U7 snRNAs: A Computational Survey

Manja Marz[1], Axel Mosig[2,3], Bärbel M.R. Stadler[3], and Peter F. Stadler[1,4,5,6,7]*

[1] *Bioinformatics Group, Department of Computer Science, University of Leipzig, Leipzig D-04107, Germany;* [2] *Department of Combinatorics and Geometry, CAS/MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;* [3] *Max Planck Institute for Mathematics in the Sciences, Leipzig D-04103, Germany;* [4] *Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig D-04107, Germany;* [5] *Fraunhofer Institute for Cell Therapy and Immunology, Leipzig D-04103, Germany;* [6] *Department of Theoretical Chemistry, University of Vienna, Vienna A-1090, Austria;* [7] *Santa Fe Institute, Santa Fe, NM 87501, USA.*

U7 small nuclear RNA (snRNA) sequences have been described only for a handful of animal species in the past. Here we describe a computational search for functional U7 snRNA genes throughout vertebrates including the upstream sequence elements characteristic for snRNAs transcribed by polymerase II. Based on the results of this search, we discuss the high variability of U7 snRNAs in both sequence and structure, and report on an attempt to find U7 snRNA sequences in basal deuterostomes and non-drosophilids insect genomes based on a combination of sequence, structure, and promoter features. Due to the extremely short sequence and the high variability in both sequence and structure, no unambiguous candidates were found. These results cast doubt on putative U7 homologs in even more distant organisms that are reported in the most recent release of the Rfam database.

Key words: U7 snRNA, non-coding RNA, RNA secondary structure, evolution

Introduction

The U7 small nuclear RNA (snRNA) is the smallest polymerase II transcript known to date, with a length ranging from only 57 nt (sea urchin) to 70 nt (fruit fly). Its expression level of only a few hundred copies per cell in mammals is at least three orders of magnitude smaller than the abundance of other snRNAs. It is part of the U7 small nuclear ribonucleoprotein (snRNP), which plays a crucial role in the 3′ end processing of histone mRNAs (*1*). Replication-dependent histone mRNAs in metazoa are the only known eukaryotic protein-coding mRNAs that are not polyadenylated ending but contain a conserved stem-loop sequence instead (*2*). Beyond metazoan animals, non-polyadenylated histone genes have been described in the algae *Chlamydomonas reinhardtii* and *Volvox carteri* (*3*), and *Dictyostelium discoideum* has a homolog of the histone RNA hairpin-binding protein/stem-loop-binding protein (HBP/SLBP) (DictyBase: DDB0169192). It appears that replication-dependent histone mRNAs are the only mRNAs that are processed in this way (*4*).

The 5′ region of the U7 snRNA is complementary to the "histone downstream element" (HDE), located just downstream of the conserved hairpin. The interaction of the U7 snRNP with the HDE is crucial for the correct processing of the histone 3′ elements (*1*). The 3′ part of the U7 snRNA is occupied by a modified binding domain for Sm proteins consisting of a characteristic sequence motif followed by a conserved stem-loop secondary structure motif (*5*). U7 snRNA binds five of the seven Sm proteins that are present in spliceosomal snRNAs, while the D1 and D2 subunits are replaced by the Sm-like proteins Lsm10 and Lsm11 (*6–8*). This difference is likely to be associated with the differences in the Sm-binding sequence. Recently, the U7 snRNP has not only received considerable attention from a structural biology point of view (*9*, *10*), but also has been investigated as a means of modifying splicing dys-regulation. In particular, U7 snRNA-derived constructs that target a mutant dystrophin gene were explored as a gene-therapy approach to Duchenne muscular dystrophy (*11*, *12*).

***Corresponding author.**
E-mail: studla@bioinf.uni-leipzig.de

Given the attention received by histone RNA 3′ end processing and the protein components of the U7 snRNP, it may come as a surprise that the U7 snRNA itself has received little attention in the last decades. In fact, the only two experimentally characterized mammalian U7 snRNAs are those of mouse (*13–16*) and human (*1*, *17*), while most of the earliest work on U7 snRNPs concentrated on the sea urchin *Psammechinus miliaris* (*18–21*) and two *Xenopus* species (*22–24*). More recently, the U7 snRNA sequences have been reported for *Drosophila melanogaster* (*25*) and *Takifugu rubripes* (*26*).

We are aware of only two studies that considered U7 snRNA from a bioinformatics point of view. In Lück *et al* (*27*), the U7 snRNA was used as an example for the application of the Construct tool to compute consensus secondary structures, and Bompfünewerer *et al* (*28*) briefly reported on a BLAST-based homology search that uncovered candidate sequences for chicken and two teleost fish.

The U7 snRNP-dependent mode of histone end processing is a metazoan innovation (*2*, *6*). Nevertheless, the most recent release of the Rfam database (*29*) (Version 8.0; February 2007) lists sequences from eukaryotic protozoa, plants, and even bacteria. This discrepancy prompted us to critically assess the available information on U7 snRNAs.

Results

Bona fide U7 snRNA sequences

The results of the BLAST-based searches are summarized in Table 1. In most species, only a single gene with clear snRNA-like upstream elements was found. In addition, BLAST identified several pseudogenes. Clusters of U7 snRNAs as previously described for sea urchins and frogs were otherwise only found in zebrafish (Figure 1).

