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

Andreas R. Gruber*, **Dorota Koper-Emde***, **Manja Marz***, **Hakim Tafer***, **Stephan Bernhart**, **Gregor Obernosterer**, **Axel Mosig**, **Ivo L. Hofacker**, **Peter F. Stadler**, **Bernd-Joachim Benecke**

Invertebrate 7SK snRNAs

July 13, 2007

Abstract 7SK RNA is a highly abundant non-coding RNA in mammalian cells whose function in transcriptional regulation has only recently been elucidated. Despite its highly conserved sequence throughout vertebrates, all attempts to discover 7SK RNA homologs in invertebrates species have failed so far. Here we report on a combined experimental and computational survey that succeeded in discovering 7SK RNAs in most of the major deuterostome clades and in two protostome phyla: molluscs and annelids. Despite major efforts, no candidates were found in any of the many available ecdysozoan genomes, however. The additional sequence data confirm the evolutionary conservation and hence functional importance of the previously described 3' and 5' stem-loop motifs, and provide evidence for a third structurally well-conserved domain.

Keywords

7SK RNA, Polymerase III transcription, non-coding RNA, lophotrochozoans

A. Gruber, H. Tafer, S. Bernhart, I.L. Hofacker, P.F. Stadler
Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Wien, Austria
E-mail: {agruber,htafer,berni,ivo}@tbi.univie.ac.at

D. Koper-Emde, B.-J. Benecke
Biochemistry NC6/132, Ruhr-University Bochum, Germany
E-mail: bernd.benecke@rub.de

M. Marz, S. Bernhart, Peter F. Stadler
Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany
E-mail: {manja,bstephan,studla}@bioinf.uni-leipzig.de

G. Obernosterer
Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), A-1030 Vienna, Austria E-mail: Gregor.Obernosterer@imba.oeaw.ac.at

A. Mosig
Department of Combinatorics and Geometry (DCG), MPG/CAS Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences (SIBS) Campus, Shanghai, China;
Max Planck Insitute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany
E-mail: {axel.mosig@gmail.com

P.F. Stadler
RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie — IZI Deutscher Platz 5e, D-04103 Leipzig, Germany;
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

*These authors contributed equally

1 Introduction

The 7SK snRNA is a highly abundant noncoding RNA in vertebrate cells. The Pol III transcript with a length of about 330nt [14,20] is highly conserved in vertebrates [9]. Due to its abundance it has been known since the 1960s. Its function as a transcriptional regulator, however, has only recently been discovered. 7SK mediates the inhibition of the general transcription elongation factor P-TEFb by the HEXIM1 protein and thereby represses transcript elongation by Pol II [16, 2,6,22]. Furthermore, 7SK RNA suppresses the deaminase activity of APOBEC3C and sequesters this enzyme in the nucleolus [10].

Two distinct secondary structure elements are highly conserved throughout vertebrates [6]: a 5'-terminal hairpin structure that binds both HEXIM1 and P-TEFb, and a 3'-terminal hairpin that interacts with P-TEFb only. In contrast to the nearly perfect sequence conservation in jawed vertebrates, the 7SK RNA from the lamprey *Lampetra fluviatilis* differed in more than 30% of its nucleotide positions from its mammalian counterpart [9]. The highest sequence conservation is observed in the 5' and 3' hairpin regions. The sequence conservation seems to decline rapidly outside the gnathostomes. In [9], some of us also reported on an unsuccessful attempt to finding 7SK RNA in hagfish and lancet and suggested that the 7SK RNA might be a vertebrate innovation. In this contribution we combine improved cloning strategies with systematic computational homology searches to detect highly divergent 7SK RNAs in invertebrate animals.

2 Materials and Methods

2.1 Cloning and Sequencing of 7SK RNAs

Total cellular RNA was isolated from frozen tissue minced with scissors and homogenized in buffer containing guanidinium thiocyanate [4]. Northern blots were performed with 5 μg of purified RNA separated in 2%-agarose gels containing 0.67 M formaldehyde. After transfer to nylon membranes (Hybond-N, Amersham), immobilized RNA was hybridized with labelled antisense RNA probes generated by T7 transcription of inversely cloned cDNA fragments of the previously identified 7SK RNA of *Lampetra fluviatilis* [9]. Hybridization with labelled antisense RNA (2×10^6 cpm/ml) was in: 50% formamide; 0.1% SDS; $5\times$ Denhardt's reagent; 10 μg/ml each of yeast tRNA and denatured salmon sperm DNA; $5\times$ SET (150 mM NaCl, 20 mM Tris-HCl, pH 7.9, 1 mM EDTA).

Cloning of new 7SK cDNAs was based on RT-PCR reactions. In a first step 4 μg of total RNA were reverse transcribed using the Omniscript RT-Kit (Qiagen) with a primer complementary to the 3'-end of 7SK RNA. An aliquot of that assay was taken for PCR with the same 3'-primer and a specific upstream primer. The successful combination for *Myxine* and *Branchiostoma* amplification were nucleotides 54-72 as upstream primer and nucleotides 197-215 as downstream primer, both of the human 7SK DNA sequence. For *Helix*, the same upstream primer, yet in combination with a Lamprey downstream primer (corresponding to position 299-316) was successful. Candidate fragments were cloned, sequenced and taken to deduce gene-specific "nested" primers for rapid amplification of cDNA ends (RACE) experiments [7], as described earlier [9]. Briefly, for the 3'-ends cellular RNA was first polyadenylated and reverse transcription started with with oligo(dT) carrying at its 5'-side an oligonucleotide sequence suitable for subsequent PCR with two "nested" gene-specific primers. The 5'-ends were obtained by reverse transcription with a specific primer and oligonucleotide ligation (T4-RNA ligase) to the 3'-end of first-strand cDNAs. As before, PCR amplification was achieved with nested gene-specific primers.

