spliced exons	*P*-value[a]
Average size	87	128	$P < 10^{-16}$
Percent exons whose length is a multiple of 3	73% (177/243)	37% (642/1753)	$P < 10^{-9}$
Average human–mouse exon conservation[b]	94%	89%	$P < 10^{-36}$
Percent exons with upstream intronic elements conserved in mouse[c]	92% (223/243)	45% (788/1753)	$P < 10^{-11}$
Percent exons with downstream intronic elements conserved in mouse[c]	82% (199/243)	35% (611/1753)	$P < 10^{-14}$
Percent exons with both upstream and downstream intronic elements conserved in mouse[c]	77% (188/243)	17% (292/1753)	$P < 10^{-37}$

[a]*P*-value was calculated using Fisher's exact test, except for the "average size" and "average human–mouse exon conservation", for which *P*-value was calculated using student's t-test.
[b]Average percent of matching nucleotides in global alignment of the respective exons.
[c]The 100 intronic nucleotides immediately upstream (or downstream) of the exon were locally aligned with the mouse 100 counterpart intronic nucleotides using Sim4 (Florea et al. 1998). Conservation was defined if at least 12 consecutive perfectly matching nucleotides were found in the alignment.

level when aligned to their mouse counterpart exon (Fig. 1A–E). These differences probably stem from the unique function of the alternative exons: Because these exons are cassette exons that are sometimes inserted and sometimes skipped, their size should be a multiplication of three so that their skipping would not alter the reading frame of the downstream exons. This constraint, which was also recently reported by Resch et al. (2004), does not apply to constitutively spliced exons. The higher identity level between human and mouse could be explained by the fact that alternatively spliced exons frequently contain sequences that regulate their splicing (exonic splicing enhancers and silencers, reviewed by Cartegni et al. 2002). These regulatory sequences add another level of conservation constraint on the exon sequence. The fact that alternatively spliced exons are smaller than constitutively spliced ones was observed before (Thanaraj and Stamm 2003), and might be related to suboptimal recognition of smaller exons by the spliceosome (Berget 1995).

The features described above could be used to identify exons that are skipped in the human and the mouse genomes. However, each feature by itself provides only a weak classification for exons. Our goal was to find a combination of features that would detect a substantial fraction of the alternative exons, while making near-zero false-positive detection errors. The features we have chosen are the following: (1) exon length, (2) divisible/not divisible by three, (3) percent identity when aligned to the mouse counterpart, and (4) conservation in the upstream and downstream intronic sequences. Each of the two "intronic conservation" features (upstream and downstream) were divided into two subfeatures: (1) length of best human/mouse local alignment in the 100 intronic nucleotides nearest to the exon (where only local alignments with at least 12 consecutive perfectly matching nucleotides were considered) and (2) identity level in this local alignment.

For each of the features we defined a set of thresholds (see Methods). For example, the "human/mouse exon identity" threshold can be set to 100%, at least 99%, at least 98%, and so forth. Similarly, the thresholds for "length of conserved upstream region" can be set to 100, at least 95, at least 90 and so forth. By using a threshold for each of the seven features above, one gets a *classification rule* that classifies as alternative all exons that pass all seven thresholds. Such a rule might, for example, be: "all exons that are at least 99% conserved with their mouse counterpart **and** have at least 95 conserved nucleotides upstream the exon **and** are divisible by three **and** …".

We enumerated all possible rules (about 100 million rules) and tested the quality of the resulting classification on our train-

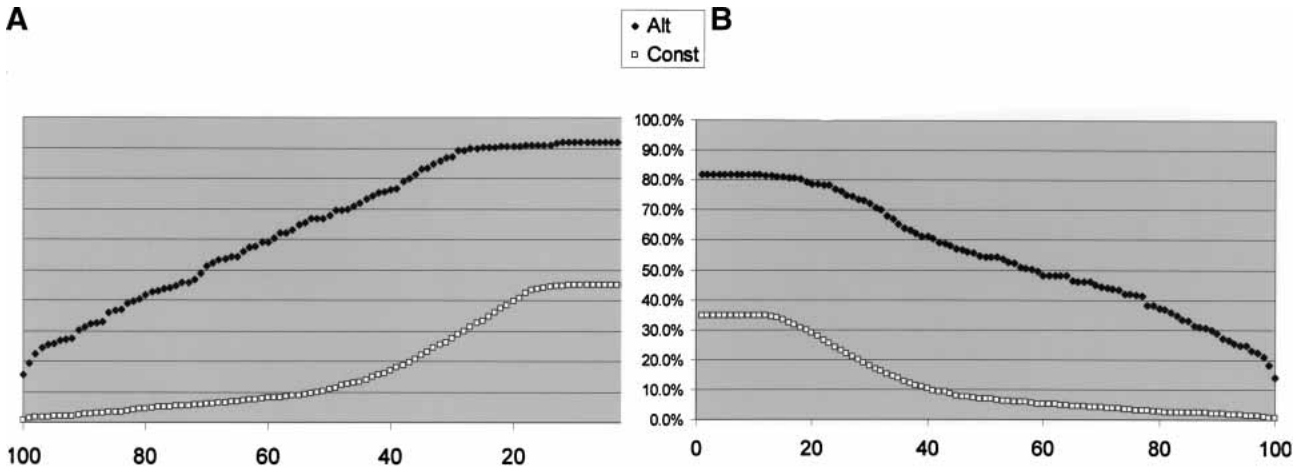


Figure 1 (Continued on next page)

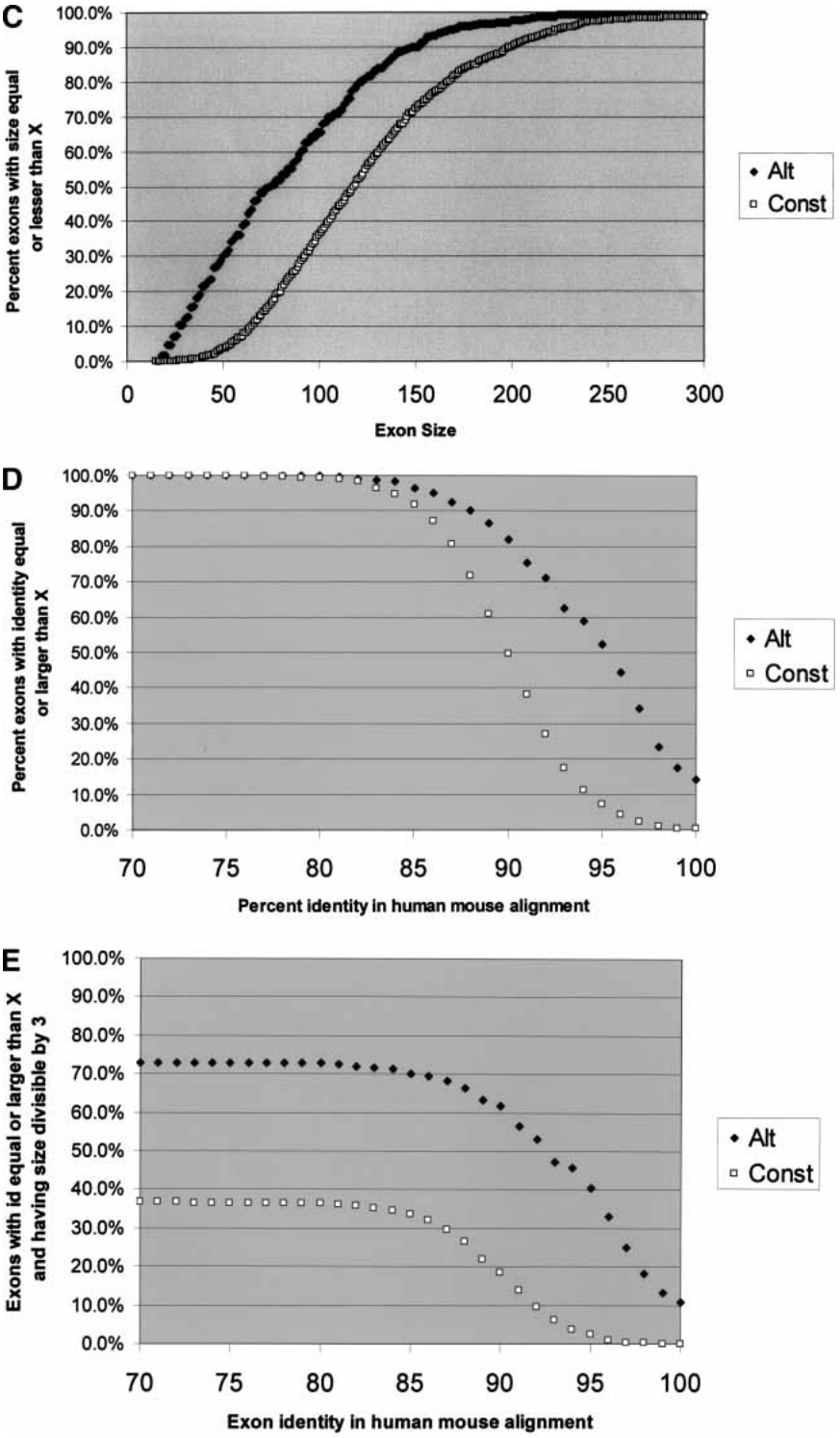


Figure 1 Graphic representation of the differences between alternative and constitutive exons. For each of the following curves, constitutive exons are in squares, and alternatives are in diamond shapes. (*A*) Length of conserved region in the nearest 100 nt of the flanking upstream intron. *x*-axis, length of conserved region (best Sim4 local alignment); *y*-axis, percent exons with upstream conserved region greater than or equal to the value in *x*. Conservation was detected using local alignment with the mouse 100 counterpart intronic nt. A minimum hit was 12 consecutive perfectly matching nt. (*B*) Length of conserved region in the nearest 100 nt of the flanking downstream intron. Axes as in *A*. (*C*) Exon size distribution. *x*-axis, exon size; *y*-axis, percent exons having size lesser or equal to the size in *x*. (*D*) Human–mouse exon identity. *x*-axis, percent identity in the global alignment of the human and the mouse exons; *y*-axis, percent exons with identity greater or equal to the value in *x*. (*E*) Human–mouse exon identity, for exons whose size is a multiple of 3. Axes as in *D*. Note that by combining two features we get better separation of the two exon-types.

ing set of 243 alternative and 1753 constitutive exons. We sought a rule that would correctly identify a maximum number of alternative exons from the training set while making no false-positive identification.

The best rule that emerged was the following: At least 95% identity with the mouse exon counterpart; exon size is a multiple of three; a best local alignment of at least 15 intronic nucleotides upstream of the exon with at least 85% identity; and a perfect match of at least 12 consecutive intronic nucleotides downstream of the exon. This combination of features identified 76 exons, or 31% of the 243 alternatively spliced exons in our training set, whereas none of the 1753 constitutively spliced exons matched these features. To check the robustness of this analysis we employed five-way cross validation (see Supplemental material for details). The average sensitivity in these five analyses was 32.3%, and the average specificity was 99.72%.

The above combination of parameters can therefore be used to identify alternatively spliced exons with very high specificity, making less than 0.3% false-positive calls. We note that because the ratio of constitutive to alternative exons in the genome is probably higher than in our training set, and because our training set may have some other unknown bias, the performance in genome-wide application of the rule may be somewhat lower.

To test this classifier in a genome-wide manner, as well as to discover novel splice variants in the human genome, we collected a large set of 108,983 human exons, for which a mouse counterpart could be identified (see Methods). To ensure the coherence of the analysis, we excluded our training exons from this analysis. For each of the exons, all classifying parameters were calculated. Out of the 108,983 human exons, 952, or ~1%, were found to comply with the above-mentioned combination of parameters. Information on these 952 exons appears as Supplemental material.

To check whether these exons are indeed alternatively spliced, we searched for human expressed sequences (ESTs or cDNAs) that skip the exons but contain the two flanking exons. For 453 (48%) of the 952 candidate alternative exons there was such skipping evidence. For comparison, only 7% (7495 exons) out of our entire set of 108,983 exons had similar skipping EST evidence. This means that our classification rule indeed substantially enriches for alternatively spliced exons.

Moreover, there is evidence that EST databases can contain spurious sequences that appear as splice variants but are, in fact, artifacts caused by aberrant splicing. Such splicing artifacts are usually characterized by low EST support, although there are

cases in which real, functional splice variants are supported by a single EST (Sorek et al. 2004). Indeed, only 17% of the 453 exons that were classified as 'alternative' by our rule had their exon-skipping supported by only one EST–the rest were supported by two or more. In comparison, skipping was supported by only a single EST in 46% of the total 7495 exons that showed skipping EST evidence. This suggests that our classification rule enriches for alternatively spliced exons with higher probability of being "real" relative to alternative exons merely supported by EST evidence. To calculate the classification sensitivity of the whole-genome analysis while eliminating the low EST coverage factor, we took only exons that were supported by at least 10 human ESTs skipping the exon. There were 873 such exons in the entire set of 108,983 exons, and 176 in our set of 453 exons classified as alternatives. This means that the sensitivity of our analysis on the whole genome is at least 20% (176/873). This is probably an underestimate, as we eliminated our training exons-set from the whole-genome analysis.

We manually examined the remaining 499 candidate alternative exons (952 − 453) for which no EST/cDNA showing an exon skipping event was found, by using the UCSC genome browser (April 2003). We found that for 190 additional exons (out of the 499) there was a human expressed sequence showing patterns of alternative splicing other than exon skipping [41 cases (22%) of alternative donor/acceptor; 33 cases (17%) of intron retention; 14 cases (7%) of mutually exclusive exons. More complicated types, such as double and triple exon skipping, comprise the remaining]. Thus, for 643 (453 + 190; 68%) of the 952 candidate alternative exons identified by our method, there was independent evidence for alternative splicing in dbEST and RefSeq.

But what about the remaining 309 candidate exons for which no EST or cDNA indicating the skipped isoform was found? These can still be rarely expressed alternatively spliced exons, or exons that are specific to a tissue, developmental stage, or condition which is underrepresented in dbEST, so that an EST representing their skipping isoform has not been sequenced yet. Indeed, although on average there were 32 supporting expressed sequences per exon in our general set of 108,983 exons (median 10), the support for the 309 candidate alternative exons was much smaller, averaging 14 sequences (median 7). This shows that the 309 candidate exons are supported by fewer ESTs than the average exon, in accordance with our hypothesis that under-representation in dbEST is the cause for not identifying them as alternatively spliced.

To test whether these candidate alternative exons for which no skipping ESTs were found are indeed alternative, we selected 5% of them (15 exons) for experimental verification (Table 2). Only exons with EST support equal to or less than the average (14 sequences) were selected for this verification, as such alternative splicing events are more likely to have been missed in dbEST due to low sampling and not due to a their appearance in a transient developmental state or in a rare condition. For each of these 15 exons, primers were designed from the two flanking exons. RT–PCR reactions were carried out with RNA extractions of 14 different tissue types (see Methods). For nine of these exons, a splice variant was detected in at least one of the 14 tissues tested (Fig. 2). In six of the nine cases the variant represented exon skipping. Interestingly, in the other three cases the exon was alternatively spliced, but in a pattern other than exon-skipping: Two cases (genes *BAZ1A* and *SMARCD1*) of alternative acceptor site, and one case (*VLDLR*) of intron retention. This is consistent with our genome-wide scan, where 453/643 (70%) cases that were identified according to the classifying parameters were exon-skipping, whereas the remaining 30% exhibited other types of alternative splicing.

The above experimental results indicate that at least 60% (9/15) of our predictions are true (although this estimate can have a relatively large variance, due to the small size of exon set tested). Some or all of the remaining six exons might also be alternatively spliced, but in a tissue other than the ones we tested, or in an early developmental stage. We therefore believe

Table 2. Experimental Validation of Predicted Alternatively Spliced Exons

Gene	Alt exon[a]	PCR confirmed[b]	Type of alternative confirmed[c]	Gene description
FGF11	2	Yes	Skip	Fibroblast growth factor 11
EFNA5	4	Yes	Skip	Ephrin-A5
NCOA1	8	Yes	Skip	Steroid nuclear receptor coactivator
PAM	22	Yes	Skip	Protein associated with Myc mRNA
GOLGA4	9	Yes	Skip	Golgi autoantigen, golgin subfamily a, 4
NPR2	9	Yes	Skip	Natriuretic peptide receptor B/guanylate cyclase B
VLDLR	9	Yes	Int Ret[d]	Very low density lipoprotein receptor
BAZ1A	12	Yes	Alt 3′ss[e]	Bromodomain adjacent to zinc finger domain protein 1A
SMARCD1	7	Yes	Alt 3′ss[f]	SWI/SNF related, matrix associated, actin-dependent regulator of chromatin, subfamily d, member 1
PRKCM	15	No		Protein kinase C, mu
TIAM2	12	No		T-cell lymphoma invasion and metastasis 2
MDA5	4	No		Melanoma differentiation associated protein-5
RNASE3L	15	No		Nuclear RNase III
HAT1	7	No		Histone acetyltransferase 1
DICER1	6	No		Dicer1, Dcr-1 homolog (Drosophila)

[a]Serial number of exon (out of gene's exons) identified as alternative.
[b]For each predicted exon, primers were designed from its flanking exons and RT–PCR was conducted using total RNA from 14 different tissue types: cervix, uterus, ovary, placenta, breast, colon, pancreas, liver + spleen, brain, prostate, testis, kidney, thyroid, and assorted cell-lines. Products were sequenced, and alternative splicing was searched.
[c]Type of alternative splicing: Skip, exon-skipping; Alt 3′ss, alternative 3′ splice site (acceptor); Int Ret., intron retention.
[d]Retention of intron 8 (size 103 nucleotides) was detected in *VLDLR*.
[e]Deletion of 86 nucleotides was detected on the 3′ end of exon 12 7 of *BAZ1A*.
[f]Extension of 44 nucleotides was detected on the 3′ end of exon 12 of *SMARCD1*.

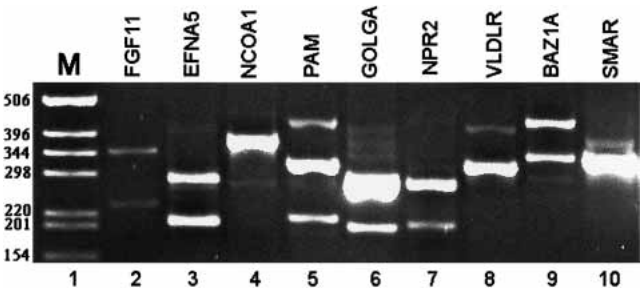


Figure 2 Experimental validation for the existence of alternative splicing in selected predicted exons. RT–PCR for 15 exons (detailed in Table 2), for which no EST/cDNA indicating alternative splicing was found, was conducted over 14 different tissue types and cell lines (see Methods). Detected splice variants were confirmed by sequencing. For nine of these exons a splice isoform was detected in at least one of the tissues tested. Only a single tissue is shown here for each of these nine exons. Lane *1*, DNA size marker. Lane *2*, exon 2 skipping in *FGF11* in ovary tissue (the 344-nt and 233-nt products are exon inclusion and skipping, respectively). Lane *3*, exon 4 skipping in *EFNA5* gene in ovary tissue (exon inclusion 287 nt; skipping 199nt). Lane *4*, exon 8 skipping in *NCOA1* gene in placenta tissue (exon inclusion 377 nt; skipping 275 nt). Lane *5*, exon 22 skipping in *PAM* gene in cervix tissue (exon inclusion 323 nt; skipping 215 nt). Additional upper band contains a novel exon in *PAM*. Lane *6*, exon 9 skipping in *GOLGA4* gene in uterus tissue (exon inclusion 288 nt; skipping 213 nt). Lane *7*, exon 9 skipping of *NPR2* gene in placenta tissue (282nt inclusion; 207nt skipping). Lane *8*, intron 8 retention in *VLDLR* gene in ovary tissue (wild type 324 nt; intron retention 427 nt). Lane *9*, alternative acceptor site in exon 12 of *BAZ1A* in ovary tissue (wild type 351 nt; alternative acceptor variant 265 nt). The uppermost band represents a new exon in *BAZ1A*, inserted between exons 12 and 13. Lane *10*, alternative acceptor site in exon 7 of *SMARCD1* in uterus tissue (wild type 353 nt; exon 7 extension 397 nt).

that the actual prediction rate of this method may be even higher.

The classification rule that was chosen for the experimental verification retrieves alternatively spliced exons with a very high specificity (less than 0.3% false-positive rate) but at the price of a relatively low sensitivity (20%–32%). Other rules can be chosen in which sensitivity is higher, but naturally this would increase the false-positive rate of the prediction. Figure 3 presents a sensitivity versus false-positive rate plot (ROC curve) for different rules selecting for increasing number of alternative exons from our test set of 243 exons. As shown in the figure, it is possible to employ a rule that would identify up to 73% of the alternative exons, but this rule would also retrieve 36% of the constitutively spliced exons (the upper limit of 73% is due to the Boolean nature of the "divisibility by 3" feature). Note that because most of the exons in the human genome are constitutive, such a rule would have low predictability for exon skipping: Assuming, for example, that ~10%, or 20,000 of the ~200,000 predicted exons in the human genome are alternative, the probability that an exon identified by the 73% : 36% rule would really be alternative is only 18% (0.73*20,000/[0.73*20,000 + 0.36*180,000]). This is why we preferred a rule with close to zero false positives. The curve in Figure 3 presents a variety of alternatives, and allows the selection of a rule for a desired target specificity or sensitivity. For example, 50% sensitivity is achievable at an ~1.8% false-positive rate.

Our method is able to identify alternative splicing ab initio. Other computational approaches to detect alternative splicing were previously described, but most of them used ESTs and/or cDNAs, or information from transcripts predicted using ESTs, to predict alternative splicing (e.g., Clamp et al. 2003; Haas et al. 2003; for review, see Modrek and Lee 2002). There was also an attempt to predict alternatively spliced exons using suboptimally scored exons in the gene structure prediction software GENSCAN (Burge and Karlin 1997; see http://genes.mit.edu/GENSCANinfo.html), but as far as we know this prediction method was not tested experimentally.

We have described a novel computational method for prediction of alternative splicing. A possible improvement of the method could be the addition of more classifying features. One such feature could be the comparison of the flanking intronic sequences between the human and other genomes. For example, we were able to locate in the chicken genome 72 and 328 exons from our original alternative and constitutive training sets, respectively. Of the 72 alternatively spliced exons, 34 (47%) had conserved sequences in both their upstream and downstream introns when human and chicken genomes were compared; only 10 (3%) of the 328 constitutively spliced exons that could be found in the chicken genome had such intronic conservation (data not shown).

Currently, our classifier mainly identifies exon-skipping events in exons conserved between human and mouse. In the future, it could develop into a more general alternative splicing predictor that would identify other types of alternative splicing. The ultimate goal of such a predictor would be genome-based prediction of all splice variants, including their pattern of alternative splicing (i.e., in which tissue would the exon be inserted). This could set the foundations for understanding the absolute number of exons that are alternatively spliced and might ultimately lead to narrowing the gap between the genome and the proteome, and thereby advance toward revealing the full extent of our proteome's complexity.

METHODS

Enumeration Over Features in Training Set

Training sets of alternatively spliced internal exons and constitutively spliced internal exons were taken from our previous study (Sorek and Ast 2003). For the present analysis we eliminated from our constitutive exons' set, exons for which novel evidence for alternative splicing appeared in the newer version of GenBank, 136. This left us with 1753 constitutive exons.

The thresholds used in the enumeration of classification rules were as follows: Exon identity thresholds were 100%, at least 99%, at least 98%, and so forth until 80%; exon lengths were below 18 bp, 23 bp, 28 bp, . . . , 198 bp and 1000 bp; length of human/mouse local alignment of the 100 nearest upstream (or downstream) intronic nucleotides using Sim4 (Florea et al. 1998)

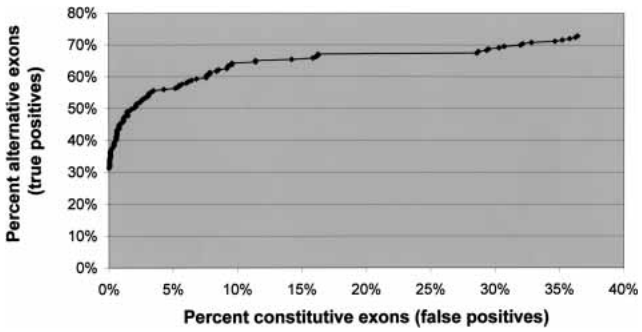


Figure 3 Sensitivity vs. false-positive rate in classification rules. Each square on the curve represents the performance of a single classification rule. *x*-axis, 1-specificity, i.e., percent constitutive exons (false positives) retrieved by the rule. *y*-axis, sensitivity, i.e., percent alternative exons (true positives) identified by the rule. Values were computed relative to the training set. Rules that were used for this plot are provided as Supplemental material.

was at least 100, 95, 90, … 0; minimum identity level in the locally aligned segment of the upstream (or downstream) region was 100%, 97%, 94%, …, 67%; exon divisibility by three had two categories, 'yes' or 'no'. Overall we enumerated more than 100 million different combinations of features.

Genome-Wide Retrieval of Human and Mouse Orthologous Exons

For the genome-wide compilation of human exons, human ESTs and cDNAs were obtained from NCBI GenBank version 136 (June 2003) (www.ncbi.nlm.nih.gov/dbEST) and were mapped to the human genome (April 2003 assembly, www.ncbi.nlm.nih.gov/genome/guide/human) using the spliced alignment module described (Sorek et al. 2002; Sorek and Safer 2003). For each expressed sequence, all mappings of internal exons on the human genome were retrieved. Only exons flanked by AG/GT or AG/GC splice sites were allowed. Thus, 185,799 human exons mapped to the human genome were retrieved.

