Am. J. Hum. Genet. 67:788–792, 2000

INVITED EDITORIAL
Genomic Sequence, Splicing, and Gene Annotation

Stephen M. Mount

Department of Cell Biology and Molecular Genetics, University of Maryland, College Park

Introduction

The sequence of the human genome is at hand. Most scientists who use the sequence will rely on annotations that provide information about the number and location of genes and about their inferred protein products. Traditionally, genes have been annotated by scientists with a particular interest in them. However, annotation of the complete human genome sequence will have to be at least partially automated. Gene annotation incorporates cDNA data (including expressed sequence tags [ESTs]), sequence similarity, and computational predictions based on the recognition of probable splice sites and coding regions (Stormo 2000; also see David Haussler's Web site, Computational Genefinding). The state of the art was recently surveyed by the Genome Annotation Assessment Project-GASP1 and must be regarded as imperfect (Bork 2000; Reese et al. 2000).

This review enumerates aspects of pre-mRNA splicing that limit our ability to predict gene structure from genomic sequence, drawing on the recently annotated complete genome of *Drosophila melanogaster* (Adams et al. 2000) as an example. In particular, the following four facts will be discussed. First, splice sites do not always conform to consensus. Second, noncoding exons are common. Third, internal exons can be arbitrarily small, and small internal exons confound not only gene finding but also the alignment of cDNA and genomic sequences. Fourth, splice sites are not recognized in isolation, and nucleotides that are far from splice sites can affect splicing. This list and the accompanying analysis should make molecular geneticists aware of the ways in which gene annotations can be wrong and should encourage recourse to the primary data. In addition, the same considerations indicate that inherited disease can be caused by mutations remote from splice sites that nevertheless affect splicing.

Received August 3, 2000; accepted for publication August 15, 2000; electronically published September 8, 2000.

Address for correspondence and reprints: Dr. Stephen M. Mount, Department of Cell Biology and Molecular Genetics, H. J. Patterson Hall, University of Maryland, College Park, MD 20742-5815. E-mail: sm193@umail.umd.edu

This article represents the opinion of the author and has not been peer reviewed.

0002-9297/2000/6704-0003$02.00

Discussion

Splice Sites Do Not Always Conform to Consensus

It is well established that nearly all splice sites conform to consensus sequences (Mount 1982; Senapathy et al. 1990; Zhang 1998). These consensus sequences include nearly invariant dinucleotides at each end of the intron—GT at the 5′ end of the intron and AG at the 3′ end of the intron. Most gene-finding software and most human annotators will find only introns that begin with a GT and end with an AG. However, nonconsensus splice sites have been described, and I will discuss three classes, in decreasing order of frequency.

The most common class of nonconsensus splice sites consists of 5′ splice sites with a GC dinucleotide. Senapathy et al. (1990) listed 17 examples among 3,724 5′ splice sites, suggesting a frequency of ~0.5%. Jackson (1991) listed a total of 26 GC sites, whereas Wu and Krainer (1999) cited an additional 18 examples. GC 5′ splice sites are consistent with the experimental observation that, of the six possible point mutations within the GT dinucleotide, mutation of T to C in position 2 has the smallest effect on in vitro splicing (Aebi et al. 1986). At other positions within the consensus, GC sites conform extremely well to the standard consensus; for example, 42 of the 44 sites cited above have a consensus G residue at both position −1 and position +5. It is reasonable to assume that GC sites are recognized by the standard (U2-dependent) spliceosome.

The second class of exception to splice-site consensus is U12 introns, a minor class of rare introns with splice-site sequences that are very different from the standard consensus but that are very similar to each other. The existence of this class was first pointed out by Jackson (1991) and was considered in more detail by Hall and Padgett (1994). It was subsequently discovered that U12 introns are removed by a minor spliceosome containing the rare U11, U12, U4atac, and U6atac snRNPs, in place of U1, U2, U4, and U6 (Tarn and Steitz 1997; Burge et al. 1998). Some U12 introns have AT and AC in place of GT and AG and are known as "AT-AC" introns. However, terminal intron dinucleotide sequences do not

distinguish between U2- and U12-dependent introns (Dietrich et al. 1997). Rather, U12 introns can be identified by highly conserved sequences at the 5′ splice site (RTATCCTY; R = A or G, and Y = C or T) and branch site (TCCTRAY). U12 introns are found in many eukaryotes, including *Drosophila melanogaster* (Adams et al. 2000) and *Arabidopsis thaliana* (Shukla and Padgett 1999) but not *Caenorhabditis elegans*.