2.2 Computational Homology Search

Homology search was performed as a stepwise procedure. In the first stage, we started with the sequence of the functional human 7SK sequence (***X05490***, ***X04236*** [26, 19, 14, 25]) and performed a `blast` search against the genome assemblies available in `ensembl` (version 42). In this way, we identified candidates in other vertebrate genomes, including the following, previously published sequences: *Mus musculus* (***M63671*** [17]), *Rattus norvegicus* (***K02909*** [23]), *Takifugu rubripes* (***AJ890104***, [21, 6]), *Tetraodon nigroviridis* (***AJ890103***, [6]), *Danio rerio* (***AJ890102***, [6]), and *Gallus gallus* (***AJ890104***, [6]). In addition we searched the shotgun traces of a selection of unfinished mammalian genomes

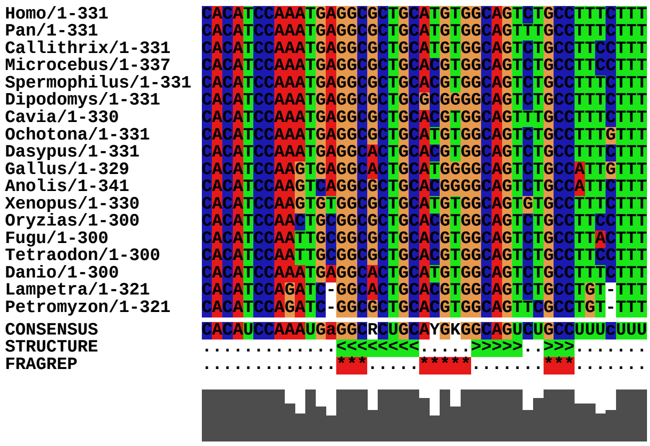

Fig. 1 Alignment and consensus sequence and structure of the 3' hairpin of vertebrate 7SK RNAs.

as well as all unfinished non-mammalian animals. Beyond jawed vertebrates, this initial `blast` search recovered a single candidate in the genome of the lamprey `Petromyzon marinus`, which turned out the be very closely related to the published sequence of the *Lampetra fluviatilis* 7SK RNA [9]. The match to a single shotgun trace (1047111637562) from the nematode *Brugia malayi* was disregarded since it exactly matched the human sequence and hence is certainly a contamination.

We then created a multiple sequence alignment and used `aln2pattern` to extract a search pattern for the `fragrep2` tool [18]. This tool searches fragmented approximate sequence patterns in genomic DNA sequences. The construction of the sequence patterns was guided by the functional interpretation of the structural domains of the 7SK in [6].

First efforts focused on the 3'-terminal hairpin structure, Fig. 1. Both the loop motive and the GGC-GCC stem have been shown to be crucial for P-TEFb binding and have just little or no sequence variation. Search patterns for `fragrep` are specified as partial sequence patterns (using IUPAC notation), annotated by the minimal and maximal number of unspecified nucleotides separating the sequence motifs and the maximal number of mismatches and in/dels that are acceptable in each of the sequence motifs. The initial search pattern for the 3' hairpin is indicated in the last line of Fig. 1.

By manual filtering we selected candidate sequences from *Ciona intestinalis* and *Ciona savignyi*, which showed also high sequence conservation in the part of the 5'-terminal hairpin structure responsible for HEXIM1 and P-TEFb binding [6]. We then extended our pattern to search for highly conserved regions in both the 5' and 3' hairpin structures. The pattern was refined in an iterative way every time a new candidate sequences was found. The final search pattern is given in Fig. 2

```
6
  0   0 GAUCGC  0 0
  4  16 GGC     0 0
  1   2 GAUCUG  0 0
150 250 UCCA    0 0
  0   5 GGC     0 0
  0  19 GCC     0 0
```

Fig. 2 Final `fragrep` search pattern for 7SK RNAs, specifying 6 sequence motifs, their minimal and maximal separation and the number of acceptable substitutions and in/dels. Here exact matches of the short motifs are required.

2.3 Computational Identification of Putative 7SK Promoters

It is well known that the promoters of RNA Polymerase III transcribed 7SK, U6, and U6atac snRNA contain three common elements: the *proximal sequence element* (PSE) about 50nt upstream of the gene, a TATA box-like element and distal enhancer elements [25,5]. The PSE of the pol-III snRNAs is very similar to that on the snRNA transcribed by Polymerase II (U1, U2, U4, U5, U11, U12, U4atac).

In order to distinguish functional 7SK genes from pseudogenes we investigated their upstream regions for snRNA-specific promoter elements. Since these sequence motifs can vary significantly [11] between species, we searched the genomes also for the spliceosomal snRNAs, extracted 100nt upstream regions, and used `meme` (version 3.5.4) [1] to identify the PSE consensus separately for each species. For the study reported here we used the applet available from `http://meme.nbcr.net/downloads/` with default options and parameters `-nmotifs 5 -minw 10 -maxw 30`. The PSE patterns obtained in this way were then used to identify those 7SK candidates that have an snRNA-like PSE. The results of the homology search for the spliceosomal RNA genes will be reported elsewhere [15].