To find the mouse ortholog for each human exon, we first aligned the mouse expressed sequences from GenBank version 136 to the human genome, as described (Sorek and Ast 2003). Mouse sequences exactly spanning human exons were aligned to the mouse genome as well, and the corresponding sequence on the mouse genome was declared as the orthologous mouse exon, if it was flanked by AG/GT or AG/GC legal splice sites.

Human exons for which no spanning mouse expressed sequence was detected were aligned directly to the mouse genome. Hits spanning the full length of the exon, and were flanked by AG/GT or AG/GC legal splice sites, were declared as the orthologous mouse exons.

Altogether, these searches retrieved 108,983 pairs of exons in the human and mouse genomes (this set does not contain the exons from our two training sets). For each such exon, all classifying parameters were calculated as follows. Conservation between exons was calculated from aligning the human and mouse exons using the global alignment program "GAP" of the GCG software package with default parameters (Womble 2000). Conservation in the flanking intronic sequences was calculated by Sim4 as described (Sorek and Ast 2003). Sim4 detects exact matches of length 12 and extends them in both directions with a score of 1 for a match and 5 for a mismatch, stopping when extensions no longer increase the score (Florea et al. 1998). Exon size and divisibility by three were retrieved from the exon sequence itself.

Reverse Transcription of mRNA Samples

cDNA was obtained by reverse transcription of total RNA from the following human tissue samples: (1) Brain pool, a pool of brain-derived RNA samples (Biochain – Normal); (2) Prostate pool, a pool of prostate-derived RNA samples (Biochain – Normal); (3) Testis pool, a pool of testis-derived RNA samples (Biochain – Normal); (4) Kidney pool, a pool of kidney-derived RNA samples (Biochain – Normal); (5) Thyroid pool, a pool of thyroid-derived RNA samples (Biochain – Normal); (6) Assorted cell-line pool, a pool of cell line-derived RNA samples from the cell lines: DLD, MiaPaCa, HT29, THP1, MCF7 (ATCC); (7) Cervix pool, a pool of three cervix-derived RNA samples, mixed origin (Tumor and Normal, in-house tissue samples); (8) Uterus pool, a pool of three uterus-derived RNA samples (Biochain – Normal), mixed origin (Tumor and Normal); (9) Ovary pool, a pool of five ovary-derived RNA samples (Biochain – Normal), combined with two samples of mixed origin (Tumor and Normal); (10) Placenta, one sample of placenta-derived RNA (Biochain – Normal); (11) Breast pool, a pool of three breast-derived RNA samples of mixed origin (two from tumor and one from normal in-house tissue samples); (12) Colon and intestine, a pool of five colon-derived RNA of mixed origin (Tumor and Normal), combined with one intestine (Normal) -derived RNA sample (in-house tissue samples); (13) Pancreas, one sample of pancreas-derived RNA (Biochain – Normal); (14) Liver and spleen, one sample of liver-derived RNA (Biochain – Normal), one sample of spleen-derived RNA (Biochain – Normal), combined with one sample of HepG2 cell line (liver tumor – ATCC) derived RNA.

RNA was incubated with a random hexamer primer mix (Invitrogen), denatured at 70°C for 5 min, and transferred to 4°C for hexamer annealing. Reverse transcription was done by Superscript II Reverse transcriptase (Invitrogen) in the presence of RNAsin (Promega) at 37°C for 1 h. Reaction was terminated by enzyme deactivation on beads (Promega).

Amplification of Splicing Products

For each exon tested, oligonucleotide primers were designed from its flanking exons (Supplemental Table 1). Amplification was performed for 35 cycles, consisting of 94°C for 45 sec, annealing at a primer-specific temperature (4°C below the primer's TM) for 45 sec, and extension at 72°C for 1 min. The cycle was ended by one stage of gap filling at 72°C for 10 min. The products were resolved on 2% agarose gel and confirmed by sequencing.

ACKNOWLEDGMENTS

We thank Amos Tanay, Irit Gat-Viks, and Gideon Dror for fruitful discussion, and Kinneret Savitsky, Dvir Dahary, and Pini Akiva for critical reading.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Berget, S.M. 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270:** 2411–2414.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474:** 83–86.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Cartegni, L., Chew, S.L., and Krainer, A.R. 2002. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nat. Rev. Genet.* **3:** 285–298.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., et al. 2003. Ensembl 2002: Accommodating comparative genomics. *Nucleic Acids Res.* **31:** 38–42.

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.

Graveley, B.R. 2001. Alternative splicing: Increasing diversity in the proteomic world. *Trends Genet.* **17:** 100–107.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr., R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31:** 5654–5666.

Johnson, J.M., Castle, J., Garrett-Engele, 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.

Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J. 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.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Maniatis, T. and Tasic, B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418:** 236–243.

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.

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. *Nat. Genet.* **34:** 177–180.
Modrek, B., Resch, A., Grasso, C., and Lee, C. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29:** 2850–2859.
Nurtdinov, R.N., Artamonova, I.I., Mironov, A.A., and Gelfand, M.S. 2003. Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.* **12:** 1313–1320.
Resch, A., Xing, Y., Alekseyenko, A., Modrek, B., and Lee, C. 2004. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.* **32:** 1261–1269.
Sorek, R. and Ast, G. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13:** 1631–1637.
Sorek, R. and Safer, H.M. 2003. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.* **31:** 1067–1074.
Sorek, R., Ast, G., and Graur, D. 2002. Alu-containing exons are alternatively spliced. *Genome Res.* **12:** 1060–1067.
Sorek, R., Shamir, R., and Ast, G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20:** 68–71.
Thanaraj, T.A. and Stamm, S. 2003. Prediction and statistical analysis of alternatively spliced exons. *Prog. Mol. Subcell. Biol.* **31:** 1–31.
Womble, D.D. 2000. GCG: The Wisconsin Package of sequence analysis programs. *Methods Mol. Biol.* **132:** 3–22.

WEB SITE REFERENCES

http://genes.mit.edu/GENSCANinfo.html; GENSCAN.
www.ncbi.nlm.nih.gov/dbEST; GenBank version 136 (June 2003).
www.ncbi.nlm.nih.gov/genome/guide/human; Human genome (April 2003 assembly).

Received March 14, 2004; accepted in revised form June 2, 2004.

BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 7 2005, pages 897–901
doi:10.1093/bioinformatics/bti132

Genome analysis

Accurate identification of alternatively spliced exons using support vector machine

Gideon Dror[1,*], Rotem Sorek[2,3] and Ron Shamir[4]

[1]The Academic College of Tel-Aviv-Yaffo, Tel Aviv 4044, Israel, [2]Department of Human Genetics, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel, [3]Compugen, Tel Aviv 69512, Israel and [4]School of Computer Science, Tel Aviv University, Tel Aviv 69073, Israel

Received on September 7, 2004; revised and accepted on November 1, 2004
Advance Access publication November 5, 2004

ABSTRACT

Motivation: Alternative splicing is a major component of the regulatory action on mammalian transcriptomes. It is estimated that over half of all human genes have more than one splice variant. Previous studies have shown that alternatively spliced exons possess several features that distinguish them from constitutively spliced ones. Recently, we have demonstrated that such features can be used to distinguish alternative from constitutive exons. In the current study, we used advanced machine learning methods to generate robust classifier of alternative exons.

Results: We extracted several hundred local sequence features of constitutive as well as alternative exons. Using feature selection methods we find seven attributes that are dominant for the task of classification. Several less informative features help to slightly increase the performance of the classifier. The classifier achieves a true positive rate of 50% for a false positive rate of 0.5%. This result enables one to reliably identify alternatively spliced exons in exon databases that are believed to be dominated by constitutive exons.

Availability: Upon request from the authors.

Contact: gideon@mta.ac.il

1 INTRODUCTION

Alternative splicing is a process through which one gene can generate several distinct proteins. It occurs by the alternative usage of exons or parts of exons within pre-mRNA transcripts, and can be specific to a tissue, developmental stage or a condition such as stress (Maniatis and Tasic, 2002).

Computational prediction of alternative splicing usually involves the usage of expressed sequences, i.e. expressed sequence tags (ESTs) or cDNAs [reviewed in Graveley (2001) and Modrek and Lee (2002)]. Through such predictions, in addition to microarray analyses, several studies have estimated that alternative splicing occurs in 35–74% of all human genes (Brett *et al*., 2000; Kan *et al*., 2001, 2002; Lander *et al*., 2001; Mironov *et al*., 1999; Modrek *et al*., 2001; Johnson *et al*., 2003). However, ESTs and microarrays produce only a snapshot of the tissue they sample, in the specific time and condition it was sampled. Exons that are alternatively spliced in conditions other than the ones sampled will evade detection.

Recently, we have described several features in which alternative exons differ from constitutive ones. These features include the size of the exon, its divisibility by 3, the identity level when aligned to its mouse ortholog exon and the human/mouse conservation in the intronic sequences flanking the exon (Sorek and Ast, 2003; Sorek *et al*., 2004a). Using brute-force enumeration, we demonstrated that a combination of these features could be used to classify alternative exons with a true positive rate of ∼30% for a false positive rate <1%, regardless of their representation in ESTs (Sorek *et al*., 2004b).

In the current study we use state-of-the-art machine learning methods, along with additional sequence features, to generate a robust classifier of alternative exons. We achieved better sensitivity for a similar specificity performance—a true positive rate of 50% for a false positive rate of 0.5%. Moreover, not only is our performance measure more robust, but also we get much higher area under the ROC curve, which provides a proper measure for the quality of ranking of a classifier (Ling *et al*., 2003). We also report on the merit of many additional sequence features extracted from the vicinity of the exon.

2 METHODS

2.1 Dataset

The dataset comprises of 243 alternative and 1753 constitutive exons that are conserved between human and mouse. The data are described in detail in our previous studies (Sorek and Ast, 2003; Sorek *et al*., 2004b). Briefly, alternative exons in this set are exons that were found to be skipped both in the human and mouse transcriptomes; and constitutive exons are exons that are supported by at least four expressed sequences, with no evidence of ESTs skipping them, both in human and in mouse.

2.2 Data representation

For the current study, we used the seven features described in our previous study, as well as 221 additional new sequence features. The original features were (1) exon length, (2) exon divisibility by 3 (a Boolean feature), (3) percent identity when aligned to the mouse counterpart and (4) conservation in the upstream and downstream intronic sequences. Each of the two 'intronic conservation' features (upstream and downstream) was divided into two sub-features: (1) length of the best human/mouse local alignment in the 100 intronic nucleotides nearest to the exon (where only local alignments with at least 12 consecutive perfectly matching nucleotides were considered) and (2) identity level in this local alignment. Local alignments were performed using sim4 (Florea *et al*., 1998) as described by Sorek *et al*. (2004b).

Additional features tested here include 3-tuple counts, computed separately for the sequence of the exon, the 100 bases of the intron upstream of the exon

*To whom correspondence should be addressed.

(called 'pre') and the 100 bases of the intron downstream of the exon (called 'post'), adding up to $64 \times 3 = 192$ features.

We also used information from the 5′ splice site (5′ss, also called donor site) sequence. The nucleotide composition of the 5′ss reflects its base-pairing with small nuclear RNAs such as U1 (Zhuang and Weiner, 1986). It was previously shown that the composition of the 5′ss differs between alternative and constitutive exons (Clark and Thanaraj, 2002). It was also demonstrated that the alteration of 5′ss sequences can result in transition from alternative to constitutive splicing, or vice versa (Sorek *et al*., 2004b). We therefore used position-dependent single base counts at the 5′ss sequence, ranging from the -3 to the $+6$ position relative to the splice site (not including invariant positions $+1$ and $+2$). This added up to $4 \times 7 = 28$ features.

The last feature used was the intensity of the poly-pyrimidine tract (PPT), which was defined as the number of pyrimidines (Cs and Ts) in a window of 15 bases in the last 19 nt of the upstream intron (not including the last 4 nt of the intron).

We also examined position-dependant base combinations of three bases at the splice site, that were shown to be the highly discriminating features for a similar task (Zhang *et al*., 2003). However, preliminary analysis has indicated that the 3-base 5′ss combinations are not as informative for the present task and were therefore not included in the sequel.

We concatenated all features into one vector representation in $\mathbb{R}^N$ where $N = 7 + 192 + 28 + 1 = 228$. Since the features have very different distributions (binary, integer and real numbers), we standardized them such that each feature has a zero mean and variance of one. We denote the i-th standardized vector by $\mathbf{x}^i = (x_1^i, \ldots, x_N^i)$. Each example is labeled by $y^i = -1$ or $y^i = +1$, depending on whether it represents a constitutive or alternative exon, respectively.

2.3 Data partitioning

In the experiments reported here, we randomly split the dataset entries into a training and a testing set at a ratio of 2:1. Feature vectors as described above were used as examples for training various classifiers, while the testing examples were not exposed to the system during learning, feature selection and hyper-parameter selection phases.

2.4 Support vector machines

Support vector machine (SVM) learning is an area of statistical learning, subject to extensive research (Vapnik, 1998; Schölkopf *et al*., 1999; Smola *et al*., 2000). SVM has been used extensively for a wide range of applications in science, medicine and engineering and has shown excellent empirical performance. Recent bioinformatic investigations utilizing SVM include Brown *et al*. (1999), Zien *et al*. (1999), Jaakkola *et al*. (2000) and Leslie *et al*. (2004). More recently, SVM was used for the detection of splicing sites (Yamamura and Gotoh, 2003; Sun *et al*., 2003; Zhang *et al*., 2003). SVM has the following advantages for the present task:

(1) SVM is based on the principle of risk minimization and thus provides good generalization control. This allows one to work with datasets that contain many irrelevant and noisy features.

(2) Using non-linear kernels, SVM can model non-linear dependences among features and the target, which may prove advantageous for the problem at hand.

(3) SVM allows natural control on the relative cost of false positives and false negatives.

In the present research we used soft-margin SVM implemented in SVM$^{\text{light}}$ (Joachims, 1999). The latest version of this software is available at http://svmlight.joachims.org/.

2.5 Hyper-parameter selection

SVM training involves fixing several hyper-parameters. The values of these hyper-parameters determine the function that SVM optimizes and therefore have a crucial effect on the performance of the trained classifier. To identify an optimal hyper-parameter set we used 10-fold cross-validation on the training set, which is a robust method for hyper-parameter tuning (Duan *et al*., 2003). The cross-validation was used also to tune the number of features used by the classifier, as discussed in the next subsection.

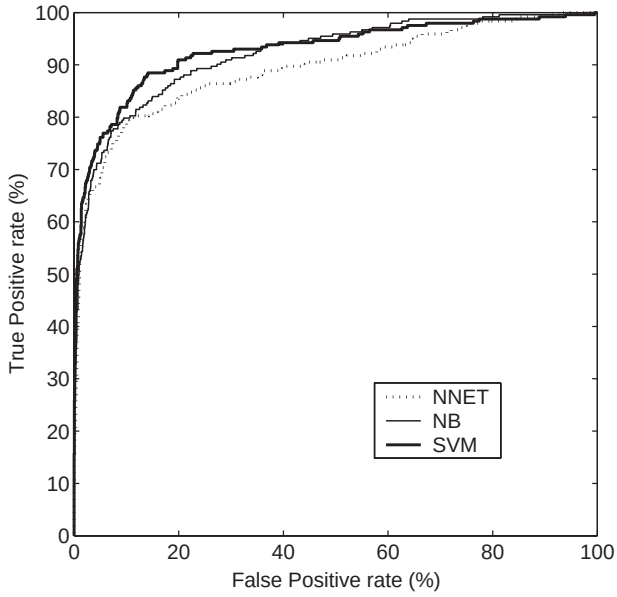


Fig. 1. The ROC curves of the three classifiers. The AUC for neural network (NNET), the Naive-Bayes (NB) and the SVM are 0.92, 0.89 and 0.93, respectively. The optimal performance of the first two classifiers was obtained with 11 features. SVM classifier uses a linear kernel with hyper-parameters $c = \sqrt{10}$ and $j = 0.5$. The SVM ROC with these hyper-parameters is quite insensitives to the number of features, as is shown below.

We used several kernels: linear, polynomials of degrees 2 and 3 and Gaussian kernel. For each kernel, we performed a grid search over the values of the slack parameter c, and the cost factor j, by which training errors on positive examples (false negatives) outweigh errors on negative examples (false positives). For the Gaussian kernel, we repeated the search for several values of r, the parameter that controls the width of the kernel.

For each hyper-parameter combination we measured the 10-fold cross-validation area under the ROC curve (AUC). AUC (Agarwal *et al*., 2004) is a global performance measure since it is integrated over all threshold values. However, for the task of identifying alternative exons within a population in which the vast majority of exons are constitutive, one specifically needs high discrimination power at low false-positive rate. To this end, we also measured the true positive rate for a small value $0 < \alpha \ll 1$ of the false positive rate. We denote this performance measure as TP_α. For small values of α, TP_α is very sensitive to the minute details in the distribution of examples (e.g. the details of split between the training set and test set). Therefore we did not directly try to maximize it, so as to reduce the risk of severe overfitting. We selected the kernel and hyper-parameter set that gave the highest value of $\lambda\text{AUC} + (1-\lambda)\text{TP}_\alpha$, where $0 < \lambda < 1$. We found that for the whole range $0.1 < \lambda < 0.9$ there was very good generalization, and that the final results varied only insignificantly.

The best cross-validation performance, for a value of $\lambda = 0.5$, was obtained by the Gaussian kernel, with intermediate slack parameter $c = \sqrt{10}$ and cost factor $j = 1/2$.

In addition to SVM, we also used Naive-Bayes and neural network classifiers. For training the neural network we used the Levenberg–Marquardt algorithm with Bayesian regularization. For both Naive-Bayes and neural network, we performed a search in hyper-parameter space and among several architectures to optimize performance. Figure 1 shows the ROC curves of the best SVM, Naive-Bayes and neural network classifiers. The values of AUC are quite close to each other.

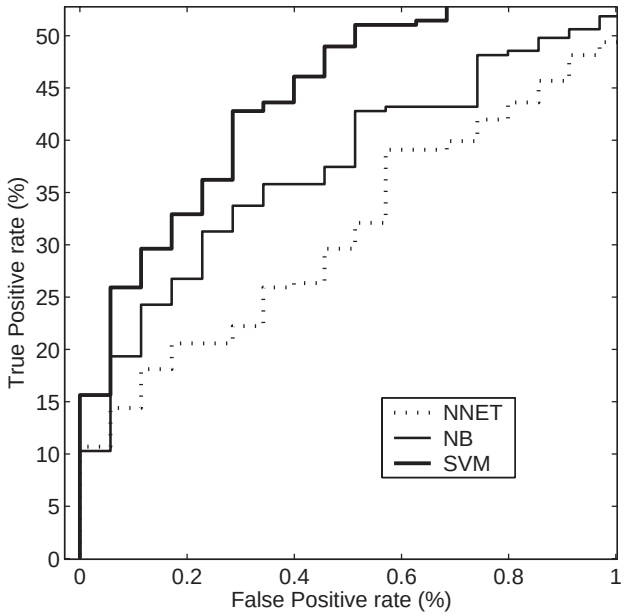


Fig. 2. The ROC curve of the three classifiers in the region of small false positive rate, FP $< 1\%$. It is evident that SVM considerably outperforms the neural network (NNET) and the Naive-Bayes (NB) classifiers.

Figure 2 depicts the ROCs of the three classifiers at the region of low false positive rate. It is clear that for all the ranges shown, $0 < \alpha < 1\%$, SVM achieves considerably higher TP_α and therefore better resolution in identifying alternative exons.

2.6 Feature selection

The potential benefits of feature selection are 3-fold: improving the performance of the classifier, producing a cost-effective classifier, and providing better understanding of the problem at hand. In our case, we used feature selection primarily for the purpose of enhancing the classifier's performance. Although state-of-the-art classifiers such as SVM and neural networks that incorporate regularization techniques can accommodate situations where many of the features are redundant or noisy, removing non-informative features can considerably enhance their performance (Guyon and Elisseeff, 2003).

Preliminary analysis of the data has shown that the seven features used in the original paper by Sorek *et al*. (2004c) are much more informative for the classification task than the vast majority of the remaining features. However, a χ^2-test showed that for several features, the distributions of the positive and negative examples are significantly different. Namely, they potentially convey useful information for the task of classification.

Our feature selection criterion is that used by Golub *et al*. (1999). For each feature x_j, $j = 1, \ldots, N$, we calculated the mean $\mu_j^+(\mu_j^-)$ and standard deviation $\sigma_j^+(\sigma_j^-)$ using only positive (negative) examples. The score

$$F(x_j) = \left| \frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ + \sigma_j^-} \right| \quad (1)$$

serves as a simple heuristic for ranking the features according to how well they discriminate the positive and negative examples.

To avoid overfitting, we used the feature selection within the cross-validation loop. In other words, to estimate the performance of a classifier which uses n features, where $n \leq N$, we used Equation 1 on each split of the training set and simply took the n features with the highest $F(x_j)$ scores. Needless to say that this procedure produces a unique feature set for each split.

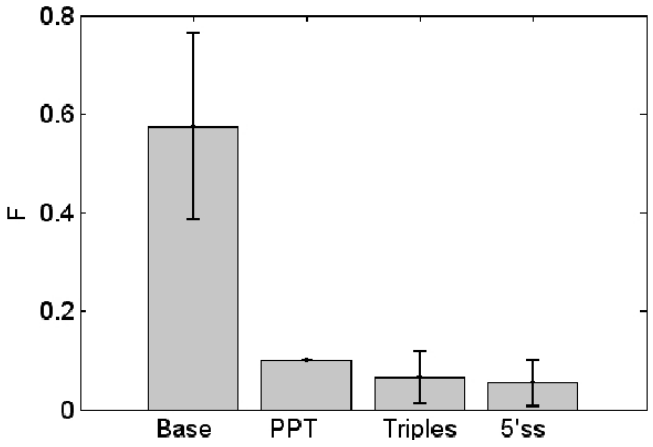


Fig. 3. The discriminative power of different feature types. For each set of features we plot the average values of F. Feature sets are: the original features of Sorek *et al*. (2004b) (Base—7 features), intensity of poly-pyrimidine tract (PPT—1 feature), triple counts (Triples—192 features) and position-dependent single base counts at the 5′ss (5′ss—24 features). The standard deviation of F within each set is expressed by the error bars.

Table 1. Most informative triples

Triple	Location	$\mu^+(\sigma^+)$	$\mu^-(\sigma^-)$	F	P-value
TTC	Pre	0.033 (0.021)	0.026 (0.016)	0.215	5.77E−7
AGG	Post	0.014 (0.017)	0.022 (0.020)	0.212	1.13E−9
GAG	Pre	0.008 (0.012)	0.014 (0.015)	0.210	5.94E−9
AGG	Pre	0.010 (0.014)	0.015 (0.016)	0.186	3.30E−7
GGA	Post	0.012 (0.015)	0.018 (0.017)	0.185	1.21E−7
GAG	Post	0.013 (0.016)	0.020 (0.019)	0.181	3.31E−7
TTT	Post	0.056 (0.055)	0.039 (0.042)	0.178	2.38E−6
TTT	Pre	0.070 (0.053)	0.052 (0.047)	0.178	1.99E−6
GTG	Exon	0.014 (0.016)	0.019 (0.015)	0.168	7.29E−7
AAG	Post	0.015 (0.014)	0.019 (0.014)	0.168	4.54E−6

The 10 most informative triples ranked by their F-value. For each triple, we specify its location relative to the exon (pre, exon, post) and its mean frequency among alternative and among constitutive exons, μ^+ and μ^-, respectively. The standard deviations of the latter quantities are listed in parentheses. For each feature we also list its F-value and the χ^2-P-value, which represents the probability that the distributions of the positive and negative class are sampled from a single distribution.

Figure 3 demonstrates the relative importance of the four parts comprising the feature vectors. It is evident that the set of seven features used by Sorek *et al*. (2004b) (Base) has a much higher discriminative power than the other sets. The F-values within this set fall between 0.287 and 0.834. It should be noted that although the average values of F of the remaining three sets of features, intensity of the PPT, triple counts (Triples) and position-dependent single base counts at the 5′ss are similar, the latter two sets contain many features that are much more informative than the single PPT feature.

In addition to the seven features reported by Sorek *et al*. (2004b), we discovered many features that convey useful information for the task of identifying alternative exons. Table 1 lists the 10 most informative features (all of them triples), together with their mean frequencies among alternative and constitutive exons, their F-score, and their significance level, as measured by χ^2-test. Interestingly, only one of these ten features is a tuple within the exon body, possibly indicating the significance of flanking intronic sequences in the regulation of alternative splicing. This tendency prevails also when inspecting

Table 2. Most informative single base features within the 5′ss region

Base	Position	$\mu^+(\sigma^+)$	$\mu^-(\sigma^-)$	F	P-value
A	4	0.56 (0.50)	0.71 (0.45)	0.163	8.41E−7
G	5	0.65 (0.48)	0.79 (0.41)	0.157	1.37E−6
T	4	0.21 (0.41)	0.12 (0.32)	0.129	3.73E−5
G	3	0.24 (0.42)	0.33 (0.47)	0.101	4.35E−3
T	5	0.13 (0.33)	0.07 (0.25)	0.098	1.45E−3
G	4	0.16 (0.37)	0.10 (0.30)	0.095	2.82E−3
T	3	0.05 (0.23)	0.02 (0.15)	0.078	8.38E−3
C	5	0.09 (0.28)	0.05 (0.21)	0.078	1.18E−2
A	−2	0.70 (0.46)	0.63 (0.48)	0.074	3.45E−2

Informative positions at the 5′ss, ranked by their F-value. For each feature we specify the base, its position relative to the actual splice site and its mean frequency among alternative and among constitutive exons, μ^+ and μ^-, respectively. The standard deviations of the latter quantities are listed in parentheses. For each feature we also list its F and $\chi^2 P$-value.

