# The Fungi snoRNAome

Sebastian Canzler[*,a], Peter F. Stadler[a,b,e,d,g,f,h], Jana Hertel[c]

[a] *Bioinformatics Group, Department of Computer Science, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[b] *Interdisciplinary Center for Bioinformatics, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[c] *Young Investigators Group Bioinformatics and Transcriptomics, Department Proteomics, Helmholtz Centre for Environmental Research – UFZ, Permoserstraße 15, D-04318 Leipzig, Germany*
[d] *Department of Diagnostics, Fraunhofer Institute for Cell Therapy and Immunology – IZI, Perlickstraße 1, D-04103 Leipzig, Germany*
[e] *Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany*
[f] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[g] *Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark*
[h] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

**Abstract**

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`.

*Key words:* small nucleolar RNAs, snoRNA, fungi, evolution, target switch, snoRNA target, conservation

## 1. Introduction

Small nucleolar RNAs (snoRNAs) are non-protein-coding RNAs (ncRNAs) that guide the modification of single nucleotides in other RNA molecules. Localized in the nucleolus of eukaryotic (and some archaean) cells they associate with at least four proteins in a small nucleolar ribonucleoprotein (snoRNP) complex. The target RNA molecule is hold in the correct position by base pairing to short unpaired region(s) within the snoRNA. These regions are referred to as antisense elements (ASE). At a specific position, the target is then modified. Ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs) constitute the main class of targets, although several snoRNAs have been shown to target residues in other RNA molecules [1–8]. 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 modification they guide. Box C/D snoRNAs form a stem-loop structure with a rather long loop which is stabilized by the associated proteins. They guide the 3'OH-methylation of nucleotides. Box H/ACA snoRNAs, on the other hand, are longer, fold into a thermodynamically more stable double stem-loop structure. These genes guide pseudouridylation of uracil residues in the target RNA. In addition, there are chimeric snoRNAs that share features of both classes, are much longer and/or are described to have different functions [9]. Like other small ncRNAs, snoRNAs are functional due to their secondary structure *and* characteristic (short) regions in their nucleotide sequence. In consequence, selective pressure acts on these short sequence motifs (boxes and ASEs) and the preservation of the stem-loop(s) during evolution. This led to a low conservation of the *overall* sequence, an increased number of compensatory mutations in structured regions and high sequence conservation at the protein and target RNA binding sites. The annotation and classification of snoRNAs according to sequence homology only, e.g., by using `NCBI-blast` [10] is therefore not feasible. Our recently introduced computational annotation pipeline `snoStrip` [11] is specifically designed to take care of all specific characteristics of snoRNAs. It can be used to relibaly annotate the snoRNA complement in a group of species based on an initial set of confirmed snoRNA genes of a

[*] Corresponding author
*Email addresses:*
`sebastian@bioinf.uni-leipzig.de` (Sebastian Canzler),
`studla@bioinf.uni-leipzig.de` (Peter F. Stadler),
`jana.hertel@ufz.de` (Jana Hertel)

related organism.

Here, we analysed 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. Further, we studied the evolutionary conservation of those snoRNAs and the co-evolution of snoRNAs and their targets. We provide a comprehensive set of fungal snoRNAs, their detailed description with respect to genomic location, box motifs and position, potential/confirmed target information (including observed target switches), family assignment and a suggestion of the evolutionary history of individual snoRNA families. All data can be viewed in and downloaded from our electronic supplement.

## 2. Materials and Methods

### *2.1. Genome and snoRNA Data*

Genome sequences from 147 fungal species were downloaded from Ensembl Genomes [12], JGI [13], Broad Institute (Fungal Genome Initiative) , and Candida Genome Database [14]. An `NCBI`-based taxonomic tree displaying the relationship, genome source, and version for all fungal organisms in this evolutionary survey is shown in the supplementary Figures S1.