2.4 Structural Alignments

A structural alignment of the vertebrate sequences based on the experimentally determined structure for the human 7SK snRNA [14,25] was constructed manually using the RALEE mode [8] for the `emacs` editor. The model was iteratively improved upon addition of new candidate sequences during analysis with the help of consensus structure predictions using `RNAalifold` [12].

The 5' stem sequence of the basal deuterostomes and lophotrocozoa diverged too much from the vertebrate consensus, so that they cannot be aligned based on sequence similarity alone. We therefore used the absolutely conserved GATC-GATC stem in the center of this region as a anchor since it defined both sequence and structure constraints. The alignment was then edited so as to maximize the number of base pairs in the consensus structure.

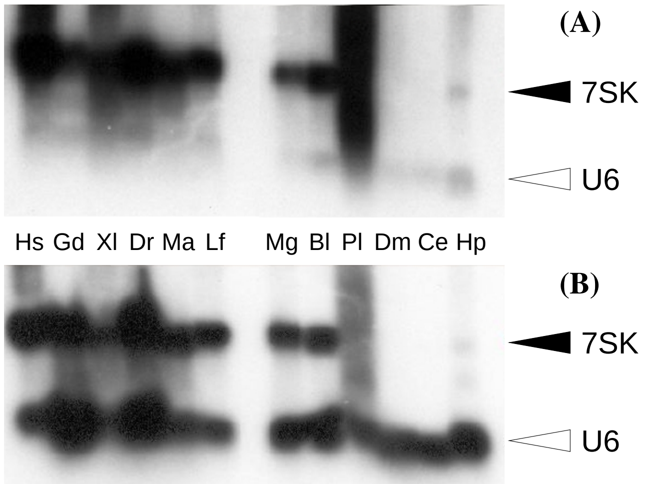

Fig. 3 Northern blot of cellular RNA from 12 different organisms using **(A)** 3'- and **(B)** 5'-terminal 7SK antisense RNA for hybridization. Hs *Homo sapiens*, Gd *Gallus domesticus*, Xl *Xenopus laevis*, Dr *Danio rerio*, Ma *Mustelus asterias*, Lf *Lampetra fluviatilis*, Mg *Myxine glutinosa*, Bl *Branchistoma lanceolatum*, Pl *Paracentrotus lividus*, Dm *Drosophila melanogaster*, Ce *Caenorhabditis elegans*, Hp *Helix pomatis*.
For normalization among samples, a labelled full length antisense U6 snRNA has been included in both hybridizations. In the upper panel, weaker U6 signals are observed throughout. This is due to the re-hybridization of the stripped blot. In both rounds, U6 antisense RNA hybridizes to the same target sequences whereas the two 7SK probes bind to different areas of the 7SK RNA. The broad smear observed with sea urchin RNA is due to cross contamination of the 7SK antisense probe with ribosomal RNA and degradation products thereof.

3 Results

3.1 Northern Blot Verification of 7SK Sequences

Numerous attempts to identify 7SK RNA in invertebrate phyla have remained unsuccessful in the past. Neither RT-PCR experiments with mammalian primers nor northern blot analysis with oligonucleotide-primed cDNA probes were successful [9]. Therefore, we decided to increase northern blot sensitivity by using radioactively labelled antisense RNA probes. For this, the two most conserved (in comparison with the human sequence) regions of the lamprey (*Lampetra fluviatilis*) 7SK RNA gene [9] were subcloned in inverted orientation under control of the T7 RNA polymerase promoter. Thus, these two clones allowed the synthesis *in vitro* of labelled transcripts with very high specific activity. The resulting two antisense RNA probes were complementary to regions 1-94 **(A)** and 283-316 **(B)** of lamprey 7SK RNA, respectively.

A northern blot obtained with cellular RNA of six vertebrate and six invertebrate organisms was consecutively hybridized with the aforementioned probes, Fig. 3. As is evident from the left section of Fig. 3, all vertebrate phyla analyzed (man to lamprey) showed a clear hybridization signal with both the 3-terminal (upper panel) and the 5-terminal (lower panel) 7SK antisense RNA. In contrast, only hagfish

(*Myxine glutinosa*) and two of the five invertebrates in the r.h.s. section were found positive with respect to the 7SK antisense probes: amphioxus (*Branchiostoma lanceolatum*) and snail (*Helix pomatia*). Even longer exposure did not give a hint to any 7SK hybridization signal of either of the two probes with sea urchin (*Paracentrotus lividus*), fruitfly (*Drosophila melanogaster*) or nematode (*Caenorhabditis elegans*) RNA.

3.2 cDNA Cloning of Novel 7SK RNAs

Cloning of invertebrate 7SK cDNA was performed by RT-PCR and primers deduced from the most conserved elements of vertebrate 7SK RNA. Routinely, about ten different primer combinations had to be tested. In many cases, PCR fragments with the expected lengths were obtained. After subcloning and sequencing, however, most fragments were found to represent pieces of ribosomal DNA. Only a single previously unidentified sequence was amplified from hagfish RNA. Two identical sequences, yet differing in length, were obtained from amphioxus. A single new clone was obtained from snail. These clones showed a limited but significant sequence homology with the corresponding central sections of the vertebrate 7SK DNA. Therefore, the 3- and 5-flanking sequences of those clones were amplified by RACE experiments. After subcloning and sequencing composite 7SK RNA sequences were obtained for *Myxine glutinosa* (329 nt), *Branchiostoma lanceolatum* (304 nt), and *Helix pomatia* (303 nt).