Position	-3	-2	-1	1	2	3	4	5	6
Consensus	C	A	G	G	T	A/G	A	G	T
A		Alt					Con		
G						Con	Alt	Con	
T						Alt	Alt	Alt	
C								Alt	
	Exon			Intron					

Fig. 4. Different compositions of 5′ss in alternative and constitutive exons. Shown are positions −3 to +6 relative to the 5′ss. Positions −3 to −1 depict the end of the exon, and positions 1–6 are the beginning of the intron. Also shown is the consensus of the 5′ss. Each shaded frame indicates an informative nucleotide in the specific position, which is either over-represent in alternative exons (Alt) or constitutive exons (Con). Dark gray, alternative/constitutive difference is significance to $\alpha \leq 0.01$; light gray, $\alpha \leq 0.05$. For example, in position 4, A is over-represented in constitutive exons, while G and T are more pronounced in alternative ones.

a considerably larger number of top ranking triples, and is therefore a real characteristic of the data.

Importantly, the analysis has revealed biologically significant details. As seen in Table 1, 9 out of 10 informative triples were stretches of purines or pyrimidines in the upstream ('pre') or downstream ('post') introns. From these data, it is clear that there appears to be an under-representation of poly-purine stretches in the intronic sequence proximal to alternatively spliced exons, both upstream and downstream of the exon, and over-representation of poly-pyrimidine stretches in these same regions. Indeed, poly-purine stretches within exons are known to be composed of sequences that regulate splicing (both alternative and constitutive) (Cartegni *et al*., 2002). Therefore, it is possible that some of these discriminative features are parts of splicing regulatory motifs.

We were also able to identify informative features within the 5′ss sequence. Table 2 lists the most informative 5′ss features, ranked by their F-values. As shown in Figure 4, the most informative features lie in positions 3, 4 and 5 of the 5′ss. Such differences in the 5′ss composition of alternative versus constitutive exons were noted previously (Clark and Thanaraj, 2002).

To improve the results we also tried Recursive Feature Elimination (RFE), suggested by Guyon *et al*. (2002). In contrast to the ranking based on F, that considers each feature in isolation, RFE is capable of taking into account dependencies between features, and is therefore considered more

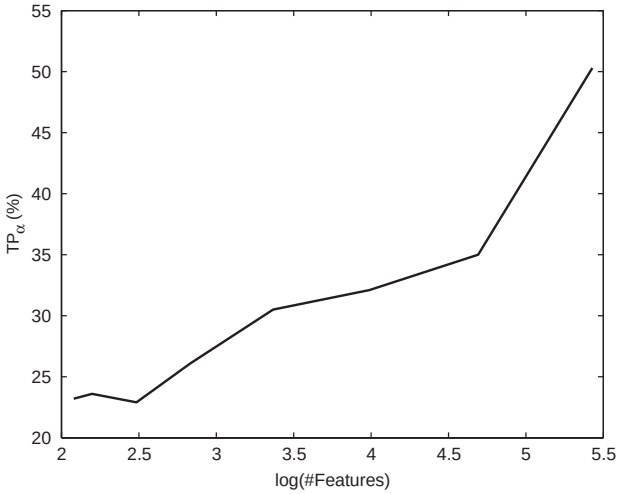


Fig. 5. The behavior of the true positive (TP) rate at a fixed false positive rate $\alpha = 0.5\%$ as a function of the logarithm of the number of features selected. The classifier uses a Gaussian kernel with $c = \sqrt{10}$ and $j = 0.5$.

sophisticated. However, no significant improvement has been observed in either AUC or the value of TP_α. One possible explanation for this is the fact that each input vector x is actually a concatenation of several parts with significantly different distributions. This non-homogeneity introduces a bias which reduces the effectiveness of RFE.

2.7 Performance versus the number of features

To see the effect of the number of features, we used the optimal SVM hyper-parameters obtained by cross-validation and constructed eight classifiers. Each classifier was trained on a different feature subset, where the number of features was one of 8, 9, 12, 17, 29, 54, 109 and 228. The features were selected by their F-value. The performance (AUC, TP_α) of each classifier was measured on the test set, to get an estimate of the performance of the SVM classifier as a function of the number of features selected. Figure 5 shows the dependency of TP_α on the number of features selected for $\alpha = 0.5\%$. Similar analysis of the AUC shows that it varies irregularly between 0.92 and 0.94, with no clear tendency, a behavior that probably originates from finite sample effects.

3 DISCUSSION AND CONCLUSION

Our aim in this paper was to build a classifier that robustly discriminates between constitutively and alternatively spliced conserved exons. To this end, we used a dataset comprising of constitutive and alternative exons in a 7:1 ratio, to train an SVM classifier.

Our feature selection procedure identified several new features whose alternative and constitutive distributions are significantly different. Those features might be involved in splicing regulation.

Using hyper-parameter selection and feature selection combined with cross-validation, a classifier with an AUC score of 0.93 was obtained. More importantly, this classifier is capable of rejecting constitutive exons very effectively at reasonable acceptance rates for true alternative exons. For example, with a false positive rate of 0.5% our classifier empirically achieved ∼50% true positive rate on an untouched test set.

It is important to note that our method is only capable of detecting exon-skipping in exons conserved between human and mouse genomes, because of its heavy reliance on conservation-based features. It is believed that a large proportion of functional alternative splicing is

of the conserved type, but functional species-specific splice variants were also documented (Sorek *et al*., 2004a; Modrek and Lee, 2003). In our method, species-specific alternative splicing event will skip detection, as no conservation-based features can be calculated for them. Therefore, this set of exon-skipping events deserves specific solutions other than ours.

The results of this study are an improvement over our previous study, in which we used only seven features (five of them being conservation-based) to achieve a sensitivity of 30% at false positive rates similar to the ones in this study. The performance of the current study would enable effective scan of the exon database in search for novel alternatively spliced exons, in the human or other genomes.

ACKNOWLEDGEMENT

R.S. was supported by a fellowship from the Clore Israel Foundation.

REFERENCES

Agarwal,S., Graepel,T., Herbrich,R., Har-Peled,S. and Roth,D. (2004) Generalization bounds for the area under an ROC curve. *Technical Report UIUCDCS-R-2004-2433*, Department of Computer Science, UIUC, May 2004.

Brett,D., Hanke,J., Lehmann,G., Haase,S., Delbruck,S., Krueger,S., Reich,J. and Bork,P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, 83–86.

Brown,M., Grundy,W., Lin,D., Christianini,N., Sugnet,C., Ares,M. and Haussler,D. (1999) Support vector machine classification of microarray gene expression data. *Technical Report UCSC-CRL 99-09*, University of California, Santa Cruz CA, June 1999.

Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.*, **3**, 285–298.

Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451–464.

Duan,K., Keerthi,S. and Poo,A. (2003) Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, **51**, 41–59.

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.

Golub,T., Slomin,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M., Bloomfield,C. and Lander,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.

Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learning Res.*, **3**, 1157–1182.

Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **7**, 95–114.

Joachims,T. (1999) Making large-scale SVM learning Practical. In *Advances Kernel Methods—Support Vector Learning*, MIT-Press, Chapter 11, pp. 169–184.

Johnson,J.M., Castle,J., Garrett-Engele,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.

Kan,Z., Rouchka,E.C., Gish,W.R. and States,D.J. (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.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Leslie,C., Eskin,E., Cohen,A., Weston,J. and Noble,W. (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Ling,C., Huang,J. and Zhang,H. (2003) AUC: a better measure than accuracy in comparing learning algorithms. In: *Proceedings of the 2003 Canadian Artificial Intelligence Conference*, pp. 329–341.

Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.

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

Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.

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

Modrek,B. and Lee,C. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, **34**, 177–180.

Smola,A., Bartlett,P., Scholkopf,B. and Schuurmans,D. (eds) (2000) *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA.

Schölkopf,B., Burges,C.J. and Smola,A. (eds) (1999) *Advances in Kernel Methods*. MIT Press, Cambridge, MA.

Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631–1637.

Sorek,R., Shamir,R. and Ast,G. (2004a) How prevalent is functional alternative splicing in the human genome? *Trends Genet.*, **20**, 68–71.

Sorek,R., Shemesh,R., Cohen,Y., Basechess,O., Ast,G. and Shamir,R. (2004b) Non-EST based method for exon-skipping prediction. *Genome Res.*, **14**, 1617–1623.

Sorek,R., Lev-Maor,G., Reznik,M., Dagan,T., Belinky,F., Graur,D. and Ast,G. (2004c) Minimal conditions for exonization of intronic sequences: 5′ splice site formation in alu exons. *Mol. Cell*, **14**, 221–231.

Sun,F., Fan,D. and Li,D. (2003) Identifying splicing sites in eukaryotic RNA: support vector machine approach. *Comput. Biol. Mach.*, **33**, 17–29.

Vapnik,V. (1998) *Statatistical Learning Theory*. Wiley, NY.

Zhang,X., Heller,K., Hefter,I., Leslie,C. and Chasin,L. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res.*, **13**, 2637–2650.

Yamamura,M. and Gotoh,O. (2003) Detection of the splicing sites with Kernel method approaches dealing with nucleotide doublets. *Genome Informatics*, **14**, 426–427.

Zhuang,Y. and Weiner,A.M. (1986) A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. *Cell*, **46**, 827–835.

Zien,A., Ratsch,G., Mika,S., Scholkopf,B., Lengauer,T. and Muller,K. (1999) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.

activation threshold, but also enables the channel to be dynamically modulated by inflammatory products that activate PLC. Finally, it is interesting to note that the C-terminal domain of TRPV3, a warm-sensitive channel with an activation threshold of ~35°C (*21–23*), conspicuously lacks a region corresponding to the 777–792 domain of TRPV1, the minimal essential core of the predicted PIP_2 binding site. Thus, modification of this PIP_2 regulatory domain by genetic, biochemical, or pharmacological mechanisms may have profound effects on sensitivity of primary afferent nerve fibers to chemical and thermal stimuli under normal or pathological conditions.

References and Notes

1. D. W. Hilgemann, S. Feng, C. Nasuhoglu, *Sci. STKE* **2001**, RE19 (2001).
2. R. C. Hardie, *Annu. Rev. Physiol.* **65**, 735 (2003).
3. H. H. Chuang *et al.*, *Nature* **411**, 957 (2001).
4. M. Tominaga, M. Wada, M. Masu, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 6951 (2001).
5. L. W. Runnels, L. Yue, D. E. Clapham, *Nature Cell Biol.* **4**, 329 (2002).
6. M. J. Caterina *et al.*, *Nature* **389**, 816 (1997).
7. M. Tominaga *et al.*, *Neuron* **21**, 531 (1998).
8. M. J. Caterina *et al.*, *Science* **288**, 306 (2000).
9. J. B. Davis *et al.*, *Nature* **405**, 183 (2000).
10. E. D. Prescott, D. Julius, unpublished data.
11. H. Zhang, C. He, X. Yan, T. Mirshahi, D. E. Logothetis, *Nature Cell Biol.* **1**, 183 (1999).
12. C. L. Huang, S. Feng, D. W. Hilgemann, *Nature* **391**, 803 (1998).
13. E. Kobrinsky, T. Mirshahi, H. Zhang, T. Jin, D. E. Logothetis, *Nature Cell Biol.* **2**, 507 (2000).
14. C. Wiedemann, T. Schafer, M. M. Burger, *EMBO J.* **15**, 2094 (1996).
15. V. Vellani, S. Mapplebeck, A. Moriondo, J. B. Davis, P. A. McNaughton, *J. Physiol.* (*London*) **534**, 813 (2001).
16. L. S. Premkumar, G. P. Ahern, *Nature* **408**, 985 (2000).
17. M. Numazaki, T. Tominaga, H. Toyooka, M. Tominaga, *J. Biol. Chem.* **277**, 13375 (2002).
18. P. M. Zygmunt *et al.*, *Nature* **400**, 452 (1999).
19. T. Hofmann *et al.*, *Nature* **397**, 259 (1999).
20. S. E. Jordt, D. Julius, *Cell* **108**, 421 (2002).
21. A. M. Peier *et al.*, *Science* **296**, 2046 (2002).
22. G. D. Smith *et al.*, *Nature* **418**, 186 (2002).
23. H. Xu *et al.*, *Nature* **418**, 181 (2002).
24. Materials and methods, fig. S1, and associated references are available as supporting material on *Science* Online.

25. We thank R. Nicoll and D. Minor for many helpful suggestions and encouragement throughout this project, G. Reid for help with the Peltier system, and members of our laboratory for advice. This work was supported by predoctoral fellowships from the NSF and the University of California, San Francisco, Chancellor's Fund (E.D.P.) and by the NIH (D.J.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/300/5623/1284/DC1
Materials and Methods
Fig. S1
References

19 February 2003; accepted 15 April 2003

The Birth of an Alternatively Spliced Exon: 3′ Splice-Site Selection in *Alu* Exons

Galit Lev-Maor,[1]* Rotem Sorek,[1,2]* Noam Shomron,[1] Gil Ast[1]†

Alu **repetitive elements can be inserted into mature messenger RNAs via a splicing-mediated process termed exonization. To understand the molecular basis and the regulation of the process of turning intronic *Alu*s into new exons, we compiled and analyzed a data set of human exonized *Alu*s. We revealed a mechanism that governs 3′ splice-site selection in these exons during alternative splicing. On the basis of these findings, we identified mutations that activated the exonization of a silent intronic *Alu*.**

Alu elements are short (about 300 nucleotides in length), interspersed elements that amplify in primate genomes through a process of retroposition (*1–3*). These elements have reached a copy number of about 1.4 million in the human genome, composing more than 10% of it (*4*). A typical *Alu* is a dimer, built of two similar sequence elements (left and right arms) that are separated by a short A-rich linker. Most *Alu*s have a long poly-A tail of about 20 to 100 bases (*5*).

Parts of *Alu* elements, predominantly on their antisense orientation, can be inserted into mature mRNAs by way of splicing ("exonization"). Presumably, the exonization process is facilitated by sequence motifs that resemble splice sites, which are found within the *Alu* sequence (*6–9*) (see fig. S1 for a model of exonization). Because *Alu*s are found in primate genomes only, *Alu*-derived exons might contribute to some of the characteristically unique features of primates.

We have previously shown that more than 5% of human alternatively spliced exons are *Alu*-derived and that most, if not all, *Alu*-containing exons are alternatively spliced (*9*). We therefore hypothesized that mutations causing a constitutive splicing of intronic *Alu*s would cause genetic diseases, and indeed we found in the literature several instances in which a constitutive *Alu* insertion caused a genetic disorder (*10–12*).

To study the alternative splicing regulation of exonized *Alu*s, we compiled a data set of exonized *Alu*s from the human genome. An analysis of this data set revealed that two positions along the inverted *Alu* sequence are most commonly used as 3′ splice sites (3′SSs) in *Alu* exonizations: position 279 ("proximal AG") and position 275 ("distal AG"). The relationships between two near AGs in a 3′SS were well characterized previously in the context of constitutive splicing (*13*, *14*). To pinpoint the sequence determinants by which the spliceosome selects one of the two possible AGs in the context of alternative splicing, we aligned the exonized *Alu*s that use either of these AGs to their ancestor.

The 3′SS regions of these instances are shown in Fig. 1. Figure 1 also shows that the proximal AG is selected mostly in exonized *Alu*s of S subfamilies (9 times out of 13), whereas the distal AG is mainly selected in exonized *Alu*s belonging to J subfamilies (12 times out of 16). This differential usage of AG selection in *Alu* subfamilies is probably because of the polymorphism between the J and S subfamilies in position 277 (Fig. 1, colored yellow), which eliminates the distal AG in *Alu*s of the S subfamilies. As a result, the proximal AG is selected. Although another polymorphism at position 275 creates a new distal AG in the S subfamilies, this new AG is six nucleotides downstream from the proximal AG, a distance that was shown to be out of the effective range for selecting a distal AG in constitutive splicing (*14*). Indeed, the cases where *Alu*s of the S subfamilies used the distal AG required mutations that shortened the distance between AGs back to four nucleotides (Fig. 1, colored green). This indicates that when the range between the two AGs is four nucleotides or less the distal AG is preferred and when the distance is six nucleotides or more the proximal is preferred.

However, in five cases (Fig. 1, rows 25 to 29), the proximal AG was selected, even though a distal AG existed less than six nucleotides in range; in all these cases, the G in position –7 (colored purple) was mutated to either A (two cases) or T (three cases). Remarkably, a mutation in the same position in intron 5 of the COL4A3 gene leads to exonization of a silent intronic *Alu*. This *Alu* exon is constitutively spliced, resulting in an Alport syndrome phenotype (*10*). This implies that the G in position –7 suppresses the selection of the proximal AG, causing a shift toward selection of the distal AG. When this G is mutated, the proximal AG is preferred.

[1]Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel. [2]Compugen, 72 Pinchas Rosen Street, Tel Aviv 69512, Israel.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: gilast@post.tau.ac.il

This is supported by the finding that GAG triplets at ends of introns are poorly cleaved in vitro and extremely rare in vivo (*15*).

To examine the above hypotheses, we cloned a minigene of the ADAR2 gene (adenosine deaminase, involved in RNA editing) (Fig. 1, row 1). Previously, exon 8 [denoted as exon 5a in (*16*)] of this gene was found to be an alternatively spliced *Alu*-derived exon, adding 40 amino acids in frame to the protein (*17*). In this exon, the distal AG is used as the 3′SS. Trying to characterize the relationship between proximal and distal AGs in the context of alternative splicing, we generated a set of mutations within the 3′SS.

Whereas the *Alu* exon in the wild-type ADAR2 was included in 40% of the transcripts (Fig. 2B, lane 3), replacement of the G in position –7 to A, U, or C (Fig. 2B, lanes 10 to 12) had two effects. First, as predicted from Fig. 1, the replacement shifted the selection from the distal AG to the proximal one. Second, the replacement resulted in a shift from alternative splicing of the *Alu* exon toward a nearly constitutive inclusion of the exon in the mature transcript. Our results point to the important role of the G in position –7 in shifting the selection toward the distal AG, thus maintaining the alternative splicing of the *Alu*-containing exon. Mutation of that G will likely result in a constitutive inclusion of the *Alu* exon and thus might cause a disease, as occurs in the case of Alport syndrome (*10*).

To check whether the proximal AG affects the selection of the distal AG, we mutated the proximal AG to UC or GA (Fig. 2B, lanes 8 and 9, respectively). The GA mutation resulted in a higher ratio of exon inclusion, reaching more than 85% inclusion instead of 40% in the wild type. The UC mutation caused the splicing of the *Alu* exon to become constitutive, possibly because it strengthened the polypyrimidine tract (PPT) that was originally 18 bases long (on average, the PPT length in exonized *Alu*s was 19 bases ±3). These findings indicate that the proximal AG presumably weakens the selection of the distal AG and is therefore required for maintaining alternative rather than constitutive splicing of the *Alu* exon. To summarize, when the distal 3′SS is used, the G at position –7 suppresses the selection of the proximal AG, and the proximal AG maintains the alternative splicing.

We then sought to understand whether the nucleotide composition between the two adjacent AGs affects 3′SS selection and ratio of alternative splicing. The two AGs are separated by an AC dinucleotide (Fig. 2A). A deletion of both these nucleotides (position –3 and –4) or only the C (Fig. 2B, lanes 5 and 7) resulted in an exon skipping, pointing to the importance of the C in position –3. Deletion or mutations of the A in position –4 to G or C changed the ratio between the two isoforms (Fig. 2B, lanes 6, 13, and 14). This indicates that position –4 also affects the inclusion ratio.

To test whether increased distance between the two AGs shifts the selection toward the proximal AG, we introduced additional nucleotides between the two AGs (Fig. 3A). Increasing the distance between the proximal and distal 3′SS to six or eight nucleotides resulted in *Alu* exon skipping (Fig. 3B, lanes 7 and 8). However, when the distance between the two AGs grew to 10 nucleotides, a residual exon inclusion was recovered in a little more than 3% of the spliced transcripts (Fig. 3B,

Gene name	Alu exon number	subfam	position relative to distal 3'ss 290	289	288	287	286	285	284	283	-7 282	-6 281	-5 280	-4 279	-3 278	-2 277	-1 276	275	274	273	272	271
		J	T	T	T	T	T	T	T	T	G	A	G	A	C	A	G	G	G	T	C	T
		S	T	T	T	T	T	T	T	T	G	A	G	A	C	G	G	A	G	T	C	T
1 ADAR2	8	Jb	T	T	T	T	T	T	T	T	G	A	G	A	C	A	G	G	G	T	C	T
2 TFB2M	4	Jb	T	C	T	T	T	T	T	T	G	A	G	A	C	A	G	G	G	T	C	T
3 MVK	4	Jb	T	T	T	T	T	T	T	T	G	A	G	A	C	A	G	G	G	T	C	T
4 CBFA2T2	3	Jo	T	T	T	T	T	T	T	T	G	A	G	A	C	A	G	G	G	T	C	T
5 NPD002	7	Jb	T	T	T	C	T	T	T	T	G	A	G	A	C	A	G	G	G	T	C	C
6 MOG	3	Jb	T	T	T	T	T	T	T	T	G	A	G	A	C	A	G	T	G	T	C	T
7 n/a	5	Jb	T	T	T	A	T	T	T	T	G	A	G	A	C	A	G	A	G	T	C	T
8 PTGES	2	Jb	T	T	T	T	T	T	T	G	G	A	G	A	C	A	G	A	G	T	C	T
9 DAF*	10	Jb	T	T	T	T	T	T	T	T	G	A	G	A	C	A	G	G	T	T	C	T
10 STK2*	n/a	Jb	T	T	T	T	T	T	T	T	G	A	G	A	C	A	G	A	G	T	C	T
11 MLANA	4	Jb	T	T	T	T	T	C	T	T	G	A	A	A	C	A	G	G	A	A	A	T
12 n/a	24	Jb	T	T	T	T	T	T	T	T	G	A	A	A	C	A	G	C	G	T	C	T
13 ITGB1*	7	Sx	T	T	T	A	T	T	T	T	G	A	G	A	C	A	G	-	-	T	C	T
14 n/a	2	Sg	T	C	T	T	T	T	T	T	T	T	G	A	C	A	G	A	G	T	C	C
15 MBD3	12	Sx	T	T	T	T	T	T	T	T	G	T	G	A	C	A	G	A	G	T	C	T
16 CNN2	6	Sx	T	T	T	A	T	T	T	T	G	A	G	A	T	A	G	G	A	T	C	T
17 PGT	12	Sp	T	T	T	T	T	T	T	T	G	A	G	A	C	G	G	A	G	T	T	T
18 n/a	2	Sg	C	T	T	T	T	T	T	T	G	A	G	A	T	G	G	A	G	T	C	T
19 RES4-22	18	Sg	T	T	T	A	T	T	T	C	G	A	G	A	T	G	G	A	G	T	T	T
20 LOC51193	5	Sg	T	T	T	T	T	T	T	T	G	A	G	A	T	G	G	A	G	T	C	T
21 n/a	3	Sx	T	T	T	T	T	T	T	T	G	A	G	A	T	G	G	A	G	T	C	C
22 CHRNA3*	6	Sx	T	T	T	T	T	T	T	T	G	A	G	A	T	G	G	A	G	T	C	T
23 PTD011	2	Sx	T	T	T	T	T	T	T	T	G	A	G	A	C	G	C	C	C	A	G	G
24 HCA66	18	Sg	T	T	T	T	T	T	T	T	T	A	G	A	C	G	G	A	G	T	C	T
25 CYP3A43	8	Sg	T	T	T	T	T	T	T	T	T	A	G	A	C	A	G	A	G	T	C	T
26 LCAT*	6	Jo	T	T	T	T	T	T	T	G	T	A	G	A	G	A	C	A	G	G	G	T
27 KIAA1169	24	Jo	T	T	T	T	G	T	T	T	T	A	G	A	G	A	T	G	G	T	A	T
28 SLC3A2	6	Jb	T	T	G	T	T	T	T	T	A	A	G	A	C	A	G	C	A	T	T	T
29 ICAM2	0	Jb	T	T	T	G	T	T	T	T	A	A	G	A	C	A	G	G	G	T	C	T
30 COL4A3	6	Sx	T	T	T	T	T	C	T	T	T	A	G	A	T	G	G	A	G	T	C	T
31 GUSB	9	Sg/x	A	T	T	T	T	T	T	G	G	A	T	A	T	G	C	A	G	T	C	T
32 OAT	4	Jo	T	T	T	T	T	T	T	T	G	A	G	A	C	A	G	A	G	T	T	T

proximal AG

distal AG

Fig. 1. The selection of AGs in the 3′SSs of *Alu*-derived exons. Alignment is shown for the region near the two most prevalent 3′SSs in the right arm of exonized *Alu* sequences (in the antisense orientation). Data for 29 exonized *Alu*s, compiled from the results of our previous study (*9*) as well as newly collected data from the literature (*22*–*26*), are shown. The 20 nucleotides presented are positions 290 to 271 in the *Alu* sequence, according to the numbering in (*27*). The two possible AG dinucleotides (distal and proximal to the PPT) are marked in red. The selected AG dinucleotide, defining the end of the intron, is underlined for each exonized *Alu*. Selected AG dinucleotides were inferred with the use of alignments of expressed sequences to the human genome (*9*) (table S1). Those marked by an asterisk next to the gene name are additional *Alu* exons found in the literature scan (*22*–*26*). Consensus sequences of subfamilies S and J appear in the first two rows, with positions differing between subfamilies marked in yellow. Rows 30 to 32 represent the 3′SSs of *Alu* sequences whose constitutive exonization was shown to cause a genetic disease [Alport syndrome (COL4A3), Sly syndrome (GUSB), and OAT deficiency (OAT)]. The mutation causing Alport syndrome is marked light blue (position –7 G to T); exonization in Sly syndrome and OAT deficiency resulted from mutations in the 5′SS. Numbers on top mark the position relative to the distal 3′SS as referred to in this article. Gene names are as given in RefSeq conventions, the *Alu* exon number is the serial number of the *Alu*-containing exon in the gene, and the subfam is the *Alu* subfamily type, inferred with the use of RepeatMasker (*28*).

