The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae)[i]

J Vierna[1ii], K T Jensen[2], A Martínez-Lage[1] & A M González-Tizón[1iii]

[1] *Department of Molecular and Cell Biology, Evolutionary Biology Group (GIBE), Universidade da Coruña, La Coruña, Spain*

[2] *Marine Ecology, Department of Biological Sciences, Aarhus University, Ole Worms Allé 1, Aarhus C, Denmark*

Heredity Volume 107, Issue 2, pages 127–142, August 2011
Received 12 July 2010, Revised 18 October 2010, Accepted 08 November 2010, Published 02 March 2011

How to cite:
Vierna J, Jensen K T, Martínez-Lage A, González-Tizón A M (2011). The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae). *Heredity* **107**: 127–142. https://doi.org/10.1038/hdy.2010.174

Abstract

The linkage between 5S ribosomal DNA and other multigene families has been detected in many eukaryote lineages, but whether it provides any selective advantage remains unclear. In this work, we report the occurrence of linked units of 5S ribosomal DNA (5S rDNA) and U1 small nuclear DNA (U1 snDNA) in 10 razor shell species (Mollusca: Bivalvia: Pharidae) from four different genera. We obtained several clones containing partial or complete repeats of both multigene families in which both types of genes displayed the same orientation. We provide a comprehensive collection of razor shell 5S rDNA clones, both with linked and nonlinked organisation, and the first bivalve U1 snDNA sequences. We predicted the secondary structures and characterised the upstream and downstream conserved elements, including a region at −25 nucleotides from both 5S rDNA and U1 snDNA transcription start sites. The analysis of 5S rDNA showed that some nontranscribed spacers (NTSs) are more closely related to NTSs from other species (and genera) than to NTSs from the species they were retrieved from, suggesting birth-and-death evolution and ancestral polymorphism. Nucleotide conservation within the functional regions suggests the involvement of purifying selection, unequal crossing-overs and gene conversions. Taking into account this and other studies, we discuss the possible mechanisms by which both multigene families could have become linked in the Pharidae lineage. The reason why 5S rDNA is often found linked to other multigene families seems to be the result of stochastic processes within genomes in which its high copy number is determinant.

Keywords: birth-and-death evolution; regulatory regions; 5S ribosomal RNA; U1 small nuclear RNA; linkage; *Ensis*

Introduction

The 5S ribosomal RNA molecule (5S rRNA) is a component of the large subunit of ribosomes, encoded by the 5S ribosomal DNA (5S rDNA) and transcribed by RNA polymerase III. The eukaryote 5S rDNA is a multigene family, typically composed of hundreds of repeats of an approximately 120 nucleotides (nts) RNA coding region (hereafter, 5S) and an intergenic spacer (IGS) usually referred to as nontranscribed spacer

(NTS). The first nts downstream the 5S are transcribed as part of the primary RNA and deleted during RNA maturation (Sharp *et al*., 1984; Sharp and Garcia, 1988), but they are considered as part of the NTS.

The 5S rDNA is characterised by a flexible organisation, as it has been found in clusters composed of similar or divergent tandemly arranged repeats (differences mainly occur within the NTS; for example, Shippen-Lentz and Vezza, 1988), and in clusters of 5S rDNA repeats tandemly linked to other multigene families (for example, Cross and Rebordinos, 2005; Freire *et al*., 2010; Cabral-de-Mello *et al*., 2010). A dispersed organisation of 5S rDNA has also been reported (Morzycka-Wroblewska *et al*., 1985 and references therein), and some species were found to have more than one type of organisation within the genome (Little and Braaten, 1989).

The 5S rDNA multigene family was thought to be characterised by low levels of intragenomic divergence in virtually all species because of the concerted evolution of ribosomal multigene families (see Eickbush and Eickbush, 2007 for a review). Nevertheless, the ocurrence of divergent variants of 5S rDNA within a genome has been described in animals, plants and fungi (for example, Fernandez *et al*., 2005; Rooney and Ward, 2005; Caradonna *et al*., 2007), and in some cases, differences in the RNA coding regions were found to correspond to tissue-specific variants (Peterson *et al*., 1980). Therefore, recent studies have pointed out to a more complex evolutionary scenario in which birth-and-death processes generate new 5S rDNA variants that may be homogenised by unequal crossing-overs and gene conversions. For instance, in *Ensis* razor shells (Schumacher, 1817), the long-term evolution of 5S rDNA was found to be driven by birth-and-death processes and selection, and it was suggested that homogenising mechanisms were also taking part within each variant in each species (Vierna *et al*., 2009). Later on, it was proposed that the levels of intragenomic divergence—much higher within the 5S rDNA than within the major ribosomal genes—were due to the more flexible organisation of 5S rDNA, meaning that homogenisation processes were more efficient within the array(s) of major ribosomal genes, as they may occur in a smaller number. The long-term evolution of both rDNA regions was then proposed to be driven by a mixed process of concerted evolution, birth-and-death evolution and purifying selection, as described by Nei and Rooney (2005) (Vierna *et a*l., 2010).

Most eukaryotic genes are transcribed into precursor messenger RNAs that must undergo splicing, an essential step of gene expression. During precursor messenger RNA splicing, introns are removed from the precursor messenger RNA and exons are ligated together to form mRNA (Will and Lührmann, 2005). Splicing is performed by the spliceosomes, ribonucleoprotein complexes consisting of small nuclear RNAs and several proteins. The U1 small nuclear RNA molecule is a component of the major spliceosome, essential for the interaction with the 5′ splice site of introns (Zhuang and Weiner, 1986). This molecule is encoded by the U1 small nuclear DNA (U1 snDNA), which consists of an RNA coding region (hereafter, U1) and an IGS (when it is organised in tandem repeats). U1 snDNA, transcribed by RNA polymerase II, is a multigene family with a variable number of repeats in each genome (around tens of repeats in the metazoan species studied by Marz *et al*., 2008). Although not much information is available about the organisation of U1 snDNA, it was found to be linked to other multigene families, such as 5S rDNA (Pelliccia *et al*., 2001), other spliceosomal snDNA families (Marz *et al*., 2008) and organised in the same array together with 5S rDNA repeats and other spliceosomal snDNA (Manchado *et al*., 2006). In general, however, clustered copies of distinct or the same small nuclear RNA coding genes are not common in metazoan genomes (Marz *et a*l., 2008).

The evolution of spliceosomal snDNA has been recently studied in two different surveys, covering insect species (Mount *et al*., 2007) and several other metazoan groups (Marz *et al*., 2008), and appears not to be a simple issue. In insects, it is governed by several concurrent forces, namely purifying selection, unequal crossing-overs, gene conversions and birth-and-death processes (Mount *et al*., 2007). Distinguishable U1 snDNA paralogs differentially expressed throughout development have been described in some species (for example, Lo and Mount, 1990), but the snDNA paralog groups seem not to be stable over a long evolutionary time, although they appear independently in several clades (Marz *et al*., 2008).

The linkage between 5S rDNA and U1 snDNA has only been reported in one crustacean (Pelliccia *et al*., 2001) and in one fish (Manchado *et a*l., 2006). In this survey, we report linked units of 5S rDNA and U1 snDNA in 10 razor shell species (Mollusca: Bivalvia: Pharidae) from four different genera. We obtained new data about the genomic organisation of both multigene families in these animals, and studied the genesis and evolution of the 5S rDNA–U1 snDNA linked units. Using the *Ensis* sequences available from DDBJ/EMBL/GenBank and the new sequences obtained, we provide a comprehensive collection of razor shell 5S rDNA variants, including their secondary structures and the characterisation of putative pseudogenes. We also report the first Bivalvia U1 snDNA sequences, including their predicted secondary structures. Finally, several putative regulatory regions of both multigene families were studied in detail.

Materials and methods

Animals

We selected 11 species belonging to family Pharidae (Adams and Adams, 1858; Mollusca: Bivalvia, Table 1). Though a greater sampling effort was made on genus *Ensis*, we tried to represent the whole family by selecting species from its different subtaxons. Thus, from subfamily Cultellinae (Davies, 1935), we studied eight *Ensis* species and one *Ensiculus* (Adams, 1860). The species *Siliqua patula* (Dixon, 1789) was also included in the analysis, as genus *Siliqua* (Mühlfeld, 1811) may represent a separate subfamily from the Cultellinae (see Cosel, 1993). From the other subfamily, Pharinae (Adams and Adams, 1858), we took into consideration the species *Pharus legumen* (Linné, 1758). Two homonymous species, *Ensis minor* (Chenu, 1843) and *E. minor* (Dall, 1899) were studied in this survey, and hereafter they will be referred to as *E. minor* (Chenu) and *E. minor* (Dall). All taxon names follow Cosel (1993) and Cosel (2009), when applicable. Razor shells were provided by several colleagues and preserved in 100% ethanol until species identification, except the *Ensiculus cultellus* (Linné, 1758) sample that consisted of an ethanol-preserved piece of muscle tissue, and *Ensis goreensis* (Clessin, 1888) from which only dried tissue was available.

DNA extraction, PCR, cloning and sequencing

DNA extractions were done from muscle tissue using the NucleoSpin Tissue kit (Macherey-Nagel, North Rhine- Westphalia, Germany). Using the primers 5S-Univ-F and 5S-Univ-R (Vierna *et al*., 2009), we serendipitously amplified complete U1 snDNA sequences flanked by two partial 5S rDNA repeats in the species *Ensis magnus* (Schumacher, 1817) and *P. legumen*. From these sequences, different primer pairs annealing at the 5S and U1 regions of razor shells were designed using GeneFisher (Giegerich *et al*., 1996) (Table 2). PCR reactions were conducted in a final volume of 20 µl using the 2 × Taq Master Mix RED (VWR/Ampliqon, Skovlunde, Denmark), applying the following conditions: an initial denaturation step at 94 °C for 3 min followed by 40 cycles of denaturation at 94 °C for 20 s, annealing at the temperatures indicated in Table 2 for 20 s, extension at 72 °C for 1 min, and a final extension at 72 °C for 5 min. Amplification products were run on 1% agarose gels, stained with a 0.5 µg/ml solution of ethidium bromide, and imaged under UV light. They were cloned using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA). A subset of transformant colonies from each cloning reaction was analysed by PCR in order to check the insert size. From each PCR, we selected one clone per species when only one band was retrieved (that is, in all cases except in one of the PCRs of *Ensis macha* (Molina, 1782) and *E. cultellus* individuals, in which case we obtained two slightly different bands, so two clones were sequenced). Sequencing was performed at Macrogen (Seoul, South Korea) using both T3 and T7 primers (forward and reverse) included in the cloning kit.

