Spliceosomal small nuclear RNA genes in 11 insect genomes

STEPHEN M. MOUNT,^{1,2,5} VALER GOTEV,^{3,5} CHIAO-FENG LIN,³ KRISTINA HERNANDEZ,³ and WOJCIECH MAKALOWSKI^{3,4}

¹Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742-5815, USA

²Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA

³Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

⁴Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

ABSTRACT

The removal of introns from the primary transcripts of protein-coding genes is accomplished by the spliceosome, a large macromolecular complex of which small nuclear RNAs (snRNAs) are crucial components. Following the recent sequencing of the honeybee (*Apis mellifera*) genome, we used various computational methods, ranging from sequence similarity search to RNA secondary structure prediction, to search for putative snRNA genes (including their promoters) and to examine their pattern of conservation among 11 available insect genomes (*A. mellifera*, *Tribolium castaneum*, *Bombyx mori*, *Anopheles gambiae*, *Aedes aegypti*, and six *Drosophila* species). We identified candidates for all nine spliceosomal snRNA genes in all the analyzed genomes. All the species contain a similar number of snRNA genes, with the exception of *A. aegypti*, whose genome contains more U1, U2, and U5 genes, and *A. mellifera*, whose genome contains fewer U2 and U5 genes. We found that snRNA genes are generally more closely related to homologs within the same genus than to those in other genera. Promoter regions for all spliceosomal snRNA genes within each insect species share similar sequence motifs that are likely to correspond to the PSEA (proximal sequence element A), the binding site for snRNA activating protein complex, but these promoter elements vary in sequence among the five insect families surveyed here. In contrast to the other insect species investigated, Dipteran genomes are characterized by a rapid evolution (or loss) of components of the U12 spliceosome and a striking loss of U12-type introns.

Keywords: snRNA; spliceosomal snRNA; secondary structure; U12 introns; honeybee; Insecta

INTRODUCTION

Protein-coding genes in most eukaryotes are interrupted by introns that must be removed in a process known as RNA splicing (Berget et al. 1977; Chow et al. 1977; Gilbert 1978). This complex process is carried out by the spliceosome, a large macromolecular assembly consisting of perhaps as many as 100 proteins (Jurica and Moore 2003; Hochleitner et al. 2005) and a set of five small nuclear RNAs (snRNAs) that are found complexed with proteins in the form of small nuclear ribonucleoproteins (snRNPs). Most introns are removed by the major spliceosome, which includes the U1, U2, U4, U5, and U6 snRNAs. Spliceosomal components, including the snRNAs, are highly conserved throughout

eukaryotes (Mount and Salz 2000; Barbosa-Morais et al. 2006). The snRNAs are encoded by moderately repeated genes that show some variation within a species. Here we take advantage of the recently sequenced honeybee (*Apis mellifera*) (Honeybee Genome Sequencing Consortium 2006) and 10 other insect genomes to examine the pattern of conservation of spliceosomal snRNA genes within the class Insecta.

snRNAs and snRNA variants

Past work in *Drosophila melanogaster* has described variant U1 (Lo and Mount 1990) and U5 (Chen et al. 2005) snRNAs and their distinct patterns of expression throughout development. In each case, individual variants with distinct sequences showed unique patterns of developmental and tissue-specific expression. Similar observations have been made in pea (Hanley and Schuler 1991) and silk moth (Sierra-Montes et al. 2005). The differential expression of snRNA variants could be due to functional differences. However, differential expression could also be explained if

⁵These authors contributed equally to this work.

Reprint requests to: Wojciech Makalowski, 514 Mueller Laboratory, The Pennsylvania State University, University Park, PA 16802, USA; e-mail: wojtek@psu.edu; fax: (814) 865-9366.

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.259207>.

sequence variants without functional significance are associated by chance with genes showing distinct patterns of expression driven by other factors (most likely, transcriptional regulation of the amount of snRNA produced during different developmental stages). Here we have addressed this question by examining the evolutionary stability of snRNA variants.

The minor spliceosome

In the case of $\sim 0.2\%$ of mammalian introns (Burge et al. 1998; Levine and Durbin 2001), splicing is carried out by a minor (U12) spliceosome, which contains U11, U12, U4atac, and U6atac in place of U1, U2, U4, and U6 snRNAs, respectively (Tarn and Steitz 1996). The U5 snRNP is shared between the major and minor spliceosomes (Luo et al. 1999; Schneider et al. 2002). The snRNPs that are unique to this U12 spliceosome contain, in addition to common core snRNP proteins, a number of unique proteins, most of which are homologous to proteins in the U2 spliceosome (Will et al. 2004). The U12 spliceosome was originally identified as required for the removal of introns carrying the noncanonical AT-AC terminal dinucleotides in place of GT-AG (Hall and Padgett 1994; Tarn and Steitz 1996). It has since become clear that most U12 introns carry standard GT and AG dinucleotides (Dietrich et al. 1997; Levine and Durbin 2001), but that the U12 introns can nevertheless be recognized by distinct splice signal sequences (Sharp and Burge 1997). The U12 spliceosome appears to be ancient since U12 introns are found, together with genes for components of the minor spliceosome, in plants (Wu et al. 1996), insects, and chordates (Burge et al. 1998). However, many species (including *Caenorhabditis elegans* and most, if not all, unicellular eukaryotes) have lost the U12 spliceosomal machinery altogether.