Finally, there are a small number of nonconsensus sites that fit into neither of the two categories mentioned above. Many reports of such variant splice sites can be traced to errors in annotation or interpretation, polymorphic differences between the sources of cDNA and genomic sequence, inclusion of pseudogene sequences, or failure to account for somatic mutation (author's unpublished data; for examples, see Jackson 1991). However, there are many examples of sites that match the consensus very poorly, and experimental work has established that 5′ splice sites do not absolutely require GT—and that 3′ splice sites do not absolutely require AG—in order to be recognized in vivo (Aebi et al. 1986; Roller et al. 2000, and references therein). In yeast, an intron that is within the HAC1 mRNA and that has no similarity to the standard nuclear pre-mRNA intron consensus sequence is spliced by a specific, regulated, endonuclease and tRNA ligase (Sidrauski et al. 1996). This intron provides a precedent for introns in protein-coding genes with completely novel splice sites.

Noncoding Exons Are Common

There is considerable confusion between exons and coding regions. The term "exon" was coined by Gilbert (1978) to refer to what is left when introns are removed by splicing, and RNAs that are entirely noncoding (such as tRNAs) are sometimes spliced. However, the term exon is often misused to refer to a stretch of coding information. In reality, however, noncoding exons are quite common, occurring in >35% of human genes (Zhang 1998). Gene-finding software generally detects sequence features characteristic of coding regions rather than of exons and does not even attempt to identify noncoding exons, or noncoding portions of exons. This is because the statistical biases introduced by protein-coding are in fact a very powerful tool for the identification of coding DNA, and no similar tool has been developed for the identification of noncoding exons.

A similar problem can arise in genes without noncoding exons. If the first intron occurs near the initiator AUG, then the coding information in the first exon can be very short and difficult to identify by measures of coding tendency. Furthermore, the first intron tends to be longer than average (Maroni 1996), and such an arrangement can separate promoter function (perhaps including downstream transcriptional enhancer elements lying in the first intron) from the bulk of the coding information downstream. In these cases, investigators have no way of knowing how much information is missing—or where the 5′ end of the gene is likely to reside—without experimental data such as a cDNA sequence or a 5′ EST.

Internal Exons Can Be Arbitrarily Small

A less frequent but perhaps more serious problem for gene-discovery methods is posed by small internal exons. Vertebrate internal exons have an average size of ~130 nucleotides (Hawkins 1988; Zhang 1998), and roughly 65% of internal human exons are 68–208 nucleotides in length (Maroni 1996). This size distribution reflects a functional constraint. Optimal splicing efficiency requires exons with sizes of ~50–300 nucleotides (Robberson et al. 1990; Dominski and Kole 1991; see review by Berget 1995). However, a considerable number, >10%, of exons are <60 nucleotides in length, and it is these exons that can be difficult to identify by measures of coding tendency.

Just how small can internal exons be? There appears to be no lower limit, and many cases of exons <10 nucleotides have been described (for examples, see Stamm et al. 1994; also see the author's Web site, Gene Annotation and Splice Site Selection). An illustrative case is the *invected* gene of *D. melanogaster* (also listed in GadFly as CG17835). This gene encodes a homeodomain protein that is similar to *engrailed*, and these two genes are adjacent. One of four *invected* exons is only 6 nucleotides long and is flanked by introns of 27,659 and 1,134 nucleotides. Significantly, this exon is not recognized by cDNA alignment software such as SIM4 (Florea et al. 1998), and the gene is incorrectly annotated (GenBank accession number AE003825.1). As a result, the protein sequence predicted by annotation of the genome (Adams et al. 2000; GenBank accession number AAF58640) differs from that predicted from the cDNA (Coleman et al. 1987; GenBank accession number CAA28885), because of a frameshift affecting the entire carboxyl-terminal coding exon, a highly conserved region of the protein. This is despite the fact that the microexon sequence, GTCGAA, is flanked by intron sequences that perfectly match the splice-site consensus. Use of this microexon provides perfect agreement between the cDNA and genomic sequences when consensus splice sites are used, whereas the annotation predicts an RNA with several discrepancies relative to the cDNA. The frameshift is due to the predicted use of a 5′ splice site 10 nucleotides downstream of the true 5′ splice site, which was apparently selected to account for the microexon. It seems clear that the protein sequence predicted by the cDNA is correct. Why was it not incorporated into the annotation? The alignment problem

arises because a pattern-matching algorithm that locates exons by similarity between the cDNA and the genomic sequence cannot find exons of this size unless its stringency is reduced to an unacceptable level (Florea et al. 1998).