Table 1. DNA sequences studied and specimen details.

Species		***sbf.***	***Identification***	***Museum code***	***Sampling site***	***Primer pair names and accession numbers***			
						5S-Univ	***5S-U1***	***U1-5S***	***U1-U1***
Ensis magnus	Schumacher, 1817	Cul.	Cosel, 2009	MNHN 40042	Bonden, Sweden	FN908876*a*	FN908883*a*	FN908894*a*	FN908904*a*
E. magnus	Schumacher, 1817	Cul.	Cosel, 2009		Ortigueira, Spain	FM201454-56*b*			
E. siliqua	(Linné, 1758)	Cul.	Cosel, 2009	MNHN 40047	Vigo, Spain	FM201457-62*b*, FM211689*b*	FN908890*a*	FN908900*a*	FN908908*a*
E. ensis	(Linné, 1758)	Cul.	Cosel, 2009	MNHN 40044	La Capte, France	FM211690-91*b*	FN908885*a*	FN908896*a*	FN908905*a*
E. goreensis	(Clessin, 1888)	Cul.	Cosel, 2009	MNHN 17948	Gorée, Senegal	FM211692*b*			
E. minor	(Chenu, 1843)	Cul.	Cosel, 2009	MNHN 40045	La Capte, France		FN908886*a*	FN908897*a*	FN908906*a*
E. directus	(Conrad, 1843)	Cul.	Cosel, 2009	MNHN 40049	Long Pond, Canada		FN908884*a*	FN908895*a*	
E. directus	(Conrad, 1843)	Cul.	Cosel, 2009		Various localities, Denmark	AM904878-933*b*			
E. macha	(Molina, 1782)	Cul.	Cosel, 2009	MNHN IM-2009-8446	Puerto Lobos, Argentina		FN908887*a*, FN908888*a*	FN908898*a*	FN908907*a*
E. macha	(Molina, 1782)	Cul.	Cosel, 2009	MNHN 40048	Playa Dichato, Chile	FM201452*b*			
E. macha	(Molina, 1782)	Cul.			Concepción, Chile	AM940998-1009*b/c* AM906171-80*b/c*, AM906203-8*b/c*			
E minor	Dall, 1899	Cul.	Cosel, 2009	MNHN IM–2009–8447	Christmas Bay, USA		FN908889*a*	FN908899*a*	
Ensiculus cultellus	(Linné, 1758)	Cul.	John Taylor	BMNH 20070223	Moreton Bay, Australia		FN908881*a*, FN908882*a*	FN908893*a*	FN908903*a*
Siliqua patula	(Dixon, 1789)	Cul.	Dan Ayres	MNHN IM–2009–8448	Ocean City, USA		FN908892*a*	FN908902*a*	FN908910*a*
Pharus legumen	Linné, 1758	Phar.	Cosel, 2009	MNHN 40051	Bandol, France	FN908877-80*a*	FN908891*a*	FN908901*a*	FN908909*a*

Abbreviations: *a*, new sequences; *b*, sequences previously studied by Vierna *et al.* (2009); *c*, sequences previously studied by Fernández-Tajes and Méndez, (2009); Cul., Cultellinae; Phar., Pharinae; sbf., subfamily.

Table 2. Primer pairs used in this survey

Sequence/reference		*T*	*a.r.*	*a.p.*
5S-Univ-F	Vierna et al. (2009)	50 °C	5S	13–32
5S-Univ-R	Vierna et al. (2009)	50 °C	5S	36–55
5S-U1-F	5′ GTCTACGGCCATATCACGTT	61 °C	5S	1–20
5S-U1-R	5′ GTTAGCGCGAACGCAGVC	61 °C	U1	142–159
U1-5S-F	5′ VCTGCGTTCGCGCTAVCC	65 °C	U1	143–160
U1-5S-R	5′ GGTATTCCCAGGCGGTCAC	65 °C	5S	87–105
U1-U1-F	5′ GCAATGGAAGGGCCTCCTCCT	61 °C	U1	49–69
U1-U1-R	5′ TTCGGTTGGGCTGATGCCTG	61 °C	U1	72–91

Abbreviations: a.p., annealing position within each RNA coding region; a.r., annealing region; T, annealing temperature; U1, U1 small nuclear RNA coding region; 5S, 5S ribosomal RNA coding region.

Bioinformatic analyses

Electropherograms were inspected in BioEdit 7.0.9.0 (Hall, 1999). The Blast 2 sequences tool (available at www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi) was used to compare the ends of both the forward and reverse sequences obtained from each clone, which were subsequently overlapped by hand. Sequences obtained were subjected to a sequence-similarity search against the DDBJ/EMBL/GenBank nucleotide collection databases using the blastn algorithm. Sequences similar to other 5S, U1 and their intergenic spacers were deposited in the DDBJ/EMBL/GenBank databases under the accession numbers specified in Table 1. The pair-wise comparisons were also performed in the Blast 2 sequences tool and multiple sequence alignments were carried out in ClustalW 2.0 (Larkin *et al*., 2007), and manually adjusted for local optimisation in MEGA 4.0.2 (Tamura *et al*., 2007). The number of polymorphic sites was retrieved from DnaSP 5.10.0 (Librado and Rozas, 2009). Lengths and p-distances were obtained from MEGA 4.0.2 (Tamura *et al*., 2007). In p-distance calculation, gaps were not considered, and 1000 bootstrap replicates were performed for the estimation of standard errors.

In order to search for putative regulatory conserved elements, sequences upstream and downstream the 5S and U1 regions were analysed. Searches were performed considering the first 100 nt upstream and downstream the RNA coding regions. In the case of U1 upstream analyses, two sequences from the gastropod molluscs *Aplysia californica* and *Lottia gigantea* (provided by Manja Marz, Philipps-Universität, Marburg, Germany) were selected and included in the analyses. Conserved motifs were identified by MEME (Bailey and Elkan, 1994) and they were manually compared with published regulatory elements.

5S and U1 sequences were folded in RNAstructure 5.02 (Reuter and Mathews, 2010) at 15 °C, and we used the efn2 function (Mathews *et al*., 1999) to recalculate the ΔG values. The consensus secondary structures were obtained from the RNAalifold webserver (Hofacker, 2003).

We used PALM (Chen *et al*., 2009) to select nucleotide substitution models and to infer maximum likelihood phylogenies. The best-fit model of nucleotide substitution was directly selected using Modeltest 3.7 (Posada and Crandall, 1998), applying the Akaike information criterion. Phylogenies were constructed by PALM using PhyML (Guindon and Gascuel, 2003). Starting trees were obtained by the BioNJ algorithm (Gascuel, 1997) and gaps were treated as unknown characters. The number of substitution rate categories employed was eight, and the bootstrap test (Felsestein, 1985) was used to estimate node support (1000 replicates).

Maximum parsimony phylogenies were obtained from PAUP*4.0b10 (Swofford, 2002) as detailed in Vierna *et al*. (2010). Following Marz *et al*. (2008), we calculated phylogenetic networks in addition to phylogenetic trees, using the neighbour-net algorithm (Bryant and Moulton, 2004), implemented as part of the SplitsTree4 package (Huson and Bryant, 2006).

Different gene tandem arrangements were drawn using pDRAW32 (AcaClone sofware, http://www.acaclone.com/) and we edited all phylogenetic trees in FigTree 1.2.2 (Andrew Rambaut, http://tree.bio.ed.ac.uk/software/figtree/).

Results

Sequence characterisation

The identification of 5S, U1 and spacer sequences was performed by comparing them against the DDBJ/EMBL/GenBank nucleotide collection databases, as explained above. For the sake of clearness, all spacer sequences downstream a 5S will be referred to as NTS, and all spacers downstream a U1, as IGS. All complete 5S sequences were 120 nts and NTS ranged between 283 and 986 nts. All complete U1 sequences were 164 nts except the ones obtained from *S. patula*, that had a nucleotide insertion at position 37. IGS ranged between 222 and 422 nts. The DDBJ/EMBL/GenBank accession numbers of the sequences studied are listed in Table 1.

Average GC contents were 55.1% for the 5S region, 54.8% for the U1 region, 38.8% for the NTS and 41.9% for the IGS. The number of polymorphic sites in the RNA coding regions was S=32 for the 5S region and S=20 for the U1 region.

Hereafter, clones containing partial or complete repeats of both multigene families will be referred to as mixed clones.

Alignments

An initial alignment of the NTS region showed that the NTSs of razor shells were highly divergent, so sequences had to be grouped separately, according to their similarity. After performing several combinations, we divided the NTSs into seven supergroups and 17 groups. Each supergroup was named using a Roman numeral and each group was denoted by a Greek letter following Vierna *et al*. (2009). Supergroups and groups contained sequences belonging to one or more species. Similarily, IGS sequences were divided into two groups, one containing all *Ensis* and *Ensiculus* and the other one containing *Pharus* and *Siliqua* IGSs. The species composition, lengths and mean *P*-distances for each spacer group and supergroup were recorded in Table 3.

Let's now consider only the spacer sequences from mixed clones. We were able to align all IGSs from *Ensis, Ensiculus, Pharus* and *Siliqua* individuals, but the divergence among them was evident; however, the last part of the alignment (containing the upstream region of the next 5S repeat) revealed a more conserved region. Quite the opposite, the analysis of the NTSs from mixed clones (upstream U1 sequences) revealed that these spacers were less conserved than the IGSs and could not be aligned at once. In this case, we were able to align all *Ensis* sequences (from supergroup II), except an NTS from the species *E. macha* (from supergroup V). The NTSs from the species *P. legumen* and S. patula, belonging to supergroup IV, could also be aligned together. However, E. cultellus NTSs could not be aligned to *Ensis*, *P. legumen* or *S. patula* sequences.

