Published in final edited form as:

Adv Exp Med Biol. 2007 ; 623: 50–63.

Complex Alternative Splicing

Jung Woo Park and Brenton R. Graveley

Department of Genetics and Developmental Biology University of Connecticut Health Center 263 Farmington Avenue Farmington, CT 06030-3301

Abstract

Alternative splicing is a powerful means of controlling gene expression and increasing protein diversity. Most genes express a limited number of mRNA isoforms, but there are several examples of genes that use alternative splicing to generate hundreds, thousands, and even tens of thousands of isoforms. Collectively such genes are considered to undergo complex alternative splicing. The best example is the *Drosophila* Down syndrome cell adhesion molecule (Dscam) gene, which can generate 38,016 isoforms by the alternative splicing of 95 variable exons. In this review, we will describe several genes that use complex alternative splicing to generate large repertoires of mRNAs and what is known about the mechanisms by which they do so.

Introduction

Alternative splicing affords eukaryotes with the opportunity to produce multiple proteins from a single gene¹⁻³. This allows organisms to maximize the coding capacity of their genomes. To illustrate this, let us consider two different ways in which evolution can produce two highly related proteins starting from a single gene. Take, for instance, a representative human gene that contains ~10 exons, produces a single mRNA isoform, and encompasses ~28,000 bp^{4,5}. At least two different scenarios can give rise to a new variant of the encoded protein that contains an additional 10 amino acids. First, the gene could be duplicated and diverge such that either a pre-existing exon is extended by 30 nucleotides or a new 30 nt exon is created. Alternatively, a single 30 nucleotide cassette exon could be inserted into, or arise within an intron in the original gene allowing for two mRNA isoforms that either contain or lack the exon to now be produced. While these two scenarios have a similar outcome - the production of a new protein 10 amino acids longer than the original protein - they can have drastically different consequences on the size of the genome. Gene duplication requires expanding the genome by at least 28,000 bp. In contrast, creating the same protein by simply adding an alternative exon to the original gene would only increase the size of the genome by ~30 nt. The difference in efficiency between gene duplication or the evolution of new alternative exons becomes more pronounced as the number of new isoforms increases. For example, in the same amount of genome space required to generate a single new isoform of our hypothetical gene by gene duplication, hundreds of new isoforms could be created by evolving new alternative exons. Thus, alternative splicing is an extremely economic means of increasing protein diversity.

Alternative splicing is prevalent in metazoan genomes. For example, current estimates suggest that at least 42% of *Drosophila* genes⁶ and over two thirds of mouse and human

genes⁷ encode alternatively spliced pre-mRNAs. These numbers have been increasing at a brisk pace over the past several years and are likely to still be underestimates as many low abundance, tissue-specific or developmentally regulated isoforms almost certainly remain to be characterized. Thus, it is now fair to say that the majority of metazoan genes encode alternatively spliced pre-mRNAs. Alternative splicing is clearly the rule, not the exception.

An issue that is quite distinct from the number of genes that encode alternatively spliced pre-mRNAs is the number of isoforms generated per gene. Figure 1 depicts the number of mRNA isoforms per gene in *D. melanogaster*, which is perhaps the best annotated metazoan genome. There are several conclusions that can be drawn from this graph. First, many genes encode only a single mRNA isoform. Second, very few genes encode a large number of mRNA isoforms - the greatest number of mRNA isoforms encoded by a gene in this dataset is 26 for the *longitudinals lacking (lola)* and *modifier of mdg4 (mod(mdg4))* genes. Thus, the number of genes encoding multiple mRNA isoforms decreases as the number of isoforms increases, yet it would appear that alternative splicing is rarely used to create a tremendously diverse set of mRNA isoforms from a single gene.

There are, however, many caveats that should be kept in mind when interpreting this dataset as it significantly underestimates both the number of mRNA isoforms expressed per gene and the number of genes encoding multiple mRNAs. The current annotation of the *D. melanogaster* genome (version 4.1) lists only 21.7% of the genes as encoding more than one mRNA isoform. This is because this dataset includes only *annotated* mRNA isoforms most of which have been manually curated. Despite the overall high quality of the annotation, many mRNA isoforms that are well-documented in the literature are not present in this annotation of the genome. Moreover, recent microarray analyses suggests that at least 42% of genes encode multiple mRNA isoforms and few, if any of the isoforms detected by this method are present in this dataset⁶. As the annotation and our understanding of the entire repertoire of mRNAs expressed by the *Drosophila* genome improves, the slope of this graph will change such that both the number of genes that express multiple isoforms and the number of isoforms per gene will increase.

In this review we will consider in detail a few genes that express a large repertoire of mRNA isoforms and represent the outliers of this graph. These genes represent what we refer to as complex alternative splicing events. Genes in this class typically have both a complex genomic organization and appear to use unique mechanisms in the expression of their mRNA repertoire. Throughout this review, we will focus on *Drosophila* genes as they represent some of the most unusual and best characterized complex alternative splicing events.