The notion that exons can be arbitrarily small is supported by the observation of exons with length 0. Of course, such sites are not exons at all but, rather, are resplicing sites (see fig. 1). This phenomenon has been demonstrated in the case of the *Drosophila Ultrabithorax* locus (Hatton et al. 1998), which has a region of 60 kb containing two alternatively spliced exons, and may be a general feature of long introns (J. Burnette and A. J. Lopez, personal communication). The existence of resplicing sites not only illustrates the lack of a lower limit to exon size (which has implications for gene annotation) but also has implications for the analysis of hereditary mutations. A mutation at one of these sites could potentially create a frozen intermediate such as that diagrammed in figure 1. This partially spliced RNA would probably be unstable, because of nonsense-mediated decay (Culbertson 1999), and the apparent result would be no RNA (rather than aberrantly spliced RNA). Such mutations would be very hard to identify.

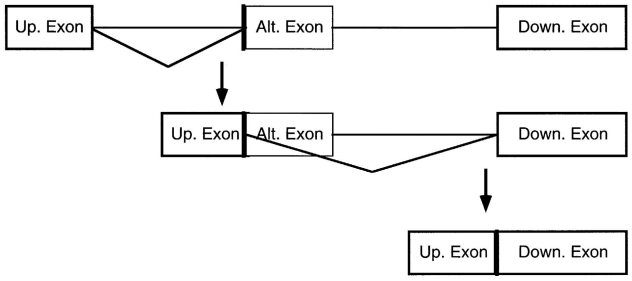

Figure 1 Small internal exons and resplicing. This schematic figure indicates the pathway of resplicing demonstrated for the *Drosophila Ubx* locus (Hatton et al. 1998). The thicker vertical line indicates a resplicing site, which does not contribute any nucleotides to the final mRNA product. The same pathway could be followed in the case of a microexon, in which case an arbitrarily small number of nucleotides would remain in the mRNA product. "Up. Exon" and "Down. Exon" denote the exons upstream and downstream of the resplicing site, respectively. In the case of *Ubx*, the sequence immediately downstream of the resplicing site is an alternatively spliced exon (here designated "Alt. Exon"), but resplicing sites are not always accompanied by such alternatively spliced exons (J. Burnette and A. J. Lopez, personal communication).

Nucleotides Far from Splice Sites Can Affect Splicing

No method of evaluating potential splice sites that is based on sequence alone can be 100% reliable. One can be sure of this because many sequences that are not splice sites are capable of acting as splice sites, and vice versa. Perhaps the clearest demonstration of this is provided by the activation of cryptic splice sites. These are splice sites that are used, sometimes with 100% efficiency, when a natural splice site has been mutationally inactivated. The activation of cryptic sites occurs in approximately one-third of splicing mutations (Nakai and Sakamoto 1994). The phenomenon shows that the cryptic sites are perfectly capable of being recognized by the splicing machinery. Clearly, the sequence of such cryptic sites is compatible with splicing, and context is important for splice-site choice.

Two contextual elements that contribute to splice-site selection are the location of splice sites relative to each other and splicing-enhancer sequences. The exon-size preferences described above are widely understood in terms of an exon-definition model that includes the interaction of splicing factors bound at either end of an exon (Berget 1995). The requirement for productive interactions among splicing factors, including U1 snRNPs at the 5′ splice site and U2 snRNP auxiliary factor (U2AF) at the 3′ splice site, are thought to give rise to preferred exon lengths because of steric constraints and geometry favoring interactions. In the case of small introns, a similar model of intron bridging has been proposed (Guo and Mount 1995; McCullough and Berget 1997). In combination, these models suggest that, in order to be recognized, a splice site must have a partner an appropriate distance away, so that either exon definition or intron definition is facilitated by the spacing. One experimental distinction between exon definition and intron definition is the result of mutations that inactivate the splice site. Failure to undergo exon definition results in exon skipping, whereas failure to undergo intron definition results in intron retention.

