Evolution of 7SK RNA and Its Protein Partners in Metazoa

Manja Marz,[1] Alexander Donath,*[1] Nina Verstraete,† Van Trung Nguyen,†*
Peter F. Stadler, ‡ § ‖ ¶ and Olivier Bensaude†*

*Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany; †CNRS UMR 8541 Ecole Normale Superieure, Paris, France; ‡Max-Planck Institute for Mathematics in the Sciences, Leipzig, Germany; §Fraunhofer Institut für Zelltherapie und Immunologie (IZI), Leipzig, Germany; ‖Department of Theoretical Chemistry, University of Vienna, Wien, Austria; and ¶Santa Fe Institute, Santa Fe, New Mexico

7SK RNA is a key player in the regulation of polymerase II transcription. 7SK RNA was considered as a highly conserved vertebrate innovation. The discovery of poorly conserved homologs in several insects and lophotrochozoans, however, implies a much earlier evolutionary origin. The mechanism of 7SK function requires interaction with the proteins HEXIM and La-related protein 7. Here, we present a comprehensive computational analysis of these two proteins in metazoa, and we extend the collection of 7SK RNAs by several additional candidates. In particular, we describe 7SK homologs in *Caenorhabditis* species. Furthermore, we derive an improved secondary structure model of 7SK RNA, which shows that the structure is quite well-conserved across animal phyla despite the extreme divergence at sequence level.

Introduction

Vertebrate 7SK small nuclear RNA (snRNA) is a highly abundant noncoding RNA (ncRNA) with a length of approximately 330 nt (Krüger and Benecke 1987; Murphy et al. 1987). It is involved in the regulation of the activity of the positive transcription elongation factor b (P-TEFb) (Peterlin and Price 2006). It mediates the inhibition of the general transcription elongation factor P-TEFb by the HEXIM1/2 proteins (also known as CLP1, MAQ1, and EDG1), thereby repressing transcript elongation by RNA polymerase II (Michels et al. 2003, 2004; Yik et al. 2003; Blazek et al. 2005; Egloff et al. 2006). A highly specific interaction with La-related protein 7 (LARP7, also known as PIP7S), on the other hand, regulates its stability (He et al. 2008; Krueger et al. 2008; Markert et al. 2008). 7SK RNA is capped at its $5'$ end by a highly specific methylase methylphosphate capping enzyme (MePCE), also known as bicoid-interacting protein 3 (BCDIN3) (Jeronimo et al. 2007).

The sequence of the 7SK snRNA is extremely well-conserved across jawed vertebrates. In contrast, the sequence of the lamprey 7SK is highly divergent (Gürsoy et al. 2000), and invertebrate 7SK RNAs were recently found only using specialized sophisticated homology search techniques (Gruber et al. 2008a,b). The latter study made extensive use of the fact that the 7SK genes feature a canonical class-3 pol-III promoter structure (Sürig et al. 1993). Despite considerable efforts, phylogenetic distribution and evolutionary age of 7SK RNA remain uncertain because no homologs have been found so far in basal metazoan lineages or in important invertebrate phyla such as Platyhelminthes and Nematoda.

Because 7SK RNA interacts specifically with HEXIM and LARP7, we survey here the phylogenetic distribution of these proteins to determine in which organisms we can also expect a 7SK gene. Because the primary interaction sites with HEXIM and LARP7 are among the few well-conserved features of the invertebrate 7SK snRNAs (Gruber et al. 2008a), we reevaluate and refine the secondary structure model of Wassarman and Steitz (1991). This in turn forms the basis for the detection of additional invertebrate 7SK RNAs.

Materials and Methods

Bioinformatics

Sequence Data

Accession number and sources of genomes are listed in the electronic supplementary material (www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-010).

Homology Search for Proteins

For HEXIM1/2, LARP7, and MePCE/BCDIN3, we used TBlastN (Altschul et al. 1990) with $E \leqslant 10^{-3}$ to search genomic DNA sequences starting from the human protein sequences. This search was complemented by a PsiBlast (Altschul et al. 1997) search using the NCBI Web interface. Nematode genome sequences were in addition searched with PsiTBlastN using the CLP1 protein of *Brugia malayi* as query (XP_001897213). Possible hits were conceptually translated and searched for introns with prosplign (http://www.ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html). Translated protein sequences were searched for domain annotations using the batch sequence search of Pfam (Finn et al. 2008) and aligned with ClustalW 2.0.9 (Larkin et al. 2007) and MUSCLE version 3.7 (Edgar 2004). A phylogenetic network analysis of HEXIM was performed using the NeighborNet algorithm (Bryant and Moulton 2004) with Hamming distances implemented in SplitsTree version 4.10 (Huson and Bryant 2006).

Homology Search for 7SK RNAs

Homology search was performed by Blast, GotohScan (Hertel et al. 2009), and Fragrep (Mosig et al. 2007). Based on our experience with these approaches, and the previously known 7SK snRNAs, we constructed a specialized automaton to recognize 7SK RNAs.

The automaton combines four separate rnabob (Eddy 1992–1996) searches of the target genome (fig. 1) and

[1] These authors contributed equally.

Key words: 7SK RNA, HEXIM, LARP7, MePCE, polymerase III transcription, noncoding RNA, secondary structure.

E-mail: manja@bioinf.uni-leipzig.de.

Mol. Biol. Evol. 26(12):2821–2830. 2009
doi:10.1093/molbev/msp198
Advance Access publication September 4, 2009

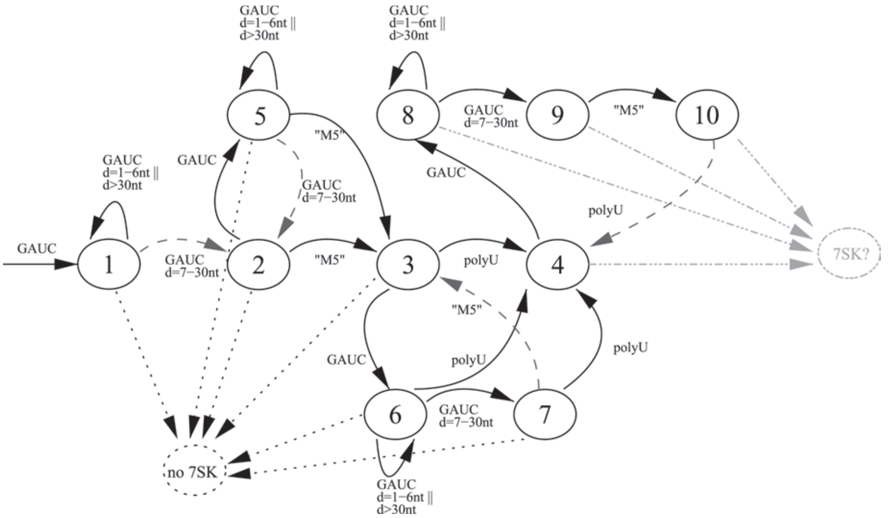


FIG. 1.—The "7SK-automaton". rnabob-hits for GAUC, "M5," and polyU within the whole genome or 500 nt downstream of potential polymerase III candidates are searched separately. The automaton evaluates the correct order and distances between these motifs and discards all entries if state 4 is not reached within a distance $d > 500$ nt (dotted transitions). Distance constraints are also enforced between the individual motifs and the last GAUC pair that has been encountered (dashed transitions). If all motifs are encountered in the correct order and with acceptable mutual distances, the candidate is assumed to be a 7SK candidate (dash-dotted transitions). The numbers indicate finite states of the automaton; arrows indicate transition of states.

requires some target-specific training that determines the stringency with which individual motifs are searched.