Table 3. Intergenic spacer groups and supergroups

NTS group	*Species*	*Clade*	*N*	*Length*	*Mean P-distance*
Supergroup I			*72*	*286–329*	*0.135±0.010*
α	*Ensis directus*	A	41	321–329	0.011±0.003
β	*Ensis macha*	A	18	314–318	0.019±0.004
ζ	*Ensis magnus, E. siliqua, E. ensis, E. goreensis*	E	13	286–315	0.042±0.006
Supergroup II			28	407–965	0.240±0.012
γ	*Ensis directus*	A	11	444–654	0.023±0.004
δ	*Ensis directus*	A	4	407	0.002±0.002
η*	*Ensis magnus, E. siliqua, E. ensis, E. minor (Chenu)*	E	9	893–965	0.046±0.004
θ*	*Ensis directus, E. macha, E. minor (Dall)*	A	4	926–960	0.127±0.008
Supergroup III			14	405–620	0.141±0.009
ε 1	*Ensis macha*	A	6	603	0.011±0.003
ε 2	*Ensis macha*	A	5	618–620	0.004±0.002
ξ*	*Ensiculus cultellus*		3	405	0.010±0.004
Supergroup IV			8	355–550	0.241±0.013
μ	*Pharus legumen*		3	548–550	0.005±0.002
ο*	*Siliqua patula*		2	355	0
λ*	*Pharus legumen*		3	419–420	0
Supergroup V					
ι*	*Ensis macha*	A	1	776	
Supergroup VI			2	209–369	0.077±0.018
π*	*Siliqua patula*		1	369	
ρ*	*Siliqua patula*		1	209	
Supergroup VII			2	283–332	0.366±0.029
κ	*Pharus legumen*		1	332	
ν*	*Ensiculus cultellus*		1	283	
IGS group	***Species***		**n**	**Length**	**Mean P-distance**
Supergroup Ensis–Ensiculus			15	222–422	0.203±0.015
Ensis spp.	*Ensis directus, E. macha, E. minor (Dall, 1899)*		13	225–231	0.177±0.014
	Ensis magnus, E. siliqua, E. ensis, E. minor (Chenu, 1843)				
Ensiculus cultellus	*Ensiculus cultellus*		2	421–422	0.002±0.002
Supergroup Pharus–Siliqua			5	236–342	0.193±0.017
Siliqua patula	*Siliqua patula*		2	236	0.064±0.015
Pharus legumen	*Pharus legumen*		3	342	0.007±0.004

Abbreviations: A, American clade; E, European clade (*Ensis* phylogenetic clades according to Vierna *et al.* (unpublished data)); IGS, intergenic spacer (downstream a U1 small nuclear RNA coding region); n, sample size, NTS, nontranscribed spacer (intergenic spacer downstream a 5S ribosomal RNA coding region).
Asterisks (*) indicate nontranscribed spacers linked to U1 small nuclear DNA;

In the alignment of *Ensis* U1–U1 clones (Supplementary File S1), all Ensis IGSs displayed a region of similarity with δ- and γ-NTSs, from the species *E. directus* (Conrad, 1843). This region was located at the end of the IGS (just upstream the 5S region) and resembled the last portion of δ- and γ-NTSs. Downstream this 5S region, in the NTS, we found another region of similarity with δ- and γ-NTSs, and downstream of it there was a fragment resembling a 5S (probably an old pseudogenised copy). Even though this pattern was only found in *Ensis* species, the first portion of the alignment that corresponded to the U1–IGS–5S sequence (positions 1 to 427, Supplementary file S1), could be aligned to *E. cultellus* clones, and with more difficulties, to *P. legumen* and *S. patula* ones (as explained above).

Upstream elements

A conserved region was identified at −25 nts from both the 5S rDNA and U1 snDNA transcription start sites (Supplementary file S2) and named −25 region. It was a TATA-like motif in the 5S upstream sequences (Supplementary file S3a), and upstream the U1 region (Supplementary file S3b), it was an A/G-rich motif: AAAAG in *Ensis* and *E. cultellus*, GGGGA in gastropods, AAATG in *P. legumen* and GTAAG upstream *S. patula* putative-pseudogenised U1 sequences (see U1 predicted secondary structures). Another motif (AAAGC, Supplementary file S2) was identified just upstream the U1 snDNA transcription start site, identical to the one found in *Drosophila melanogaster* (Lo and Mount, 1990) and in other organisms (see Discussion), but it only occurred in some of the razor shell sequences. Finally, a less conserved region was found upstream the −25 region in U1 snDNA upstream sequences (Supplementary file S2), centred at −44 nts.

Although it was not possible to align all *Ensis* NTSs at once, we were able to align the 100 nt upstream the transcription start site of 5S rDNA of *Ensis* species. These stretches were the last part of either NTS or IGS sequences. We failed to include the other Pharidae species in this alignment, as sequences were not conserved among genera.

Internal regulatory regions

5S internal control regions (ICR I to IV) were compared with those described in *D. melanogaster* (Sharp and Garcia, 1988). As some ICRs coincided with the primer-annealing regions, some sequences were excluded from the comparisons, and sequences amplified with the 5S-Univ primers (Table 2) were only included in the ICR IV analysis. Results were as follows: 12/16 matches within ICR I (positions 3–18); 7/8 matches within ICR II (positions 37–44); 11/14 matches within ICR III (positions 48–61); and 14/21 matches within the ICR IV region (positions 78–98). The degree of conservation of these elements within razor shells was of 14/16, 8/8, 13/14 and 15/21 matches, respectively. Similarly, positions 50–61 (Box A), 80–89 (Box C) and 62–79 (intermediate sequence) were compared with those described by Pieler *et al.* (1987) in *Xenopus laevis*, obtaining 6/12, 6/10 and 12/18 matches. Within razor shells, the matches obtained were 9/12, 7/10 and 14/18.

Six U1 internal regions that appear to be conserved accross metazoa (Zhuang and Weiner, 1986; Marz *et al*., 2008) were analysed in all razor shell sequences. They were compared with the two gastropod sequences (see above), the ones from the insect *D. melanogaster* (Lo and Mount, 1990), and those from crustaceans *Asellus aquaticus* and *Proasellus coxalis* (Barzotti *et al*., 2003). Considering as a reference the *E. magnus* U1 sequence (see U1 predicted secondary structures), they correspond to the following positions: the 5′ end (includes the 5′ splice site, Zhuang and Weiner, 1986 and references therein); 28–33 (within the U1–70 K protein binding site, Query *et al*., 1989); the stem-loop II positions 53–55, 65–72 and 84–86 (U1-A protein binding region, Scherly *et al*., 1989); and positions 124–132 (include the Sm protein binding region, named ‘domain A’ by Branlant *et al*., 1982). The most conserved region was the 5′ end (11 nt) that was identical in

all sequences. Positions 28–33 were identical in all sequences, but *L. gigantea* had an additional G inserted between the first and the second nt. Positions 65–72 were also identical, except in the last nt. Finally, the 124–132 region was also conserved with the exception of the sixth and last nt. The remaining two regions were conserved at positions 54–55 and 84–85 in all molluscs and arthropods.

Termination signals

One or more TTTT stretches (required for 5S rDNA transcription termination, Bogenhagen and Brown, 1981; Huang and Maraia, 2001; Richard and Manley, 2009) occurred within the first 20 nt of all NTSs, except for those belonging to the δ-group, for which the first perfect TTTT was located at positions 96–99. All NTSs except those from the α-group had a TTT motif within the first six nts, and 124/125 sequences had a T residue in the first position.

The analysis of the first portion of the IGS revealed that a TAAAA motif occurred in all *Ensis* species and *E. cultellus*, contiguous to the 3′ end of the U1. Sequences from *S. patula* had a TCCAT and those from *P. legumen*, an ATATA motif. All sequences displayed between two and four AAT stretches within the first 88 sites downstream the U1 region. However, no other evidence of conserved regions that could be involved in the formation of the 3′ end (as the 3′ box, Hernandez, 1985) were found. The first 50 sites downstream the U1 region were very rich (44.5%) in adenines.

Genomic organisation

Mixed clones of 5S rDNA and U1 snDNA were retrieved from all species analysed, except from *E. goreensis* (the quality of the extracted DNA was very low) and both multigene families displayed the same orientation. Both U1–5S and 5S–U1 primer pairs yielded PCR products, and amplifications using U1–U1 primers were successful in eight of them (see primer details in Table 2). Tandemly-arranged 5S rDNA repeats (a partial 5S, an NTS and a partial 5S) were retrieved from *P. legumen* and six *Ensis* species: *E. goreensis*, *E. magnus*, *E. siliqua* (Linné, 1758), *E. ensis* (Linné, 1758), *E. directus* and *E. macha*; two 5S rDNA repeats flanked by U1 snDNA were sampled from the species *S. patula* and *E. cultellus* (Figures 1a and b) and tandemly-arranged U1 snDNA repeats were not found in any of the species studied, as clones obtained with the U1–U1 primers always had one or two 5S rDNA repeats in between. The sequence analysis of clone ends permitted us to determine which clones could be overlapped, assuming that identical spacer sequences retrieved from different clones from the same individual were, in fact, the same copy. Therefore, a sequence of 2217 nts was obtained from *P. legumen* clones (Figure 1c). In all *Ensis* species, the organisation of mixed clones was very similar and consisted of two partial U1 snDNA repeats flanking one complete 5S rDNA repeat and/or vice-versa (for example, Figure 1d).

One of the two 5S–U1 clones from the species *E. macha* (Supplementary file S4) was different from all other *Ensis* clones. It consisted of a partial 5S followed by a divergent NTS (from group ι, supergroup V). This NTS contained a region of similarity with ε-2 NTSs (from the species *E. macha*), a 50 nts truncated 5S copy, 95 nts very similar to the previous ε-2-similar region and a region that matched to a sequence associated to a *Taenia solium* spliced leader and spliced leader mini-exon (from Brehm *et al.*, 2002) that appeared to be a silent DNA (Klaus Brehm, personal communication). At the end of the clone, we found the type A U1 sequence (see U1 predicted secondary structures) that was somewhat divergent, with respect to the other *Ensis* U1s (see Phylogenetic trees and networks).

