Published online 4 April 2008

Nucleic Acids Research, 2008, Vol. 36, No. 9 ***3001–3010***
doi:10.1093/nar/gkn142

Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components

Marcela Dávila López[1], Magnus Alm Rosenblad[2] and Tore Samuelsson[1,*]

[1]Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy, Box 440 and [2]Department of Cell and Molecular Biology, University of Gothenburg, Box 462, SE-405 30 Göteborg, Sweden

Received February 8, 2008; Revised March 13, 2008; Accepted March 14, 2008

ABSTRACT

The RNA molecules of the spliceosome are critical for specificity and catalysis during splicing of eukaryotic pre-mRNA. In order to examine the evolution and phylogenetic distribution of these RNAs, we analyzed 149 eukaryotic genomes representing a broad range of phylogenetic groups. RNAs were predicted using high-sensitivity local alignment methods and profile HMMs in combination with covariance models. The results provide the most comprehensive view so far of the phylogenetic distribution of spliceosomal RNAs. RNAs were predicted in many phylogenetic groups where these RNA were not previously reported. Examples are RNAs of the major (U2-type) spliceosome in all fungal lineages, in lower metazoa and many protozoa. We also identified the minor (U12-type) spliceosomal U11 and U6atac RNAs in *Acanthamoeba castellanii*, where U12 spliceosomal RNA as well as minor introns were reported recently. In addition, minor-spliceosome-specific RNAs were identified in a number of phylogenetic groups where previously such RNAs were not observed, including the nematode *Trichinella spiralis*, the slime mold *Physarum polycephalum* and the fungal lineages Zygomycota and Chytridiomycota. The detailed map of the distribution of the U12-type RNA genes supports an early origin of the minor spliceosome and points to a number of occasions during evolution where it was lost.

INTRODUCTION

An essential step of gene expression in eukaryotes is the removal of introns from the pre-mRNA and the ligation of exons to form the mature RNA. It occurs by two sequential *trans*-esterification reactions and is catalyzed by a multicomponent complex, the spliceosome (1). To date, two intron classes are known, a U2-type and a low-abundance U12-type. Splicing of U2-type introns is catalyzed by the U2-dependent (major) spliceosome, which includes the U1, U2, U4, U5 and U6 spliceosomal RNAs as well as multiple protein factors. The U12-dependent (minor) spliceosome, responsible for the excision of the U12-type introns, is structurally similar to the U2-type spliceosome. It contains protein subunits and the U5 RNA as well as the U11, U12, U4atac and U6atac spliceosomal RNAs that are functionally and structurally related to the U1, U2, U4 and U6 RNAs of the major spliceosome.

All spliceosomal RNAs, except U6 and U6atac, are synthesized by Pol II (2) and contain a conserved single-stranded region, referred to as the Sm site, with the consensus $PuAU_{4-6}GPu$ that is normally flanked by two hairpins and serves as the binding site for the Sm proteins (3).

For U2-type introns, spliceosome assembly is initiated by the interaction of U1 snRNP with the 5′ splice site and U2 snRNP with the branch site. Here, the U1 and U2 RNAs play important roles as they pair with 5′ splice site and branch site sequences, respectively. A U4–U5–U6 tri-snRNP complex, where U4 and U6 RNA are associated by base-pairing, associates with U1, U2 and the pre-mRNA to form a spliceosome. Structural rearrangements then take place such that U6 separates from U4 to allow pairing between U6 and U2. U6 also interacts with the 5′ splice site and U1 is displaced from the spliceosome. The U6/U2 complex plays an important role in the catalytic reaction (4). Assembly of the U12-dependent spliceosome is similar to that of the U2-dependent spliceosome but a major difference is that U11 and U12 snRNPs form a highly stable di-snRNP that binds cooperatively to the 5′ splice site and branch site (5,6).

U2-type introns are ubiquitous in eukaryotes while U12-type introns have so far been demonstrated only in vertebrates, insects, cnidarians (7), *Rhizopus oryzae*, *Phytophthora* and *Acanthamoeba castellanii* (8). They are

*To whom correspondence should be addressed. Tel: +46 31 786 3468; Fax: +46 31 41 6108; Email: Tore.Samuelsson@medkem.gu.se

absent from the yeast *Saccharomyces cerevisiae* (9) and from the nematode *Caenorhabditis elegans* (7). In order to understand the evolution of the splicing machinery and of spliceosomal RNAs, we wanted to systematically examine the phylogenetic distribution of these RNAs. In general ncRNAs are poorly conserved in sequence but each class of ncRNA is typically characterized by a specific secondary structure. This is also true for spliceosomal RNAs, although many spliceosomal RNAs are conserved also in sequence, like U2 and U6 RNAs (10). Nevertheless, for some spliceosomal RNAs the primary sequence is highly variable. In the case of U1 RNA also the secondary structure is subject to variation, as observed in yeast (11) and in *Trypanosoma* (12). Therefore, the computational identification of spliceosomal RNA genes, as with many other noncoding RNA genes, is challenging. A large number of spliceosomal RNAs from different organisms have been identified experimentally as well as computationally (13) and have been deposited in sequence databases. For instance, a large number of spliceosomal RNA sequences are available in the Rfam database (13), aimed at prediction of ncRNAs using covariance models (14). However, there are phylogenetic groups where spliceosomal RNAs have not been identified and it is not clear whether this is due to poor performance of prediction methods or because such RNAs are lacking in these organisms. In order to improve on this situation we have developed a simple protocol for computational identification of spliceosomal RNA, based on local alignment methods, profile HMMs and covariance models (14). Our method is efficient as we are able to present a large number of previously unrecognized spliceosomal RNA orthologues.

MATERIALS AND METHODS

Sources of genomic and protein sequences

Genomic sequences were obtained from NCBI (http://www.ncbi.nlm.nih.gov/entrez/; ftp.ncbi.nih.gov/genomes), EMBL (http://www.ebi.ac.uk), ENSEMBL (http://www.ensembl.org), TraceDB (ftp.ncbi.nlm.nih.gov/pub/TraceDB), TIGR (ftp://ftp.tigr.org/pub/data/), the U.S. Department of Energy Joint Genome Institute (http://www.jgi.doe.gov), the WU Genome Sequencing Center (http://genome.wustl.edu/), the Sanger Institute (http://www.sanger.ac.uk), the HGSC at Baylor College (http://www.hgsc.bcm.tmc.edu/projects/) as well as specific Genome Project Databases: CryptoDB (http://www.cryptodb.org/cryptodb/), PlasmoDB (http://www.plasmodb.org), GiardiaDB (http://www.jbpc.mbl.edu/Giardia-HTML/index2.html), ToxoDB (http://www.toxodb.org/toxo/home.jsp), DictyBase (http://dictybase.org/), the *Cyanidioschyzon merolae* Genome Project (http://merolae.biol.s.u-tokyo.ac.jp) and the *Galdieria sulphuraria* Genome Project (http://genomics.msu.edu/galdieria/). Access to the provisional 4× assembly of *Mucor circinelloides* genome was granted by the DoE Joint Genome Institute and the Mucor genome project (http://mucorgen.um.es/). More details on database versions are in Supplementary Data 4. Protein sequences were retrieved from Uniprot (http://beta.uniprot.org/).