The short length and the substantial divergence of the U7 snRNA sequences make it impossible to distinguish functional U7 snRNAs from pseudogenes based on the U7 sequence alone. To make this distinction, it is necessary to analyze the flanking sequences as well. *Bona fide* snRNA genes are accompanied by characteristic promoter elements (*30*, *31*). Figure 2 displays the consensus sequence motifs of the presumably functional amniote U7 snRNAs.

In human and mouse, several pseudogenes have been described in detail in addition to the functional genes (*16*, *32*). Notably, several variant U7 snRNA sequences from human HeLa cells were reported in Yu *et al* (*17*). This might indicate that the human genome, in apparent contrast to mouse, also contains more than one functional U7 snRNA gene, or that some of the pseudogenes are transcribed at low levels. Table 1 therefore lists the number of U7-associated loci obtained by BLAST searches that use the presumably functional gene from the same species as query. This number can be fairly large in some mammalian lineages, reaching almost 100 loci in primates. In contrast, in most species there are only a few U7-associated sequences, most of which are readily recognizable as retrogenes by virtue of poly-A tails.

In several genomes, we were not able to find an unambiguous candidate for a functional U7 snRNA, although we found sequences that are clearly derived from U7 but are not accompanied by a recognizable proximal sequence element (PSE). Examples include *Sorex araneus* and platypus. Most likely, these BLAST hits are pseudogenes, although many of them are annotated with Ensembl gene IDs. This annotation derives from sequence homology with the examples stored in the Rfam database. In Figure 3 and Table 1, we compile the results of our BLAST-based homology search, which contain only sequences that are either experimentally known to be expressed or are predicted to be functional genes based on the presence of conserved upstream elements.

Separate multiple sequence alignments of amniotes, teleosts, frogs, sea urchins, and flies reveal strong conservation of the Sm-binding motif, consisting of the deviant Sm-binding site RUUUNUCYNG and the 3′ hairpin structure. Furthermore, the histone-binding region contains a universally conserved box UCUUU (*33*). Using these features as anchors, we obtained the alignment in Figure 3, which highlights the differences between major clades. Notable variations within the vertebrates are in particular the A-rich 5′ and the reduced stem in teleosts, and their A-rich sequences in the hairpin loop. The hairpin region is very poorly conserved at the sequence level between vertebrates, sea urchins, and flies, although its structural variation is limited in essence to the length of the stem and a few short interior loops or single-nucleotide bulges.

More distant homologs?

The U7 snRNA sequences evolve rather fast. Together with the short sequence length, this limits the power

Table 1 Trusted U7 snRNA sequences*

Species	Assembly	Sequence	Start	Stop	Ori.	Database ID	ψ
Mus musculus	Ensembl 43	Chr.6	124,706,844	124,706,905	−	ENSMUSG00000065217	27
Rattus norvegicus	Ensembl 43	Chr.X	118,163,804	118,163,865	−	ENSRNOG00000034996	31
Rattus norvegicus	Ensembl 43	Chr.4	160,870,934	160,870,995	−	ENSRNOG00000035016	31
Homo sapiens	Ensembl 43	Chr.12	6,923,240	6,923,302	+	ENSG00000200368	91
Macaca mulatta	Ensembl 43	Chr.11	7,125,496	7,125,557	+	ENSMMUG00000027525	95
Otolemur garnettii	PreEnsembl 43	Scaffold_102959	117,572	117,633	−		0
Oryctolagus cuniculus	Ensembl 43	GeneScaffold_1693	111,485	111,546	+		3
Procavia capensis	NCBI TRACE	175719230	275	336	+		–
Loxodonta africana	Ensembl 43	Scaffold_60301	4,254	4,314	−		2
Echinops telfairi	Ensembl 43	GeneScaffold_2204	10,742	10,803	+	ENSETEG00000020899	57
Felis catus	Ensembl 43	GeneScaffold_69	192,907	192,968	+		7
Canis familiaris	Ensembl 43	Chr.27	41,131,749	41,131,810	−	ENSCAFG00000021852	2
Myotis lucifugus	PreEnsembl 43	Scaffold_168837	32,294	32,356	−		0
Equus caballus	PreEnsembl 43	Scaffold_58	7,463,562	7,463,623	+		0
Bos taurus	Ensembl 43	Chr.5	10,349,126	10,349,187	−	AAFC03061782	8
Tursiops truncatus	NCBI TRACE	194072802	598	659	+		–
Dasypus novemcinctus	Ensembl 43	GeneScaffold_1944	24,469	24,530	+		16
Spermophilus tridec.	PreEnsembl 43	Scaffold_139061	45,428	45,489	−		0
Erinaceus europaeus	Ensembl 43	GeneScaffold_2232	5,133	5,194	+		30
Monodelphis domestica	Ensembl 43	Un	131,411,333	131,411,393	+	ENSMODG00000022029	1
Gallus gallus	Ensembl 43	Chr.1	80,484,148	80,484,212	+	ENSGALG00000017891	1
Taeniopygia guttata	NCBI TRACE	TGAB-afg09c06.b1	683	748	−		–
Anolis carolinensis	NCBI TRACE	G889P8207RM16.T0	106	171	−		–
Xenopus tropicalis	Ensembl 43	Scaffold_883	Cluster: ∼20 copies from 272,500 to end				
Xenopus laevis	GenBank	X64404	Cluster (partial)				
Xenopus borealis	GenBank	Z54313	Cluster (partial)				
Danio rerio	Ensembl 43	Chr.16	Cluster: 4 copies from 13,708,000 to 13,723,000				
Takifugu rubripes	Ensembl 43	Scaffold_205	229,679	229,736	+		0
Tetraodon nigroviridis	Ensembl 43	Chr.8	9,059,483	9,059,541	+		(1)
Gasterosteus aculeatus	Ensembl 43	GroupXX	11,616,333	11,616,392	−		0
Oryzias latipes	Ensembl 43	Chr.16	17,393,002	17,393,059	+		0
Strongylocentrotus p.	BCM_Spur_v2.1	Cluster: 2 sequences each on scaffolds 83935 and 88560					
Psammechinus miliaris	GenBank	Cluster: 5 genes, 1 sequence=M13311.1					
Drosophila melanogaster	UCSC	3L	3,577,355	3,577,425	+	CR33504	0
Drosophila ananassae	CAF-1	CH902618.1	9,849,345	9,849,414	−		0
Drosophila erecta	CAF-1	CH954178.1	6,292,889	6,292,959	+		1
Drosophila grimshawi	CAF-1	CH916366.1	10,347,991	10,348,062	+		1
Drosophila mojavensis	CAF-1	CH933809.1	2,924,982	2,925,053	−		1
Drosophila persimilis	CAF-1	CH479328.1	89,311	89,383	−		0
Drosophila pseudoobscura	CAF-1	CH379070.2	5,738,714	5,738,786	+		1
Drosophila simulans	CAF-1	CM000363.1	3,136,652	3,136,582	−		1
Drosophila virilis	CAF-1	CH940647.1	4,512,836	4,512,907	−		1
Drosophila willistoni	CAF-1	CH964101.1	1,418,210	1,418,280	+		0
Drosophila yakuba	CAF-1	CM000159.2	4,146,836	4,146,905	+		0

