
The fungal snoRNAome

SEBASTIAN CANZLER,¹ PETER F. STADLER,^{1,2,3,4,5,6,7} and JANA SCHOR⁸

¹Bioinformatics Group, Department of Computer Science, Leipzig University, D-04107 Leipzig, Germany

²Interdisciplinary Center for Bioinformatics, German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions, and Leipzig Research Center for Civilization Diseases, Leipzig University, D-04103 Leipzig, Germany

³Department of Diagnostics, Fraunhofer Institute for Cell Therapy and Immunology – IZI, D-04103 Leipzig, Germany

⁴Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany

⁵Department of Theoretical Chemistry, University of Vienna, A-1090 Wien, Austria

⁶Center for non-coding RNA in Technology and Health, University of Copenhagen, DK-1870 Frederiksberg C, Denmark

⁷Santa Fe Institute, Santa Fe, New Mexico 87501, USA

⁸Young Investigators Group Bioinformatics and Transcriptomics, Department of Proteomics, Helmholtz Centre for Environmental Research – UFZ, D-04318 Leipzig, Germany

ABSTRACT

Small nucleolar RNAs (snoRNAs) are essential players in the rRNA biogenesis due to their involvement in the nucleolytic processing of the precursor and the subsequent guidance of nucleoside modifications. Within the kingdom Fungi, merely a few species-specific surveys have explored their snoRNA repertoire. However, the wide range of the snoRNA landscape spanning all major fungal lineages has not been mapped so far, mainly because of missing tools for automatized snoRNA detection and functional analysis. For the first time, we report here a comprehensive inventory of fungal snoRNAs together with a functional analysis and an in-depth investigation of their evolutionary history including innovations, deletions, and target switches. This large-scale analysis, incorporating more than 120 snoRNA families with more than 7700 individual snoRNA sequences, catalogs and clarifies the landscape of fungal snoRNA families, assigns functions to previously orphan snoRNAs, and increases the number of sequences by 450%. We also show that the snoRNAome is subject to ongoing rearrangements and adaptations, e.g., through lineage-specific targets and redundant guiding functions.

Keywords: small nucleolar RNAs; snoRNA; fungi; evolution; target switch; snoRNA target; conservation

INTRODUCTION

Small nucleolar RNAs (snoRNAs) are non-protein-coding RNAs (ncRNAs) that guide the chemical modification of single nucleotides in other RNA molecules. Localized in the nucleolus of eukaryotic (and some archaeal) cells, they associate with at least four proteins to form the small nucleolar ribonucleoprotein (snoRNP) complex (Reichow et al. 2007). The target RNA molecule is held in the correct position by base-pairing to short unpaired region(s) within the snoRNA usually referred to as the antisense elements (ASE). The base-pairing completely specifies the target nucleotide. Known modifications are mostly located in ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs) (Darzacq et al. 2002; Decatur and Fournier 2002; Bratkovič and Rogelj 2011). Some snoRNAs have been shown to target residues in other RNA molecules such as transfer RNAs (Clouet d'Orval et al. 2001; Dennis et al. 2001), spliced leader RNAs (Uliel et al. 2004), or brain-specific messenger RNAs (Cavaillé et al. 2000; Kishore and Stamm 2006). Furthermore, snoRNAs are known

to be involved in the nucleolytic processing of rRNA precursors, the synthesis of telomeric DNA, genomic imprinting, and alternative splicing (Maxwell and Fournier 1995; Tollervey and Kiss 1997; Kiss 2002; Matera et al. 2007).

There are two distinct classes of snoRNAs: box C/D and box H/ACA snoRNAs. They are distinguished by their secondary structure, sequence features, and the chemical modifications they are guiding (Balakin et al. 1996; Tollervey and Kiss 1997). Box C/D snoRNAs form a stem-loop structure with a rather long loop, which is stabilized by the associated proteins and guide the 2'-O-methylation of ribose groups. Box H/ACA snoRNAs are longer, fold into a thermodynamically more stable double stem-loop structure, and guide the pseudouridylation of uracil residues in the target RNA. Additionally, there are chimeric snoRNAs that share features of both classes. They are much longer and are described to have different functions (Darzacq et al. 2002). Similar to other small ncRNAs, snoRNAs require both specific secondary

Corresponding author: sebastian@bioinf.uni-leipzig.de

Article is online at <http://www.rnajournal.org/cgi/doi/10.1261/rna.062778.117>.

© 2018 Canzler et al. This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

structures and characteristic sequence motifs to perform their function. These features are, therefore, preserved during evolution and are clearly recognizable by comparative methods (Ganot et al. 1997; Tollervey and Kiss 1997). While the sequence motifs involved in protein binding are common to all members in each of the two classes, the ASEs are conserved only among members of snoRNA families with the same target. These limited restrictions on the snoRNAs primary sequence allowed for a rapid evolution, albeit retaining the secondary structure elements. This hinders the identification of snoRNA genes by purely sequence-based methods such as blast (Altschul et al. 1990). To overcome this limitation, we introduced the computational annotation pipeline snoStrip (Bartschat et al. 2014) that is specifically designed to track all specific characteristics of snoRNAs.

So far, the topic of fungal snoRNAs has mainly been approached by species-specific analysis leading to an exceptionally sparse snoRNA landscape. Here, we use the snoStrip approach to analyze a large set of fungal species with genomes that are available in decent quality for their snoRNA abundance. We started with experimentally verified snoRNAs in five fungi. We subsequently studied their evolutionary conservation and the coevolution of snoRNAs with their targets across the whole kingdom to gain insights into the evolutionary history of this ncRNA class and to understand the dynamics and processes to which it is still subjected to. We provide a comprehensive set of fungal snoRNAs, their detailed description with respect to genomic location, box motifs, potential/confirmed target information (including observed target switches), family assignment and suggestions of the evolutionary history of individual snoRNA families. All data can be viewed in and downloaded from our [Supplemental Material](#). We submit manually curated snoRNA family alignments to the Rfam database (Nawrocki et al. 2015).

MATERIALS AND METHODS

Genome and snoRNA data

Genome sequences from 147 fungal species were downloaded from Ensembl Genomes (Kersey et al. 2016), JGI (Nordberg et al. 2014), Broad Institute (Fungal Genome Initiative), and Candida Genome Database (Skrzypek et al. 2017). An NCBI-based taxonomic tree displaying the relationship, genome source, and genome version for all fungal organisms in this evolutionary survey is shown in the [Supplemental Figure S1](#). For 63 out of the 147 species, most snoRNA sequences have already been retrieved in a previous study, primarily to test snoStrip (Bartschat et al. 2014). In this earlier work, we started with experimentally detected snoRNAs extracted from five surveys for *Neurospora crassa* (Liu et al. 2009), *Aspergillus fumigatus* (Jöchl et al. 2008), *Candida albicans* (Mitrovich et al. 2010), *Saccharomyces cerevisiae* (Piekna-Przybylska et al. 2007), and *Schizosaccharomyces pombe* (Li et al. 2005). An overview of these snoRNAs and the corresponding publications is compiled in [Supplemental Table S2](#). The nomenclature of snoRNAs used in this contribution is consistent across

different species. [Supplemental Table S3](#) contains a dictionary that combines the species-specific snoRNA names, taken from their original publications with internal snoRNA family designations. Here we use the results of Bartschat et al. (2014) as our starting point. The initial set comprises 3564 snoRNA sequences assigned to 123 snoRNA families in the 63 species. It includes 231 experimentally validated snoRNA genes taken from the five publications.

Homology search

We applied the snoStrip pipeline (Bartschat et al. 2014) to the set of collected snoRNAs and the 147 fungal species in an iterative manner, starting with Pezizomycotina, followed by Saccharomycotina, and other lineages toward the root of the phylogenetic tree. Each time new (plausible) homologous snoRNAs were detected, the procedure was repeated to decrease the number of false negatives until no novel homologs were found anymore.

Data curation

The candidate snoRNAs were curated regarding the automatically identified box motifs, class-specific sequence lengths, and the overall fit of each snoRNA sequence within its respective family. To identify incorrectly annotated box motifs, the conservation of all predicted boxes was checked by comparing their start positions in the snoRNA family alignment. Motifs that start at non-conserved positions are most probably false annotations and were readjusted to fit the snoRNA family-specific box pattern and box position. Sequences with the re-adjusted C- or D-boxes that did not agree with the canonical box motif pattern were removed from further analysis. Candidate sequences that are either too long or too short were mostly the consequence of misannotated box motifs since snoStrip cuts snoRNA genes based on their box motif positions. For these candidates, box motifs were analyzed with respect to their conserved start positions. Sequences with re-adjusted box motifs were automatically trimmed or enlarged, respectively.

Box motifs, sequence, and structure

Box motifs were generated from all snoStrip-derived snoRNA candidates and compared to canonical box motifs of yeast and vertebrate snoRNAs. Sequence lengths and distances between all box motifs were collected and compared. RNAfold and RNAalifold, both part of the Vienna RNA Package (Hofacker et al. 1994), were utilized to predict the secondary structure.

Phylogenetic analysis

We used the ePoPE software (Hertel and Stadler 2015) to follow the evolution of the snoRNA families along the phylogenetic tree. It implements a variant of Sankoff's parsimony algorithm using the Dollo variant that excludes the loss and re-gain of a gene family along the same lineage during evolution. Innovation and deletion/loss/divergence events are deduced and mapped to the branches of the phylogenetic tree. The ePoPE results are combined for *all* snoRNA

families using the ePoPE_summarize.pl tool that comes with the ePoPE distribution.

Target prediction and analysis

Target prediction is part of the snoStrip pipeline. The computational tools PLEXY and RNAsnoop are used to predict targets for box C/D snoRNAs and box H/ACA snoRNAs, respectively (Tafer et al. 2010; Kehr et al. 2011). SnoRNAs are investigated for single or double guiding potentials based on these predictions and/or confirmed target interactions. SnoRNAs that remain without target association are considered orphan. SnoRNAs that are assigned to the same family but differ in their associated targets are manually investigated for a potential target switch.

Lineage-specific conservation of target interactions

To study the conservation of interactions, single sequence snoRNA targets are initially predicted and, subsequently, their conservation in other species is evaluated. Kehr et al. (2014) developed the interaction conservation index (ICI) to formally investigate the conservation of such interactions. In brief, the conservation of the modification and the conservation in a specific snoRNA family are calculated as follows:

$$\begin{aligned} \text{ICI}_{\text{mod}}(t, s) &= \frac{1}{|O(s)|} \left(\sum_{k \in O(t,s)} \frac{\varepsilon(t, s, k)}{\bar{\varepsilon}(t, k)} \right) \\ \text{ICI}_{\text{sno}}(t, s) &= \frac{1}{|O(s)|} \left(\sum_{k \in O(t,s)} \frac{\varepsilon(t, s, k)}{\hat{\varepsilon}(s, k)} \right) \end{aligned} \quad (1)$$

Here, $\varepsilon(t, s, k) = \min_{x \in X(t,s,k)} E_{\text{mfe}}[x, y_{t,k}]$ is the most negative interaction minimum free energy between a snoRNA x of family s and the target t in species k . The normalizations

$$\begin{aligned} \bar{\varepsilon}(t, k) &= \sum_{s \in S(t,k)} \varepsilon(t, s, k) / |S(t, k)| \\ \hat{\varepsilon}(s, k) &= \sum_{t \in T(s,k)} \varepsilon(t, s, k) / |T(s, k)| \end{aligned} \quad (2)$$

are obtained by averaging over all predictions of target t in species k or all targets t of snoRNA s in species s , respectively. Their normalized parameters are then summed up over all species $k \in O(t,s)$ in which a prediction of target t is found for snoRNA family s . The sum is then normalized w.r.t. the number of species $|O(s)|$ in which the snoRNA family s is present. This approach is particularly suitable for modification sites that are present in a large set of analyzed organisms. In cases where a potential target appears to be lineage-specific, the ICI score will drop to rather low values due to the normalization score $1/|O(s)|$ that represents all organisms sharing a homologous snoRNA of family s .