For 63 out of the 147 species, snoRNA sequences were already retrieved in a previous `snoStrip` study [11] that started with experimentally detected snoRNAs extracted from five surveys for *Neurospora crassa* [15], *Aspergillus fumigatus* [16], *Candida albicans* [17], *Saccharomyces cerevisiae* [18], and *Schizosaccharomyces pombe* [19]. An overview of the exerimentally verified snoRNAs and the corresponding publications can be seen in supplementary Table S2. An entire mapping of the species-specific snoRNA names as they were used in the original publications and the internal `snoStrip` names is shown in supplementary Tables S3. In total, our starting set of snoRNAs comprises 3564 snoRNA sequences assigned to 123 snoRNA families in the 63 species including 231 snoRNA genes taken from the five publications.

### *2.2. Homology search*

`snoStrip`[11] - our snoRNA annotation pipeline - was applied to the set of collected snoRNAs and the 147 fungal species in an iterative manner. Starting with Pezizomycotina, followed by Saccharomycotina, and other lineages towards 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.

### *2.3. Data curation*

The `snoStrip` retrieved homologs were curated regarding the automatically identified box motifs, correct molecule lengths, and the overall fit of each snoRNA sequence in its respective family. To identify falsely annotated box motifs, the conservation of all automatically selected boxes was checked by a comparison of the start positions within the snoRNA family alignment. Motifs that have unconserved start positions were re-adjusted to fit the snoRNA family specific box pattern and and box position. what is the criteria for beeing conserved or not? Sequences that were too large or too short were automatically adjusted according to box motifs, sequence and structure conservation.

### *2.4. 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. Secondary structure prediction was done using the `RNAfold` and `RNAalifold` programs from the Vienna RNA Package[20].

### *2.5. Phylogenetic analysis*

To follow the evolution of the snoRNA families along the phylogenetic tree we used the software `ePoPE`[21]. It implements a variant of Sankoffs 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.

### *2.6. Target prediction and analysis*

Target prediction is part of the `snoStrip` pipeline. There, the computational tools `PLEXY` and `RNAsnoop` are employed to predict targets for box C/D snoRNAs and box H/ACA snoRNAs, respectively. SnoRNAs are investigated for single or double guide potential 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 show variance in their associated target are investigated manually for a potential target switch.

### 2.7. *Lineage specific conservation of target interactions*

To study the conservation of interactions, the targets for each individual snoRNA sequence are initially predicted and subsequently their conservation in other species is evaluated. To formally investigate the conservation, the Interaction Conservation Index (ICI) was developed by [22]. Briefly, the conservation of the modification and the conservation in a specific snoRNA family are calculated as follows:

$$ICI_{mod}\,(t,s) = \frac{1}{|O(s)|} * \left( \sum_{k \in O(t,s)} \frac{\varepsilon(t,s,k)}{\bar{\varepsilon}(t,k)} \right) \quad (1)$$

$$ICI_{sno}\,(t,s) = \frac{1}{|O(s)|} * \left( \sum_{k \in O(t,s)} \frac{\varepsilon(t,s,k)}{\hat{\varepsilon}(s,k)} \right) \quad (2)$$

Where $\varepsilon(t,s,k) = \min_{x \in X(t,s,k)} E_{\mathrm{mfe}}[x, y_{t,k}]$ is scored on family level searching for the best interaction, i.e. with the lowest minimum free energy, between a snoRNA $x$ of family $s$ and the target $t$ in species $k$.

Averaging over all predictions of target $t$ in species $k$

$$\bar{\varepsilon}(t,k) = \sum_{s \in S(t,k)} \varepsilon(t,s,k)/|S(t,k)| \quad (3)$$

and averaging over all targets $t$ of snoRNA $s$ in species $k$

$$\hat{\varepsilon}(s,k) = \sum_{t \in T(s,k)} \varepsilon(t,s,k)/|T(s,k)| \quad (4)$$

allows the calculation of normalized parameters. These are then summarized over all species where a prediction of $t$ is found for snoRNA family $s$ in species $k$ ($k \in O(t,s)$) and normalized over all species where snoRNA family $s$ is present ($|O(s)|$).