Not only is the use of one splice site dependent on the presence of its partner across the exon, but weakness in one partner can be compensated by strength in the other, as seen with second-site revertants of splice-site mutations that cause exon skipping. In an analysis of splicing mutations at the dihydrofolate reductase locus, Carothers et al. (1993) found that a mutation at the 5′ splice site of exon 5 (G to C in the third position of the intron) could be partially reversed by mutations that increased the strength of the 3′ splice site upstream of the same exon (AAAG| to TTAG|, ACAG|, or ATAG|). Although reversion was not complete, these data provide a strong argument that whether a sequence functions as a splice site depends not only on its intrinsic strength but also on its context. Similarly, there are mutations that create splice sites within introns, activating cryptic exons by recruitment of appropriately placed partners (e.g., see Bagnall et al. 1999).

Splicing enhancers are sequences that stimulate splicing at nearby sites. A family of non-snRNP splicing factors known as "SR proteins" appear to be important for the recognition of splicing enhancers in

exons (Blencowe 2000). A splicing difference between SMN1 and SMN2, which explains their differential effects on spinal muscular atrophy, has been attributed to a translationally silent substitution within the coding sequence that affects splicing (Lorson et al. 1999). Similarly, H.-X. Liu, L. Cartegni, M. Q. Zhang, and A. R. Krainer (personal communication) have shown that a nonsense mutation causing the skipping of BRCA1 exon 18 affects splicing in vitro and that a missense mutation at the same position can also cause exon skipping. There are also splicing-enhancer sequences in introns—and examples of mutations that affect them (Cogan et al. 1997). Although general mechanisms for their function have yet to be defined, there is some evidence that at least some splicing enhancers in introns may act by facilitating exon definition in the case of small exons (Carlo et al. 2000).

Outlook

This review has presented aspects of pre-mRNA splicing that pose special problems for gene annotation. However, even though the best gene finders predict genes exactly right less than half the time, 95% of total coding nucleotides are predicted accurately, and <5% of genes are completely missed (Reese et al. 2000; Genome Annotation Assessment Project-GASP1). When cDNA and homology data are available, annotations will tend to be even better. Thus, one would be wrong to conclude from this review that the gene annotations attending the human genome sequence will not provide an extremely valuable resource. Nevertheless, molecular geneticists will want to have an understanding of the kinds of errors that are likely to occur—and to carefully review the available evidence for genes that matter to them. Annotators are likewise obligated to make the source of each specific aspect of their annotation an integral part of the annotation; for example, if part of the annotation is supported by a EST whereas the rest of it is based on the prediction of a gene finder, then the limits of the cDNA should be indicated, and the accession number of the EST should be part of the annotation.

A related but distinct point is that these same factors are also relevant when candidate mutations are evaluated during the analysis of hereditary disease. Mutations that lie within splicing enhancers, at resplicing sites, or at cryptic splice sites can affect splicing even when they lie some distance from the splice sites actually used in the generation of the affected mRNA. The problem is further compounded by alternative splicing and the interplay between splicing and polyadenylation, topics that are beyond the scope of the present review.

In summary, gene annotations will be a valuable resource. However, they will not substitute for expertise in molecular genetics.

Acknowledgments

Support by National Institutes of Health grant GM37991-11 is gratefully acknowledged. I thank Doug Black for helpful comments on the manuscript. I thank James Burnette, A. Javier Lopez, and Adrian Krainer for providing information prior to publication.

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

Computational Genefinding, http://www.cse.ucsc.edu/~haussler/genefindingpaper

GadFly: Genome Annotation Database of Drosophila, http://www.fruitfly.org/annot/index.html

GenBank, http://www.ncbi.nlm.nih.gov/ (for incorrect annotation of *invected* [accession number AE003825.1] and predicted protein sequence [accession numbers AAF58640 and CAA28885])

Gene Annotation and Splice Site Selection, http://www.wam.umd.edu/~smount/Annotation.html

Genome Annotation Assessment Project-GASP1, http://www.fruitfly.org/GASP1/index.html

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, et al (2000) The genome sequence of Drosophila melanogaster. Science 287:2185–2195