*ψ gives the number of paralog loci, most likely U7 pseudogenes, defined by a BLAST E-value less than 0.001 compared with the functional copy. CAF-1 refers to the genome freezes provided by the *Drosophila* Comparative Genomics Consortium. These sequences were retrieved from http://rana.lbl.gov/drosophila/caf1.html in December 2006. The *D. melanogaster* sequence is the one used by the UCSC Genome Browser (http://genome.ucsc.edu/) (Release 4; Apr. 2004, UCSC version dm2). The sea urchin genome BCM_Spur_v2.1 was obtained from ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Spurpuratus/fasta/Spur_v2.1/linearScaff.

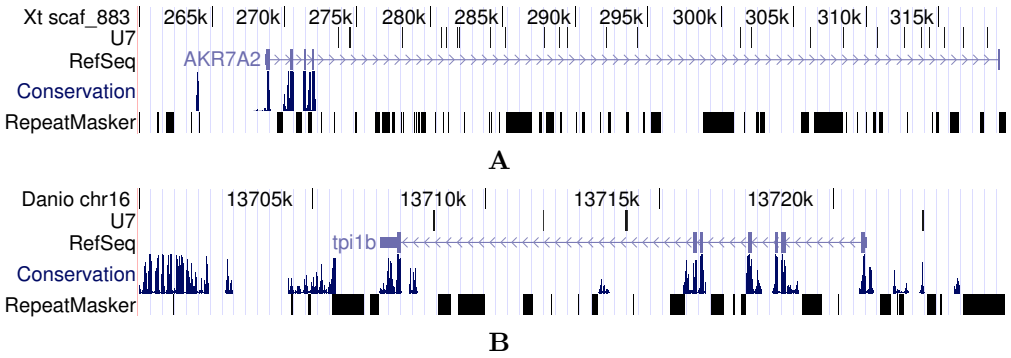


Fig. 1 Clusters of U7 snRNA genes in *Xenopus tropicalis* (**A**) and *Danio rerio* (**B**) taken from the USCS Genome Browser (http://genome.ucsc.edu/). The "U7" track shows BLAST matches of the U7 snRNA sequences; "RepeatMasker" refers to annotated repetitive sequence elements; the "RefSeq" track shows the intron/exon structure of protein-coding genes; the "Conservation" panel displays PhastCons score measuring sequence conservation across vertebrates. We refer to the data track description at the USCS Genome Browser for technical details.

