Preprint manuscript No.
(will be inserted by the editor)

Manuela Marz, Toralf Kirsten, Peter F. Stadler

Evolution of Spliceosomal snRNA Genes in Metazoan Animals

March 20, 2008

Abstract While studies of the evolutionary histories of protein families are common place, little is known on noncoding RNAs beyond microRNAs and some snoRNAs. Here we investigate in detail the evolutionary history of the 9 spliceosomal snRNA families (U1, U2, U4, U5, U6, U11, U12, U4atac, and U6atac) across the completely or partially sequenced genomes of metazoan animals.
Representatives of the five major spliceosomal snRNAs were found in all genomes. None of the minor splicesomal snRNAs was detected in Nematodes and in the shotgun traces of *Oikopleura dioica*, while in all other animal genomes at most one of them is missing. Although snRNAs are present in multiple copies in most genomes, distinguishable paralog groups are not stable over long evolutionary times, although they appear independently in several clades. In general, animal snRNA secondary structures are highly conserved, albeit in particular U11 and U12 in insects exhibit dramatic variations. An analysis of genomic context of snRNAs reveals that they behave like mobile elements, exhibiting very little syntenic conservation.

M. Marz
Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany
E-mail: manja@bioinf.uni-leipzig.de

T. Kirsten
Interdisziplinäres Zentrum für Bioinformatik, Härtelstrasse 16-18, D-04107 Leipzig, Germany
E-mail: kirsten@izbi.uni-leipzig.de,

P.F. Stadler
Bioinformatics Group, Department of Computer Science, and Interdisziplinäres Zentrum für Bioinformatik, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany;
RNomics Group, Fraunhofer Institute for Immunology and Cell Therapy, Leipzig;
Institute for Theoretical Chemistry and Structural Biology, University of Vienna, Währingerstrasse 17, A-1090 Wien, Austria; and
The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501
E-mail: studla@bioinf.uni-leipzig.de

1 Introduction

In most eukaryote lineages, introns are spliced out of protein-coding mRNAs by the spliceosome, a huge RNP complex consisting of about 200 proteins and five small non-coding RNAs [58]. These snRNAs exert crucial catalytic functions in the process [86,88,87] in three distinct splicing machineries. The *major spliceosome*, containing the snRNAs U1, U2, U4, U5 and U6, is the dominant form in metazoans, plants, and fungi, and removes introns with GT-AG (as well as rarely AT-AC and GC-AG) boundaries. Another class of "non-canonical" introns with AT-AC (and rarely GT-AG [71]) boundaries is excised by the *minor spliceosome* [61], which contains the snRNAs U11, U12, U4atac, U5, and U6atac. Just as the major spliceosome, the minor spliceosome is present across most eukaryotic lineages and traces back to an origin very early in the eukaryote evolution [9,44,65]. Recently it was found that the minor spliceosome can also act outside the nucleus and controls cell proliferation [35]. Functional and structural differences of two spliceosomes are reviewed in [89]. The third type of splicing the *SL-trans-splicing*, in which a "miniexon" derived from the non-coding spliced-leader RNA (SL) is attached to each protein-coding exon. The corresponding spliceosomal complex requires the snRNAs U2, U4, U5, and U6, as well as an SL RNA [24]. Due to the high sequence variation of the short SL RNAs, and the patchy phylogenetic distribution of SL-trans-splicing, the evolutionary origin(s) of this mechanism, which is active at least in chordates, nematodes, cnidarians, euglenozoa, and kinetoplastids, is still unclear.

Previous studies on the evolutionary origin of the spliceosomes have been performed predominantly based on homology of the most important spliceosomal proteins. Thus relatively little detail is known on the evolution of the snRNA sequences themselves beyond the homology of nine famlilies of snRNAs across all eukaryotes studies so far [73,69,10,44,9,65]. This may come as a surprise since it has been known for more than a decade that at least all of the snRNAs of the major spliceosome appear in multiple copies and that these paralogs are differentially regulated in at least some species, see e.g. [43,80,79,5,52]. Very recently, however, some of these variants have been studied in more details, see e.g. [64,8,39,77,29,78] and the references therein. The only systematic study that we are aware of is the recent comprehensive analysis of 11 insect genomes [53] which reported that phylogenetic gene trees of insect snRNAs do not provide clear support for discernible paralog groups of U1 and/or U5 snRNAs that would correspond to the variants with tissue-specific expression patterns. Instead, the analysis supports a concerted mode of evolution and/or extreme purifying selection, a scenario previously described for snRNA evolution [42,40,57].

In this contribution we extend the detailed analysis of the nine spliceosomal snRNAs to metazoan animals. In particular in mammals, the analysis is complicated by high copy number of snRNAs of the major spliceosome and an associated large number of pseudogenes [13]. We focus here on four questions: (1) Is there evidence for discernible paralog groups of snRNAs in some clades? A dominating mode of concerted evolution does not necessarily prevent this, as demonstrated by the existence of two highly diverged copies of both LSU and SSU rRNA in Chaetognatha [83,60], which is probably associated with a duplication of the entire rDNA cluster. (2) Are there clades with deviant snRNA structures? The prime example for a highly divergent snRNA is the U11 in a subset of the insects [69]. (3) Are there interpretable trends in the copy number of snRNAs across metazoa? (4) How mobile are snRNA genes relative to the “background” of protein coding genes? In other words, to what extent are some or all of the snRNA genes off-springs of a locus that remains stably linked to its context over large time-scales.

2 Materials and Methods

2.1 Sequence Data

Known snRNA sequences were retrieved from `Genbank` [4], `Rfam` [23], and in some cases extracted directly from the literature. Genomic DNA sequences were downloaded from the websites of `ensembl`, the Joint Genome Institute, the Sanger Institute, WormBase, the Genome Sequencing Center, UCSC, CAF1, Broad Institute, BGI, and the NCBI trace archive. For some species, we also performed non-exhaustive searches in the `NCBI Trace Archive` using `megablast`. Details on the dataset can be found in the Electronic Supplement.[1]

Over all, the published experimental evidence on metazoan snRNAs is very unevenly distributed. For example, a large and phylogentically diverse set of U2 snRNA sequences is reported in [20], while most other snRNAs have mostly been reported for a few model organisms only. A recent experimental screen for snRNAs in *Takifugu rubripes* [55] resulted in copies of eight snRNAs families. U4atac was missing, but a plausible candidate can easily be found by `blast`. Only a few sequences of minor spliceosomal snRNAs have been reported so far, mostly in a few model mammals [82] and in Drosophilids [69,53].

2.2 Homology Search

In a first automatic step we used a local installation of `NCBI blast` (v.2.2.10) with default parameters and $E < 10^{-6}$ to find candidate sequences in closely related genomes. If successful, the results of this search were aligned to the query sequence using `clustalw` (v.1.83). After a manual inspection using `clustalx`, the consensus sequence of the alignment was again used as a blast query with the same E-value cutoff.

If this automatic search was not successful, the best `blast` hit(s) were retrieved and aligned to a set of known snRNAs from related species. Candidate sequences were retained only when a visual inspection left no doubt that they were true homologs. This manual analysis step included a check whether the phylogenetic position of the candidate sequence in a neighborjoining tree was plausible, taking into account that the sequences are short and some parts of the alignments are of low quality.

In cases where no snRNA homologs were found as described above, we searched the genome again with a much less stringent cutoff of $E < 0.1$ (or even larger in a few cases) and extracted all short hits together with 200nt flanking sequence. We used Sean Eddy’s `rnabob` with a manually constructed structure model to extract a structure-based match within the selected regions and attempted to align the candidate sequences manually to a structure-annotated alignment of snRNAs in the `emacs` editor using the `ralee` mode [22].

Finally, the resulting alignments of snRNAs where used to derive search patterns for `RNAmotif` [45] and `erpin` [19]. To this end, the consensus structure of the alignment was computed using `RNAalifold` [30] and converted into a form suitable as input for the two search programs.

2.3 Structure Models

Structure annotated sequence alignments were manually modified in the `emacs` text editor using the `ralee` mode [22] to improve local sequence-structure features based on secondary structure predictions for the individual sequences obtained from `RNAfold` [31]. Consensus structures were then computed using `RNAalifold` [30]. The structure models are compiled in the Electronic Supplement.