1. Promoter search. Promoter sequences were obtained by aligning the 100 nt upstream flanking sequences of pol-III transcripts (U3 small nucleolar RNA, snRNA U6, snRNA U6atac, RNAse mitochondrial RNA processing (MRP), and RNAse P). For the search in nematodes, for instance, we used the *Caenorhabditis remanei* proximal sequence element (PSE) motif GGCGGAACCCGnnnnnTGTCGG, allowing three mismatches, and searched the University of California Santa Cruz rhabditina alignment, obtaining 92 hits. The 500 nt downstream of these hits was extracted and passed to the next stage.
2. GATC search locates the highly conserved pattern GATC.
3. poly-T search locates stretches of five thymidines within 7 nt, which might constitute a termination signal.
4. Stem M5 search searches for a GC-rich stem–loop that could constitute stem M5.

The hits obtained in steps 2–4 are sorted by location and then filtered with respect to distance constraints and secondary structure constraints as summarized in figure 1. In particular, a stem–loop structure is required not more than 20 nt upstream of the terminator, and two of the GATC motifs must form an additional hairpin.

In order to evaluate the candidates, we attempted to incorporate them into the sequence/structure alignment described below. In addition, the promoter regions were compared with those of other known pol-III transcripts of the same organism, in particular, U3, U6, U6atac, RNase MRP, and RNase P RNAs.

Structural Alignments of 7SK snRNAs

Structural alignments of 7SK snRNAs were constructed manually in the Emacs editor using the RNA-specific ralee mode (Griffiths-Jones 2005). Pairing possibilities were evaluated using the toolkit provided by the Vienna RNA package (Hofacker 2004). In particular, we employed RNAsubopt to determine energetically plausible alternative foldings, RNAduplex to find possible intramolecular pairings, and RNAalifold to compute consensus structures. For closely related organisms and local regions of 7SK gene, we furthermore used ClustalW 2.0.9 to obtain sequence alignments and locarnate (Otto et al. 2008) to construct structural alignments.

Molecular Biology

Polymerase Chain Reaction Amplification

A 161 nt DNA fragment corresponding to the predicted 7SK homolog was amplified by polymerase chain reaction (PCR) from *Caenorhabditis elegans* genomic DNA (sense primer: tatgatatcTTCAGTATGGGTCAATCTC; reverse primer: tatagatatcAAAAGAGTCTTATGTTTCC). The resulting PCR amplicon was further used for nick end translation. Its DNA sequence was found to be identical to the expected sequence in the worm genome.

Northern Blot

RNAs were electrophoresed in 6% polyacrylamide/urea gels, transferred onto Hybond N (GE Healthcare), and hybridized to nick end–translated 32P-labeled DNA probes in Church buffer. The membrane was washed in 0.1X SSC buffer (0.015 M NaCl, 0.015 M sodium citrate) at 65°C and autoradiographed. The membrane was hybridized first with the ce7SK probe then autoradiographed a first time and rehybridized with the U4 probe then autoradiographed again.

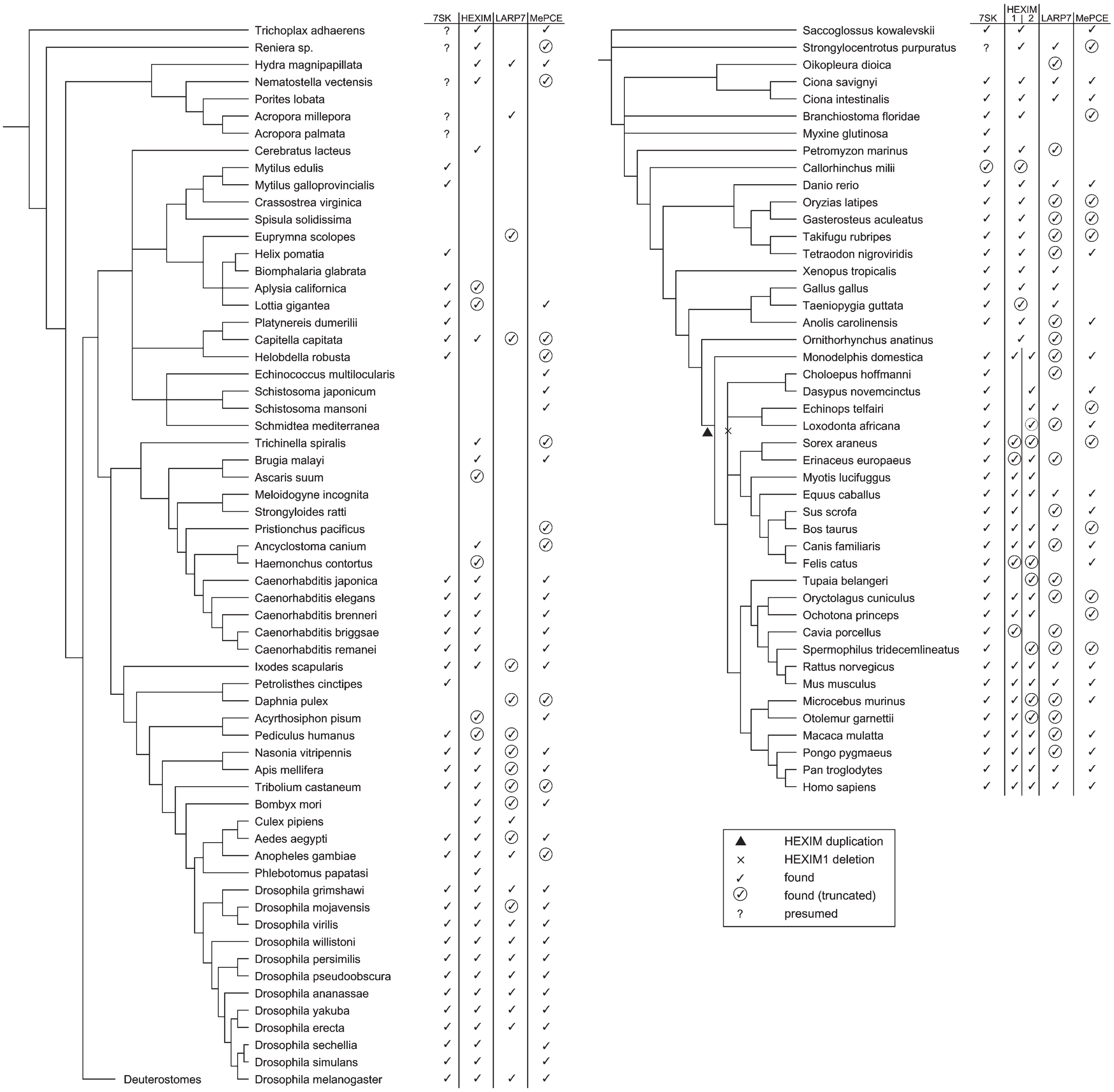


FIG. 2.—Distribution of HEXIM1/2, LARP7, MePCE/BCDIN3, and 7SK RNA. Complete proteins or 7SK RNA is indicated by a tick. Cases in which only a truncated protein sequence was found but both alignments and the HMM analysis suggested a true candidate are marked with a circled tick. 7SK candidate sequences for which not all features are present are indicated by a question mark. The black filled triangle indicates the HEXIM duplication event. The secondary HEXIM1 loss for Afrotheria and Xenartha is marked with a cross within the taxonomic tree. The underlying tree follows a combination of the NCBI taxonomy and Blaxter et al. (1998), Mitreva et al. (2005), Holterman et al. (2006), Webster et al. (2006), and Dunn et al. (2008).