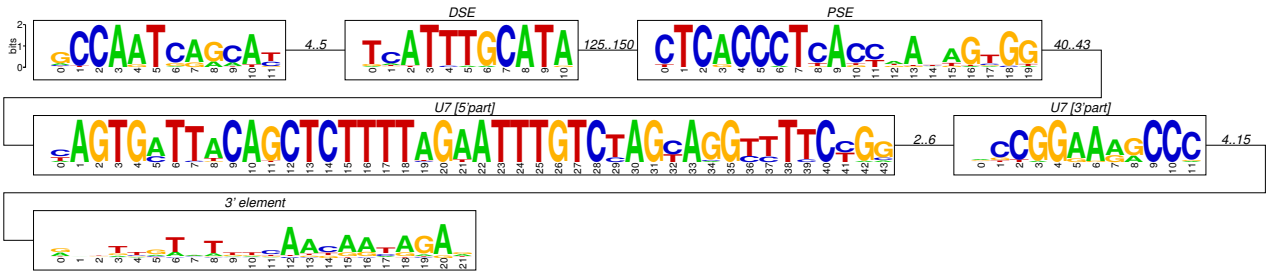
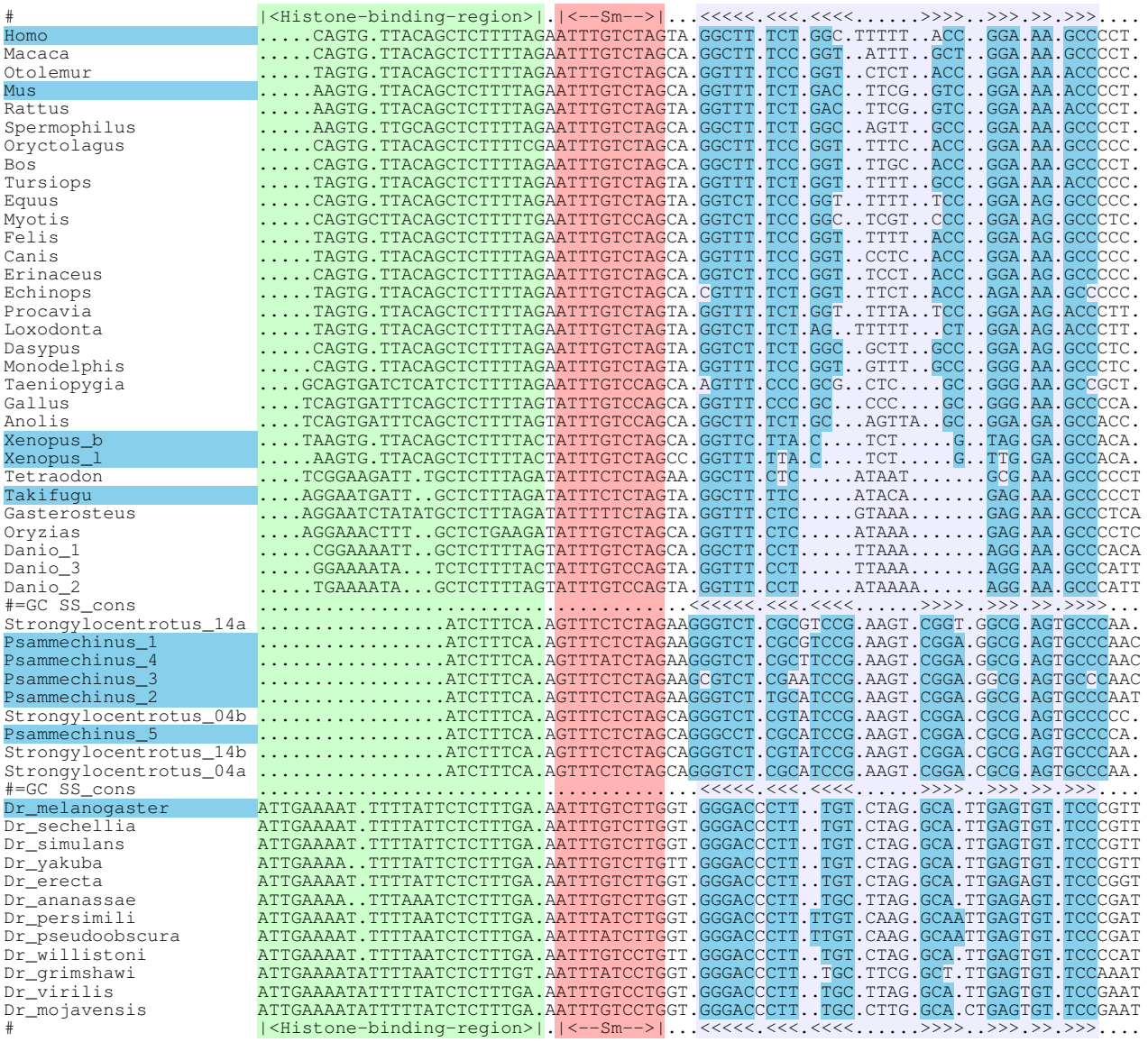


Fig. 2 Conserved elements in functional U7 snRNA genes. The consensus pattern is the amniote sequences from Table 1. The classical distal sequence element (DSE), proximal sequence element (PSE), and 3′ element of pol-II spliceosomal RNA genes are clearly discernible. The U7 sequence itself is interrupted by a short variable region with substantial length variation.