Identification of spliceosomal RNA orthologues

Sequences of RNAs annotated as spliceosomal RNAs (U1, U2, U4, U5, U6, U11, U12, U4atac and U6atac) were assembled (Supplementary Data 1) from Rfam (13). These sequences were used as initial queries with BLASTN (15) and FASTA (16) against genomic sequences of the organisms listed in Supplementary Data 4. The *E*-value threshold was set to 10, while the word size was 7 and 6 for BLAST and FASTA, respectively. Hits including 200 nt upstream and downstream sequences were retrieved and analyzed with cmsearch of the Infernal package (14) using the relevant covariance model from Rfam. For yeast U1 RNA the Rfam model specific to this group was used.

A threshold was then set for each one of the spliceosomal RNAs that was based on the initial query RNA giving rise to the lowest score from cmsearch. These threshold values were for U1, U2, U4, U5, U6, U11, U12, U4atac and U6atac 55.74, 52.75, 40.63, 60.60, 54.24, 36.80, 51.74, 40.03 and 39.23, respectively. All sequences above these threshold values were considered as reliable predictions. For species where sequences with scores above threshold were found, all sequences below the threshold were discarded. For the remaining species sequences with a score below the threshold but greater than 15 were considered for further analysis. Considering relatively low scores was in this case motivated by the fact that Rfam covariance models tend to be phylogenetically biased as sequences from mammals and other well studied species are overrepresented. Relatively few sequences (between 1% and 12%, depending on the RNA family) belonged to this category of low-scoring sequences. They were evaluated using a procedure where the presence of specific conserved primary sequence motifs as well secondary structure was examined. We required exact matches to the primary sequence motifs listed in Supplementary Data 4. The cmsearch output was used to produce structure plots based on the covariance model used. These plots were manually browsed to verify the presence of secondary structure elements according to the consensus secondary structure of the specific spliceosomal RNA. If relevant primary and secondary structure features were present the sequences were considered as reliable predictions.

The resulting predicted sequences were then used as queries in a second round of searches to retrieve homologues in species where that particular RNA orthologue was not identified. The resulting hits were analyzed as described earlier and any reliable predictions obtained were used in yet another round of searches. This procedure was repeated until no more significant hits were retrieved.

In species where we were not able to find a reliable spliceosomal RNA, WU-BLAST blastn was used with word size 2. Sequences identified in such searches were analyzed with the respective covariance model and using the same criteria as described earlier in order to identify reliable predictions, which were used in a second round of searches. In addition, we performed hmmsearch searches

using HMM models based on Rfam alignments against the set of reliable spliceosomal RNA sequences and against genomes where we did not previously find a specific RNA. Sequences with *E*-values lower than 10 were retrieved and examined with cmsearch and processed as described earlier. All 17136 sequences considered as reliable candidates identified in this study, together with the 356 sequences used as initial queries, are in Supplementary Data 1.

Multiple alignments of sequences were created using ClustalW 1.83 (17) or T-Coffee (18). Alignments obtained with cmalign of Infernal are shown in Supplementary Data 3. Secondary structure was predicted with MFOLD (19) as well as with Infernal.

RESULTS AND DISCUSSION

Spliceosomal RNAs may be efficiently identified using a combination of high-sensitivity local alignment methods, profile HMMs and covariance models

Genomic sequences from 149 eukaryotic organisms (Figure 1) were analyzed with respect to spliceosomal RNAs. The Infernal software (14) to identify ncRNAs using covariance models is effective to identify members of a specific RNA family but is computationally demanding and not practical for the analysis of large genomes. Therefore, a first step to filter sequences is necessary. In our method, we used NCBI BLAST (wordsize 7) and FASTA (wordsize 6). RNA sequences represented in the Rfam database and annotated as spliceosomal RNAs (U1, U2, U4, U5, U6, U11, U12, U4atac and U6atac) were first assembled (a total of 356 sequences, see Supplementary Data 1) and used as queries in BLAST and FASTA searches against genomic sequences. To maximize sensitivity, we considered all hits, irrespective of *E*-value, identified in these searches for analysis with covariance models of spliceosomal RNAs collected from Rfam.

The predictions that represented novel spliceosomal RNA sequences and that were considered reliable (see under Materials and Methods section for details) were used in a second round of searches to search against organisms where we were missing the respective spliceosomal RNA. Novel hits were analyzed as described with covariance models and this procedure was repeated until no further reliable predictions could be obtained. For genomes where we were not able to find a spliceosomal RNA orthologue using BLAST or FASTA we also made use of WU-BLAST (wordsize 2) [Gish, W. (1996–2004) http://blast.wustl.edu/] and HMMER searches (http://hmmer.janelia.org/) using profile HMM models based on Rfam alignments. For comparison, we also used Infernal to analyze the following genomes with selected covariance models without any initial filtering of sequences; *Trichinella spiralis* (U11, U12 and U4atac), *A. castellanii* (U11, U4atac), *G. sulphuraria* (U1, U2, U4, U5 and U6), *Giardia lamblia* (U1, U2, U4, U5 and U6), *Physarum polycephalum* (U11, U12 and U4atac), *Naegleria gruberi* (U1), *Trichomonas vaginalis* (U1), *Batrachochytrium dendrobatidis* (U1, U11 and U12), *Antonospora locustae* (U1), *Encephalitozoon cuniculi* (U1), *Phycomyces blakesleeanus* (U12) and *Cyanidioschyzon merolae* (U1, U2, U4, U5 and U6).

We thus obtained 17136 sequences predicted as spliceosomal RNAs. All these sequences are distributed among 147 species as shown in Figure 1 and Supplementary Data 1 and 2. It should be noted that many animals and plants have numerous copies of each RNA gene and a fraction of these are fragmented genes or pseudogenes. As it is very difficult to distinguish a true gene from a pseudogene using computational methods a fraction of our candidates in animals and plants are presumably pseudogenes. In some phylogenetic groups such as fungi, heterokonts and Apicomplexa each of the spliceosomal RNAs are represented by one or a few genes and in this case the predicted sequences are more likely to be bona fide spliceosomal RNA genes.