Results

Phylogenetic Distribution of HEXIM

Homologs of HEXIM were found across the metazoan tree, using known HEXIM1 protein sequences and TblastN. In particular, we identified clear homologs in the poriferan *Reniera* sp., the placozoan *Trichoplax adhaerens*, and the cnidarians *Nematostella vectensis* and *Hydra magnapapillata* implying that HEXIM was present in the metazoan ancestor. On the other hand, no homologs were detected in fungi, plants, and the choanoflagellate *Monosiga brevicollis*, suggesting with the current methods that HEXIM is a bona fide animal innovation (fig. 2). Full alignments are available here: www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-010.

Interestingly, the WKPY motif of HEXIM's 7SK-binding domain is modified to YPXWK in flies. Overall the motif is more degenerate in Protostomia compared with Deuterostomia but remains recognizable.

The only metazoan phylum with fully sequenced genomes in which we did not succeed in finding HEXIM are the notoriously fast-evolving Platyhelminthes (Philippe et al. 2005; Lartillot et al. 2007). In nematodes, HEXIM was easily recognized in basal lineages such as *Trichinella*

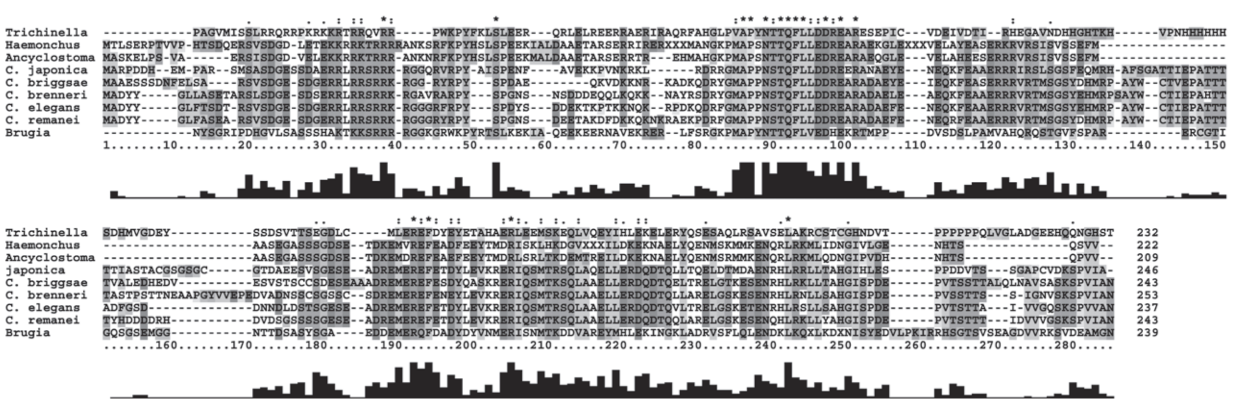


FIG. 3.—Alignment of the 7SK associated protein HEXIM in nematodes. Conserved boxes known to be involved in RNA–protein or protein–protein interaction are located at alignment positions 36, 43–50 (NLS), and around position 90 (P-TEFb-binding site).

spiralis and *B. malayi*. Using TblastN, no homologs were found in Chromadorea, including *C. elegans*. A PsiBlast search starting from the *Brugia* HEXIM (XP_001897213), however, detected a single unannotated putative *C. elegans* protein Y39E4B.6 (NP_499710). A multiple sequence alignment (see www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-010) strongly suggests that Y39E4B.6 is indeed the *C. elegans* HEXIM/CLP1 homolog. We then used TblastN to search for homologs in the other *Caenorhabditis* species and used the results to construct a position-specific scoring matrix (Altschul et al. 1997) for a psiTBlastN search of additional nematode genomes. An alignment is shown in figure 3. We remark that the NCBI EST database contains several HEXIM homologs from an additional nematode (see www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-010).

Eutheria are well known to carry two HEXIM paralogs (Byers et al. 2005). Marsupials (*Monodelphis domestica*) also have clearly recognizable orthologs of both HEXIM1 and HEXIM2. On the other hand, there is only a single copy of HEXIM in the genome of platypus. (*Ornithorhynchus anatinus*). In several eutherian genomes, only one of the two HEXIM paralogs was found (fig. 2). This could be an artifact caused by the low coverage and incomplete assemblies of the genomes in question. It is surprising, however, that neither Afrotheria (*Echinops telfairi*, *Loxodonta africana*) nor Xenarthra (*Dasypus novemcinctus*, *Choloepus hoffmanni*) has a copy of HEXIM1. We conclude that HEXIM was duplicated before the divergence of Metatheria and Eutheria, with a possible secondary loss of HEXIM1 in some eutherian clades. Because the phylogenetic relationships of the major Eutherian groups are under intense discussion (Nishihara et al. 2009), it remains unclear whether the loss in Afrotheria and Xenarthra was independent or whether these are sister groups whose ancestor already lost HEXIM1.

HEXIM1 and HEXIM2 are always located very close to each other (from ∼10.000 nt in *Canis familiaris* up to ∼26.000 nt in *Myotis lucifugus*) on the same chromosome (where sequence assembly allows such observations). Comparing mammalian HEXIM1/2 proteins to HEXIMs of birds, frogs, and fish (Gnathostomes), a much higher similarity of HEXIM to HEXIM1 is apparently observed (fig. 4). HEXIM2, as well as protostome HEXIM/CLP1, contains several introns (conserved at least from mice to humans; Michels et al. 2003). In contrast, there are no introns in the HEXIM1 gene, suggesting that HEXIM1 derived from reverse transcription of HEXIM2. Surprisingly, the intron-less copy evolved much more conservatively than the ancestral intron–containing template. The fact that all mammals show this pattern indicates that a functional separation of the two HEXIM variants occurred soon after the retroposition event before the radiation of the crown group mammals.

A comparison of all metazoan HEXIM proteins shows the high conservation of three motifs: Motif 1 consists of a basic domain that has been identified as the 7SK-binding domain in human HEXIMs (Michels et al. 2004; Yik et al. 2004) and overlaps a nuclear localization signal (NLS) (Ouchida et al. 2003; fig. 5*a*). The latter contains a C-terminal WKPY that is interestingly modified to YPXWK in flies (fig. 5*b*). Motif 2 (APYNTTQFLM) is conserved in all metazoa (fig. 5*c*). It has been shown that replacement of tyrosine or threonine in the PYNT motif in the human HEXIM1 suppresses its capacity of binding P-TEFb (Michels et al. 2004; Byers et al. 2005). Motif 3 overlaps helix $\alpha 2$ that is known to be involved in HEXIM dimerization and cyclin T binding (Dames et al. 2007; fig. 5*d*).