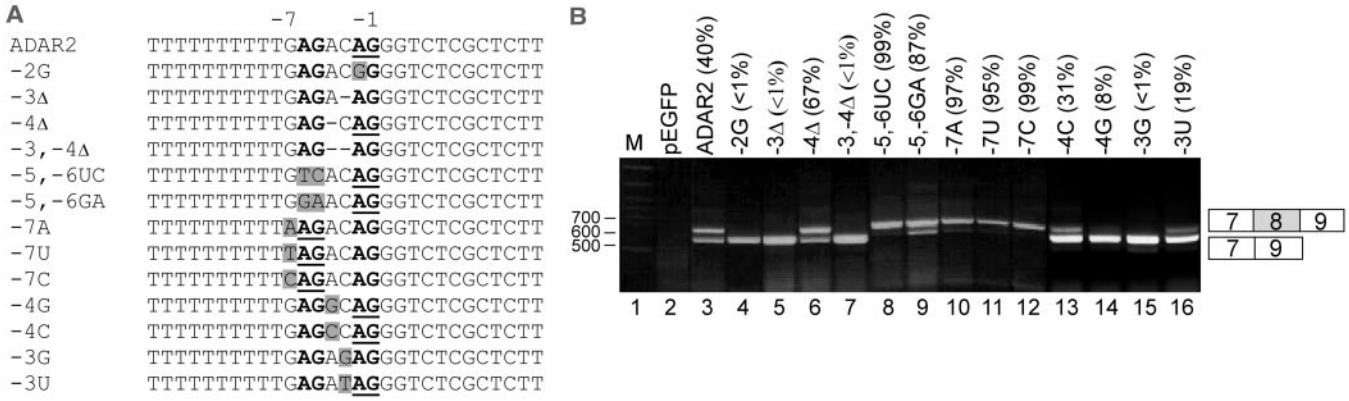


Fig. 2. Splicing assays on ADAR2 minigene mutants. (**A**) The 3′SS sequence of the wild type and 13 mutants of ADAR2. The proximal and distal AGs are in bold, mutations are shaded, and the selected AGs are boldfaced and underlined. (**B**) The indicated plasmid mutants were introduced into 293T cells by transfection, total cytoplasmic RNA was extracted, and splicing products were separated in 2% agarose gel after reverse transcription polymerase chain reaction (RT-PCR) (*18*). Lane 1, DNA size marker; lane 2, vector only (pEGFP); lane 3, splicing products of wild-type (wt) ADAR2; and lanes 4 to 16, splicing products of mutated ADAR2 minigenes, corresponding to the sequences in (A). The two possible minigene mRNA isoforms are shown on the right. Numbers in parentheses indicate percentages of the *Alu*-containing mRNA isoform as determined by quantified RT-PCR (100% corresponds to the total of both mRNA isoforms). Identical results were also obtained with HeLa cells.

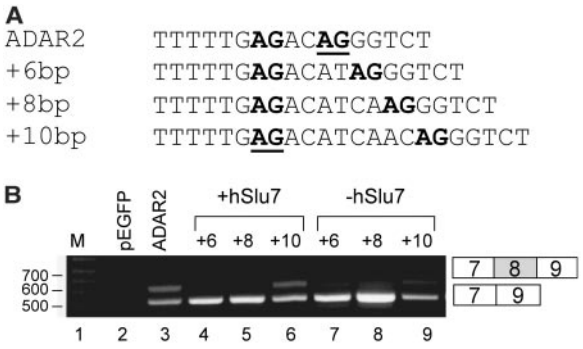


Fig. 3. The effect of hSlu7 on AG selection. (**A**) The sequence of the 3′SS of ADAR2 and the three insertion mutants. Both potential AGs are marked in bold, and the selected AG is boldfaced and underlined. bp, base-pair (**B**) The indicated plasmid mutants were treated as described for Fig. 2B (*18*). Lanes 4 to 6 represent cotransfection of the insertion mutants with a plasmid expressing hSlu7; lanes 7 to 9 represent the insertion mutants without additional hSlu7.

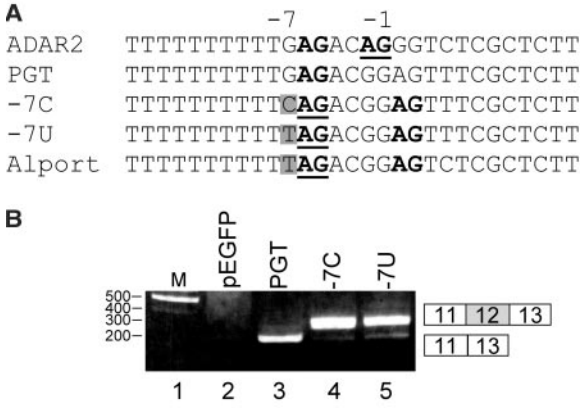


Fig. 4. Splicing assays on wt and mutated PGT. (**A**) Sequences from top to bottom are as follows: wt *Alu* 3′SS of ADAR2; wt *Alu* 3′SS of PGT; mutant *Alu* 3′SS of PGT; and mutant sequence of COL4A3, which causes Alport syndrome. Both potential AGs are marked in bold, and the selected 3′SS is boldfaced and underlined. The mutated position is shaded. (**B**) Transfection was performed in HT1080 cell lines. Total RNA and RT-PCR was performed as mentioned in Fig. 2B (*18*). Lane 1, DNA size marker; lane 2, vector only (pEGFP); lane 3, splicing product of wt PGT; lanes 4 and 5, splicing products of mutated PGT minigenes, corresponding to the sequences in (A). The two possible minigene mRNA isoforms are shown on the right. The results were reproducible in 293T cell lines (*19*) as well.

lane 9). In these transcripts, the proximal AG was selected even though it was preceded by G (Fig. 3A).

We further examined whether hSlu7 (human synergistic lethal with U5 small nuclear RNA), a second-step splicing factor, might be involved in the activation of the proximal AG. This protein is known to be required for correct AG identification when more than one possible AG exists in the 3′SS region (*13*). Cotransfection of 293T cells with plasmids containing the insertion mutants and with hSlu7 (10-fold higher than endogenous hSlu7 concentrations) (*18*) led to an increase in the selection of the proximal AG by 10-fold, reaching 32% inclusion when the distance between the proximal and distal AGs was 10 bases (Fig. 3B, lane 6). Presumably, hSlu7 activation of the weak splice site may depend on the existence of a distal AG, because elimination of the distal AG (mutant –2G, Fig. 2) resulted in an exon skipping that was not reversed by increasing the concentration of hSlu7 (*19*). These results propose that the distal AG can affect the selection of the proximal one negatively when the proximal is preceded by a G nucleotide. The proximal 3′SS can be selected when hSlu7 is present, and the efficiency of this selection is increased when the distal AG is found far enough from the splice site (in our case, 10 nucleotides into the exon). This observation, therefore, indicates that activation of the weak 3′SS (GAG) depends on hSlu7 concentration and suggests a possible role for hSlu7 concentration in alternative-splicing regulation.

Rows 17 to 22 in Fig. 1 show instances in which the proximal AG is selected even though the distal AG is found six nucleotides downstream. However, the +6 base-pair mutant (Fig. 3B, lane 6) resulted in a total exon skipping. The above results suggest that these exonization instances might occur with high hSlu7 concentrations within certain cell types or with high local concentration of hSlu7 within the subregion of the nucleus. From this, we further assumed that in normal conditions these *Alu* exons would be skipped. We therefore chose one of these genes, one encoding a putative glucosyltransferase (PGT) (Fig. 1, row 17), and cloned a minigene of its exons 11 to 13, including the introns in between (the *Alu* exon being exon 12). Indeed, when the PGT minigene was transfected into HT1080 and 293T cell lines, only a single mRNA isoform appeared, corresponding to *Alu*-exon skipping (Fig. 4B, lane 3). Repeating the same experiment with the use of endogenous PGT mRNA also showed *Alu*-exon skipping (*19*).

To test if, as predicted from our results, a mutation in position –7 of a completely silent intronic *Alu* element would result in exoniza-

tion, we mutated this position in the PGT minigene. As seen in Fig. 4B (lanes 4 and 5), this point mutation was enough to activate the nearly constitutive inclusion of the *Alu* exon in the mature transcript. As indicated above, the same mutation in the COL4A3 gene activates a constitutive exonization of a silent intronic *Alu*, resulting in Alport syndrome (*10*). To assess the importance of our findings, we analyzed the entire content of *Alu*s in the human genome and found that there are at least 238,000 antisense *Alu*s located within introns in the human genome (*20*). Of these, 52,935 *Alu*s carry a potential ADAR2-like 3′SS, and 23,012 carry a potential PGT-like 3′SS. Our results suggest that many of these silent intronic *Alu* elements might be susceptible to exonization by the same single point mutation and are thus under strict selective pressure. Such point mutations in human genomic antisense *Alu*s may, therefore, be the molecular basis for predisposition to so-far uncharacterized genetic diseases.

Because all *Alu*-containing exons are alternatively spliced (*9*), they add splice variants to our transcriptome while maintaining the original proteins intact. Exonized *Alu*s can, thus, acquire functionality and become exapted, i.e., adapted to a function different than their original (*21*). When the splicing of an *Alu* exon is constitutive, however, the transcript encoding to the original protein is permanently disrupted, which could provide the basis for a genetic disorder. Identification of genomic *Alu*s that are one point mutation away from exonization might therefore enable the screening for predisposition for genetic diseases that involve *Alu* exonization.

References and Notes

1. A. J. Mighell, A. F. Markham, P. A. Robinson, *FEBS Lett.* **417**, 1 (1997).
2. D. J. Rowold, R. J. Herrera, *Genetica* **108**, 57 (2000).
3. C. W. Schmid, *Prog. Nucleic Acid Res. Mol. Biol.* **53**, 283 (1996).
4. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
5. A. M. Roy-Engel *et al.*, *Genome Res.* **12**, 1333 (2002).
6. W. Makalowski, G. A. Mitchell, D. Labuda, *Trends Genet.* **10**, 188 (1994).
7. W. Makalowski, *Gene* **259**, 61 (2000).
8. A. Nekrutenko, W. H. Li, *Trends Genet.* **17**, 619 (2001).
9. R. Sorek, G. Ast, D. Graur, *Genome Res.* **12**, 1060 (2002).
10. B. Knebelmann *et al.*, *Hum. Mol. Genet.* **4**, 675 (1995).
11. G. A. Mitchell *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 815 (1991).
12. R. Vervoort, R. Gitzelmann, W. Lissens, I. Liebaers, *Hum. Genet.* **103**, 686 (1998).
13. K. Chua, R. Reed, *Nature* **402**, 207 (1999).
14. K. Chua, R. Reed, *Mol. Cell. Biol.* **21**, 1509 (2001).
15. R. E. Manrow, S. L. Berger, *J. Mol. Biol.* **234**, 281 (1993).
16. D. Slavov, K. Gardiner, *Gene* **299**, 83 (2002).
17. F. Lai, C. X. Chen, K. C. Carter, K. Nishikura, *Mol. Cell. Biol.* **17**, 2413 (1997).
18. Materials and methods are available as supporting material on *Science* Online.
19. G. Lev-Maor, R. Sorek, N. Shomron, G. Ast, unpublished data.
20. R. Sorek, R. Shalgi, G. Ast, D. Graur, unpublished data.
21. J. Brosius, S. J. Gould, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10706 (1992).
22. M. Mihovilovic *et al.*, *Biochem. Biophys. Res. Commun.* **197**, 137 (1993).
23. M. Miller, K. Zeller, *Gene* **190**, 309 (1997).
24. W. G. Cance, R. J. Craven, T. M. Weiner, E. T. Liu, *Int. J. Cancer* **54**, 571 (1993).
25. I. W. Caras *et al.*, *Nature* **325**, 545 (1987).
26. G. Svineng, R. Fassler, S. Johansson, *Biochem. J.* **330**, 1255 (1998).
27. J. Jurka, A. Milosavljevic, *J. Mol. Evol.* **32**, 105 (1991).
28. RepeatMasker is available online at http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker.

29. We thank M. Kupiec for a critical reading; R. Reed for the hSlu7 plasmid; and also F. Belinky, R. Shalgi, T. Dagan, and E. Sharon for assistance in *Alu* data analysis. Supported by a grant from the Israel Science Foundation and, in part, by a grant from the Israel Cancer Association and the Indian-Israeli Scientific Research Corporation to G.A.

Supporting Online Material

www.sciencemag.org/cgi/content/full/300/5623/1288/DC1
Materials and Methods
Fig. S1
Table S1
References and Notes

21 January 2003; accepted 9 April 2003

Essential Role of Fkbp6 in Male Fertility and Homologous Chromosome Pairing in Meiosis

Michael A. Crackower,[1]*† Nadine K. Kolas,[2]* Junko Noguchi,[4] Renu Sarao,[1] Kazuhiro Kikuchi,[4] Hiroyuki Kaneko,[4] Eiji Kobayashi,[5] Yasuhiro Kawai,[6] Ivona Kozieradzki,[1] Rushin Landers,[1] Rong Mo,[7] Chi-Chung Hui,[7] Edward Nieves,[3] Paula E. Cohen,[2] Lucy R. Osborne,[8] Teiji Wada,[1] Tetsuo Kunieda,[6] Peter B. Moens,[9] Josef M. Penninger[1]‡

Meiosis is a critical stage of gametogenesis in which alignment and synapsis of chromosomal pairs occur, allowing for the recombination of maternal and paternal genomes. Here we show that FK506 binding protein (Fkbp6) localizes to meiotic chromosome cores and regions of homologous chromosome synapsis. Targeted inactivation of *Fkbp6* in mice results in aspermic males and the absence of normal pachytene spermatocytes. Moreover, we identified the deletion of *Fkbp6* exon 8 as the causative mutation in spontaneously male sterile *as/as* mutant rats. Loss of Fkbp6 results in abnormal pairing and misalignments between homologous chromosomes, nonhomologous partner switches, and autosynapsis of X chromosome cores in meiotic spermatocytes. Fertility and meiosis are normal in *Fkbp6* mutant females. Thus, Fkbp6 is a component of the synaptonemal complex essential for sex-specific fertility and for the fidelity of homologous chromosome pairing in meiosis.

Meiosis is a fundamental process in sexually reproducing species that allows genetic exchange between maternal and paternal genomes (*1*, *2*). Defects in high-fidelity meiotic chromosome alignment or in genome segregation in germ cells result in aneuploidies such as trisomy 21 in Down syndrome. Aneuploidy is a leading cause of spontaneous miscarriage in humans and a hallmark of many human cancer cells (*2*). Once homologs are paired, the chromosomes are connected by a specific structure: the synaptonemal complex (SC) (*3*). SCs are zipperlike structures assembled along the paired meiotic chromosomes during the prophase of the first meiotic division (*3*). Although SCs were first discovered more than 45 years ago (*4*, *5*), only very few structural meiosis-specific components of the SC have been identified in mammals, such as SC proteins 1, 2, and 3 [Scp1 (also known as Syn1/Sycp1), Scp2, and Scp3 (also known as Cor1)] (*3*). Genetic inactivation of the mouse *Scp3* gene results in male infertility due to a failure to form chromosome synapsis in meiotic prophase (*6*). Female *Scp3*$^{-/-}$ mice have reduced fertility, and embryos from *Scp3*$^{-/-}$ mothers have increased incidents of aneuploidy (*7*). To our

[1]Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), c/o Dr. Bohrgasse 7, 1030, Vienna, Austria. [2]Department of Molecular Genetics, and [3]Department of Biochemistry, Laboratory for Macromolecular Analysis and Proteomics, Albert Einstein College of Medicine (AECOM), 1300 Morris Park Avenue, Bronx, NY 10461, USA. [4]Germ Cell Conservation Laboratory, National Institute of Agrobiological Sciences, Kannondai, Tsukuba, Ibaraki 305-8602, Japan. [5]National Livestock Breeding Center, Odakura, Nishigo, Fukushima 961-851, Japan. [6]Graduate School of Natural Science and Technology, Okayama University, Okayama 700-0082 Japan. [7]Program in Developmental Biology, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada. [8]Departments of Medicine and Molecular and Medical Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada. [9]Department of Biology, York University, Toronto, ON M3J 1P3, Canada.

*These authors contributed equally to this work.
†Present address: Department of Protein Sciences, Amgen, Thousand Oaks, CA 91320, USA.
‡To whom correspondence should be addressed. E-mail: josef.penninger@oeaw.ac.at

Molecular Cell, Vol. 14, 221–231, April 23, 2004, Copyright ©2004 by Cell Press

Minimal Conditions for Exonization of Intronic Sequences: 5′ Splice Site Formation in *Alu* Exons

Rotem Sorek,[1,2,5] Galit Lev-Maor,[1,5] Mika Reznik,[1,5] Tal Dagan,[3] Frida Belinky,[3] Dan Graur,[4] and Gil Ast[1,*]

[1]Department of Human Genetics and Molecular Medicine
Sackler Faculty of Medicine
Tel Aviv University
Ramat Aviv 69978
Israel

[2]Compugen
72 Pinchas Rosen Street
Tel Aviv 69512
Israel

[3]Department of Zoology
George S. Wise Faculty of Life Sciences
Tel Aviv University
Ramat Aviv 69978
Israel

[4]Department of Biology and Biochemistry
University of Houston
Houston, Texas 77204

Summary

***Alu* exonization, which is an evolutionary pathway that creates primate-specific transcriptomic diversity, is a powerful tool for studying alternative-splicing regulation. Through bioinformatic analyses combined with experimental methodology, we identified the mutational changes needed to create functional 5′ splice sites in *Alu*. We revealed a complex mechanism by which the sequence composition of the 5′ splice site and its base pairing with the small nuclear RNA U1 govern alternative splicing. We show that in *Alu*-derived GC introns the strength of the base pairing between U1 snRNA and the 5′ splice site controls the skipping/inclusion ratio of alternative splicing. Based on these findings, we identified 7810 *Alu*s within the human genome that are prone to exonization. Mutations in these *Alu*s may cause genetic disorders or contribute to human-specific protein diversity.**

Introduction

The availability of the complete human genome sequence has made it clear that gene number is not the sole determinant of proteome complexity. Sequencing of the human genome showed that humans possess only ~26,000 protein coding genes, which is only slightly larger than the number of genes in *C. elegans*. This surprisingly low number contrasts with the number of human proteins, estimated to be more than 90,000 (Harrison et al., 2002; Lander et al., 2001).

By producing more than one type of mRNA from a single gene, alternative splicing is a significant contributor to proteome diversification (Brett et al., 2002; Graveley, 2001). Bioinformatic analyses indicate that 35%–74% of all human genes participate in alternative splicing, which contributes significantly to human proteome complexity and explains the numerical disparity between genes and proteins (Black, 2000; Johnson et al., 2003; Modrek and Lee, 2002). Alternative splicing is often regulated according to cell type, developmental stage, sex, or in response to an external stimulus (Cartegni et al., 2002; Stoilov et al., 2002). Aberrant regulation of alternative splicing has been implicated in an increasing number of human diseases, including cancer (Hastings and Krainer, 2001; Modrek and Lee, 2002; Nissim-Rafinia and Kerem, 2002; Stoilov et al., 2002; Xu and Lee, 2003).

Approximately 75–130 million years may have passed since the human and mouse common ancestor speciated into two separate lineages (Waterston et al., 2002; Yang and Yoder, 1999). Most of the genes (99%) are orthologous, and the majority of these genes (86%) share the same intron/exon arrangement as well as a high degree of conservation (88%) in homologous exon sequences (Waterston et al., 2002). If most of the genes are highly conserved between human and mouse, what are the genomic elements that contribute to some of the unique features of humans and primates? Human-mouse comparative analysis revealed that alternative splicing is often associated with recent exon creation and/or loss (Modrek and Lee, 2003). Thus, alternative splicing has the potential of creating species-specific cassette exons.

Retrotransposons are short sequences of DNA that produce new copies of themselves by the reverse transcription of an RNA intermediate. These mobile DNA elements have had a profound influence in shaping eukaryotic genomes. As much as 46% of the human genome is made up of transposable elements, the most abundant of which is a primate-specific dimeric retrotransposon called *Alu*. The human genome contains approximately 1.4 million *Alu* copies, all derived from a single 7SL RNA-specifying gene (Brosius, 1999; Kazazian, 2000; Lander et al., 2001). *Alu* elements, which are 300 nucleotides long, currently amplify at a rate of about one insertion in every 100–200 new live births (Dewannieux et al., 2003). The transcription of *Alu* elements is usually suppressed by a large number of hypermethylated CG dinucleotides, which block a bipartite RNA polymerase III promoter (Deininger and Batzer, 2002; Makalowski et al., 1994).

We have recently shown that more than 5% of the alternatively spliced internal exons in the human genome are derived from *Alu* (Sorek et al., 2002; reviewed in Kreahling and Graveley, 2004). As far as we know, all alternatively spliced *Alu* exons (*AEx*) were created exclusively via the exonization of intronic elements. These *AEx*s enrich the transcriptome and enhance the coding capacity and regulatory versatility of primate genomes with new isoforms without compromising the integrity and original repertoire of the transcriptome and its resulting proteome. In contrast, each newly created

*Correspondence: gilast@post.tau.ac.il
[5]These authors contributed equally to this work.

constitutively spliced *AEx* will generate a new product at the expense of the original product, and such a loss may be deleterious (Sorek et al., 2002; Lev-Maor et al., 2003).

Through molecular evolutionary methodology, it was possible to align each *AEx* to its inferred ancestral sequence. We could, therefore, identify the changes in the *Alu* sequences that were most probably responsible for the exonization. This methodology has recently allowed us to identify the delicate interplay between two AG dinucleotides that maintain a weak 3′ splice site (3′ss) responsible for alternative splicing (Lev-Maor et al., 2003). We have further demonstrated that an activation of an intronic *Alu* sequence (silent *Alu* sequence never spliced in as an exon) can be induced either by point mutations in certain positions along the *Alu* sequence or by changing the concentration of the splicing regulatory proteins in the cell (Lev-Maor et al., 2003). Thus, *Alu* exonization is an evolutionary pathway that creates primate-specific genomic diversity. Here, we reveal the mutational steps that create 5′ splice site (5′ss) in alternatively spliced *AEx*s and examine how the spliceosome selects this 5′ss.

Results

We compiled a data set of exonized *Alu*s in which the prevalent 5′ss of these exons is selected (Figure 1, *Alu* position 157 being the first position of the intron) and compared their adjacent regions with homologous positions in the *Alu* antisense consensus sequence (Jurka and Milosavljevic, 1991). This allowed us to identify nucleotide positions along the *Alu* sequence that need to be changed or remain conserved in order to enable *Alu* exonization. Two types of 5′ss were found: introns that start with GC (Figure 1, cases 1–4) and introns that start with GT (Figure 1, cases 5–25). In the human genome, more than 98% of all introns begin with GT (Farrer et al., 2002; Lander et al., 2001; Thanaraj and Clark, 2001). The minority GC introns (0.7% of all introns) were claimed to be frequently involved in alternative splicing (Thanaraj and Clark, 2001).

The most significant change observed between exonized *Alu*s and their ancestral sequences was at position 156 (position two of the intron), where a mutation from C to T generates a canonical GT 5′ss at positions 157-156. This occurred in 21 of the 25 (84%) exonized *Alu*s. In those cases in which positions 157-156 remained GC as in the ancestral sequence (rows 1–4), position 155 (representing position 3 of the intron) was found to be mutated from G to A.

In the ancestral sequence, CG dinucleotides are found in positions 156-155, 154-153, and 152-151 (Figure 1, upper row). Since CG dinucleotides in *Alu* are frequently hypermethylated (Kunkel and Diaz, 2002) and mutate 9.2 times more frequently than non-CG positions (Batzer et al., 1990; Kunkel and Diaz, 2002), one may attribute the formation of the GT 5′ss to the propensity of these sites to mutate from CG to TG. The prevalent mutations in positions 156, 154, and 152 are from C to T (yellow and blue), whereas in positions 155 and 151 the prevalent mutations are from G to A (green).

Some of the changes observed in Figure 1 seem to be random (purple); others are presumably due to CG substitutions (yellow, blue, and green). However, it is unclear which of these changes are important to the exonization of *Alu*s and which represent inconsequential intronic substitutions. To pinpoint the positions that are important for exonization, we compiled a data set of 166,276 full-length intronic *Alu*s that are found in the antisense orientation in introns of human genes and compared them to exonized *Alu*s (Dagan et al., 2004) (see also Experimental Procedures). By aligning each of these *Alu*s to its ancestral sequence, we examined the percentage of each of the four nucleotides in each position along the *Alu* sequence (Figure 2A; positions 172 to 146 of the antisense *Alu* consensus sequence are shown). We similarly examined the 25 *AEx*s shown in Figure 1 (Figure 2B). Finally, we compared the nucleotide distribution for each position of the intronic *Alu*s to those of the *AEx*s and looked for statistically significant deviations between the distributions (Figure 2B).

Remarkably, in only two positions along the entire exonized-*Alu*s sequence did we find a distribution significantly different from that in the homologous positions of the intronic *Alu*s. In position 156 (position two of the intron) C changes predominantly to T to create the aforementioned GT 5′ss; in position 153 (position 5 of the intron) G is conserved over the expected frequency (G was found in 18 instead of 10.9 expected sequences). Since sense- and antisense-oriented *Alu*s can be under different selective pressures in the human genome, we also compiled a data set of 136,151 intronic *Alu*s that are found in the sense orientation in introns and repeated the statistical comparison to exonized *Alu*s. This yielded the same results, indicating positions two and five of the intron as significantly different between exonized and nonexonized *Alu*s (Figures 2C and 2D). This indicates that these are the most important positions in the creation of functional *Alu*-derived 5′ss. Therefore, *Alu*s in the antisense orientation in introns are in fact "preexons" whose exonization requires only a small number of mutations.