The genome of *D. melanogaster* has a divergent U12 spliceosome, and was originally thought to lack U11 snRNA (Adams et al. 2000), but subsequent analysis including

affinity purification revealed a highly divergent U11 snRNA (Schneider et al. 2004). We find that the divergence of the minor spliceosome appears to be limited to Diptera.

snRNA promoters

With the exception of U6 and U6atac, the spliceosomal snRNA genes are transcribed by RNA polymerase II and have a cap structure resembling that found on messenger RNAs but with a characteristic trimethylation of the cap guanosine. U6 and U6atac genes are transcribed by RNA polymerase III and have a distinct monomethyl phosphate cap. However, all snRNAs are abundant and are transcribed with the help of a common multisubunit transcription factor known as the snRNA activating protein complex (SNAPc) (Hernandez 2001; Lai et al. 2005) that binds to the proximal sequence element A (PSEA). We have identified conserved sequences upstream of snRNAs in each of the insect species that are likely to correspond to the PSEA.

RESULTS AND DISCUSSION

Putative spliceosomal snRNAs genes in the honeybee genome

We searched for snRNA genes in the honeybee genome using a combination of BLAST with customized parameters and covariance model approaches (see Materials and Methods). The honeybee genome has five putative U1 snRNA genes, three U2, two U4, three U5, three U6, and one putative gene for each of the minor spliceosomal snRNAs (Table 1). A putative U1 pseudogene located on Chromosome 1 (NCBI gi 63053347, position 4437-4276) carries two mutations in the essential conserved 5' terminus and lacks the conserved U1-specific start-site motif AAGC (which is present immediately upstream of all other

TABLE 1. Approximate (see Materials and Methods) gene numbers by snRNA type and species

Species	U gene type/reference gene ^a length								
	U1 165	U2 192	U4 143	U5 124	U6 108	U11 275	U12 238	U4atac 160	U6atac 97
<i>Drosophila melanogaster</i>	5	6	3	7	3	1	1	1	1
<i>Drosophila simulans</i>	8	6	3	6	3	1	1	1	1
<i>Drosophila sechellia</i>	7	6	3	7	3	1	1	1	1
<i>Drosophila yakuba</i>	9	7	3	8	3	1	1	1	1
<i>Drosophila pseudoobscura</i>	7	8	2	7	3	1	1	1	1
<i>Drosophila persimilis</i>	7	7	3	7	3	1	1	1	1
<i>Anopheles gambiae</i>	8	7	1	7	2	2	1	1	1
<i>Aedes aegypti</i>	18	10	3	9	3	1	1	2	1
<i>Bombyx mori</i>	8	8	3	8	4	1	1	1	1
<i>Apis mellifera</i>	5	3	2	3	3	1	1	1	1
<i>Tribolium castaneum</i>	5	5	2	6	3	1	1	1	1

^a*D. melanogaster* gene sequences were used as references and queries in BLAST searches.

putative U1 snRNA genes in the insect species examined here). However, this gene has the PSEA signal and relatively few other substitutions. We conclude that this gene was functional until recently and cannot exclude the possibility that it is still expressed and functional. In addition, there are four tandemly repeated U2 pseudogenes adjacent to the U2 snRNA gene on Chromosome 8. Each of them is missing 25 nucleotides (nt) from the 5' end, which affects formation of the first loop in the U2 snRNA secondary structure (see Supplemental Fig. S1 at http://warta.bio.psu.edu/htt_doc/Projects/snRNA/). They also lack the PSEA signal upstream, which likely affects their transcription.

The 20 snRNA genes in *A. mellifera* are spread across 11 chromosomes, two of them being located, however, in contigs yet unmapped. Because of this, there is little clustering of the snRNA genes, apart from two U1 genes that lie ~1 kb apart and within 100 kb of a third U1 gene on Chromosome 16. This is in contrast to *D. melanogaster*, which has four clusters of snRNA genes (Mount and Salz 2000), including one with two U2, one U4, and two U5 genes spread over 6 kb of Chromosome 2L.

A comparison of the number of snRNA genes among the 11 insect species analyzed (Table 1) reveals that U1, U2, and U5 genes tend to be the most abundant in each species, while U4 and U6 genes are intermediate in abundance. *A. mellifera* appears to be an exception to this rule, because it only has three copies for each of the U2 and U5 genes, just as many as there are U6 genes. The genes for snRNAs found in the minor spliceosome, U4atac, U6atac, U11, and U12, are single copy, with the exception of *Aedes aegypti* U4atac and *Anopheles gambiae* U11, which have two copies each. Copy number polymorphism for U1 snRNA (five to seven genes) is documented in *D. melanogaster* (Lo and Mount 1990), and such minor variations in gene number are plausibly common for insect snRNA genes. These relative abundances of gene numbers reflect the overall abundance of snRNAs within cells (Mount and Steitz 1981).