This approach is especially 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 harbours all organisms that share prediction of target $t$ in snoRNA family $s$. Assume the overall taxonomic tree is represented by a tree $T = (V, E)$ with root $\gamma$. The minimal subtree $U_\tau = (V_\tau, E_\tau)$ with root $\tau$ shares the node set $V_\tau = \{\, v \mid \forall (v,u)\ ,\ u \in V_\tau\ :\ LCA_T(v,u) \in V_\tau \,\}$ where $LCA_T(v,u)$ 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 $\tau$ can be calculated as follows:

$$ICI_{mod,\tau}\,(t,s) = \frac{1}{|O_\tau(s)|} * \left( \sum_{k \in O_\tau(t,s)} \frac{\varepsilon(t,s,k)}{\bar{\varepsilon}(t,k)} \right) \quad (5)$$

$$ICI_{sno,\tau}\,(t,s) = \frac{1}{|O_\tau(s)|} * \left( \sum_{k \in O_\tau(t,s)} \frac{\varepsilon(t,s,k)}{\hat{\varepsilon}(s,k)} \right) \quad (6)$$

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

In this work, snoRNA families are mostly denoted with their internal `snoStrip` name. Original names are given in parentheses in cases where previously annotated snoRNAs are present, e.g., the internal `snoStrip` name CD_22 maps the experimentally detected *S.cerevisiae* sequence snR62. A complete mapping of `snoStrip` derived snoRNA names with their species specific names is provided in the supplement. A similar notion is chosen when target positions are described. In most cases, the alignment position is given and the sequence specific position of selected organisms are provided in parentheses.

## 3. Results

In this work, the naming pattern of snoRNAs mostly follows the convention established by the original snoRNA sources. Normally, the

predominant order of naming is: *S.cerevisiae*, *N.crassa*, *A.fumigatus*, *C.albicans*, and *S.pombe*, i.e., snoRNA families are first of all adressed with their budding yeast name, in case there is no such sequence in this particular family, the *N.crassa* name will be used, etc. In either way, the internal `snoStrip`-names will be given in parentheses. A complete mapping of `snoStrip` names with their species specific names is shown in supplement S3. In case of target positions we will denote the alignment position and the sequence specific position of selected organisms are written in parentheses.

### *3.1. Expanded fungi snoRNA complement*

A starting set of 67 box C/D snoRNA and 56 box H/ACA snoRNA families were searched for additional homologs in 147 fungal organisms. The U3 snoRNA family is considered separately due to its special function and characteristics and published elsewhere [23]. All `snoStrip` retrieved snoRNA sequences were carefully cross-checked in all species to reduce the number of false negatives and exclude potentially wrong annotations. This resulted in a total amount of 5595 and 2331 box C/D snoRNA and box H/ACA snoRNA sequences, respectively. Hence, we expanded the fungi snoRNA complement by more than 120% and immensely increased its resolution.

### *3.2. Observed fungi snoRNA characteristics*

***Box motifs***. Sequence motifs were created from all `snoStrip`-annotated snoRNAs. They can be downloaded from the supplement section S5. In general these motifs follow the propagated rules for canonical snoRNA box motifs known from yeast and animals:

Box C (RUGAUGA) and D (CUGA) match the consensus sequence motifs almost perfectly. Box C shows a purine (R) in 92% of all cases. The 5' GA di-nucleotide is conserved to 100%. The 5' nucleotide (C) of box D shows few mutations (4.2%) mostly towards an adenine while the remaining positions are highly conserved (≥99.7%).

As expected from yeast and other animal snoRNAs the situation is different for the prime box motifs. In box C', merely the 5' UG di-nucleotide and, to a lesser extent, the trailing GA di-nucleotide show high conservation. This might indicate a role in the binding of snoRNP associated proteins. In box D', variations of the canonical nucleotides occur quite frequently (between 15% and 45%) in each position.