The results using the different methods NCBI BLAST, FASTA, WU-BLAST and HMMER are compared in Figure 2. As expected, the sensitivity of FASTA, WU-BLAST and HMMER was much greater than that of NCBI BLAST (W7) and HMMER is the most sensitive method of the four. As there are speed disadvantages to WU-BLAST and HMMER, we did not systematically examine every possible genome with these methods. Instead we searched with these methods only genomes where we did not find a specific spliceosomal RNA with FASTA or BLAST. We also analyzed with WU-BLAST and HMMER all the RNA sequences found using FASTA and BLAST. In the results shown in Figure 2, therefore, the efficiency of WU-BLAST and HMMER is probably underestimated. In general, the results obtained in our searches are consistent with previous results (20) where different software, including programs used here, were tested against a set of previously known ncRNAs.

In summary, our results demonstrate that sensitive sequence alignment methods, including profile HMMs, are important as a first filtering step to identify ncRNA candidates. In addition, a combination of the different methods maximizes sensitivity. There are two RNAs, from *G. sulphuraria* and *A. locustae* (Figure 1), that could only be identified using an Infernal search against the complete genome. This finding illustrates that we could be lacking more orthologues not identified by HMMER, WU-BLAST or FASTA. However, it is our impression that very few RNA genes are missed by the initial screen.

During the production of this article, spliceosomal RNA sequences from *Plasmodium* (21), *Entamoeba histolytica* (22) as well as *Candida albicans* and other hemiascomycetous yeasts (11) were identified and characterized. All these sequences are identical to the sequences identified in the present study, providing support to the reliability of our prediction method.

RNAs of the major spliceosome

The spliceosomal RNAs as identified here are summarized in Figure 1 (actual sequences are in Supplementary Data 1). In 107 species, we are able to report one or more of spliceosomal RNAs where that particular RNA had not been reported before, as highlighted in the Figure 1 with blue boxes. As a guide to the phylogenetic relationships

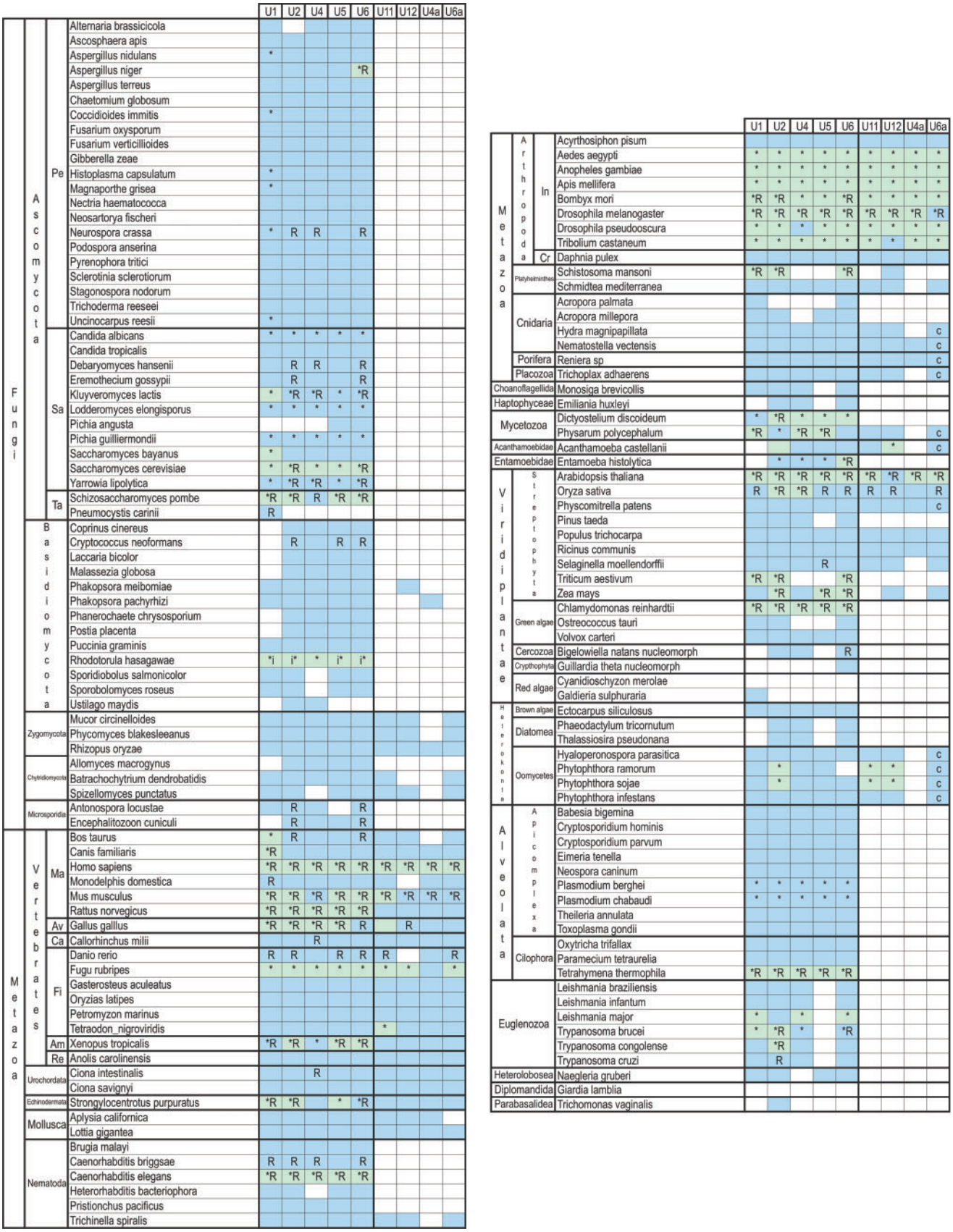

Figure 1. Phylogenetic distribution of the major and minor spliceosomal RNAs. Results of computational prediction of spliceosomal RNAs. Green boxes show instances where a sequence was previously known and it was used as query in searches. Blue boxes show RNAs predicted in this work. A star (*) indicates that an RNA was previously described in the literature (Supplementary Data 5) and 'R' shows that the sequence was in Rfam (13). Also indicated are sequences known to have introns ('i') and U6atac RNAs of the CC variant type ('c', see Results and Discussion section).