Significantly, the same mutation, from C to T at position 156, of an antisense *Alu* sequence in intron six of the CTDP1 gene was found to be the cause of CCFDN (congenital cataracts, facial dysmorphism, and neuropathy) syndrome (Figure 1, row 26; Varon et al., 2003). In this gene, the G in position 153 is indeed conserved as predicted by our analysis. Previously, only mutations leading to the constitutive exonization of *Alu* elements were known to be deleterious, e.g., in Alport-syndrome type I, Sly syndrome, and ornithine aminotransferase (OAT) deficiency (Vervoort et al., 1998; Mitchell et al., 1991; Knebelmann et al., 1995). CCFDN syndrome was the first reported case in which a mutation leading to the creation of an alternatively spliced *AEx* resulted in a genetic disease. Thus, the appearance of an aberrant *Alu*-containing spliced form may result in genetic disease even when the normal mRNA continues to be synthesized.

Following these results we wished to understand how the alternative splicing of these exonized *Alu*s is regulated. In mRNA splicing, the 5′ss is recognized by three small nuclear RNPs (complexes of snRNA and proteins), one of which (U1) base pairs across the 5′ss junction (potential base pairing between positions −3 to +6;

				exon			intron																				
				Position relative to 5'ss			-3	-2	-1	1+	2+	3+	4+	5+	6+	7+											
				172	171	170	169	168	167	166	165	164	163	162	161	160	159	158	157	156	155	154	153	152	151	150	149	148	147	146
	Gene name	Alu exon#	subfamily	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	C	G	C	G	C	G	C	C	A	C	C
1	MLANA	4	J	T	A	G	C	T	G	G	G	A	C	C	A	C	A	G	G	C	A	C	G	T	G	C	C	A	C	C
2	RES4-22	18	S	T	A	G	T	T	G	G	A	A	T	T	A	C	A	G	G	C	A	A	G	C	A	A	A	A	C	C
3	HPK1*	31	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	C	A	T	G	T	G	C	C	A	C	C
4	ADAR2	8	J	G	A	G	C	T	G	G	G	A	C	T	A	C	A	G	G	C	A	T	G	T	A	C	C	A	C	T
5	KIAA1169	24	J	C	A	G	C	T	G	G	A	A	C	T	A	C	A	G	G	T	G	T	G	C	A	C	T	G	T	G
6	LOC51193	5	S	T	A	G	C	T	G	G	G	G	T	T	A	C	A	G	G	T	G	T	G	C	G	C	C	A	C	C
7	ZFX	2	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	T	G	C	G	C	C	C	C	C
8	MVK	4	J	T	A	G	C	T	G	G	G	A	C	C	A	C	A	G	G	T	G	T	G	C	G	C	C	A	C	C
9	PTGES	2	J	T	A	G	C	T	G	G	G	A	C	C	A	C	A	G	G	T	G	T	G	T	A	T	C	A	C	C
10	N/A	6	J	T	A	G	C	T	G	G	G	A	C	T	A	C	A	G	G	T	G	T	G	T	G	C	C	A	C	C
11	EVI5	3	J	T	G	G	C	T	G	G	G	A	C	T	A	C	A	G	G	T	G	T	G	T	G	C	C	A	T	C
12	C20orf26	9	J	T	A	G	C	T	G	.	G	A	C	T	A	T	A	G	G	T	A	T	G	T	G	C	C	A	C	C
13	N/A	4	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	A	T	G	T	G	C	C	A	C	C
14	BRCA2	20	S	T	A	G	C	T	G	G	G	A	C	T	A	C	A	G	G	T	G	C	G	T	G	C	C	A	C	C
15	CNN2	6	S	T	A	G	C	T	G	G	G	A	C	T	A	C	A	G	G	T	G	C	A	T	G	C	T	G	C	C
16	BIRC3	2	S	T	A	G	C	T	G	G	A	A	A	T	A	C	A	G	G	T	G	C	G	T	G	C	C	A	C	C
17	CYP3A43	8	S	G	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	A	C	A	C	A	C	C	A	C	C
18	N/A	2	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	C	C	C	G	C	C	A	C	C
19	MBD3	12	S	T	A	G	C	T	G	G	G	A	T	T	T	C	A	G	G	T	A	C	C	C	G	T	C	A	C	A
20	PLA2G4B	2	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	C	C	C	A	C	C	A	C	C
21	BCAS4	5	S	T	A	G	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	C	G	C	G	C	T	A	C	C
22	ICAM2	2	J	T	A	G	C	T	G	G	G	A	T	C	A	C	A	G	G	T	G	A	G	A	G	C	C	A	C	C
23	TGM4*	2	FAM	T	A	A	C	C	G	G	G	A	T	T	A	C	A	G	G	T	A	T	G	T	G	A	C	T	C	C
24	Integrin β1*	7	S	T	A	C	C	T	G	G	G	A	T	T	A	C	A	G	G	T	G	C	C	T	G	C	C	T	C	A
25	CHRNA3*	5	S	T	G	T	C	T	G	G	G	A	C	T	A	C	A	G	G	T	A	C	C	C	G	C	C	C	G	C
26	CTDP1*	7	Y	G	T	G	T	T	G	G	G	A	T	T	A	C	A	G	G	T	A	T	G	A	G	C	C	A	T	T

Figure 1. Selection of 5′ss of *Alu*-Derived Exons

Alignment is shown for the region near the most prevalent 5′ss on the right arm of exonized *Alu* sequences (in the antisense orientation). Data for 26 exonized *Alu*s (Dubbink et al., 1998; Hu et al., 1996; Mihovilovic et al., 1993; Miller and Zeller, 1997; Sorek et al., 2002; Svineng et al., 1998) are shown. The 27 nucleotides spanning positions 146–172 according to the numbering in Jurka and Milosavljevic (1991) are shown. The dinucleotides GT or GC that are selected as the 5′ss (defining the beginning of the intron) are in red. The 5′ss were inferred by alignment of expressed sequences to the human genome (Sorek et al., 2002) (Supplemental Table S1 available on *Molecular Cell*'s website). The *Alu* consensus sequence appears in the first row; the position differing between *Alu* subfamilies S and J is marked in gray. Nucleotides that differ from the *Alu* consensus sequence are marked in purple; mutations that changed the dinucleotide CG to TG are marked in yellow; those for which this mutation creates a splice site are marked in blue; mutations that changed CG to CA are marked in green. Row 26 represents the 5′ss of an antisense *Alu* sequence in intron six of CTDP1 gene in which a mutation from C to T at position 156 resulted in CCFDN syndrome (Varon et al., 2003). This mutation (marked red) led to the exonization of an intronic *Alu* sequence by activation of the 5′ss and the creation of an alternatively spliced *AEx*. Numbers on the top mark positions relative to the 5′ss. Gene names are as in RefSeq convention. The *AEx* number is the serial number of the *Alu*-containing exon in the related gene, and the *Alu* subfamily type was inferred with the use of RepeatMasker (http://www.repeatmasker.org/).

for clarity, "−" and "+" indicate positions upstream or downstream of the 5′ss, respectively). This base pairing is a prerequisite step for splicing in most introns (Brow, 2002). Although the importance of the U1 snRNA:5′ss base pairing is well established in constitutive splicing, the function of this base pairing in alternative splicing is only partially understood (Cohen et al., 1993). We, therefore, set to understand the manner in which U1 affects the alternative splicing of *AEx*s.

We used a minigene containing the genomic sequence of adenosine deaminase gene, ADAR2, from exon seven to nine, in which exon eight is an alternatively spliced *AEx* (Lev-Maor et al., 2003). The 5′ss of this *AEx* is of the GC type (Figure 1, row 4). To examine the function of the base pairing between U1 and the 5′ss in alternative splicing, we transfected 293T cells with the ADAR2 minigene containing mutations in the 5′ss and complemented these mutations with cotrasfected U1 gene containing the appropriate compensatory mutation. Therefore, the cells contained exogenous and endogenous U1s that competed with each other to bind the 5′ss (Zhuang and Weiner, 1986). Following transfection, cytoplasmic RNA was collected, and the splicing pattern of the ADAR2 minigene was examined by RT-PCR (see Experimental Procedures). We tested the effect of serial mutations on the splicing of the ADAR2 minigene when the first nucleotides of intron eight are GC (Figure 3A, representing GC 5′ss) or when the C is mutated to T creating GT (Figure 3B, representing GT 5′ss).

Our results indicate that the GC 5′ss maintains alternative rather than constitutive splicing directly because the C in position two of the 5′ss unpairs with U1. This conclusion is supported by the fact that a compensatory mutation in U1 (A to G in position seven), which restores the base pairing with position two, results in the constitutive splicing of the exon (Figure 3A, lanes 1–3; see Figure 3C for wild-type [wt] U1:5′ss base pairing).

As expected from our bioinformatic analysis (Figure 1, rows 1–4), when the 5′ss is of the GC type, an A in position three is essential for its proper selection; mutations to C, T, or G led to *AEx* skipping (Figure 3A, lanes 5, 6, and 8). This finding is in agreement with the weight metric of GC 5′ss where A is the prevalent nucleotide in position three (Thanaraj and Clark, 2001) and may suggest that an A at position three of the GC intron that forms A:Ψ pairing (Ψ, pseudo-uridine) with U1 is required to avoid two consecutive positions that

A

Position	Alu consensus	Genomic, non-exonized Alus in introns: Count per position A	G	T	C	Percent per position A	G	T	C
172	T	2517	1613	155796	4583	1.5%	1.0%	94.7%	2.8%
171	A	158590	3870	1515	649	96.3%	2.4%	0.9%	0.4%
170	G	5360	153227	2419	3471	3.3%	93.2%	1.5%	2.1%
169	C	1171	2025	6161	155132	0.7%	1.2%	3.7%	94.3%
168	T	1354	1734	159323	2179	0.8%	1.1%	96.8%	1.3%
167	G	9873	152484	1317	630	6.0%	92.8%	0.8%	0.4%
166	G	9879	151737	1819	1122	6.0%	92.2%	1.1%	0.7%
165	G	9679	151024	2073	1426	5.9%	92.0%	1.3%	0.9%
164	A	157893	3364	2273	890	96.0%	2.0%	1.4%	0.5%
163	T	1627	1465	82293	78985	1.0%	0.9%	50.1%	48.1%
162	T	2408	1151	150771	9988	1.5%	0.7%	91.8%	6.1%
161	A	155739	5616	1169	1817	94.8%	3.4%	0.7%	1.1%
-3 160	C	2274	1642	14755	145456	1.4%	1.0%	9.0%	88.6%
-2 159	A	158275	3715	1356	958	96.3%	2.3%	0.8%	0.6%
-1 158	G	6277	155408	1463	877	3.8%	94.7%	0.9%	0.5%
1 157	G	8282	150508	2520	2201	5.1%	92.0%	1.5%	1.3%
2 156	C	3200	2603	49508	108038	2.0%	1.6%	30.3%	66.1%
3 155	G	70291	86492	3762	3155	42.9%	52.8%	2.3%	1.9%
4 154	C	2302	2337	61547	97440	1.4%	1.4%	37.6%	59.6%
5 153	G	32027	71147	4427	55774	19.6%	43.5%	2.7%	34.1%
6 152	C	4294	2941	66256	89861	2.6%	1.8%	40.6%	55.0%
7 151	G	55629	102805	1849	2736	34.1%	63.1%	1.1%	1.7%
150	C	2039	1551	7413	152348	1.2%	0.9%	4.5%	93.3%
149	C	2779	860	9448	150848	1.7%	0.5%	5.8%	92.0%
148	A	156184	4269	1394	2387	95.1%	2.6%	0.8%	1.5%
147	C	2190	2055	8776	151244	1.3%	1.3%	5.3%	92.1%
146	C	3526	1782	11503	147531	2.1%	1.1%	7.0%	89.8%

B

Position	Alu consensus	Exonized Alus: Observed A	G	T	C	Expected A	G	T	C	E value
172	T	0	2	22	1	0.4	0.2	23.7	0.7	
171	A	23	2	0	0	24.1	0.6	0.2	0.1	
170	G	1	22	1	1	0.8	23.3	0.4	0.5	
169	C	0	0	1	24	0.2	0.3	0.9	23.6	
168	T	0	0	24	1	0.2	0.3	24.2	0.3	
167	G	0	25	0	0	1.5	23.2	0.2	0.1	
166	G	0	24	0	0	1.4	22.1	0.3	0.2	
165	G	3	22	0	0	1.5	23.0	0.3	0.2	
164	A	24	1	0	0	24.0	0.5	0.3	0.1	
163	T	1	0	13	11	0.2	0.2	12.5	12.0	
162	T	0	0	21	4	0.4	0.2	22.9	1.5	
161	A	24	0	1	0	23.7	0.9	0.2	0.3	
-3 160	C	0	0	1	24	0.3	0.3	2.2	22.2	
-2 159	A	25	0	0	0	24.1	0.6	0.2	0.1	
-1 158	G	0	25	0	0	1.0	23.7	0.2	0.1	
1 157	G	0	25	0	0	1.3	23.0	0.4	0.3	
2 156	C	0	0	21	4	0.5	0.4	7.6	16.5	1.82E-07
3 155	G	10	15	0	0	10.7	13.2	0.6	0.5	
4 154	C	2	0	12	11	0.4	0.4	9.4	14.9	
5 153	G	2	18	0	5	4.9	10.9	0.7	8.5	3.66E-02
6 152	C	1	0	13	11	0.7	0.5	10.1	13.8	
7 151	G	6	19	0	0	8.5	15.8	0.3	0.4	
150	C	2	0	2	21	0.3	0.2	1.1	23.3	
149	C	1	0	3	21	0.4	0.1	1.4	23.0	
148	A	19	2	2	2	23.8	0.6	0.2	0.4	
147	C	0	1	2	22	0.3	0.3	1.3	23.0	
146	C	1	1	1	22	0.5	0.3	1.7	22.4	

C

Position	Alu consensus	Genomic, non-exonized Alus in introns: Count per position A	G	T	C	Percent per position A	G	T	C
172	T	2330	1806	124457	6047	1.7%	1.3%	92.4%	4.5%
171	A	129427	3570	1175	523	96.1%	2.7%	0.9%	0.4%
170	G	4257	125856	1623	2867	3.2%	93.5%	1.2%	2.1%
169	C	1219	1859	5021	126540	0.9%	1.4%	3.7%	94.0%
168	T	1175	1451	130082	1973	0.9%	1.1%	96.6%	1.5%
167	G	9088	123515	1060	649	6.8%	92.0%	0.8%	0.5%
166	G	8556	123319	1471	1253	6.4%	91.6%	1.1%	0.9%
165	G	8357	123186	1428	1347	6.2%	91.7%	1.1%	1.0%
164	A	129679	2432	1687	682	96.4%	1.8%	1.3%	0.5%
163	T	1620	1149	65209	66505	1.2%	0.9%	48.5%	49.5%
162	T	2192	1045	122476	8682	1.6%	0.8%	91.1%	6.5%
161	A	129406	2789	846	1436	96.2%	2.1%	0.6%	1.1%
-3 160	C	2130	1538	11200	119468	1.6%	1.1%	8.3%	88.9%
-2 159	A	129674	2864	1156	791	96.4%	2.1%	0.9%	0.6%
-1 158	G	5837	126192	1271	930	4.3%	94.0%	0.9%	0.7%
1 157	G	7014	123257	1831	1774	5.2%	92.1%	1.4%	1.3%
2 156	C	3451	2278	41202	86796	2.6%	1.7%	30.8%	64.9%
3 155	G	59038	70000	2515	2396	44.1%	52.3%	1.9%	1.8%
4 154	C	2395	1961	49199	80356	1.8%	1.5%	36.7%	60.0%
5 153	G	27422	58181	3225	44883	20.5%	43.5%	2.4%	33.6%
6 152	C	4119	2361	52120	75135	3.1%	1.8%	39.0%	56.2%
7 151	G	47023	82898	1313	2121	35.3%	62.2%	1.0%	1.6%
150	C	2411	1408	6842	123034	1.8%	1.1%	5.1%	92.0%
149	C	2555	821	7406	123412	1.9%	0.6%	5.5%	92.0%
148	A	128819	2953	996	1763	95.8%	2.2%	0.7%	1.3%
147	C	2409	1646	8104	122367	1.8%	1.2%	6.0%	91.0%
146	C	3178	1604	9520	120289	2.4%	1.2%	7.1%	89.4%

D

Position	Alu consensus	Exonized Alus: Observed A	G	T	C	Expected A	G	T	C	E value
172	T	0	2	22	1	0.4	0.3	23.1	1.1	
171	A	23	2	0	0	24.0	0.7	0.2	0.1	
170	G	1	22	1	1	0.8	23.4	0.3	0.5	
169	C	0	0	1	24	0.2	0.3	0.9	23.5	
168	T	0	0	24	1	0.2	0.3	24.1	0.4	
167	G	0	25	0	0	1.7	23.0	0.2	0.1	
166	G	0	24	0	0	1.5	22.0	0.3	0.2	
165	G	3	22	0	0	1.6	22.9	0.3	0.3	
164	A	24	1	0	0	24.1	0.5	0.3	0.1	
163	T	1	0	13	11	0.3	0.2	12.1	12.4	
162	T	0	0	21	4	0.4	0.2	22.8	1.6	
161	A	24	0	1	0	24.1	0.5	0.2	0.3	
-3 160	C	0	0	1	24	0.4	0.3	2.1	22.2	
-2 159	A	25	0	0	0	24.1	0.5	0.2	0.1	
-1 158	G	0	25	0	0	1.1	23.5	0.2	0.2	
1 157	G	0	25	0	0	1.3	23.0	0.3	0.3	
2 156	C	0	0	21	4	0.6	0.4	7.7	16.2	2.87E-07
3 155	G	10	15	0	0	11.0	13.1	0.5	0.4	
4 154	C	2	0	12	11	0.4	0.4	9.2	15.0	
5 153	G	2	18	0	5	5.1	10.9	0.6	8.4	3.60E-02
6 152	C	1	0	13	11	0.8	0.4	9.7	14.0	
7 151	G	6	19	0	0	8.8	15.5	0.2	0.4	
150	C	2	0	2	21	0.5	0.3	1.3	23.0	
149	C	1	0	3	21	0.5	0.2	1.4	23.0	
148	A	19	2	2	2	23.9	0.5	0.2	0.3	
147	C	0	1	2	22	0.4	0.3	1.5	22.7	
146	C	1	1	1	22	0.6	0.3	1.8	22.3	

Figure 2. Comparison of Exonized *Alu*s to Nonexonized Intronic *Alu*s

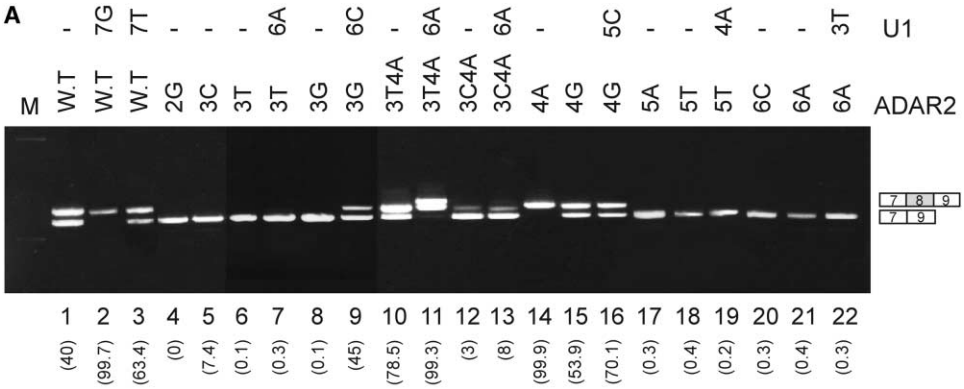
(A) Profile of 166,276 full-length antisense intronic *Alu*s. For each position, the number of appearances of each nucleotide (count per position) is shown as well as the frequency of each base in the position ("percent per position").

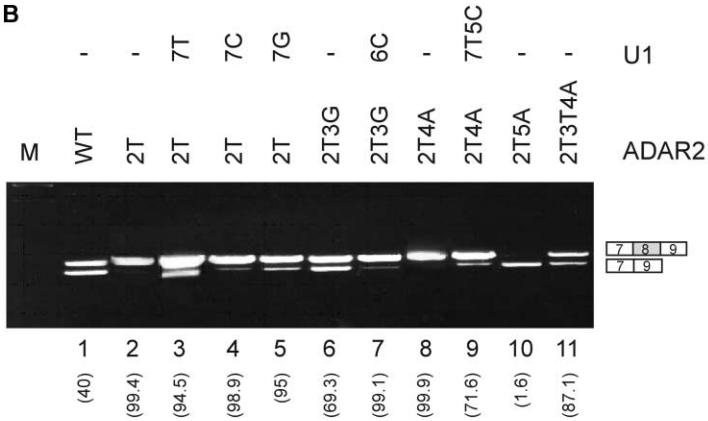
(B) Per position comparison between the 25 exonized *Alu*s (observed) and nonexonized intronic antisense *Alu*s. Expected profile was calculated by multiplying the profile matrix of nonexonized antisense intronic *Alu*s from (A) by 25. Significant deviations between the observed and expected profiles are apparent in positions 156 and 153. In position 156, T is expected 7.6 times (marked gray) but observed 21 times (gray), indicating a strong tendency of this position to change from C to T in exonized *Alu*s. In position 153, G is expected 10.9 times but appears 18 times, indicating the importance of G in position 5 of the splice site.

(C) Profile as in (A) but for 136,151 full-length sense intronic *Alu*s.

(D) Per position comparison between the 136,151 full-length sense intronic *Alu*s and 25 exonized *Alu*s. Comparison as in (B), indicating positions 156 and 153 as having significant deviations between the observed and expected profiles, with only slight differences between the E values in (B) and (D).

unpair with U1 (see also Freund et al., 2003). A U1 containing a compensatory mutation restoring the base pairing in position three of the 5′ss restores *AEx* inclusion (Figure 3A, lane 9). In contrast, a T:A compensatory mutation in the same position failed to restore *AEx* inclusion (Figure 3A, lane 7). This suggests that G:C rather than A:U pairing in position three of GC 5′ss contains the sufficient energy for U1 binding to the 5′ss. The failure of this A:T pairing to promote *AEx* inclusion is in contrast to the Ψ:A U1:5′ss pairing that is required for





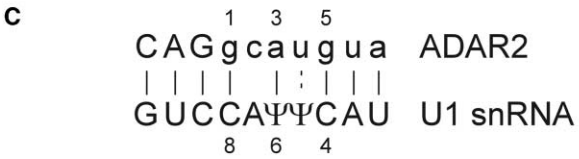


Figure 3. Splicing Assays on ADAR2 Minigene Mutants Using Compensatory U1 Mutants

(A) Analysis of GC introns. The top line shows the mutants in the U1 gene, and the second line shows the mutants in the ADAR2 5′ss sequence. The mutations are numbered according to the numbers indicated in (C). The indicated plasmid mutants were transfected or cotransfected into 293T cells, total cytoplasmic RNA was extracted, and splicing products were separated in 2% agarose gel after reverse transcription-polymerase chain reaction (RT-PCR). The leftmost lane is the DNA size marker. Lane 1, splicing products of wt ADAR2, and lanes 2–22, splicing products of the indicated ADAR2 minigene mutated at the 5′ss. The two mRNA isoforms are shown on the right. Numbers in parentheses in the bottom indicate percentage of the *Alu*-containing mRNA isoform as determined by TINA2 (100% corresponds to the total of both mRNA isoforms). Lane 11 contains two closely joined bands; the sequences of both were found to be identical and correspond to exon inclusion. This phenomenon may be attributed to migration of the overexpressed U1 together with the splicing product.

(B) Analysis of GT introns. Leftmost lane is the DNA size marker. Lane 1, splicing products of wt ADAR2, and lanes 2–11 are the splicing products of the indicated mutants.

(C) Schematic illustration of the base paring between the 5′ss of exon/intron eight of ADAR2 and U1. ADAR2 and U1 positions are numbered forward and reverse, respectively. Watson-Crick and non-Watson-Crick base pairing are marked by solid or dashed line, respectively.

AEx inclusion. This might be related to the stabilizing effect of Ψ on the backbone and the stem structure (Arnez and Steitz, 1994) or to the dynamics it allows for a noncanonical interaction at the base of a stem (Sundaram et al., 2000).

Position four of the 5′ss affects the level of alternative splicing of the *AEx*. Mutation of that position from T to A, which strengthens the base pairing with position five in U1, resulted in constitutive inclusion of the *AEx* (Figure 3A, lane 14). Mutation of the same position from T to G increased the level of the *AEx* inclusion from 40% to 54%. The inclusion level increased to 70% following cotransfection with U1 containing a compensatory mutation that base pairs with that G (Figure 3A, lanes 15 and 16).