2.4 Upstream Region Analysis

With `MEME` (v.3.5.0) we discovered motifs upstream of the sequences for analysis of regulators and other possible depencies. They were manually compared with previously published sequence elements. We visually compared the `MEME` -patterns with the upstream elements in related species from

[1] `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-001/`

the following literature sources: [26] (general motifs), [14, 82, 2, 38] (human), [36, 5] (chicken), [53] (insects), [77] (*Bombyx mori*), [81] (*Strongylocentrotus purpuratus*), [84] (*Caenorhabditis elegans*).

2.5 Phylogenetic Analysis

Since the snRNA sequences are short and in addition there are several highly variable regions, we use split decomposition [1] and the neighbor net [7] algorithm (as implemented as part of the `SplitsTree4` package [33]) to construct phylogenetic networks rather than phylogenetic trees. The advantage of these method is that they are very conservative and that the reconstructed networks provide and easy-to-grasp representation of the considerable noise in the sequence data.

2.6 Synteny Information

In order to assess whether snRNA genes are mobile in the genome, we determined their flanking protein-coding genes. We used the `ensembl compara` annotation [17] to retrieve homlgous proteins in other genomes and compared whether these homologs also have adjacent snRNAs. For consistency, this analysis is performed based on `ensembl` (release 46) [32] using the data integration platform `BioFuice` [34]. More precisely, for each human snRNA G we examined that the relation of the left homologous $L_H(G)$ and right homologous $R_H(G)$ of flanking protein coding genes $L(G)$ and $R(G)$ on both sides of G. We only considered annotations in $L_H(G)$ and $R_H(G)$, resp., if the sequence distance between G_H and $L_H(G)$ and $R_H(G)$ was not more than twice (five times for mammals) the distance between G and $L(G)$ and $R(G)$.

3 Results

3.1 Homology Search

Tab. 1 summarizes the results of the sequence homology search. We find that, with few exceptions, `blast`-based homology search strategies are in general sufficient to find homologs of all nine spliceosomal snRNAs in most metazoa genomes. The procedure is hard to automatize, however, since in many cases the initial `blast` hits have poor E-values, while a multiple sequence alignment then leaves little doubt that a true homolog has been found. This is in particular true for searches bridging large evolutinary distances, in particular when the search extends beyond bilateria.

With very few exceptions we find multiple copies of all five major spliceosomal RNAs that exhibit the typical snRNA-like promoter elements and are hence mostly likely functional copies of the genes. The snRNA copy numbers vary substantially between different clades. The genus *Caenorhabditis*, for example, is set apart from other nematodes by a two to threefold increase in the number of major spliceosomal snRNAs. In contrast, the snRNAs of the minor spliceosome are in most cases single-copy genes.

Many genomes, most notably mammalian genomes, contain a sizeable number of major snRNA pseudogenes. Table 1 therefore lists only candidates that have plausible snRNA-like promoter structure, that fit the secondary structures of snRNAs in related species, and that exhibit strong sequence similarity in the unpaired regions of the molecule. These are rather restrictive criteria. In the Electronic Supplement, we therefore provide a corresponding table that is based only on sequence homology.

It is surprisingly difficult to compare the present snRNA survey with previous reports on vertebrate snRNAs. The main reason for discrepancies in the count of snRNAs is that distinguishing functional snRNAs from pseudogenes is still an unsolved problem. In this contribution, we use a very stringent criterion by insisting on a recognizable promoter structure. In some cases, however, it is known that snRNAs have internal promoters only [85]. These cases constitute false negatives in Tab. 1. On the other hand, much of the published literature considers sequence similarity to the known functional genes as the only criterion, thus most likely leading to the inclusion of a substantial fraction of pseudogenes. For instance, ref. [67] counts 16 U1, 6 U2 and 44 U6 snRNAs in the human genome (compared to our 8, 3, and 7, resp.), while [14] report 5-9 U6 snRNA genes, consistent with our list. Similarly, only a fraction of the major spliceosomal snRNAs reported for the chicken genome in [27] pass our promoter analysis.

For Drosophilids, on the other hand, our analysis is almost identical to the results of [53, Tab.1] and the data reported in [77]. Furthermore, we come close the results of a comparative genomics screen for non-coding RNAs in *C. elegans* [49], which reported 12 U1, 19 U2, 5 U4, 13 U5, and 23 U6, i.e., only a few more candidates than our present purely homology-based approach. A comparative screen of the two *Ciona* species for evolutionary conserved structured RNAs [48] missed a small number of snRNA genes that we indentified as most likely functional ones.

In a few species we failed to identify individual major spliceosomal snRNAs. Minor spliceosomal snRNAs are more often missing. In those cases where only some of the major or minor snRNAs remain undetected, the missing family member most likely escaped our detection procedure for one of several reasons:

(1) in the case of unassembled incomplete genomes for which only shotgun reads were searched, the snRNA may be located in the not yet sequenced fraction of the genome or it might not be completely contained within at least one single shotgun read.

(2) The snRNA in question may be highly derived in sequence. (For instance, the U11 snRNA in Drosophilids [69] cannot be found by be a simple `blast` search starting from non-insect sequences. It can be found however, by the combination of very un-specific blast and subsequent structure search as described in section 2.2.)

Table 1 Approximate copy number of snRNA genes.
We list here only those sequences that (1) are consistent with the secondary structures of related snRNAs, (2) show substantial sequence conservation in the unpaired regions of these structures, and (3) have recognizable promoter motifs. In some cases none of the candidates satisfies all these criteria. If there are nevertheles clear homologuous sequences. Entries of the form $^{S}0$ and $^{P}0$ indicate that there is homolgous sequence which however lacks structural similarity or recognizable promoter elements. The quality of the genome assembly is marked by the following sysmbols: △ – Traces, □ – Contigs, ◇ – Scaffolds, ♠ – Chromosoms.

Coverage	Species	U1	U2	U4	U5	U6	U11	U12	U4atac	U6atac
◇	*M. brevicollis*	0	0	0-1	0-2	1	0	0	0	0
△	*Reniera sp*	2	0-1	2	3	2	1	1	0	3
◇	*Trichoplax adhaerens*	1	1	1	1	2	1	1	1	1
◇	*N. vectensis*	2	2	4	5	3	3	3	1	2
△ 7.45-8.33X	*H. magnipapillata*	4	2	5	7	4	1	1	0	2
△ 0.05X	*A. millepora*	0	2	0	2	2	0	0	0	0
△ 0.047X	*A. palmata*	1	0	0	0	1	0	0	0	0
◇	*S. mansoni*	3	3	1	2	9	0	1	0	0
□	*S. mediterannea*	2	$^{P}0$	3	2	2	0	0	0	0
△ 13.03X	*L. gigantea*	3	8	11	2	7	2	1	0	2
△ 0.05X	*B. glabrata*	$^{S}0$	2	0	1	$^{S}0$	0	0	0	0
△ 0.54X	*P. lobata*	1	1	1	0	0	0	0	0	0
△ 0.012X	*E. scolopes*	$^{SP}0$	1	0	0	0	0	0	0	0
△ 4.48X	*A. californica*	4	2	4	10	8	1	1	0	1
◇	*C. capitata*	5	2	1	4	2	1	1	1	1
◇	*H. robusta*	6	8	4	7	4	0	1	1	1
△ 0.23X	*H. bacteriophora*	2	2	0	2	1	0	0	0	0
△ 11.33X	*B. malayi*	3	3	1	1	2	1	0	0	—
△ 12.15X	*T. spiralis*	1	5	2	3	1	1	0	0	0
△ 11.24X	*P. pacificus*	2	2	4	4	7	1	0	0	0
□	*C. brenneri*	19	19	10	19	25	0	0	0	0
□	*C. remanei*	14	11	5	13	15	0	0	0	0
△ 10.18X	*C. japonica*	16	15	4	14	7	0	0	0	0
♠	*C. elegans*	10	17	4	9	15	0	0	0	0
♠	*C. briggsae*	9	10	4	10	22	0	0	0	0
△ 3.29X	*D. pulex*	5	6	4	9	8	1	1	$^{PS}0$	1
△ 11.81X	*P. humanus*	3	4	1	2	1	1	1	0	1
□	*N. vitripennis*	7	4	3	5	5	1	2	1	2
△ 2.58X	*I. scapularis*	4	4	3	4	3	0	1	0	1
△ 1.6X	*A. pisum*	2	3	0	2	3	1	1	0	1
◇	*A. mellifera*	5	3	2	3	3	1	1	1	1
◇	*B. mori*	5	6	3	5	4	1	1	1	2
△ 0.75X	*T. castaneum*	5	5	2	6	3	1	1	0	1
♠	*A. gambiae*	7	7	2	5	2	2	1	1	1
♠	*D. melanogaster*	5	6	3	7	3	1	1	1	1
♠	*D. ananassae*	9	8	2	4	2	1	1	1	1
♠	*D. erecta*	8	9	3	7	4	1	1	1	1
♠	*D. grimshawi*	7	6	3	7	3	1	1	1	2
♠	*D. mojavensis*	6	8	3	6	3	1	1	1	1
♠	*D. persimilis*	7	7	3	7	3	1	1	1	1
♠	*D. pseudoobscura*	7	7	3	6	3	1	1	1	1
♠	*D. sechellia*	7	6	3	7	3	1	1	1	1
♠	*D. simulans*	8	6	3	8	3	1	1	0	1
♠	*D. virilis*	6	8	3	6	2	1	1	2	1
♠	*D. willistoni*	8	9	3	8	$^{P}0$	1	1	1	0
♠	*D. yakuba*	8	7	3	8	3	1	1	1	1