Evolution of spliceosomal snRNA genes

As mentioned above, it has been shown that different U1 and U5 variants have tissue-specific expression patterns in *D. melanogaster*. We investigated the relationship between different variants by phylogenetic analysis to see if the pattern of conservation is consistent with expression differentiation. We found that neither the phylogenetic tree of U1 genes nor that of U5 genes clearly supports the functional differentiation of different variants. In fact, the trees of all snRNA genes reveal a pattern that is more consistent with a concerted mode of evolution or extreme purifying selection (Piontkivska et al. 2002; Nei and Rooney 2005). This is because different variants are more similar to variants within a genus than between genera (see a neighbor joining tree of U5 genes in Fig. 1) suggesting that snRNA genes are being constantly homogenized by

processes such as gene conversion, unequal crossover, or replacement through birth and death.

Even though the concerted mode of evolution appears as a convenient explanation at first, especially because it applies to other RNA gene families as well, it would require clustering of the genes that are being homogenized by gene conversion or unequal crossover (Nei and Rooney 2005). In the case of the spliceosomal snRNA genes, we observe only partial clustering, mostly in the eight Dipteran species that have a small number of chromosomes as compared to the other five insect species included in this analysis. Therefore, the concerted mode of evolution can only partially explain the snRNA phylogenies. We investigated additional forces acting on these gene families by looking at their functional constraints.

The case of U6 snRNA genes is particularly interesting due to the high level of conservation observed across all species. All U6 genes are 108 nt long, of which only 13 are variable sites, the rest (95 nt; 88%) being perfectly conserved across all 33 U6 genes detected. It is remarkable that all *Drosophila* U6 genes are identical, with the exception of the *Drosophila yakuba* gene Dyak[U6|84681803|81773-81666, which differs only by 2 nt (singleton mutations) in the 5'-loop region and is otherwise perfectly conserved. Had we looked only at *Drosophila*, where the three U6 genes are found within a 1.5-kb region, the concerted evolution scenario would have seemed highly possible, in spite of protein-coding genes CG6643 and CG13624 flanking the triplet (U6:96Aa is, in fact, located in the last intron of CG6643). This scenario is contradicted, however, by low sequence conservation outside of genes and by the lack of clustering in non-Dipteran species, possibly due to the higher fragmentation of their genomes. And yet, the four *Bombyx mori* U6 genes are identical to each other, as are two of the three *Tribolium castaneum* genes, and two of the three *A. mellifera* genes (in the case of both *T. castaneum* and *A. mellifera*, only the first nucleotide is different in the third copy). Out of the 13 variable sites, four more are singletons that all belong to one *A. aegypti* gene (Aaeg[U6|78152160|82903-82796). Twelve of the 13 variable sites are concentrated in the 18-nt segment of the 5' end of the gene, the rest being almost invariant (one of the *A. aegypti* singletons is found at position 43). This conservation pattern agrees well with expected stringent functional constraints of the U6 gene: it binds Lsm proteins and base pairs with U4 (Supplemental Fig. S2 at http://warta.bio.psu.edu/htt_doc/Projects/snRNA/), U2, and the donor splice sites of introns. Based on these facts, one can say that purifying selection rather than concerted evolution is the more likely explanation for the high conservation of U6 genes.

Within the spliceosome, U6 snRNA pairs with U4 snRNA (Supplemental Fig. S1 at http://warta.bio.psu.edu/htt_doc/Projects/snRNA/), which, however, does not interact as intimately with the other components of the spliceosome as U6 does. Consequently, U4 genes are subject

to more relaxed functional constraints. Interestingly, the U4 phylogeny reveals a mixed evolutionary pattern. Within the genus *Drosophila*, U4 genes appear to be subject to divergent evolution, as the orthology relationship between genes located in regions of conserved synteny is unambiguously solved by the phylogenetic tree. This indicates that U4 genes are not homogenized by gene conversion or unequal crossover in spite of their being located on the same chromosome (2L in the case of *D. melanogaster*). For the rest of the species, genes are more similar within a genus than among genera, which is more consistent with a mechanism of concerted evolution. The lack of conservation outside the genic region and the multichromosomal distribution of U4 genes (e.g., *A. mellifera* U4 genes are located on Chromosomes 1 and 15) indicate that gene conversion or unequal crossover is unlikely to homogenize U4 genes. Instead, it could be that a common functional constraint, such as interacting with U6 genes that are highly conserved (see above), determines all genes to evolve in a concerted manner. The same mechanism could very well apply to functional elements in the promoter region, such as the PSEA (see below).

In the case of the other major spliceosomal components, U1, U2, and U5, a complex of evolutionary forces appears to have generated the spectrum of different snRNA variants that we were able to detect (see U5 snRNA example in Fig. 2). Higher clustering in Diptera allows gene conversion and unequal crossover to occur. For example, nine of the 18 *A. aegypti* U1 genes are separated by 1–3 kb, and six of the seven U1 genes in *Drosophila pseudoobscura* form two three-gene clusters that are 17 kb apart. The potential of U1 for recombination events is demonstrated by a U1-mediated translocation event in *Drosophila* (Gonzalez et al. 2004). An increased number of U1 genes in *A. aegypti* (Table 1) and identification of pseudogenes for all snRNA types (data not shown) both agree with the death-and-birth model of evolution (Nei and Rooney 2005). So does duplication of genes, which can be