To appropriately investigate alternative or additional targets that merely appear in a particular subset of organisms, the ICI score calculation has to be adapted to take the particular phylogenetic distribution of a target interaction into account. Therefore, the normalization is restricted to the smallest phylogenetic or taxonomic subtree that harbors all organisms that share prediction of target t in snoRNA family s . Assume the overall taxonomic tree is represent-

ed by a tree $T = (V, E)$ with root γ . The minimal subtree $U_\tau = (V_\tau, E_\tau)$ with root τ shares the node set

$$V_\tau = \{v | \forall (v, u), u \in V_\tau : \text{LCA}_T(v, u) \in V_\tau\},$$

where $\text{LCA}_T(v, u)$ is the lowest common ancestor in tree T of both nodes v and u . More precisely, the LCA is the lowest node, i.e., the farthest node from the root, that has both v and u as descendants. Hence, the ICI scores in a particular subtree rooted at τ can be calculated as follows:

$$\begin{aligned} \text{ICI}_{\text{mod}, \tau}(t, s) &= \frac{1}{|O_\tau(s)|} \left(\sum_{k \in O_\tau(t,s)} \frac{\varepsilon(t, s, k)}{\bar{\varepsilon}(t, k)} \right) \\ \text{ICI}_{\text{sno}, \tau}(t, s) &= \frac{1}{|O_\tau(s)|} \left(\sum_{k \in O_\tau(t,s)} \frac{\varepsilon(t, s, k)}{\hat{\varepsilon}(s, k)} \right), \end{aligned} \quad (3)$$

where $O_\tau(s) = \{k | \exists t : X(t, s, k) \neq \emptyset \text{ and } v_k \in V_\tau\}$ denotes the set of organisms that are contained in the subtree τ and share at least one snoRNA of family s . v_k is the leaf that denotes organism k .

Data validation

To verify our snoRNA annotations, we compared our data with Rfam-annotated families and cross-checked with available Ribo-seq archives.

Rfam database

The extensive collection of snoRNA families reported here is intended to be integrated into Rfam. We, therefore, carefully compared our results with the previous Rfam-annotated snoRNA s . The current Rfam (version 12.3) covers 755 snoRNA models. Of note, 116 of these families contain at least one fungal snoRNA from the 147 organisms investigated here. In total, these 116 models contain 1621 snoRNA sequences found in our set of organisms, 457 of which are included in the Rfam seed alignments, identifying them as very high confidence sequences, typically with direct experimental support.

Ribosome footprinting

Despite its natural purpose of visualizing translation, Ribo-seq data are also known to include ncRNAs that are not part of ribosomal protein-protected complexes in their high-throughput sequencing libraries (Ingolia et al. 2014; Ji et al. 2016). We, therefore, used publicly available Ribo-seq data from four different fungi (*S. cerevisiae* [Ingolia et al. 2009], 14 libraries; *S. pombe* [Duncan and Mata 2014], three libraries; *C. albicans* [Muzzey et al. 2014], three libraries; and *Ajellomyces capsulatus* [Gilmore et al. 2015], four libraries) to support the identification of our novel snoRNAs. Importantly, all snoRNAs in *S. cerevisiae* were based on independent experimental studies and will serve here as a positive control for the use of Ribo-seq data.

In addition to the raw read data, we also evaluated the localization of these reads when overlapping with our snoStrip annotation. In contrast to mRNAs, where one would expect a rather uniformly distributed pattern with a clearly visible 3-nt periodicity, nonribosomal protein-associated RNAs show a highly localized read pattern. We, therefore, used the percentage of maximum entropy (PME) to

quantify the uniformity of read distribution across the snoRNA-annotated regions, using the Rfoot tool (Ji et al. 2016).

RESULTS

So far, there is no generally accepted nomenclature of snoRNA families across different fungal species, and thus quite a few snoRNA genes of distinct organisms are named differently, although these genes belong to the same snoRNA family. Here, we present the first complete and reliable mapping of snoRNA names across the kingdom of fungi, that is entirely based on sequence and functional homologies. In the following, we will use established gene names to designate snoRNA families where possible. In cases where homologs have different names in different species, we use the preferred order *Saccharomyces cerevisiae*, *Neurospora crassa*, *Aspergillus fumigatus*, *Candida albicans*, and *Schizosaccharomyces pombe*. To simplify cross-referencing with machine-readable data we also list the snoStrip family designations in parentheses. A complete dictionary of nomenclature correspondences can be found in [Supplemental Table S3](#). Similarly, we pragmatically identify target positions with their position in the multiple sequence alignments of the target RNAs. Coordinates for reference sequences from selected organisms are given in parentheses. Single-sequence target RNAs and target RNA alignments are provided in [Supplemental Table S4](#).

Expanded complement of fungal snoRNAs

We used snoStrip to search for additional homologs of the initial set of 67 box C/D snoRNA and 56 box H/ACA snoRNA families in 147 fungal species. The U3 snoRNA family is published separately due to its special function, various splice variants and characteristics (Canzler et al. 2017). All snoStrip candidates were carefully cross-checked in all species to reduce the number of false negatives and to exclude potentially incorrect annotations. In total, we found 5593 box C/D snoRNA and 2255 box H/ACA snoRNA sequences, expanding the collection of annotated fungal snoRNAs by more than 200% compared to our previously available snoRNA set (Bartschat et al. 2014) and over 450% compared to Rfam-annotated snoRNAs. This massive amount of snoRNA data substantially increases both the phylogenetic depth and the resolution of the snoRNA annotation.

Characteristics of fungal snoRNAs

Box motifs

Sequence motifs were extracted from all snoStrip-annotated snoRNAs. The complete collection is available for download from [Supplemental Section S6 \(A\)](#). In general, these motifs are consistent with the published rules (Xia et al. 1997; Watkins et al. 2000, 2002; Cahill et al. 2002) for canonical snoRNA box motifs known from both yeast and animals.

Box C (RUGAUGA) and D (CUGA) match the consensus sequence motifs almost perfectly. Box C shows an initial purine (R) in 92% of all cases. The first GA dinucleotide is absolutely conserved. In 4.2% of the cases, the 5' nucleotide (C) of box D is substituted, usually by an adenine. The remaining positions are nearly perfectly conserved ($\geq 99.7\%$). As expected from yeast and other animal snoRNAs the situation is different for prime box motifs (Kiss-László et al. 1998; Cahill et al. 2002). In box C', merely the first UG dinucleotide and, to a lesser extent, the trailing GA dinucleotides are highly conserved. This might indicate a role in the binding of snoRNP-associated proteins. In box D', variations of the canonical nucleotides occur quite frequently in each position (between 15% and 45%).

In box H/ACA snoRNAs, we observe that the sequence of box ACA is highly conserved with rare variations in its middle position. The adenine residues of box H (ANANNA) are highly conserved at the first and third position, while the trailing adenine (sixth position) is more variable. The second position of this motif is a guanine in nearly 80% of the box H/ACA snoRNAs, whereas the fourth and fifth N positions do not show a significantly overrepresented nucleotide. Again, these results are in accordance with previously published motif constraints (Normand et al. 2006).

Sequence length

Consistent with the published lengths of box C/D snoRNAs, 90% of the novel snoStrip-annotated snoRNAs are 80–135 nt in length, with a median of 93 nt ([Supplemental Fig. S6.2](#)). Family Nc_CD_53 (*N. crassa*, CD_53 in snoStrip) is the only exception since its members share sequences with lengths between 200 and 300 nt. Crucial features are the distances between box C and the potential box D' as well as between box C' and D since these stretches harbor the target binding sites. These regions provide sufficient space to harbor a potential ASE in all detected snoRNA candidates, see [Supplemental Section S6 \(B\)](#).

Box H/ACA snoRNAs are usually longer than box C/D snoRNAs. Their median sequence length is 188 nt. The shortest sequence covers 115 nt, while 90% of all sequences are between 148 and 266 nt. Both single hairpin sequences share a similar length distribution. For boxplots and more details, see [Supplemental Section S6 \(B\)](#).

Secondary structure

Due to its specific post-transcriptional processing by exonucleases, both trailing ends of box C/D snoRNAs are cut precisely 5 nt away from the C and D boxes, respectively (Kishore et al. 2013). Because of these rather short ends, only a small subset of snoRNA sequences was predicted to fold a short closing stem (1208 out of 5593). If we enlarge the trailing ends to 10 nt instead, a stem could be predicted for nearly 60% (3317). The fact that more than 40% of the box C/D snoRNAs do not form a terminal stem strongly

suggests that the terminal helix is not required for their function, and hence the snoRNP-associated proteins themselves may be in charge to bring the RNA molecule and the assembled proteins into the correct functional conformation.

In contrast, box H/ACA snoRNAs are required to develop a specific secondary structure to function appropriately. Only 15% (395 out of 2255) of all box H/ACA snoRNAs were not predicted to fold into a stem-loop structure for both hairpins.

In general, snoRNA-specific characteristics such as box motifs, lengths, and secondary structures are very similar between Fungi and Metazoa (Kehr et al. 2014).

Phylogenetic distribution of fungi snoRNAs

The comprehensive snoRNA data set reported here makes it feasible to thoroughly examine the phylogenetic fundament of fungal snoRNAs. Figure 1 depicts a heatmap of the distribution of fungal box C/D snoRNA families. Higher resolution heatmaps of both C/D and H/ACA snoRNA families are included in Supplemental Section S7.

In general, fungal snoRNA families encompass exactly one snoRNA sequence per organism. Exceptions to this rule are the snoRNA “clans” CD_5 and CD_19, which typically have two or three members per species. This is explained by several target switches and major rearrangements between different snoRNA families that forced snoStrip to merge the previously separate snoRNA families. We will return to this point below in the context of target switches.

Individual species often encode multiple paralogs of one or several families. Good examples are *Postia placenta*, *Atractiellales sp.*, and *Nadsonia fulvescens*. In some cases, paralogs persist in larger clades, such as AM921940 (CD_41) in Leotiomycetes and Nc_CD_28 (CD_28) in Sordariomycetes.

Almost half of the box C/D snoRNA families are traceable down to the root of fungi (32/68), i.e., at least one early branching fungal lineage is attested to carry this snoRNA family, such as Microsporidia, Mucoromycotina, Chytridiomycota, or Blastocladiomycota. On the other hand, several families appear to be lineage-specific, e.g., seven in Saccharomycotina (see box “A” in Fig. 1), nine in Pezizomycotina (box “B”), and six in Sordariomycetes (box “C”).

In addition to lineage-specific families there are families that are absent in specific clades only. Basidiomycota, for example, do not seem to contain orthologs of families snR48 (CD_8), snR190 (CD_16), or U14 (CD_37), while there is no trace of family AM921940 (CD_41) in Saccharomycotina. Members of Nc_CD_40 (CD_40) are not detected in Eurotiomycetes, while Sordariomycetes are attested to miss homologs of families snR39/b (CD_47) and snR58 (CD_68). In some other cases, one or two representatives are found in lineages where the other species carry no detectable homologs. In these cases, only a more detailed analysis of target interaction might answer the question whether this single

snoRNA is a true member of the family or whether it might be an artifact.

In contrast to the broad distribution of box C/D snoRNAs, only seven box H/ACA snoRNA families (out of 50) are detected in early branching fungi and Dikarya. None of these are detected in Microsporidia leaving this clade completely without any annotated box H/ACA snoRNA. Our data show that box H/ACA snoRNAs show substantially more lineage-specific innovation and deletion events than observed in box C/D snoRNAs, see Supplemental Figure S7. In total, 22 out of the 50 box H/ACA families are found only in a small subset of species. Moreover, several families are found in two or more lineages but seem to be completely lost in others, e.g., snR42 (HACA_33), AJ632014 (HACA_56), and snR33 (HACA_24) that are present in Taphrinomycotina and Saccharomycotina but cannot be found in Pezizomycotina.