Coverage	Species	U1	U2	U4	U5	U6	U11	U12	U4atac	U6atac
◇	*S. purpuratus*	5	7	9	8	3	2	3	1	1
△ 3.77X	*S. kowalevski*	7	4	4	5	4	1	2	0	3
◇	*C. savignyi*	3	2	3	7	2	1	1	1	1
◇	*C. instestinalis*	1	1	3	5	2	1	1	1	1
△ 7.8X	*O. dioica*	1	6	2	7	4	0	0	0	0
◇	*B. floridae*	8	3	5	9	4	1	1	0	1
△ 6.19X	*P. marinus*	6	5	8	9	5	1	2	$^{PS}0$	3
♠	*D. rerio*	5	4	4	7	3	1	1	1	1
♠	*O. latipes*	4	2	2	4	4	1	1	1	1
♠	*G. aculeatus*	6	2	4	7	3	1	1	1	1
◇	*F. rubripes*	5	5	3	6	4	1	1	1	1
♠	*T. nigroviridis*	4	5	3	5	2	1	1	0	1
◇	*X. tropicalis*	5	1	3	2	5	1	1	1	2
♠	*G. gallus*	1	1	1	2	4	1	1	1	1
△ 8.34X	*T. guttata*	2	5	2	3	2	1	1	0	1
△ 8.24X	*A. carolinensis*	14	6	2	6	5	1	2	1	1
♠	*O. anatinus*	5	2	2	4	6	1	1	1	1
♠	*M. domestica*	7	4	2	5	6	1	$^{PS}0$	1	1
♠	*M. musculus*	7	5	1	6	7	1	2	1	2
♠	*R. norvegicus*	4	10	1	4	5	4	1	1	1
♠	*C. familiaris*	6	5	2	4	5	1	1	1	1
♠	*B. taurus*	7	8	2	5	6	2	1	1	1
♠	*P. tropicalis*	7	2	2	7	8	1	1	3	1
♠	*H. sapiens*	8	3	2	5	7	1	1	3	1

(3) In some cases we list a "0" in Tab. 1 even though there is recognizable sequence homology in the genome. In these cases we were not able to identify the snRNA-like promoter elements and/or the secondary does not fit the expectation. These cases marked in the table.
(4) It is conceivable that some species have lost a particular snRNA and replaced it by corresponding snRNA from the other spliceosome. The observation that U4 may function in both the major and minor spliceosomes [74] shows that such a replacement mechnism might indeed be evolutionarily feasible.

In our data set, we most frequently were unable to find a U4atac homolog. We cannot know, of course, whether we missed these cases due to poor sequence conservation or due to loss of the gene. For instance, we did not recover a plausible U4atac candidate for the hemichordate *Saccoglossus kowaleski* despite the fact that the U4atac sequence of the sea urchin *Strongylocentrotus purpuratus* was easily retrieved.

Surprisingly, we found neither a canonical U6 nor a canonical U6atac in *Drosophila willistoni*. A highly derived U6 homolog has no recognizable snRNA-like promoter structure and exhibits substantial deviations from the consensus structure, see section 3.5. Similarly, the U4atac candidate from *Daphnia pulex* deviates substantially from other arthropod sequences. It is possible that in some or all of these cases the snRNA is present in the genome but is not contained in the currently available genomic sequence data. This is most likely the case for the missing minor spliceosomal snRNAs of *Ixodes scapularis*, *Pediculus humanus*, or *Drosophilia willistoni*.

In some cases, however, we failed to identify all four minor spliceosomal snRNAs. Consistent with previous work [61] we found no convincing homologs of the minor spliceosomal snRNAs U11, U12, U4atac, or U6atac in any of the nematode genomes, suggesting that the minor spliceosome was lost early in the nematode lineage. Nevertheless, we find some `blast` hits for minor spliceosomal snRNAs in some nematode genomes.

Our analysis furthermore suggests the possible loss of the minor spliceosome in *Oikopleura dioica*, while a complete complement of minor spliceosomal snRNAs was found in the genus *Ciona*. It is unclear, however, whether this is an artifact due to limiations of available shotgun traces.

Our survey provides evidence that most metazoan clades for which genomic sequences are available have retained the minor spliceosome. For many groups, such as Annelida or Cnidaria, we are not aware of earlier references to the existence of minor spliceosome.

3.2 Specific Upstream Elements

The classical snRNA-specific PSE and TATA elements that have been described in detail for several vertebrates [26, 14] are highly conserved. This appers to be an exception rather than the rule, however: the snRNA upstream elements are highly diverse across metazoa. Our analysis agrees with the recent observation that in Drosophilids there is a rapid turnover in the upstream sequences. Even though the PSE is fairly well-conserved within Drosophilids, it already differs substantially between the major insect groups [53]. Similarly, within the nematodes conservation of upstream elements is limited to the genus level. In general, the PSE of U11, U12 and U4atac is much less conserved than their counterpart in major spliceosomal snRNA genes. For the purpose of this study, the relatively well-conserved elements were used to discriminate functional snRNAs from likely pseudogenes. We concentrated on PSE and TATA elements for this purpose because other snRNA-associated upstream elements, such as SPH, OCT, CAAT-box, GC-box, -35-element and *Inr* are even less well conserved:

A GC-box was identified in *Caenorhabditis* at a non-canonical position (about -68nt). These elements are different for each single snNRA class: U1 `GGACGG` (44/52 sites), U2 `TGGCCG` (38/60 sites) and for U5 `CGGCCG` (39/46 sites). However, also among a single snRNA this element varies a lot: insects have a U1 GC-box `GCGCTG` at about -75nt (15/39 sites). About half of the U6 sequences of basal deuterostomes show the CAAT-box motif `TGCCAAGAA` at the known position of -70nt. Interestingly, we find related motifs in the upstream region of Drosophilids U11 (`GACCAATAT`, -33nt) and other insects U5 snRNA (`TTCCAATCA`, -28nt) and . The Octamer motif (OCT, `ATTTGCAC`) was found in 6 of 7 sequences of basal deuterostomes at the known position of -54nt upstream of U6atac. However, in 12 of 14 Drosophilids sequences, the closely related motif `ATTTGCTT` was found at position -33nt. About 35nt upstream of U11 and U12 snRNAs of teleosts we found the motif `GTGACA` and `TGCACA`, respectively. The *Inr* element of U1 snRNA was found in each species. For teleost fishes and Drosophilids we found a complete set of this element for all snRNAs. However, the element show substantial sequence variations both between different genes in the same species and between homologous genes in different species. We refer to the Electronic Supplement for further details and lists of identified sequence elements.