Aebi M, Hornig H, Padgett RA, Reiser J, Weissman C (1986) Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. Cell 47:555–565

Bagnall RD, Waseem NH, Green PM, Colvin B, Lee C, Giannelli F (1999) Creation of a novel donor splice site in intron 1 of the factor VIII gene leads to activation of a 191 bp cryptic exon in two haemophilia A patients. Br J Haematol 107:766–771

Berget SM (1995) Exon recognition in vertebrate splicing. J Biol Chem 270:2411–2414

Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25:106–110

Bork P (2000) Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res 10:398–400

Burge CB, Padgett RA, Sharp PA (1998) Evolutionary fates and origins of U12-type introns. Mol Cell 2:773–785

Carlo T, Sierra R, Berget SM (2000) A 5′ splice site-proximal enhancer binds SF1 and activates exon bridging. Mol Cell Biol 20:3988–3995

Carothers AM, Urlaub G, Grunberger D, Chasin LA (1993) Splicing mutants and their second-site suppressors at the dihydrofolate reductase locus in Chinese hamster ovary cells. Mol Cell Biol 13:5085–5098

Cogan JD, Prince MA, Lekhakula S, Bundrey S, Futrakul A, McCarthy EM, Phillips JA III (1997) A novel mechanism of aberrant pre-mRNA splicing in humans. Hum Mol Genet 6:909–912

Coleman KG, Poole SJ, Weir MP, Soeller WC, Kornberg T

(1987) The invected gene of Drosophila: sequence analysis and expression studies reveal a close kinship to the engrailed gene. Genes Dev 1:19–28

Culbertson MR (1999) RNA surveillance: unforeseen consequences for gene expression, inherited genetic disorders and cancer. Trends Genet 15:74–80

Dietrich RC, Incorvaia R, Padgett RA (1997) Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol Cell 1:151–160

Dominski Z, Kole R (1991) Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol 11:6075–6083

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA. Genome Res 8:967–974

Gilbert W (1978) Why genes in pieces? Nature 271:501

Guo M, Mount SM (1995) Localization of sequences required for size-specific splicing of a small *Drosophila* intron. J Mol Biol 253:426–437

Hall SL, Padgett RA (1994) Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J Mol Biol 239:357–365

Hatton AR, Subramaniam V, Lopez AJ (1998) Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol Cell 2:787–796

Hawkins JD (1988) A survey on intron and exon lengths. Nucleic Acids Res 16:9893–9905

Jackson IJ (1991) A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res 19:3795–3798

Lorson CL, Hahnen E, Androphy EJ, Wirth B (1999) A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Proc Natl Acad Sci USA 96:6307–6311

Maroni G (1996) The organization of eukaryotic genes. Evol Biol 29:1–19

McCullough AJ, Berget SM (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol Cell Biol 17:4562–4571

Mount SM (1982) A catalogue of splice junction sequences. Nucleic Acids Res 10:459–472

Nakai K, Sakamoto H (1994) Construction of a novel database containing aberrant splicing mutations of mammalian gene. Gene 141:171–177

Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE (2000) Genome annotation assessment in Drosophila melanogaster. Genome Res 10:483–501

Robberson BL, Cote GL, Berget SM (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol 10:84–94

Roller AB, Hoffman DC, Zahler AM (2000) The allele-specific suppressor sup-39 alters use of cryptic splice sites in Caenorhabditis elegans. Genetics 154:1169–1179

Senapathy P, Sharpiro MB, Harris NL (1990) Splice junctions, branch point sites and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol 183:252–278

Shukla GC, Padgett RA (1999) Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. RNA 5:525–538

Sidrauski C, Cox JS, Walter P (1996) tRNA ligase is required for regulated mRNA splicing in the unfolded protein response. Cell 87:405–413

Stamm S, Zhang MQ, Marr TG, Helfman DM (1994) A sequence compilation and comparison of exons that are alternatively spliced in neurons. Nucleic Acids Res 22:1515–1526

Stormo GD (2000) Gene-finding approaches for eukaryotes. Genome Res 10:394–397

Tarn WY, Steitz JA (1997) Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. Trends Biochem Sci 22:132–137

Wu A, Krainer AR (1999) AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. Mol Cell Biol 19:3225–3236

Zhang MQ (1998) Statistical features of human exons and their flanking regions. Hum Mol Genet 7:919–932