Phylogenetic Distribution of LARP7

LARP7 has similarities to autoantigen La (genuine La protein, Sjogren syndrome antigen B). Among the various La-related protein families, the LARP7 family is characterized by its domain arrangement (Bousquet-Antonelli and Deragon 2009). LARP7 proteins contain a La-domain (PFAM ***PF05383***) at their N-terminus followed by a well-conserved RNA recognition motif of type 1 (RRM1, PFAM ***PF00076***) and an atypical RRM3 domain (PFAM ***PF08777***), which is much less well conserved. A TblastN search for LARP7 revealed its existence in all major metazoan phyla, including basal lineages such as porifera, placozoa, and cnidaria. The LARP7 protein sequences of protostomes and deuterostomes, respectively, are clearly

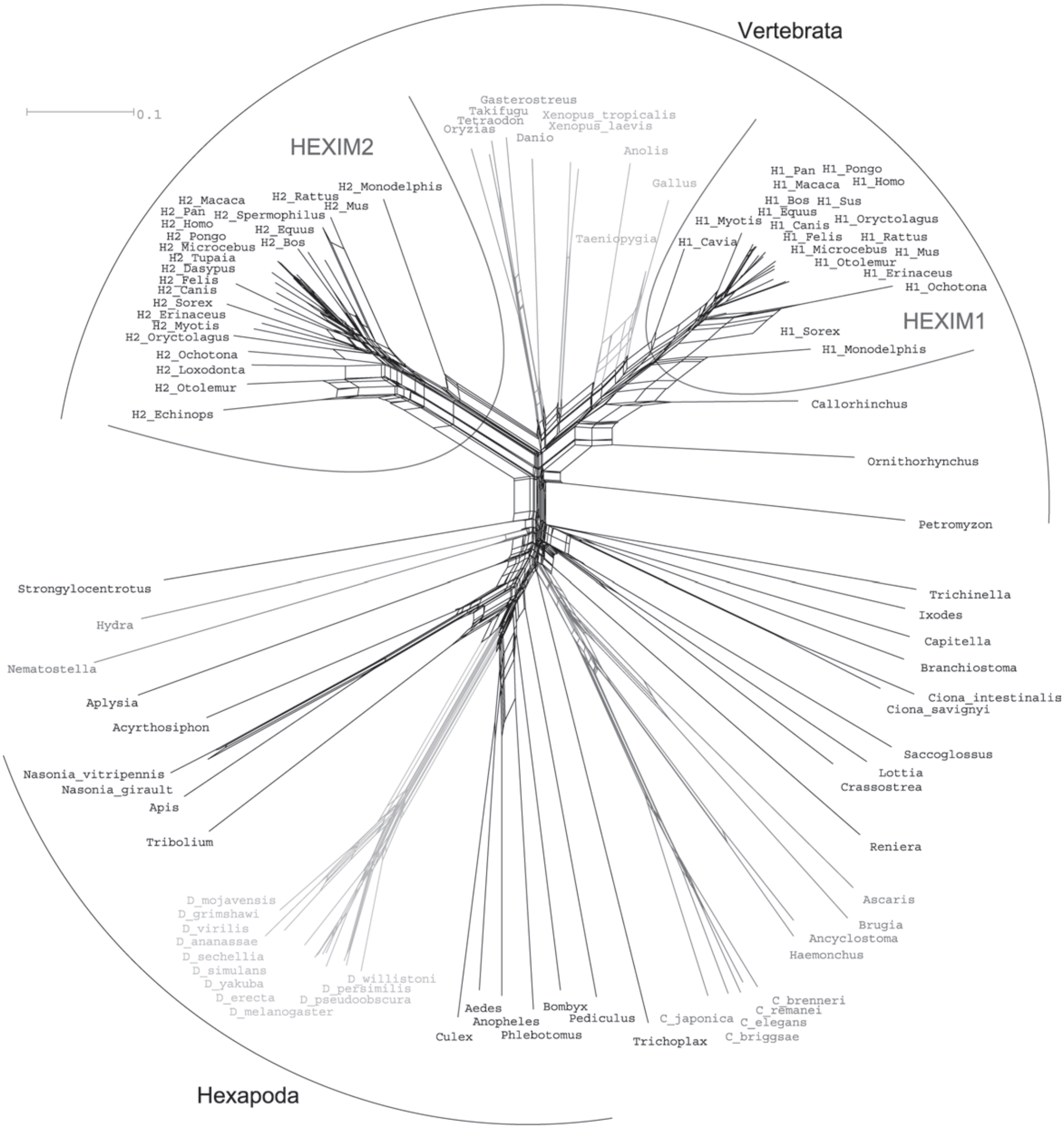


FIG. 4.—Phylogenetic network of all metazoan HEXIMs computed by the NeighborNet algorithm with Hamming distances as implemented in SplitsTree (version 4.10) (Huson and Bryant 2006). The paralogous HEXIM1 and HEXIM2 groups, as well as the close relationship of the protostomia HEXIMs, are very well supported. The split illustrating the closer relationship of HEXIM of fishes (magenta), amphibians (light blue), and birds (orange) to HEXIM1 of mammals is clearly identifiable. Additional well-supported groups are those of nematodes (red), drosophilids (green), and cnidarians (dark blue). Due to their very basal position, all other sequences cannot be resolved any further. See supplement for ClustalW source alignment and a higher resolution of this figure.

distinguishable. Within each group, they can be aligned unambiguously; for details see www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-010.

Unambiguous LARP7 homologs were found in species in which we also found 7SK and/or HEXIM (fig. 2), including the Cnidaria *H. magnapapillata* and *Acropora millepora*. Except for the previously known La domain proteins, we were not able to identify an LARP7 homolog in nematodes. In several lophotrochozoans, we did not find a LARP7 homolog. This can presumably be attributed to the incomplete status of the genome assemblies.

An alignment of metazoan LARP7 proteins is provided here: www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-010.

Phylogenetic Distribution of MePCE/BCDIN3

The MePCE has been described as a possibly 7SK-specific capping enzyme (Jeronimo et al. 2007). It is homologous to BCDIN3 of *Drosophila*, which in turn has homologs in both plants (e.g., *Arabidopsis thaliana*) and fungi (*Schizosaccharomyces pombe* and *Laccaria bicolor*, ***XP_001879607***, but not *Saccharomyces cerevisae*).

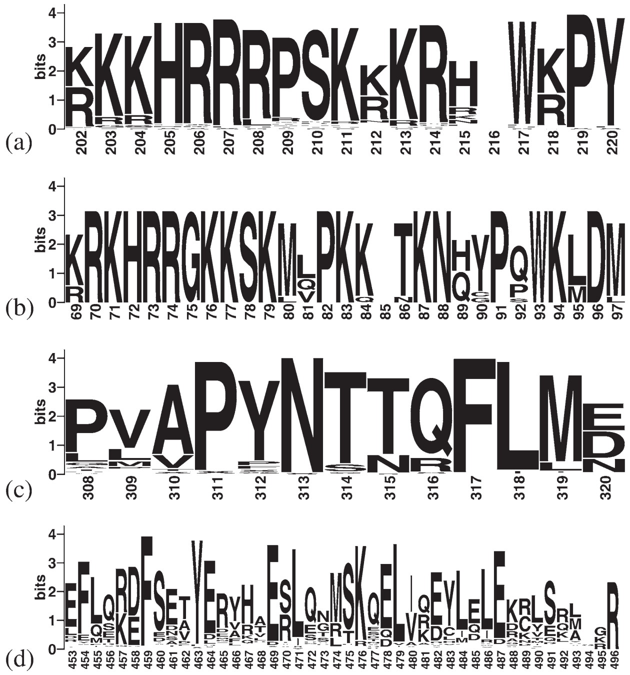


FIG. 5.—Consensus sequence logo (Crooks et al. 2004) of motifs 1–3 of all metazoan HEXIM/CLP1 sequences. (*a*) The 7SK-binding domain of deuterostomes that overlaps an NLS. (*b*) The WKPY within the NLS is rearranged to YPXWK in flies. (*c*) The logo shows the sequence motif that contains the domain PYNT for which has been shown that replacement of tyrosine or threonine in the motif of the human HEXIM1 suppresses its capacity of binding P-TEFb. (*d*) Motif 3 overlaps helix $\alpha2$ that is known to be involved in HEXIM dimerization and cyclin T binding. All alignments were created using ClustalW.