We remark that not a single box H/ACA snoRNA is found in *Pyrenophora tritici-repentis* (marked with an asterisk in the Supplemental Fig. S7.2). This observation is in sharp contrast to box C/D snoRNA sequences, where *P. tritici-repentis* orthologs are found in nearly all families that are present in the *P. tritici-repentis*-containing Dothideomycetes lineage.

Evolutionary events in snoRNA history

With the help of the ePoPE software, we identified the last common ancestor of each individual snoRNA family and found the most parsimonious estimate for the number of paralogs at the inner nodes of the tree. We deduced potential gain and loss events of individual paralogs of each snoRNA family and summarized this information for all analyzed snoRNA families to retrieve a full picture of the evolution of snoRNAs in fungi.

Relative innovation and deletion events mapped to the pre-ordered nodes of the NCBI-derived taxonomic tree up to species level are shown in Figure 2; see Supplemental Figure S8.1 for a version with absolute values. We observe a large number of snoRNA families that emerged at each major branch point along the backbone of the taxonomic tree. A total of 32 box C/D snoRNA families could be traced to the root of fungi, indicating an even more ancient origin. At the root of Dikarya, Ascomycota, Saccharomyceta, and Pezizomycotina, a total of 9, 3, 6, and 10 families seem to have emerged, respectively. A similar picture is drawn in the case of box H/ACA snoRNAs where seven families could be traced to the root of fungi. An additional 7, 10, 4, and 3 families were gained at the root of Dikarya, Ascomycota, Saccharomyceta, and Pezizomycotina, respectively. According to our methods, we could only detect innovations of snoRNA families at branches leading to the five starting species.

Microsporidia seem to have lost almost the entire snoRNA complement that has been present before their split during the evolution. Only two box C/D snoRNA families seem to be conserved in this lineage. Gardner et al. (2010) already mentioned the remarkable absence of snoRNA genes in

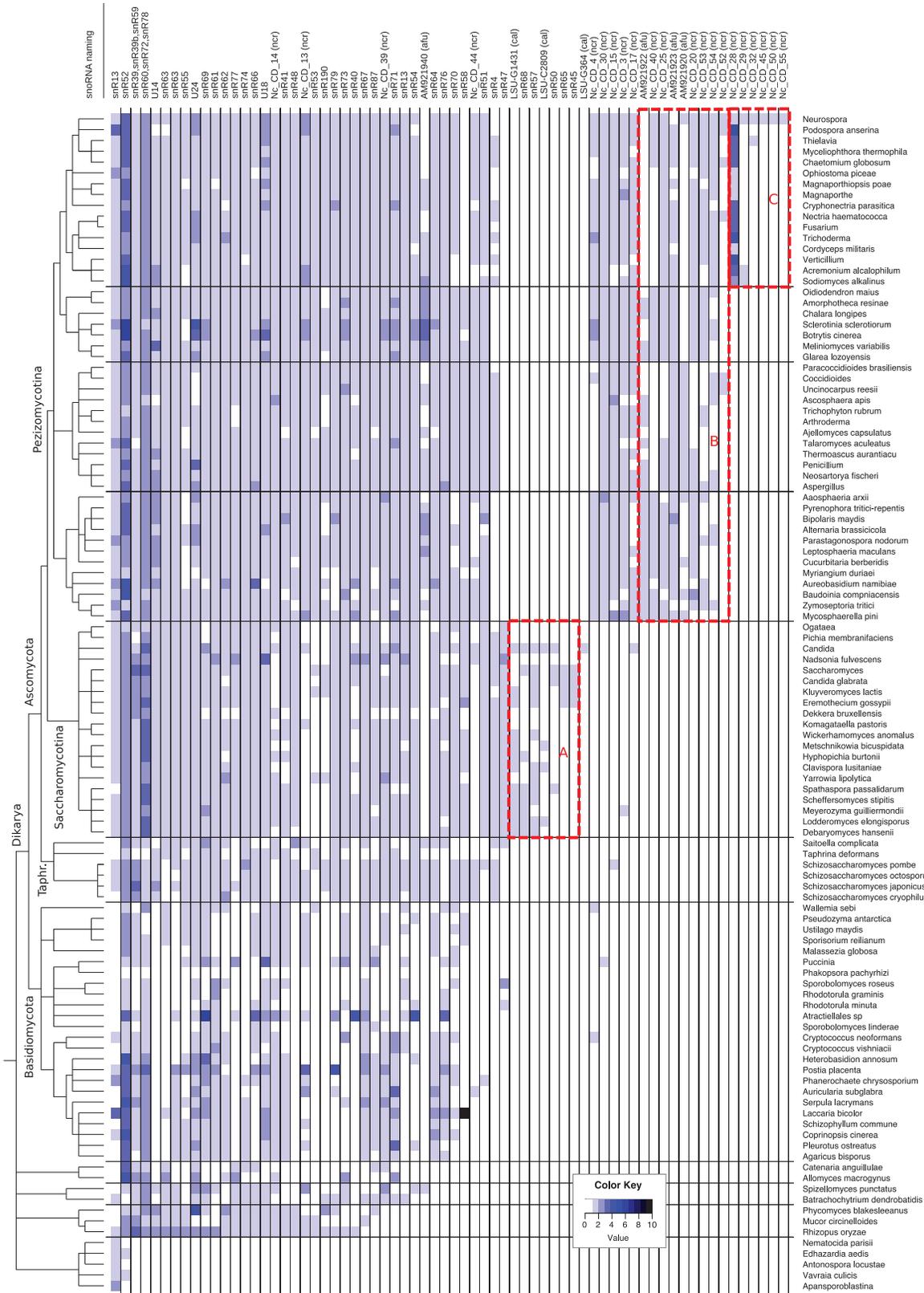


FIGURE 1. The heatmap shows the phylogenetic distribution of box C/D snoRNAs. Each column represents a specific snoRNA family, while each row either represents a certain species or genus. A taxonomic classification is shown on the *left-hand* side. The number of snoRNAs detected in a particular species and snoRNA family is encoded in a blue color scheme. Lineage-specific families are boxed (A, Saccharomycotina; B, Pezizomycotina; C, Sordariomycetes).

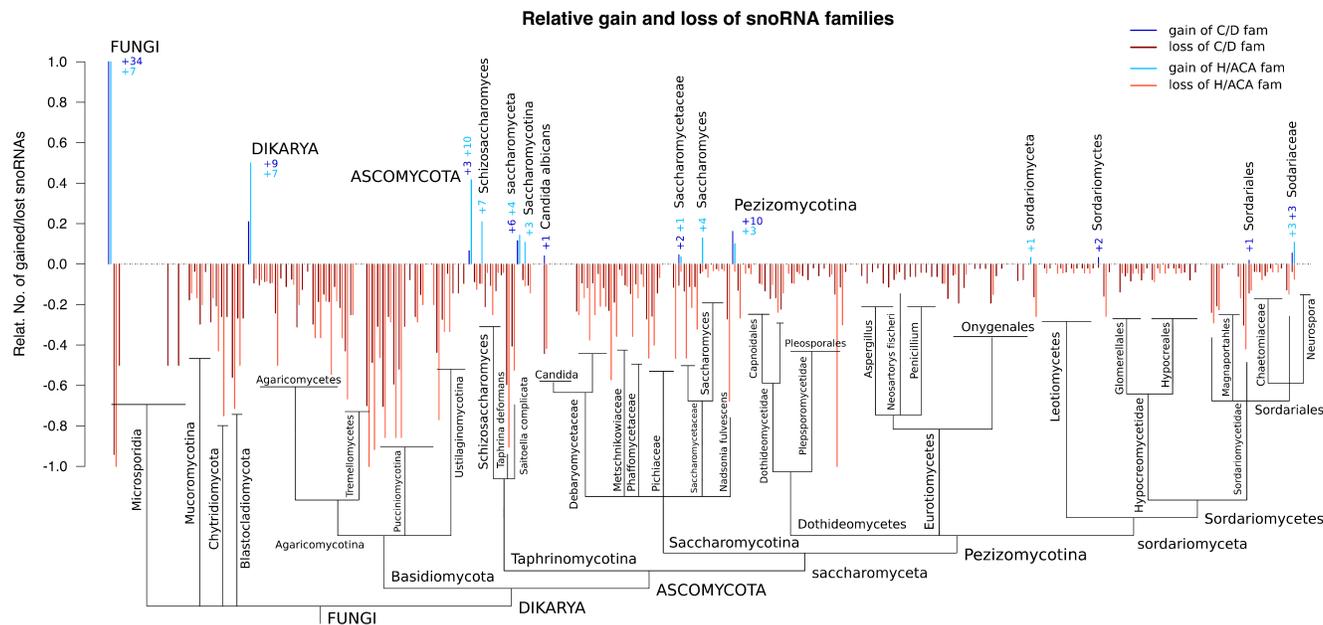


FIGURE 2. Relative numbers of gains and losses of entire snoRNA families during fungal evolution. The relative gain is the number of gained snoRNA families compared to the observed number of snoRNA families. The relative loss describes the number of lost snoRNA families compared to the number of snoRNA families in the parent node of the phylogenetic tree.

this clade, although all components of the snoRNA machinery are clearly present. We agree with these researchers that without further experimental investigations in these fungi, we cannot state a true loss or a rearrangement of their snoRNA repertoire.

Focusing on species level, we frequently observe that individual organisms seem to have lost a substantial number of their snoRNAs, i.e., in the Basidiomycota lineage. In particular, *Walleimia sebi* and several Pucciniomycota seem to have lost nearly their entire set of box H/ACA snoRNAs (*W. sebi*: 92%, *Rhodotorula minuta*: 86%, or *Phyllozymba linderiae*: 86%). The impact on box C/D snoRNAs is more moderate (26% on average). A potential correlation with significantly smaller genome sizes in Pucciniomycota was not detected (data not shown). The previously mentioned loss of the entire box H/ACA snoRNA set in *Pyrenophora tritici-repentis* is also clearly visible. Other organisms such as *Podospira anserina* and *Ophiostoma piceae* also show an increased loss rate (*P. anserina*: 15% C/D and 13% H/ACA; *O. piceae*: 30% C/D and 42% H/ACA).

Novel *Candida albicans* snoRNAs are lineage-specific

Mitrovich and colleagues identified four novel snoRNA candidates among their set of 40 snoRNA genes in *C. albicans* that showed no high sequence similarity toward already annotated budding yeast sequences (Mitrovich et al. 2010). One of these sequences is found to share a homologous target binding region with a known *N. crassa* snoRNA (Nc_CD_39). Families LSU-C2809 and LSU-G1431 in Mitrovich et al. (2010) (snoStrip: CD_69 and CD_71) are exclusively present in

Saccharomycotina except for Saccharomycetaceae. They are also found to share an extraordinarily conserved target-interaction with ICI scores of 1.813 (25S-4055; *C. albicans*: 25S-3118) and 1.289 (25S-2490; *C. albicans*: 25S-1740), respectively. The remaining family LSU-G364 (CD_72) is merely found in two closely related species: *Candida dubliniensis* and *Candida tropicalis*.

Fission yeast-specific snoRNAs

Similar to *C. albicans*, several snoRNAs published in the fission yeast (Li et al. 2005) are found to be lineage- or even species-specific. In the original publication, 12 sequences have not been mapped to budding yeast snoRNAs and seven of them have no predicted target interaction. By means of snoStrip, AJ632008 (HACA_46) and AJ632011 (HACA_47) have been detected to be functional homologs to snR86 (HACA_36) and snR5 (HACA_27), respectively. The first one includes a switch of the ASE from the first (*S. pombe*) to the second hairpin (*S. cerevisiae*), while the latter two families share far too little sequence similarity to be denoted as homologous sequences. Families AJ632018 (HACA_9), AJ632010 (HACA_48), AJ632016 (HACA_53), and AJ632012 (HACA_54) are found to be conserved outside of Taphrinomycotina. The first two families map to families with an annotated target, while the latter families lack such a finding. The remaining sequences are either specifically detected in *Schizosaccharomyces* (AJ632009 [HACA_50], AJ632017 [HACA_51], and AJ632013 [HACA_55]) or exclusively found in *S. pombe* (AJ632015 [HACA_45], AJ632019 [HACA_49], and AJ632014 [HACA_56]).