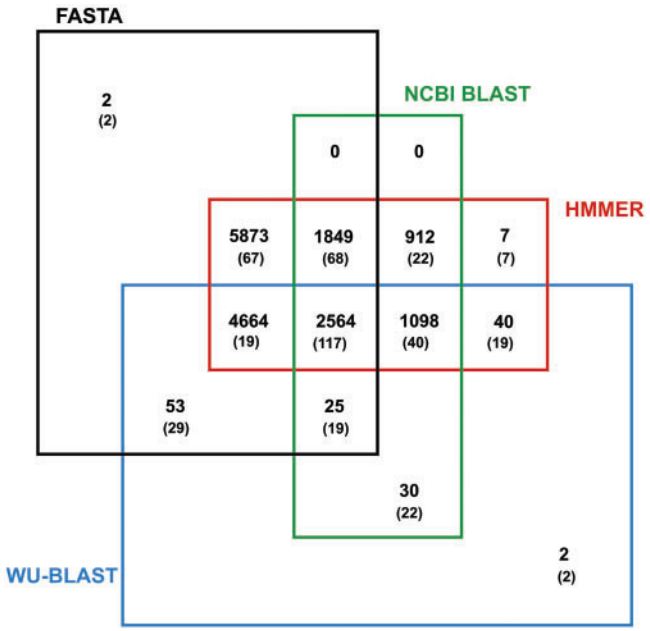

Figure 2. Comparison of local alignment and profile HMM methods to identify spliceosomal RNAs. Venn diagram showing the number of spliceosomal RNA genes found by NCBI BLAST (W7), WU-BLAST (W2), FASTA (W6) and HMMER. The number of species where the RNAs are distributed is shown within parentheses.

between the species investigated here, a schematic phylogenetic tree is shown in Figure 3.

We have analyzed RNAs of the U2-type spliceosome as well as those of the minor U12-type spliceosome (Figures 1 and 3). As to the U2-type, we have identified such RNAs in virtually every species examined. The only exceptions are the red alga *C. merolae* and the deeply branching protist *G. lamblia* where we could not identify any spliceosomal RNAs. Of spliceosomal RNAs in *T. vaginalis* only the U2 RNA was identified. *T. vaginalis* possesses a gene encoding the essential spliceosomal component PRP8 (23) as well as many putative introns (24). *G. lamblia* possesses three introns to date (25,26) and ~27 spliceosomal proteins (27). In *C. merolae*, introns as well as conserved U2 and U5 snRNP-protein specific subunits are known to be present (28). Therefore, splicing is likely to occur in these organisms, and it is puzzling that we fail to identify spliceosomal RNAs, particularly in *C. merolae*, where the genome sequence is complete (29). This could mean that spliceosomal RNAs are lacking and have been replaced by protein functions. But these organisms could also have spliceosomal RNAs very different from most other species, or these genes could be present in a part of the genome for some reason not yet covered by the genome sequencing. A U1 RNA was the only spliceosomal RNA that we identified in *G. sulphuraria*, another red alga. Additional spliceosomal RNAs might be found once its genome is fully sequenced.

RNA components of the major spliceosome are known to be present in fungi and recently the evolution of such RNAs in the hemiascomycetous yeasts was examined (11). Major spliceosomal RNAs have previously been reported in the Basidiomycota *Rhodotorula* (30,31) and *Cryptococcus neoformans* (Rfam). Here, we show that such RNAs are ubiquitous in the Basidiomycota lineage (Figure 1). More deeply branching in the fungi tree are Zygomycota and Chytridiomycota (Figures 1 and 3). We show for the first time that spliceosomal RNAs are present in the Zygomycota *P. blakesleeanus*, *R. oryzae* and *M. circinelloides*, as well as in the Chytridiomycota *B. dendrobatidis*, *Spizellomyces punctatus* and *Allomyces macrogynus*.

The microsporidia are believed to be positioned close to the root of the fungal branch. They have been reduced severly in genome size as compared to other fungi. In *A. locustae* and *E. cuniculi* U2 and U6 orthologues are reported in Rfam. Here, we have also identified the U4 and U5 RNA orthologues in both of these Microsporidia (Figure 1) and U1 RNA in *A. locustae* (Supplementary Data 3). We may therefore conclude that the major spliceosomal RNAs are ubiquitous in all fungal groups, including Microsporidia.

The nucleomorphs of *Guillardia theta* (a cryptomonad) and *Bigelowiella natans* (a chlorarachniophyte) represent the smallest eukaryotic genomes known. It is interesting to note that also in these two genomes spliceosomal RNA genes are identified (Figure 1). With respect to *G. theta*, the results of our predictions (only a U6 RNA) are completely consistent with available annotation (Douglas *et al.*; John Archibald and Paul Gilson, personal communication), whereas there are differences with respect to published annotation for the *B. natans* genome (32).

Minor-spliceosome-specific RNAs are identified in the worm *Trichinella spiralis*, in *Physarum polycephalum* and in the fungal lineages Zygomycota and Chytridiomycota

U12-type introns were previously identified in plants, in most of the metazoan taxa including vertebrates, insects and cnidarians (7), and more recently in *R. oryzae* of Zygomycota, in *Acanthamoeba* and in the heterokont *Phytophthora* (8). Minor spliceosomal RNAs have been found in metazoa, *Acanthamoeba*, plants and *Phytophthora* (for references see Supplementary Data 5). A small number of organisms that have been well studied seem to lack the U12-type splicing, such as *S. cerevisiae, S. pombe and C. elegans* (7,33).

In this investigation, we discovered many novel minor spliceosomal RNA orthologues (Figure 1). More importantly, phylogenetic groups are represented where such RNAs were not previously reported. These are nematodes (*T. spiralis*), mycetozoa (*P. polycephalum*) and the fungal lineages Basidiomycota, Zygomycota and Chytridiomycota as discussed in more detail subsequently.

Trichinella spiralis. The major spliceosomal RNAs of the nematode *C. elegans* have been characterized (34). Previous analyses have failed to identify minor spliceosomal components, including U12-type introns, in this organism (7). In this investigation, we analyzed different species of the Rhabditida branch and *Brugia malayi* of the Chromadorea branch. In neither of these species U12-type RNAs were identified. However, we identified