FIGURE 1. (Legend on next page)

observed in the case of head-to-head U2-U5 gene pairs found in the genus *Drosophila*. In the case of *D. melanogaster*, three such pairs are found on Chromosome 2L (two in region 38AB and one in region 34A) and one on Chromosome X (region 14B). Since no such pair appears to exist outside of the genus *Drosophila*, it is reasonable to believe they are the result of segmental duplication events rather than independent association of U2 and U5 genes in four genomic places. In spite of these apparent duplication events, the number of U2 and U5 genes in *Drosophila* is not significantly greater than in other species (Table 1), indicating that other genes were lost, in agreement with the birth-and-death model of evolution. In *Drosophila* we also observe conservation of certain variants across the genus. A good example is a U5 variant (U5:63BC in *D. melanogaster*) that has conserved the 3' end in all *Drosophila* species. It is not clear, however, if the conservation of this variant is due to functional constraints or because it is located in a region that does not favor recombination events, making it impossible for its sequence to be homogenized.

In conclusion, we should re-emphasize that the evolution of spliceosomal snRNA is governed by several concurrent forces. Purifying selection is certainly one major force. In Dipteran species, which have a small number of chromosomes, concerted evolution by gene conversion or unequal crossover may play an important role as well. In non-Dipteran species, where snRNA genes do not form clusters (perhaps because of the higher fragmentation of genomes), concerted evolution by recombination is likely to be extremely rare. It is thus possible that apparently concerted evolution results from coordinated changes in gene sequences as a result of selection and birth-death processes. This scenario could also explain why different species have slightly different PSEA consensi (see below), but it requires further investigations.

Small nuclear RNA promoters

The PSEA motifs upstream of snRNA genes are remarkably similar within species (Hernandez 2001). This is true for species as diverse as *Arabidopsis thaliana*, *D. melanogaster*, and human. We investigated the pattern of conservation of the PSEA elements within the 11 insect genomes available at the time of this study (see Materials and Methods and Fig. 3). A conserved promoter motif was found in every species, and was shared among genes for all of the spliceosomal snRNAs. Sequence logos summarizing these conserved elements are presented in Figure 3. In each case, the motif is present at a similar distance of ~55 nt upstream of each gene. The honeybee PSEA signal can be represented by the consensus TTCTC (the underlined nucleotide denotes a variable position). All *Drosophila* species share a slightly longer conserved signal (ATTCCCAA), whereas the two mosquitoes share ATCGCTA (see Fig. 3). It is important to note that in each case, the longer consensus contains a common TCNC motif. In the case of *T. castaneum*, however, the signal appears to be much less conserved, and MEME did not report any conserved motif at the expected position. A more careful analysis showed that *Tribolium* genes had the TCNC motif, but the flanking sequences are much less conserved, with the exception of U6 and U6atac genes (Fig. 3). Therefore, it is possible that the conservation of the short TCNC motif across all species and U gene types is due to functional constraints, being crucial for SNAPc binding, whereas the species-specific conservation of flanking residues is due to coevolution of the transcription factor with a small number of sites.

Divergence of the minor spliceosome within Diptera

The secondary structure of *Drosophila* U11 snRNA, a component of the minor spliceosome, was described as highly diverged from other known U11 sequences (Schneider et al. 2004), and we expected that the honeybee U11 would have a structure similar to that of *Drosophila*. To our surprise, we found that the structure of honeybee U11 resembles more closely that of human or plant U11 genes (Fig. 4). In fact, the dramatic elongation of the second stem-loop of U11, which is a characteristic of *Drosophila* species, is not seen in either *Apis*, *Bombyx*, or *Tribolium*, each of which has a simple stem with 12 or fewer base pairs. Interestingly, the mosquito U11 genes contain stems of intermediate length in this position (the *A. gambiae* structure is shown in Fig. 4). This finding suggests that the conservation of the minor spliceosome observed from plants to mammals was lost in Dipteran species.

FIGURE 1. Phylogenetic tree of insect U5 snRNA molecules. The tree was inferred using the neighbor-joining method. However, the maximum parsimony tree gave essentially the same topology. Bootstrap values based on 1000 replicas are given as percentage numbers. Leaf labels follow our standard name code: four-letter species abbreviation, type of snRNA, NCBI gi number of the sequences where the snRNA is located, and snRNA gene coordinates. Species abbreviations are as follow: Aaeg, *Aedes aegypti*; Agam, *Anopheles gambiae*; Amel, *Apis mellifera*; Bmor, *Bombyx mori*; Dmel, *Drosophila melanogaster*; Dper, *Drosophila persimilis*; Dpse, *Drosophila pseudoobscura*; Dsec, *Drosophila sechellia*; Dsim, *Drosophila simulans*; Dyak, *Drosophila yakuba*; and Tcas, *Tribolium castaneum*. Some leaves represent identical multiple sequences denoted by a species abbreviation, followed by the number of identical sequences and a single letter code. (a) Agam|U5|19612093|3732939-3732816 || Agam|U5|19612093|3738569-3738446; (b) Dsim|U5|61711086|937-1062 || Dsec|U5|76494084|16278-16403; (c) Dpse|U5|40362012|24310-24185 || Dpse|U5|55845750|73589-73714; (d) Dpse|U5|40362012|29122-29247 || Dpse|U5|55845750|68777-68652 || Dper|U5|76469969|49677-49552; (e) Dpse|U5|40362012|59086-59211 || Dpse|U5|55845750|36653-36528; (f) Dpse|U5|40361995|65738-65614 || Dpse|U5|55845375|2085-2209; (g) Dmel|U5|27232088|2013728-2013604 || Dsim|U5|61738723|6567-6443 || Dsec|U5|76494655|18549-18673; (h) Dpse|U5|40362439|55864-55989 || Dpse|U5|55845417|57627-57502 || Dper|U5|76465975|1936-2061; (i) Dpse|U5|40360900|2979-2852 || Dpse|U5|55845704|65681-65554; (j) Bmor|U5|54108152|25278-25156 || Bmor|U5|54109135|218-340; (k) Bmor|U5|46790838|623-745 || Bmor|U5|54096763|3978-3856; (l) Bmor|U5|46585312|891-1014 || Bmor|U5|54046501|217-340 || Bmor|U5|54064195|209-332 || Bmor|U5|54081478|592-469.