Conservation of target interaction

Despite some single-species studies, a thorough functional analysis of the entire fungal snoRNAome has not been done before but is essential to sort and clarify the snoRNA landscape (identification of functional homologous sequences) and to further investigate functional rearrangements and other peculiarities.

In accordance with their conserved function, each snoRNA family can either be classified as single guide, double guide, or orphan snoRNA. Single guide sequences share a conserved and functional antisense element either upstream of box D or D' in box C/D snoRNA or either in hairpin 1 (HP1) or hairpin 2 (HP2) in box H/ACA snoRNAs. Double guide snoRNAs exhibit functional target binding regions in both positions. Orphan snoRNAs have no known and conserved target interaction. Normally, each individual snoRNA is predicted to be capable of binding several regions in different target RNAs. But target predictions that are based on single sequence predictions are not overly convincing in a biological point of view.

Among the 68 box C/D snoRNA families, the majority (40) are “true” single guides meaning that these families share exactly one conserved target region (28 families share a functional D' target and 12 a conserved D target). Another 14 box C/D snoRNA families are “predominantly” single guides, i.e., these families share exactly one strongly conserved target binding region (three families share a conserved D target, 11 families share a functional D' target), while the other target region is only found to be functional in a restricted subset of taxa. Only eight families harbor two functional target binding regions that are conserved throughout all the lineages in which the snoRNA families are present. The remaining six families stay orphan, i.e., no potential interaction has been published so far.

In case of box H/ACA snoRNAs, 23 families are true single guides (eight families share a conserved pseudouridylation pocket in HP1, 15 families in HP2). Six families exhibit a lineage-specific HP2 target in addition to the globally conserved target in HP1. The reverse situation can be seen in three box H/ACA snoRNA families. Eleven families are double guides, and seven families remain orphan. A summary of the snoRNA classification can be seen in Figure 3. Detailed information about each family and the snoStrip-assigned target interactions, e.g., alignment position of the modification site, ICI scores, and mean minimum free energy values, can be found in the [Supplemental Sections S12–S23](#).

Only a minority of box C/D snoRNAs is found to contain two overly conserved target regions upstream of box D and D'. However, except for the “snoRNA clans” CD_5 and CD_19, none of the remaining six families is traceable among all major fungal lineages. Two families, Nc_CD_17 (CD_17) and AM921920 (CD_35), are found in Pezizomycotina, whereas snR47 (CD_67) is exclusively found in Saccharomycotina. The remaining families are either found in Sordariales Nc_CD_32 (CD_32), a subgroup of Sordariomycetes, or in Glomerellales and Neurospora Nc_CD_29 (CD_29).

Double guide box H/ACA snoRNA families occur more frequently. Eleven families are originally annotated as double guides, and most of their targets are convincingly confirmed by snoStrip. Furthermore, double guided box H/ACA snoRNAs are commonly traceable across a wide range of fungal organisms. Four families have their origin at the root of Dikarya or even further back: Nc_HACA_2 (HACA_2), snR3 (HACA_3), snR8 (HACA_6), snR80 (HACA_37). Two more families are traced to the root of Ascomycota: snR5 (HACA_27), snR49 (HACA_29), whereas the remaining five families are lineage-specific (two found in Saccharomycotina, snR82 [HACA_31], snR161 [HACA_39]) or genus-specific (two found in *Saccharomyces*, snR81

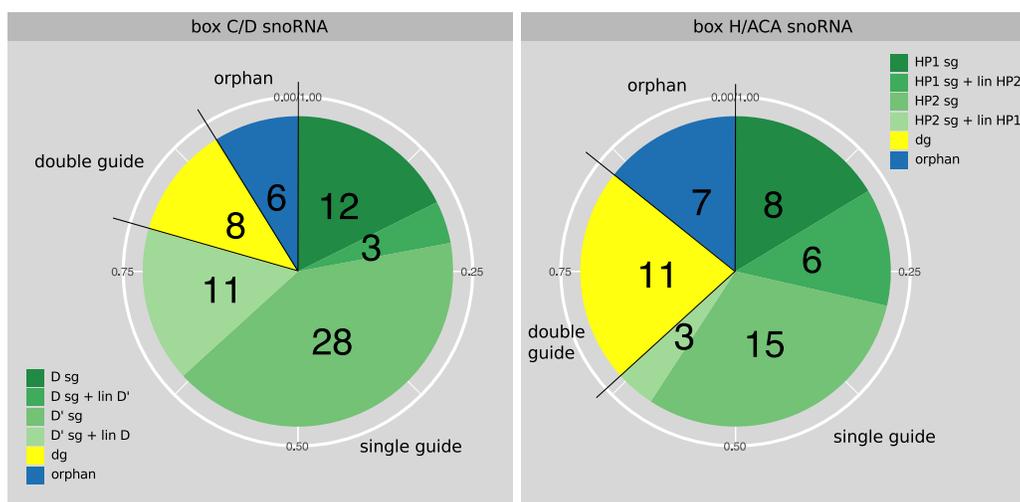


FIGURE 3. Pie chart of both major snoRNA classes. Based on their conserved target prediction, snoRNA families are either classified as single guide (sg), single guide with a lineage-specific target in its nonconserved target region (lin), double guide (dg), or orphan.

[HACA_26], snR83 [HACA_30]; one found in *Schizosaccharomyces*, AJ632008 [HACA_46]). We note that family HACA_2, despite its early origin, is absent in Saccharomycotina. The function of HP2 to guide the modification 25-3541 (25S-2351 in *S. cerevisiae*) is shifted toward the Saccharomycotina-specific HACA_31 (HP2). Both families show convincing sequence homology in their second hairpin but are fairly diverged in their first one.

Family snR3 is known (Schattner et al. 2004) to guide three targets in both the budding yeast and fission yeast (annotated as AJ632000 in *S. pombe*, HACA_3 in snoStrip); HP1 is known to guide modification at position 25S-3311 (25S-2129 and 25S-2216 in the budding and fission yeast, respectively), while there are two targets in HP2; 25S-3449 and 25S-3315 (*S. cerevisiae* 25S-2264 and 25S-2133, *S. pombe* 25S-2351 and 25S-2220). All three targets are found to be conserved across Dikarya. In the original *Neurospora* publication (Liu et al. 2009), however, HP1 is annotated to guide the isomerization at position 25S-1200 (25S-401 in *Neurospora crassa*). This guiding capability is not found to be conserved throughout the members of this family, unlike the yeast annotated target, which is also convincingly predicted in *Neurospora* species even with a lower interaction energy.

Orphan snoRNA

Orphan snoRNAs are sequences without a known target interaction on both potential antisense elements. In the originally published snoRNA data sets of five different fungi, orphan box C/D snoRNAs were annotated for *S. cerevisiae* (two sequences), *N. crassa* (2), and *A. fumigatus* (9). In addition to these sequences, 11 *N. crassa* snoRNAs were published with predicted targets based on single sequence target prediction only. Since there is usually more than just one valuable prediction for a single snoRNA, these predictions might be misleading until they are evaluated under the light of evolutionary conservation or the original snoRNA sequences are mapped to species with verified targets.

A detailed summary of these sequences and their predicted targets with respect to evolutionary conservation is shown in Supplemental Table S16. Table 1 shows a summary of highly conserved target interactions that are predicted by snoStrip.

For both orphan *N. crassa* snoRNAs, no unambiguous targets were identified by snoStrip. The best prediction yields an ICI_{sno} score of 0.71 for family Nc_CD_53 (CD_53) and is loosely found in several Pezizomycotina species (25S-3500, mean mfe: -11.56). The second family Nc_CD_55 (CD_55) is exclusively found in *Neurospora* preventing a functional analysis of potential targets based on conservation aspects.

In the case of both budding yeast snoRNAs (snR4, snR45), no potential target is found across canonical target sequences, although family snR4 is found to be present in several fungal lineages such as Taphrinomycotina, Saccharomycotina, and

TABLE 1. Assigning putative targets to previously orphan box C/D snoRNAs

Original name	Box	Target position	ICI score	snoStrip name
Nc_CD_10	D'	18S-479	1.13	CD_10
Nc_CD_26	D'	25S-3836	0.86	CD_26
Nc_CD_53	D'	25S-3500	0.71	CD_53*
Nc_CD_54	D'	U60-70	1.43	CD_54*
AM921936	D'	25S-4198	1.50	CD_36
AM921937	D'	18S-479	1.13	CD_31
AM921938	D'	25S-3474	1.19	CD_7
AM921939	D'	18S-179	1.09	CD_15*
AM921940	D	18S-849	1.21	CD_41*
AM921941	D'	18S-630	1.36	CD_24
AM921942	D	18S-456	1.71	CD_37
AM921944	D'	18S-1083	1.57	CD_49
AM921945	D'	25S-3836	0.86	CD_26

Families that do not contain sequences with experimentally verified targets are marked with an asterisk.

several Pezizomycotina species. Family snR45, on the other hand, is exclusively found in Saccharomycetaceae.

The picture looks much better in the case of *A. fumigatus* orphan snoRNAs. The snoStrip pipeline was able to map seven out of nine orphan box C/D snoRNAs to families with experimentally validated targets. These target interactions are also predicted in *A. fumigatus*. Both remaining families (marked by "*" in Table 1) are traceable in the majority of Pezizomycotina species, and putative target sites are also conserved making the snoStrip results plausible despite a missing experimental verification.

The set of 11 *N. crassa* snoRNAs with single sequence predictions without homology with other known snoRNAs comprised 16 distinct targets published in the original publication (Liu et al. 2009). Ten of these targets were confirmed through a conserved prediction using snoStrip. Three targets were annotated as tRNA modification sites, and hence they are not checked in this study. However, these target regions show no conserved and obvious base-pairing capabilities to canonical target RNAs such as rRNAs or snRNAs. The remaining three target sites were predicted based on falsely detected D' box motifs and, thus, are neither biologically correct nor conserved across species. In two cases, evolutionary conserved box motifs are identified, and convincing target sites are predicted by snoStrip (Nc_CD_10, D' target, ICI: 1.13; Nc_CD_26, D' target, ICI: 0.86), see Table 1.

Family Nc_CD_54 (CD_54) was originally published to guide modification at 25S-1648 (*N. crassa* 25S-667; D target) (Liu et al. 2009). By means of snoStrip, family Nc_CD_54 is detected among all Pezizomycotina lineages, and a highly conserved target region is clearly visible upstream of box D' that was originally denoted as orphan. This region shows convincing base-pairing capabilities to U6-70 (*N. crassa* U6-55) in virtually all organisms. The high ICI_{sno} score of 1.43 and the low mean mfe of -18.10 kcal/mol further

promote the correctness of this prediction, see Table 1. The initially annotated target for box D, on the other hand, is not found to be conserved outside of *Neurospora*.

Within the initial box H/ACA snoRNA data sets, orphan sequences were published for *N. crassa* (six sequences), *A. fumigatus* (1), and *S. pombe* (8). A detailed summary of these sequences can be seen in Supplemental Table S22.

By means of snoStrip, eight orphan sequences are found to be conserved on sequence level, and five of them include budding yeast sequences, providing experimentally validated target sites (Nc_HACA_11 matches snR11, Nc_HACA_12 matches snR30, Nc_HACA_13 matches snR10, AM921943 matches snR32, and AJ632018 matches snR43). The three remaining snoRNA families comprise a conserved target in HP2, see Table 2. Family Nc_HACA_7 is found to be a distant homolog to family snR86 (HACA_36), which is merely detected in Saccharomycetes organisms. Nonetheless, both families are sufficiently predicted to guide the validated isomerization of uridine at position 25S-3500. Due to substantial differences in sequence lengths (HACA_36 is ~1 kb in length; Nc_HACA_7 is ~180 nt in length), snoStrip was unable to detect a potential common origin. Family AJ632012 (HACA_54) is exclusively found in Schizosaccharomyces, Candida, and Debaryomycetaceae. All species with a sufficient LSU sequence are predicted to guide the pseudouridylation at position 25S-3439 (*S. cerevisiae* 25S-2254). This position is not known to be modified in the budding yeast, which may explain why homologs in this clade are missing. Family AJ632016 (HACA_53) is found across Taphrinomycotina and Pezizomycotina and is convincingly predicted to accompany target binding at position 18S-1302. However, this position is not yet known to be modified in yeast or human.