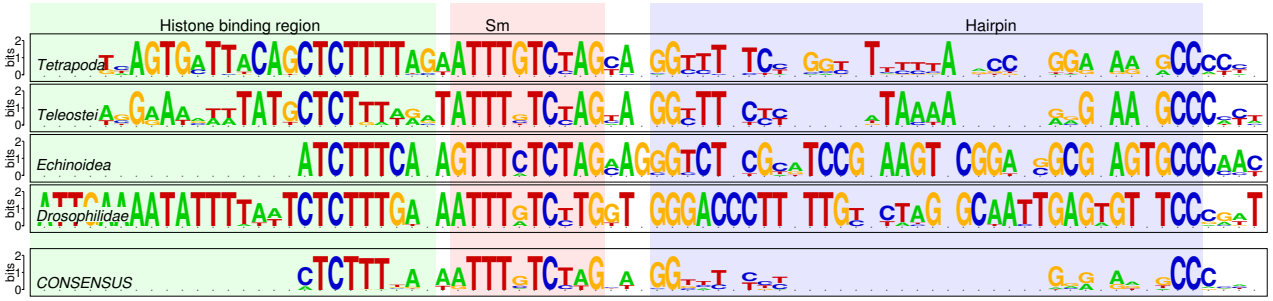
of sequence-based approaches to distant homology search. The consensus pattern in Figure 3 indicates quite clearly that such methods are bound to fail outside the four groups with experimentally known sequences (tetrapods, teleosts, sea urchins, and flies). Indeed, both BLAST and Fragrep (*34*) did not provide additional candidates that could be unambiguously classified as U7 snRNAs based on sequence information alone.

The comparison of the U7 hairpins in the different clades (Figure 4) reveals significant differences in the secondary structures between invertebrates and vertebrates: vertebrates have smaller stem-loop structures with smaller or no interior loops or bulges. The stem in teleosts, furthermore, is systematically shorter than that in tetrapods. These structural differences between clades have to be taken into account for homology search. In fact, as a consensus rule, we can only deduce that the stem-loop structure has a total of 8–15 bp, which is nearly symmetric, and it is enclosed by an uninterrupted stem at least 5 bp in length with 2 GC pairs at its base.

Even combined with the conserved sequence motifs in the 5′ part of the molecule, it yields only a rather loose definition of the U7 snRNA. Release 8.0 of the Rfam database (*29*) lists several sequences in its U7 RNA section that are surprising. Neither contained in the literature nor contained in the manually curated U7 "seed-set", these candidate sequences were found using a homology search based on Infernal (*35*) and the seed alignment. While the *Danio rerio* sequences are identical with the sequences we identified in work starting from the much closer homolog in *T. rubripes*, the candidates reported for *Caenorhabditis elegans* and *Girardia tigrina* raise serious doubts. The *C. elegans* sequence, although ostensibly well conserved in comparison with the deuterostome sequences, has no recognizable homologs in any one of the other three sequenced *Caenorhabditis* species, *C. briggsae*, *C. remanei*, and *C. reinhardii*. The *G. tigrina* sequence is located in the 3′ UTR of the *DthoxE-Hox* gene (X95413). Both sequences furthermore do not share even the core UUUNUC of the consensus Sm-binding motif.



A



B

Fig. 3 Manually curated alignment of functional U7 snRNA sequences. **A**. The 3′ stem-loop, the Sm-binding site, and the histone-binding regions are highlighted. The 5′ most part of the histone-binding region is not aligned between vertebrate and drosophilid sequences. **B**. Sequence logos for the partial alignment comprising only tetrapods, teleosts, sea urchins, and flies, respectively, as well as the consensus pattern arising from combining all the data.