For completeness, we recorded MePCE genes across metazoa (fig. 2). The widespread phylogenetic distribution, however, suggests that MePCE/BCDIN3 also has other functions and thus may not be a good indicator for the presence of the 7SK system.

Revised Secondary Structure Model of 7SK RNA

A complete alignment of all 79 7SK sequences, including the five nematode homologs, can be found here: www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-010; unclear candidates are excluded. The expanded collection of sequences provides sufficient information for the construction of a global multiple sequence alignment. In contrast, previous studies (Gruber et al. 2008a,b) were content with local alignments of the best-conserved regions. Based on this alignment, a much more comprehensive consensus structure model can be derived (fig. 6).

A comparison of the structure proposed for the human 7SK RNA based on chemical probing (Wassarman and Steitz 1991) shows that most of our structure model is consistent with the previous proposal. There are, however, several novel features that provide new insights into the function of 7SK RNA. Most parts of the stems M3, M5, M7, and M8 were described previously and correspond to the stems 1, 3, 5, and 6 of Wassarman and Steitz (1991) (fig. 6). Our reevaluation of the invertebrate data demonstrates that these stems are conserved and can be identified in all organisms.

1. Stem M1 is the best-conserved feature of 7SK RNAs. It is recognizable in all known homologs (Gruber et al. 2008a,b). Corresponding to stem 1 of Wassarman and Steitz (1991), it contains the HEXIM-binding site, an absolutely conserved helical region with the sequence GAUC:GAUC.
2. The additional stem M4 is highly conserved at the structural level in all organisms except nematodes, even though it does not exhibit recognizable sequence similarities at phylum level (fig. 8).
3. Drosophilids have an expansion domain between M5 and M4, which forms a stem–loop structure covering about 90 nt. They also have an extended loop M4.
4. M5, corresponding to stem 3 of Wassarman and Steitz (1991), is not only conserved in its structure but also in its sequence (see fig. 7), which contains the motif CGNNGC pairing with GCNNCG in all known 7SK RNAs. The M5 stem is slightly shorter in deuterostomes compared with other metazoa.
5. Most species have an additional stem, M6, located between M4/M5 and M7. It is missing, however, in many insects (drosophilids, *Tribolium*, and *Pediculus*) and in the two *Ciona* species.
6. Stem M8, corresponding to stem 6 of Wassarman and Steitz (1991), is also very well conserved. Therefore, M8 might harbor the LARP7-binding site.
7. The most interesting part of the structure is region M2 with a length of 15–25 nt, which is located between stems M1 and M3. Surprisingly, it can form three distinct structural alternatives in all known cases, as shown in figure 6 and the left-hand side panel of figure 9.

   M2a It can form a local hairpin. This local hairpin is much smaller in vertebrates.

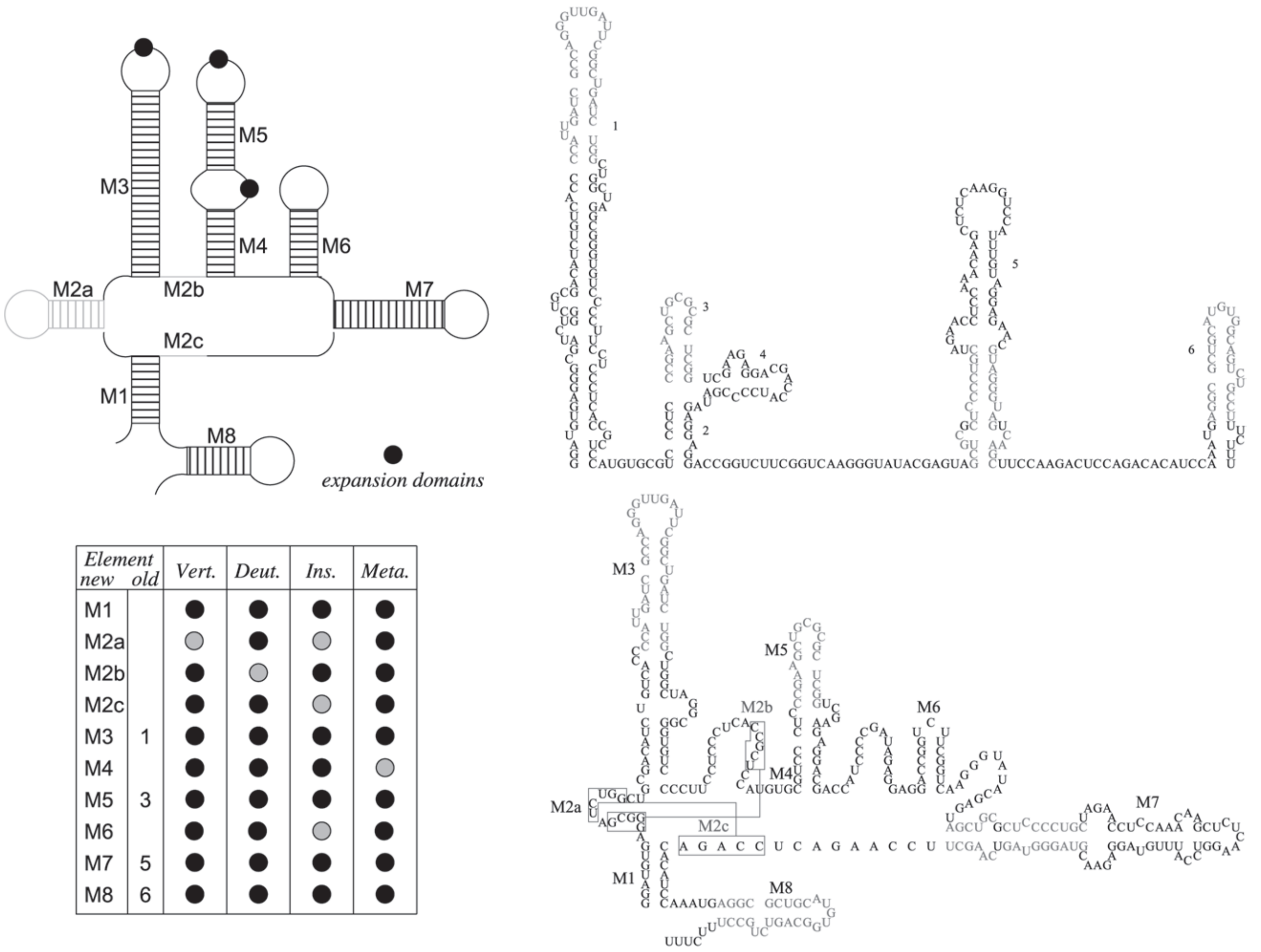
   M2b It can bind downstream of M3, as published previously by Gruber et al. (2008a), resulting in an extension of stem M3. *Ciona* is the only case in which this structural alternative seems to be absent.

   M2c It can bind upstream of M1, resulting in an extension of M1.

   The conservation of this flexible arrangement suggests that refolding the M2 region between the three structural alternatives is part of the core functionality of 7SK, implying that 7SK RNA might be an RNA switch.