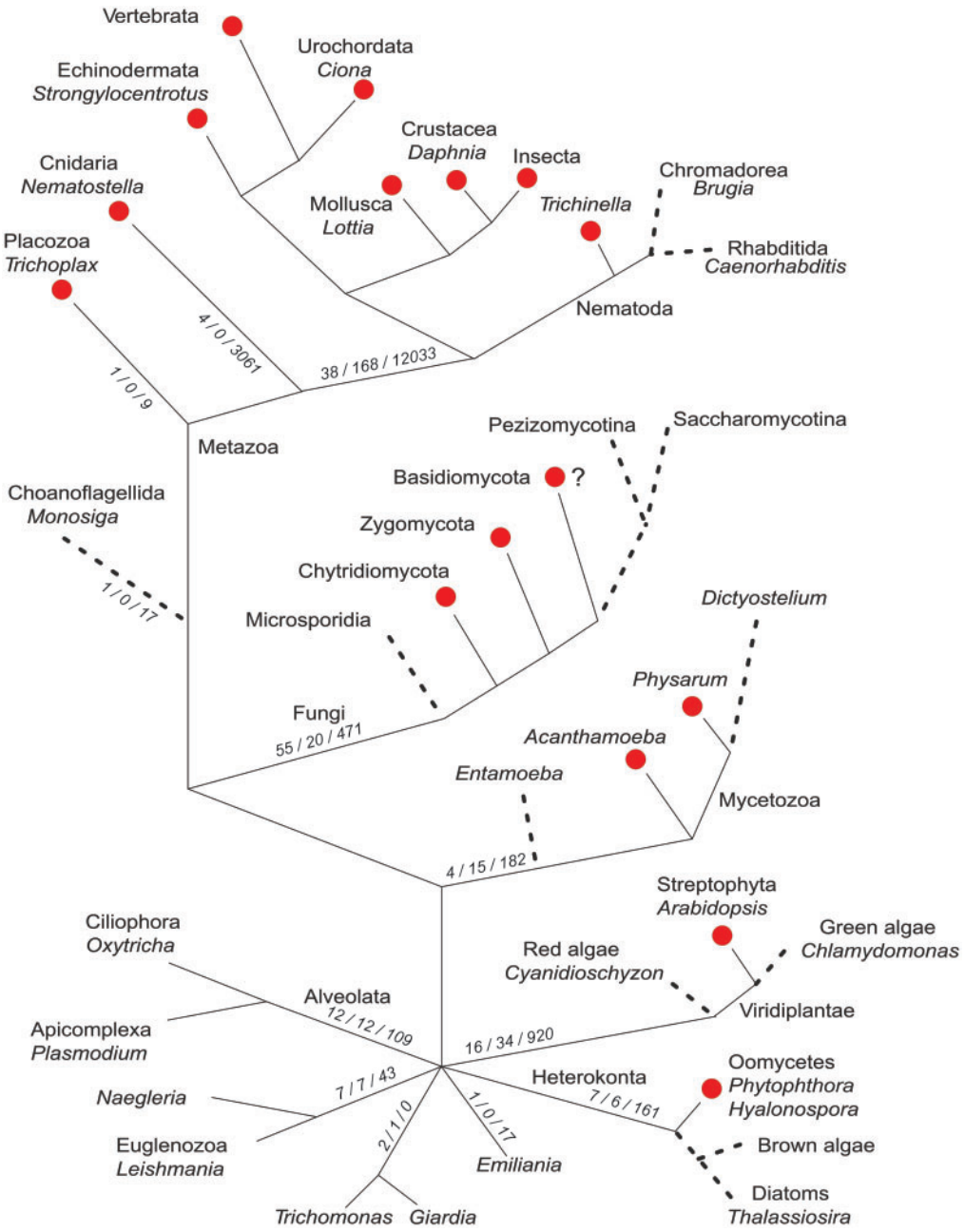

Figure 3. Schematic phylogenetic tree. Phylogenetic groups and their relationships are shown together with example species (genus in italics). Species where one or more U12-type spliceosomal RNAs were found are highlighted (red circles) as well as branches where the U12-type RNAs seem to have been lost (dotted lines). In the case of Basidiomycota, only two different U12-type RNAs have been identified and for this reason there is only weak evidence of a minor spliceosome. Numbers at branches indicate 1) number of genomes analyzed, 2) number of query sequences used and 3) number of new sequences identified.

U11, U12 and U6atac RNAs in another nematode, *T. spiralis* (Figure 1). Predicted secondary structures of these RNAs are shown in Figure 4. We also found evidence of U11/U12 specific proteins in *T. spiralis* (Supplementary Data 4), providing further support of a minor spliceosome in this organism.

Basidiomycota, Zygomycota and Chytridiomycota. No minor spliceosomal components have been described in fungi except for minor spliceosomal proteins and potential U12-type introns in *R. oryzae* (8), a species in the fungal Zygomycota lineage. However, we here identified minor spliceosomal RNA components in the Zygomycota *P. blakesleeanus*, *R. oryzae*, *M. circinelloides* and in the Chytridiomycota *B. dendrobatidis*, *S. punctatus* and *A. macrogynus*. Secondary structure predictions of *R. oryzae* RNAs are shown in Figure 4 and further structures are shown in Supplementary Data 3. These results provide strong evidence of a U12-type spliceosome in these phylogenetic groups.

In the Basidiomycota phylum, there was previously no evidence of a U12-type spliceosome. However, we