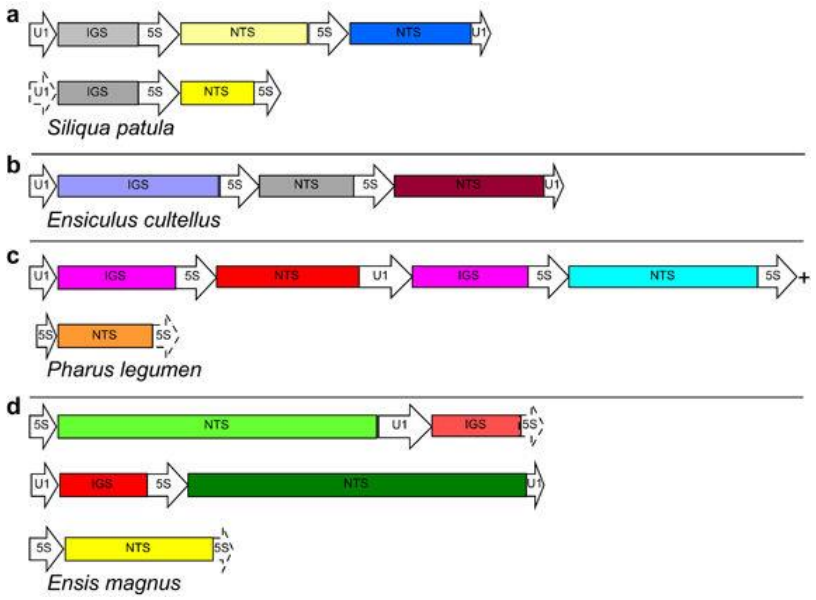

Figure 1. Different 5S ribosomal DNA and U1 small nuclear DNA tandem arrangements sampled from razor shell species. For each species, drawings were constructed using sequences retrieved from the same individual. Drawings are done to scale (except dash lined boxes). (**a**) Grey IGSs are very similar (identities=90%, gaps=2%, E value=1×10^{-93}); yellow NTSs are similar but the darker one has a deletion of 160 nts (identities=90% in both aligned regions, gaps=4% in the first region and gaps=1% in the second region, E value=4×10^{-58}) . Blue and yellow NTSs are very divergent and could not be aligned. Blue, *o*-NTS; light yellow, π-NTS; dark yellow, ρ-NTS. (**b**) NTSs are very divergent and could not be aligned. Grey, ν-NTS; brown, ξ-NTS. (**c**) Both IGS (same colour) are identical. The three NTS are very divergent and could not be aligned. Red, λ-NTS; light blue, μ-NTS; orange, κ-NTS. (**d**) Green NTSs (both η) are very similar (identities=82%, gaps=9%, E value=0), red IGSs are also very similar (identities=92%, no gaps, E value=1×10^{-93}). Yellow NTS corresponds to ζ-group. 5S, 5S ribosomal RNA coding region; U1, U1 small nuclear RNA coding region; NTS, nontranscribed spacer; IGS, intergenic spacer. (+) Reconstructed by overlapping clones from the same individual (see main text).

5S predicted secondary structures

Fourteen different sequences that contained a complete 5S after excluding the primer-annealing regions were considered for the secondary structure prediction. The predicted structures of 11 sequences (Supplementary file S5 a-k) were consistent with the general secondary structure of 5S rRNA (Barciszewska *et al*., 2000). The presence of two fixed thymines (uracils in the RNA molecule) in positions 80 and 96 caused the formation of an additional mini-loop in helix IV (Figure 2, Supplementary file S5).

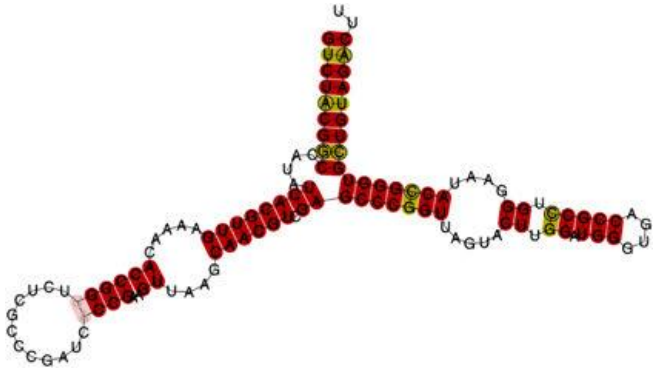

Figure 2. Predicted consensus secondary structure of razor shell 5S ribosomal RNA. Helices are named with Roman numerals and letters correspond to loops, following Barciszewska et al. (2000). Red indicates one type of base pair and ochre indicates two types of base pairs. Pale colours indicate pairs that cannot be formed by all sequences.

The other three sequences (from the species *E. ensis*, *E. siliqua* and *E. directus*) may be pseudogenised sequences, because they displayed abnormal secondary structures (Supplementary file S5 l-n). The ΔG values (Table 4) of these putative pseudogenes were the most positive ones, indicating that the structures are less stable.

Compensatory changes C–G → A–T occurred at paired positions 8–111 in the *E. cultellus* sequence (Supplementary file S5 i). A C → T change in position 45 (Supplementary file S5 j and k) made helix III to be one nt smaller in both *P. legumen* predicted structures.

U1 predicted secondary structures

After excluding the primer-annealing regions, 30 razor shell U1 sequences ranging between 48 and 164 nts were obtained. Two of them, from the species *E. magnus* and *P. legumen*, were complete sequences. The species *E. ensis* and *E. siliqua* had identical sequences to *E. magnus*. Complete U1s were build up by overlapping sequences obtained with primer pairs U1–U1 and 5S–U1 (Table 2) from *E. minor* (Chenu), *E. cultellus* and *S. patula* individuals. Three different clones containing partial U1s were retrieved from the *E. macha* individual (two 5S–U1 and one U1–U1) and two different U1 types, named A and B, were characterised. *E. macha* type A U1 was retrieved from a 5S-U1 clone and lacked 23 nt from its 3′ end, which were completed for secondary structure prediction with the corresponding 3′ end of *E. magnus* U1. *E. macha* type B U1 was built up by overlapping the sequence obtained with primers U1–U1 with the sequence obtained from the other 5S–U1 clone. The *E. directus* and *E. minor* (Dall) sequences were also completed in order to predict their secondary structures. Complete U1s from the gastropod molluscs *A. californica* and *L. gigantea* were included, as explained above, in the analyses.

U1 predicted secondary structures were in agreement with previously proposed ones (for example, that of *D. melanogaster*, Lo and Mount, 1990; and consensus structures for several metozoan groups, Marz *et al*., 2008) (Figure 3; Supplementary file S6). The secondary structure of stem-loop IV was conserved in razor shells and gastropods and consisted of one hairpin loop, one internal loop, two stems and a central nonpaired region containing the 5′ splice site (Zhuang and Weiner, 1986 and references therein) and the Sm proteins binding region (‘domain A’, Branlant *et al*., 1982). The predicted secondary structure of stem-loop I was similar in all U1s, except *E. macha* type A U1 and *L. gigantea* sequences (Supplementary file S6 c and i), whose internal loops were 2–3 nts bigger. *E. macha* type B U1, *E. directus*, *A. californica* and *E. minor* (Dall) sequences had two internal loops in stem-loop III (Supplementary file S6 d, e, h, and j), whereas the rest of them had three. Stem-loop II had two internal and one hairpin loop in all structures, except in *L. gigantea* and *E. minor* (Dall) (Supplementary file S6 i and j), which lacked an internal one. However, this sequence and the one from *S. patula* were considered as putative pseudogenised copies because of their abnormal predicted secondary structures and their ΔG values (Supplementary file S6 j and k; Table 4).

Table 4. ΔG values calculated at 15 °C using the efn2 function for each predicted secondary structure

ΔG (kcal mol−1)	***Species***	***NTS group***	***Linked to U1 snDNA?***	***Complete sequence?***	***m.s. (nts)***
5S rRNA					
−46.5	*Ensis ensis*	ζ	No	Yes	120
−47.6	*Ensis siliqua*	η	Yes	Yes	120
−49.6	*Ensis directus*	α	No	Yes	120
−52.0	*Pharus legumen*	λ	Yes	Yes	120
−52.2	*Pharus legumen*	λ	Yes	Yes	120
−54.0	*Ensiculus cultellus*	ν	Yes	Yes	120
−54.9	*Ensis directus*	α	No	Yes	120
−55.2	*Siliqua patula*	ρ	Yes	Yes	120
−55.2	*Siliqua patula*	π	Yes	Yes	120
−55.2	*Siliqua patula*	ο	Yes	Yes	120
−55.2	*Ensis directus*	α	No	Yes	120
−55.2	*Ensis directus*	α	No	Yes	120
−56.6	*Ensis directus*	δ	No	Yes	120
−56.6	*Ensis minor* (Chenu)	η	Yes	Yes	120
−56.6	*Ensis macha*	θ	Yes	Yes	120
−56.6	*Ensis magnus*	η	Yes	Yes	120
−56.6	*Ensis directus*	γ	No	Yes	120
−56.7	*Ensis ensis*	η	Yes	Yes	120

ΔG (kcal mol−1)	***Species***	***Linked to 5S rDNA?***	***Complete sequence?***	***m.s. (nts)***
U1 snRNA				
−70.1	*Siliqua patula*	Yes	Yes	165
−70.3	*Ensis minor* (Dall)	Yes	No	164
−83.9	*Aplysia californica*	No	Yes	161
−84.7	*Ensis macha* A	Yes	No	164
−85.4	*Ensis macha* B	Yes	Yes	164
−85.9	*Ensis minor* (Chenu)	Yes	Yes	164
−86.1	*Ensis directus*	Yes	No	164
−86.2	*Ensis siliqua*	Yes	Yes	164
−86.2	*Ensis magnus*	Yes	Yes	164
−86.2	*Ensis ensis*	Yes	Yes	164
−87.9	*Ensiculus cultellus*	Yes	Yes	164
−87.9	*Pharus legumen*	Yes	Yes	164
−91.3	*Lottia gigantea*	No	Yes	166

Abbreviations: IGS, intergenic spacer (downstream a U1 small nuclear RNA coding region); m.s., molecule size; NTS, non-transcribed spacer (intergenic spacer downstream a 5S ribosomal RNA coding region); nts, nucleotides; U1 snDNA, U1 small nuclear DNA; U1 snRNA, U1 small nuclear RNA; 5S rDNA, 5S ribosomal DNA; 5S rRNA, 5S ribosomal RNA.

Most positive ΔG values correspond to less stable structures. Bold values correspond to putative pseudogenised copies (see main text).