3.3 Clusters of snRNA genes

In Mammalia, we observe linkeage of tandem copies of U2 snRNAs, see also [41,62], while there there are no clusters of distinct snRNAs. In *Drosophila*, there are surprisingly constant patterns of snRNA clusters: (a) U2-U5 clusters are observed 4-6 times per genome, (b) there are one or two U1-U2 clusters, and (c) 3-9 tandem copies of snRNAs. Two species deviated therefrom. In *D. ananassae*, we find no U2-U5 cluster, but instead 7 U1-U2, one U4-U5 cluster and 4 other tandem copies, while the *D. willistoni* lacks the U4-U5 cluster but contains 10 U2-U5 pairs and 6 tandem copies. Teleost fishes also have a common pattern: there are one or two U1-U2 pairs and 2-6 tandem copies. In general, however, snRNA do not appear in clusters throughout metazoan genomes.

In several species, linkeage of snRNAs with 5S rRNA has been observed [42,40,16,63,11,46]. We found only one further example of this type: in *Daphnia pulex* 5S and U5 snRNA are separated by only 308bp.

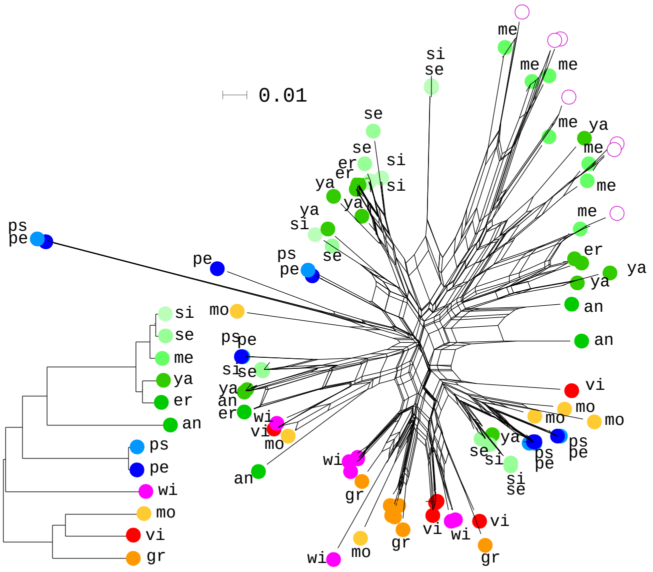

Fig. 1 Phylogenetic network of Drosophilid U5 snRNAs. The eight U5 snRNA reported by [8] are shown by white dots. me – *D. melanogaster*, er – *D. erecta*, si – *D. simulans*, se – *D. sechellia*, ya – *D. yakuba*, wi – *D. willistoni*, gr – *D. grimshawi*, mo – *D. mojavensis*, vi – *D. virilis*, pe – *D. persimilis*, ps – *D. pseudoobscura*, an – *D. ananassae*. The phylogenetic tree is adapted from ref. [15].

3.4 Phylogenetic Analysis and Paralogs

Like ribosomal RNAs, spliceosomal RNAs are subject to *concerted evolution* [28,68,21], i.e., one observes that paralogous sequences in the same species are more similar than orthologous sequences of different species. Multiple molecular mechanisms may account for this phenomenon: gene conversion, repeated unequal crossover, and gene amplification (frequent duplications and losses within family), see [40] for a review. In some cases, however, paralogs can escape from the concerted evolution mechanisms as exemplified by the two paralog groups of SSU rRNA in Chaetogatha [60].

Distinguishable snRNA paralogs that are often differentially expressed have previously been reported for a diverse collection of major spliceosmal snRNAs including U1 snRNAs in insects [43,64,77], Xenopus [12], and human [39], U2 snRNAs in *Dictyostelium* [29], sea urchin [80] and silk moth [77], U5 snRNAs in human [79], sea urchin [52], and Drosophilids [8], U6 snRNAs in silk moth [78] and human [85,14].

A phylogenetic analysis of the individual snRNA families nevertheless does not show widely separated paralog groups that are stable throughout larger clades. Fig. 1, for example shows that the U5 variants described by [8] do not form clear paralog groups beyond the closest relatives of *Drosophila melanogaster*. On the other hand, there is some evidence for distinguishable paralogs outside the melanogaster subgroup. The situation is much clearer for the Drosophilid U4 snRNAs, where three paralog groups can be distinguished, see Fig. 2. One group is well separated from the other two

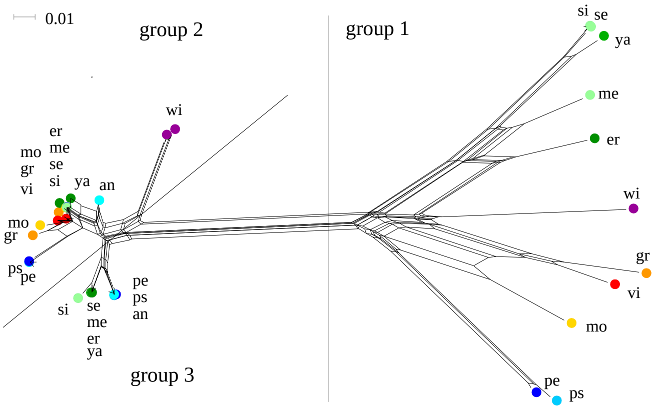

Fig. 2 Phylogenetic tree of insect U4 snRNAs. In this case we can distinguish three paralog groups within the Drosophilids. me – *D. melanogaster*, er – *D. erecta*, si – *D. simulans*, se – *D. sechellia*, ya – *D. yakuba*, wi – *D. willistoni*, gr – *D. grimshawi*, mo – *D. mojavensis*, vi – *D. virilis*, pe – *D. persimilis*, ps – *D. pseudoobscura*, an – *D. ananassae*.

Table 2 Paralog groups of major splicesomal snRNAs recognizable within major animal clades. The symbol •denotes clearly distinguishable paralog groups and refers to the supplemental material for details, ? indicates ambigous cases, = means that all paralogous genes have identical sequences.

Clade	U1	U2	U4	U5	U6
Annelids	–	–	–	–	=
Nematods	–	–	–	–	=
Caenorhabditis	–	–	–	•	=
Insects	–	–	–	–	=
Drosophilids	?	–	Fig.2	[8]	=
Teleosts	–	Fig.3a	Fig.3b	Fig.3c	–
Tetrapoda	–	–	–	–	–
Mammalia	–	–	–	•	–

and internally rather diverse. The other two groups are very clear distinguishable for the melanogaster and obscura group (see [15]). For *D. virilis*, *D. mojavensis*, *D. grimshawi* and *D. willistoni* we have two nearly identical copies instead of two different groups of genes.

Table 2 summarized the presence of recognizable paralog groups within major animal groups. Within the genus *Caenorhabitis* we find evidence for the formation of U5 paralog groups in *C. remanei*, *C. brenneri*, and *C. briggsae* to the exclusion of *C. elegans* and *C. japonica*. Evidence for paralog groups of U1 snRNA in Drosophilids remains ambiguous due to the small sequence differences.

In teleost fishes we find clearly recognizable paralog groups for U2, U4, and U5 snRNAs. Surprisingly, the medaka *Oryzias latipes* has only a single group of closely related sequences, despite the fact that for U4, the split of the paralogs appear to predate the last comman ancestor of zebrafish and fugu, Fig. 3.

Neither the two rounds of genome duplications at the root of the vertebrates nor the teleost-specific genome duplication has lead to recognizable paralog groups of snRNAs. In particular, minor snRNA genes are single-copy genes in teleosts.

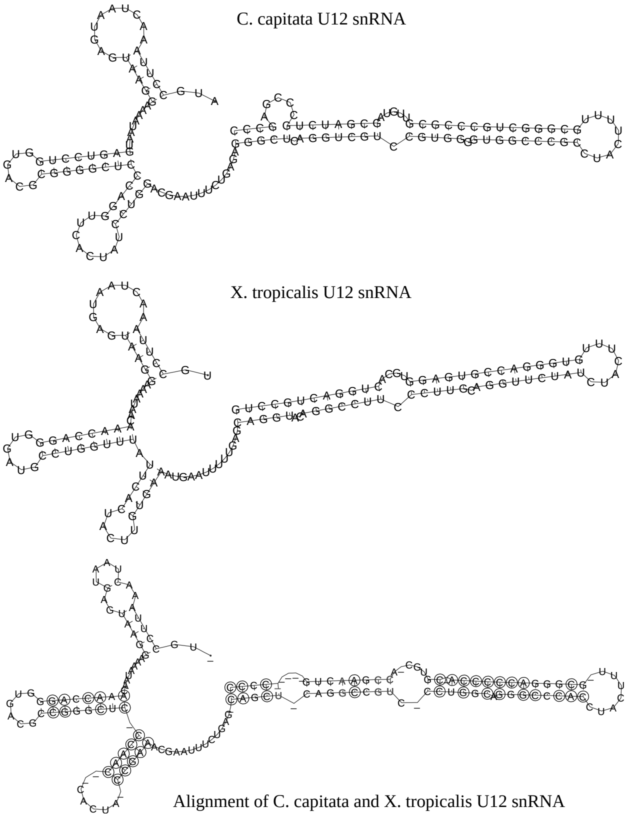

Fig. 4 Predicted secondary structures of *Capitella capitata*, *Xenopus tropicalis* and an alignment created with `RNAalifold` of both. Circles represent different bases and therewith compensatory mutations.