FIGURE 2. Secondary structure of *A. mellifera* U5 snRNA as predicted using the covariance model approach (see Materials and Methods for details). (A) snRNA structure based on the gene sequence found in Group 16, position 448405–448287; variable positions up to the Sm protein-binding site (boxed) are denoted in gray and italics. (B,C) 3' ends of two other genes found in the honeybee genome: Group 16, position 390994–390874 (B), and Group 3, position 3078080–3078200 (C).

This observation is reinforced by another feature of the divergent *Drosophila* U11 snRNP. Some of the U11/U12 snRNP proteins first characterized in the human U11/U12 snRNP (Will et al. 2004) and conserved broadly through eukaryotes (Lorkovic et al. 2005) are absent from *Drosophila* (Schneider et al. 2004). We searched for genes encoding these U11/U12 proteins in the 11 insect genomes using translated BLAST searches (Table 2). All six *Drosophila* species seem to lack the 31K and 35K proteins, yet these same genes are present in the other five insect species, including the two mosquito genomes. The 25K protein is also likely to be missing specifically from the dipteran species. Homologs can be identified in *Apis*, *Tribolium*, and *Bombyx* but not in *Drosophila* or mosquito species. However, because the 25K protein is not well conserved, this particular negative result is not compelling.

U12-dependent introns

To investigate whether structural divergence of the U11/U12 snRNP is coupled with a loss of U12 introns, we examined the *A. mellifera* and 15 other metazoan genomes for the distribution of U12 introns. Using the statistical criteria described in Materials and Methods, we found that 57 out of 103,861 *A. mellifera* introns belong to the minor-type category, 49 of which are of the GT-AG type. These 57 introns are present in

56 genes, as one of the genes (ENSAPMT00000031526; a member of voltage-sensitive calcium channels) contains two U12 introns (introns 1 and 15) (Supplemental Table S1 at http://warta.bio.psu.edu/htt_doc/Projects/snRNA/). This large gene family is known to harbor many U12 introns (Wu and Krainer 1999), and our preliminary analysis suggests that the two introns present in the *Apis* gene are well conserved in different metazoan lineages (data not shown). The list of all the honeybee genes containing U12 introns is presented in Supplemental Table S2 at http://warta.bio.psu.edu/htt_doc/Projects/snRNA/. This count is fourfold higher than the number observed in Diptera (*D. melanogaster* and *A. gambiae*), but similar to that of the basal chordate *Ciona intestinalis* (Table 3).

Although the number of U12 introns in vertebrates is higher, this most likely reflects extensive gene duplications in the vertebrate lineage. The honeybee U12 introns are present in 55 unique gene families. Interestingly, while 41 of these gene families have at least one U12-intron-containing homolog in a vertebrate genome, only six of them have a homolog with a U12 intron in *D. melanogaster* (see Table 4). While these data are subject to minor errors due to inconsistencies in the annotation of genes between species,

FIGURE 3. Sequence logos of insect snRNA gene promoters. They were created with WebLogo (Crooks et al. 2004) with the promoters corresponding to the genes presented in Table 1 for each species.

FIGURE 4. U11 snRNA secondary structures from five insect species and humans. *Drosophila* and human structures were redrawn after Schneider et al. (2004). All other structures were inferred based on covariance models as described in Materials and Methods.

they indicate that >63% of the honeybee U12 introns are present in vertebrates but absent from *Drosophila*. Altogether, these data imply that most U12 introns present in

the last common ancestor of flies, bees, and vertebrates have been lost in the Dipteran lineage.

One of the most interesting metazoan introns resides between exons 2 and 3 of the *prospero* gene of *D. melanogaster*. This U12 intron contains active splice sites for a U2 spliceosome that are alternatively used, an arrangement referred to as a “twintron.” The alternative splicing of the *prospero* gene is temporally regulated during fly development (Scamborova et al. 2004). The U12 intron appears to be ancestral and is present in vertebrates, where it is not alternatively spliced (Oliver et al. 1993), and the U2 intron signals are not present. We

TABLE 2. Observation of identified U11/U12 snRNP proteins in various insect species

Protein	Query gi ^a	<i>Tribolium</i>	<i>Bombyx</i>	<i>Apis</i>	Mosquito	<i>Drosophila</i>
20K	9506863	Yes	Yes	Yes	Yes	Yes
25K	13443018	Yes	Yes	Yes	No	No
31K	51243065	Yes	No	Yes	Yes	No
35K	5902144	Yes	Yes	Yes	Yes	No
65k	40538732	Yes	Yes	Yes	Yes	Yes

Either *D. melanogaster* (20K, 65K) or human (25K, 31K, and 35K) proteins were used as queries in a BLAST search (TBLASTN) against 11 insect genomes.