These three 7SK RNA clones revealed sequence homology with lamprey 7SK DNA in the range between 44% (snail) and 59% (hagfish). The identification of several interspersed elements (7 to 11 nucleotides in length) with a perfect match to the vertebrate 7SK RNA sequence strongly supported the notion that 7SK cDNA has been successfully cloned from two basal chordates (hagfish and amphioxus) and a single non-deteuterostome invertebrate. In addition, report here sequences for *Gadus morrhua* and *Mustelus asterias*. All sequences have been deposited in GenBank, accession numbers ******-******. Multiple sequence alignments can be found in an electronic supplement available at `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/07-021/`.

3.3 Homology Search

3.3.1 Vertebrates

Within vertebrates, homology search turned out to be rather straightforward. Simple `blastn` searches were sufficient. In eutherians, however, it is a problem to identify the functional 7SK gene among a larger number 7SK-derived pseudogenes. In fact, searching ENSEMBL v.44 with an E-value cutoff of 10^{-4} return more than 100 hits in all eutherian genomes. In contrast, there are only 31 hits in *Monodelphis domestica* and 11 hits in the chicken genome. The current assembly of the genome of *Xenopus tropicalis* features two adjacent copies. These are identical also in an extended flanking sequence, indicating a recent segmental duplication of the locus or an assembly artifact, see e.g. [3]. In each of the five sequenced teleost fishes, only a single copy of the 7SK is present. Three `blast` hits were found in the `pre-ensembl` release of the sea lamprey genome. Only a single one, located on Contig17254, matches the published sequence from *Lampetra fluviatilis* over its full length.

In three vertebrate species, however, we failed to find a complete 7SK gene. Only a single partial hit was recovered from the low-coverage genome elphant shark *Callorhynchus milli*. All good `blast` hits of the chicken 7SK sequence against the available *Taeniopygia guttata* shotgun reads seem to belong to a single locus. The corresponding sequence very well matches the chicken sequence but shows a 398nt insert, which we interpret as an artifact. To our surprise, only a single `blastn` hit was found in the Platypus genome. The corresponding sequence significantly deviates from the vertebrate consensus in both the first $\sim$ 8 nt and in the last $\sim$ 100nt, and it is not located in a region that is syntenic to the functional 7SK genes in other vertebrates. This `blast` hit thus is most likely a pseudogene. Since the locus around the platypus ICK homolog is incompletely assembled, it is reasonable to assume that we fail to find the platypus 7SK due to missing data, not because platypus has lost its functional 7SK RNA.

The functional 7SK genes can be recognized by the characteristic PSE sequence in their promotor regions, see Fig. 4. Furthermore, their genomic location between the Glutathione S-Transferase Alpha GTA4 on the 5' side and Serine/Threonine protein kinase ICK on the 3' side is conserved throughout vertebrates. We remark that in *Xenopus* no GTA4 homolog is annotated.

3.3.2 Basal Deuterostomes

While the vertebrate 7SK RNA are very well conserved at sequence level [9], `blast` searches soon reached their limits outside of this clade. A weak blast hit of the human query sequence in the *Branchiostoma floridae* genome was easily verified by comparison with the experimentally determined *Branchiostoma lanceolatum* 7SK RNA sequence. In total, we find six nearly identical 7SK candidates on five different scaffolds. These sequences are also nearly identical at least 100nt upstream of the 7SK. It is unclear whether there are really multiple functional copies of 7SK RNA dispersed in the amphioxus genome.

Beyond amphioxus, three further candidates were found by means of `fragrep` only: two closely related sequences from the urochordates *Ciona intestinalis* and *Ciona savignyi*, and a single candidate from the hemichordate *Saccoglossus kowalevskii*. In the *C. intestinalis* genome there is only a single 7SK locus. In contrast, the current assembly of *C. savignyi* features four nearly identical copies within 8kb on reftig_107.