An Overview of Alternative Splicing - from Simple to Complex

As revealed in the plot shown in Figure 1, there are thousands of *Drosophila* genes that encode pre-mRNAs that are alternatively spliced to generate only two mRNA isoforms. A classic example of this is *doublesex (dsx)* which functions as a key regulatory gene in the sex-determination pathway (Figure 2)⁸. In males, the *dsx* pre-mRNA is spliced to include exons 1-3, 5, and 6 and to skip exon 4. In contrast, in females, the same pre-mRNA is spliced to include exons 1-4 and a poly(A) site within exon 4 is used. The male and female-

specific mRNAs encode male- and female-specific DSX proteins that function as transcription factors to regulate the expression of genes that control the sexual differentiation pathway⁹. This is an excellent example where alternative splicing is used to create two mRNA isoforms that encode proteins that function as a binary switch to control an extremely important aspect of biology. While many genes that encode only two mRNA isoforms are of obvious interest and biological importance, for the remainder of this review, we will describe genes and regulatory mechanisms that generate truly phenomenal numbers of mRNA isoforms by virtue of alternative splicing.

Mhc

An excellent example of a gene that undergoes complex alternative splicing in *Drosophila* is the *Myosin heavy chain (Mhc)* gene which encodes a protein that plays a critical role in the function of muscle cells¹⁰. The *Mhc* gene contains 30 exons, 17 of which are alternatively spliced (Figure 2). With the exception of exon 18, which is represented at the genomic level by a single alternative exon, the other alternatively spliced exons are organized into 5 separate clusters which contain from 2 to 5 exons each. The alternative exons within each cluster are included in the mRNA in a mutually exclusive manner – only one exon from each cluster is included in the final mRNA. The consequence of this gene organization is that 480 different mRNAs can be generated from this gene by alternative splicing.

A few RNA sequence elements have been identified that play roles in controlling the splicing of the exon 11 cluster of *Mhc*, which contains five exons (exons 11a to 11e)¹¹. The regulatory elements were all initially identified based on their evolutionary conservation and are called conserved intronic elements (CIEs). One of these elements, CIE3, is located in the last intron of the exon 11 cluster. Deletion of CIE3 results in skipping of exon 11e in the indirect flight muscle of the fly, and exon 10 is spliced directly to exon 12. However, the splicing of the other 4 exon 11 variants is unaffected. Though the mechanism by which this element functions is entirely unknown, it is clear that the CIE3 plays an important role in regulating splicing of the exon 11 cluster.

Para

Even greater numbers of distinct proteins can be synthesized from other *Drosophila* genes when other RNA processing events, such as RNA editing, are combined with alternative splicing. One type of RNA editing is the post-transcriptional conversion of adenosine to inosine – a modification that can alter the identity of a single amino acid in the encoded protein. The *Drosophila paralytic (para)* gene, which encodes the major voltage-gated action potential sodium channel, produces a pre-mRNA that is processed by both alternative splicing and RNA editing^{12,13}. The *para* gene contains 13 alternative exons and can potentially synthesize 1,536 different mRNAs utilizing alternative splicing alone (Figure 2). However, at least 11 adenosines are edited to inosine in *para* transcripts. Considering both RNA editing and alternative splicing, 1,032,192 different *para* transcripts can theoretically be synthesized from this single gene. These examples serve to illustrate how alternative splicing, and in some cases RNA editing, can significantly expand protein diversity.

Dscam

The *Dscam* gene is by far the most extreme case in which alternative splicing alone can generate an extraordinarily diverse repertoire of mRNAs and proteins. This gene, which is essential in *Drosophila*, contains 115 exons, 95 of which are alternatively spliced. The alternative exons are organized into four distinct clusters – the exon 4, 6, 9, and 17 clusters – that contain 12, 48, 33, and 2 variable exons each (Figure 2)^{14,15}. Importantly, the exons within each cluster are alternatively spliced in a mutually exclusive manner. As a result, *Dscam* potentially encodes 38,016 different isoforms. The *Drosophila* Dscam protein is most similar to the human Down Syndrome Cell Adhesion Molecule Protein. However, the human gene does not appear to undergo any of the alternative splicing that occurs in *Drosophila*.

The *Drosophila* Dscam protein contains an extracellular domain, composed of 10 immunoglobulin domains and four fibronectin type III domains, connected to a trans-membrane domain and an intracellular domain^{14,15}. Alternative splicing of the exon 4, 6, and 9 clusters alters the sequence of three of the Ig domains, while alternative splicing of the exon 17 cluster alters the transmembrane domain. These splicing events have important consequences on the function of the protein. *Dscam* is conserved in all insects¹⁶⁻¹⁸ and has recently been identified in *Daphnia pulex*, a crustacean (B.R.G. unpublished data), indicating that the gene first evolved in its current form at least 600 million years ago. In each of these organisms, *Dscam* expresses thousands of forms. Thus a diverse repertoire of *Dscam* isoforms must be important for its function.

The function and biochemical properties of *Dscam* are as remarkable as its organization. One important function of *Dscam*, which is an essential gene, is in specifying the wiring of the nervous system. For example, *Dscam* is required for the proper wiring of Bolwigs nerve¹⁴, olfactory receptor neurons¹⁹, projection neurons of the olfactory system²⁰, mushroom body neurons²¹⁻²³, and mechanosensory neurons²⁴. Essentially, in every case that has been examined, mutations in *Dscam* result in neural wiring defects.

More recently, Dscam has been shown to play an important role in the insect immune system. In both *Drosophila*¹⁸ and the mosquito *Anopheles gambiae*²⁵, the malaria vector, *Dscam* is required for the animal to mount an effective immune response. Moreover, loss of *Dscam* function impairs the ability of hemocytes to phagocytose pathogens.