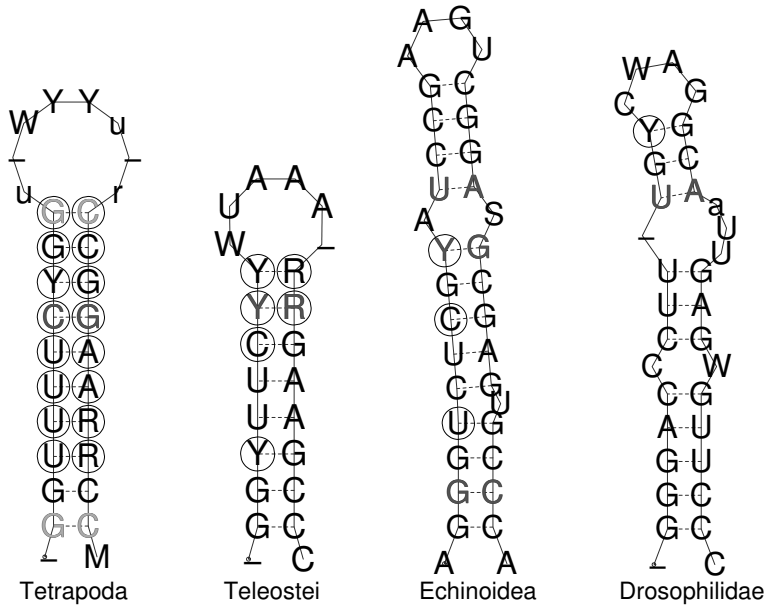


Fig. 4 Comparison of U7 hairpin structures. Consensus secondary structures are computed using RNAalifold program (*39*) on the manually improved alignments of tetrapods, teleosts, sea urchins, and flies, respectively. Circles indicate consistent and compensatory mutations that leave the structure intact. Gray letters indicate that one or two of the aligned sequences cannot form the base pair.

```
#                 |<Histone-binding-region>|.|<--Sm-->|....<<<<<.<<<.<<<<.....>>>>..>>>.>>.>>>....
Homo              .....CAGTG.TTACAGCTCTTTTAGAATTTGTCTAGTA..GGCTT.TCT.GGC.TTTTT..ACC..GGA.AA.GCCCCT.
Mus               .....AAGTG.TTACAGCTCTTTTAGAATTTGTCTAGCA..GGTTT.TCT.GAC..TTCG..GTC..GGA.AA.ACCCCT.
Xenopus_l         .....AAGTG.TTACAGCTCTTTTACTATTTGTCTAGCC..GGTTT.TTA.C....TCT.....G..TTG.GA.GCCACA.
Takifugu          ....AGGAATGATT..GCTCTTTAGATATTTCTCTAGTA..GGCTT.TTC.....ATACA.......GAG.AA.GCCCCCT
Petromyzon-c1     .........ATTGAGGATCTTTGAC.TTTTGTCTTTGTGTGGTGCACC.......GAAA........GGAGC.ACC....
Branchiostoma-c1  .....ACTGG.TAAC.GCTCTTTCAC.CTTTATCCGCG...GGGTA.A........CCT..........T.TA.TCCGTA.
Branchiostoma-c2  .....GAGTG.TAAC.GTTCTTTCAC.CTTTATCCGCG...GGGTA.........ACCTA..........TA.TCCGTT.
Psammechinus_1    ................ATCTTTCA.AGTTTCTCTAGAA.GGGTCT.CGCGTCCG.AAGT.CGGA.GGCG.AGTGCCCAAC
Bombyx_mori-c1    TCCATCAAT.ATGTTCTATCTTTTA..ATTTATCGAAAA.CGGTCA.AG.A....ACTAGTC....G.CT.TG.GCC....
Bombyx_mori-c2    AAGATTTTG.GTGTGTAATCTTTAACTGTTTATCTTTTG.CGGTAGG...T.AGCGGCTTGGCT.......CT.GCC....
Dr_melanogaster   ATTGAAAAT.TTTTATTCTCTTTGA.AATTTGTCTTGGT..GGGACCCTT..TGT.CTAG.GCA.TTGAGTGT.TCCCGTT
#                 |<Histone-binding-region>|.|<--Sm-->|....<<<<<.<<<.<<<<.....>>>>..>>>.>>.>>>....
```

Fig. 5 Best candidates from searches using RNABOB in *Petromyzon marinus*, *Branchiostoma floridae*, and *Bombyx mori*. In addition to the putative U7 snRNA sequences shown here, these candidate sequences also have a putative PSE associated with them.

Several additional candidates were reported in the Rfam database for higher plants and even bacteria. Higher plants apparently do not have the replication-dependent metazoan-style histone 3′ end processing machinery (*2*, *6*), and bacteria do not even have proper histones. It is very unlikely that these sequences are real U7 snRNAs. No conclusive argument can be given at this point for the few isolated U7 snRNA candidates listed in the Rfam database. These examples show once again that at least for very short non-coding RNAs, the results from homology searches have to be taken with caution, in particular when they are not corroborated by additional supporting evidence.

The poor sequence conservation between major groups highlighted in Figure 3 suggests that purely sequence-based homology searches have little chance of success in insect or basal deuterostome genomes. Indeed, neither BLAST nor Fragrep found convincing candidates. We therefore resorted to structure-based approaches and explicitly included the PSE in the search procedure (see Materials and Methods for details). We used RNABOB software with a non-restrictive pattern to find plausible initial candiates, which were then manually compared with the alignment in Figure 3. The most plausible candidates are shown in Figure 5, albeit none of them is unambiguous. No convincing candidates were found in the mosquito *Anopheles gambiae* and in the honeybee *Apis mellifera*.

Discussion

Since U7 snRNA has its primary function in histone 3′ maturation, it is virtually certain that this class of non-coding RNAs is restricted to metazoan animals—after all, the process in which they play a crucial role is unknown outside multicellular animals. With its length of 70 nt or less, U7 snRNA is the smallest known polymerase II transcript. Each of its three major domains, the histone-binding region, the Sm-binding sequence, and the 3′ stem-loop structure exhibits substantial variation in both sequence and structural details, as can be seen from the detailed sequence alignments (Figure 3) and the structural models of the terminal stem-loop structure (Figure 4). As a consequence, our computational survey not only compiles a large number of previously undescribed U7 homologs from vertebrates and drosophilids, but also stresses the limits of current approaches to RNA homology search.

While BLAST already fails to unambiguously recognize teleost fish homology from mammalian queries and *vice versa*, even more sophisticated (and computationally expensive) methods have limited success when applied to basal deuterostome or insect genomes. On the other hand, not only the limited sensitivity of current approaches poses a problem; conversely, the most sensitive methods are fooled by false positives, as exemplified by the plant and bacterial sequences in Rfam.

In summary, thus, this study calls both for more experimental data on U7 snRNAs—Which, if any, of our U7 candidate sequences in lamprey or silkworm are really U7 snRNAs in these species?—and for improved bioinformatics approaches for homology search that can deal with such small and rapidly evolving genes.

Materials and Methods

The experimentally known U7 snRNA sequences were retrieved from GenBank database (http://www.ncbi.nlm.nih.gov/entrez). Starting from the known functional mouse gene (GenBank X54748.4), we used the built-in BLAST search function of Ensembl (http://www.ensembl.org; Release 43) to retrieve homologous regions in other mammalian genomes and the chicken genome. Parameters were set to "*distance homologies*" and repeat-masking was disabled. The resulting sequences were downloaded and aligned using both DIALIGN2 (*36*) and ClustalW (*37*) to determine whether the characteristic up- and downstream elements were present. In order to check for consistency, we compared these alignments with the Ensembl genomic alignments of the homologous human locus. In all cases, Ensembl data and our own search gave consistent results. The *T. rubripes* U7 snRNA sequence described in Myslinksi *et al* (*26*) was used as starting point for searching the teleost fish genomes.