In box H/ACA snoRNAs, we observe that the sequence of box ACA is highly conserved with slight variations in its middle position. The adenine residues of box H (ANANNA) are highly conserved at the $1^{st}$ and $3^{rd}$ position while the trailing adenine ($6^{th}$ position) is more variable. The $2^{nd}$ position of this motif is a guanine in nearly 80% of the box H/ACA snoRNAs, whereas the other "N" associated positions ($4^{th}$ and $5^{th}$ position) do not show a favoured nucleotide.

***Sequence***. In accordance with the published box C/D snoRNA length, 90% of the novel `snoStrip`-annotated snoRNAs are 80-135nt in length, with a median of 93nt (see supplementary Figure S5.2). Family Nc_CD_53 (*N.crassa*, CD_53 in `snoStrip`) is the only exception since its members share sequences with lengths between 200 and 300nt. 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 Figure S5.2.

Box H/ACA snoRNAs are reasonably longer. Their median sequence length is 188nt. The shortest sequence comprises 115nt, while 90% of all sequences are between 148 and 266nt. Both single hairpin sequences share a similar length distribution. For boxplots and more details see supplement section S5.

***Structure***. Due to its specific post-transcriptional processing by exonucleases, both trailing ends of box C/D snoRNAs are cut not farther than 5 nucleotides apart from box C and D, respectively [24]. Owed to these rather short ends, only a small subset of snoRNA sequences were predicted to fold a short closing stem (1208 out of 5595). If we enlarge the trailing ends to 10nt instead, a stem could be predicted for nearly 60% (3317). These observations and the rather large fraction of snoRNAs that are still unable to fold into a characteristic stem loop structure indicate that a specific naturally occurring secondary structure is probably not needed for box C/D snoRNAs to fulfil their function. In consequence, snoRNP-associated proteins may take charge of bringing the RNA molecule and the assembled proteins into the correct functional conformation.

In contrast, box H/ACA snoRNAs are required to develop a significant and specific secondary structure to function appropriately. Only 15% (395 out of 2269) of all snoRNAs were not predicted to fold into a stem-loop structure for both hairpins.

In general, snoRNA-specific characteristics like box motifs, lengths, and secondary structures are highly comparable between Fungi and Metazoa [22].

### *3.3. Evolution of fungi snoRNAs*

***Phylogenetic distribution of snoRNAs***. A heatmap depicting the distribution of fungal box C/D snoRNA families is shown in figure 1. A similar illustration of box H/ACA snoRNA families can be seen in supplementary Figure S6. In both figures, the amount of snoRNA sequences belonging to a particular organism and snoRNA family is color encoded.

In general, fungal snoRNA families encompass exactly one snoRNA sequence per organism. Exceptions of this rule are given by snoRNA 'clans' CD_5 and CD_19 whose coverage number mainly lies between two and three. This is explained due to several target switches and major rearrangements between three different snoRNA families which promted `snoStrip` to merge the previously separate snoRNA familes. Details are explained later when target switches are discussed.

Besides an enlarged snoRNA coverage in specific families, it frequently happens that certain species encode an increased amount of paralogs to one or many snoRNA families, e.g., *Postia placenta*, *Atractiellales sp* or *Nadsonia fulvescens*. Even whole lineages share increased copy numbers in certain families, e.g., Leotiomycetes in AM921940 (CD_41) or Sordariomycetes in Nc_CD_28 (CD_28).

Almost half of all 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. Additionally, several families are found to be lineage-specific, e.g., seven in Saccharomycotina (see box 'A' in Figure 1), nine in Peiziomycotina (box 'B'), and six in Sordariomycetes (box 'C'). These lineages map exactly to the clades where original snoRNA data originated from.

In contrast to lineage-specific families, lineage-specific losses of snoRNAs are also detectable. Basidiomycota, for example, are not found to contain orthologs of families snR48 (CD_8), snR190 (CD_16), or U14 (CD_37), while in Saccharomycotina, no trace is found of snoRNAs of family AM921940 (CD_41). 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 remaining species do not contain this particular family. In these lineages, only the analysis of target interaction might answer the question whether this single snoRNA is a true member of the family or merely an artifact.