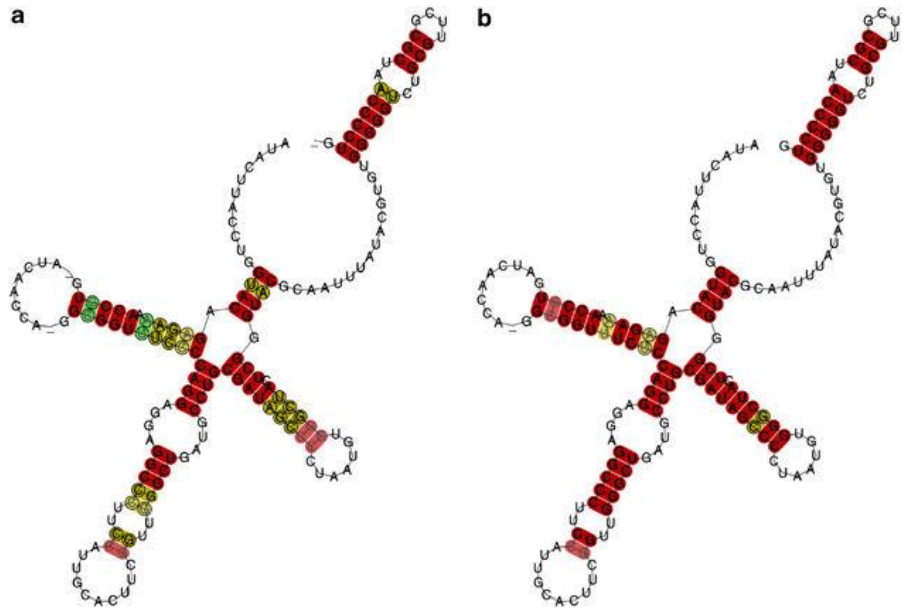

Figure 3. Predicted consensus secondary structure of U1 small nuclear RNA. Stem-loops are indicated by Roman numerals, following Lo and Mount (1990). Red, ochre and green indicates one, two and three types of base pairs, respectively. Pale colours indicate pairs that cannot be formed by all sequences. (a) Including razor shell and gastropod sequences. (b) Including only razor shell sequences.

Phylogenetic trees and networks

Phylogenetic trees were constructed both under maximum likelihood and maximum parsimony criteria and the resulting tree topologies were consistent in all cases. All trees shown were constructed with no outgroups.

The phylogenetic trees and the network performed with the 5S sequences of razor shells did not show a clear clustering by species. However, they also failed to show a well-supported clustering by 5S variants (Supplementary file S7).

We were able to align the upstream region of *Ensis* 5S rDNA, and the phylogenies performed showed that 5S rDNA can be divided into four different groups according to their upstream sequences (see Figure 4).

Because of the impossibility of aligning all NTSs at once due to their high degree of divergence, NTS phylogenies were performed considering each one of the supergroups with $n \geq 3$ (see Table 3). In supergroup I phylogeny (Figure 5a), each NTS group was recovered as monophyletic with high bootstrap support, and α- and β-sequences were included in a highly supported clade. In supergroup II phylogeny (Figure 5b), θ- and η-sequences (from mixed clones) were included in the same clade (bootstrap value of 100), with respect to a clade formed by γ- and δ-sequences. However, θ-sequences were very similar to γ-and δ-sequences in some regions of the alignment, and may represent an intermediate state between η- and γ-/δ-NTSs. The alignment of supergroup III sequences displayed an unexpected similarity between *E. macha* and *E. cultellus* NTSs (which appeared to be somewhat conserved among different genus). However, in the corresponding phylogeny, each NTS group was highly supported (Figure 5c). Supergroup IV alignment also revealed a certain degree of conservation among NTSs retrieved from different genus, and the phylogeny supported each NTS group with the highest value (Figure 5d).

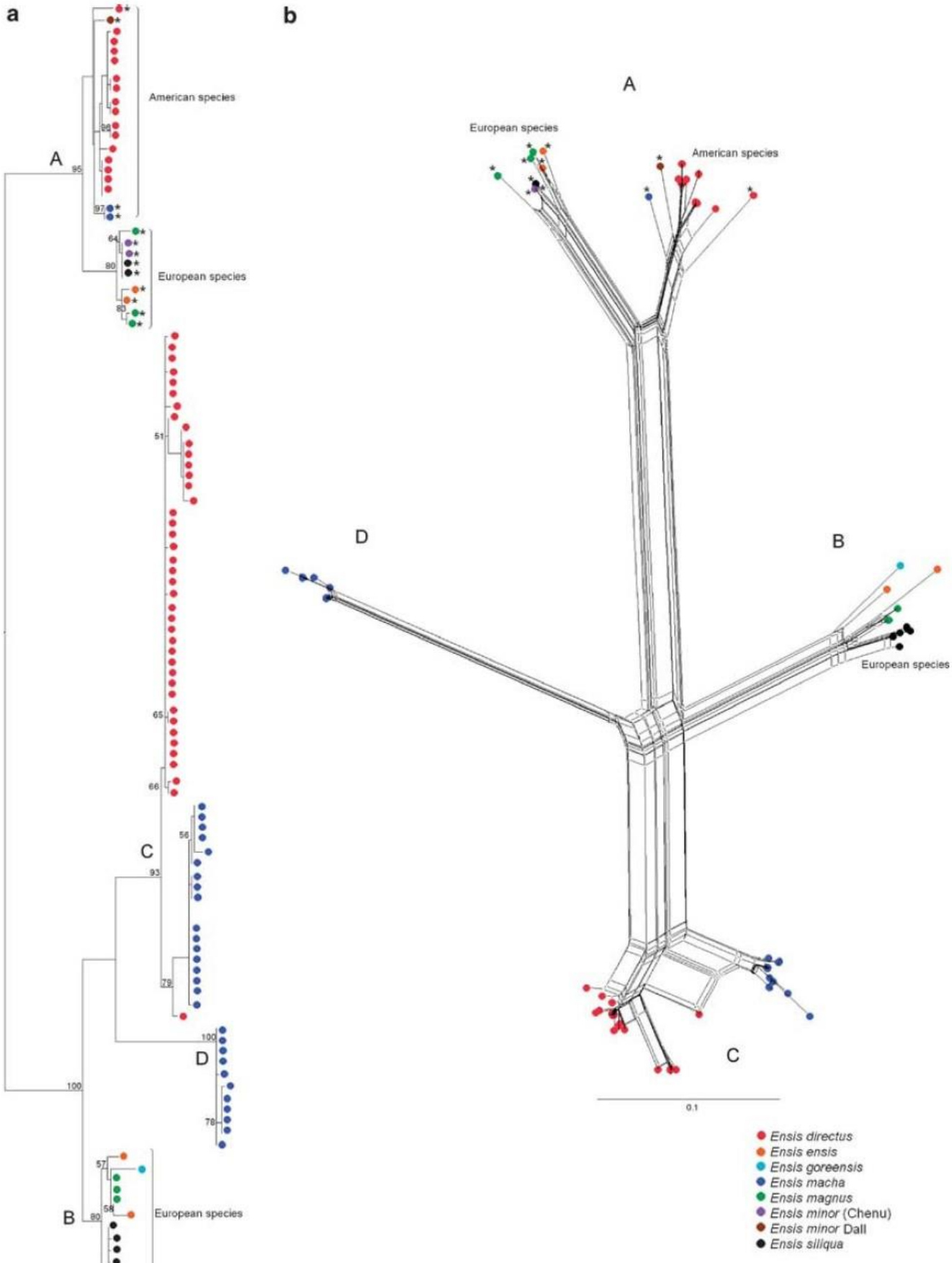
a
American species
A
European species
B
C
D
0.2
b
A
European species
American species
D
B
European species
C
0.1
Ensis directus
Ensis ensis
Ensis goreensis
Ensis macha
Ensis magnus
Ensis minor (Chenu)
Ensis minor Dall
Ensis siliqua

Figure 4. Phylogenetic relationships of the 100 nucleotides upstream the 5S ribosomal DNA transcription start site of *Ensis* species. Four different types of upstream regions (A–D) are identified. Upstream region (A) includes sequences from mixed clones of 5S ribosomal DNA and U1 small nuclear DNA of the European species, the American species and sequences from γ- and δ-NTSs from *E. directus*; (B) includes sequences from nonmixed clones of the European species; (C) sequences from the American species (*E. directus* α-NTSs and *E. macha* β-NTSs); and (D) sequences from *E. macha* ε-I and ε-II NTSs. The relationships among the different sequences are consistent with the phylogenetic history of the genus, as European and American species are reciprocally monophyletic (Vierna *et al.* unpublished). However, the phylogenetic pattern must be understood in the light of a birth-and-death evolutionary scenario (see main text). Asterisks (*) indicate the repeats retrieved from mixed clones. (**a**) Maximum likelihood phylogenetic tree constructed using the K80+G model. Numbers on the tree correspond to nonparametric bootstrap supports (1000 replicates) and they are reported only for nodes with values ≥ 50. Each upstream region type is indicated at the most external node common to all its sequences. (**b**) Phylogenetic network constructed using the neighbour-net algorithm and uncorrected P-distances. For NTS types, see Table 3.

Figure 5. Maximum likelihood phylogenies of the nontranscribed spacers (NTSs) downstream the 5S ribosomal RNA coding regions of razor shell species. Numbers on the trees correspond to nonparametric bootstrap supports (1000 replicates) and they are reported only for nodes with values ≥ 50. NTSs groups (Table 3) are indicated. Asterisks (*) indicate NTS sequences retrieved from mixed clones of 5S ribosomal DNA and U1 small nuclear DNA. (**a**) Phylogeny of supergroup I NTSs reconstructed by the TVM+G model. (**b**) Phylogeny of supergroup II NTSs reconstructed by the TVM+G model. (**c**) Phylogeny of supergroup III NTSs reconstructed by the K81uf+G model. (**d**) Phylogeny of supergroup IV NTSs reconstructed by the GTR+G model.

Sequences considered for the U1 secondary structure prediction were subjected to phylogenetic analyses, excluding putative pseudogenised copies (Figure 6). *Ensis* U1 sequences were included in a nonsupported clade, and all of them, except the divergent *E. macha* type A U1 were recovered as monophyletic with a bootstrap support of 70. However, if the divergent sequence was excluded from the analysis (tree not shown), then the clade containing all remaining *Ensis* U1s decreased its bootstrap value. European *Ensis* sequences were grouped together with a bootstrap value of 91. Razor shell and gastropod sequences were reciprocally monophyletic with the highest support.