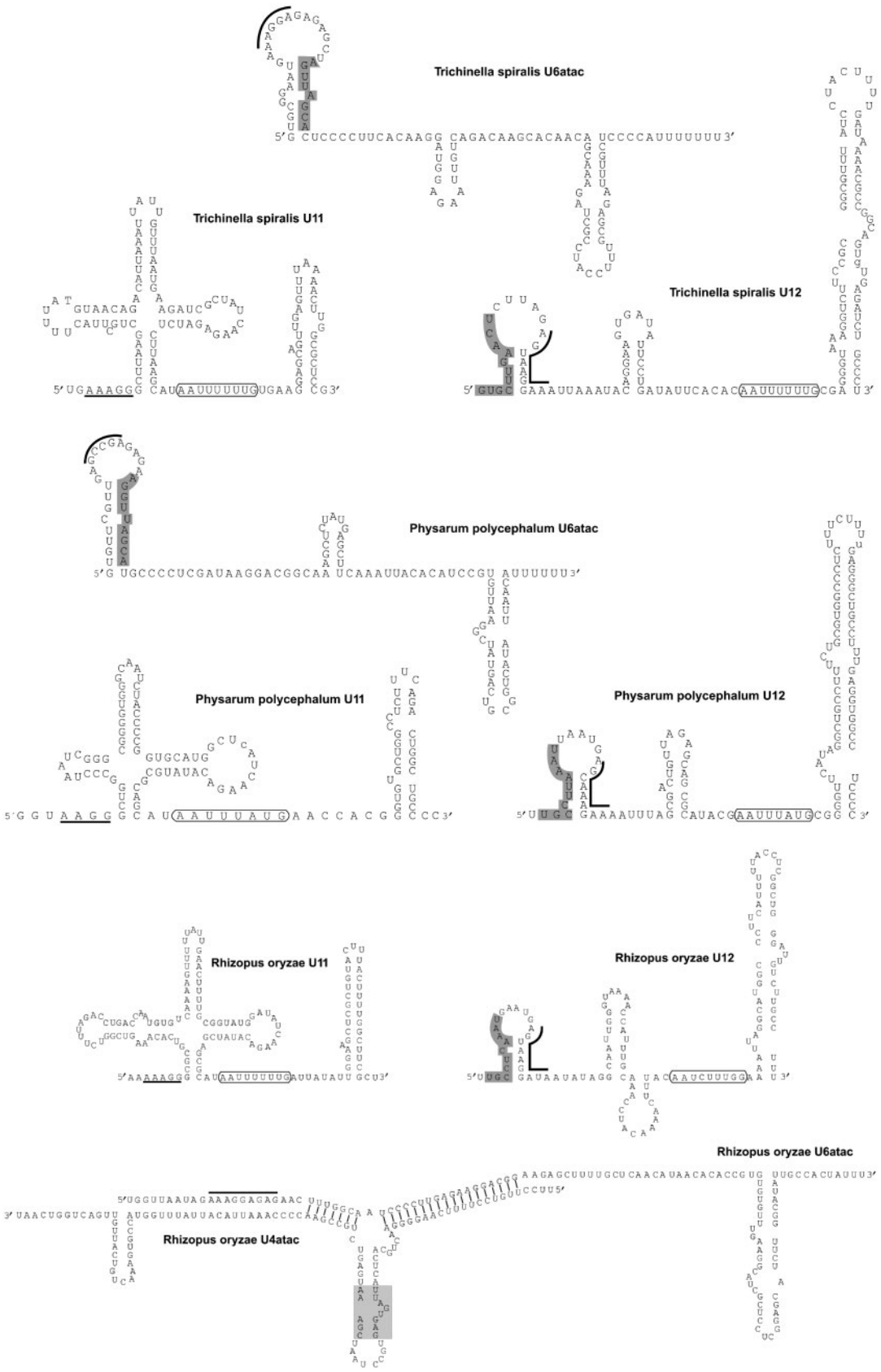

Figure 4. Structures of selected spliceosomal RNAs. Highlighted regions are the U12 site pairing to the branch site (underlined), regions of U11 and U6atac proposed to pair to the 5′ splice (underlined), U6atac-U12 interaction (shaded background), Sm-site (box with rounded corners) and K-turn (box) in U4atac RNA. Organisms represented are *T. spiralis*, *P. polycephalum* and *R. oryzae*. Structures of additional RNAs are in Supplementary Data 3.

here identified a U12 RNA in *Phakopsora meibomiae* and a U4atac RNA in *Phakopsora pachyrhizi*. At the same time, there is so far no evidence of U12-type introns or of U12-specific proteins. Therefore, it is possible that the spliceosomal RNAs that we observe are pseudogenes and remnants from a U12 machinery that was present in an ancestral lineage.

Acanthamoeba castellanii and Physarum polycephalum. In *A. castellanii* a U12 spliceosomal RNA as well as minor introns have been reported recently (8). These observations provided evidence of a minor spliceosome in this organism. Consistent with these results we identified two additional U12-type RNAs, U11 and U6atac, in this species (Figure 1).

Acanthamoeba is believed to share a common ancestor with the Mycetozoa, where spliceosomal components or U12-type introns were not previously reported. However, we here identified U11, U12 and U6atac spliceosomal RNA genes in *P. polycephalum* (Figure 1 and Supplementary Data 1).

The minor spliceosome was lost at multiple instances during evolution

The fact that we identified U12-type RNAs in a range of species representing very diverse phyla, such as Fungi, *Acanthamoeba*/Mycetozoa, Streptophyta and Heterokonta support the notion that the minor spliceosome was an early invention in eukaryotic evolution (8). In fact, we cannot exclude that such a spliceosome was present in the last common ancestor of the eukaryotes. The detailed map of the phylogenetic distribution of U12-type RNAs also allows us to identify a number of occasions during evolution where it seems that the minor spliceosome was lost (Figure 3, dashed lines).

U12-type RNAs were found in *T. spiralis* but not in the other nematodes examined here. *T. spiralis* belongs to a clade that is probably deeply branching within nematodes and is distant to the Rhabditida and Chromadorea groups of nematodes (35). A mode of evolution therefore seems likely where the minor spliceosome was present at an early stage in nematode evolution but was lost in many branches (Figure 3).

There are other examples where the U12-type RNAs are missing in the fungi/metazoan lineage. Minor spliceosomal RNAs are present in a majority of metazoa, including *Trichoplax*, the simplest known species of the metazoan branch. An exception is *Acropora millepora*, a coral of the phylum Cnidaria. In addition, we failed to identify such RNAs in *Monosiga brevicollis*, a choanoflagellate and close relative of Metazoa. A minor spliceosome is probably present in the fungal phyla Zygomycota and Chytridiomycota as discussed earlier. In Ascomycota and Microsporidia on the other hand, these components seem to be lacking. It is likely that a minor spliceosome was present at an early stage in the evolution of fungi, but was lost in the development of Ascomycota and Microsporidia. It is not clear why the minor spliceosome was lost in the Ascomycota but in the case of Microsporidia it could be a consequence of the strong pressure to reduce genome size.

We identified minor spliceosome-type RNAs in *P. polycephalum* and *A. castellanii* but not in the evolutionary related *Entamoeba* or *Dictyostelium* (Figure 3). This would suggest that the minor spliceosome was lost in the development of *Entamoeba* as well as in the *Dictyostelium* branch.

The analysis of Streptophyta (plant) genomes revealed the presence of minor spliceosomal RNAs, whereas only U2-dependent spliceosomal RNAs were found in green and red algae. Finally, in Heterokonta we found U12-type RNAs in the Oomycetes *Phytophthora* and *Hyalonospora* but not in any diatoms or brown algae.

In summary, our results point to a large number of instances where the minor spliceosome was lost during evolution of fungi/metazoa, Mycetozoa, Streptophyta and heterkonts. We are not able to reach a conclusion as to other phyla such as Euglenozoa and Alveolata because we do not know whether the common ancestor to these lineages had a minor spliceosomal machinery.

All U4 and U4atac RNAs have a K-turn motif

K-turn motifs have previously been identified in a large number of RNA families (36,37), including the U4 and U4atac RNAs (38,39). We found such a motif in all novel U4 and U4atac RNAs reported here. Examples are *R. oryzae* (Figure 4) and *Phakopsora* U4atac RNA (Supplementary Data 3) with characteristic noncanonical G-A and A-G pairs and a 3-nt loop. This would suggest that this motif is compulsory in U4 RNAs and that prediction accuracy may be improved by updating the covariance model in this respect.

Identification of a large number of novel U6atac orthologues

The U6atac RNA was previously identified in vertebrates, insects and plants (for references see Supplementary Data 5). Here, we identified orthologues in a majority of metazoan species, in Zygomycota, *Physarum*, *Acanthamoeba* and in Oomycetes (Heterokonta). A multiple alignment of U6atac sequences was constructed and selected sequences from that alignment are shown in Figure 5.

U6atac RNA pairs with U12 as well as with U4atac RNA (40). The novel sequences of U12, U6atac and U4atac that we have identified are consistent with these base-pairing interactions. Thus, the U6atac and U4atac sequences are all consistent with the formation of the stems 1 and 2 in the complex of these RNAs (Figure 4). In addition, the U6atac/U12 helices 1a and 1b are phylogenetically supported.