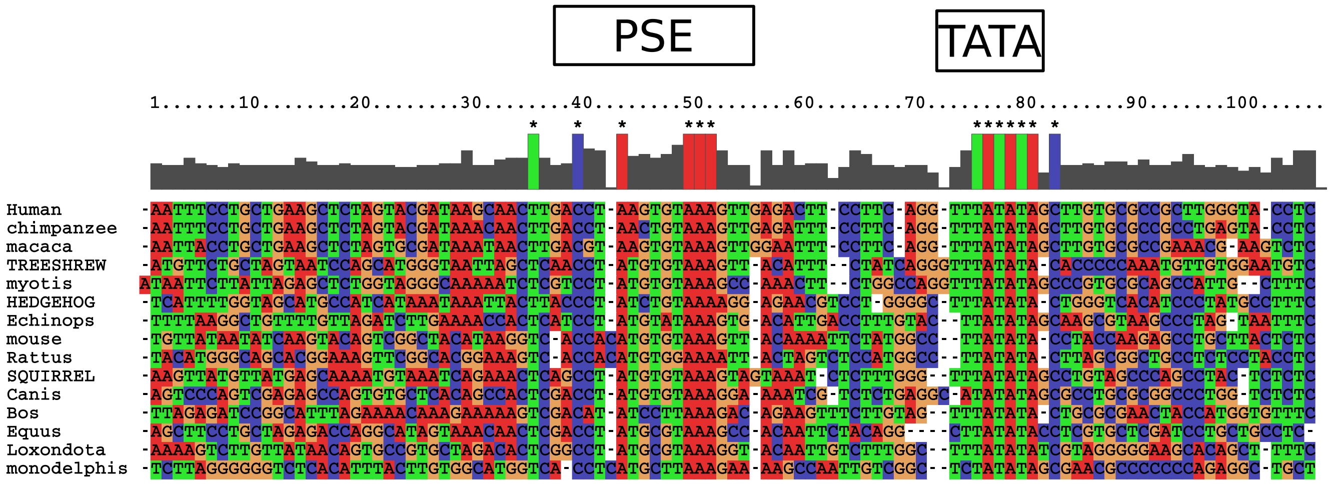

Fig. 4 Conservation of the the 7SK promoter region of the 7SK candidates located between GTA4 and ICK genes. In all mammals, both the "proximal sequence element" (PSE) and the TATA-box are easily recognizable.

The PSE of basal deuterostomes, 63-48nt upstream of 7SK, is well conserved with the one of U6 snRNA for all investigated organisms. The TATA-box is located 32-25nt upstream. An exception is *Ciona intestinalis*, which has an insertion of 15nt between TATA-box and 7SK, and a 9nt deletion between the PSE and TATA. The *Ciona* species shows no TATA-box conservation with other snRNAs, nevertheless there is a slightly modified TATA box. *Branchiostoma* has a canonical TATA-box.

Despite significant efforts we did not find credible candidates in the genome of the sea urchin *Strongylocentrotus purpuratus*. The three best candidates in this case lack the 3' hairpin structure. In addition, the 5' hairpin region can hardly be aligned with other deuterostome 7SK sequences. Our search also failed for the shotgun traces of the urochordate *Oikopleura dioica*. In this case we found a good candidate for the 3' stem-loop structure, but the 300nt upstream of this hit do not match other 7SK sequences. We suspect that these negative result might in these cases be the result of the incomplete genomic data .

3.3.3 Protostome 7SK RNAs

The `fragrep` search was successful in three protostome genomes: the mollusc *Lottia gigantea* and the two annelids *Capitella capitella* and *Helobdella robusta*. All three sequences are easily recognizable as homologs of the 7SK sequence that was cloned from the escargot *Helix pomatia*. In addition, a partial sequence from *Aplysia californica* was found by `blast` using the experimentally determined escargot 7SK sequence as query.

The PSE of *Aplysia californica* is located 67nt-41nt upstream of 7SK snRNA, the sequence motif TGTATAGA matches the typical TATA-box sequence 35nt-28nt upstream. In *Lottia gigantea* we find CTTATATA (position -31 to -24) and the PSE 15nt upstream of the TATA box. In *Capitella* we find TATACA at positions -27 to -21 and a possible PSE, although it does not match well with the upstream sequence of the U6 snRNA in this species. The single shotgun read from *Helobdella robusta* does not show a recognizable TATA-box region but an alignable PSE region. It is not clear whether this sequence is a functional gene or a 7SK-derived pseudogene.

Despite extensive efforts, on the other hand, no 7SK candidate was found in any of the many available insect and nematode genomes. A search in the genomes of the two platyhelminths *Schmidtea mediterranea* and *Schistosoma mansoni* also remained unsuccessful. Among protostomes, thus, the 7SK RNA can be found only among lophotrochozoans.

Given the lack of success on ecdysozoan genomes and the four highly derived lophotrochozoan sequences we were not surprised that searches in the genomes of diploblastic animals and in the choanoflagellate *Monosiga brevicollis* were also not successful.

3.4 Refined Structural Models of 7SK RNAs

The conservered structural features of the 7SK sequences are summarized in Figure 5. Three structural elements are common to all known 7SK RNAs: the 5'stem, the 3'stem, and a short hairpin structure (Stem A) that should probably be considered as part of the 5' structure. Vertebrate 7SK in addition share Stem B.

The secondary structures for the 5' and 3' stems were proposed already in previous publications. Wassarman and Steitz [25] derived a model for the human 7SK snRNA based on chemical probing data. Egloff *et al.* [6] used site-directed mutategenesis to demonstrate that both the 5' and 3' stems are functionally relevant. The structural model in Fig. 5 is derived from a sequence alignment that takes into account both sequence covariation and thermodynamic considerations. Our consensus model is in agreement with the pre-