3.5 Secondary Structures

The spliceosomal snRNAs have evolutionarily well-conserved secondary structures [73]. These structures have received substantial interest in the past, as explified by the following non-exhaustive list of references covering a diverse set of animal species: *Homo sapiens* U1 [54], U2 [25], U4 [37], U5 [6,79], U6 [25], U11 [66,51,82], U12 [66,51,82] and U4atac [72], *Rattus norvegicus* U1 [37], U4 [37], U5 [37], *Gallus gallus* U4 [37], U5 [6], *Xenopus laevis* U1 [18], U2 [47], *Caenorhabditis elegans* U1, U2, U5, U4/U6 [84], *Drosophila melanogaster* U1 [54,56], U2 [56], U4 [56], U5 [56], U4atac/U6atac, U6atac/U12 [59], *Bombyx mori* U1 [76], U2 [75], *Asselus aquaticus* U1 [3], *Ascaris lumbricoides* U1, U2, U5, U4/U6 [70]. Large changes in snRNA structures over evolutionary time were recently reported for hemiascomycetous yeasts [50]. The comprehensive survey of snRNA sequences throughout metazoa set the stage for a comparably detailed analysis of metazoan snRNA structures. In order to asses structural variations, we contructed structure annotated sequence alignments of all snRNA families. These are provided as part of the electronic supplement.

In general we find that snRNA sequences vary more in paired regions than in the loops. The sequence variations almost exclusively comprises compensatory mutations that

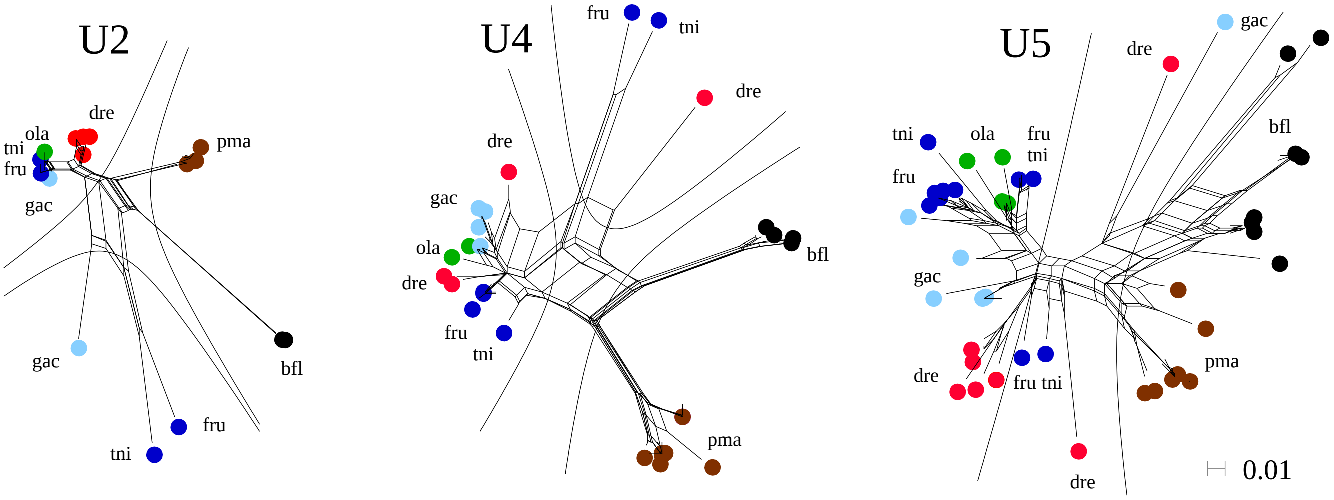

Fig. 3 Phylogenetic networks of teleost fish snRNAs. Species abbreviations: fru – *Fugu rubripes*, tni – *Tetraodon nigrovidis*, gac – *Gasterosteus aculeatus*, ola – *Oryzias latipes*, dre – *Danio rerio*, pma – *Petromyzon marinus*, bfl – *Branchiostoma floridae*.

leave the secondary structures intact. As an example, Fig. 4 shows the structures of the U12 snRNA of *Xenopus tropicalis* and *Capitella capitata*. The sequences have few paired nucleotides in common.

Structural variations are typically limited. In Fig. 5 we use the U1 snRNAs as a typical example for the evolutionary variation of snRNAs across the metazoa. Overall the structures are extremely well conserved with small variations in the length of the individual stems. With several notable exceptions this is true for all metazoan snRNAs.

As reported previously [8], the second stem of U5 snRNA shows some variations. More interestingly, the minor spliceosomal snRNAs tend to be derived in insects. This has been reported previously in particular for U11 in Drosophilids [69, 53]. We found substantial structural variations also for drosophilid U12 snRNAs: there are massive insertions in and after Stem III, while Stem I and II show mispairings. Furthermore, Stem II of U6atac is completely deleted in all examined insects. Details are compiled in the electronic supplement.

Most surprisingly, *Acyrthosiphon pisum* exhibits highly derived structures for all four minor spliceosomal snRNAs, Fig. 6.

The U2 snRNA of *Schmidtea mediterannea* does fit well to the structural alignment of the other U2 snRNAs. In *Schistosoma mansoni* we found a canonical U12 snRNA, while the sequences of the candidates for minor spliceosomal snRNAs do not fit well to the consensus secondary structure models. Details can be found in the Electronic Supplement.

3.6 Syntenic Conservation

In order to assess the conservation of the genomic positions of the snRNAs we retrieved the protein coding genes adjacent to the 31 human snRNAs (8 U1, 3 U2, 2 U4, 5 U5, 7 U6, 1 U11, 1 U12, 3 U4atac and 1 U6atac) and compared the position of their homologs in 14 vertebrate genomes (teleosts, frog, chicken, platypus, opossum, rodents, cow, dog, and chimp) with the 234 snRNA genes that were found in these genomes. We found syntenic conservation of snRNA and flanking genes in only 36 cases, of which 20 belong to the human-chimp comparison and 9 pairs are conserved between

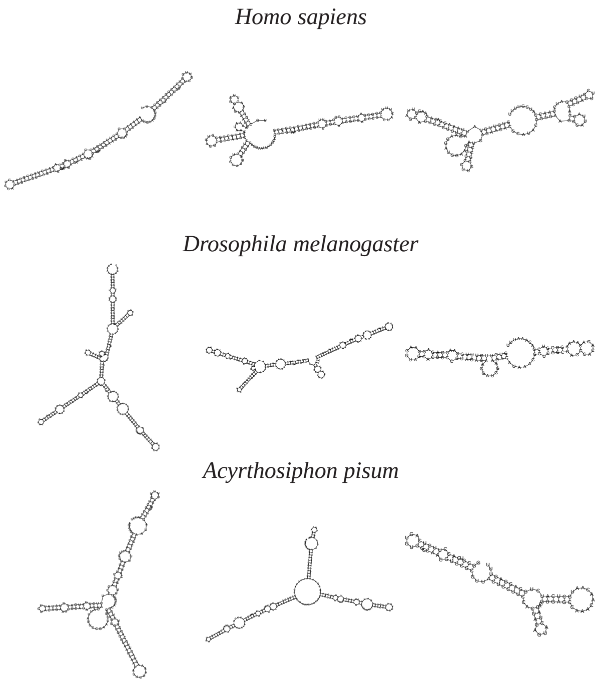

Fig. 6 Secondary structures of U11 (left), U12 (center), U6atac (right) in *Acyrthosiphon pisum*, *Drosophila melanogaster* and *Homo sapiens*. Drosophilids derived far from all other minor spliceosome structures (e.g. human). Moreover, *Acyrthosiphon pisum* built an autonomous structure group for all minor snRNAs.