To understand whether or not a cooperative effect between positions three and four exists, we produced double mutants and tested their inclusion ratios. Mutations that introduced a mispairing between U1 and position three (A to T) but allowed better base pairing of U1

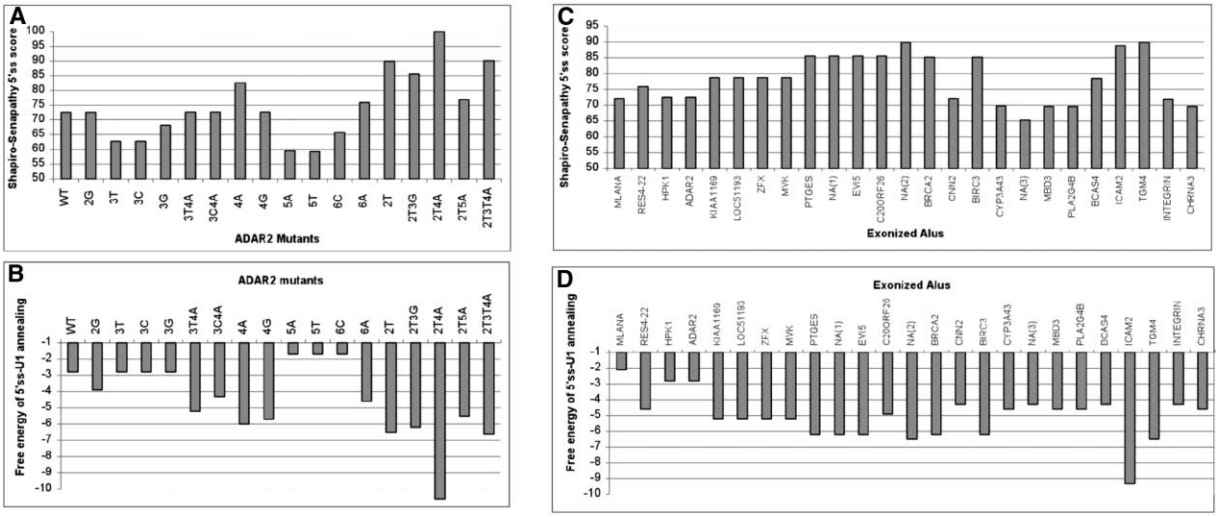


Figure 4. Strengths of 5′ss

(A) Shapiro-Senapathy score for 5′ss strength (Shapiro and Senapathy, 1987), calculated for the ADAR2 5′ss mutants indicated in Figures 3A and 3B.

(B) Free energy (ΔG) for 5′ss:U1 binding in ADAR2 5′ss mutants. Free energy was calculated using the software Mfold (Zuker, 1989), version 3.1 and according to Carmel et al. (2004).

(C) Shapiro-Senapathy score for the 25 exonized *Alu*s indicated in Figure 1.

(D) Free energy (ΔG) for 5′ss:U1 binding in 5′ss of the 25 exonized *Alu*s.

with position four (T to A) restored alternative splicing of the *AEx* with a ratio of inclusion of 78% (Figure 3A, lane 10). This splicing became constitutive when the cells were cotransfected with U1 containing a compensatory mutation that base pairs with the T in position three (Figure 3A, lane 11). Our results show that positions three and four are involved in controlling the level of exon inclusion in GC 5′ss. We note that position three in GC 5′ss may be T if position four is A; however, no such case was found in the set of exonized *Alu*s, probably because such an event would require two sequential transversions (from GC to TA), a combination of events that have a very low probability of occurrence in the genome.

The cooperative effect between positions three and four did not occur in double mutants in which position three was mutated to C and position four to A (Figure 3A, lanes 12 and 13), implying that U but not C, in position three of the 5′ss, can form a noncanonical Ψ:U pairing with Ψ in position six of U1. Indeed, Ψ:U base pairing was recently reported to be important in position four of 5′ss in yeast splicing (Libri et al., 2002). The above results suggest a hierarchy in the strength of the pairing between U1 and position three of the 5′ss: Ψ:A > Ψ:U > Ψ:C. This hierarchy seems to determine the level of alternative splicing.

We further studied the importance of positions five and six in GC 5′ss. Mutations in these positions resulted in the total skipping of the *AEx*, which compensatory mutations in U1 failed to restore (Figure 3A, lanes 17–22). This suggests that other splicing factors also recognize these positions in 5′ss.

To study the regulation of *AEx*s with GT in their 5′ss, we mutated the C in position 156 of the ADAR2 *AEx* to T, creating a GT 5′ss (Figure 3B). This mutation resulted in a shift from alternative to constitutive inclusion of the *AEx* (Figure 3B, lanes 1 and 2), further supporting our conclusion that the weak base pairing with U1 maintains the alternative splicing of the GC 5′ss. Cotransfection with U1 that unpairs with the T at position two had only a minor effect of ~5% exon skipping (Figure 3B, compare lane 2 to lanes 3–5), presumably reflecting the effect of U5(p220) binding to that nucleotide (Reyes et al., 1996). In addition, as predicted from our bioinformatical analysis (Figure 2), G in position five of the intron is essential for the alternative splicing of *AEx*s, as a G to A mutation in that position resulted in total *AEx* skipping (Figure 3B, lane 10).

In contrast to the GC 5′ss, which required an obligatory A in position three, mutation of that position to G in the GT 5′ss reduced the splicing from constitutive to alternative but did not eliminate exon inclusion entirely (Figure 3B, compare lane 2 to 6). This probably stems from the fact that GT introns are generally stronger 5′ss than GC introns, and a G in position three of the GT 5′ss can form a G:Ψ pairing with U1 (see also Freund et al., 2003). Cotransfection with U1 containing a compensatory mutation that base pairs with the G in position three restored a constitutive splicing, suggesting that the base pairing of that position with U1 is a main factor affecting the alternative splicing ratio (Figure 3B, lanes 6 and 7, and also Figure 3A, lane 10). A similar preference for A over G in position three was also observed in other 5′ss where the base pairing between U1 and the 5′ss was suboptimal (Burge and Karlin, 1997; Freund et al., 2003).

To test the importance of position four in the GT 5′ss, we mutated it to A, which base pairs with the wt U1. This maintained the constitutive inclusion of the *AEx* (Figure 3B, lane 8). When a mutated U1 that unpairs with position four was cotransfected, the splicing became alternative (Figure 3B, lane 9). A return to alternative splicing was also observed when an additional mutation, which unpairs with U1, was introduced in position three (Figure 3B, lane 11). This further supports the results from the GC 5′ss analysis, which showed that the nucle-

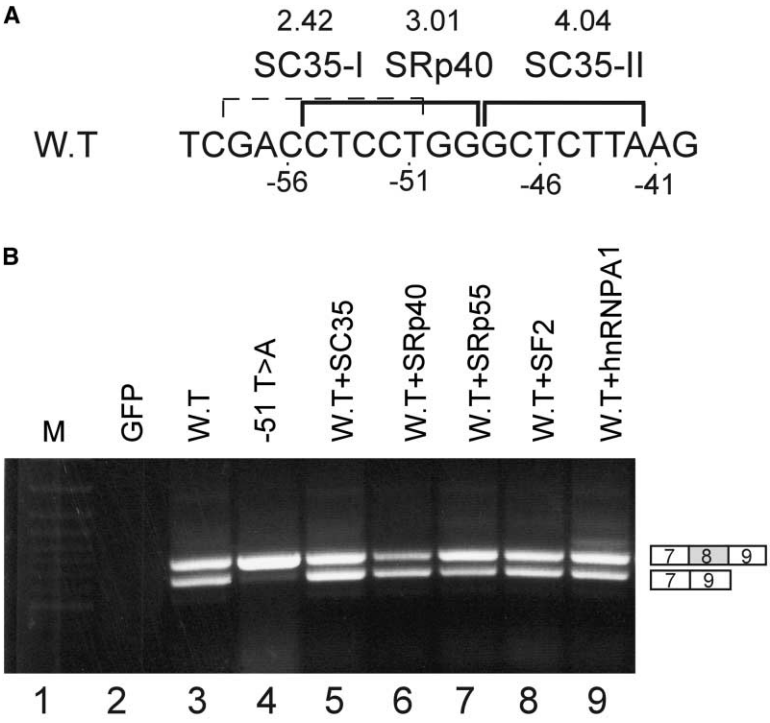


Figure 5. A Potential ESE Site in Exon Eight of ADAR2 Is Not Involved in Alternative Splicing Regulation

(A) The ADAR2 sequence from position −40 to −60 upstream of the 5′ss. SR protein potential sites are marked with a solid line; the broken line indicates a potential site that was enhanced by the mutation in position −51 from T to A. The potential sites were detected by ESEfinder (Roca et al., 2003); the type and binding score of each site is indicated.

(B) Transfection or cotransfection was performed in 293T cells. Total RNA and RT-PCR were performed as described in Figure 3. Lane 1, DNA size marker; lane 2, vector only (pEGFP-C3); lane 3, splicing products of wt ADAR2; lane 4, splicing products of the T to A mutant in position −51 that enhanced the ESE of SC35-I from a score of 2.4 to 4.2 (−51 T > A); lanes 5–9, splicing products of wt ADAR2 with the indicated SR/A1 hnRNP protein.

otide composition in positions three and four affects the delicate balance of skipping or inclusion of alternatively spliced exons.

From these serial mutations in the 5′ss and the compensatory mutations in U1 we conclude the following. (1) U1:5′ss base pairing is involved in both GC and GT 5′ss selection. (2) The alternative splicing of ADAR2 *AEx* is maintained due to the unpairing of U1 with position two of the 5′ss. (3) The nucleotide composition of positions three and four in the intron, both in GT 5′ss and in GC 5′ss, control the delicate skipping/inclusion ratio depending on the canonical/noncanonical base pairing of these positions with U1. (4) An A in position three of the intron is more important in GC 5′ss than in GT 5′ss, possibly due to the need to avoid two successive nonpairing positions. (5) G in position five is essential for the selection of 5′ss in the two types of introns.

The results for position four of the 5′ss indicate that when this position is mutated to A the ADAR2 exon becomes constitutive. However, there are two cases among the 25 alternatively spliced *Alu* exons in our compilation that contain A in position four (Figure 1, rows 2 and 22). A similar inconsistency is observed for the mutation of position six from T to C, which makes the ADAR2 exon inactive, but is found in several exonized *Alu*s (Figure 1). To further examine this inconsistency, we calculated splice site scores and free energy of U1:5′ss binding (ΔG) in exonized and mutated *Alu*s (Figure 4). The splice site score is a measure of how "close" the splice site is to the consensus sequence profile of 5′ss (Shapiro and Senapathy, 1987). The free energy is a measure of the U1:5′ss binding strength, taking also into consideration the differences between G:C, A:U, and G:U base pairing and stacking energy—lower ΔG values stand for stronger binding and might indicate higher exon inclusion/exclusion ratio (Carmel et al., 2004).

When position six in ADAR2 is mutated to C, the 5′ss score is reduced from 73 to 65 (Figure 4A, mutant 6C) and the free energy is increased from −2.8 to −1.7, which indicates that binding is inefficient (Figure 4B, mutant 6C). In exonized *Alu*s where position six is C the 5′ss score can also get as low as 65 [for example NA(3) in Figure 4C], but the U1:5′ss free binding energy in these exons is never higher than −4.0 [Figure 4D, see for example ZFX, MVK, NA(3), and MBD3]. The lower ΔG stands for more efficient U1:5′ss binding, which explains why these exons are still recognized.

When position four in ADAR2 is mutated to A, the 5′ss score increases from 73 to 83, and the ΔG becomes −6.0 (Figures 4A and 4B, mutant 4A), which is in agreement with the fact that this mutation results in a constitutively spliced exon. In the exonized *Alu* in gene RES4-22 position four is A, but the 5′ss score is only 75 (similar to the score of the wt alternative ADAR2) and the ΔG is −4.6 (Figures 4C and 4D). This can explain the fact that the *AEx* of RES4-22 is alternatively spliced. In the *AEx* of ICAM2, position four is also A, and the 5′ss score and ΔG indicate a very high affinity to U1. However, that *AEx* is alternatively spliced, which might indicate that sequence elements other than the ones in the actual 5′ss are involved in its regulation.

Overall, the 5′ss score for exonized *Alu*s was similar to that for non-*Alu* alternative cassette exons, averaging 78.3 in the 25 *AEx*s and 79.8 in the set of 243 non-*Alu* cassette exons from Sorek and Ast (2003). The free energy value was also similar between the sets, with an average of −5.08 in the 25 *AEx*s and −5.29 in the non-*Alu* cassette exons.

The results presented here and in Lev-Maor et al. (2003) indicate that the sequence composition of both 3′ss and 5′ss determines the alternative splicing ratio of *AEx*s. In addition, our statistical comparison between intronic *Alu*s and exonized ones shows that only two

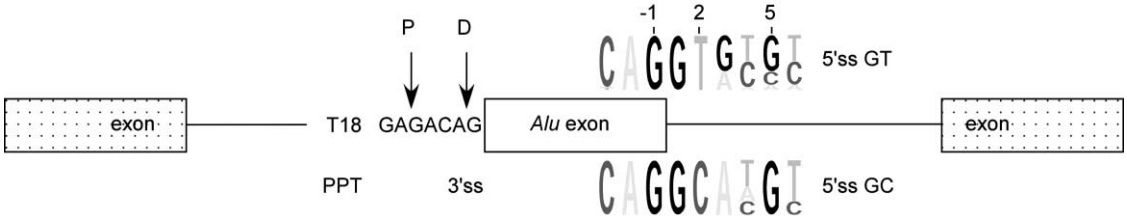


Figure 6. Sequence Elements Needed for the Creation of an Alternatively Spliced *Alu* Exon

The prevalent 3′ splice site (3′ss) is found in position 275 of the antisense *Alu* and is composed of a polypyrimidine tract (PPT) averaging 18 bases in length, followed by a 3′ss motif of GAGACAG containing a proximal and a distal AG (P and D, respectively) (Lev-Maor et al., 2003). The prevalent 5′ss can either begin with GT or with GC. Pictograms (http://genes.mit.edu/pictogram.html) depict the profiles of 21 GT-exonized and 4 GC-exonized *Alu*s, above and below the box, respectively.

positions along the *Alu* sequence substantially differ between intronic and exonized *Alu*s, both in the 5′ss sequence (Figure 2). These results imply that only the sequences of the 3′ and 5′ss of the *AEx*s are important for exonization. This conclusion is supported by the finding that, so far, all the mutations leading to exonization of intronic *Alu*s that cause genetic disorders were only found to affect 5′ss or 3′ss sequences (references in Figure 1 and in Lev-Maor et al., 2003).

Still, it is possible that the splicing of *AEx*s is regulated by splicing enhancers or silencers residing outside the splice sites. We used the ESE finder (Roca et al., 2003) to search the sequence of the *AEx* of ADAR2 for such potential sequences. Seven potential binding sites for SR proteins were found, three of them in close proximity (Figure 5A). SR proteins are known to be involved in alternative splicing regulation through binding to short sequences on the RNA molecule (Graveley, 2001). The score of all potential SR binding sites was low, but such sites can still be functional (Roca et al., 2003). To test whether these sites are functional, we serially mutated positions −48 to −54. We found no effect on the splicing pattern of the *AEx*, indicating that these potential ESEs are not functional (data not shown). Mutations in position −11, embedded within another weak potential ESE, gave similar results (data not shown).

To examine whether other sites are involved in the regulation of the *AEx* splicing, we cotransfected 293T cells with the ADAR2 minigene and various plasmids containing the most abundant splicing regulatory proteins (SR proteins and hnRNP A1, the latter was shown to promote exon skipping; Figure 5B, lanes 5–9) (Cartegni et al., 2002). In principle, if one of these proteins was involved in the regulation of *AEx* splicing, then increasing its nuclear concentrations might affect the skipping/inclusion ratio of the *AEx*. This was not the case: none of these proteins affected the alternative splicing of the *AEx*. This suggests that the *AEx* of ADAR2 has no splicing enhancers or silencers that can bind to one of these proteins and regulate the *AEx* splicing, but we cannot rule out the possibility that other splicing regulatory proteins, which are not part of the panel in Figure 5, affect the inclusion/skipping ratio of this exon. As a control to that experiment, we artificially created a strong binding site for SC35 by mutating T in position −51 to A and thus increasing the ESE score from 2.4 to 4.2. This mutation led to a shift toward exon inclusion, indicating that, in principle, SR proteins have the potential to modulate the splicing of *AEx*s but are presumably not involved in the case of the exon under study (Figure 5B, lanes 3 and 4). This further supports the analysis in Figure 2, implying that in the general exonization process of *AEx*s there is no selective pressure for the creation or loss of a splicing regulatory sequence/s located outside of the splice sites and suggesting that only the sequences of the splice sites are involved in that process.

Discussion

We have analyzed the factors influencing *Alu* exonization through a combined computational and experimental approach. We showed that 5′ss can be created de novo within *Alu* sequences in a process requiring only minimal base substitutions, involving positions two and five of the intron. By mutational analysis we extensively studied the delicate balance between the different positions of the 5′ss and U1 and showed how this balance controls the inclusion/skipping ratio of the exon in both GC and GT introns.

Our results suggest that the fast decay of CG dinucleotides to TG or CA in the human genome is a major driving force for *Alu* exonizations. This decay causes a CAGGCG motif in positions 160-154 of antisense *Alu*s to become either CAGGTG (CG becomes TG, creating functional GT 5′ss) or CAGGCA (CG becomes CA, creating functional GC 5′ss with A in position three). Ironically, CG decay is considered one of the evolutionary mechanisms promoting the silencing of retroposition of active *Alu*s in the genome, as there is a correlation between the number of CG dinucleotides in the *Alu* sequence and its retroposition activity (Deininger and Batzer, 2002). Therefore, we would expect a negative correlation between retropositional activities of *Alu*s and exonization, i.e., if an *Alu* became an exon, there is less probability for it to be transpositionally active.

The results presented in this article and the previous analysis of 3′ss formation and regulation in exonized *Alu*s (Lev-Maor et al., 2003) provide the molecular basis for *Alu* exonization. Figure 6 summarizes these findings and shows a model for the composition of the prevalent 3′ss and 5′ss in exonized *Alu*s. The prevalent 3′ss is found in position 275 of the antisense *Alu*. It is composed of a polypyrimidine tract (PPT) averaging 18 bases in length, followed by a 3′ss motif of GAGACAG. In the 3′ss motif the distal AG is selected, and there is a delicate interplay between the two AGs. The G at position −7

(the seventh nucleotide upstream of the distal AG) suppresses the selection of the proximal AG: when that G is mutated there are two effects—the *AEx* becomes constitutive and the proximal AG is selected. The proximal AG is essential to weaken the selection of the distal AG, thus maintaining alternative splicing. The four nucleotides distance between the proximal and distal AG also ensure that mode of alternative splicing; increasing that distance leads to *AEx* skipping, and *AEx* inclusion can be restored by a high concentration of the second step splicing factor hSlu7 when the distance between the two AGs is over eight nucleotides. On the other side of the *AEx* the prevalent 5′ss is in position 158. There are two types of 5′ss that can be used in that same position: GT 5′ss (more common) and GC 5′ss, in which position three has to be A (less common).

This model, as well as the results presented in Figure 5, indicate that exonization of *Alu* sequences depends almost solely on the sequence composition of the potential 3′ and 5′ splice sites. *Alu*s containing the 5′ss described here and the 3′ss described in Lev-Maor et al. (2003) are predicted to become alternatively spliced exons. We scanned the genome for *Alu*s that have the potential to undergo exonization (see Experimental Procedures). We found 244,472 *Alu*s in the antisense orientation located in introns of genes (not necessarily being full-length *Alu*s). Of these, 46,518 had GT or GC 5′ss (GT: 15,413 and GC: 31,105) and 7810 also had the ADAR2-like 3′ss with a strong polypyrimidine tract preceding it. This set represents *Alu*s that have either been exonized already or are "on the verge" of exonization (Supplemental Database S1 at http://www.molecule.org/cgi/content/full/14/2/221/DC1).

Presumably, the 7810 *Alu*s should be exonized. We manually examined several examples from this set: only a minority of these *Alu*s were found in expressed sequences (ESTs or mRNAs), and therefore, we lack evidence for exonization in the majority of cases. In one case of a "perfect" candidate *Alu* found in intron 20 of the inhibitor of κ-light polypeptide gene enhancer in B cell's kinase complex-associated protein (IKBKAP) gene, we experimentally tested the possibility of exonization. Mutations in this gene are known to lead to familial dysautonomia (FD), an autosomal recessive congenital neuropathy disorder. The major FD-causing mutation (99.5%) affects the 5′ss of exon 20 in IKBKAP, changing position +6 from T to C and leading to aberrant splicing (Fini and Slaugenhaupt, 2002). The perfect *Alu* elment in intron 20 contains a putative PPT of 18 Ts, followed by the GAGACAG motif. In the putative 5′ss, it contains a CAGgtgtgc sequence having a 5′ss score of 78.8 and ΔG of −5.2, which is similar to that of exonized *Alu*s. This *Alu* is located more than 100 nucleotides from the flanking splice sites, so that there are probably no intron size constraints on the exonization process. In spite of these facts, we could detect no exonization of this *Alu* (data not shown).

There may be several reasons for these findings. First, the inclusion ratio of most *AEx*s is 10% or lower (Sorek et al., 2002), and for many of these *Alu*s, EST sampling may be more limited than the level required for discovery of lowly expressed exonized *Alu*s. ESTs that may become available in the future may uncover additional instances of *Alu* exonization. Second, some *Alu*s might lack the proper branch point upstream of their 3′ss (although the branch point consensus sequence is degenerate). Third, other factors, such as reading frame compatibility, have been shown to affect spliceosome selection of exons (Li et al., 2002). Finally, intronic sequence elements that reside outside the *Alu* sequence itself might have an inhibitory effect on the splicing of *Alu*. Indeed, Fairbrother and Chasin (2000) demonstrated that a substantial fraction of the human genome contains sequences that have splicing-inhibition properties. In any case, revealing why some of these *Alu*s are selected as exons and some are not could be a major step toward understanding how mammalian exons are defined.

Based on the aforementioned mutation leading to CCFDN syndrome, we can predict that C-to-T mutation in position two of the GC 5′ss of some of these 7810 perfect intronic *Alu*s may, in some cases, lead to constitutive exonization and may result in genetic disease. These *Alu*s on the verge of exonization may have an additional evolutionary role: they may serve as a "reservoir" for new human-specific exons that may one day be exapted (i.e., adapted to a function different from their original) into promoting speciation of the human lineage.

Experimental Procedures

Building a Genomic *Alu* Profile

The *Alu*Gene database (http://alugene.tau.ac.il/; Dagan et al., 2004) was used to extract full-length genomic *Alu* sequences that are found within introns in the antisense and sense orientations of *Alu*. This search yielded 166,276 and 136,151 full-length *Alu*s in the antisense and sense orientations, respectively. Each of these *Alu*s was aligned to the *Alu* consensus sequence, using ClustalW 1.8 (Thompson et al., 1994). A nucleotide frequency profile was built for each site along the *Alu* consensus sequence (Figures 2A and 2C). In a similar manner, the sequences of 25 *Alu*s that were exonized using position 157 as 5′ss (Figure 1) were aligned to their consensus and an "observed" profile was built (Figures 2B and 2D). A parallel "expected" profile was calculated, by multiplying the frequency profile matrix from the nonexonized intronic *Alu*s by 25 (the number of exonized sequences). Statistical significance between the observed and expected profile was calculated using chi-square statistical test. *Alu*Gene was further queried in the same manner to find the 244,472 *Alu*s in the antisense orientation (not necessarily full-length) located in introns of known genes.

Calculation of Splice Site Strength

Splice site scores in Figures 4A and 4C were calculated using the matrix from Shapiro and Senapathy (1987). The free energy in Figures 4B and 4D was calculated using the Mfold software (Zuker, 1989) version 3.1, which predicts the free energy (ΔG) of a single folded RNA strand. To predict the ΔG in U1:5′ss annealing, we concatenated their sequence strands into one RNA strand as follows: "NN"-5′ss-"NNNNN"-reversed U1-"NN." This sequence was given as the input for Mfold. Since this software does not hybridize the "N" sequence because it is considered neutral, the resulting calculated free energy was that of the U1:5′ss hybrid. A more detailed description is found in Carmel et al. (2004). For both score and energy calculations, the splice site was set to be at positions −3 to +6 relative to the 5′ss.

Plasmid Constructs

Oligonucleotide primers were designed to amplify a minigene that contains the exons 7, 8, and 9 of the gene adenosine deaminase (ADAR2). Each primer contained an additional sequence encoding a restriction enzyme. The PCR product (2.2 kb) was restriction de-

signed and inserted between the KpnI/BglII sites in the pEGFP-C3 plasmid (Clonthech). The U1 gene was cloned in the pCR vector.

Site-Directed Mutagenesis

Oligonucleotide primers containing the desired mutations were used to amplify the mutation-containing replica of either the wt ADAR2 minigene plasmid or the U1 gene, respectively. The products were treated with DpnI restriction enzyme (12 U) (New England Biolabs) at 37°C for 1 hr. The mutant DNA (1–4 μL) was transformed into *E. coli* DH5α strain. Colonies were picked, followed by miniprep (QIAgene) and midiprep (BRL). All plasmids were confirmed by sequencing.

Transfection, RNA Isolation, and RT-PCR Amplification

293T cell line was cultured in Dulbecco's Modification of Eagle medium, supplemented with 4.5 gr/mL glucose (Renium) and 10% fetal calf serum (Biological Industries). Cells were cultured in 60 mm dishes under standard conditions at 37°C with 5% CO_2. Cells were grown to 50% confluence, and transfection was performed using 12 μL FuGENE6 (Roche) with 4 μg of plasmid DNA. After 48 hr, cells were harvested. Total cytoplasmic RNA was extracted using Tri Reagent (Sigma), followed by treatment with 2 U DNase RNase-free (Ambion). Reverse transcription (RT) was performed on 2 μg total cytoplasmic RNA for 1 hr at 42°C, using the pEGFP-C3-specific reverse primer and 2 U reverse transctiptase of avian myeloblastosis virus (RT-AMV, Roche).

The spliced cDNA products derived from the expressed minigene were detected by PCR, using the pEGFP-C3-specific reverse primer and an exon seven forward primer. Amplification was performed for 30 cycles, consisting of 94°C for 30 s, 61°C for 45 s, and 72°C for 1 min using high-fidelity Taq (Roche). The products were resolved on 2% agarose gel.

Acknowledgments

We thank M. Kupiec for critical reading, we thank Alan Weiner for the U1 gene, and we thank Adrian Krainer and Stefan Stemm for the SR plasmids. We also thank the Bioinformatics Unit at Tel Aviv University for providing technical assistance and computation facilities, and we thank Amir Goren for generating the website. R.S., G.L.-M., M.R., and G.A. were supported by a grant from the Israel Science Foundation, FD Hope, Israel Cancer Association, Chief Scientist of Israel Health Ministry, and the MOP India-Israel. R.S., T.D., F.B., and D.G. were supported by the Norman and Rose Lederer Chair of Biology at Tel Aviv University.