One attractive hypothesis is that each Dscam isoform would interact with different axon guidance cues in the nervous system or pathogens/antigens in the immune system¹⁵. In an elegant series of experiments, the Zipursky lab reported that the extracellular portion of Dscam (including the region encoded by the exon 4, 6, and 9 clusters) is capable of homodimerization²⁶. These homophilic interactions are strikingly specific as isoforms that differ by a single alternative exon fail to interact with one another, but strongly bind to themselves²⁶. Even altering as few as three amino acids in one of the variable domains can significantly reduce binding. Thus, it appears as though 19,008 Dscam extracellular domains exist that interact in an isoform-specific manner. In the immune system, it has also been shown that some Dscam isoforms can interact with *E. coli* while others do not¹⁸. This suggests that Dscam may function as the insect equivalent of antibodies. In support of this,

infecting *A. gambiae* with different pathogens results in the expression of specific repertoires of *Dscam* isoforms²⁵.

The striking interaction properties of *Dscam* raise the possibility that in the nervous system, individual neurons would express different repertoires of *Dscam* isoforms and that these would dictate the wiring pattern of that neuron. In support of this idea, there is clear evidence that the splicing of at least some of the alternative exons are regulated in a developmental and tissue-specific manner^{27,28}. However, these results most likely vastly underestimate the degree of regulation as they were performed on either entire animals or large, complex tissues. It will be necessary to analyze *Dscam* expression at single cell resolution to adequately address this issue. Some attempts have been made to determine the number of isoforms expressed in individual cells by combining single-cell RT-PCR with microarray analysis^{23,28}. These studies suggest that individual neurons express a limited collection of isoforms (less than 50). However, this issue again could benefit from a detailed analysis at the single cell level in the fly. Nonetheless, it is clear that individual neurons express different collections of isoforms.

Recent evidence also suggests that the precise collection of isoforms expressed in an individual cell is critical to specify the wiring pattern of the neuron²⁴. To address this, two *Dscam* alleles were generated that each lacked 5 of the exon 4 variants - one lacked exons 4.2 through 4.6 while the other lacked exon 4.4 through 4.8. This reduces the *Dscam* repertoire from 38,016 to 22,176 potential isoforms. They found that axonal targeting of mechanosensory neurons was disrupted in flies homozygous for both of these alleles. More importantly, however, the pattern of axon branching was different for these two alleles, yet highly consistent in multiple individuals carrying each allele. Because the repertoire of potential isoforms that can be expressed from the two alleles is different, yet the overall number of isoforms is identical, these results provide strong evidence that the sequences encoded by the exon 4 variants have non-redundant functions. In other words, the identity of the isoforms expressed in an individual neuron is critical to specify the correct wiring pattern. The implication of this is that the splicing of *Dscam* must be precisely regulated.

Mechanisms used for Mutually Exclusive Splicing in *Dscam*

One of the most intriguing aspects of *Dscam* is the fact that the exons within each variable cluster are spliced in a strictly mutually exclusive manner. Work on other genes has revealed several mechanisms that are used to ensure that the splicing of pairs of alternative exons is strictly mutually exclusive. First, the splice sites in the intron separating the two alternative exons can be spatially arranged such that when splicing factors recognize one splice site, they prevent the binding of splicing factors to the other splice site through steric hinderance. This mechanism has been shown to occur in the mammalian α -tropomyosin²⁹ and α -actinin³⁰ genes. Alternatively, splicing of the two exons would be mutually exclusive if the intron separating the exons is too small to be efficiently spliced. For example, in *Drosophila*, introns smaller than 59 nucleotides cannot be removed by the spliceosome³¹. Second, mutually exclusive splicing can also be ensured if a gene has a unique arrangement of splice sites that are recognized by the major (which consists of U1, U2, U4, U6, and U5 snRNPs) and minor (which consists of the U11, U12, U4atac, U6atac and U5 snRNPs)

spliceosomes³². Because, neither spliceosome can remove an intron containing a mixture of major and minor splice sites³³, the splicing of such genes is by definition, mutually exclusive. The human stress-activated protein kinase (*JNK 1*) gene contains an alternatively spliced region with this type of organization³⁴. Finally, the splicing of two cassette exons may not be truly mutually exclusive, but rather may simply appear to be so. This could occur if the two alternative exons are not a multiple of three nucleotides. If neither or both exons are included, premature termination codons will be introduced into the mRNA and those isoforms will be subject to degradation by the non-sense mediated decay pathway³⁵. As a result, mRNAs containing one, and only one exon will be stable and the splicing of such premRNAs will appear to be mutually exclusive, though in reality it is not.