Compared to 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 is detected in Microsporidia leaving this clade completely without any annotated snoRNA candidate. It is apparent, however, that box H/ACA snoRNAs shows substantially more lineage specific innovation and deletion events than observed in box C/D snoRNAs, see supplementary Figure S6.

In total, 22 out of 50 H/ACA families are merely found in a small subset of species. Moreover, several families are found in two or more lineages but seem to be completely lost in others, such as snR42 (HACA_33), AJ632014 (HACA_56), and snR33 (HACA_24). They are present in Taphrinomycotina and Saccharomycotina but cannot be found in Pezizomycotina.

Another noticeable observation is that not a single box H/ACA snoRNA is found in *Pyrenophora tritici-repentis* (marked with an asterisk in the supplement figure). This stands in sharp contrast to C/D snoRNA sequence, where *P.tritici-repentis* orthologs are found in nearly all families that are present in the *P.tritici-repentis*-containing Dothideomycetes lineage.

***Evolutionary event 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 gain and loss event of individual paralogs of each snoRNA family and summarized this information for all analysed 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 supplementary Figure S7.1 for a version with absolute values. We observe a large amount of snoRNA families that emerged at each major branch point along the backbone of the taxonomic tree. A total of 34 box C/D snoRNA families could be traced to the root of fungi, indicating an even more ancient origin. At the root of Dikarya, As-