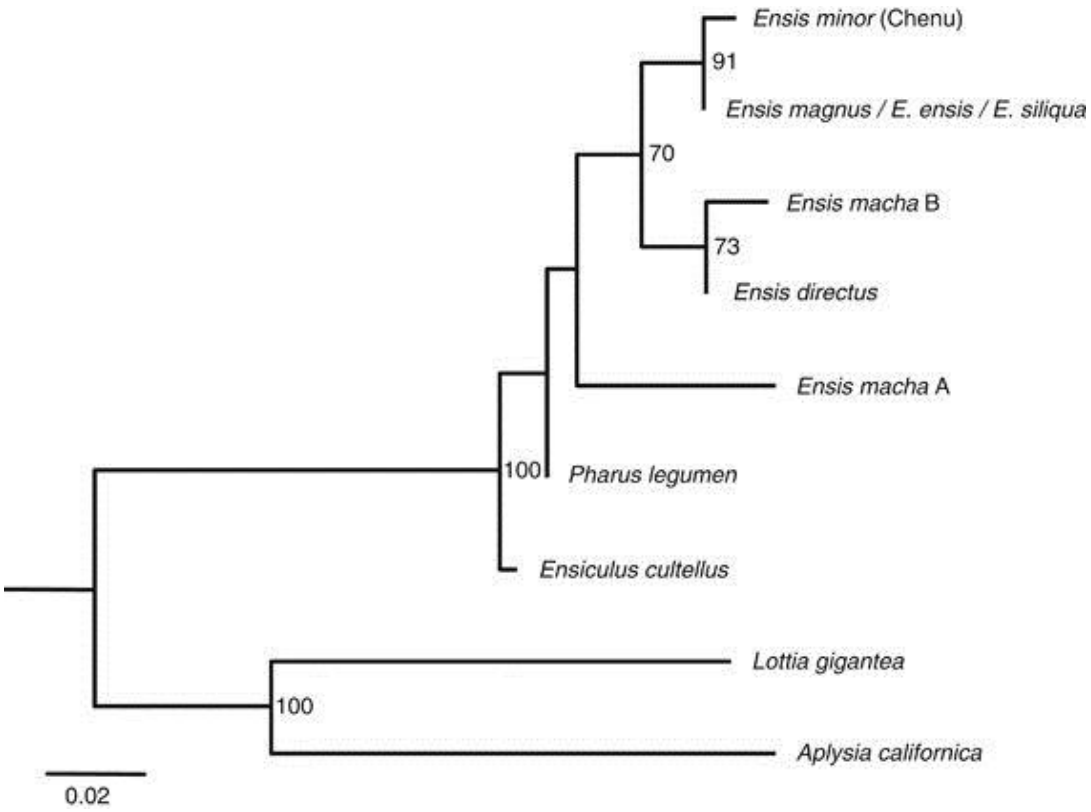

Figure 6. Maximum likelihood phylogeny of the U1 small nuclear RNA coding region of the razor shell and gastropod species, reconstructed using the K81+I model. Sequences analysed correspond to putative functional copies on the basis of their predicted secondary structures and free energies. Numbers on the tree correspond to nonparametric bootstrap supports (1000 replicates). They are reported only for nodes with values ≥ 50. *Ensis macha* type A and *E. directus* sequences were completed with the last 23 nts of the *E. magnus* sequence (see main text).

Two different phylogenies of IGS sequences were performed (Figure 7), one including supergroup *Ensis—Ensiculus* sequences, and another one including supergroup *Pharus—Siliqua* ones. The phylogeny of supergroup *Ensis—Ensiculus* (Figure 7a) recovered American and European *Ensis* sequences, as reciprocally monophyletic with high bootstrap support; the same happened with *Ensis* and *Ensiculus* sequences. In this tree, IGSs from the species *E. macha* were the ones located downstream type B U1s (no IGS downstream the type A U1 was sampled). The phylogeny of supergroup *Pharus—Siliqua* (Figure 7b) also recovered sequences from each species as monophyletic.

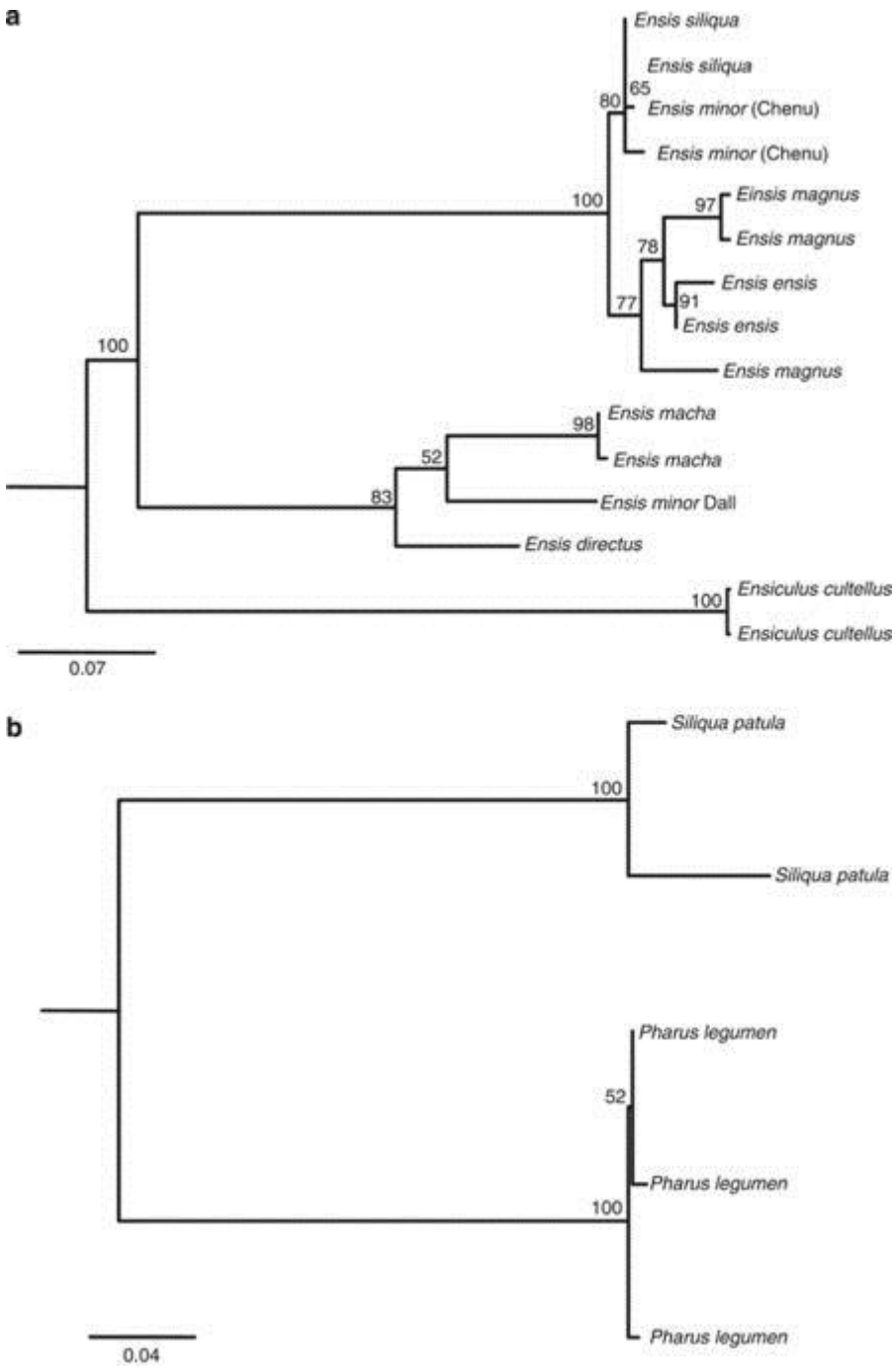

Figure 7. Maximum likelihood phylogenies of the intergenic spacers (IGS) downstream the U1 small nuclear RNA coding region. Numbers on the tree correspond to nonparametric bootstrap supports (1000 replicates). They are reported only for nodes with values ≥ 50. (**a**) Phylogeny of supergroup *Ensis—Ensiculus* IGSs reconstructed following the HKY+G model. (**b**) Phylogeny of supergroup *Pharus—Siliqua* IGSs reconstructed following the HKY+G model. For IGS types, see Table 3.

Discussion

Long-term evolution of 5S rDNA in Pharidae

The 5S region of razor shells was not very polymorphic but the last three sites varied widely when considering the whole dataset. This means that the real number of variants in each species is higher than the number of predicted secondary structures obtained (because we only used complete 5Ss, after excluding the primer-annealing regions, for secondary structure prediction). We characterised several 5S sequences and found that a single species could have more than one 5S variant and that some of these variants were shared among species. Similarily, some NTSs were more closely related to NTSs from other species (and genera) than to NTSs from the species they were retrieved from. Therefore, the existence of divergent NTS sequences predates the speciation of the group. Several variants likely already occurred in the most recent common ancestor of the Pharidae (ancestral polymorphism) and some of them were retained until the present time.

The presence of pseudogenes in a multigene family strongly suggests that it evolves under a birth-and-death process (Rooney and Ward, 2005 and references therein). In this survey, we have found putative pseudogenised and truncated 5S copies. However, the long-term evolution of 5S rDNA in Pharidae appears to be a more complex issue. New variants arise through gene duplication, some of them are retained in the genome and others accumulate mutations and become pseudogenes (birth-and-death process). The action of purifying selection seems to be important to mantain the integrity of the RNA-coding regions and the upstream and downstream elements, and unequal crossing-overs and gene conversions should be also taking part and may be responsible for some of the sequence homogeneity (at least among 5S rDNA repeats located in the same array). We could suppose that divergent NTSs are located at different arrays where they have evolved independently, but we have shown that some species (*S. patula* and *E. cultellus*) have divergent NTS organised in tandem. In the same way, a few clones containing α- and δ-, and α- and γ- NTSs were characterised in *E. directus* (Vierna *et al*., 2009), and other studies found an intermixed organisation of 5S rDNA variants in grey mullets and *E. macha* (Gornung *et al*., 2007; Fernández-Tajes and Méndez, 2009). These findings support the idea that 5S rDNA frequently moves on within genomes (see below). If we consider that a 5S rDNA variant is the result of independent (nonconcerted) evolution in a given genomic location, this variant may have later been transposed into another array containing a different variant. Then, the intermixed organisation would be the result of duplications involving both variants, so they could eventually spread throughout the array. From our results, however, it is not clear whether divergent variants located in the same array are being more and more homogenised through the mechanisms typically involved in the concerted evolution of ribosomal multigene families. In conclusion, the long-term evolution of 5S rDNA in Pharidae has been mainly driven by birth-and-death processes and purifying selection. Nevertheless, homogenising mechanisms, such as unequal crossing-overs (favoured by the tandem organisation of 5S rDNA repeats) and gene conversions have been probably taking part, in agreement with what was previously reported on *Ensis* species (Vierna *et al*., 2009, 2010). Interestingly, recent studies in various animal groups have come to the same conclusion (Fujiwara *et al*., 2009; Freire *et al*., 2010; Úbeda-Manzanaro *et al*., 2010). Other techniques, such as fluorescent *in situ* hybridisation may provide interesting data regarding the chromosomal locations of 5S rDNA arrays in razor shells and this should be an issue for further research.