Seven of 15 orphan box H/ACA snoRNAs are found to be conserved only on genus or species level, i.e., two orphan *N. crassa* sequences are exclusively found in the two other *Neurospora* organisms, while five *S. pombe* snoRNAs are either found in all *Schizosaccharomyces* species (2) or in the fission yeast only (3). Such a small set of species that share a homologous snoRNA sequence makes an appropriate target prediction impossible. Hence, a sufficient conclusion about their true function and, further on, about their genuine

existence in terms of viability and biological necessity, remains elusive.

Lineage-specific targets

Several box C/D snoRNA families harbor a highly conserved target either at their D or D' position. It may be the case, however, that many of these families exhibit additional lineage-specific target binding capabilities on their "nonfunctional" ASE. Such a functionality might have evolved at a specific time point during evolution, and because of a potential benefit, it is retained in all of today's organisms descending from this ancestor.

Interesting box C/D snoRNA families with previously annotated functional D' targets and lineage-specific D targets are shown in Figure 4. Detailed information about all snoRNA families with an additional, lineage-specific target is shown in Supplemental Table S14.

Family snR87 (CD_10), for example, with its experimentally verified target 18S-479 (18S-436; D' target) (Davis and Ares 2006), is detected in all analyzed fungal lineages except for Microsporidia. Besides the functional D' region, all Pezizomycotina species, whose large subunit rRNA is available, are also predicted to guide an additional target upstream of their D box. The target 25S-2066 (*N. crassa* 25S-1042) has an ICI_{sno} score of 1.21 among members in the Pezizomycotina subtree. The mean mfe is -13.19 kcal/mol. Family snR53 (CD_11) was shown to guide the methylation at position 18S-894 (18S-796; D' target) in the budding yeast (Lowe and Eddy 1999). The snoStrip-analysis confirmed the snoRNA and this specific target interaction in a wide range of fungi. An additional D' target, U6-62 (*S. cerevisiae* U6-45), was originally published in *N. crassa* (Liu et al. 2009) based on single sequence prediction. This interaction is also convincingly confirmed by snoStrip in all snoRNAs that were previously found to guide the 18S-894 target, except for Saccharomycetaceae, see Figure 4. Position 45 in U6 snRNA was not found to be modified in the budding yeast (Massenet et al. 1998; Machnicka et al. 2013). Due to missing analyses, no such statement can be made in most other fungal species. Since the ICI score for the U6 target is only marginally smaller than for the 18S target, 0.89 to 0.94, respectively, and the mean mfe value is found to be -13.78 kcal/mol (18S-894: -17.34 kcal/mol), it is not unlikely that this snoRNA is capable of modifying both targets. Two additional targets can be found for the ASE upstream of box D: 25S-1153 and 25S-1796 (*N. crassa* 25S-359 and 25S-790). Both candidates are predicted throughout all Pezizomycotina species and, surprisingly, *Taphrina deformans*, a relative to the fission yeast. The first interaction is additionally found in *Yarrowia lipolytica*, which is a close relative to the budding yeast. Because of its extraordinary low mean minimum free energy of -21.12 kcal/mol, this target is assigned a high ICI value of 1.66. The second putative interaction has an ICI score of 0.83 and a mean mfe of -11.50 kcal/mol.

TABLE 2. Assigning putative targets to previously orphan box H/ACA snoRNAs

Original name	Box	Target position	ICI score	snoStrip name
Nc_HACA_7	HP2	25S-3500	1.26	HACA_7
AM921943	HP2	25S-3374	1.12	HACA_21*
AJ632012	HP2	25S-3439	1.22	HACA_54
AJ632016	HP2	18S-1302	0.82	HACA_53
AJ632018	HP1	25S-1962	1.17	HACA_9*

Families that do contain sequences with experimentally verified targets are marked with an asterisk.

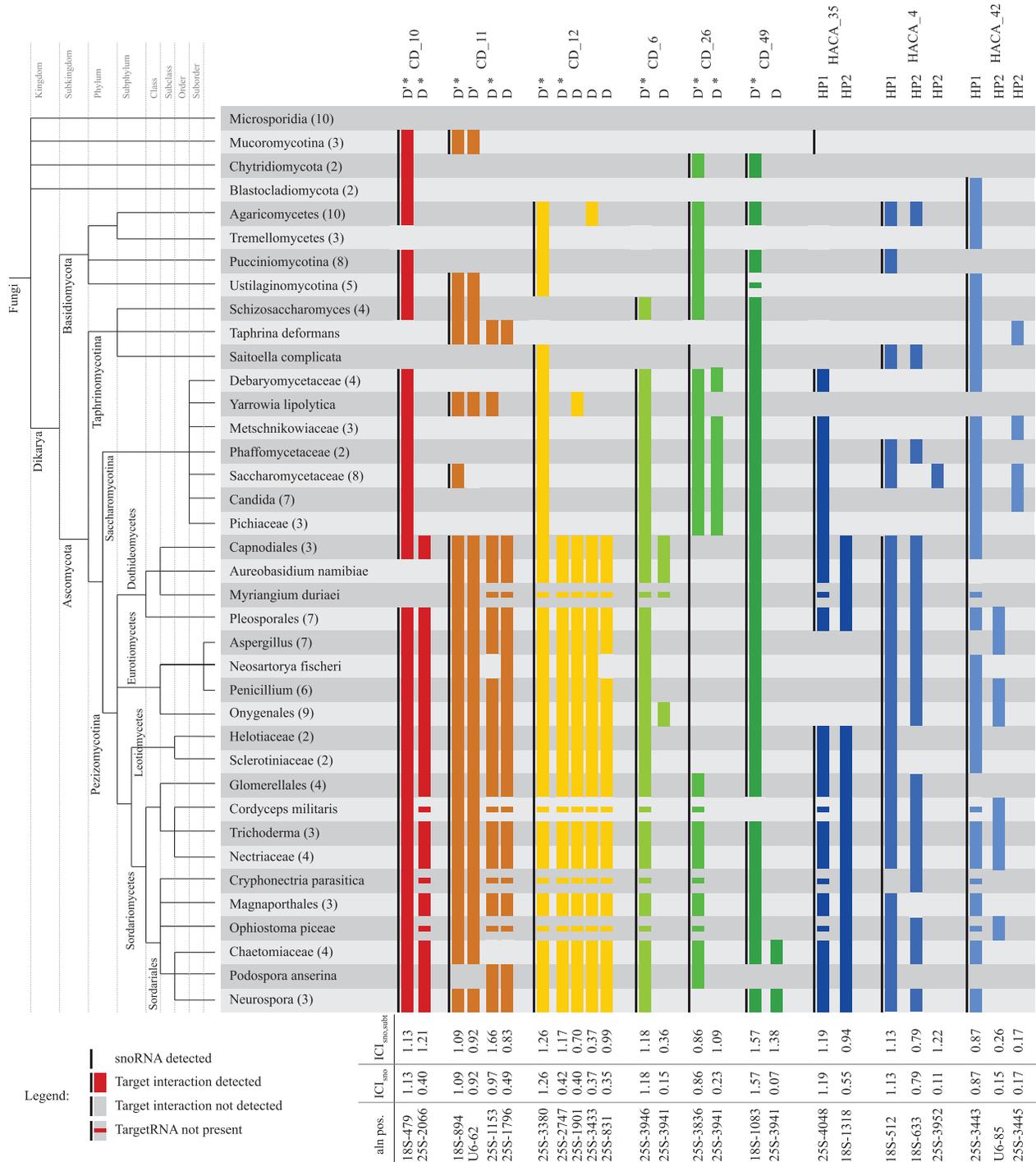


FIGURE 4. The conservation of predicted target interactions is shown for interesting single-guide snRNA families that exhibit an additional functional target at their “nonfunctional” ASE. Each family is depicted in a different color. The black bar in front of each family shows the presence of the family in a certain lineage or organism. The color bar shows that at least one target interaction was predicted in that lineage. The respective family name and target site can be seen on top, while the alignment position and the corresponding ICI score are shown at the bottom. Experimentally confirmed interactions are denoted with an asterisk.

A very interesting modification site is 25S-3941 (*S. cerevisiae* 25S-2724), whose actual methylation and the guidance by snR67 (CD_26) was experimentally shown (Lowe and Eddy 1999). The conserved interaction of this position is

traceable in at least three different families, each in another fungal lineage. Family snR67 is present in all Dikarya lineages and Chytridiomycota and shares a conserved D' target 25S-3836 (*S. cerevisiae* 25S-2619) that is predictable in all

Dikarya with the exception of Dothideomycetes, Eurotiomycetes, and Leotiomycetes (ICI: 0.86, mean mfe: -23.03 kcal/mol). The D target 25S-3941, on the other hand, is solely found in Saccharomycotina (ICI: 1.09, mean mfe: -15.34 kcal/mol). Family snR51 (CD_6) is found to share this target as a conserved D box interaction in Onygenales and in a part of Dothideomycetes (ICI: 0.36, mean mfe: -15.46 kcal/mol). In a third family, snR54 (CD_49), the modification at 25S-3941 is predicted in Sordariales (ICI: 1.38, mean mfe: -14.14 kcal/mol, D target).

Similar to the box C/D snoRNA class, several box H/ACA snoRNAs have a functional and highly conserved target guiding region in one hairpin and show lineage-specificity in the other, see Figure 4. A detailed summary can be found in the Supplemental Table S20. Some of these functions might already be annotated, in particular for snoRNA sequences of the budding yeast, e.g., snR189 (HACA_4) and snR191 (HACA_42), which are in fact officially denoted as double guides in *S. cerevisiae* (Badis et al. 2003; Schattner et al. 2004). HP1 is highly conserved in both families, and the corresponding target binding capability is at least present in Dikarya. In their second hairpin, however, they developed two different guiding functions that are predictable in separate lineages. Family snR189, for example, is known to guide the pseudouridylation at 25S-3952 in Saccharomycetaceae, while outside of this clade the snoRNA is mostly predicted to guide modification at 18S-633. In snR191, on the other hand, the separation of both target guiding functions becomes even more conspicuous. The budding yeast annotated modification site is predicted in Saccharomycotina and *Taphrina deformans* (25S-3445), whereas the position U6-85 is predicted in a wide range of Pezizomycotina.

Family snR32 (HACA_21) is predicted to guide the modification at position 57 (*N. crassa* 54, *S. cerevisiae* 54) in the 5.8S rRNA with its first hairpin in a large number of Pezizomycotina species (ICI_{sub} = 0.73). This particular modification is not present in budding yeast 5.8S molecules, which undoubtedly explains the missing predictions in this subtree. On the contrary, the corresponding human position is found to be pseudouridylated raising the possibility for this predicted interaction to be an authentic and biologically correct modification. Based on the ICI_{sub} score, a potential, alternative target at position 25S-2813 is convincingly predicted with 1.07 in 19 out of 27 Saccharomycetales organisms. Since experimental evidence for this precise position is missing, the prediction remains hypothetical.

Target switches

Occasionally during evolution, novel guiding interactions are acquired or ancestral ones are lost in different species or lineages. It is, however, much more uncommon that some target interactions are translocated from one snoRNA to another. Therein, the position of the ASE within the snoRNA se-

quence, upstream of box D'/D or in HP1/HP2, is mostly preserved, but it happens sporadically that this position is also shifted. In the following, we will present an in-depth description of highly complex rearrangements including target translocations and duplications between several snoRNA families that have been automatically uncovered by snoStrip.