Importantly, none of these mechanisms can explain how the exon 4, 6, and 9 clusters of *Dscam* are spliced in a mutually exclusive manner. First, even if steric hinderance could prevent adjacent exons from being spliced together, it would be unable to prevent non-adjacent exons from being spliced. Second, none of the introns separating the variable exons are below the size limit for *Drosophila* introns, and again, this size limit would not effect non-adjacent exons. Third, none of the splice sites in *Dscam* conform to the minor spliceosome consensus sequence. Finally, the NMD mechanism cannot operate. The exons within the exon 4 cluster are all multiples of 3 such that the reading frame would not change regardless of how many exons are included. For the exon 6 and 9 clusters, while the inclusion of more than one exon would result in a frameshift, the inclusion of multiples of three exons (i.e., 4 exons, 7 exons, etc.) would not. However, such products have not been observed. Thus, it would appear that a novel mechanism(s) must exist to ensure that the splicing of the exon 4, 6, and 9 clusters of *Dscam* occurs in a mutually exclusive manner. Surprisingly, in each cluster that has been studied, it has been shown that RNA secondary structures play a critical role in the mutually exclusive splicing.

Mutually exclusive splicing of the exon 4 cluster

Our laboratory recently identified sequences located in the intron between exon 3 and the first exon 4 variant that plays an important role in the splicing of the exon 4 cluster. These sequences were shown to form an RNA secondary structure we call the inclusion stem, or iStem, that is required for the efficient inclusion of all 12 variable exons in the exon 4 cluster³⁶. The iStem is formed by basepairing interactions between a 20 nt sequence located immediately downstream of the 5' splice site of exon 3 with a second 20 nt sequence located 300 nt downstream (Figure 3A). The structural aspects of the iStem are supported by experiments showing that mutations in either half of the iStem that disrupt basepairing result in skipping of all 12 exon 4 variants and that compensatory mutations that restore the structure, but not the sequence, also restore the function of the iStem. The iStem is conserved in all 12 *Drosophila* species that have been sequenced, though the precise sequence varies considerably and there are several examples of compensatory changes. Finally, several observations suggest that only the base-paired portion of the iStem is critical for it's function. First, deleting large portions of the intervening sequence have little effect on the function of the iStem. More importantly, however, is the fact that the sequence and the distance between the regions that engage in basepairing interactions is highly variable

among all *Drosophila* species. Thus, the iStem is a structural rather than a sequence-specific element that governs inclusion of all of the exons within a single cluster.

The fact that the structure, but not the sequence of the iStem is required for its function raises some interesting issues regarding the mechanisms by which it may act. One possibility is that the RNA structure of the iStem somehow promotes exon inclusion in a manner that does not require additional protein factors. Alternatively, a protein, or protein complex interacts with the iStem in a sequence non-specific manner - perhaps a double-stranded RNA binding protein, or an RNA helicase - and this complex somehow promotes exon inclusion. An intriguing property of the iStem is that despite the fact that it controls the inclusion of all of the exons within the exon 4 cluster, it does not play a significant role in determining which variable exon is selected. Thus, although the mechanism by which the iStem functions is not known, it is clear that the iStem is a novel type of regulatory element that simultaneously controls the splicing of multiple alternative exons.

In addition to the iStem, a second RNA secondary structure exists in the exon 4 cluster and is located within the intron between exon 4.12 and exon 5. Like the iStem, this RNA structure promotes the inclusion of all 12 exon 4 variants and is therefore called the iStem2 (Kreahling and Graveley, unpublished). The iStem2 is also evolutionarily conserved, consists of two distantly located sequences that basepair with one another, and mutations in either half of the iStem2 results in skipping of the exon 4 variants. Interestingly, the iStem and iStem2 appear to function together, as simultaneously disrupting both structures more strongly affects exon inclusion than does either structure alone.

These results clearly show that RNA secondary structures play an important role in the mutually exclusive splicing of the exon 4 cluster of the *Dscam* pre-mRNA. As we shall see, this will be a recurring theme as we consider the other alternatively spliced clusters in *Dscam*.

Mutually exclusive splicing of the exon 6 cluster

RNA secondary structures also play an important function in the mutually exclusive splicing of the *Dscam* exon 6 cluster. While searching the entire *Dscam* gene for sequence elements that were conserved in all *Drosophila* species, and could therefore potentially function as binding sites for splicing regulatory proteins, two classes of sequence elements in the exon 6 cluster were identified – the docking site, which is located in the intron downstream of constitutive exon 5, and the selector sequences, which are located upstream of each of the 48 exon 6 variants^{16,37}. The most striking aspect of these elements is that each selector sequence is complementary to a portion of the docking site (Figure 3B). As a result, the interaction between a selector sequence and the docking site juxtaposes one, and only one, alternative exon to the upstream constitutive exon. Moreover, because each selector sequence interacts with the docking site, only one selector sequence at a time can bind to the docking site. The mutually exclusive nature of the docking site-selector sequence interactions immediately suggests that the formation of these competing RNA structures is a central component of the mechanisms guaranteeing that only one exon 6 variant is included in each *Dscam* mRNA^{16,37}.

Though the formation of these structures remain to be experimentally proven, the evidence that they do is extremely compelling. First, these sequence elements are highly conserved - both the docking site and the selector sequences are conserved in all 20 of the sequenced arthropod genomes, and therefore first evolved at least 600 million years. Second, each exon 6 variant in all of the sequenced insect genomes contains an upstream selector sequence. Third, each selector sequence has the potential to base-pair with the docking site. Fourth, several double compensatory mutations have been identified in the honeybee (*Apis mellifera*) *Dscam* gene¹⁶. The docking site in *A. mellifera* contains two nucleotides that are different from the docking sites of all other insects. However, several of the *A. mellifera* selector sequence have the potential to form base pairing interactions with the variable docking site nucleotides. Finally, the high degree of conservation of the docking site is expected of an element that interacts with a large number of elements as mutations in the docking site would affect the splicing of all of the exon 6 variants. Thus, all evidence to date strongly suggests that these RNA secondary structures form and that they play a key role in the mutually exclusive splicing mechanism of the exon 6 cluster.