Homology Search for 7SK snRNAs

Due to the high sequence conservation across jawed vertebrates, the 7SK genes of newly sequenced genomes such as *Tupaia belangeri*, *Equus caballus*, *Tribolium castaneum*, and *Acyrthosiphon pisum* were easily retrieved by Blast. In addition, several partial 7SK sequences were identified in the NCBI sequences databases using the interactive Blast facility: *Platynereis sp.* ***CT030666*** (EMBL), *Mytilus edulis* ***AM880723***, *Mytilus galloprovincialis* ***EH663179.1***, and *Petrolisthes cinctipes* ***CAYF7296.g3***.



Element new old		Vert.	Deut.	Ins.	Meta.
M1		●	●	●	●
M2a		○	●	○	●
M2b		●	○	●	●
M2c		●	●	○	●
M3	1	●	●	●	●
M4		●	●	●	○
M5	3	●	●	●	●
M6		●	●	○	●
M7	5	●	●	●	●
M8	6	●	●	●	●

FIG. 6.—Revised secondary structure model of 7SK RNA. Left panel: Scheme of the 7SK RNA secondary structure and its revised nomenclature. Stems are labeled M1 to M8. M3 refers to hairpin 1, M5 to stem 3, M7 to stem 5, and M8 to the last stem of Wassarman and Steitz (1991). M1 was published recently by Gruber et al. (2008b). M6 is not present in *Drosophila*; however, this species shows an expansion domain between M5 and M4. Stem M2 has one of three possibilities to base pair. M2a: M2 builts a hairpin as drawn in the picture, which is absent in *Drosophila* sp. M2b: M2 binds downstream of M3, constructing an extended M3 stem, which is absent in *Ciona*. Therefore, in *Ciona* M2 binds upstream to M1 (M2c) extremely well. This extension form of M1 is absent in some insects. "old" refers to Wassarman and Steitz, whereas "new" is referring to this work. Vert, vertebrates; Deut, basal deuterostomes; Ins, insects; Meta, other metazoans. Black circles denote the presence of an element in all species. Gray circles denote the presence of an element in the majority of species within a group. Right panel: Revision of the human 7SK secondary structure. Top: proposal of Wassarman and Steitz (1991). Below: revised model. Elements that coincide between the structures are highlighted. Stem 5 of Wassarman and Steitz and M7 show similar RNAfold energies: −20.32 kcal/mol (stem 5) versus −22.00 kcal/mol kcal/mol (M7).

Nematoda

Using the promoter-based approach, we obtained a hit in *Caenorhabditis briggsae* that warranted detailed analysis. A simple Blast search then showed that the sequence is well conserved across the genus *Caenorhabditis*. Sequence and secondary structure of *C. elegans* are shown in the left-hand side panel of figure 9. Although it is significantly shorter than other 7SK RNAs, it bears the hallmarks of a true 7SK homolog: 1) M1 is structurally highly conserved. 2) It can form all three alternative helices M2a/M2b/M2c. 3) The stem M3 contains the highly conserved GAUC sequence. 4) M5 is usually a GC-rich stem. 5) The poly-T is indicative of a pol-III terminator.

Another Blast search showed that nearly identical copies of the 7SK candidates reside at two distinct loci in the genomes of *C. elegans*, *C. briggsae*, *C. remanei*, and *Caenorhabditis brenneri*. Locus A (7SK-A) is linked to the hypothetical protein ***T26A8.4***, and locus B (7SK-B) is associated with the gene str-170 (***T08B6.7***). Both are annotated in WormBase Genome Browser by RNAz (***RNAz-514034*** and ***RNAz-513873***). The current genome assemblies of the *C. remanei* and *C. brenneri* genomes contain two or three copies of both loci, which are most likely technical artifacts. *Caenorhabditis japonica* only exhibits two closely spaced copies of the 7SK RNA gene at locus A. Despite the substantial evolutionary distances between the *Caenorhabditis* species, the 7SK RNA genes at the two loci do not form two clearly distinguishable paralog groups, whereas the flanking sequences clearly reflect a duplication event before the last common ancestor of *C. elegans*, *C. briggsae*, *C. remanei*, and *C. brenneri*, see www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-010. This indicates that the two 7SK

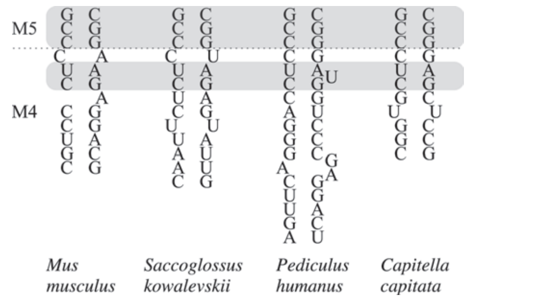


FIG. 7.—Stem M4 is structurally conserved across all metazoan clades except Nematoda, even though the sequence is highly variable.

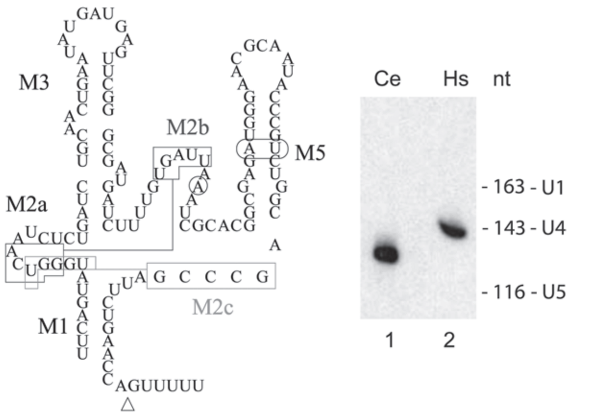


FIG. 9.—*Caenorhabditis* 7SK RNA. Left-hand side: The secondary structure model of the *Caenorhabditis elegans* 7SK-B RNA candidate. In contrast, the 7SK-A RNA candidate shows an insert (AGUUGCCUA-GAAAUAUUUUGGAAACAUAAGACUC) that folds into a hairpin at the position of the triangle. Both fold with the typical highly conserved elements: GAUC (M3) and the GC-rich stem M5. These elements are likely to interact with HEXIM and substrate (DNA/RNA). The 5′-part binding with the 3′-region (M1) is conserved among all known 7SK. A characteristic poly-T pol III termination signal is present at the expected position. In addition, the *Caenorhabditis* sequences can all form the three structural alternatives M2a/M2b/M2c. Mutations between 7SK-A and 7SK-B are indicated with circles. Right-hand side: 10 μg of total RNAs from worms (lane 1) and human cells (lane 2), respectively, were analyzed by northern blot. The figure shows the autoradiogram following sequential hybridation of the same membrane with *C. elegans* 7SK-A and human U4 snRNA probes. The *C. elegans* 7SK-A probe hybridized to a single band in lane 1. Rehybridization with U1 and U5 human snRNA probes lit up additional single bands in lane 2 that were used for sizing.

loci are subject to some form of concerted evolution, a phenomenon that is not uncommon among snRNAs (Marz et al. 2008).