^aNCBI's protein gi number.

TABLE 3. Number of genes and introns that have been analyzed in the genomes investigated, and results of identification of U12 introns

Species name	Total number of genes	Total number of introns	Number of U12 introns	U12-intron containing genes	U12 intron frequency (%)
<i>Anopheles gambiae</i>	13,046	41,041	15	15	0.037
<i>Apis mellifera</i>	13,448	103,861	57	56	0.055
<i>Ciona intestinalis</i>	10,706	114,226	81	80	0.071
<i>Drosophila melanogaster</i>	11,399	48,251	14	14	0.029
<i>Homo sapiens</i>	21,190	202,715	422	395	0.208

examined the *A. mellifera* gene in order to explore the origins of the twintron arrangement. The *Apis* gene perfectly conserves the U12 splice sites, but not the U2 sites. Although it is possible that nonorthologous cryptic sites are used for U2 splicing in the bee (and ancestral insects), we hypothesize that the twintron arrangement is a recent development in the Dipteran lineage consistent with the switch from U12 to U2 splicing that must be occurring for other introns.

Coordinated divergence of the U12 spliceosome and loss of U12-dependent introns

Representative Lepidopteran (*Bombyx*), Coleopteran (*Tribolium*), and Hymenopteran (*Apis*) genomes all conserve genes for components of the minor spliceosome that are missing in *Drosophila*. In addition, the honeybee genome has many more U12 introns than do Dipteran (fly or mosquito) species, indicating that divergence of the U12 spliceosome in Diptera is associated with the loss of U12 introns. These coordinated changes indicate that U12 introns are being lost from genomes that remove them inefficiently, as has been described for *Drosophila* (Patel et al. 2002). For a number of other genes, we have noted that bee U12 introns are either cleanly lost or replaced by U2 introns. In the case of *prospero*, the twintron arrangement may represent an evolutionary intermediate where a poorly spliced U12 intron serves as an alternative to a poorly placed U2 intron. The *Drosophila* genome has a total of only 14 U12 introns, and the genes for three U11/U12 spliceosomal proteins have been lost. It is tempting to speculate that this state represents an intermediate on the path toward the complete loss of U12 introns observed for species such as *C. elegans*.

MATERIALS AND METHODS

Sequence data

The genome assembly Amel_3.0 provided by the Human Genome Sequencing Center at Baylor College of Medicine (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Amelifera/) was used for finding and annotating the honeybee snRNA genes and their promoters.

For interspecies comparison, NCBI's "BLAST with arthropoda genomes" server was used (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=insects). At the time of this analysis (January 2006), 11 insect genomes were available: *Aedes aegypti*, *Anopheles gambiae* str. PEST*, *Apis mellifera**, *Bombyx mori*, *Drosophila melanogaster**, *Drosophila persimilis*, *Drosophila pseudoobscura*, *Drosophila sechellia*, *Drosophila simulans*, *Drosophila yakuba*, and *Tribolium castaneum* (*denotes completed genomic sequence).

Annotation of the snRNA genes

The annotation of the honeybee snRNA genes involved two steps. First, sequences of all snRNAs (U1, U2, U4, U4atac, U5, U6, U6atac, U11, and U12) of *D. melanogaster* were used as queries against the *A. mellifera* genome assembly 3.0. NCBI's BLAST was used with the following parameters: -r 5 -q -4 -G 10 -E 6 -W 7 -FF -X 40 -y 20 -Z 100 -e 0.1. In the next step, the nucleotide sequences around each hit were extracted and used as input for the INFERNAL software (Eddy 2002).

The snRNA gene number in 11 insect genomes was estimated based on BLAST hits using *Drosophila* snRNAs as queries and NCBI's insect genomes server (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=insects). The BLAST parameters used for finding *Apis* genes were used again here. In the case of very short hits or no hit at all, other insect snRNA sequences were used as queries. Functionality of the determined genes was assessed by

TABLE 4. Conservation pattern of homologous U12 containing genes among *A. mellifera* and other taxa

<i>Apis mellifera</i>	<i>Anopheles gambiae</i>	<i>Drosophila melanogaster</i>	<i>Ciona intestinalis</i>	<i>Homo sapiens</i>	Number of gene families
+	—	—	—	—	14
+	—	—	—	+	27
+	—	+	—	+	3
+	—	+	+	+	3
+	—	—	+	+	5
+	+	—	—	+	2
+	+	—	+	+	1
Total					55

Homology information among U12-containing genes was based on each gene's gene family ID, assigned by ENSEMBL. (+/—) Presence/absence of U12-containing genes in that taxon. Each row represents a pattern where U12-containing genes are present in a certain combination of taxa. The last column is the number of gene families that shows that particular pattern.

the integrity of the gene and by the presence of the PSEA element at the expected distance from the transcription start site of a gene.