U1 snDNA variation

We have characterised U1 snDNA for the first time in Bivalvia. Some of the species shared the same U1 variant, many others had a single U1 (not shared) variant and one of them (*E. macha*) had two different U1s, named type A and type B. The phylogeny of the U1 region places the *E. macha* type A sequence outside the clade formed by the other *Ensis* sequences. Taking into consideration that the relationships between this clade and the *E. cultellus*, *P. legumen* and *E. macha* type A sequences were not resolved, the latter one could

be an old copy that diverged before the speciation of *Ensis*. However, it could well be a pseudogenised copy too (for example, derived from the type B U1) because the –25 region was different in two sites compared with the other *Ensis* sequences. We cannot be sure whether the predicted secondary structure was functional or not, as it was somewhat different compared with the other *Ensis* structures but had an intermediate ΔG value.

In the survey by Marz *et al.* (2008), discernible paralogs of spliceosomal snDNA multigene families were not uncommon within genera or families, but no dramatically different paralogs were found. We should take into account that we have only searched for tandemly repeated U1 snDNA (not found) or U1 snDNA linked to 5S rDNA. This means that dispersed U1 snDNA and U1 snDNA linked to other multigene families may occur in the genomes of Pharidae species, and these (hypothetical) copies and the ones linked to 5S rDNA would be paralogs.

We should be cautious regarding U1 snDNA long-term evolution because the number of repeats we obtained from each species was small. In any way, it is clear that duplication events and purifying selection have been involved.

Upstream elements, internal regulatory regions and downstream elements

The upstream elements, the internal regulatory regions and the termination signals are essential in 5S rDNA transcription, but epigenetic mechanisms were also found to be involved in transcription regulation (Douet and Tourmente, 2007). A TATA-like motif located at around –30 to −25 nt is essential for efficient transcription *in vitro* in *Caenorhabditis elegans* and *C. briggsae* (Nelson *et al*., 1998), *Neurospora crassa* (Tyler, 1987) and *D. melanogaster* (Sharp and Garcia, 1988). In razor shells, the TATA-like −25 region that we found upstream the 5S rDNA transcription start site is likely to be analogous to that of the mentioned organisms. Among the 5S internal regulatory regions, the ICR II was the most conserved one in the comparisons with *D. melanogaster* ICRs.

The transcription termination signal of 5S rDNA has been studied in various organisms and seems to be quite conserved (a TTTT stretch). We have found this element in almost all razor shell NTSs, in agreement with previous findings in other eukaryotes (Bogenhagen and Brown, 1981; Huang and Maraia, 2001).

According to Marz *et al*. (2008), the classical snDNA-specific proximal sequence elements (PSE) and TATA boxes that have been described in detail for several vertebrates and were highly conserved (Hernandez, 2001; Domitrovich and Kunkel, 2003) are the exception rather than the rule, as the snDNA promoters are highly diverse across metazoa. In *Drosophila*, there are two elements essential for the efficient initiation of transcription of snDNA families transcribed by RNA polymerase II: they are the PSEA (−61 to −41 nt), analogous to the vertebrate PSE and the PSEB (from −32 to −25 nt; consensus sequence C/TATGGAA/GA, Lo and Mount, 1990) (Zamrod *et al*., 1993). In razor shells and gastropods, we have identified an A/G-rich region (−25 region; from −27 to −23 nts), which was conserved in location and quite conserved in sequence that could correspond to the PSEB. The region centred at around −40 nts upstream the U1 snDNA transcription start site was not very conserved, so it does not seem analogous to the PSEA/PSE. Interestingly, an AAAGC motif was found just upstream the U1 snDNA transcription start site of only some of our razor shell sequences (and not in the gastropod sequences). This pentanucleotide is shared with *D. melanogaster*, the slime mold *Physarum polycephalum* and some vertebrates (Lo and Mount, 1990 and reference therein). According to our data, it is not conserved in molluscs, but the occurrence of this motif in the same location and shared among distantly related taxa suggests it may have a function.

The internal regulatory regions within the U1 seemed to be somewhat more conserved than the ones within the 5S, as some of them were identical in the bivalve, gastropod, crustacean and insect species considered. Our data is consistent with the results by Marz *et al*. (2008), except in the positions 86, 131 and 132 (from the reference sequence, see Internal regulatory regions) that were not conserved in molluscs.

The snDNA transcription termination is more variable and different genes appear not to use a common process (Richard and Manley, 2009). For instance, transcription of human U1 snDNA terminates close to the 3′ box (Cuello *et al.*, 1999), but transcription of U2 snDNA extends about 800 sites beyond it (Medlin *et al*., 2003). The human 3′ box (Hernandez, 1985) is a 16 nt stretch, located 10 sites downstream the U1. In razor shells, we have not found a conserved region analogous to the 3′ box; however, the first five nts of the IGS were identical in *Ensis* species and *E. cultellus* and similar in *P. legumen*. Sequences from *S. patula* were somewhat different, but this could be related to the fact that their preceeding U1s were likely to be pseudogenised copies.

One or more linkage events throughout evolution?

In order to study whether the linkage happened once or more throughout the evolution of the Pharidae lineages, we have constructed several phylogenies and carefully studied the alignments performed. By mapping the 5Ss from mixed clones on the phylogenetic trees and on the network performed (Supplementary file S7), we tried to detect whether the linkage between the multigene families emerged once or more throughout the evolution of razor shells, but unfortunately, the phylogenies were not resolved.

The alignment of the IGS region supports that the linkage between both multigene families is homologous in these Pharidae species, with the exception of *E. macha* type A U1. In this case, as we did not sample its downstream IGS, we do not know how similar it would be compared with the other *Ensis* IGSs. The origin of this clone is unclear, as it could represent a new linkage between both multigene families, or it could be a descendant of the original linkage in which the NTS was replaced.

The most parsimonious explanation for our data is an evolutionary scenario in which the linkage happened only once, in a common ancestor to all the Pharidae species studied. Subsequently, there were duplications involving either the entire linked unit, or any of the RNA coding regions explaining why we have found the different genomic organisations recorded in Figure 1. Sequences started to accumulate mutations and diverged, but purifying selection and, perhaps, other homogenising mechanisms, maintained the integrity of the functional regions. Finally, different units continued to be duplicated and/or deleted across the different Pharidae lineages.

How 5S rDNA and U1 snDNA can become linked and why?

There are two possible alternatives for both multigene families to become linked: the linkage was the consequence of the insertion of one or more 5S rDNA repeats next to one or more U1 snDNA repeats, or vice-versa. However, how could this happen? Several surveys have suggested that rDNA (both 5S rDNA and the major ribosomal genes) frequently moves on from one location to another in the eukaryote genome (Rooney and Ward, 2005; Datson and Murray, 2006; Veltos *et al*., 2009; Nguyen *et al*., 2010), and several mechanisms have been proposed to explain this apparent mobility. Drouin and Moniz de Sá (1995) hypothesised that a 5S rDNA transposition could be produced at the DNA level mediated by extrachromosomal circular DNA or by an RNA intermediate. Interestingly, recent surveys have given support to both hypothesis (Kalendar *et al*., 2008; Cohen *et al*., 2010). Similarily, Rooney and Ward, (2005) hypothesised that 5S rDNA was capable of multiplying and integrating into other areas of the genome through a process the same as, or similar to, retroposition, in filamentous fungi. In a survey concerning the major ribosomal genes, it has been proposed that ectopic recombination (homologous recombination between repetitive sequences of nonhomologous chromosomes) was the primary motive force in the repatterning of these genes in lepidopteran species (Nguyen *et al*., 2010). Similar to what has been reported for rDNA, Marz *et al*., (2008) concluded that metazoan spliceosomal snDNA families behave like mobile genetic elements because they barely appear in syntenic positions, as measured by their flanking regions. Therefore, in theory, there are a few possible ways by which 5S rDNA and U1 snDNA could have become linked, but why?

Several examples have been reported in which 5S rDNA and U1 snDNA were found linked to each other or to other multigene families in virtually all eukaryote groups. Interestingly, the linkage between 5S rDNA and other multigene families have been repeatedly established and lost throughout evolution in several lineages, but this lack of conservation and the diversity of the linkages make it unlikely that they provide any selective advantage (for example, transcriptional co-regulation, (Drouin and Moniz de Sá, 1995). In the same way, Marz *et al*., (2008) concluded that tandem repeats of different spliceosomal snDNA families, or of a spliceosomal snDNA family and 5S rDNA, are not conserved over long evolutionary timescales in metazoans. So, even though the linkages between multigene families may provide a benefit that has not been reported yet, they rather seem to us to be the result of stochastic processes within genomes. The high copy number of 5S rDNA would make it quite likely to establish a linkage with another multigene family.

References

Bailey TL, Elkan C (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Second Int Conf Intell Syst Mol Biol* **2**: 28–36. AAAI Press, Menlo Park, California.

Barciszewska MZ, Szymanski M, Erdmann VA, Barciszewski J (2000). 5S Ribosomal RNA. *Biomacromolecules* **1**: 297–302.

Barzotti R, Pelliccia F, Rocchi A (2003). Identification and characterization of U1 small nuclear RNA genes from two crustacean isopod species. *Chromosome Res* **11**: 365–373.

Bogenhagen DF, Brown DD (1981*).* Nucleotide sequences in *Xenopus* 5S DNA required for transcription termination. *Cell* **24**: 261–270.

Branlant C, Krol A, Ebel JP, Lazar E, Haendler B, Jacob M (1982). U2 RNA shares a structural domain with U1, U4, and U5 RNAs. *EMBO J* **1**: 1259–1265.

Brehm K, Hubert K, Sciutto E, Garate T, Frosch M (2002). Characterization of a spliced leader gene and of trans-spliced mRNAs from *Taenia solium*. *Mol Biochem Parasitol* **122**: 105–110.

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

Cabral-de-Mello DC, Moura RC, Martins C (2010). Chromosomal mapping of repetitive DNAs in the beetle *Dichotomius geminatus* provides the first evidence for an association of 5S rRNA and histone H3 genes in insects, and repetitive DNA similarity between the B chromosome and A complement. *Heredity* **104**: 393–400.

Caradonna F, Bellavia D, Clemente AM, Sisino G, Barbieri R (2007). Chromosomal localization and molecular characterization of three different 5S ribosomal DNA clusters in the sea urchin *Paracentrotus lividus*. *Genome* **50**: 867–870.