Uncertain Invertebrate Candidates

The same approach that was successful for nematodes was also used to search for all invertebrates species for which consensus pol-III motifs could be derived from known ncRNA loci. Several candidates were found, which however lack at least one of the seven features described above. The promoter of *T. adhaerens* is very clearly recognizable with the upstream regions of RNase MRP, RNase P, and snRNA U6 and U6atac. Only 65 regions distributed over the whole genome were found with at most three point mutations. These candidates were observed in detail for 7SK features. The best candidates are compiled at www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-010 for *Trichoplax*, *Reniera sp.*, *N. vectensis*, and *Strongylocentrotus purpuratus*. If the latter candidate is a real 7SK, it diverged drastically from other deuterostomes. M1 would have an unexpectedly low MFE, M3 changed in sequence and structure, and M4–M7 are not conserved compared with other deuterostomes. On the other hand, the ultraconserved GAUC:GAUC base pairing is present and a typical polymerase III terminator (poly-T) directly after a hairpin with the proper length of 8 nt was found.

A 7SK RNA Homolog Is Expressed in *C. elegans*

To establish that the putative *C. elegans* 7SK RNA homolog corresponds to an expressed RNA species, total worm RNA was analyzed by northern blot (fig. 9, right-hand side panel, lane 1). Total human RNA was loaded in lane 2 as a hybridation control and to serve as a molecular weight scale. Human snRNA probes only detected bands in lane 2 that had been loaded with human total RNA; the U4 autoradiogram is the only one shown on figure 9. The putative 7SK-A and 7SK-B genes were predicted to generate two transcript of distinct sizes, 142 nt and 130 nt, respectively. However, although 7SK-A DNA was used for probing, a single band matching 130 nt was detected in lane 1. Hence, 7SK-B is likely to be expressed at much higher levels than 7SK-A. As both U4 and 7SK probes had similar specific activities and the same amount of RNA was loaded in lanes 1 and 2, we conclude that their expression levels are comparable. As U4 snRNA is an abundant RNA, the putative 7SK homolog in *C. elegans* is a highly expressed ncRNA species.

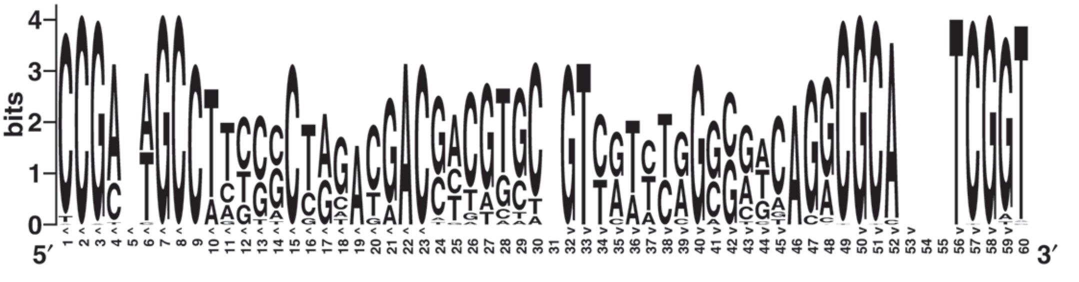


FIG. 8.—Consensus of the M5 hairpin. Left-hand side: Consensus sequence created by Weblogo (Crooks et al. 2004) and expanded by corresponding base pairings. Nucleotide 1–8 and 50–59 conserved for all 74 known 7SK sequences including lophotrochozoans, arthropods, and deuterostomes. Base pairing from nucleotide 10–45 is observed in drosophilids only. Right-hand side: Secondary structure consensus. The brackets indicates the stem extension observed for insects.

Discussion

The regulation of transcriptional elongation by 7SK involves at least two protein partners that interact with the 7SK snRNA in a highly specific fashion: HEXIM and the La-protein LARP7. According to our data, both proteins are metazoan innovations: LARP7 homologs are found in most metazoan clades, with the exception of the two most basal lineages (sponges and placozoans) and two rapidly evolving protostome phyla (Nematoda and Platyhelminthes). On the other hand, we did not find clear LARP7 homologs outside the metazoa. The two La proteins p43 and p65 in ciliates show similarities with LARP7 (Bousquet-Antonelli and Deragon 2009) but probably evolved independently. The picture is even clearer in the case of HEXIM. With the exception of Platyhelminthes, unambiguous homologs, in most cases as single copy, are found in animals and only in animals. In contrast, MePCE/BCDIN3 has a much broader phylogenetic distribution, indicating that it has other important functions beyond its interaction with 7SK RNA.

Our refined searches were successful in identifying previously undescribed HEXIM and 7SK RNA homologs in nematodes. In *C. elegans*, we verified the expression of an ncRNA species with a length of only ∼130 nt that still shows all hallmarks of the much larger 7SK RNAs in vertebrates (∼320 nt) and in drosophilids (>400 nt). Interestingly, *Caenorhabditis* species have this miniature 7SK RNAs while at the same time lacking an LARP7 ortholog. Further studies will be necessary to determine whether another La protein takes on this role. Interestingly, drosophilids have both the largest 7SK RNA and the largest LARP7 homologs, which are extended by 140AA C-terminal domain that is absent in other protostomes.

In Platyhelminthes, none of the components of the 7SK-based regulation system have been identified, suggesting that the entire mechanism might have been lost in this phylum. Due to the rapid evolution of Platyhelminthes and the incompleteness of the genome assemblies, however, it is possible that HEXIM, LARP7, and 7SK RNA are present in the genomes of these species but have diverged beyond recognition.

The drastic length variations of 7SK RNA across the metazoa is reminiscent of the same phenomenon for telomerase RNA (Xie et al. 2008). The comparison of, in particular, the invertebrate 7SK RNA structures shows that this RNA has a structural core which shows rapid sequence turnover, whereas the intervening parts rapidly expand or contract in different lineages. This type of behavior appears to be a common theme for many ncRNA families. For the 7SK RNA, it has been impossible to identify this core structure until a diverse collection of invertebrate homologs was available. The core structure contains known interaction domains (such as the HEXIM-binding site in stem M1) and a few unexplained highly conserved features (such as stem M4). The ambiguity of stem M2 is particularly interesting. Given the few sequence constraints, it is highly unlikely that evolution has retained the three alternatives by chance. We suggest therefore that structural transitions between the three states are involved in 7SK RNA function.

Taken together, our data provide compelling evidence that 7SK-mediated regulation of transcriptional elongation is a metazoan innovation and is (nearly) ubiquitously employed in the animal clade. Given that eubacteria have 6S RNA as an ubiquitous RNA regulator of transcription (Trotochaud and Wassarman 2005), this raises the question whether different small RNA species play a role in transcription control also in the other eukaryotic kingdoms.

Acknowledgments

We thank Dr Jean Louis Bessereau for crude *C. elegans* nucleic acids and Anne Catherine Dock-Bregeon for critical reading of the manuscript. This work was supported in part by Landesstipendium Sachsen, DFG SPP-1174 “Deep Metazoan Phylogeny”, Association pour la Recherche sur le Cancer (ARC 8/2/1010), Agence Nationale pour la Recherche (ANR 06-BLAN-0072). Many thanks to Petra Pregel and Jens Steuck for making work so much easier.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Altschul SF, Madden TL, Schffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Blaxter ML, De Ley P, Garey JR, et al. (12 co-authors). 1998. A molecular evolutionary framework for the phylum Nematoda. Nature 392:71–75.

Blazek D, Barboric M, Kohoutek J, Oven I, Peterlin BM. 2005. Oligomerization of HEXIM1 via 7SK snRNA and coiled-coil region directs the inhibition of P-TEFb. Nucleic Acids Res. 33:7000–7010.

Bousquet-Antonelli C, Deragon JM. 2009. A comprehensive analysis of the La-motif protein superfamily. RNA 15:750–764.