Received: January 26, 2004
Revised: March 17, 2004
Accepted: March 23, 2004
Published: April 22, 2004

References

Arnez, J.G., and Steitz, T.A. (1994). Crystal structure of unmodified tRNA(Gln) complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. Biochemistry *33*, 7560–7567.

Batzer, M.A., Kilroy, G.E., Richard, P.E., Shaikh, T.H., Desselle, T.D., Hoppens, C.L., and Deininger, P.L. (1990). Structure and variability of recently inserted Alu family members. Nucleic Acids Res. *18*, 6793–6798.

Black, D.L. (2000). Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell *103*, 367–370.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. Nat. Genet. *30*, 29–30.

Brosius, J. (1999). RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene *238*, 115–134.

Brow, D.A. (2002). Allosteric cascade of spliceosome activation. Annu. Rev. Genet. *36*, 333–360.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Carmel, I., Tal, S., Vig, I., and Ast, G. (2004). Comparative analysis detects dependencies among the 5′ splice-sites positions. RNA *10*, 828–840.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. *3*, 285–298.

Cohen, J.B., Broz, S.D., and Levinson, A.D. (1993). U1 small nuclear RNAs with altered specificity can be stably expressed in mammalian cells and promote permanent changes in pre-mRNA splicing. Mol. Cell. Biol. *13*, 2666–2676.

Dagan, T., Sorek, R., Sharon, E., Ast, G., and Graur, D. (2004). AluGene: a database of Alu elements incorporated within protein-coding genes. Nucleic Acids Res. *32*, D489–D492.

Deininger, P.L., and Batzer, M.A. (2002). Mammalian retroelements. Genome Res. *12*, 1455–1465.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. *35*, 41–48.

Dubbink, H.J., de Waal, L., van Haperen, R., Verkaik, N.S., Trapman, J., and Romijn, J.C. (1998). The human prostate-specific transglutaminase gene (TGM4): genomic organization, tissue-specific expression, and promoter characterization. Genomics *51*, 434–444.

Fairbrother, W.G., and Chasin, L.A. (2000). Human genomic sequences that inhibit splicing. Mol. Cell. Biol. *20*, 6816–6825.

Farrer, T., Roller, A.B., Kent, W.J., and Zahler, A.M. (2002). Analysis of the role of Caenorhabditis elegans GC-AG introns in regulated splicing. Nucleic Acids Res. *30*, 3360–3367.

Fini, M.E., and Slaugenhaupt, S.A. (2002). Enzymatic mechanisms in corneal ulceration with specific reference to familial dysautonomia: potential for genetic approaches. Adv. Exp. Med. Biol. *506*, 629–639.

Freund, M., Asang, C., Kammler, S., Konermann, C., Krummheuer, J., Hipp, M., Meyer, I., Gierling, W., Theiss, S., Preuss, T., et al. (2003). A novel approach to describe a U1 snRNA binding site. Nucleic Acids Res. *31*, 6963–6975.

Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet. *17*, 100–107.

Harrison, P.M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res. *30*, 1083–1090.

Hastings, M.L., and Krainer, A.R. (2001). Pre-mRNA splicing in the new millennium. Curr. Opin. Cell Biol. *13*, 302–309.

Hu, M.C., Qiu, W.R., Wang, X., Meyer, C.F., and Tan, T.H. (1996). Human HPK1, a novel human hematopoietic progenitor kinase that activates the JNK/SAPK kinase cascade. Genes Dev. *10*, 2251–2264.

Johnson, S., Halford, S., Morris, A.G., Patel, R.J., Wilkie, S.E., Hardcastle, A.J., Moore, A.T., Zhang, K., and Hunt, D.M. (2003). Genomic organisation and alternative splicing of human RIM1, a gene implicated in autosomal dominant cone-rod dystrophy (CORD7). Genomics *81*, 304–314.

Jurka, J., and Milosavljevic, A. (1991). Reconstruction and analysis of human Alu genes. J. Mol. Evol. *32*, 105–121.

Kazazian, H.H., Jr. (2000). Genetics. L1 retrotransposons shape the mammalian genome. Science *289*, 1152–1153.

Knebelmann, B., Forestier, L., Drouot, L., Quinones, S., Chuet, C., Benessy, F., Saus, J., and Antignac, C. (1995). Splice-mediated insertion of an Alu sequence in the COL4A3 mRNA causing autosomal recessive Alport syndrome. Hum. Mol. Genet. *4*, 675–679.

Kreahling, J., and Graveley, B.R. (2004). The origins and implications of Aluternative splicing. Trends Genet. *20*, 1–4.

Kunkel, T.A., and Diaz, M. (2002). Enzymatic cytosine deamination: friend and foe. Mol. Cell *10*, 962–963.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. (2003). The birth

of an alternatively spliced exon: 3′ splice-site selection in Alu exons. Science *300*, 1288–1291.

Li, B., Wachtel, C., Miriami, E., Yahalom, G., Friedlander, G., Sharon, G., Sperling, R., and Sperling, J. (2002). Stop codons affect 5′ splice site selection by surveillance of splicing. Proc. Natl. Acad. Sci. USA *99*, 5277–5282.

Libri, D., Duconge, F., Levy, L., and Vinauger, M. (2002). A role for the Psi-U mismatch in the recognition of the 5′ splice site of yeast introns by the U1 small nuclear ribonucleoprotein particle. J. Biol. Chem. *277*, 18173–18181.

Makalowski, W., Mitchell, G.A., and Labuda, D. (1994). Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. *10*, 188–193.

Mihovilovic, M., Mai, Y., Herbstreith, M., Rubboli, F., Tarroni, P., Clementi, F., and Roses, A.D. (1993). Splicing of an anti-sense Alu sequence generates a coding sequence variant for the alpha-3 subunit of a neuronal acetylcholine receptor. Biochem. Biophys. Res. Commun. *197*, 137–144.

Miller, M., and Zeller, K. (1997). Alternative splicing in lecithin:cholesterol acyltransferase mRNA: an evolutionary paradigm in humans and great apes. Gene *190*, 309–313.

Mitchell, G.A., Labuda, D., Fontaine, G., Saudubray, J.M., Bonnefont, J.P., Lyonnet, S., Brody, L.C., Steel, G., Obie, C., and Valle, D. (1991). Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. Proc. Natl. Acad. Sci. USA *88*, 815–819.

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

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. Nat. Genet. *34*, 177–180.

Nissim-Rafinia, M., and Kerem, B. (2002). Splicing regulation as a potential genetic modifier. Trends Genet. *18*, 123–127.

Reyes, J.L., Kois, P., Konforti, B.B., and Konarska, M.M. (1996). The canonical GU dinucleotide at the 5′ splice site is recognized by p220 of the U5 snRNP within the spliceosome. RNA *2*, 213–225.

Roca, X., Sachidanandam, R., and Krainer, A.R. (2003). Intrinsic differences between authentic and cryptic 5′ splice sites. Nucleic Acids Res. *31*, 6321–6333.

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.

Sorek, R., and Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res. *13*, 1631–1637.

Sorek, R., Ast, G., and Graur, D. (2002). Alu-containing exons are alternatively spliced. Genome Res. *12*, 1060–1067.

Stoilov, P., Meshorer, E., Gencheva, M., Glick, D., Soreq, H., and Stamm, S. (2002). Defects in pre-mRNA processing as causes of and predisposition to diseases. DNA Cell Biol. *21*, 803–818.

Sundaram, M., Durant, P.C., and Davis, D.R. (2000). Hypermodified nucleosides in the anticodon of tRNA(Lys) stabilize a canonical U-turn structure. Biochemistry *39*, 15652.

Svineng, G., Fassler, R., and Johansson, S. (1998). Identification of beta1C-2, a novel variant of the integrin beta1 subunit generated by utilization of an alternative splice acceptor site in exon C. Biochem. J. *330*, 1255–1263.

Thanaraj, T.A., and Clark, F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids Res. *29*, 2581–2593.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Varon, R., Gooding, R., Steglich, C., Marns, L., Tang, H., Angelicheva, D., Yong, K.K., Ambrugger, P., Reinhold, A., Morar, B., et al. (2003). Partial deficiency of the C-terminal-domain phosphatase of RNA polymerase II is associated with congenital cataracts facial dysmorphism neuropathy syndrome. Nat. Genet. *35*, 185–189.

Vervoort, R., Gitzelmann, R., Lissens, W., and Liebaers, I. (1998). A mutation (IVS8+0.6kbdelTC) creating a new donor splice site activates a cryptic exon in an Alu-element in intron 8 of the human beta-glucuronidase gene. Hum. Genet. *103*, 686–693.

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.

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.

Yang, Z., and Yoder, A.D. (1999). Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. *48*, 274–283.

Zhuang, Y., and Weiner, A.M. (1986). A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. Cell *46*, 827–835.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. Science *244*, 48–52.

Nucleic Acids Research, 2004, Vol. 32, Database issue ***D489–D492***
DOI: 10.1093/nar/gkh132

*Alu*Gene: a database of *Alu* elements incorporated within protein-coding genes

Tal Dagan*, Rotem Sorek[1,2], Eilon Sharon, Gil Ast[1] and Dan Graur

Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel, [1]Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel and [2]Compugen, 72 Pinchas Rosen Street, Tel Aviv 69512, Israel

Received August 20, 2003; Revised and Accepted October 27, 2003

ABSTRACT

***Alu* elements are short interspersed elements (SINEs) ~300 nucleotides in length. More than 1 million *Alu*s are found in the human genome. Despite their being genetically functionless, recent findings suggest that *Alu* elements may have a broad evolutionary impact by affecting gene structures, protein sequences, splicing motifs and expression patterns. Because of these effects, compiling a genomic database of *Alu* sequences that reside within protein-coding genes seemed a useful enterprise. Presently, such data are limited since the structural and positional information on genes and *Alu* sequences are scattered throughout incompatible and unconnected databases. *Alu*Gene (http://Alugene.tau.ac.il/) provides easy access to a complete *Alu* map of the human genome, as well as *Alu*-associated information. The *Alu* elements are annotated with respect to coding region and exon/intron location. This design facilitates queries on *Alu* sequences, locations, as well as motifs and compositional properties via a one-stop search page.**

INTRODUCTION

Alu sequences are short interspersed elements (SINEs), typically 300 nucleotides in length, which account for more than 10% of the human genome (1). *Alu*s have a dimeric structure and are ancestrally derived from the gene specifying 7SL RNA, an abundant cytoplasmic component of the signal recognition particle that mediates the translocation of secreted proteins across the endoplasmic reticulum (2). *Alu* elements multiply within the genome through RNA polymerase III-derived transcripts in a process termed retroposition.

Alu sequences can be divided into five subfamilies of related elements based upon key diagnostic nucleotide positions shared by subfamily members (3). Several overlapping subfamilies of *Alu* repeats of different evolutionary ages have been identified. These observations have led to the suggestion that *Alu* subfamilies have originated through successive waves of fixation from sequential small subsets of active *Alu* sequences. The oldest *Alu*-related elements are the monomeric *FAM*, *FRAM* and *FLAM* sequences. The oldest *Alu* dimeric subfamilies are *Alu*-Jo and *Alu*-Jb, estimated to be ~80 million years old. The intermediately aged *Alu* subfamilies belong to the *Alu*-S class, which is divided into subfamilies Sx, Sp, Sq, Sg and Sc. These subfamilies are estimated to be 30–50 million years old. The youngest subfamilies belong to the *Alu*-Y class, which are less than 15 million years old (2). Because of their newness, some *Alu*-Y elements have neither reached fixation nor been lost, and they exist in a polymorphic state. Most *Alu* repeats in the human genome belong to the *Alu*-S class, with *Alu*-Sx being the commonest (2).

Despite their being genetically functionless, recent findings suggest that *Alu* elements have a broad evolutionary impact. Parts of *Alu* elements may become inserted into mature mRNAs by way of splicing in a process called 'exonization'. Presumably, the exonization process is facilitated by sequence motifs within *Alu* that resemble splice sites (4–6). Indeed, more than 5% of the alternatively spliced exons in the human genome are *Alu* derived. All *Alu*-containing internal exons studied so far were found to be alternatively spliced (6). It was, thus, concluded that mutations resulting in constitutively spliced exonic *Alu*s would result, in the vast majority of cases, in the creation of defective genes causing deleterious effects on fitness. An example of such an occurrence is the addition of a new *Alu*-derived exon in conjunction with exon skipping in the β-glucuronidase gene resulting in a mild form of Sly syndrome (7). Another example of splicing-mediated diseases caused by *Alu* is the insertion of an *Alu* element into intron 18 of the Factor VIII gene, which leads to exon 19 skipping and results in a severe form of hemophilia A (8). *Alu*-mediated homologous unequal recombination may also result in genetic defects, as in the case of iduronate-2-sulfatase in which an *Alu*-mediated exon 8 deletion results in Hunter syndrome (9).

Alu insertions may sometimes create a new function or modify an existing one. One such example concerns tissue localization of casein kinase 2 (CK2) (10). CK2 is a $2\alpha + 2\beta$ tetrameric enzyme that phosphorylates serine and threonine

*To whom correspondence should be addressed. Tel: +972 3 6408646; Fax: +972 3 6409403; Email: tali@kimura.tau.ac.il
Present address:
Dan Graur, Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5501, USA

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

residues, and is essential for the viability of eukaryotic cells. A novel isoform of the α subunit was recently found to be highly expressed in the liver. Examination of the isoform sequence CK2α′, revealed a translated *Alu*-containing cassette exon incorporated into the mature mRNA. This C-terminal sequence was found to be essential in the determination of the nuclear localization of the CK2α″ isoform (10). *Alu*s were also found to be involved in expression regulation. For instance, *Alu* repeats in the distal promoter region of the human colesteryl ester transfer protein (CETP) were found to act as repressive regulatory elements of the activity of the promoter (11). *Alu*s have also been found to be involved in apoptosis. An alternatively spliced *Alu*-like exon was found to be essential for the ability of the Bcl-*rambo* β protein to promote etoposide- and taxol-induced cell death (12). Interestingly, *Alu*s are found to be highly clustered in genes that are involved in metabolism, transport and signaling processes, while they are less abundant in genes encoding structural proteins or information-pathway components. This non-random distribution was claimed to support the hypothesis that *Alu*s may not always be useless (junk) DNA (13,14). An alternative explanation may be that *Alu* insertions (or other mutations) may be deleterious, and hence more strongly selected against if they affect genes involved in information storage and processes.

Given the 'anecdotal' evidence concerning *Alu* involvement in gene structure and expression, it seems worthwhile to construct a genomic compilation of *Alu*s that reside within protein-coding genes. Such a database may be important in our efforts to elucidate the rules governing *Alu* exonization, gene regulation by *Alu* sequences and *Alu*-associated risk factors for mutation and pathogenesis. Presently, such data are limited since the information on gene and *Alu* location, as well as their characteristic and relative positions, is disjointed and scattered throughout incompatible and unconnected files in the various human genome databases. *Alu*Gene aims to provide an easy access to the complete *Alu* map within the human genome, as well as associated information, such as GC content and subfamily affiliation.

DATABASE DESIGN

The *Alu*Gene database was implemented using the Select Query Language (SQL) from the MySQL database server (http://www.mysql.com/). The database merges three main constituents: (i) a map of mRNAs juxtaposed on the human genome, (ii) a map of *Alu* sequences and (iii) a comparison (alignment) of each *Alu* element with the consensus sequence of the subfamily to which it belongs. Currently, the database relies on the May 2003 version of the NCBI human genome sequence (http://www.ncbi.nlm.nih.gov/), and will be updated with each new release. We used Perl scripts to extract positions and sequences of genes and their coding sequence from the GenBank files. The contig coordinates and orientations were stored as mapping data for each gene. Additional descriptive data for the mRNAs, such as LocusLink entry and protein entry, were also stored in the database. The locations of exons and intron were stored in distinct tables.

We used the RepeatMasker software (http://ftp.genome.washington.edu/RM/RepeatMasker.html) to search for *Alu* sequences in human genomic contigs. For each *Alu* entry, its location on the contig, orientation and sequence were stored in the database. Genomic locations of *Alu*s and genes were calculated from their position in the contigs and the relative location of the contig within the chromosome according to the seq_contig.md file in the NCBI. By using ClustalW (15), each *Alu* sequence was aligned to the consensus sequence of its subfamily (A. F. A. Smit and P. Green, unpublished data). Indels were interpreted as insertions or deletions, respectively, according to the absence or presence of non-null nucleotides in the aligned positions in the consensus sequence.

Data

The *Alu*Gene database contains a map of the human transcriptome, and a map and properties of *Alu* sequences in the human genome. The transcriptome is split into three sets (tables): mRNA, intron and exon. The mRNA records are linked to other genetic databases through four different accession keys provided by the NCBI: Refseq ID, GI number, Interim ID and LocusLink ID. In addition, each mRNA record was identified by its genomic location. Intron and exon records were linked to their corresponding mRNA through the GI number.

The data for each *Alu* entry include DNA sequence, genomic location, *Alu* subfamily to which the *Alu* entry belongs, an alignment of the entry to the consensus sequence of the subfamily and a list of differences from the consensus sequence of the subfamily. The *Alu* data also include GC content and length of the poly(A) tail, i.e. features that have been shown to affect the role that *Alu* may play in the genome (16,17). The identification of the poly(A) tail might be problematic since in addition to the terminal poly(A) sequence, *Alu* elements contain an internal poly(A). Thus, as far as partial *Alu* insertions are concerned, the internal poly(A) may be confused with the terminal one. Using pairwise alignments of each *Alu* to its subfamily consensus sequence, we were able to ascertain that the poly(A) at the 3′ of an *Alu* sequence instance is indeed a tail.

It is, of course, not our purpose here to provide an all-exhaustive statistical description of the '*Alu*ome'. In the following, we provide a couple of illustrative enumerative statistics that can be gleaned from *Alu*Gene. The total number of *Alu* elements in the currently sequenced human genome is 1 169 291. Forty five percent (45%) of all *Alu*s are contained within genes; the rest lie within intergenic regions. There are 28 049 transcripts in *Alu*Gene, of which 17 781 (63%) contain at least one *Alu* element. Within 1 kb regions upstream of transcription initiation sites, there are 9212 *Alu*s. These *Alu*s are found in locations that potentially may affect expression levels of the downstream genes.

Search

The *Alu*Gene database can be accessed freely at http://Alugene.tau.ac.il/. Its main goal is to facilitate the search for *Alu*s that are located either within genes or in their immediate proximity. By merging the genomic locations of genes and *Alu*s, it is possible to find overlapping areas between the two, such as *Alu*s residing within exons. Moreover, using additional information concerning *Alu*s, it is possible to study the *Alu* subset that affects processes such as exonization, expression regulation, etc. *Alu*Gene enables specific queries about certain genes or loci, as well as a wide range of queries designed by

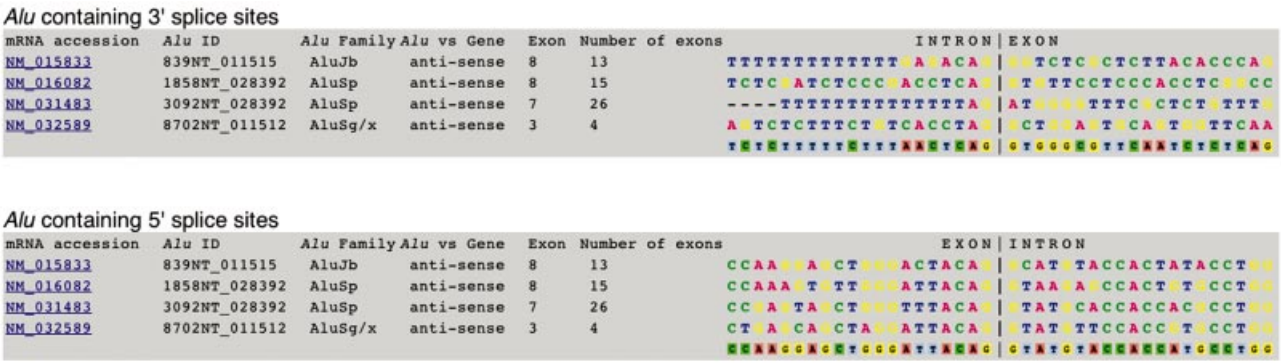


Figure 1. Output of an 'exonization' query. The query was conducted using genes NM_015833, NM_032589 from chromosome 21 and NM_016083, NM_031483 from chromosome 20. *Alu*s containing 3′ splice sites are shown at the top; *Alu*s containing 5′ splice sites are at the bottom. Twenty nucleotides either side of splice sites are shown (the window width was a parameter of the query). The results include (from left to right): accession number, *Alu* ID in *Alu*Gene, *Alu* subfamily, orientation of the *Alu* versus the gene, exon number, total exons in the gene and sequences of *Alu*s at the splice sites. The consensus sequences for each multiple sequence alignment are shown in the bottom-most rows.

the user. The outcome of the specific queries can be either a schematic map, or an alignment of the nucleotide sequences. In the alignment queries, the *Alu*s are aligned either to the genomic locus in which they are embedded or to the consensus sequence of their subfamily. Thus, *Alu*s can be searched by their location in the genome, or by location in genes (in exons, or introns, or in splice sites), and by properties such as GC content and length of A-tail. Moreover, *Alu*Gene enables search of *Alu* by sequence motifs, i.e. retrieve all *Alu*s that contain a certain sequence pattern. The schematic map generated by *Alu*Gene presents the extent and locations of genes and *Alu*s in the area defined by the user query. Viewing extended information about the elements, i.e. exons, *Alu*s, etc., can be done by placing the cursor on the map symbols.

ILLUSTRATIVE EXAMPLES

An interesting option in *Alu*Gene is the 'Do it yourself' query. One such example concerns the observation that *Alu* elements may contain active splice sites (4–6). In such cases, part of the *Alu* element will be found in an intron, and the other part in an exon. Let us use this option, for instance, to identify all *Alu*s that contain splice sites on chromosome 21. The query in SQL can be viewed in the 'Examples' section of the 'Do it yourself' query by clicking on the button '*Alu*s spanning splice-sites in chromosome 21'. In other words, for all the *Alu*s on chromosome 21, we look for those that begin upstream of a splice site (either 5′ or 3′) and end downstream of it.

The result of this query was the identification of four transcripts on chromosome 21 for which evidence for *Alu* exonization exists. Of these, one is annotated as a hypothetical gene with no additional information, and three were identified. These are: (i) NM_015833 (ADARB1), a 13-exon gene coding the enzyme responsible for pre-mRNA editing; (ii) NM_015834, a transcript variant of ADARB1; and (iii) NM_032589 (DSCR8), a 4-exon gene the Down syndrome critical region.

If we repeat the same query for chromosome 20 we obtain three transcripts, one of which is an unannotated open reading frame (ORF) and two are identified. These are: (i) NM_031483 (ITCH), a 26-exon gene encoding a protein that interacts with atrophin-1, and (ii) NM_016082 (CDK5RAP1), a 15-exon gene encoding a neuronal CDC2-like kinase that is involved in the regulation of neuronal differentiation.

The above-described *Alu*s can be used to study sequences that mediate *Alu* exonization. *Alu*Gene provides a specific query type for this purpose, the 'Exonization' alignment query. The result of this query is a splice-site alignment of the *Alu*s in the search criteria. Figure 1 shows such an alignment, produced for the four *Alu*s from the annotated genes in chromosomes 20 and 21. As seen in the alignment of the 3′ splice sites of these *Alu*s (top), an AG dinucleotide always appears at the end of the intron upstream *Alu* exons. Indeed, such an AG pair has recently been reported to be essential for *Alu* exonization (4–6). The alignment of the 5′ splice sites (bottom) also shows several conserved nucleotide positions. For example, the last two nucleotides of the exon (AG), as well as the first, third and fifth positions of the intron (G, A and G, respectively) are conserved in all four *Alu*-derived splice sites. These positions are in agreement with the general consensus of 5′ splice sites (18).

FUTURE DEVELOPMENTS

*Alu*Gene is an ongoing project whose scopes and uses will be extended in the future. One of the first extensions will involve the addition of information concerning gene products. Proteins will be classified according to biochemical pathways or genetic disorders caused by their incapacitation. Such a task may be accomplished by the linking of *Alu*Gene to the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim/). Such a link may be especially useful to study *Alu* insertions that result in pathological manifestation. Currently the exons in *Alu*Gene are derived only from RefSeq, and do not contain exons that are supported solely by EST data. Therefore, an additional expansion may include EST-based exons using dbEST (http://www.ncbi.nlm.nih.gov/dbEST/). Information about *Alu*s within RNA-specifying genes is expected to be included

within *Alu*Gene as well. Other possible additions may include single nucleotide polymorphisms within *Alu*s, as well as a taxonomic expansion into simian and scandentian genomes as these become available. The taxonomic expansion is expected to add v*Alu*able information on the evolution of *Alu* sequences.