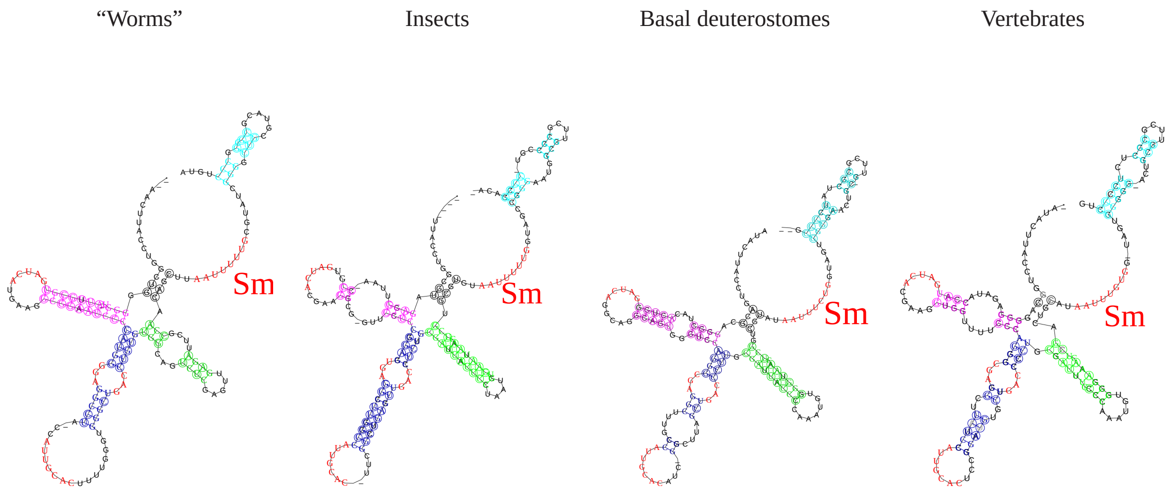

Fig. 5 Secondary structure prediction of U1 snRNA, folded by `RNAalifold`. From left to right: protostomia without insects, insects, deuterostomes without vertebrates, vertebrates. Red: Conserved sequences in all organisms, which possibly bind to proteins. Sm binding site marked separately.

human and mouse. Only a single pair is conserved between human and opossum and no syntenic conservation can be traced back further in evolutionary history. Including the pseudogenes increases the numbers of conserved pairs to 499 of 1609. Again most of these (453) are human/chimp pairs. The data clearly show that snRNA locations are not syntenically conserved, i.e., snRNA behave like mobile elements in their genomic context.

3.7 Pseudogenes

As mentioned above, snRNAs are frequently the founders of families of pseudogenes. This is a property that they share with most other small RNA classes such as 7SL RNA, Y RNA, tRNAs etc. Such families of pseudogenes are easily recognized as a by-product of `blast`-based homology searches as a large set of hits with intermediate E-values. Fig. 7 summarizes such data, more details are provided in the Electronic Supplement.

Spliceosomal snRNA pseudogenes families are very unevenly distributed across distinct phylogenetic groups and have clearly arisen in independent burst multiple times across animal evolution. Within deuterostomes, almost all sequenced genomes, whith the notable exception of teleosts and chicken, contain at least one large family of snRNA-derived pseudogenes.

The genus *Caenorhabditis* shows no pseudogenes, whereas other nematods show nearly such a high number of pseudogenes as primates. Annelids, molluscs and plathelminths behave similarly. The *Trichoplax adhaerens* genome, on the other hand, contains a single copy of each of the nine spliceosomal snRNAs.

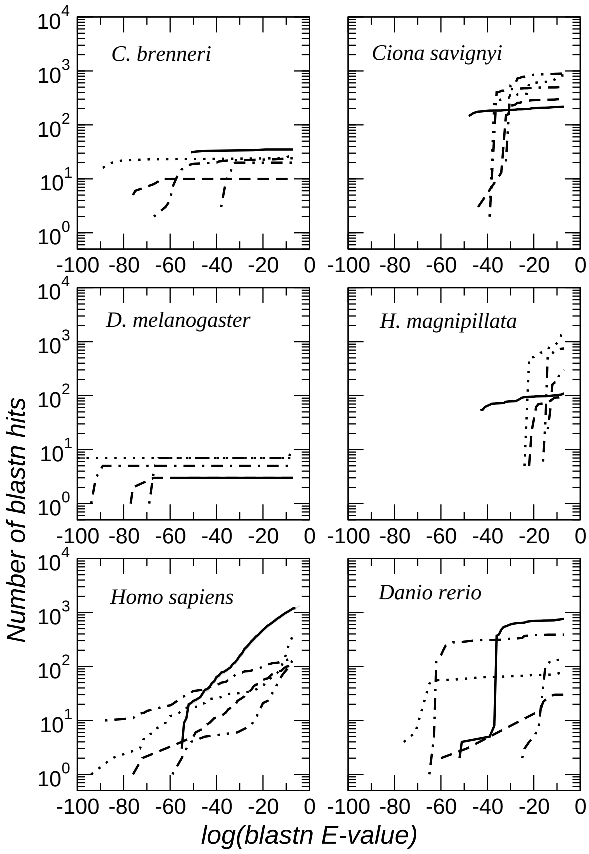

Fig. 7 Double-logarithmic plot of the number of blast hits versus cut-off E-value for 6 different genomes. Pseudogene families appear as a slowly increasing curve, while genes without a “cloud” of pseudogene have a flat distribution for $E < 10^{-5}$. Dashdotted line – U1; dotted line –U2; dashed line – U4; dashdotdotted line – U5; continuous line – U6.

4 Discussion

We have reported here on a comprehensive computational survey of spliceosomal snRNA in all currently available metazoan genomes. We thus provide a comparable and nearly complete collection of animal snRNA sequences. The dense taxon sampling allowed us to verify homology of candidate sequences. Both the major and the minor spliceosome are present in almost all metazoan clades, nematodes (and possibly *Oikopleura*) being the only notable exception. For many of the metazoan families we report here the first evidence on their spliceosomal RNAs.

Using restrictive filtering of the candidates by both secondary structure and canonical promoter structure leaves us with a high-quality data set that was then used to construct secondary structure models. This is useful in particular for the snRNAs of the minor spliceosome for which very few sequences are reported in databases; indeed, the `Rfam 7.0` [23] lists only the U11 and U12 families with a meager set of seed sequences from few model organisms. The sequence and secondary structure data compiled in this study provide a substantially improved databasis and set the stage for systematic searches of even more distant homologs.

The analysis of the genomic distribution of snRNAs reveals that discernible paralogs are not uncommon within genera or families. However, no dramatically different paralogs have been found. Spliceosomal snRNAs are prone to spawning large pseudogene families, which arose independently in many species. They behave like mobile genetic elements in that they barely appear in syntenic positions as measured by their flanking genes. While in some genomes snRNAs appear in tandem and/or associated with with 5S rRNA genes, these clusters are not conserved over longer evolutionary time-scales. Taken together, the data are consistent with a dominating duplication-deletion mechanism of concerted evolution for the genomic evolution and proliferation of snRNA.

Acknowledgments

This work was supported in part by the *Graduierten-Kolleg Wissensrepräsentation* and by the Bioinformatics Initiative of the *Deutsche Forschungs-Gemeinschaft* (DFG). Special thanks to Petra Pregel and Jens Steuck for making work much easier.