Each of these two “snoRNA clans” comprises two, three, or even more snoRNA sequences in each organism with distinct target interactions. Due to target switches during fungal evolution, these previously independent snoRNA sequences became connected. Table 3 summarizes the target interactions that are convincingly predicted in the snoRNA clans CD_5 (containing budding yeast sequences snR60, snR72, and snR78) and CD_19 (snR52, snR56).

In the following, we will focus on the description of the snoRNA clan CD_5. The potential evolutionary history of CD_19 is illustrated and discussed in detail in Supplemental Section S9.

The snoRNA clan CD_5 comprises three distinct budding yeast snoRNA sequences (snR60, snR72, and snR78), which at first sight do not share a common evolutionary background. SnR60 was verified to guide methylations at 25S-1898 (single sequence 25S-908, D target) and 25S-1806 (25S-817, D' target), snR72 guides the methylation at 25S-1866 (25S-876, D target), and snR78 was shown to direct the modification at position 25S-3615 (25S-2421, D' target) (Lowe and Eddy 1999). The methylations at positions 25S-1806, 25S-1898, and 25S-3915 map to known and verified modifications in human large subunit ribosomal RNAs. Hence, they are most likely ancient, which suggests that both the methylations and their guiding snoRNAs also existed at the root of fungi. However, through individual target switches in the cause of fungal evolution, the history of these sequences became connected. A phylogenetic tree displaying a potential evolutionary history involving snoRNAs that are predicted to guide the aforementioned modifications is shown in Figure 5. The putative ancestral state probably consisted of two individual snoRNA sequences guiding the three ancient methylations. Deletion

TABLE 3. Interaction properties of four LSU modifications of CD_5 are shown

	Target position	ICI _{sno}	Ø mfe	Detected interactions
CD_5	25S-1806	0.79	-16.46	23.08%
	25S-1866	0.90	-19.49	24.61%
	25S-1898	1.20	-25.80	25.38%
	25S-3615	1.00	-18.48	25.77%
CD_19	18S-462	1.52	-20.62	34.49%
	18S-602	1.11	-15.30	34.18%
	18S-1580	1.75	-20.76	34.49%
	25S-2574	0.48	-22.85	9.49%
	25S-4143	0.28	-15.49	7.59%

Properties for three SSU and two LSU methylations are given for clan CD_19.

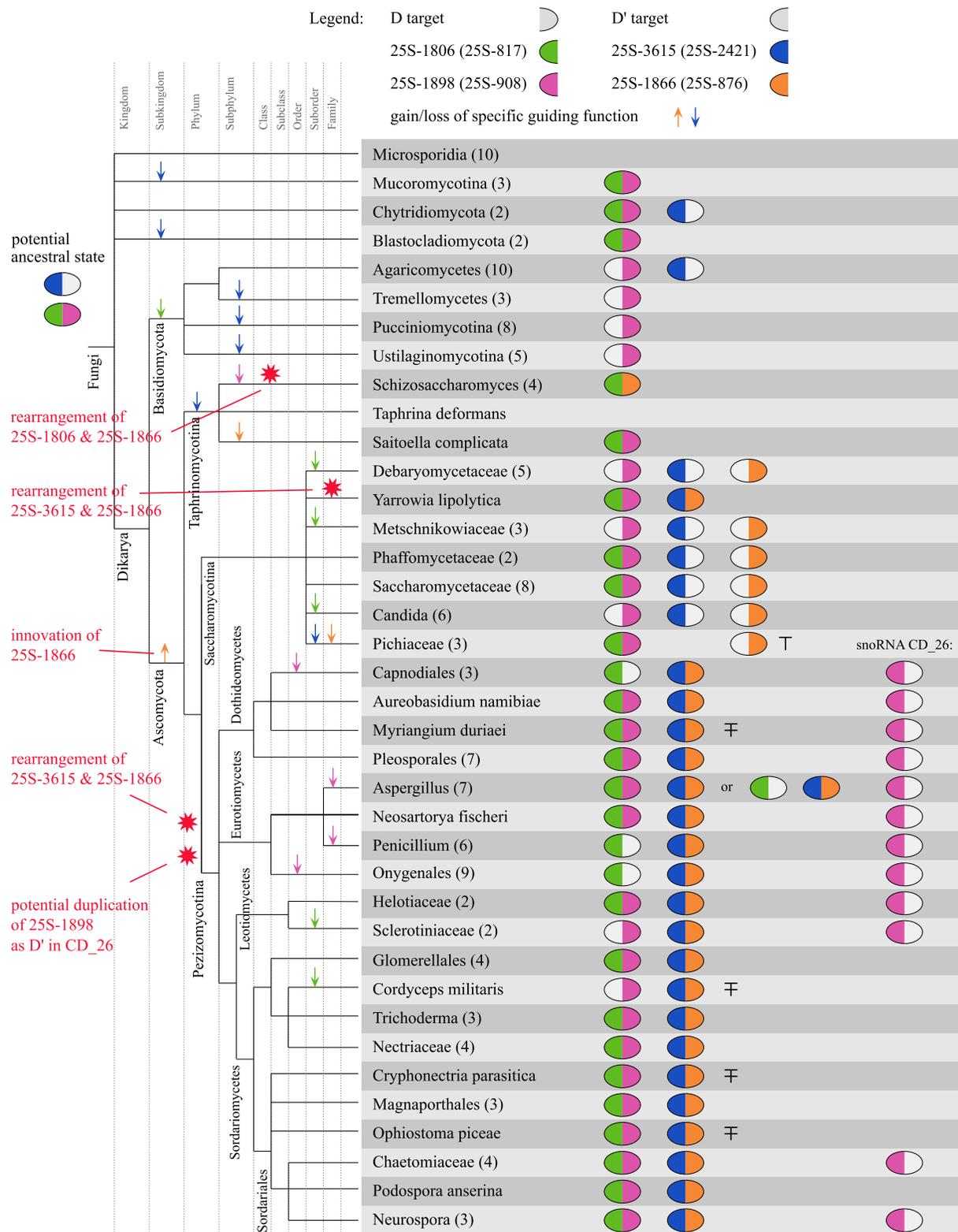


FIGURE 5. Potential evolutionary history of snoRNA clan CD₅ involving four modification sites on the LSU rRNA. Gain/loss events are displayed with arrows, while potential rearrangements are shown with red stars. (⊠) 25S-1866 is solely found in Pichia. (⊠) Putative since LSU sequences are missing; snoRNAs show convincing ASE conservation.

and innovation events of target interactions (inferred assuming parsimony) are marked accordingly. The emergence of the fourth modification, 25S-1866, is predicted at the root of Ascomycota since all diverging lineages are either predicted or verified to target this specific site. The loss of any of the four guiding functions occurred rather frequently in several lineages, e.g., Basidiomycota are supposed to have lost the guiding potential for 25S-1806, while different Basidiomycota lineages are further predicted to have lost the ability to guide methylation at 25S-3615.

In addition to gain and loss events, target interactions responsible for these four modifications switched between different snoRNAs several times during fungal evolution. However, the target sites within the snoRNAs (D' or D target) are mostly preserved. Within the Taphrinomycotina lineage, including the fission yeast, target guiding functions at 25S-1806 (D' target) and 25S-1866 (D target) were incorporated into one snoRNA sequence after the original guidance of 25S-1898 (D target) has been lost in this family.

At the root of Ascomycota, a polycistronic snoRNA transcript harbored the snoRNA sequences of snR77 (CD_24), snR76 (CD_12), snR75 (CD_7), snR74 (CD_21), and snR73 (CD_31), ordered in 5'-3' direction, see Figure 6. All these snoRNA families have been present already at the root of Dikarya, distributed over large distances or over different chromosomes. After the formation of this cluster, the precise order and the length of ~1.5 kb remained highly conserved throughout all Ascomycota.

Maybe a snoRNA of clan CD_5 guiding methylation at 25S-3615 was already present at the 5' end of this cluster when it emerged. However, there are several possibilities how this snoRNA cluster evolved after the innovation of guiding function for 25S-1866. One hypothesis (blue stars in Fig. 6) is the initial incorporation of 25S-1866 into the snoRNA that already guides 25S-3615, creating a double guide snoRNA at the 5' end of the polycistronic transcript. In Taphrinomycotina, the loss of guiding function for 25S-3915 and 25S-1898 might have caused the rearrangement of 25S-1806 and 25S-1866 and the exclusion from the snoRNA cluster. At the root of Saccharomycotina, the double guide snoRNA might have split up leaving a single guide at the 5' end (25S-3615) and a novel single guide at the 3' end of the cluster (25S-1866). In this scenario, the original arrangement is conserved only in *Yarrowia lipolytica*. An alternative is outlined by the green stars in Figure 6. Assuming that the innovation of 25S-1866 led to a novel single guide snoRNA located at the 3' end of the snoRNA cluster, as seen in Saccharomycetaceae, *Y. lipolytica* would be the only organism in Saccharomycotina where a rearrangement is detected. As a consequence, the previously single guide sequences are reorganized into a double guide sequence with guiding ability for 25S-3615 as D' target and 25S-1866 as D target. This novel double guide is now located at the 5' end of the cluster. Coincidentally, the same reorganization can be observed at the root of Pezizomycotina, where the first snoRNA of the cluster is found to guide modifications at position 25S-3615 (D')

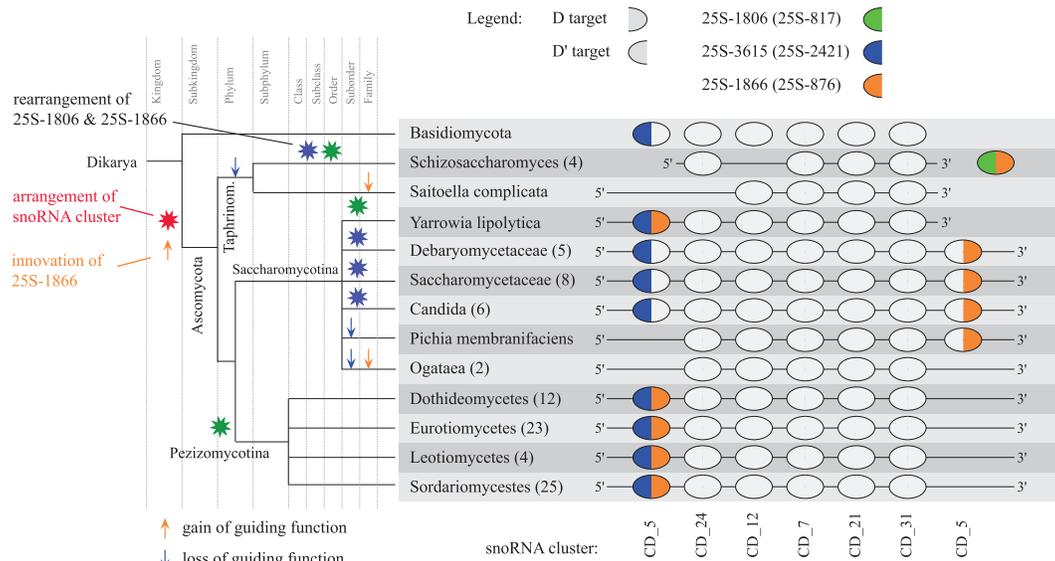


FIGURE 6. Sequences of the CD_5 snoRNA family are incorporated into a polycistronic transcript that harbors up to seven snoRNA genes. This cluster with its highly conserved structure and size occurred at the root of Ascomycota, but most of its genes arose at least at the root of Dikarya. There are different potential histories regarding the evolution of the cluster depending on how the newly innovated target guiding function at position 25S-1866 (orange) was initially introduced in this polycistronic transcript. (A) Evolutionary history under the assumption that 25S-1866 is incorporated as a second guiding function into the snoRNA guiding 25S-3615. (B) History under the hypothesis that a novel single guide snoRNA is introduced at the 3' end of the snoRNA cluster. The most parsimonious rearrangement events that led to the observed cluster organization are depicted in blue and green stars, according to hypotheses A and B, respectively.