A sequence 'AAGGA' near the 5′ end of U6atac has been proposed to pair with a region at the 5′ splice site (40,41). This sequence is present in a majority of U6atac RNAs, i.e. in vertebrates, urochodates, *S. purpuratus*, *Lottia gigantea*, insects and plants as well as in the fungi *R. oryzae* and *P. blakesleeanus*. The RNA of the lycophyte *Selaginella moellendorffii* (spikemoss) has the sequence 'ATGGA'. However, we observed a different motif, '[GU]CCGA' (in the following referred to as the CC variant, as opposed to the normal AG sequence) in *Trichoplax*, cnidarians, *Reniera sp.*, *A. castellanii*,

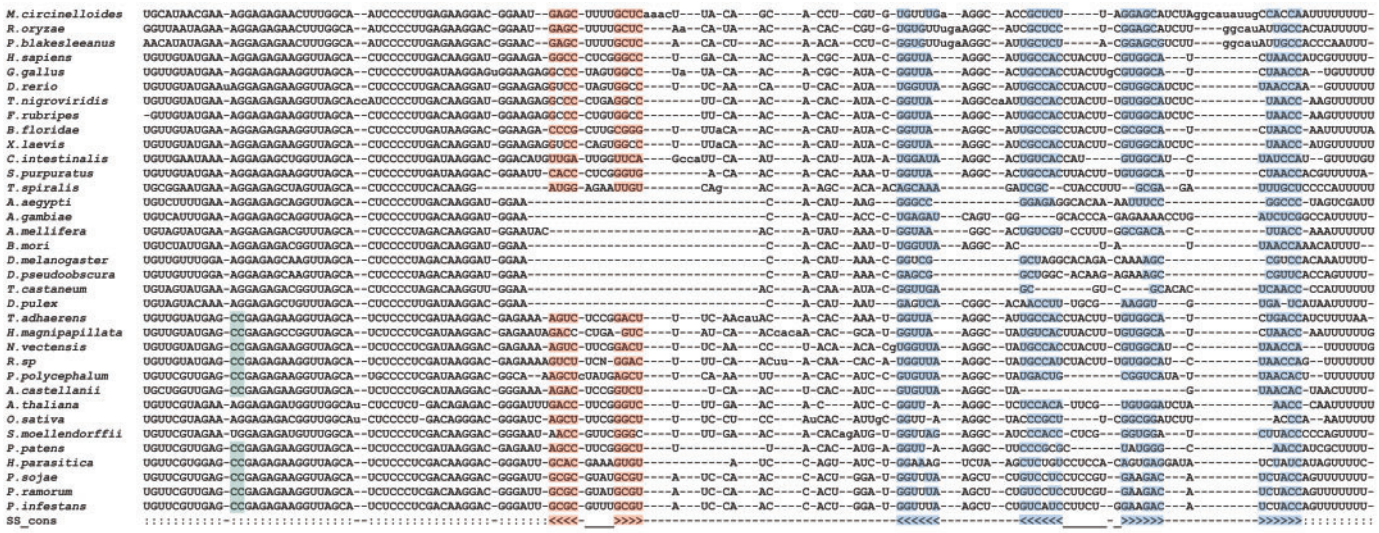

Figure 5. Alignment of U6atac spliceosomal RNA genes. Secondary structure elements are shown in a bracket notation at the bottom of the alignment. CC motifs in region supposed to pair to 5′ splice site, as well as regions involved in base-pairing are highlighted with color.

P. polycephalum, *Physcomitrella patens* and Oomycetes (Figure 5). We think that the predictions of U6atac RNA in these species are highly reliable because of sequence similarity, covariance model scores and ability to pair with U4atac and U12. Furthermore, it would seem that the CC variant does not represent a 'paralogue' of U6atac, as in all species examined either the AG or the CC homolog is present. It is intriguing that if the AG motif was the ancestral version a change to the CC variant occurred more than once during evolution. Examples are in the land plant branch and in the development of the lower metazoa *Trichoplax* and *Nematostella*. However, the possibility that the CC variant represents the ancestral version of the gene cannot be excluded. Also, in such a case, there must have been multiple independent transitions from CC to AG. The presence of the variant CC sequence is difficult to explain if the sequence AAGGA is important in pairing to the 5′ splice site (40,41). U12-type introns of *A. castellanii* and *Phytophthora* with a sequence able to pair with AAGGA are presented in Russell *et al.* (8) but these introns are not able to pair well with U6atac RNAs with the CC motif.

CONCLUSIONS

We have described a method to identify spliceosomal RNAs where in a first step candidates are identified using sensitive similarity searches or by profile HMM searches. These candidates are then more rigorously examined using covariance models to arrive at a final prediction of spliceosomal RNA. New spliceosomal RNA sequences found are used as queries in similar searches until no further genes are identified.

The results of this procedure clearly illustrate that highly sensitive local alignment searches and profile HMMs are important in the identification of spliceosomal RNAs. These RNAs tend to be conserved in sequence during evolution as compared to many other RNAs and perhaps the protocol used here is particularly suited for this category of ncRNAs. Ideally a combination of methods should be used to maximize sensitivity. At the same time, the covariance models are critical in order to evaluate the hits found in the initial searches. We have here relied to a large extent on the specificity of these models to predict ncRNAs and regard all RNAs reported here as strong predictions.

A large number of novel RNAs are identified in this work. Most noteworthy is the identification of RNAs being components of the minor U12 spliceosome in phylogenetic groups that previously were not known to have these RNAs or any U12-type spliceosomal components or introns. Examples are *Trichinella*, a nematode which in contrast to other nematodes like *C. elegans* contains minor spliceosomal RNA genes. We also have shown that minor spliceosomal RNAs are present in the deeply branching fungal branches Zygomycota and Chytridiomycota.

In summary, therefore, these results confirm previous studies of Russell *et al.* (8) that demonstrate an early origin of the minor spliceosome, as U12-type spliceosomal RNAs are present in a variety of evolutionary distant phyla. At the same time, our results do not allow us to conclude that a minor spliceosome was present in the ancestor of eukaryotes. Our results also point to multiple instances in evolution where the minor spliceosome seem to have been lost. One example is in the development of nematodes where the loss of U12-type RNAs occurred after the divergence of Pseudocoelomata. In the case of fungi, such RNAs were present in very early fungal evolution while they were lost in the development of Microsporidia and Ascomycota. Furthermore, minor spliceosomal RNAs were lost in the development of Dictyostelium of the Mycetozoa branch and in the development of the heterokont and plant branches. From this it would seem that U12-type splicing has a comparatively marginal role and may be disposed of in many phylogenetic groups.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