References

1. Bandelt, H.J., Dress, A.W.M.: A canonical decomposition theory for metrics on a finite set. Adv. Math. **92**, 47 (1992)
2. Bark, C., Weller, P., Zabielski, J., Pettersson, U.: Genes for human U4 small nuclear RNA. Gene **50**, 333–344 (1986)
3. Barzotti, R., Pelliccia, F., Rocchi, A.: Identification and characterization of U1 small nuclear RNA genes from two crustacean isopod species. Chromosome Res **11**, 365–373 (2003)
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L.: `GenBank`. Nucleic Acids Res. **35**, D21–D25 (2007)
5. Bhathal, H.S., Zamrod, Z., Tobaru, T., Stumph, W.E.: Identification of proximal sequence element nucleotides contributing to the differential expression of variant U4 small nuclear RNA genes. J. Biol. Chem. **270**, 27,629–27,633 (1995)
6. Branlant, C., Krol, A., Lazar, E., Haendler, B., Jacob, M., Galego-Dias, L., Pousada, C.: High evolutionary conservation of the secondary structure and of certain nucleotide sequences of U5 RNA. Nucleic Acids Res **11**, 8359–8367 (1983)
7. Bryant, D., Moulton, V.: Neighbor-net: An agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. **21**, 255–265 (2004)
8. Chen, L., Lullo, D.J., Ma, E., Celniker, S.E., Rio, D.C., Doudna, J.A.: Identification and analysis of U5 snRNA variants in Drosophila. RNA **11**, 1473–1477 (2005)
9. Collins, L., Penny, D.: Complex spliceosomal organization ancestral to extant eukaryotes. Mol. Biol. Evol. **22**, 1053–1066 (2005)
10. Collins, L.J., Macke, T.J., Penny, D.: Searching for ncRNAs in eukaryotic genomes: maximizing biological input with `RNAmotif`. J. Integ. Bioinf. **1**, 2004–08–04 (2004). URL `http://journal.imbio.de/index.php?paper_id56`
11. Cross, I., Rebordinos, L.: 5S rDNA and U2 snRNA are linked in the genome of *Crassostrea angulata* and *Crassostrea gigas* oysters: does the $(ct)_n.(ga)_n$ microsatellite stabilize this novel linkage of large tandem arrays? Genome **48**, 1116–1119 (2005)
12. Dahlberg, J.E., Lund, E.: The genes and transcription of the major small nuclear RNAs. In: M.L. Birnstiel (ed.) Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles, pp. 38–70. Springer-Verlag, Berlin (1988)
13. Denison, R.A., Van Arsdell, S.W., Bernstein, L.B., Weiner, A.M.: Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. Proc. Natl. Acad. Sci. USA **78**, 810–814 (1981)
14. Domitrovich, A.M., Kunkel, G.R.: Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies. Nucleic Acids Res. **31**, 2344–2352 (2003)
15. Drosophila 12 Genomes Consortium: Evolution of genes and genomes on the Drosophila phylogeny. Nature **450**, 203–218 (2007)
16. Ebel, C., Frantz, C., Paulus, F., Imbault, P.: Trans-splicing and cis-splicing in the colourless euglenoid, *Entosiphon sulcatum*. Curr Genet **35**, 542–550 (1999)
17. Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K.L., Howe, K., Johnson, N., Jenkinson, A., Kähäri, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A.J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J.P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., Searle, S.: Ensembl 2008. Nucleic Acids Res. **36**, D707–D714 (2008)
18. Forbes, D.J., Kirschner, M.W., Caput, D., Dahlberg, J.E., Lund, E.: Differential expression of multiple U1 small nuclear RNAs in oocytes and embryos of *Xenopus laevis*. Cell **38**, 681–689 (1984)
19. Gautheret, D., Lambert, A.: Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J. Mol. Biol. **313**, 1003–1011 (2001)
20. Giribet, G., Edgecombe, G.D., Wheeler, W.C.: Arthropod phylogeny based on eight molecular loci and morphology. Nature **413**, 157–161 (2001)
21. Gonzalez, I.L., Sylvester, J.E.: Human rDNA: Evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. Genomics **73**, 255–263 (2001)
22. Griffiths-Jones, S.: `RALEE`—rna alignment editor in `Emacs`. Bioinformatics **21**, 257–259 (2005)
23. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A.: `Rfam`: annotating non-coding RNAs in complete genomes. Nucleic Acids Res **33**, D121–D124 (2005)

24. Hastings, K.E.: SL trans-splicing: easy come or easy go? Trends Genet. **21**, 240–247 (2005)
25. Hausner, T.P., Giglio, L.M., Weiner, A.M.: Evidence for base-pairing between mammalian U2 and U6 small nuclear ribonucleoprotein particles. Genes Dev **4**, 2146–2156 (1990)
26. Hernandez, N.: Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. J. Biol. Chem. **276**, 26,733–26,736 (2001)
27. Hillier, L.W., Miller, W., Birney, E., *** authors: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**, 695–716 (2004)
28. Hillis, D.M., Dixon, M.T.: Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. **66**, 411–453 (1991)
29. Hinas, A., Larsson, P., Avesson, L., Kirsebom, L.A., Virtanen, A., Söderbom, F.: Identification of the major spliceosomal RNAs in *Dictyostelium discoideum* reveals developmentally regulated U2 variants and polyadenylated snRNAs. Eukaryotic Cell **5**, 924–934 (2006)
30. Hofacker, I.L., Fekete, M., Stadler, P.F.: Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. **319**, 1059–1066 (2002)
31. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P.: Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125**, 167–188 (1994)
32. Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X.M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., Birney, E.: `Ensembl` 2005. Nucleic Acids Res. **33**, D447–D453 (2005)
33. Huson, D.H., Bryant, D.: Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23**, 254–267 (2006)
34. Kirsten, T., Rahm, E.: `BioFuice`: Mapping-based data intergation in bioinformatics. In: U. Leser, F. Naumann, B. Eckman (eds.) Proceedings of the 3rd International Workshop on Data Integration in the Life Sciences (DILS), vol. 4075, pp. 124–135. Springer Verlag, Berlin, Heidelberg (2006)
35. König, H., Matter, N., Bader, R., Thiele, W., Müller, F.: Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation. Cell **131**, 718–729 (2007)
36. Korf, G.M., Stumph, W.E.: Chicken U2 and U1 RNA genes are found in very different genomic environments but have similar promoter structures. Biochemistry **25**, 2041–2047 (1986)
37. Krol, A., Branlant, C., Lazar, E., Gallinaro, H., Jacob, M.: Primary and secondary structures of chicken, rat and man nuclear U4 RNAs. Homologies with U1 and U5 RNAs. Nucleic Acids Res **9**, 2699–2716 (1981)
38. Kunkel, G.R., Pederson, T.: Upstream elements required for efficient transcription of a human U6 RNA gene resemble those of U1 and U2 genes even though a different polymerase is used. Genes Dev **2**, 196–204 (1988)
39. Kyriakopoulou, C., Larsson, P., Liu, L., Schuster, J., Söderbom, F., Kirsebom, L.A., Virtanen, A.: U1-like snRNAs lacking complementarity to canonical 5' splice sites. RNA **12**, 1603–1611 (2006)
40. Liao, D.: Concerted evolution: molecular mechanism and biological implications. Am J Hum Genet **64**, 24–30 (1999)
41. Liao, D., Pavelitz, T., Kidd, J.R., Kidd, K.K., Weiner, A.M.: Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. EMBO J. **16**, 588–598 (1997)
42. Liao, D., Weiner, A.M.: Concerted evolution of the tandemly repeated genes encoding primate U2 small nuclear RNA (the RNU2 locus) does not prevent rapid diversification of the $(CT)_n.(GA)_n$ microsatellite embedded within the U2 repeat unit. Genomics **30**, 583–593 (1995)
43. Lo, P.C., Mount, S.M.: *Drosophila melanogaster* genes for U1 snRNA variants and their expression during development. Nucleic Acids Res **18**, 6971–6979 (1990)
44. Lorković, Z.J., Lehner, R., Forstner, C., Barta, A.: Evolutionary conservation of minor U12-type spliceosome between plants and humans. RNA **11**, 1095–1107 (2005)
45. Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A., Sampath, R.: RNAMotif, an RNA secondary structure definition and search algorithm. Nucl. Acids Res. **29**(22), 4724–4735 (2001)
46. Manchado, M., Zuasti, E., Cross, I., Merlo, A., Infante, C., Rebordinos, L.: Molecular characterization and chromosomal mapping of the 5S rRNA gene in *Solea senegalensis*: a new linkage to the U1, U2, and U5 small nuclear RNA genes. Genome **49**, 79–86 (2006)
47. Mattaj, I.W., Zeller, R.: *Xenopus laevis* U2 snRNA genes: tandemly repeated transcription units sharing 5' and 3' flanking homology with other RNA polymerase II transcribed genes. EMBO J **2**, 1883–1891 (1983)
48. Missal, K., Rose, D., Stadler, P.F.: Non-coding RNAs in *Ciona intestinalis*. Bioinformatics **21 S2**, i77–i78 (2005)
49. Missal, K., Zhu, X., Rose, D., Deng, W., Skogerbø, G., Chen, R., Stadler, P.F.: Prediction of structured non-coding RNAs in the genome of the nematode *Caenorhabitis elegans*. J. Exp. Zool.: Mol. Dev. Evol. **306B**, 379–392 (2006)
50. Mitrovich, Q.M., Guthrie, C.: Evolution of small nuclear RNAs in *S. cerevisiae*, *C. albicans*, and other hemiascomycetous yeasts. RNA **13**, 2066–2080 (2007)
51. Montzka, K.A., Steitz, J.A.: Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. Proc Natl Acad Sci U S A **85**, 8885–8889 (1988)
52. Morales, J., Borrero, M., Sumerel, J., C., S.: Identification of developmentally regulated sea urchin U5 snRNA genes. DNA Seq. **7**, 243–259 (1997)
53. Mount, S.M., Gotea, V., Lin, C.F., Hernandez, K., Makałowski, W.: Spliceosomal small nuclear RNA genes in 11 insect genomes. RNA **13**, 5–14 (2007)
54. Mount, S.M., Steitz, J.A.: Sequence of U1 RNA from *Drosophila melanogaster*: implications for U1 secondary structure and possible involvement in splicing. Nucleic Acids Res **9**, 6351–6368 (1981)
55. Myslinksi, E., Krol, A., Carbon, P.: Characterization of snRNA and snRNA-type genes in the pufferfish *Fugu rubripes*. Gene **330**, 149–158 (2004)
56. Myslinski, E., Branlant, C., Wieben, E.D., Pederson, T.: The small nuclear RNAs of *Drosophila*. J. Mol. Biol. **180**, 927–945 (1984)
57. Nei, M., Rooney, A.P.: Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. **39**, 121–152 (2005)
58. Nilsen, T.W.: The spliceosome: the most complex macromolecular machine in the cell? Bioessays **25**, 1147–1149 (2003)
59. Otake, L.R., Scamborova, P., Hashimoto, C., Steitz, J.A.: The divergent U12-type spliceosome is required for pre-mRNA splicing and is essential for development in *Drosophila*. Mol Cell **9**, 439–446 (2002)
60. Papillon, D., Perez, Y., Caubit, X., Le Parco, Y.: Systematics of chaetognatha under the light of molecular data, using duplicated ribosomal 18S DNA sequences. Mol Phylogenet Evol. **38**, 621–634 (2006)
61. Patel, A.A., Steitz, J.A.: Splicing double: insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. **4**, 960–970 (2003)
62. Pavelitz, T., Liao, D., Weiner, A.M.: Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. EMBO J **18**, 3783–3792 (1999)
63. Pelliccia, F., Barzotti, R., Bucciarelli, E., Rocchi, A.: 5S ribosomal and U1 small nuclear RNA genes: a new linkage type in the genome of a crustacean that has three different tandemly repeated units containing 5S ribosomal DNA sequences. Genome **44**, 331–335 (2001)
64. Pereira-Simon, S., Sierra-Montes, J.M., Ayesh, K., Martinez, L., Socorro, A., Herrera, R.J.: Variants of U1 small nuclear RNA assemble into spliceosomal complexes. Insect Molecular Biology **13**, 189–194 (2004)