ACKNOWLEDGEMENTS

We thank Ran Blekhman, Giddy Landan and Itay Mayrose for their help. T.D. was supported in part by a scholarship in Complexity Science from the Yeshaia Horvitz Association.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Mighell,A.J., Markham,A.F. and Robinson,P.A. (1997) *Alu* sequences. *FEBS Lett.*, **417**, 1–5.
3. Kapitonov,V. and Jurka,J. (1996) The age of *Alu* subfamilies. *J. Mol. Evol.*, **42**, 59–65.
4. Lev-Maor,G., Sorek,R., Shomron,N. and Ast,G. (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in *Alu* exons. *Science*, **300**, 1288–1291.
5. Nekrutenko,A. and Li,W.H. (2001) Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.*, **17**, 619–621.
6. Sorek,R., Ast,G. and Graur,D. (2002) *Alu*-containing exons are alternatively spliced. *Genome Res.*, **12**, 1060–1067.
7. Vervoort,R., Gitzelmann,R., Lissens,W. and Liebaers,I. (1998) A mutation (IVS8+0.6kbdelTC) creating a new donor splice site activates a cryptic exon in an *Alu*-element in intron 8 of the human β-glucuronidase gene. *Hum. Genet.*, **103**, 686–693.
8. Ganguly,A., Dunbar,T., Chen,P., Godmilow,L. and Ganguly,T. (2003) Exon skipping caused by an intronic insertion of a young *Alu* Yb9 element leads to severe hemophilia A. *Hum. Genet.*, **113**, 348–352.
9. Ricci,V., Regis,S., Di Duca,M. and Filocamo,M. (2003) An *Alu*-mediated rearrangement as cause of exon skipping in Hunter disease. *Hum. Genet.*, **112**, 419–425.
10. Hilgard,P., Huang,T., Wolkoff,A.W. and Stockert,R.J. (2002) Translated *Alu* sequence determines nuclear localization of a novel catalytic subunit of casein kinase 2. *Am. J. Physiol. Cell Physiol.*, **283**, C472–C483.
11. Le Goff,W., Guerin,M., Chapman,M.J. and Thillet,J. (2003) A CYP7A promoter binding factor site and *Alu* repeat in the distal promoter region are implicated in regulation of human CETP gene expression. *J. Lipid Res.*, **44**, 902–910.
12. Yi,P., Zhang,W., Zhai,Z., Miao,L., Wang,Y. and Wu,M. (2003) Bcl-rambo β, a special splicing variant with an insertion of an *Alu*-like cassette, promotes etoposide- and Taxol-induced cell death. *FEBS Lett.*, **534**, 61–68.
13. Makalowski,W. (2003) Not junk after all. *Science*, **300**, 1246–1247.
14. Grover,D., Majumder,P.P., Rao,C.B., Brahmachari,S.K. and Mukerji,M. (2003) Non-random distribution of *Alu* elements in genes of various functional categories: Insight from analysis of human chromosomes 21 and 22. *Mol. Biol. Evol.*, **20**, 1420–1424.
15. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
16. Roy-Engel,A.M., Salem,A.H., Oyeniran,O.O., Deininger,L., Hedges,D.J., Kilroy,G.E., Batzer,M.A. and Deininger,P.L. (2002) Active *Alu* element 'A-tails': Size does matter. *Genome Res.*, **12**, 1333–1344.
17. Jurka,J., Krnjajic,M., Kapitonov,V.V., Stenger,J.E. and Kokhanyy,O. (2002) Active *Alu* elements are passed primarily through paternal germlines. *Theor. Popul. Biol.*, **61**, 519–530.
18. Horowitz,D.S. and Krainer,A.R. (1994) Mechanisms for selecting 5′ splice sites in mammalian pre-mRNA splicing. *Trends Genet.*, **10**, 100–106.

3. DISCUSSION

In the course of this thesis I have studied evolutionary and regulatory aspects of alternative splicing and developed novel computational methods to predict alternatively spliced exons. To achieve this I used two approaches in parallel: The first was comparative genomics, and the second was the study of primate specific Alu-derived exons. This section will discuss the results from both these approaches and their significance.

3.1 Conservation of introns flanking alternative exons

We started by aligning human exons to their mouse orthologs, and compared alternatively spliced exons to constitutively spliced ones. We found that intronic sequences that flank alternatively spliced exons are frequently highly conserved between human and mouse. This was surprising because intronic sequences are usually non-functional, and therefore tend to evolve very rapidly between the two lineages.

Following this result we hypothesized that these conserved intronic elements are responsible for regulation of alternative splicing. Indeed, constitutively spliced exons were only rarely flanked by such conserved intronic sequences. Moreover, we were able to locate sequence motifs that were over-abundant in the conserved introns near alternatively spliced exons, but did not appear over-represented near constitutively spliced ones. A literature search showed that at least one of these motifs was indeed documented previously to regulate alternative splicing of specific exons [53]. These results therefore strongly indicate that much of the regulation of internal alternatively spliced exons is carried out through the introns flanking them [36].

The rule that we formulated, of “conservation of flanking introns”, seems as a global feature of alternative splicing. Following our publication, others have shown that this rule also holds for other types of alternative splicing in human, such as alternative 5’ splice sites and alternative 3’ splice sites – where the conservation is restricted to the intron flanking the alternative splice site [17]. Moreover, high conservation of intronic sequences near alternative exons was also reported in insects [37], indicating that this rule widely holds in various metazoan phyla.

3.2 A method for de-novo identification of alternative exons

Apart from the implication on understanding the regulation of alternative splicing, the above results had another important consequence: They revealed a valuable feature that could help us computationally discriminate between alternatively spliced exons and constitutively spliced ones using genome sequences only, and no other experimental data (e.g., ESTs, microarray results). Such discrimination was not possible before. To this end, we discovered some additional features in which alternative exons differ from constitutive exons, such as by their size, their tendency to preserve the frame and more, and found a combination of these features that could classify alternative exons with very high specificity (over 99%) and ~30% sensitivity. This allowed the identification of novel alternatively spliced exons in the human genome, exons that were not known previously to be alternatively spliced. The prediction that these exons are alternative was validated through laboratory experiments [38]. In a follow-up study, jointly with Dr. Gideon Dror, we improved the performance of our classifier by adding more classifying features and employing Support Vector Machine machine-learning principles. This resulted in improved sensitivity of ~50% at the same specificity and higher classifier robustness[39].

Our study was the first to present a reliable method for computational identification of alternatively spliced exons. By that, it revealed hundreds of putative splice variants that could not have been found before. Following our publications, several other groups have employed the approach on different datasets of alternative exons and presented some modifications of our method [37, 40-42].

3.3 Functional versus non-functional alternative splicing

While comparing between human and mouse exons, we also observed that only a minority (~25%) of alternative splicing predicted from EST data is conserved in the mouse genome. This was unexpected, as more than 90% of the human exons are conserved in the mouse genome, and since alternative splicing is thought to be a mechanism that evolved very early in evolution. We therefore hypothesized that some of the alternative splicing predicted from EST databases is in fact aberrant splicing caused by somatic mutations in specific individuals or from experimental errors (e.g., protocol problems in EST production). To show that, we compared between conserved (and hence probably functional) and non-conserved alternative splicing, and showed that the non-conserved type do not share the features that characterize the conserved exons. We concluded that a significant portion of the alternative exons evident in EST databases is not functional, and may result from aberrant rather than regulated splicing [43].

Our publication in 2004, which was since cited more than 40 times in the works of others, has initiated a debate whether the vast amount of splice variants seen in EST data are “real” or spurious. On one hand, some studies report on rapid evolution of alternative splicing between species [44], which could explain the abundance of non-conserved (and yet functional) splice variants. On the other hand, non-conserved

alternative splicing was recently reported to be more prevalent in testis and cancerous cell lines, possibly indicating that it is the result of cellular stress and rapid proliferation and hence not necessarily functional [45].

3.4 Evolution of primate-specific Alu exons

To understand the molecular basis and the regulation of the process turning intronic Alus into new exons (which we term ‘exonization’), we compiled and analyzed a dataset of human exonized Alus. We compared the 3’ splice-sites of these exonized Alus to their consensus sequence, and predicted which nucleotide positions are important for the creation of functional 3’ splice site within Alus. In collaboration with Galit Lev-Maor, we revealed a mechanism that governs 3' splice-site selection in these exons during alternative splicing, and verified our predictions using *in-vivo* model system. On the basis of these findings we identified mutations that activated the exonization of silent intronic Alus. We further showed that such mutations could be the molecular basis for previously unidentified genetic disorders [46].

We then wished to find other nucleotide positions that need to be changed in order to enable Alu exonization. In collaboration with Tal Dagan and Eilon Sharon we compiled a database of all Alus in the human genome, annotated with respect to coding regions and exon/intron location, as well as aligned to their ancestral sequence [47]. We queried this database for silent, intronic Alus that did not become exons, and prepared a profile (PSSM) out of them. This profile was compared to the profile of exonized Alus. Except for the 3’ splice site, the only other positions whose distribution was different between exonized and non-exonized Alus were in the 5’ splice site of the Alu exons, indicating that only the two splice site sequences are important for the birth of new Alu exons.

In collaboration with Galit Lev-Maor and Mika Reznik we revealed a complex mechanism by which the sequence composition of the 5’ splice site and its base pairing with the small-nuclear RNA U1 govern alternative splicing. We showed that in Alu-derived exons the strength of the base-pairing between U1 snRNA and the 5' splice-site controls the skipping/inclusion ratio of alternative splicing [48]. With this, we achieved a rather complete picture of the sequences needed for the creation and regulation of alternatively spliced Alu-exons (see figure 6 in ref [48]).

Combined together, our publications (which were jointly cited over 70 times in the works of others) have revealed the mechanism by which Alu elements become new exons in primate genomes, and highlighted Alu exonization as a major driving force in the evolution of the human lineage. Follow-up studies by others have shown that in certain Alu exons, exonic splicing enhancers and silencers play a regulatory role [49, 50]. In addition, it was shown that transposable elements other than Alu can also contribute to the creation of new exons in other genomes [51, 52].

3.5 Regulation of alternative splicing

We have found that alternatively spliced exons are flanked by conserved sequences that extend, on average, for 100 bases on each side of the exon. We also provided evidence that these conserved sequences are probably involved in regulation of alternative splicing. Our results are consistent with several specific studies, which described long intronic regulatory elements near individual alternative exons. For example, Modafferi and Black [54] showed that for the neuron-specific exon N1 in the mouse c-src gene, an intronic splicing enhancer activity resides between intronic bases 17-142 downstream to the exon. Any part of this sequence was able to reconstitute

splicing enhancing activity only partially. What are the implications of our results on current understanding of regulation of alternative splicing?

Regulation of alternative splicing involves *cis*-acting components, which are sequences found in the pre-mRNA. These are binding sites to splicing regulatory proteins (SR and hnRNP proteins) that modulate the binding of the basal splicing machinery to the appropriate splice sites [55]. Binding of splicing regulatory proteins to their recognition sites on the pre-mRNA can either promote splicing or repress it, depending on the context of the binding site (Ast G., personal communications). Additional levels of regulation, such as signal transduction and cooperation between the transcription and splicing machineries, can also affect patterns of alternative splicing [55].

Most of the *cis*-acting factors which form the recognition sites for splicing regulatory proteins (as described to date) are relatively small, and their sizes range between 4-10 nucleotides [23]. For example, the consensus binding site for the splicing factor SRp55 is URCRUC (where R stands for A or G) [23]; the consensus site for SRp40 is UGCGUC [23]; and the consensus binding site for the brain specific splicing regulator Nova is merely UCAY [56]. In light of this data, how our results, that ~100 bases from each side of most alternatively spliced exons are conserved, can be explained?

First, it is possible that for most of the alternatively spliced exons, multiple regulatory binding sites exist in the nearby introns. Indeed, for several extensively studied exons, it was shown that more than one *cis*-acting factor is involved in their regulation. For example, the alternative splicing of exon 10 in the MAPT gene is regulated by a combination of several cis-acting elements, including an exonic splicing enhancer, an exonic splicing silencer, and an intronic splicing silencer [23]. Our results

might suggest that multiple binding sites regulating the alternative splicing of a single exon is the rule rather than the exception.

In addition, alternative splicing was also shown to be regulated by intronic secondary structures in the pre-mRNA, which can bring *cis*-acting sequence elements closer together, or sequester them [57]. For example, the regulation of the mutually exclusive splicing of exons IIIb and IIIc in FGFR2 depends on stem-forming sequences, IAS2 and ISAR, which bring a distal intronic splicing silencing element in close proximity to exon IIIb, thereby repressing its splicing [57-58]. Secondary structure forming elements in introns are expected to impose additional evolutionary constraints on the intronic sequences, and be another source for the extensive conservation we observed in our study. In this light, a reasonable follow-up for our study might be to search for patterns of RNA secondary structures in the conserved intronic sequences flanking alternatively spliced exons. Results suggesting the existence of such conserved secondary structures near alternatively spliced exons were recently published [59].

Since tissue-specific alternative splicing contributes to the differences in protein variants distributions between different cell types [13-14], the regulation of tissue specific alternative splicing is of particular interest. How is a regulatory signal that is sequence based used differently in different tissues? The answer for this probably lies in the differential expression of the *trans*-acting splicing regulatory proteins. While the *cis*-acting sequence elements reside in all tissues in which the specific pre-mRNA is transcribed, a splicing regulatory protein which is tissue specific will only bind to the pre-mRNA in that tissue (sometimes only following a specific posttranslational modification such as phosphorylation), thereby driving tissue-specific alternative splicing. For example, the brain specific protein Nova is a regulator of alternative

splicing, which binds to a consensus sequence of UCAY [56]. This protein regulates the neuron-specific alternative splicing of >30 different proteins, most of which having specific roles in the synapse [60]. Binding sites of Nova can occur within the alternative exon or in the flanking intron, and Nova binding can either activate or repress exon inclusion, depending on the context [56].

An additional example is of the brain-and-muscle-specific protein Fox-1, which was shown to specifically bind the RNA hexamer UGCAUG [61]. In our study, this hexamer was found to be over-represented in the conserved intronic elements downstream to alternatively spliced exons, suggesting abundance of Fox-1 binding sites in these conserved introns [36]. Interestingly, a recent study reported that the element UGCAUG is over-represented near brain-specific exons that are conserved from fish to mammals [62]. Moreover, the position and sequence context of intronic UGCAUG elements were also highly conserved, suggesting that Fox-1 is a brain-specific regulator of alternative splicing [62]. This might indicate that many of the alternatively spliced exons in our set undergo brain-specific alternative splicing.

As detailed above, tissue-specific regulators of alternative splicing will attach to the pre-mRNA only in the tissue where they are expressed. This notion has implications on the way experimental alternative splicing studies are interpreted. For example, gel-shift assays and RNA-protein cross-linking approaches can be used to determine the regions in the pre-mRNA on which proteins attach [56, 63]. Such regions might be regarded as the functional *cis*-acting elements in the pre-mRNA, and regions on which proteins are not found to bind might be considered as *cis*-acting-elements free. However, protein-binding regions depend, as indicated above, in the tissue/cell-type in which the experiment was done. Therefore, intronic regions that are not found to be functional/protein-binding in one tissue could still be functional in regulating

alternative splicing in another tissue or at a specific developmental stage. This “tissue-splicing code” might be an additional explanation for the extensive conservation we observed in introns near alternatively spliced exons – perhaps these long intronic regulatory elements are composed of a mosaic of tissue-specific *cis*-acting elements, each of them coming into action in a different tissue or developmental stage. The exact nature of these elements remains to be discovered.

Summary

The initial goal of my research was to understand how alternative splicing is being regulated. Indeed, both the computational genomics and the Alu-based approaches led to the identification of sequences that regulate alternative splicing. Both approaches also yielded unanticipated results, such as the invention of a novel method to predict alternative splicing events and the characterization of the Alu-exonization-based mechanism underlying several human genetic disorders. Altogether, these results demonstrate the strength of the combination between computational and experimental biology to study complex biological phenomena.

4. REFERENCES

1. Newman, A., *RNA splicing.* Current Biology, 1998. **8**(25): R903-905.
2. Makalowski, W., Mitchell, G.A. and Labuda, D., *Alu sequences in the coding regions of mRNA: a source of protein variability.* Trends in Genetics, 1994. **10**(6): 188-193.
3. Zhou, Z., Licklider, L.J., Gygi, S.P. and Reed, R., *Comprehensive proteomic analysis of the human spliceosome.* Nature, 2002. **419**(6903): 182-185.
4. Muller, S., Wolpensinger, B., Angenitzki, M., Engel, A., Sperling, J. and Sperling, R., *A supraspliceosome model for large nuclear ribonucleoprotein particles based on mass determinations by scanning transmission electron microscopy.* Journal of Molecular Biology, 1998. **283**(2): 383-394.
5. Chabot, B., *Directing alternative splicing: cast and scenarios.* Trends in Genetics, 1996. **12**(11): 472-478.
6. Blencowe, B.J., *Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases.* Trends in Biochemical Sciences, 2000. **25**(3): 106-110.
7. Reed, R., *Mechanisms of fidelity in pre-mRNA splicing.* Current Opinion in Cell Biology, 2000. **12**(3): 340-345.
8. International Human Genome Sequencing Consortium, *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): 860-921.
9. International Mouse Genome Sequencing Consortium. *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): 520-562.

10. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J. and Bork, P., *EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.* FEBS Letters, 2000. **474**(1): 83-86.

11. Mironov, A.A., Fickett, J.W. and Gelfand, M.S., *Frequent alternative splicing of human genes.* Genome Research, 1999. **9**(12): 1288-1293.

12. Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P. and Mattick, J.S., *ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome.* Nature Genetics, 2000. **24**(4): 340-341.

13. Maniatis, T. and Tasic, B., *Alternative pre-mRNA splicing and proteome expansion in metazoans.* Nature, 2002. **418**(6894): 236-243.

14. Black, D.L., *Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology.* Cell, 2000. **103**(3): 367-370.

15. Lopez, A.J., *Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation.* Annual Review of Genetics, 1998. **32**: 279-305.

16. McKeown, M., *Alternative mRNA splicing.* Annual Review of Cell Biology, 1992. **8**: 133-155.

17. Sugnet, C.W., Kent, W.J., Ares, M., Jr. and Haussler, D., *Transcriptome and genome conservation of alternative splicing events in humans and mice.* Pacific Symposium in Biocomputing, 2004. **1**(1): 66-77.

18. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F. et al., *Complementary DNA sequencing: expressed sequence tags and human genome project.* Science, 1991. **252**(5013): 1651-1656.

19. Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M., *dbEST--database for "expressed sequence tags".* Nature Genetics, 1993. **4**(4): 332-333.

20. Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. and Shoemaker, D.D., *Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.* Science, 2003. **302**(5653): 2141-2144.

21. Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D., Frey, B.J. and Blencowe, B.J., *Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform.* Molecular Cell, 2004. **16**(6): 929-941.

22. Tacke, R. and Manley, J.L., *Determinants of SR protein specificity.* Curr Opinion in Cell Biology, 1999. **11**(3): 358-362.

23. Cartegni, L., Chew, S.L. and Krainer, A.R., *Listening to silence and understanding nonsense: exonic mutations that affect splicing.* Nat Reviews Genetics, 2002. **3**(4): 285-298.

24. Fairbrother, W.G., Yeh, R.F., Sharp, P.A. and Burge, C.B., *Predictive identification of exonic splicing enhancers in human genes.* Science, 2002. **297**(5583): 1007-1013.

25. Cramer, P., Caceres, J.F., Cazalla, D., Kadener, S., Muro, A.F., Baralle, F.E. and Kornblihtt, A.R., *Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer.* Molecular Cell, 1999. **4**(2): 251-258.

26. Modrek, B. and Lee, C., *A genomic view of alternative splicing.* Nature Genetics, 2002. **30**(1): 13-19.

27. Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B. and Lander, E.S., *Human and mouse gene structure: comparative analysis and application to exon prediction.* Genome Research, 2000. **10**(7): 950-958.

28. Hardison, R.C., Oeltjen, J. and Miller, W., *Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome.* Genome Research, 1997. **7**(10): 959-966.

29. Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., Flegel, V., Bucher, P., Jongeneel, C.V. and Antonarakis, S.E., *Numerous potentially functional but non-genic conserved sequences on human chromosome 21.* Nature, 2002. **420**(6915): 578-582.

30. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D., *Ultraconserved elements in the human genome.* Science, 2004. **304**(5675): 1321-1325.

31. Mighell, A.J., Markham, A.F. and Robinson, P.A., *Alu sequences.* FEBS Letters, 1997. **417**(1): 1-5.

32. Rowold, D.J. and Herrera, R.J., *Alu elements and the human genome.* Genetica, 2000. **108**(1): 57-72.

33. Schmid, C.W., *Alu: structure, origin, evolution, significance and function of one- tenth of human DNA.* Progress in Nucleic Acid Research Molecular Biology, 1996. **53**: 283-319.

34. Roy-Engel, A.M., Salem, A.H., Oyeniran, O.O., Deininger, L., Hedges, D.J., Kilroy, G.E., Batzer, M.A. and Deininger, P.L., *Active Alu element "A-tails": size does matter.* Genome Research, 2002. **12**(9): 1333-1344.

35. Sorek, R., Ast, G. and Graur, D., *Alu-containing exons are alternatively spliced.* Genome Research, 2002. **12**(7): 1060-1067.

36. Sorek, R. and Ast, G., *Intronic sequences flanking alternatively spliced exons are conserved between human and mouse.* Genome Research, 2003. **13**(7): 1631-1637.

37. Philipps, D.L., Park, J.W. and Graveley, B.R., *A computational and experimental approach toward a priori identification of alternatively spliced exons.* RNA, 2004. **10**(12): 1838-1844.

38. Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G. and Shamir, R., *A non-EST-based method for exon-skipping prediction.* Genome Research, 2004. **14**(8): 1617-1623.

39. Dror, G., Sorek, R. and Shamir, R., *Accurate identification of alternatively spliced exons using support vector machine.* Bioinformatics, 2005. **21**(7): 897-901.

40. Hiller, M., Huse, K., Platzer, M. and Backofen, R., *Non-EST based prediction of exon skipping and intron retention events using Pfam information.* Nucleic Acids Research, 2005. **33**(17): 5611-5621.

41. Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T. and Burge, C.B., *Identification and analysis of alternative splicing events conserved in human and mouse.* Proceedings of the National Academy of Science U S A, 2005. **102**(8): 2850-2855.

42. Ratsch, G., Sonnenburg, S. and Scholkopf, B., *RASE: recognition of alternatively spliced exons in C.elegans.* Bioinformatics, 2005. **21** (Suppl 1): i369-i377.

43. Sorek, R., Shamir, R. and Ast, G., *How prevalent is functional alternative splicing in the human genome?* Trends in Genetics, 2004. **20**(2): 68-71.

44. Pan, Q., Bakowski, M.A., Morris, Q., Zhang, W., Frey, B.J., Hughes, T.R. and Blencowe, B.J., *Alternative splicing of conserved exons is frequently species-specific in human and mouse.* Trends in Genetics, 2005. **21**(2): 73-77.

45. Kan, Z., Garrett-Engele, P.W., Johnson, J.M. and Castle, J.C., *Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles.* Nucleic Acids Research, 2005. **33**(17): 5659-5666.

46. Lev-Maor, G., Sorek, R., Shomron, N. and Ast, G., *The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons.* Science, 2003. **300**(5623): 1288-1291.

47. Dagan, T., Sorek, R., Sharon, E., Ast, G. and Graur, D., *AluGene: a database of Alu elements incorporated within protein-coding genes.* Nucleic Acids Research, 2004. **32**(Database issue): D489-492.

48. Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D. and Ast, G., *Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons.* Molecular Cell, 2004. **14**(2): 221-231.

49. Lei, H., Day, I.N. and Vorechovsky, I., *Exonization of AluYa5 in the human ACE gene requires mutations in both 3' and 5' splice sites and is facilitated by a conserved splicing enhancer.* Nucleic Acids Research, 2005. **33**(12): 3897-3906.

50. Lei, H. and Vorechovsky, I., *Identification of splicing silencers and enhancers in sense Alus: a role for pseudoacceptors in splice site repression.* Molecular and Cellular Biology, 2005. **25**(16): 6912-6920.

51. Wang, W., Zheng, H., Yang, S., Yu, H., Li, J., Jiang, H., Su, J., Yang, L., Zhang, J., McDermott, J., Samudrala, R., Wang, J., Yang, H., Yu, J., Kristiansen, K., Wong, G.K.S. and Wang, J., *Origin and evolution of new exons in rodents.* Genome Research, 2005. **15**(9): 1258-1264.

52. Wang, W. and Kirkness, E.F., *Short interspersed elements (SINEs) are a major source of canine genomic diversity.* Genome Research, 2005. **15**(12): 1798-1808.

53. Lim, L.P. and Sharp, P.A., *Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats*. Molecular and Cellular Biology, 1998. **18**: 3900–3906.

54. Modafferi, E.F. and Black, D.L., *A complex intronic splicing enhancer from the c-src pre-mRNA activates inclusion of a heterologous exon.* Molecular and Cellular Biology, 1997. **17**(11): 6537-6545

55. Matlin, A.J., Clark, F. and Smith, C.W., *Understanding alternative splicing: towards a cellular code.* Nature Reviews Molecular Cell Biology, 2005. **6**(5): 386-398.

56. Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A. and Darnell, R.B., *CLIP identifies Nova-regulated RNA networks in the brain.* Science, 2003. **302**(5648):1212-1215.

57. Buratti, E. and Baralle, F. E., *Influence of RNA secondary structure on the pre-mRNA splicing process.* Molecular and Cellular Biology, 2004. **24**: 10505–10514.

58. Graveley, B.R., *Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures.* Cell, 2005. **123**(1): 65-73.

59. Miriami, E., Margalit, H. and Sperling, R. *Conserved sequence elements associated with exon skipping.* Nucleic Acids Research, 2003. **31**(7): 1974-1983.

60. Ule, J. et al. *Nova regulates brain-specific splicing to shape the synapse.* Nature Genetics, 2005, **37**(8): 844-852.

61. Nakahata, S. and Kawamoto, S. *Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities.* Nucleic Acids Research, 2005. **33**(7): 2078-89

62. Minovitsky, S., Gee, S.L., Schokrpur, S., Dubchak, I. and Conboy, J.G. *The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons.* Nucleic Acids Research, 2005. **33**(2): 714-24.

63. Moraes, K.C., Quaresma, A.J., Maehnss, K. and Kobarg J., *Identification and characterization of proteins that selectively interact with isoforms of the mRNA binding protein AUF1 (hnRNP D).* Biological chemistry, 2003. **384**(1):25-37.