M.D.L. was supported by a grant from CONACYT, The National Council for Science and Technology, Mexico. Funding to pay the Open Access publication charges for this article was provided by Swedish Research School of Genomics and Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. Nilsen,T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, **25**, 1147–1149.
2. Kiss,T. (2004) Biogenesis of small nuclear RNPs. *J. Cell Sci.*, **117**, 5949–5951.
3. Will,C.L., Schneider,C., MacMillan,A.M., Katopodis,N.F., Neubauer,G., Wilm,M., Luhrmann,R. and Query,C.C. (2001) A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J.*, **20**, 4536–4546.
4. Madhani,H.D. and Guthrie,C. (1992) A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell*, **71**, 803–817.
5. Wassarman,K.M. and Steitz,J.A. (1992) The low-abundance U11 and U12 small nuclear ribonucleoproteins (snRNPs) interact to form a two-snRNP complex. *Mol. Cell Biol.*, **12**, 1276–1285.
6. Frilander,M.J. and Steitz,J.A. (1999) Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions. *Genes Dev.*, **13**, 851–863.
7. Burge,C.B., Padgett,R.A. and Sharp,P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol. Cell*, **2**, 773–785.
8. Russell,A.G., Charette,J.M., Spencer,D.F. and Gray,M.W. (2006) An early evolutionary origin for the minor spliceosome. *Nature*, **443**, 863–866.
9. Mewes,H.W., Albermann,K., Bahr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G. *et al.* (1997) Overview of the yeast genome. *Nature*, **387**, 7–65.
10. Guthrie,C. and Patterson,B. (1988) Spliceosomal snRNAs. *Annu. Rev. Genet.*, **22**, 387–419.
11. Mitrovich,Q.M. and Guthrie,C. (2007) Evolution of small nuclear RNAs in S. cerevisiae, C. albicans, and other hemiascomycetous yeasts. *RNA*, **13**, 2066–2080.
12. Palfi,Z., Schimanski,B., Gunzl,A., Lucke,S. and Bindereif,A. (2005) U1 small nuclear RNP from Trypanosoma brucei: a minimal U1 snRNA with unusual protein components. *Nucleic Acids Res.*, **33**, 2493–2503.
13. Griffiths-Jones,S. (2007) Annotating noncoding RNA genes. *Annu. Rev. Genom. Hum. Genet.*, **8**, 279–298.
14. Eddy,S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
16. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
17. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
18. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
19. Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
20. Freyhult,E.K., Bollback,J.P. and Gardner,P.P. (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, **17**, 117–125.
21. Chakrabarti,K., Pearson,M., Grate,L., Sterne-Weiler,T., Deans,J., Donohue,J.P. and Ares,M,Jr. (2007) Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA*, **13**, 1923–1939.
22. Davis,C.A., Brown,M.P. and Singh,U. (2007) Functional characterization of spliceosomal introns and identification of U2, U4, and U5 snRNAs in the deep-branching eukaryote Entamoeba histolytica. *Eukaryot. Cell*, **6**, 940–948.
23. Fast,N.M. and Doolittle,W.F. (1999) Trichomonas vaginalis possesses a gene encoding the essential spliceosomal component, PRP8. *Mol. Biochem. Parasitol.*, **99**, 275–278.
24. Vanacova,S., Yan,W., Carlton,J.M. and Johnson,P.J. (2005) Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis. *Proc. Natl Acad. Sci. USA*, **102**, 4430–4435.
25. Russell,A.G., Shutt,T.E., Watkins,R.F. and Gray,M.W. (2005) An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of Giardia lamblia. *BMC Evol. Biol.*, **5**, 45.
26. Nixon,J.E., Wang,A., Morrison,H.G., McArthur,A.G., Sogin,M.L., Loftus,B.J. and Samuelson,J. (2002) A spliceosomal intron in Giardia lamblia. *Proc. Natl Acad. Sci. USA*, **99**, 3701–3705.
27. Collins,L. and Penny,D. (2005) Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.*, **22**, 1053–1066.
28. Misumi,O., Matsuzaki,M., Nozaki,H., Miyagishima,S.Y., Mori,T., Nishida,K., Yagisawa,F., Yoshida,Y., Kuroiwa,H. and Kuroiwa,T. (2005) Cyanidioschyzon merolae genome. A tool for facilitating comparable studies on organelle biogenesis in photosynthetic eukaryotes. *Plant Physiol.*, **137**, 567–585.
29. Nozaki,H., Takano,H., Misumi,O., Terasawa,K., Matsuzaki,M., Maruyama,S., Nishida,K., Yagisawa,F., Yoshida,Y., Fujiwara,T. *et al.* (2007) A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga Cyanidioschyzon merolae. *BMC Biol.*, **5**, 28.
30. Takahashi,Y., Tani,T. and Ohshima,Y. (1996) Spliceosomal introns in conserved sequences of U1 and U5 small nuclear RNA genes in yeast Rhodotorula hasegawae. *J. Biochem.*, **120**, 677–683.
31. Takahashi,Y., Urushiyama,S., Tani,T. and Ohshima,Y. (1993) An mRNA-type intron is present in the Rhodotorula hasegawae U2 small nuclear RNA gene. *Mol. Cell Biol.*, **13**, 5613–5619.
32. Gilson,P.R., Su,V., Slamovits,C.H., Reith,M.E., Keeling,P.J. and McFadden,G.I. (2006) Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc. Natl Acad. Sci. USA*, **103**, 9566–9571.
33. Patel,A.A. and Steitz,J.A. (2003) Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.*, **4**, 960–970.
34. Thomas,J., Lea,K., Zucker-Aprison,E. and Blumenthal,T. (1990) The spliceosomal snRNAs of Caenorhabditis elegans. *Nucleic Acids Res.*, **18**, 2633–2642.
35. Blaxter,M.L., De Ley,P., Garey,J.R., Liu,L.X., Scheldeman,P., Vierstraete,A., Vanfleteren,J.R., Mackey,L.Y., Dorris,M., Frisse,L.M. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.
36. Klein,D.J., Schmeing,T.M., Moore,P.B. and Steitz,T.A. (2001) The kink-turn: a new RNA secondary structure motif. *EMBO J.*, **20**, 4214–4221.
37. Rozhdestvensky,T.S., Tang,T.H., Tchirkova,I.V., Brosius,J., Bachellerie,J.P. and Huttenhofer,A. (2003) Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Res.*, **31**, 869–877.
38. Vidovic,I., Nottrott,S., Hartmuth,K., Luhrmann,R. and Ficner,R. (2000) Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol. Cell*, **6**, 1331–1342.
39. Schultz,A., Nottrott,S., Hartmuth,K. and Luhrmann,R. (2006) RNA structural requirements for the association of the spliceosomal hPrp31 protein with the U4 and U4atac small nuclear ribonucleoproteins. *J. Biol. Chem.*, **281**, 28278–28286.
40. Tarn,W.Y. and Steitz,J.A. (1996) Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science*, **273**, 1824–1832.
41. Incorvaia,R. and Padgett,R.A. (1998) Base pairing with U6atac snRNA is required for 5' splice site activation of U12-dependent introns in vivo. *RNA*, **4**, 709–718.