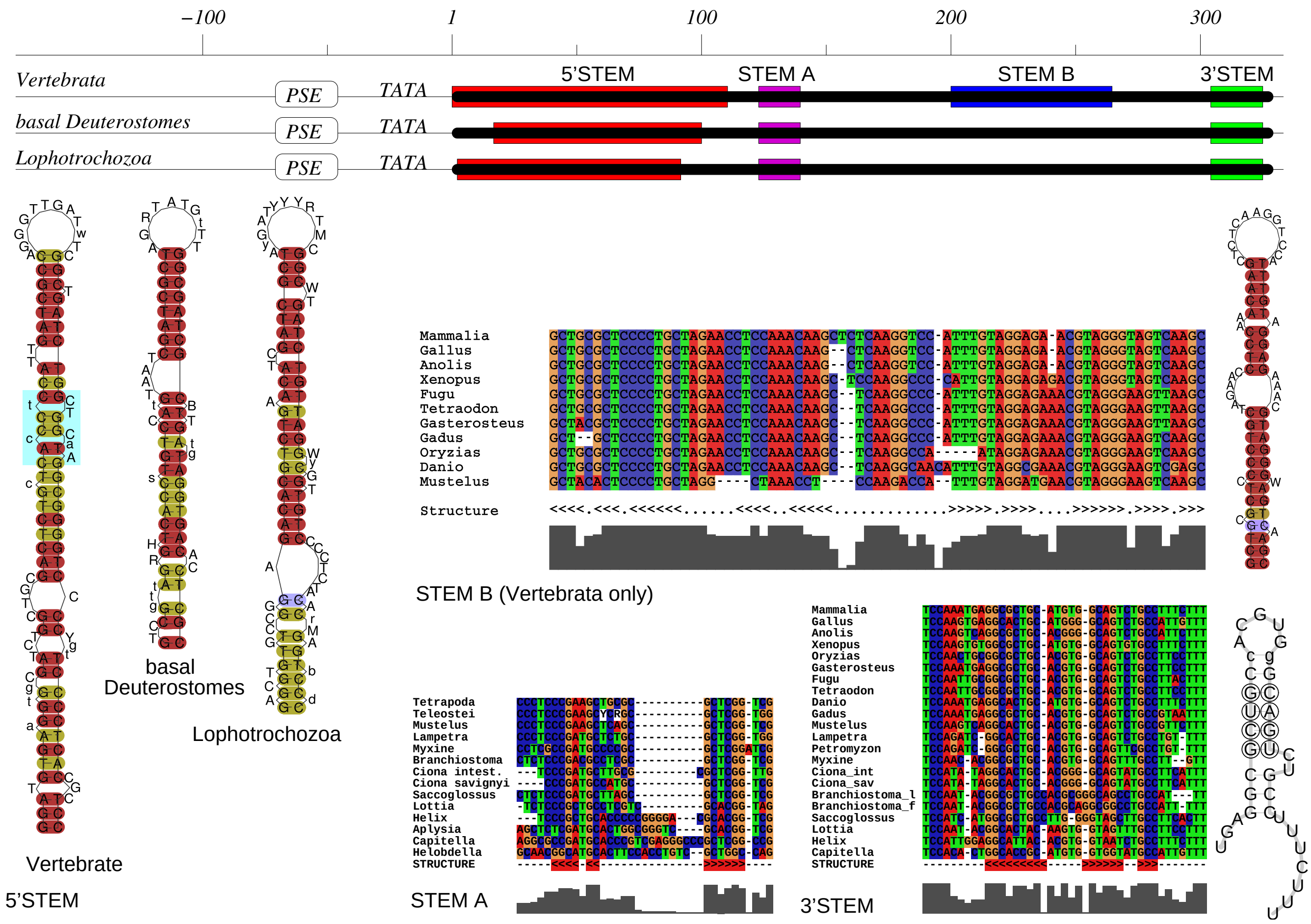

Fig. 5 Common structural elements of 7SK snRNAs. The top panel schematically compares the location of upstream elements and RNA secondary features. While the structure of the 3'stem is common to all 7SK snRNAs (except for the elongation of the stem by a GC pair in *Branchiostoma* and *Saccoglossus*), there are substantial clade-specific variations in the 5'stem. A common structure, stem B, in the "middle region", on the other hand, can be found only in vertebrates. With the expection of marginal differences in the small region marked in the vertebrate 5'stem, our consensus model is in complete agreement with previously published structures of vertebrate 7SK snRNAs [25,6]. Conserved nucleotides in stems are shown in red; ochre color (and circles in the 3'stem, resp.) indicate consistent and compensatory mutations.

viously published structures with a marginal exception: In [6] the regions marked in Fig. 5 is shown as an interior loop, while [25] shows only the terminal A-U as part of the interior loop.

The 5' stem models for both basal deuterostoma and for the lophotrochozoa are different in size sequence and structure. The only common ground between all three models is the GATC-GATC structure/sequence pattern at the beginning of the top-most stem.

The vertebrate specific Stem B, which is not necessary for P-TEFb binding [6], also fits very well with both experimental models, again with a small difference affecting a small interior loop. It does not appear to have a counterpart in basal deuterostomes and protostomes.

The central region of the 3'stem is structurally conserved in all 7SK RNAs, the only exception being an extension of the most central stem by a single GC pair in *Branchiostoma* and *Saccoglossus*. The 3' stem-loop structure can be extended by five additional base pairs in vertebrates and to a lesser extent also in the other 7SK snRNAs. The exact pairing pattern in this extended region does not seem to be very well conserved, however.

The small Stem A feature, finally, is highly conserved also in sequence across all known 7SK snRNA, although the size of the loop region is variable in the lophotrochozoan sequences. So far, no specific function has been reported for this region.

4 Discussion

Using both computational and "wet-lab" approaches we have shown that 7SK is phylogenetically much older than previously thought. While previously examples were known only from vertebrates, we found 7SK RNAs in most major deuterostome clades (chephalochordata, urochordata, and hemichordata) and in two lophotrochozoan clades (mollusca and annelida), Fig. 6.