Drosophilid sequences, with the exception of *D. melanogaster* (which was retrieved from Ensembl), were obtained from the website of the *Drosophila* Comparative Genomics Consortium (http://rana.lbl.gov/drosophila/caf1.html). The *D. melanogaster* U7 snRNA region (*25*) was used as BLAST query, resulting in a unique hit in each of the other drosophilid genomes that exhibits the characteristic upstream elements. In addition, at most one putative pseudogene was found in some species.

Sequence alignments of U7 sequences were generated separately for mammals, sauropsids, teleosts, frogs, sea urchins, and flies using ClustalW. These alignments were combined manually using the RALEE mode for Emacs (*38*). Consensus secondary structure for a given sequence alignment was computed using RNAalifold (*39*).

We expanded the tool aln2pattern, the component of the Fragrep distribution (*34*) that generates a collection of position weight matrixs as search patterns with a "Sequence-Logo" style output derived from the WebLogo PostScript code (*40*). This provides a convenient way of generating graphical representations of sequence patterns that consist of collections of local motifs from a single multiple sequence alignment.

In addition to purely sequence-based methods, we also searched for more distant homologies based on combined sequence/structure patterns using Sean Eddy's RNABOB software (downloaded from ftp://ftp.genetics.wustl.edu/pub/eddy/software/rnabob-2.1.tar.Z). We constructed search patterns comprising the most conserved motifs of the histone-binding site, the Sm-binding motif, and the stem-loop structure at the 3′ end that is enclosed by two GC pairs. In order to increase specifity, we additionally included a species-specific model of the PSE, which was derived from the upstream regions of the spliceosomal snRNAs U1, U2, U4, U5, U4atac, U11, and U12. These snRNAs are larger and better conserved than the U7 snRNAs. Hence they were straightforward to find in most of the metazoan genomes where they were not annotated previously. The RNABOB descriptors are listed in

Supporting Online Material (http://www.bioinf.unileipzig.de/Publications/SUPPLEMENTS/07-010/).

Acknowledgements

BMRS and PFS thank the PICB in Shanghai for its hospitality, where much of this work was performed in spring 2007. This work was supported by the DFG-funded *Graduiertenkolleg Wissensrepräsentation* to MM and the DFG Bioinformatics Initiative to PFS. We thank an anonymous referee for bringing evidence for the special histone 3′ end processing mechanism outside metazoa to our attention.

Authors' contributions

All authors collaborated in data analysis and homology search as well as in the interpretation of the data. AM and PFS conceived the study and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Mowry, K.L. and Steitz, J.A. 1987. Identification of the human U7 snRNP as one of several factors involved in the 3′ end maturation of histone premessenger RNAs. *Science* 238: 1682-1687.
2. Marzluff, W.F. 2005. Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. *Curr. Opin. Cell Biol.* 17: 274-280.
3. Fabry, S., *et al.* 1995. The organization structure and regulatory elements of *Chlamydomonas* histone genes reveal features linking plant and animal genes. *Curr. Genet.* 28: 333-345.
4. Townley-Tilson, W.H., *et al.* 2006. Genome-wide analysis of mRNAs bound to the histone stem-loop binding protein. *RNA* 12: 1853-1867.
5. Golembe, T.J., *et al.* 2005. Specific sequence features, recognized by the SMN complex, identify snRNAs and determine their fate as snRNPs. *Mol. Cell. Biol.* 25: 10989-11004.
6. Azzouz, T.N. and Schümperli, D. 2003. Evolutionary conservation of the U7 small nuclear ribonucleoprotein in *Drosophila melanogaster*. *RNA* 9: 1532-1541.
7. Pillai, R.S., *et al.* 2003. Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing. *Genes. Dev.* 17: 2321-2333.
8. Schümperli, D. and Pillai, R.S. 2004. The special Sm core structure of the U7 snRNP: far-reaching significance of a small nuclear ribonucleoprotein. *Cell. Mol. Life Sci.* 61: 2560-2570.
9. Kolev, N.G. and Steitz, J.A. 2006. *In vivo* assembly of functional U7 snRNP requires RNA backbone flexibility within the Sm-binding site. *Nat. Struct. Mol. Biol.* 13: 347-353.
10. Jaeger, S., *et al.* 2006. Binding of human SLBP on the 3′-UTR of histone precursor H4-12 mRNA induces structural rearrangements that enable U7 snRNA anchoring. *Nucleic Acids Res.* 34: 4987-4995.
11. Brun, C., *et al.* 2003. U7 snRNAs induce correction of mutated dystrophin pre-mRNA by exon skipping. *Cell. Mol. Life Sci.* 60: 557-566.
12. Goyenvalle, A., *et al.* 2004. Rescue of dystrophic muscle through U7 snRNA-mediated exon skipping. *Science* 306: 1796-1799.
13. Soldati, D. and Schümperli, D. 1988. Structural and functional characterization of mouse U7 small nuclear RNA active in 3′ processing of histone pre-mRNA. *Mol. Cell. Biol.* 8: 1518-1524.
14. Gruber, A., *et al.* 1991. Isolation of an active gene and of two pseudogenes for mouse U7 small nuclear RNA. *Biochim. Biophys. Acta* 1088: 151-154.
15. Phillips, S.C. and Turner, P.C. 1992. A transcriptional analysis of the gene encoding mouse U7 small nuclear RNA. *Gene* 116: 181-186.
16. Phillips, S.C. and Turner, P.C. 1991. Sequence and expression of a mouse U7 snRNA type II pseudogene. *DNA Seq.* 1: 401-404.
17. Yu, Y.T., *et al.* 1996. More Sm snRNAs from vertebrate cells. *Exp. Cell Res.* 229: 276-281.
18. Strub, K., *et al.* 1984. The cDNA sequences of the sea urchin U7 small nuclear RNA suggest specific contacts between histone mRNA precursor and U7 RNA during RNA processing. *EMBO J.* 3: 2801-2807.
19. De Lorenzi, M., *et al.* 1986. Analysis of a sea urchin gene cluster coding for the small nuclear U7 RNA, a rare RNA species implicated in the 3′ editing of histone precursor mRNAs. *Proc. Natl. Acad. Sci. USA* 83: 3243-3247.
20. Gilmartin, G.M., *et al.* 1988. Functional analysis of the sea urchin U7 small nuclear RNA. *Mol. Cell. Biol.* 8: 1076-1084.
21. Southgate, C. and Busslinger, M. 1989. *In vivo* and *in vitro* expression of U7 snRNA genes: cis- and trans-acting elements required for RNA polymerase II-directed transcription. *EMBO J.* 8: 539-549.
22. Phillips, S.C. and Birnstiel, M.L. 1992. Analysis of a gene cluster coding for the *Xenopus laevis* U7 snRNA. *Biochim. Biophys. Acta* 1131: 95-98.