Chen S-H, Su S-Y, Lo C-Z, Chen K-H, Huang T-J, Kuo B-H *et al*. (2009). PALM: a paralleled and integrated framework for phylogenetic inference with automatic likelihood model selectors. *PLoS ONE* **4(12)**: e8116.

Cohen S, Agmon N, Sobol O, Segal D (2010). Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. *Mobile DNA* **1**: 11.

Cosel von R (1993). The razor shells of the eastern Atlantic, part 1: Solenidae and Pharidae I. *Arch Moll* **122**: 207–321.

Cosel von R (2009). The razor shells of the eastern Atlantic, part 2. Pharidae II: the genus *Ensis* Schumacher, 1817 (Bivalvia, Solenoidea). *Basteria* **73**: 1–48.

Cross I, Rebordinos L (2005). 5S rDNA and U2 snRNA are linked in the genome of *Crassostrea angulata* and *Crassostrea gigas oysters*: does the (CT)n·(GA)n microsatellite stabilize this novel linkage of large tandem arrays? *Genome* **48**: 1116–1119.

Cuello P, Boyd DC, Dye MJ, Proudfoot NJ, Murphy S (1999). Transcription of the human U2 snRNA genes continues beyond the 39 box *in vivo*. *EMBO J* **18**: 2867–2877.

Datson PM, Murray BG (2006). Ribosomal DNA locus evolution in *Nemesia*: transposition rather than structural rearrangement as the key mechanism? *Chrom Res* **14**: 845–857.

Domitrovich AM, Kunkel GR (2003). Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies. *Nucleic Acids Res* **31**: 2344–2352.

Douet J, Tourmente S (2007). Transcription of the 5S rRNA heterochromatic genes is epigenetically controlled in *Arabidopsis thaliana* and *Xenopus laevis*. *Heredity* **99**: 5–13.

Drouin G, Moniz de Sá M (1995). The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. *Mol Biol Evol* **12**: 481–893.

Eickbush TH, Eickbush DG (2007). Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175**: 477–485.

Freire R, Arias A, Insua AM, Méndez J, Eirín-López JM (2010). Evolutionary dynamics of the 5S rDNA gene family in the mussel *Mytilus*: mixed effects of birth-and-death and concerted evolution. *J Mol Evol* **70**: 413–426.

Felsestein J (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Fernandez M, Ruiz ML, Linares C, Fominaya A, de la Vega MP (2005). 5S rDNA genome regions of *Lens* species. *Genome* **48**: 937–942.

Fernández-Tajes J, Méndez J (2009). Two different size classes of 5S rDNA units coexisting in the same tandem array in the razor clam *Ensis macha*: is this region suitable for phylogeographic studies? *Biochem Genet* **47**: 775–788.

Fujiwara M, Inafuku J, Takeda A, Watanabe A, Fujiwara A, Kohno S *et al*. (2009*)*. Molecular organization of 5S rDNA in bitterlings (Cyprinidae). *Genetica* **135**: 355–365.

Gascuel O (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14**: 685–695.

Giegerich R, Meyer F, Schleiermacher C (1996). GeneFisher—software support for the detection of postulated genes. *Proc Int Conf Intell Syst Mol Biol* **4**: 68–77.

Gornung E, Colangelo P, Annesi F (2007). 5S ribosomal RNA genes in six species of Mediterranean grey mullets: genomic organization and phylogenetic inference. *Genome* **50**: 787–795.

Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41**: 95–98.

Hernandez N (1985). Formation of the 3′ end of U1 snRNA is directed by a conserved sequence located downstream of the coding region. *EMBO J* **4**: 1827–1837.

Hernandez N (2001). Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. *J Biol Chem* **276**: 26733–26736.

Hofacker IL (2003). Vienna RNA secondary structure server. *Nucleic Acids Res* **31**: 3429–3431.

Huang Y, Maraia RJ (2001). Comparison of the RNA polymerase III transcription machinery in *Schizosaccharomyces pombe, Saccharomyces cerevisiae* and human. *Nucleic Acids Res* **29**: 2675–2690.

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

Kalendar R, Tanskanen J, Chang W, Antonius K, Sela H, Peleg O *et al*. (2008). *Cassandra* retrotransposons carry independently transcribed 5S RNA. *PNAS* **105**: 5833–5838.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al*. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Little RD, Braaten DC (1989). Genomic organization of human 5 S rDNA and sequence of one tandem repeat. *Genomics* **4**: 376–383.

Lo PCH, Mount SM (1990). *Drosophila melanogaster* genes for U1 snRNA variants and their expression during development. *Nucleic Acids Res* **18**: 6971–6979.

Manchado M, Zuasti E, Cross I, Merlo A, Infante C, Rebordinos L (2006). Molecular characterization and chromosomal mapping of the 5S rRNA gene in *Solea senegalensis*: a new linkage to the U1, U2, and U5 small nuclear RNA genes. *Genome* **49**: 79–86.

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

Mathews DH, Sabina J, Michael Zuker M (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**: 911–940.

Medlin JE, Uguen P, Taylor A, Bentley DL, Murphy S (2003). The C-terminal domain of pol II and a DRB-sensitive kinase are required for 3′ processing of U2 snRNA. *EMBO J* **22**: 925–934.

Morzycka-Wroblewska E, Selker EU, Stevens JN, Metzenberg RL (1985). Concerted evolution of dispersed *Neurospora crassa* 5S RNA genes: pattern of sequence conservation between allelic and nonallelic genes. *Mol Cell Biol* **5**: 46–51.

Mount SM, Gotea V, Lin C-F, Hernandez K, Makałowski W (2007). Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA* **13**: 5–14.

Nei M, Rooney AP (2005). Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**: 121–152.

Nelson DW, Linning RM, Davison PJ, Honda BM (1998). 5′-flanking sequences required for efficient transcription in vitro of 5S RNA genes, in the related nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Gene* **218**: 9–16.

Nguyen P, Sahara K, Yoshido A, Marec F (2010). Evolutionary dynamics of rDNA clusters on chromosomes of moths and butterflies (Lepidoptera). *Genetica* **138**: 343–354.

Pelliccia F, Barzotti R, Bucciarelli E, Rocchi A (2001). 5S ribosomal and U1 small nuclear RNA genes: a new linkage type in the genome of a crustacean that has three different tandemly repeated units containing 5S ribosomal DNA sequences. *Genome* **44**: 331–335.

Peterson RC, Doering JL, Brown DD (1980). Characterization of two *Xenopus* somatic 5S-DNAs and one minor oocyte-specific 5S-DNA. *Cell* **20**: 131–141.

Pieler T, Hamm J, Roeder RG (1987). The 5S gene internal control region is composed of three distinct sequence elements, organized as two functional domains with variable spacing. *Cell* **48**: 91–100.

Posada D, Crandall KA (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.

Query CC, Bentley RC, Keene JD (1989). A specific 31-nucleotide domain of U1 RNA directly interacts with the 70K small nuclear ribonucleoprotein component. *Mol Cell Biol* **9**: 4872–4881.

Reuter JS, Mathews DH (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**: 129.

Richard P, Manley JL (2009). Transcription termination by nuclear RNA polymerases. *Genes Dev* **23**: 1247–1269.

Rooney AP, Ward TJ (2005). Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. *PNAS* **102**: 5084–5089.

Scherly D, Boelens W, van Venrooij WJ, Dathan NA, Hamm J, Mattaj IW (1989). Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA. *EMBO J* **8**: 4163–4170.

Sharp S, Garcia A, Cooley L, Söll D (1984). Transcriptionally active and inactive gene repeats within the *D. melanogaster* 5S RNA gene cluster. *Nucleic Acids Res* **20**: 7617–7632.

Sharp SJ, Garcia AD (1988). Transcription of the *Drosophila melanogaster* 5S RNA gene requires an upstream promoter and four intragenic sequence elements. *Mol Cell Biol* **8**: 1266–1274.

Shippen-Lentz DE, Vezza AC (1988). The three 5S rRNA genes from the human malaria parasite *Plasmodium falciparum* are linked. *Mol Biochem Parasit* **27**: 263–273.

Swofford DL (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version.* Sinauer Associates, Sunderland, MA, USA.

Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: molecular Evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596–1599.

Tyler BM (1987). Transcription of *Neurospora crassa* 5 S rRNA genes requires a TATA box and three internal elements. *J Mol Biol* **196**: 801–811.

Úbeda-Manzanaro M, Merlo MA, Palazón JL, Sarasquete C, Rebordinos L (2010). Sequence characterization and phylogenetic analysis of the 5S ribosomal DNA in species of the family Batrachoididae. *Genome* **53**: 723–730.

Veltos P, Keller I, Nichols RA (2009). Geographically localised bursts of ribosomal DNA mobility in the grasshopper *Podisma pedestris. Heredity* **103**: 54–61.

Vierna J, González-Tizón AM, Martínez-Lage A (2009). Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia). *Biochem Genet* **47**: 635–644.

Vierna J, Martínez-Lage A, González-Tizón AM (2010). Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers. *Genome* **53**: 23–34.

Will CL, Lührmann R (2005). Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem* **386**: 713–724.

Zamrod Z, Tyree CM, Song Y, Stumph WE (1993). *In vitro* transcription of a *Drosophila* U1 small nuclear RNA gene requires TATA box-binding protein and two proximal cis-acting elements with stringent spacing requirements. *Mol Cell Biol* **13**: 5918–5927.

Zhuang Y, Weiner AM (1986). A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. *Cell* **46**: 827–835.

Acknowledgements

We thank Jane Frydenberg and Camilla Håkansson, who were very kind and helpful during lab work in Aarhus Manja Marz for providing gastropod U1 snDNA sequences, her support during bioinformatic analyses and her comments on the paper; Rudo von Cosel for his help regarding razor shell taxonomy and the identification of many of the specimens studied in this survey; Miguel Vizoso for his comments on the manuscript; and Klaus Brehm for providing information about the *Taenia* spliced leader sequence. This work would have not been possible without the help of the following colleagues (alphabetical order) who kindly provided razor shell specimens: Dan Ayres, Emile Egea, John Havenhand, Iben Heiner, Inés Naya, Lobo Orensanz, Roberto Portela-Míguez, Anja Schulze, John Taylor, Ray Thompson, and Katrine Worsaae. JV has been supported by a 'María Barbeito' fellowship and a travel grant, both from the *Consellería de Economía e Industria*, *Xunta de Galicia* (Spain) and the European Social Fund.

Conflict of interest

The authors declare no conflict of interest.

[i] Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)

[ii] jvierna@udc.es

[iii] hakuna@udc.es