Phylogenetic analysis

We investigated the mode of evolution of spliceosomal snRNA genes by constructing a phylogenetic tree of genes specifying each of the U snRNAs found in the major spliceosome. For that purpose, we aligned all genes assessed to be functional (see above) using ClustalW (Thompson et al. 1994) and adjusted the alignment manually where necessary. In the case of U5 genes, we used the secondary structure of the second stem-loop as a guide for aligning the variable 3' end for *Drosophila* sequences. The neighbor-joining method with 1000 bootstrap replicates was then used to construct the trees.

snRNA secondary structure prediction

snRNA secondary structures were drawn based on the INFERNAL alignment with manual adjustment when necessary. In some cases, for example, the *Anopheles* U11 snRNA, the structure was drawn with the aid of Mfold modeling software (<http://www.bioinfo.rpi.edu/applications/mfold/old/rna/>; Zuker 2003) run on fragmented sequence. Simply, the sequence was divided into known functional domains (defined by conserved segments), and domains were folded separately.

Detection of proximal sequence element A (PSEA)

One hundred nucleotides upstream of each snRNA gene were subjected to an initial motif search with MEME software (<http://meme.sdsc.edu/meme/intro.html>; Bailey and Elkan 1994). Next, every promoter sequence was analyzed individually to overcome the limitations of MEME software, given that the expected conserved motifs are short. Sequences of the promoters from genes that are believed to be functional (Table 1) were used to construct the PSEA profiles using WebLogo (Crooks et al. 2004).

U12-dependent introns

Intron position information for 16 metazoan genomes (*Anopheles gambiae*, *Apis mellifera*, *Bos taurus*, *Canis familiaris*, *Ciona intestinalis*, *Drosophila melanogaster*, *Danio rerio*, *Fugu rubripes*, *Gallus gallus*, *Homo sapiens*, *Monodelphis domestica*, *Mus musculus*, *Pan troglodytes*, *Rattus norvegicus*, *Tetraodon nigroviridis*, and *Xenopus tropicalis*) was downloaded from Ensembl database version 36, December 2005 (<http://www.ensembl.org/index.html>). Each intron was represented by 200-nt sequence windows centered at the 5'- and 3'-splice junctions, respectively. All introns containing an ATCC string at position +3 from the 5'-splice junction were selected as U12 intron candidates. Noncandidate introns with R (A or G) at +1 and G at +5 were used as a U2 intron training set. The U12 intron training set was composed of 46 U12 introns from a variety of higher eukaryotes (Burge et al. 1998). Weight matrices of the 5'-splice site and branch point site (BPS) for both U12 and U2 introns were created from the two training sets, and were used to calculate log-odds ratios for all candidates and introns in the two training sets. The mean and standard deviation of log-odds ratio values were calculated over all training sequences for the 5'-splice site and BPS, respectively, and then were used to normalize the

log-odds ratios of all introns that have been evaluated. Normalized BPS scores of the training sequences were plotted, and since they followed a normal distribution, we chose a Z value of 2.31, which corresponds to a 99% confidence interval, as a threshold BPS score to classify U12 intron candidates. Two additional criteria were applied: at least one adenine must be present at the branch point site and the distance from the branch point to the 3' junction has to be between 9 and 36 nt. Homology information of U12-containing genes was inferred based on the Ensembl gene family assignment, which is defined by the Markov clustering algorithm described by Enright et al. (2002).

ACKNOWLEDGMENTS

C.-F.L. and V.G. were partially supported by the Center for Comparative Genomics and Bioinformatics, part of Penn State's Huck Institutes of the Life Sciences. S.M.M. was partially supported by NSF award 0544309.

Received August 8, 2006; accepted September 26, 2006.

REFERENCES

- Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–2195.
- Bailey, T. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28–36. AAAI Press, Menlo Park, CA.
- Barbosa-Morais, N.L., Carmo-Fonseca, M., and Aparicio, S. 2006. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res.* **16**: 66–77.
- Berget, S.M., Moore, C., and Sharp, P.A. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci.* **74**: 3171–3175.
- Burge, C.B., Padgett, R.A., and Sharp, P.A. 1998. Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2**: 773–785.
- Chen, L., Lullo, D.J., Ma, E., Celniker, S.E., Rio, D.C., and Doudna, J.A. 2005. Identification and analysis of U5 snRNA variants in *Drosophila*. *RNA* **11**: 1473–1477.
- Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**: 1–8.
- Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. 2004. WebLogo: A sequence logo generator. *Genome Res.* **14**: 1188–1190.
- Dietrich, R.C., Incorvaia, R., and Padgett, R.A. 1997. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol. Cell* **1**: 151–160.
- Eddy, S.R. 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **3**: 18.
- Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**: 1575–1584.
- Gilbert, W. 1978. Why genes in pieces? *Nature* **271**: 501.
- Gonzalez, J., Casals, F., and Ruiz, A. 2004. Duplicative and conservative transpositions of larval serum protein 1 genes in the genus *Drosophila*. *Genetics* **168**: 253–264.
- Hall, S.L. and Padgett, R.A. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J. Mol. Biol.* **239**: 357–365.