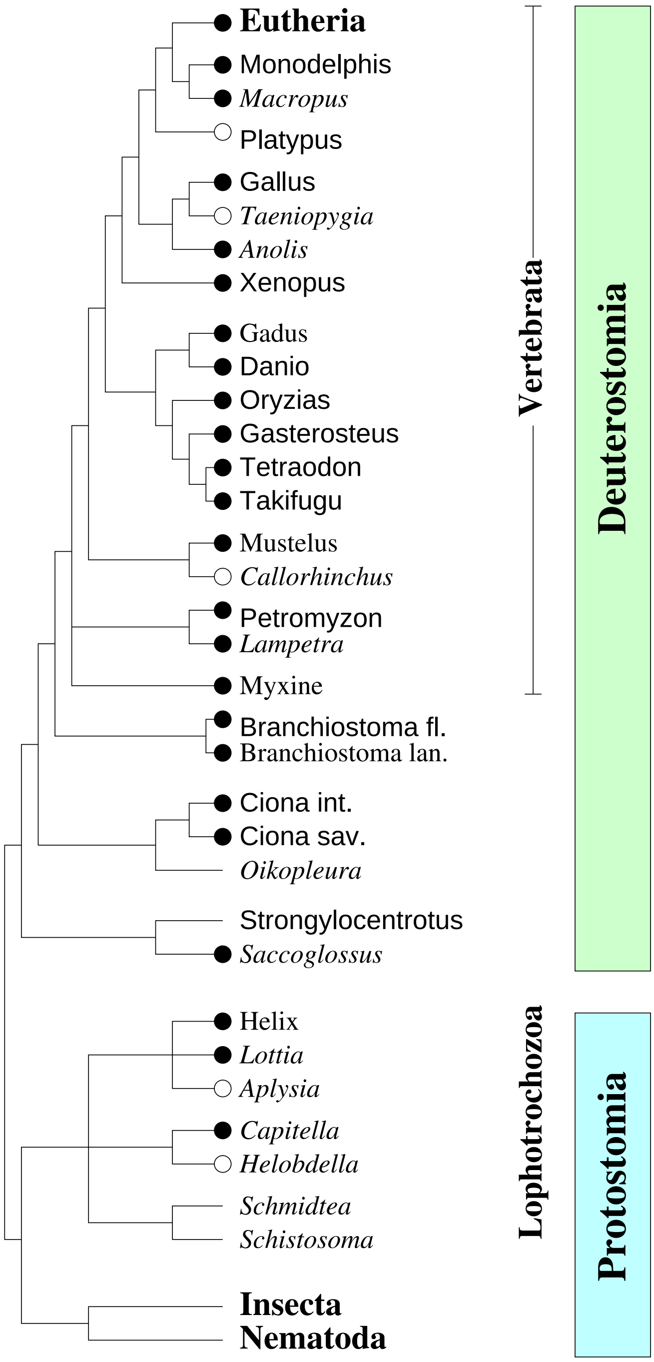

Fig. 6 Phylogenetic distribution of 7SK sequences. bold font: clades in which multiple genomes were searched with consistent results. Sans serif font: assembled genomes, *italics* font: shotgun traces, roman font: no genomic sequence available, but positive experimental results. Black bullets (●) indicate that a complete *bona fide* 7SK sequence was found, white bullets (○) mark species where we found a partial 7SK sequence. The phylogenetic tree follows the recent review [24].

Assuming that the commonly accepted sistergroup relationship of Protostomia and Deuterostomia is indeed correct, our findings imply that the 7SK originated at latest in the bilaterian ancestor. In contrast, we found no trace of a 7SK RNA candidate in either platyhelminthes or in any of the numerous ecdysozoan species for which genomic data are available.

The analysis of sequences and secondary structures revealed a striking difference between vertebrate and invertebrate sequences. While vertebrate 7SK RNAs are highly conserved in both sequence and structure, the molecule is highly variable in the other clades. Consensus structure models derived using a combination of thermodynamic folding and evaluation of compensatory mutations reveal three structural motifs that are conserved throughout all known 7SK sequences. The central domain (Stem B), however is present in vertebrates only, while elsewhere this region is so variable that our attempts to construct plausible alignments failed.

The monophyly of the Ecdysozoa is, among other arguments, also supported [24] by the shared secondary absence of large numbers of genes in euarthropods and nematodes [13]. There is no functionally described HEXIM1 ortholog in insects, although the current release of the ENSEMBL (v.44) homology annotation lists HEXIM homologs in *Drosophila melanogaster* (CG3508), *Aedes aegyptii* (AAEL013291), and *Anopheles gambiae* (AGAP002875). None of these sequences is at present associated with any functional annotation, however. It is conceivable, therefore, that an ancestral 7SK gene has been secondarily lost in this clade. Alternatively, the 7SK sequence might have diverged so far that it is not recognizable with currently available bioinformatics approaches.

Acknowledgements

This work has multiple roots: Dorota Koper-Emde's PhD thesis at the Ruhr University Bochum (2004), Manja Marz' MSc thesis at the University of Leipzig (2006), and an Advanced Bioinformatics Computer Lab Course on "RNA Homology Search" organized by PFS at the University of Vienna in Fall 2006. It was supported in part by the DFG Bioinformatics Initiative and the GK Wissensrepräsentation (Leipzig), and by the Austrian GE-NAU project "non coding RNA" (Vienna).