The exact mechanism by which the docking site-selector sequence interaction promotes the splicing of the adjacent exon is unknown. Moreover, it is not known whether these interactions serve to simply ensure that only one exon 6 variant is included or whether they play a direct role in selecting which of the 48 exons are included. For example, it is possible that selector sequences that have a high affinity for the docking site will be included more efficiently than exons having selector sequences with weaker binding affinities. Current data argues against this idea as there is no correlation between the ΔG of the interaction of each selector sequence with the docking site and the frequency of inclusion observed in microarray²⁸ and sequencing^{14,18,19,23} experiments. However, this could be complicated by factors such as distance, varying elongation rates of RNA polymerase along the gene, and the potential existence of other RNA structures that would alter the actual length or possibly the accessibility of regions within the exon 6 cluster pre-mRNA. Nonetheless, the exon 6 cluster clearly employs a novel and elegant mechanism to ensure that only one exon is included.

Mutually exclusive splicing of the exon 17 cluster

Although nothing is known about how splicing of the exon 9 cluster of *Dscam* occurs in a mutually exclusive manner, there are indications that RNA secondary structures again play a key role in this process in the exon 17 cluster. Anastassiou and colleagues³⁷ identified four conserved sequence elements (designated A, A', B, and B' in Figure 3C) in the intron between exons 16 and 17.1 that have the potential to form intriguing RNA secondary structures. These four sequences form two sets of complementary sequences - A can basepair with A' and B can basepair with B'. Though not as obviously important as the docking site-selector sequence interaction, these sequences potential function to ensure that only one of the two exon 17 variants is included. On one hand, when the A-A' stem forms, exon 17.1 would be preferentially included. In contrast, when the B-B' stem forms, the 3' splice site of exon 17.1 would be occluded and as a result, exon 17.2 would be included. These interactions are not mutually exclusive as both could potentially form at the same time. As a result, it will be important to functionally analyze the role of these sequence

elements in the splicing of the exon 17 cluster. Nonetheless, the conservation, location, and structures of these elements do suggest that they play some role in this process.

Lessons from *Dscam*

There are two striking aspects of the mechanisms of mutually exclusive splicing that have been uncovered by studying *Dscam*. First, in each cluster that has been studied, RNA secondary structures have been identified that play key roles in mutually exclusive splicing. Second, the mechanisms by which each of these structures functions appears to be distinct. This is particularly surprising as it suggests that there are multiple solutions to the common problem of how only one exon is selected for inclusion when three or more exons can be chosen from. Although *Dscam* is the most extreme case of such a genomic organization, it is not the sole member of this class of genes - there are at least 9 other *Drosophila* genes (*Mhc*, *ATPalpha*, *GluClalpha*, *slo*, *heph*, *TepII*, *wupA*, *14-3-3zeta* and *Pfk*) that contain at least one cluster of three or more mutually exclusive exons. The fact that there are multiple mechanisms by which pre-mRNAs with three or more alternative exons can be spliced in a mutually exclusive manner suggests that it is relatively easy to evolve a way to negotiate this problem. Moreover, the fact that several *Drosophila* genes have such an organization, combined with the genomic economy of exon duplication events suggests that this is a robust evolutionary strategy for generating multiple related proteins. It is therefore striking that there are no known genes in humans, or any other vertebrates for that matter, that contain more than two exons that are spliced in a mutually exclusive manner. Why is this? Have alternative mechanisms for generating protein diversity such as VDJ recombination become so successful as to plunge multiple exon mutually exclusive splicing into extinction in the vertebrate lineage? Is the vertebrate spliceosome incapable of negotiating pre-mRNAs with such a configuration? Answers to these questions will likely provide insight into the mechanism of genome evolution.

The second take home message from these findings is that each of these structures were identified by comparative genomics which was made possible by the availability of multiple genomes from closely related species. The only exception is the iStem. In this case, the 3' half of the iStem was initially identified by mutagenesis studies. However, the discovery that it was part of an RNA secondary structure rather than a protein binding site eluded detection until the sequences of several *Drosophila* genomes were available. As the ability to perform comparative genomics studies across multiple genomes has only recently become possible, it is likely that many RNA secondary structures that play an important role in alternative splicing will be discovered over the next several years and we anticipate that such elements will turn out to be quite ubiquitous.

Trans-splicing

Two genes in the *Drosophila* genome mentioned earlier, *lola* and *mod(mdg4)*, take the concept of complexity to a whole new level. These two genes utilize *trans*-splicing to generate mRNAs. *Trans*-splicing is a mechanism by which two different RNA molecules are spliced together to generate a single mRNA. *Trans*-splicing occurs frequently in organisms such as nematodes, trypanosomes and planarians³⁸⁻⁴⁰. In these cases non-coding spliced leader RNAs (SL RNAs) are spliced to the 5' end of the majority of the mRNAs in these

organisms. This serves to both add a 5' cap to the mRNAs and to process polycistronic pre-mRNAs into monocistronic mRNAs. However, as the *trans*-splicing that has been characterized in *Drosophila* involves the joining of RNAs containing coding sequence, these two processes are quite distinct.