Bryant D, Moulton V. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 21:255–265.

Byers SA, Price JP, Cooper JJ, Li Q, Price DH. 2005. HEXIM2, a HEXIM1-related protein, regulates positive transcription elongation factor b through association with 7SK. J Biol Chem. 280:16360–16377.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188–1190.

Dames SA, Schönichen A, Schulte A, Barboric M, Peterlin BM, Grzesiek S, Geyer M. 2007. Structure of the Cyclin T binding domain of Hexim1 and molecular basis for its recognition of P-TEFb. Proc Natl Acad Sci USA. 104:14312–14317.

Dunn CW, Hejnol A, Matus DQ, et al. (18 co-authors). 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.

Eddy S. 1992–1996. RNABOB: a program to search for RNA secondary structure motifs in sequence databases. Available from: http://selab.janelia.org/software.html.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Egloff S, Van Herreweghe E, Kiss T. 2006. Regulation of polymerase II transcription by 7SK snRNA: two distinct RNA elements direct P-TEFb and HEXIM1 binding. Mol Cell Biol. 26:630–642.

Finn RD, Tate J, Mistry J, et al. (11 co-authors). 2008. The Pfam protein families database. Nucleic Acids Res. 36:281–288.

Griffiths-Jones S. 2005. RALEE—RNA ALignment editor in Emacs. Bioinformatics 21:257–259.

Gruber AR, Kilgus C, Mosig A, Hofacker IL, Hennig W, Stadler PF. 2008a. Arthropod 7SK RNA. Mol Biol Evol. 25:1923–1930.

Gruber AR, Koper-Emde D, Marz M, Tafer H, Bernhart S, Obernosterer G, Mosig A, Hofacker IL, Stadler PF, Benecke BJ. 2008b. Invertebrate 7SK snRNAs. J Mol Evol. 66:107–115.

Gürsoy HC, Koper D, Benecke BJ. 2000. The vertebrate 7S K RNA separates hagfish (*Myxine glutinosa*) and lamprey (*Lampetra fluviatilis*). J Mol Evol. 50:456–464.

He N, Jahchan NS, Hong E, Li Q, Bayfield MA, Maraia RJ, Luo K, Zhou Q. 2008. A La-related protein modulates 7SK snRNP integrity to suppress P-TEFb-dependent transcriptional elongation and tumorigenesis. Mol Cell. 29:588–599.

Hertel J, de Jong D, Marz M, Rose D, Tafer H, Tanzer A, Schierwater B, Stadler PF. 2009. Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. Nucleic Acids Res. 37:1602–1615.

Hofacker IL. 2004. RNA secondary structure analysis using the Vienna RNA package. Curr Protoc Bioinformatics. 26:12.2.1–12.2.16.

Holterman M, van der Wurff A, van den Elsen S, van Megen H, Bongers T, Holovachov O, Bakker J, Helder J. 2006. Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. Mol Biol Evol. 23:1792–1800.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Jeronimo C, Forget D, Bouchard A, et al. (16 co-authors). 2007. Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. Mol Cell. 27:262–274.

Krueger BJ, Jeronimo C, Roy BB, et al. (12 co-authors). 2008. LARP7 is a stable component of the 7SK snRNP while P-TEFb, HEXIM1 and hnRNP A1 are reversibly associated. Nucleic Acids Res. 36:2219–2229.

Krüger W, Benecke BJ. 1987. Structural and functional analysis of a human 7SK RNA gene. J Mol Biol. 195:31–41.

Larkin MA, Blackshields G, Brown NP, et al. (13 co-authors). 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7:S4.

Markert A, Grimm M, Martinez J, Wiesner J, Meyerhans A, Meyuhas O, Sickmann A, Fischer U. 2008. The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes. EMBO Rep. 9:569–575.

Marz M, Kirsten T, Stadler PF. 2008. Evolution of spliceosomal snRNA genes in metazoan animals. J Mol Evol. 67:594–607.

Michels AA, Fraldi Q, Li A, Adamson TE, et al. (11 co-authors). 2004. Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor. EMBO J. 23: 2608–2619.

Michels AA, Nguyen VT, Fraldi A, Labas V, Edwards M, Bonnet F, Lania L, Bensaude O. 2003. MAQ1 and 7SK RNA interact with CDK9/cyclin T complexes in a transcription-dependent manner. Mol Cell Biol. 23:4859–4869.

Mitreva M, Blaxter ML, Bird DM, McCarter JP. 2005. Comparative genomics of nematodes. Trends Genet. 21:573–581.

Mosig A, Chen JL, Stadler PF. 2007. Homology search with fragmented nucleic acid sequence patterns. In: Giancarlo R, Hannenhalli S, editors. Algorithms in bioinformatics (WABI 2007). Lecture Notes in Computer Science. Vol. 4645. Berlin, Heidelberg: Springer Verlag. p. 335–345.

Murphy S, Di Liegro C, Melli M. 1987. 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.

Nishihara H, Maruyama S, Okada N. 2009. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proc Natl Acad Sci USA. 106:5235–5240.

Otto W, Will S, Backofen R. 2008. Structural local multiple alignment of RNA. In: Beyer A, Schroeder M, editors. Proceedings of the German conference on bioinformatics, (GCB 2008). Lecture Notes in Informatics. Vol. 136. GI. p. 178–187.

Ouchida R, Kusuhara M, Shimizu N, Hisada T, Makino Y, Morimoto C, Handa H, Ohsuzu F, Tanaka H. 2003. Suppression of NF-kappaB-dependent gene expression by a hexamethylene bisacetamide-inducible protein HEXIM1 in human vascular smooth muscle cells. Genes Cells. 8: 95–107.

Peterlin BM, Price DH. 2006. Controlling the elongation phase of transcription with P-TEFb. Mol Cell. 23:297–305.

Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol. 22:1246–1253.

Sürig D, Bredow S, Benecke BJ. 1993. The seemingly identical 7SK and U6 core promoters depend on different transcription factor complexes. Gene Expr. 3:175–185.

Trotochaud AE, Wassarman KM. 2005. A highly conserved 6S RNA structure is required for regulation of transcription. Nat Struct Mol Biol. 12:313–319.

Wassarman DA, Steitz JA. 1991. Structural analyses of the 7SK ribonucleoprotein (RNP), the most abundant human small RNP of unknown function. Mol Cell Biol. 11:3432–3445.

Webster BL, Southgate VR, Littlewood DT. 2006. A revision of the interrelationships of Schistosoma including the recently described Schistosoma guineensis. Int J Parasitol. 36:947–955.

Xie M, Mosig A, Qi X, Li Y, Stadler PF, Chen JJL. 2008. Size variation and structural conservation of vertebrate telomerase RNA. J Biol Chem. 283:2049–2059.

Yik JH, Chen R, Nishimura R, Jennings JL, Link AJ, Zhou Q. 2003. Inhibition of P-TEFb (CDK9/Cyclin T) kinase and RNA polymerase II transcription by the coordinated actions of HEXIM1 and 7SK snRNA. Mol Cell. 12:971–982.

Yik JH, Chen R, Pezda AC, Samford CS, Zhou Q. 2004. A human immunodeficiency virus type 1 Tat-like arginine-rich RNA-binding domain is essential for HEXIM1 to inhibit RNA polymerase II transcription through 7SK snRNA-mediated inactivation of P-TEFb. Mol Cell Biol. 24:5094–5105.

Barbara Holland, Associate Editor

Accepted August 23, 2009