25S-1866 (D). Proteins that are located up- and downstream from the previously described snoRNA cluster are not found to be conserved throughout major fungal lineages.

A further interesting observation is the potential duplication of target interaction for 25S-1898 at the root of Pezizomycotina. This ability is inserted into family snR67 (CD_26) as a D' target in the lineages Dothideomycetes, Eurotiomycetes, and Leotiomycetes (ICI_{Pezizomycotina}: 1.13, mean mfe: -18.79 kcal/mol). *Neurospora* species are also predicted to guide this methylation with its Nc_CD_26 (CD_26) snoRNA (Liu et al. 2009). In reverse, the original D' target of snR67, 25S-3836, was abolished in these organisms and is not found to be reestablished in any other snoRNA family. For more detailed information on family CD_26, please refer to Figure 4. The invention of redundant guides would explain the findings that in some of these species the original target site of 25S-1898 vanished in CD_5 snoRNAs, e.g., in Capnodiales, some *Aspergillus* species, and Onygenales. Families CD_5 and CD_26 are not merged due to a switch of the ASE (from D in CD_5 to D' in CD_26).

Multiple target interactions

In some cases, snoRNA families are not only convincingly predicted to guide one specific target modification but two or even more with the same ASE.

An outstanding example is the box C/D snoRNA family snR40 (CD_43), which is known to guide methylation at position 18S-1400 (18S-1271) with its D' target binding region (Lowe and Eddy 1999). This interaction is predicted in 67 out of 90 snoRNAs and provides an ICI score of 0.95 with a mean interaction energy of -12.96 kcal/mol. However, an even better target is predicted at position 18S-614 (18S-562) with an ICI score of 1.61 and a mean mfe of -21.69 kcal/mol. This interaction is found in 71 organisms. All 67 snoRNAs predicted to guide the first target are also predicted to guide the latter one, in the overwhelming majority of cases even with a better binding energy. However, a genuine modification of this site has not been reported in *S. cerevisiae*, *N. crassa*, or human. Albeit very compelling, the prediction thus remains a hypothesis.

An even more impressive example is the D' ASE of the snR70 (CD_61) family. As many as five potential targets are predicted with an ICI score above 1.0, a mean mfe below -11.30 kcal/mol and more than 80 single sequence predictions. Details are shown in Table 4. This time, the most persuasive prediction is experimentally confirmed (Lowe and Eddy 1999), whereas the other predicted positions have not been shown to be chemically modified so far.

Data validation

To add further reliability to our snoRNA annotations, we compared our data with Rfam-annotated families and additionally cross-checked with Ribo-seq archives of four differ-

TABLE 4. Summary of multiple target predictions of families snR40 (CD_43) and snR70 (CD_61) that are guided by the same ASE

	Target position	ICI _{sno}	Ø mfe	#ia
snR40	18S-1400	0.95	-12.96	67/90
	18S-614	1.61	-21.96	71/90
snR70	18S-1843	1.48	-17.82	86/102
	5.8S-155	1.16	-12.99	92/102
	18S-348	1.04	-12.99	83/102
	18S-1827	1.02	-12.49	85/102
	5.8S-120	1.00	-11.36	91/102

#ia is the number of predicted interactions.

ent Fungi. This should allow us to point out that our snoStrip annotations are in line with existing Rfam annotations. Novel snoRNA sequences, meaning that they are not listed anywhere else, should be confirmed by Ribo-seq to be transcribed and classified as “noncoding.”

Comparison to Rfam

The Rfam database (Nawrocki et al. 2015) is by far the most comprehensive source of well-curated data on noncoding RNAs. The 7852 snoRNAs detected by snoStrip encompass the overwhelming majority of Rfam-annotated sequences (1586 of 1621 snoRNA genes), and 441 of them are located in seed alignments. Most of the 35 Rfam snoRNA sequences that are not included in our data set exhibit box motifs with two or more mutations compared to the consensus sequence. This prevents snoStrip from including these sequences into our data set even though their detection by means of Blast or Infernal is straightforward.

Conversely, 32 of the snoStrip-annotated snoRNA families had no representative in Rfam even though they harbor at least one experimentally verified snoRNA that served as the starting point for a homology search in this study. Most of these are published for *N. crassa*, where 64 of the 75 snoRNAs proclaimed in (Liu et al. 2009) have not found their way into Rfam.

In 18 cases, snoStrip automatically combined snoRNA sequences of two different Rfam models to a common snoRNA family. A detailed inspection showed that merged Rfam models were constructed from disjoint sets of organisms and indeed share the same target guiding functionality. For example, family CD_12 (*S. cerevisiae*: snR76) combines sequences of the Rfam models RF01209 and RF01514, which were derived from Saccharomycetales and Pezizomycotina species, respectively. A pairwise comparison of the 18 families against all other Rfam models with CMcompare (Eggenhofer et al. 2013) showed that the merged pairs most likely exhibit the highest similarity. A table listing all 18 snoRNA families and their respective Rfam models can be seen in section S10 of the supplement alongside figures showing the CMcompare score distribution.

Ribosome footprinting

Ribo-seq or ribosome profiling combines ribosome footprinting with deep sequencing (Ingolia et al. 2009, 2012) to study translation *in vivo*. Not all the detected events, however, correspond to ribosome-protected mRNA fragments and hence to translated regions. As a byproduct, other nonribosomal RNA–protein complexes are also detected, including transfer RNAs, spliceosomal RNAs, and snoRNAs (Ingolia et al. 2014; Ji et al. 2016).

Considering the *S. cerevisiae* data, all 74 annotated snoRNAs show overlapping Ribo-seq reads in at least four of the 14 libraries provided by Ingolia et al. (2009), with evidence from every data set supporting 52 snoRNA s. 65 snoRNAs are classified as “noncoding” based on their percentage of maximum entropy (PME) scores calculated by the Rfoot tool (Ji et al. 2016). The complete results can be seen in Supplemental Section S11. In the case of *S. pombe*, 61 of 62 snoRNA are present in at least one of the three short read archives, with 46 snoRNAs supported in all available data sets. Based on their PME score, 49 annotated genes are classified as “noncoding.” For *Ajellomyces capsulatus* and *C. albicans*, all snoStrip-annotated sequences have Ribo-seq support from at least one experiment, while 60 (of 62) and 45 (of 67) are present in all data sets, respectively. In *A. capsulatus*, all snoRNA genes are classified as “noncoding,” in *C. albicans*, on the other hand, two sequences lack support from the PME scoring. Almost all of the snoRNAs without PME score support lack sufficient read coverage for the Rfoot analysis. Even though, the available Ribo-seq data strongly supports our snoStrip-annotated data sets and hence adds another convincing layer of reliability to our fungal snoRNAome.

DISCUSSION

In this study, we provide a comprehensive inventory of snoRNAs in fungi together with an in-depth analysis of the evolution of snoRNA families and their target specificities. The investigation of 147 different taxa provides a detailed history of potential gain, loss, and duplication events for 68 families of box C/D snoRNAs and 50 families of box H/ACA snoRNAs involving more than 7800 individual snoRNA sequences. For 18 snoRNA families previously unrecognized homology with other families has been uncovered. These data constitute a substantial extension and refinement of the accumulated knowledge on snoRNAs. Data and refined models will become available in the Rfam database and collectively form an important step toward a global understanding of the evolution of the snoRNAome. Since our approach is based on homology search, it is fundamentally limited by the seed sequences that have been observed and classified as snoRNAs in at least one organism. It is very unlikely, therefore, that this study presents a complete picture despite increasing the number of snoRNA sequences

by more than a factor of four. In addition, for 26 of 39 orphan snoRNAs (including sequences with single sequence target predictions only) a mapping to experimentally verified targets could be found, or at least a quite convincing prediction based on the interaction conservation index (ICI) could be assigned.

The processing of this amount of data is well beyond the realm of manual curation and has been possible only with the help of snoStrip, a pipeline specifically developed to investigate the evolution of snoRNA families across a broad phylogenetic range (Bartschat et al. 2014). The in-depth analysis of potential target interactions adds a new layer of information. We have demonstrated here that the coevolution of snoRNAs and their targets can be traced with high resolution based on the functional characteristics of the snoRNAs as determined by snoStrip together with a quantitative assessment of predicted RNA–RNA interactions based on the interaction conservation index (ICI) (Kehr et al. 2014).

Similar to Metazoa, fungal box H/ACA snoRNAs show a higher loss-ratio compared to box C/D snoRNAs. This might have both a technical and a biological explanation that manifests itself on two different levels. Since box H/ACA snoRNAs do not share long ASEs but rather short bipartite pseudouridylation pockets, it becomes considerably harder to detect homologous snoRNAs over large evolutionary time-scales. This effect may limit the scope of the homology search procedure. The short interacting regions make these molecules also more vulnerable to mutations that disrupt the snoRNA–target interaction. At the same time, the presence of the second, independent ASE in the other hairpin may be a sufficient cause to retain mutated genes.

In general, fungal snoRNAs have well-preserved target interactions, and most families are found to contain exactly one highly conserved anti-sense element. The remaining target region is in turn free to evolve or to adapt to new lineage-specific or even species-specific targets. Here, we introduced a variation on the ICI measure adapted to subclades, allowing a much more detailed quantitative assessment of target turnover. Many of the predictions made here, of course, await experimental validation, given that experimental evidence for RNA–target interactions as well as direct measurements of chemical modifications in the primary target molecules (rRNAs and snRNAs) are still restricted to a few model organisms.

The computational analysis reported here strongly suggests that snoRNAs not only address a highly conserved ASE but also frequently have additional, secondary targets. The possibility that a single snoRNA target site exerts two distinct guiding functions has been exemplarily reported for budding yeast box H/ACA snoRNAs. The budding yeast snoRNA family snR3 (HACA_3), for example, is verified to target two modification sites in its second hairpin (Schattner et al. 2004). Both interactions can be observed across Dikarya. Nevertheless, there is still very little experimental

data on the generality of this effect, and most of the predicted “double” target sites will still require experimental verification. Convincing examples of remarkably conserved multiple interactions are found in box C/D snoRNA families snR40 (CD_43) and snR70 (CD_61), which exhibit two and five high-scoring target-interactions at a single ASE, respectively. These findings suggest the possibility that snoRNAs are, at least under certain circumstances, able to guide different modifications with the same ASE. This might be dependent on developmental states, or more complex mechanisms involving conformational changes of the target.

In some cases of box H/ACA snoRNAs, these additional targets exhibit better ICI scores than the annotated modification sites. Since the ICI combines evidence from thermodynamic stability and evolutionary conservation, these predictions cannot be easily dismissed as false positives. The specialized ribosome hypothesis proposes distinct ribosomal conformations in different developmental stages and stress levels that might also entail different chemical modification patterns of the rRNAs; it is entirely plausible, in this scenario, that some modifications and, thus, snoRNA interaction sites have remained undetected (Xue and Barna 2012). The existence of stress-induced conditional pseudouridylations indeed has been reported for the U2 snRNA of budding yeast (Wu et al. 2011). The snR81 RNA, which is also responsible for the guidance of a constitutive U2 pseudouridylation, guides one of the novel modifications through imperfect and redundant base-pairing. The authors speculate that conditionally induced modifications in RNA may well be a rather frequent phenomenon.

We also found convincing evidence that some modifications are guided by two, three, or even more snoRNA families. First, this includes redundant guides, meaning that two snoRNA families of the same species are responsible for the same modification. Second, we observed several target sites that are addressed by different snoRNA families in different taxonomic groups. A good example of the latter situation is the predicted pseudouridine at position 5.8S-18. Although there is no direct experimental evidence that this particular position is modified *in vivo*, the site is predicted as a target for several distinct snoRNA families by RNAsnoop (see Supplemental Section S21).