65. Russell, A.G., Charette, J.M., Spencer, D.F., Gray, M.W.: An early evolutionary origin for the minor spliceosome. Nature **443**, 863–866 (2006)
66. Russell, A.G., Charette, J.M., Spencer, D.F., Gray, M.W.: An early evolutionary origin for the minor spliceosome. Nature **443**, 863–866 (2006)
67. S., L.E., H., W.R., S., C.F., *** authors: Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**, 69–87 (2005)
68. Schlötterer, C., Tautz, D.: Chromosomal homogeneity of Drosophila ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. Curr. Biol. **4**, 777–783 (1994)
69. Schneider, C., Will, C.L., Brosius, J., Frilander, M., Lührmann, R.: Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein paricle in *Drosophila*. Proc. Natl. Acad. Sci. USA **101**(26), 9584–9589 (2004)
70. Shambaugh, J.D., Hannon, G.E., Nilsen, T.W.: The spliceosomal U small nuclear RNAs of *Ascaris lumbricoides*. Mol Biochem Parasitol **64**, 349–352 (1994)
71. Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., Sachidanandam, R.: Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res **34**, 3955–3967 (2006)
72. Shukla, G.C., Cole, A.J., Dietrich, R.C., Padgett, R.A.: Domains of human U4atac snRNA required for U12-dependent splicing in vivo. Nucleic Acids Res **30**, 4650–4657 (2002)
73. Shukla, G.C., Padgett, R.A.: Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. RNA **5**, 525–538 (1999)
74. Shukla, G.C., Padgett, R.A.: U4 small nuclear RNA can function in both the major and minor spliceosomes. Proc. Natl. Acad. Sci. USA **101**, 93–98 (2004)
75. Sierra-Montes, J.M., Freund, A.V., Ruiz, L.M., Szmulewicz, M.N., Rowold, D.J., Herrera, R.J.: Multiple forms of U2 snRNA coexist in the silk moth *Bombyx mori*. Insect Mol Biol **11**, 105–114 (2002)
76. Sierra-Montes, J.M., Pereira-Simon, S., Freund, A.V., Ruiz, L.M., Szmulewicz, M.N., Herrera, R.J.: A diversity of U1 small nuclear RNAs in the silk moth *Bombyx mori*. Insect Biochem Mol Biol **33**, 29–39 (2003)
77. Sierra-Montes, J.M., Pereira-Simon, S., Smail, S.S., Herrera, R.J.: The silk moth *Bombyx mori* U1 and U2 snRNA variants are differentially expressed. Gene **352**, 127–136 (2005)
78. Smail, S.S., Ayesh, K., Sierra-Montes, J.M., Herrera, R.J.: U6 snRNA variants isolated from the posterior silk gland of the silk moth *Bombyx mori*. Insect Biochem Mol Biol. **36**, 454–465 (2006)
79. Sontheimer, E.J., Steitz, J.A.: Three novel functional variants of human U5 small nuclear RNA. Mol. Cell. Biol. **12**, 734–746 (1992)
80. Stefanovic, B., Li, J.M., Sakallah, S., Marzluff, W.F.: Isolation and characterization of developmentally regulated sea urchin U2 snRNA genes. Dev Biol. **148**, 284–294 (1991)
81. Stefanovic, B., Marzluff, W.F.: Characterization of two developmentally regulated sea urchin U2 small nuclear RNA promoters: a common required TATA sequence and independent proximal and distal elements. Mol Cell Biol **12**, 650–660 (1992)
82. Tarn, W.Y., Yario, T.A., Steitz, J.A.: U12 snRNAs in vertebrates: Evolutionary conservation of 5' sequences implicated in splicing of pre-mRNAs containing a minor class of introns. RNA **1**, 644–656 (1995)
83. Telford, M.J., Holland, P.W.H.: Evolution of 28S ribosomal DNA in chaetognaths: Duplicate genes and molecular phylogeny. J. Mol. Evol. **44**, 135–144 (1997)
84. Thomas, J., Lea, K., Zucker-Aprison, E., Blumenthal, T.: The spliceosomal snRNAs of *Caenorhabditis elegans*. Nucleic Acids Res **18**, 2633–2642 (1990)
85. Tichelaar, J.W., Wieben, E.D., Reddy, R., Vrabel, A., Camacho, P.: *In vivo* expression of a variant human U6 RNA from a unique, internal promoter. Biochemistry **37**, 12,943–12,951 (1998)
86. Valadkhan, S.: snRNAs as the catalysts of pre-mRNA splicing. Curr. Op. Chem. Biol. **9**, 603–608 (2005)
87. Valadkhan, S.: The spliceosome: caught in a web of shifting interactions. Curr. Op. Struct. Biol. **17**, 310–315 (2007)
88. Valadkhan, S., Mohammadi, A., Wachtel, C., Manley, J.L.: Protein-free spliceosomal snRNAs catalyze a reaction that resembles the first step of splicing. RNA **13**, 2300–2311 (2007)
89. Will, C.L., Lührmann, R.: Splicing of a rare class of introns by the U12-dependent spliceosome. Biol. Chem. **386**, 713–724 (2005)