- Hanley, B.A. and Schuler, M.A. 1991. Developmental expression of plant snRNAs. *Nucleic Acids Res.* **19**: 6319–6325.
- Hernandez, N. 2001. Small nuclear RNA genes: A model system to study fundamental mechanisms of transcription. *J. Biol. Chem.* **276**: 26733–26736.
- Hochleitner, E.O., Kastner, B., Frohlich, T., Schmidt, A., Luhrmann, R., Arnold, G., and Lottspeich, F. 2005. Protein stoichiometry of a multiprotein complex, the human spliceosomal U1 small nuclear ribonucleoprotein: Absolute quantification using isotope-coded tags and mass spectrometry. *J. Biol. Chem.* **280**: 2536–2542.
- The Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**: 931–949.
- Jurica, M.S. and Moore, M.J. 2003. Pre-mRNA splicing: Awash in a sea of proteins. *Mol. Cell* **12**: 5–14.
- Lai, H.T., Chen, H., Li, C., McNamara-Schroeder, K.J., and Stumph, W.E. 2005. The PSEA promoter element of the *Drosophila* U1 snRNA gene is sufficient to bring DmSNAPc into contact with 20 base pairs of downstream DNA. *Nucleic Acids Res.* **33**: 6579–6586.
- Levine, A. and Durbin, R. 2001. A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res.* **29**: 4006–4013.
- Lo, P.C. and Mount, S.M. 1990. *Drosophila melanogaster* genes for U1 snRNA variants and their expression during development. *Nucleic Acids Res.* **18**: 6971–6979.
- Lorkovic, Z.J., Lehner, R., Forstner, C., and Barta, A. 2005. Evolutionary conservation of minor U12-type spliceosome between plants and humans. *RNA* **11**: 1095–1107.
- Luo, H.R., Moreau, G.A., Levin, N., and Moore, M.J. 1999. The human Prp8 protein is a component of both U2- and U12-dependent spliceosomes. *RNA* **5**: 893–908.
- Mount, S.M. and Salz, H.K. 2000. Pre-messenger RNA processing factors in the *Drosophila* genome. *J. Cell Biol.* **150**: F37–F44.
- Mount, S.M. and Steitz, J.A. 1981. Sequence of U1 RNA from *Drosophila melanogaster*: Implications for U1 secondary structure and possible involvement in splicing. *Nucleic Acids Res.* **9**: 6351–6368.
- Nei, M. and Rooney, A.P. 2005. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**: 121–152.
- Oliver, G., Sosa-Pineda, B., Geisendorf, S., Spana, E.P., Doe, C.Q., and Gruss, P. 1993. Prox 1, a prospero-related homeobox gene expressed during mouse development. *Mech. Dev.* **44**: 3–16.
- Patel, A.A., McCarthy, M., and Steitz, J.A. 2002. The splicing of U12-type introns can be a rate-limiting step in gene expression. *EMBO J.* **21**: 3804–3815.
- Piontkivska, H., Rooney, A.P., and Nei, M. 2002. Purifying selection and birth-and-death evolution in the histone H4 gene family. *Mol. Biol. Evol.* **19**: 689–697.
- Scamborova, P., Wong, A., and Steitz, J.A. 2004. An intronic enhancer regulates splicing of the twintron of *Drosophila melanogaster* prospero pre-mRNA by two different spliceosomes. *Mol. Cell. Biol.* **24**: 1855–1869.
- Schneider, C., Will, C.L., Makarova, O.V., Makarov, E.M., and Luhrmann, R. 2002. Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. *Mol. Cell. Biol.* **22**: 3219–3229.
- Schneider, C., Will, C.L., Brosius, J., Frilander, M.J., and Luhrmann, R. 2004. Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein particle in *Drosophila*. *Proc. Natl. Acad. Sci.* **101**: 9584–9589.
- Sharp, P.A. and Burge, C.B. 1997. Classification of introns: U2-type or U12-type. *Cell* **91**: 875–879.
- Sierra-Montes, J.M., Pereira-Simon, S., Smail, S.S., and Herrera, R.J. 2005. The silk moth *Bombyx mori* U1 and U2 snRNA variants are differentially expressed. *Gene* **352**: 127–136.
- Tarn, W.Y. and Steitz, J.A. 1996. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell* **84**: 801–811.
- 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.
- Will, C.L., Schneider, C., Hossbach, M., Urlaub, H., Rahut, R., Elbashir, S., Tuschl, T., and Luhrmann, R. 2004. The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *RNA* **10**: 929–941.
- Wu, Q. and Krainer, A.R. 1999. AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Mol. Cell. Biol.* **19**: 3225–3236.
- Wu, H.J., Gaubier-Comella, P., Delseny, M., Grellet, F., Van Montagu, M., and Rouze, R. 1996. Non-canonical introns are at least 10⁹ years old. *Nat. Genet.* **14**: 383–384.
- Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**: 3406–3415.

RNA

A PUBLICATION OF THE RNA SOCIETY

Spliceosomal small nuclear RNA genes in 11 insect genomes

Stephen M. Mount, Valer Gotea, Chiao-Feng Lin, et al.

RNA 2007 13: 5-14 originally published online November 9, 2006

Access the most recent version at doi:[10.1261/rna.259207](https://doi.org/10.1261/rna.259207)

References

This article cites 40 articles, 17 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/13/1/5.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:
<http://rnajournal.cshlp.org/subscriptions>