The fact that specific modification sites are predicted to be guided by more than just one snoRNA family in the same organism has several possible reasons. SnoRNA expression was recently reported to be strongly regulated in development and between tissues or cell lines (Kapushesky et al. 2012; Jorjani et al. 2016). It may thus be necessary for the organism to compensate for snoRNAs that are lowly expressed under certain circumstances to maintain the functional modification levels of the target RNA. This may be achieved through paralogous snoRNAs or redundant target binding capabilities of other snoRNA families.

In summary, we observe that the landscape of snoRNAs keeps constantly changing in the kingdom of Fungi. We

observe both the extinction of entire snoRNA families and the innovation of new ones. Even the function of snoRNA families itself changes at these evolutionary scales, showing loss, gain, and turn-over of guiding functions that lead to target switches. The number of known snoRNA families in Fungi is lower than in animals, correlating well with the observation that animals have more (reported) modification sites in their rRNAs and snRNAs than “lower” Eukaryotes (see Modomics and the RNA Modification Database (Cantara et al. 2011; Machnicka et al. 2013) or even Bacteria (which have target-specific enzymes for each individual modification instead of the generic enzyme machinery with snoRNAs as evolutionary flexible “address labels”).

There are many similarities between the fungal and the metazoan snoRNAome. A common feature is a detectable burst in the snoRNA diversity at each major branching point in the taxonomic tree of both kingdoms. In the case of fungal box C/D snoRNAs, the distribution of orphan, single-guided, and double-guided snoRNAs is quite similar compared to animals, as reported by the human snoRNA atlas (Jorjani et al. 2016): Over 75% of the box C/D snoRNAs are found to be single guided (over 70% in Metazoa). In both human and fungi the remainder is about equally distributed among double-guided and orphan snoRNAs. The situation is somewhat different for box H/ACA snoRNAs: In human, double-guided snoRNAs comprise the largest group (47%), while in Fungi, only 22% of the box H/ACA snoRNA families target two distinct pseudouridylation sites with both hairpins.

SUPPLEMENTAL MATERIAL

An electronic supplement containing the data sets used and produced in this study is available at <http://www.bioinf.uni-leipzig.de/publications/supplements/17-001>.

ACKNOWLEDGMENTS

This work was funded in part by the European Union FP-7 project QUANTOMICS (no. 222664), the MMML-seq project of the International Cancer Genome Consortium (ICGC) funded by the German Federal Ministry of Education and Research, the CRC 1052 “Obesity,” and by the Deutsche Forschungsgemeinschaft (Project DFG STA 850/15-1). LIFE–Leipzig Research Center for Civilization Diseases is funded by the State of Saxony and the European Union.

Received June 30, 2017; accepted November 19, 2017.

REFERENCES

- Altshul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215: 403–410.
- Badis G, Fromont-Racine M, Jacquier A. 2003. A snoRNA that guides the two most conserved pseudouridine modifications within rRNA confers a growth advantage in yeast. *RNA* 9: 771–779.

- Balakin AG, Smith L, Fournier MJ. 1996. The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. *Cell* **86**: 823–834.
- Bartschat S, Kehr S, Tafer H, Stadler PF, Hertel J. 2014. snoStrip: a snoRNA annotation pipeline. *Bioinformatics* **30**: 115–116.
- Bratkovič T, Rogelj B. 2011. Biology and applications of small nucleolar RNAs. *Cell Mol Life Sci* **68**: 3843–3851.
- Cahill NM, Friend K, Speckmann W, Li ZH, Terns RM, Terns MP, Steitz JA. 2002. Site-specific cross-linking analyses reveal an asymmetric protein distribution for a box C/D snoRNP. *EMBO J* **21**: 3816–3828.
- Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, Zhang X, Vendeix FA, Fabris D, Agris PF. 2011. The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res* **39**: D195–D201.
- Canzler S, Stadler PF, Hertel J. 2017. Evolution of fungal U3 snoRNAs: structural variation and introns. *Non-Coding RNA* **3**: 3.
- Cavaillé J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachelier JP, Brosius J, Hüttenhofer A. 2000. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc Natl Acad Sci* **97**: 14311–14316.
- Clouet d'Orval B, Bortolin ML, Gaspin C, Bachelier JP. 2001. Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNA^{Trp} intron guides the formation of two ribose-methylated nucleosides in the mature tRNA^{Trp}. *Nucleic Acids Res* **29**: 4518–4529.
- Darzacq X, Jány BE, Verheggen C, Kiss AM, Bertrand E, Kiss T. 2002. Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. *EMBO J* **21**: 2746–2756.
- Davis CA, Ares M. 2006. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci* **103**: 3262–3267.
- Decatur WA, Fournier MJ. 2002. rRNA modifications and ribosome function. *Trends Biochem Sci* **27**: 344–351.
- Dennis PP, Omer A, Lowe T. 2001. A guided tour: small RNA function in *Archaea*. *Mol Microbiol* **40**: 509–519.
- Duncan CD, Mata J. 2014. The translational landscape of fission-yeast meiosis and sporulation. *Nat Struct Mol Biol* **21**: 641–647.
- Eggenhofer F, Hofacker IL, Höner Zu Siederdisen C. 2013. CMCompare webserver: comparing RNA families via covariance models. *Nucleic Acids Res* **41**: W499–W503.
- Ganot P, Caizergues-Ferrer M, Kiss T. 1997. The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev* **11**: 941–956.
- Gardner PP, Bateman A, Poole AM. 2010. SnoPatrol: how many snoRNA genes are there? *J Biol* **9**: 4.
- Gilmore SA, Voorhies M, Gebhart D, Sil A. 2015. Genome-wide reprogramming of transcript architecture by temperature specifies the developmental states of the human pathogen histoplasma. *PLoS Genet* **11**: e1005395.
- Hertel J, Stadler PF. 2015. The expansion of animal microRNA families revisited. *Life (Basel)* **5**: 905–920.
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures (The Vienna RNA Package). *Monatshefte f Chemie* **2**: 167–188.
- Ingolia NT, Ghaemmghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218–223.
- Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. 2012. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**: 1534–1550.
- Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. 2014. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**: 1365–1379.
- Ji Z, Song R, Huang H, Regev A, Struhl K. 2016. Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat Biotechnol* **34**: 410–413.
- Jorjani H, Kehr S, Jedlinski DJ, Gumienny R, Hertel J, Stadler PF, Zavolan M, Gruber AR. 2016. An updated human snoRNAome. *Nucleic Acids Res* **44**: 5068–5082.
- Jöchl C, Rederstorff M, Hertel J, Stadler PF, Hofacker IL, Schrettl M, Haas H, Hüttenhofer A. 2008. Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein synthesis. *Nucleic Acids Res* **36**: 2677–2689.
- Kapusshesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, Holloway E, Klebanov A, Kryvykh N, Kurbatova N, et al. 2012. Gene Expression Atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res* **40**: D1077–D1081.
- Kehr S, Bartschat S, Stadler PF, Tafer H. 2011. PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics* **27**: 279–280.
- Kehr S, Bartschat S, Tafer H, Stadler PF, Hertel J. 2014. Matching of soulmates: coevolution of snoRNAs and their targets. *Mol Biol Evol* **31**: 455–467.
- Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C, et al. 2016. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res* **44**: D574–D580.
- Kishore S, Stamm S. 2006. Regulation of alternative splicing by snoRNAs. *Cold Spring Harb Symp Quant Biol* **71**: 329–334.
- Kishore S, Gruber AR, Jedlinski DJ, Syed AP, Jorjani H, Zavolan M. 2013. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol* **14**: R45.
- Kiss T. 2002. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell* **109**: 145–148.
- Kiss-László Z, Henry Y, Kiss T. 1998. Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO J* **17**: 797–807.
- Li SG, Zhou H, Luo YP, Zhang P, Qu LH. 2005. Identification and functional analysis of 20 Box H/ACA small nucleolar RNAs (snoRNAs) from *Schizosaccharomyces pombe*. *J Biol Chem* **280**: 16446–16455.
- Liu N, Xiao ZD, Yu CH, Shao P, Liang YT, Guan DG, Yang JH, Chen CL, Qu LH, Zhou H. 2009. SnoRNAs from the filamentous fungus *Neurospora crassa*: structural, functional and evolutionary insights. *BMC Genomics* **10**: 515.
- Lowe TM, Eddy SR. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**: 1168–1171.
- Machnicka MA, Milanowska K, Osman Oglou O, Purta E, Kurkowska M, Olchowiak A, Januszewski W, Kalinowski S, Dunin-Horkawicz S, Rother KM, et al. 2013. MODOMICS: a database of RNA modification pathways—2013 update. *Nucleic Acids Res* **41**: D262–D267.
- Massenet S, Mouglin A, Branlant C. 1998. Post-transcriptional modifications in the U small nuclear RNAs. In *Modification and editing of RNA* (ed. Grosjean H, Benne R), pp. 201–227. ASM Press, Washington, D.C.
- Matera AG, Terns RM, Terns MP. 2007. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol* **8**: 209–220.
- Maxwell ES, Fournier MJ. 1995. The small nucleolar RNAs. *Annu Rev Biochem* **64**: 897–934.
- Mitrovich QM, Tuch BB, De La Vega FM, Guthrie C, Johnson AD. 2010. Evolution of yeast noncoding RNAs reveals an alternative mechanism for widespread intron loss. *Science* **330**: 838–841.
- Muzzey D, Sherlock G, Weissman JS. 2014. Extensive and coordinated control of allele-specific expression by both transcription and translation in *Candida albicans*. *Genome Res* **24**: 963–973.
- Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. 2015. Rfam 12.0:

- updates to the RNA families database. *Nucleic Acids Res* **43**: D130–D137.
- Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res* **42**: D26–D31.
- Normand C, Capeyrou R, Quevillon-Cheruel S, Mougou A, Henry Y, Caizergues-Ferrer M. 2006. Analysis of the binding of the N-terminal conserved domain of yeast Cbf5p to a box H/ACA snoRNA. *RNA* **12**: 1868–1882.
- Piekna-Przybylska D, Decatur WA, Fournier MJ. 2007. New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA* **13**: 305–312.
- Reichow SL, Hama T, Ferré-D'Amaré AR, Varani G. 2007. The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res* **35**: 1452–1464.
- Schattner P, Decatur WA, Davis CA, Ares M Jr, Fournier MJ, Lowe TM. 2004. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res* **32**: 4281–4296.
- Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. 2017. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res* **45**: D592–D596.
- Tafer H, Kehr S, Hertel J, Hofacker IL, Stadler PF. 2010. RNAsnoop: efficient target prediction for H/ACA snoRNAs. *Bioinformatics* **26**: 610–616.
- Tollervey D, Kiss T. 1997. Function and synthesis of small nucleolar RNAs. *Curr Opin Cell Biol* **9**: 337–342.
- Uliel S, Liang XH, Unger R, Michaeli S. 2004. Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. *Int J Parasitol* **34**: 445–454.
- Watkins NJ, Ségault V, Charpentier B, Nottrott S, Fabrizio P, Bachi A, Wilm M, Rosbash M, Branlant C, Lührmann R. 2000. A common core RNP structure shared between the small nucleolar box C/D RNPs and the spliceosomal U4 snRNP. *Cell* **103**: 457–466.
- Watkins NJ, Dickmanns A, Lührmann R. 2002. Conserved stem II of the box C/D motif is essential for nucleolar localization and is required, along with the 15.5K protein, for the hierarchical assembly of the box C/D snoRNP. *Mol Cell Biol* **22**: 8342–8352.
- Wu G, Xiao M, Yang C, Yu YT. 2011. U2 snRNA is inducibly pseudouridylated at novel sites by Pus7p and snR81 RNP. *EMBO J* **30**: 79–89.
- Xia L, Watkins NJ, Maxwell ES. 1997. Identification of specific nucleotide sequences and structural elements required for intronic U14 snoRNA processing. *RNA* **3**: 17–26.
- Xue S, Barna M. 2012. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat Rev Mol Cell Biol* **13**: 355–369.