Supporting Material

Supplemental data, in particular machine-readable sequence alignments, tables of genomic coordinates, and lists of PSE elements can be found at `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/07-021/`.

References

1. Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceed-

ings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28–36. AAAI Press, Menlo Park, CA (1994)
2. Blazek, D., Barboric, M., Kohoutek, J., Oven, I., Peterlin, B.M.: Oligomerization of HEXIM1 via 7SK snRNA and coiled-coil region directs the inhibition of P-TEFb. Nucleic Acids Res. **33**, 7000–7010 (2005)
3. Cheung, J., Estivill, X., Khaja, R., MacDonald, J.R., Lau, K., Tsui, L.C., Scherer, S.W.: Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol. **4**, R25 (2003)
4. Chomczynski, P., Sacchi, N.: Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. **162**, 156–159 (1987)
5. Dahlberg, J.E., Lund, E.: The genes and transcription of the major small nuclear RNAs. In: M. Birnstiel (ed.) Structure and function of major and minor small nuclear ribonucleoprotein particles, pp. 38–70. Springer-Verlag, Berlin, Germany (1988)
6. Egloff, S., Van Herreweghe, E., Kiss, T.: Regulation of polymerase II transcription by 7SK snRNA: two distinct RNA elements direct P-TEFb and HEXIM1 binding. Mol. Cell. Biol. **26**, 630–642 (2006)
7. Frohman, M.A., Dush, M.K., Martin, G.R.: Rapid production of full-length cDNAs from rare transcripts: Amplification using single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. USA **85**, 8998–9002 (1988)
8. Griffiths-Jones, S.: RALEE–RNA ALignment editor in Emacs. Bioinformatics **21**, 257–259 (2005)
9. Gürsoy, H.C., Koper, D., Benecke, B.J.: The vertebrate 7S K RNA separates hagfish (*Myxine glutinosa*) and lamprey (*Lampetra fluviatilis*). J. Mol. Evol. **50**, 456–464 (2000)
10. He, W.J., Chen, R., Yang, Z., Zhou, Q.: Regulation of two key nuclear enzymatic activities by the 7SK small nuclear RNA. Cold Spring Harb Symp Quant Biol. **71**, 301–311 (2006)
11. Hernandez Jr, G., Valafar, F., Stumph, W.E.: 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 (2007)
12. Hofacker, I.L., Fekete, M., Stadler, P.F.: Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. **319**, 1059–1066 (2002)
13. Hughes, A.L., Friedman, R.: Shedding genomic ballast: extensive parallel loss of ancestral gene families in animals. J. Mol. Evol. **59**, 827–833 (2004)
14. Krüger, W., Benecke, B.J.: Structural and functional analysis of a human 7 S K RNA gene. J. Mol. Biol. **195**, 31–41 (1987)
15. Marz, M., Kirsten, T., Stadler, P.F.: Evolution of spliceosomal snrna genes (2007). Manuscript in preparation
16. Michels, A.A., Fraldi A. Li, Q., Adamson, T.E., Bonnet, F., Nguyen, V.T., Sedore, S.C., Price, J.P., Price, D.H., Lania, L., , Bensaude, O.: Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor. EMBO J. **23**, 2608–2619 (2004)
17. Moon, I.S., Krause, M.O.: Common RNA polymerase i, ii, and iii upstream elements in mouse 7SK gene locus revealed by the inverse polymerase chain reaction. DNA Cell Biol. **10**, 23–32 (1991)
18. Mosig, A., Chen, J.L., Stadler, P.F.: Homology search with fragmented nucleic acid sequence patterns. In: R. Giancarlo, S. Hannenhalli (eds.) WABI 2007, *Lecture Notes in Computer Science*, vol. 4645, pp. 335–345. Springer Verlag, Berlin, Heidelberg (2007)
19. Murphy, S., Altruda, F., Ullu, E., Tripodi, M., Silengo, L., Melli, M.: DNA sequences complementary to human 7 SK RNA show structural similarities to the short mobile elements of the mammalian genome. J. Mol. Biol. **177**, 575–590 (1984)
20. Murphy, S., Di Liegro, C., Melli, M.: The *in vitro* transcription of the 7SK RNA gene by RNA polymerase III is dependent only on the presence of an upstream promoter. Cell **51**, 81–87 (1987)
21. Myslinksi, E., Krol, A., Carbon, P.: Characterization of snRNA and snRNA-type genes in the pufferfish *Fugu rubripes*. Gene **330**, 149–158 (2004)
22. Peterlin, B.M., Price, D.H.: Controlling the elongation phase of transcription with P-TEFb. Mol. Cell. **2006**, 23 (297-305)
23. Reddy, R., Henning, D., Subrahmanyam, C.S., Busch, H.: Primary and secondary structure of 7-3 (K) RNA of Novikoff hepatoma. J. Biol. Chem. **259**, 12,265–12,270 (1984)
24. Telford, M.: Animal phylogeny. Cur. Biol. **16**, R981–R985 (2006)
25. Wassarman, D.A., Steitz, J.A.: Structural analyses of the 7SK ribonucleoprotein (RNP), the most abundant human small RNP of unknown function. Mol. Cell. Biol. **11**, 3432–3445 (1991)
26. Zieve, G., Penman, S.: Small RNA species of the HeLa cell: metabolism and subcellular localization. Cell **8**, 19–31 (1976)