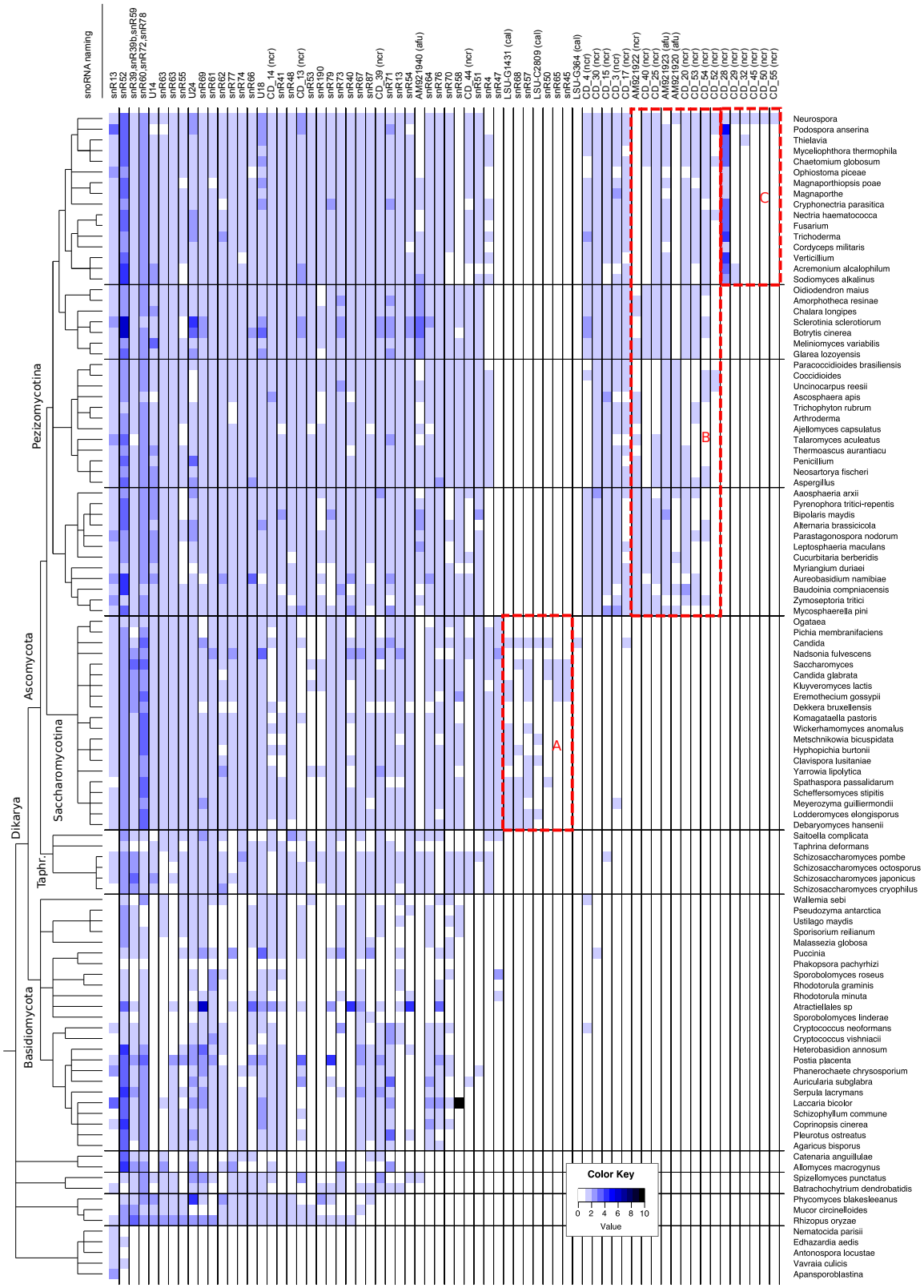


Figure 1: A heatmap of `snoStrip`-detected box C/D snoRNAs is shown on the previous site. Each column represents a specific snoRNA family, while each row either represents a certain species or genus. A taxonomic classification is shown on the left hand side. The amount of snoRNAs detected in a specific species and snoRNA family is encoded in a blue color scheme. Lineage specific families are boxed (A: Saccharomycotina, B: Pezizomycotina, C: Sordariomycetes).

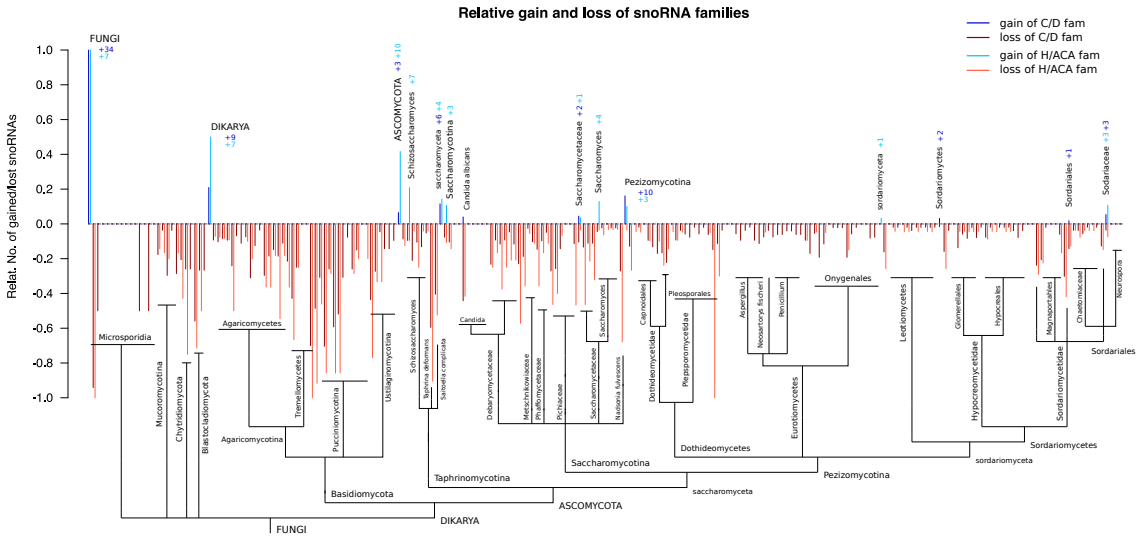


Figure 2: Relative number of gains and losses of entire snoRNA families during fungal evolution. The relative gain is the number of gained snoRNA families compared to the observed number of snoRNA families. The relative loss describes the number of lost snoRNA families compared to the number of snoRNA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

comycota, Saccharomyceta, and Pezizomycotina, a total of 9, 3, 6, and 10 families arose, respectively. A similar picture is drawn in case of box H/ACA snoRNAs where 7 families could be traced to the root of fungi and additional 7, 10, 4, and 3 families are gained at the root of Dikarya, Ascomycota, Saccharomyceta, and Pezizomycotina. 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* formerly mentioned the remarkable absence of snoRNA genes in this clade, although all components of the snoRNA machinery are clearly present [25]. 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.