23. Watkins, N.J., *et al.* 1992. The U7 small nuclear RNA genes of *Xenopus borealis*. *Biochem. Soc. Trans.* 20: 301S.
24. Wu, C.H. and Gall, J.G. 1993. U7 small nuclear RNA in C snurposomes of the *Xenopus* germinal vesicle. *Proc. Natl. Acad. Sci. USA* 90: 6257-6259.
25. Dominski, Z., *et al.* 2003. Cloning and characterization of the *Drosophila* U7 small nuclear RNA. *Proc. Natl. Acad. Sci. USA* 100: 9422-9427.
26. Myslinksi, E., *et al.* 2004. Characterization of snRNA and snRNA-type genes in the pufferfish *Fugu rubripes*. *Gene* 330: 149-158.
27. Lück, R., *et al.* 1999. Construct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res.* 27: 4208-4217.
28. Bompfünewerer, A.F., *et al.* 2005. Evolutionary patterns of non-coding RNAs. *Theory Biosci.* 123: 301-369.
29. Griffiths-Jones, S., *et al.* 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 33: D121-124.
30. Hernandez, N. 2001. Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. *J. Biol. Chem.* 276: 26733-26736.
31. Hernandez, G., Jr., *et al.* 2007. Insect small nuclear RNA gene promoters evolve rapidly yet retain conserved features involved in determining promoter activity and RNA polymerase specificity. *Nucleic Acids Res.* 35: 21-34.
32. Soldati, D. and Schümperli, D. 1990. Structures of four human pseudogenes for U7 small nuclear RNA. *Gene* 95: 305-306.
33. Dominski, Z., *et al.* 2005. Differences and similarities between *Drosophila* and mammalian 3′ end processing of histone pre-mRNAs. *RNA* 11: 1835-1847.
34. Mosig, A., *et al.* 2006. Fragrep: an efficient search tool for fragmented patterns in genomic sequences. *Genomics Proteomics Bioinformatics* 4: 56-60.
35. Nawrocki, E.P. and Eddy, S.R. 2007. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.* 3: e56.
36. Morgenstern, B. 1999. DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* 15: 211-218.
37. Thompson, J.D., *et al.* 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.
38. Griffiths-Jones, S. 2005. RALEE—RNA alignment editor in Emacs. *Bioinformatics* 21: 257-259.
39. Hofacker, I.L., *et al.* 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* 319: 1059-1066.
40. Crooks, G.E., *et al.* 2004. WebLogo: a sequence logo generator. *Genome Res.* 14: 1188-1190.

Supporting Online Material

Alignments of U7 sequences and other data:
http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/07-010/

Note added in proof

While this manuscript was in production, two relevant papers have appeared: Higuchi *et al* (U7 snRNA acts as a transcriptional regulator interacting with an inverted CCAAT sequence-binding transcription factor NF-Y. *Biochim. Biophys. Acta.* Epub ahead of print 2007 Nov 22, doi:10.1016/j.bbagen.2007.11.005) demonstrated that U7 snRNA also acts as a transcriptional regulator, and Dávila López and Samuelsson (Early evolution of histone mRNA 3′ end processing. 2008. *RNA* 14: 1-10. Epub 2007 Nov 12) reported evidence for an origin of the metazoan-like histone 3′ end processing machinery early in eukaryotic evolution. These authors also reported several computationally predicted U7 snRNA sequences, most of which agree with our results.