The first clear example of a biologically relevant *trans*-spliced coding gene was *mod(mdg4)* (*modifier of mdg4*), which encodes a BTB-domain containing transcription factor⁴¹. This was realized immediately upon obtaining sequencing the *mod(mdg4)* genomic locus and analyzing its organization. *Mod(mdg4)* encodes 28 distinct mRNA isoforms that are generated by the splicing of four common exons (1-4) located at the 5' end of the gene to a collection of alternative exons at the 3' end of the gene (Figure 4). The striking observation was that seven of the isoforms contained alternative exons transcribed from the opposite strand than the strand used to transcribed the common exons^{42,43}! Thus, the splicing of these transcripts must occur via *trans*-splicing. This was rigorously proven by placing one of the *trans*-spliced exons on a separate chromo-some from the common exons and showing that functional mRNAs were generated containing both sequences⁴². Importantly, *mod(mdg4)* is also *trans*-spliced in other *Diptera* and *Lepidoptera* species though the number and location of the variable exons differs between species^{41,44}.

The second gene shown to be *trans*-spliced is *lola* (*longitudinals lacking*), which also encodes a BTB domain-containing transcription factor essential for axon guidance^{45,46}. The *lola* gene is 60 kb in length and contains four alternative promoters and 32 exons that are alternatively spliced to generate 80 mRNAs⁴⁷ (Figure 4). Each transcript contains one of the first four exons generated by alternative promoter use, the constant exons 5-8, and one set of the variable exons at the 3' end. Unlike *mod(mdg4)*, the fact that *lola* undergoes *trans*-splicing was not obvious from the organization of the gene. Rather, this insight came from interallelic complementation studies⁴⁸. Horiuchi and colleagues found *lola* alleles with mutations in the common exons are homozygous lethal, as are alleles with mutations in the variable exons. However, what was most striking was the observation that an allele with a common exon mutation can be complemented by alleles with variable exon mutations. Moreover, the idea that this occurred via *trans*-splicing was verified by showing that chimeric mRNAs were generated in the complemented animals that contained SNPs specific to each allele⁴⁸. As with *mod(mdg4)*, at least some of the variable exons are known to be transcribed from their own promoters⁴⁸.

The identification of genes that use *trans*-splicing as their mechanism of mRNA synthesis raises numerous interesting issues. The first is why is *trans*-splicing used instead of *cis*-splicing? Perhaps this represents yet another means of ensuring mutually exclusive splicing - as the region of the pre-mRNA that is *trans*-spliced is consumed during the reaction, it is not possible to include multiple variable exons between two constant exons. Another important question is how does this type of *trans*-splicing occur? How are the two pre-mRNAs joined together and what specifies which transcripts are to be joined together? Why are the *lola* common exons not spliced to the *mod(mdg4)* variable exons and vice-versa? The observation that a *mod(mdg4)* variable exon placed on a different chromosome than the common exons suggests that this is not due to confining the transcripts to a specific nuclear location. What prevents the *trans*-spliced precursors from being spliced to other pre-mRNAs

in the cell? Finally, as the genomic organization of *lola* does not obviously suggest *trans*-splicing as the dominant synthetic mechanism, one must wonder how many other genes are *trans*-spliced? Does this phenomenon occur outside of insects? There are several examples of mammalian genes that undergo *trans*-splicing^{49,50}. However, in most cases, this occurs in only a small subset of mRNAs that are synthesized and may therefore represent noise rather than biologically significant⁵¹. Nonetheless, *mod(mdg4)* and *lola* add another intriguing example of complex alternative splicing mechanisms that raise more questions than answers.

Conclusions

The examples that have been described in this chapter serve to illustrate the fact that alternative splicing can be extraordinarily complex and is used to generate a vast repertoire of isoforms. For example, the *Drosophila* proteome potentially contains at least 1,084,500 members taking into account the diversity generated by the *Mhc*, *Dscam*, *para*, *lola*, and *mod(mdg4)* genes alone. Clearly a wide variety of regulatory mechanisms are used to precisely control the expression of these mRNAs in space and time, and the fidelity of their synthesis. As our exploration of the transcriptomes of various organisms expands, additional examples of complex alternative splicing events will most certainly be unveiled. Understanding how these remarkable events are controlled and their functional roles in biology will be an exciting and consuming endeavor for years to come.

Acknowledgments

Work in our laboratory is funded by grants from the National Institutes of Health, The Raymond and Beverly Sackler Foundation for the Arts and Sciences, and the State of Connecticut Stem Cell Initiative.