When focusing on species level, it is frequently observed that single organisms seem to have lost a large amount of their snoRNAs, i.e. in the Basidiomycota lineage. Especially *W.sebi* and several Pucciniomycota seem to have lost nearly their entire set of box H/ACA snoRNAs (*W.sebi*: 92%, *R.minuta*: 86%, or *S.linderae*: 86%). The impact on box C/D snoRNAs is more moderate (0.26 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 *P.anserina* and *O.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 *et. al* identified four novel snoRNA candidates among their set fo 40 snoRNA genes that showed no high sequence similiarity towards already annotated budding yeast sequences [17]. 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 [17] (`snoStrip`: CD_69 and CD_71) are exclusively present in Saccharomycotina except for Saccharomycetaceae. They are also found to share an extraordinary 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: *C.dubliniensis* and *C.tropicalis*.

***Fission Yeast Specific snoRNAs***. Similar to *C.albicans*, several snoRNAs published in the fission yeast [19] are found to be lineage or even

species specific. In the original publication, 12 sequences have not been mapped to budding yeast snoRNAs and 7 of them have no predicted target interaction. By means of `snoStrip`, AJ632008 in [19] (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 from a HP1 target in *S.pombe* to a HP2 target in *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)).

### *3.4. Conservation of Target Interaction*

In accordance to 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 anti sense 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 of different targetRNAs. But target predictions that are based on single sequence predictions are not overly convincing in a biological point of view.

Within the 68 box C/D snoRNA families, the large majority (40) is found to be *true* single guides. 28 families share a functional D' target and the remaining 12 families a conserved D box associated binding site. An additional amount of 14 box C/D snoRNA families are *predominantly* found to be single guides, i.e., these families share exactly one highly conserved target binding region (three families share a conserved D target while 11 families share a functional D' target), whereas the other target region is only found to be functional in subset of organisms. Eight families harbor two functional target binding regions that are conserved throughout all lineages where these families are detected. Six families are originally denoted as orphan snoRNA meaning that no potential interaction has been published thus far. In case of box H/ACA snoRNAs, 23 families are *true* single guides: 8 families share a conserved pseudouridylation pocket in hairpin 1 and 15 families share a HP2 target. Further 6 families comprise a lineage specific HP2 target besides their overly conserved target in hairpin1. The opposite situation can be seen in 3 box H/ACA snoRNA families. 11 families are found to be double guides, while 7 families are 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 supplement (section S9-S18).

Solely 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 'snoRNA clans' CD_5 and CD_19, none of the remaining six families is traceable amongst all major fungal lineages. Two families, Nc_CD_17 (CD_17) and AM921920 (CD_35), are found in Pezizomycotina while 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 frequent. 11 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 organism. 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 (HACA_26), snR83 (HACA_30); one found in Schizosaccharomyces, AJ632008 (HACA_46)).