References

1. Black DL. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*. 2000; 103:367–370. [PubMed: 11081623]
2. Blencowe BJ. Alternative splicing: new insights from global analyses. *Cell*. 2006; 126:37–47. [PubMed: 16839875]
3. Graveley BR. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*. 2001; 17:100–107. [PubMed: 11173120]
4. Lander ES, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860–921. [PubMed: 11237011]
5. Venter JC, et al. The sequence of the human genome. *Science*. 2001; 291:1304–1351. [PubMed: 11181995]
6. Stolc V, et al. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science*. 2004; 306:655–660. [PubMed: 15499012]
7. Johnson JM, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*. 2003; 302:2141–2144. [PubMed: 14684825]
8. Lopez AJ. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet*. 1998; 32:279–305. [PubMed: 9928482]
9. Burtis KC, Baker BS. *Drosophila* doublesex gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. *Cell*. 1989; 56:997–1010. [PubMed: 2493994]

10. George EL, Ober MB, Emerson CPJ. Functional domains of the *Drosophila melanogaster* muscle myosin heavy-chain gene are encoded by alternatively spliced exons. *Mol. Cell. Biol.* 1989; 9:2957–2974. [PubMed: 2506434]
11. Standiford DM, Sun WT, Davis MB, Emerson CPJ. Positive and negative intronic regulatory elements control muscle-specific alternative exon splicing of *Drosophila* myosin heavy chain transcripts. *Genetics.* 2001; 157:259–271. [PubMed: 11139507]
12. Palladino MJ, Keegan LP, O'Connell MA, Reenan RA. A-to-I premRNA editing in *Drosophila* is primarily involved in adult nervous system function and integrity. *Cell.* 2000; 102:437–449. [PubMed: 10966106]
13. Thackeray JR, Ganetzky B. Conserved alternative splicing patterns and splicing signals in the *Drosophila* sodium channel gene para. *Genetics.* 1995; 141:203–214. [PubMed: 8536968]
14. Schmucker D, et al. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell.* 2000; 101:671–684. [PubMed: 10892653]
15. Zipursky SL, Wojtowicz WM, Hattori D. Got diversity? Wiring the fly brain with Dscam. *Trends Biochem. Sci.* 2006; 31:581–588. [PubMed: 16919957]
16. Graveley BR. Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell.* 2005; 123:65–73. [PubMed: 16213213]
17. Graveley BR, et al. The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes. *RNA.* 2004; 10:1499–1506. [PubMed: 15383675]
18. Watson FL, et al. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science.* 2005; 309:1874–1878. [PubMed: 16109846]
19. Hummel T, et al. Axonal targeting of olfactory receptor neurons in *Drosophila* is controlled by Dscam. *Neuron.* 2003; 37:221–231. [PubMed: 12546818]
20. Zhu H, et al. Dendritic patterning by Dscam and synaptic partner matching in the *Drosophila* antennal lobe. *Nat Neurosci.* 2006; 9:349–355. [PubMed: 16474389]
21. Wang J, et al. Transmembrane/juxtamembrane domain-dependent Dscam distribution and function during mushroom body neuronal morphogenesis. *Neuron.* 2004; 43:663–672. [PubMed: 15339648]
22. Wang J, Zugates CT, Liang IH, Lee CH, Lee T. *Drosophila* Dscam is required for divergent segregation of sister branches and suppresses ectopic bifurcation of axons. *Neuron.* 2002; 33:559–571. [PubMed: 11856530]
23. Zhan XL, et al. Analysis of Dscam diversity in regulating axon guidance in *Drosophila* mushroom bodies. *Neuron.* 2004; 43:673–686. [PubMed: 15339649]
24. Chen BE, et al. The molecular diversity of Dscam is functionally required for neuronal wiring specificity in *Drosophila*. *Cell.* 2006; 125:607–620. [PubMed: 16678102]
25. Dong Y, Taylor HE, Dimopoulos G. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the *Anopheles gambiae* innate immune system. *PLoS Biol.* 2006; 4:e229. [PubMed: 16774454]
26. Wojtowicz WM, Flanagan JJ, Millard SS, Zipursky SL, Clemens JC. Alternative splicing of *Drosophila* Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell.* 2004; 118:619–633. [PubMed: 15339666]
27. Celotto AM, Graveley BR. Alternative splicing of the *Drosophila* Dscam premRNA is both temporally and spatially regulated. *Genetics.* 2001; 159:599–608. [PubMed: 11606537]
28. Neves G, Zucker J, Daly M, Chess A. Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nat. Genet.* 2004; 36:240–246. [PubMed: 14758360]
29. Smith CW, Nadal-Ginard B. Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell.* 1989; 56:749–758. [PubMed: 2924347]
30. Southby J, Gooding C, Smith CW. Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of alpha-actinin mutually exclusive exons. *Mol. Cell. Biol.* 1999; 19:2699–2711. [PubMed: 10082536]

31. Kennedy CF, Kramer A, Berget SM. A role for SRp54 during intron bridging of small introns with pyrimidine tracts upstream of the branch point. *Mol. Cell. Biol.* 1998; 18:5425–5434. [PubMed: 9710626]
32. Patel AA, Steitz JA. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol.* 2003; 4:960–970. [PubMed: 14685174]
33. Sharp PA, Burge CB. Classification of introns: U2-type or U12-type. *Cell.* 1997; 91:875–879. [PubMed: 9428511]
34. Letunic I, Copley RR, Bork P. Common exon duplication in animals and its role in alternative splicing. *Hum. Mol. Genet.* 2002; 11:1561–1567. [PubMed: 12045209]
35. Jones RB, et al. The nonsense-mediated decay pathway and mutually exclusive expression of alternatively spliced FGFR2IIIb and -IIIc mRNAs. *J. Biol. Chem.* 2001; 276:4158–4167. [PubMed: 11042206]
36. Krehling JM, Graveley BR. The iStem, a long-range RNA secondary structure element required for efficient exon inclusion in the *Drosophila* Dscam premRNA. *Mol. Cell. Biol.* 2005; 25:10251–10260. [PubMed: 16287842]
37. Anastassiou D, Liu H, Varadan V. Variable window binding for mutually exclusive alternative splicing. *Genome Biol.* 2006; 7:R2. [PubMed: 16507134]
38. Blumenthal T. Trans-splicing and polycistronic transcription in *Caenorhabditis elegans*. *Trends Genet.* 1995; 11:132–136. [PubMed: 7732590]
39. Nilsen TW. trans-splicing: an update. *Mol Biochem Parasitol.* 1995; 73:1–6. [PubMed: 8577317]
40. Zayas RM, Bold TD, Newmark PA. Spliced-leader trans-splicing in freshwater planarians. *Mol. Biol. Evol.* 2005; 22:2048–2054. [PubMed: 15972844]
41. Krauss V, Dorn R. Evolution of the trans-splicing *Drosophila* locus *mod(mdg4)* in several species of Diptera and Lepidoptera. *Gene.* 2004; 331:165–176. [PubMed: 15094203]
42. Dorn R, Reuter G, Loewendorf A. Transgene analysis proves mRNA trans-splicing at the complex *mod(mdg4)* locus in *Drosophila*. *Proc Natl Acad Sci U S A.* 2001; 98:9724–9729. [PubMed: 11493677]
43. Labrador M, et al. Molecular biology: Protein encoding by both DNA strands. *Nature.* 2001; 409:1000. [PubMed: 11234000]
44. Gabler M, et al. Trans-splicing of the *mod(mdg4)* complex locus is conserved between the distantly related species *Drosophila melanogaster* and *D. virilis*. *Genetics.* 2005; 169:723–736. [PubMed: 15520256]
45. Crowner D, Madden K, Goeke S, Giniger E. Lola regulates midline crossing of CNS axons in *Drosophila*. *Development.* 2002; 129:1317–1325. [PubMed: 11880341]
46. Giniger E, Tietje K, Jan LY, Jan YN. lola encodes a putative transcription factor required for axon growth and guidance in *Drosophila*. *Development.* 1994; 120:1385–1398. [PubMed: 8050351]
47. Ohsako T, Horiuchi T, Matsuo T, Komaya S, Aigaki T. *Drosophila* lola encodes a family of BTB-transcription regulators with highly variable C-terminal domains containing zinc finger motifs. *Gene.* 2003; 311:59–69. [PubMed: 12853139]
48. Horiuchi T, Giniger E, Aigaki T. Alternative trans-splicing of constant and variable exons of a *Drosophila* axon guidance gene, *lola*. *Genes Dev.* 2003; 17:2496–2501. [PubMed: 14522953]
49. Finta C, Zaphiropoulos PG. Intergenic mRNA molecules resulting from trans-splicing. *J. Biol. Chem.* 2002; 277:5882–5890. [PubMed: 11726664]
50. Wu Q, Maniatis T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell.* 1999; 97:779–790. [PubMed: 10380929]
51. Tasic B, et al. Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. *Mol Cell.* 2002; 10:21–33. [PubMed: 12150904]

Figure 1. Number of distinct mRNA isoforms per gene in *Drosophila melanogaster*

The number of distinct mRNA isoforms annotated for each gene in version 4.1 of the *D. melanogaster* genome was determined. The number of genes is plotted as a function of the number of annotated mRNA isoforms derived from each gene. Note that the Y axis is represented using the logarithmic scale.

Figure 2. Variation in the complexity of alternative splicing in *Drosophila*

(A) The *doublesex* pre-mRNA generates two distinct isoforms in a sex-specific manner. (B) The *Mhc* gene can produce 480 different mRNAs. (C) The *para* gene undergoes both alternative splicing and RNA editing (the sites and number of RNA editing events are indicated by the arrows). As a result, 1,032,192 different *para* mRNAs can potentially be synthesized. (D) The *Dscam* gene can generate 38,016 different mRNAs by virtue of alternative splicing alone.

Figure 3. RNA secondary structures important for mutually exclusive splicing in the *Drosophila Dscam* pre-mRNA

(A) The iStem functions to promote the inclusion of any one of the 12 exon 4 variants. In the absence of the iStem, the exon 4 variants are excluded. (B) The docking site-selector sequence interactions function to prevent the inclusion of multiple exon 6 variants. Due to the fact that only one selector sequence can interact with the docking site at a time, only one exon 6 variant can be included. (C) Four conserved sequence elements (represented by the boxes labeled A, A', B and B') have the potential to form RNA structures that may prevent the inclusion of both exon 17 variants in the same mRNA.

A.

Modifier of mdg4 (Mod(mdg4))

B.

longitudinals lacking (lola) **Figure 4. Trans-spliced genes in *Drosophila***

(A) The *modifier of mdg4 (mod(mdg4))* gene contains numerous variable 3' exons that are joined to a set of four common exons. Several of the variable exons are encoded on the opposite strand than the common exons and therefore are joined together by *trans*-splicing.

(B) The *longitudinals lacking (lola)* gene also contains numerous variable 3' exons that are joined to a set of common exons. In this case, the variable exons are encoded on the same strand of as the common exons. However, interallelic complementation experiments indicate that the *lola* mRNAs are synthesized by *trans*-splicing.