Family snR3 is published 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

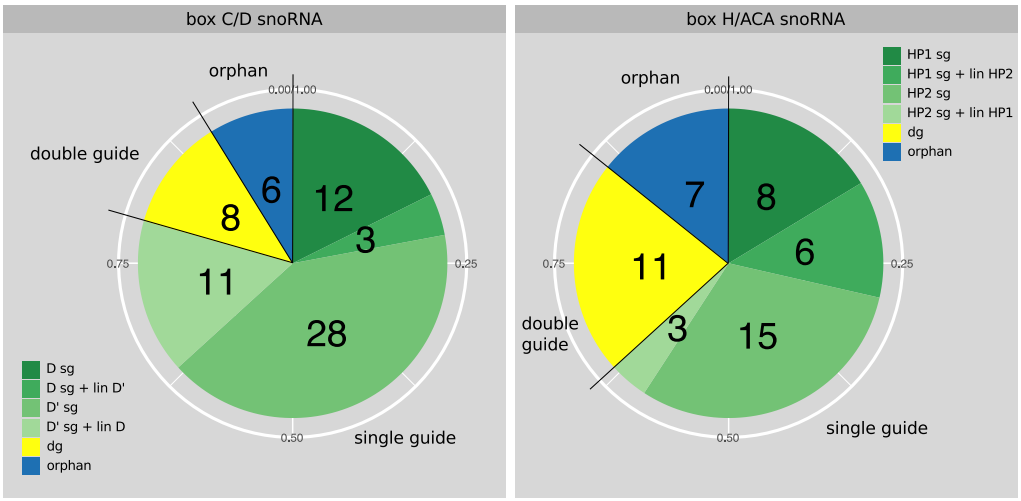


Figure 3: Pie chart of both major snoRNA classes. A snoRNA family is classified based on its conserved target prediction either as single guide (sg), single guide with a lineage specific target in its non-conserved target region (lin), double guide (dg), or orphan.

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, 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 anti sense elements. In the originally published snoRNA datasets of five different fungi, orphan box C/D snoRNAs were annotated for *S.cerevisiae* (2 sequences), *N.crassa* (2), and *A.fumigatus* (9). In addition to these sequences, 11 *N.crassa* snoRNAs have been 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 supplementary Table S13. Highly conserved target interaction that are predicted by `snoStrip` are shown in Table 1.

Table 1: Assigning putative targets to previously orphan box C/D snoRNAs. Families that do not contain sequences with experimentally verified targets are marked with '*'.

| 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 |

Unfortunately, potential targets for both orphan *N.crassa* snoRNAs are not unambiguously discovered 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 case of both budding yeast snoRNAs (snR4, snR45), no potential target is found across canon-

ical target sequences, although family snR4 is found to be present in several fungal lineages such as Taphrinomycotina, Saccharomycotina, and several Pezizomycotina species. Family snR45, on the other side, is exclusively found in Saccharomycetaceae.

The picture looks much better in 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 with '*' 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 predicted targets but without homologous relations to other known snoRNAs, comprised 16 distinct targets published in the original publication [19]. Ten of these targets were confirmed through a conserved prediction using `snoStrip`. Three targets were annotated as tRNA modification sites and hence, 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) [15]. By means of `snoStrip`, family Nc_CD_54 is detected amongst all Pezizomycotina lineages and a highly conserved target region is clearly visible upstream of box D', originally denoted as orphan. This region shows convincing base pairing capabilities to U6-70 (*N.crassa* U6-55) in virtually all identified 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 D target, on the other hand, is not found to be conserved outside of Neurospora.

Within the initial box H/ACA snoRNA datasets, orphan sequences were published for *N.crassa* (6 sequences), *A.fumigatus* (1), and *S.pombe* (8). A detailed summary of these sequence can be seen in supplementary Table S18.

Table 2: Assigning putative targets to previously orphan box H/ACA snoRNAs. Families that do contain sequences with experimentally verified targets are marked with '*'.

| 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* |

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 large differences in sequence lengths (HACA_36 is approx. 1kb long ; Nc_HACA_7 is $\sim$ 180nt 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 competently 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, explaining the missing homologs in this clade. 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 known to be modified in yeast or human by now.

Seven of 15 orphan box H/ACA snoRNAs are found to be conserved solely on genus or species level, i.e., 2 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 se-

quence makes an appropriate target prediction impossible. Hence, a sufficient conclusion about their true function and, further on, about their genuine existence in terms of a viable snoRNA molecule as well as its biological necessity remains elusive.

***Lineage-specific Targets***. Quite a few box C/D snoRNA families harbor a highly conserved target either at their D or D' position. However, in a large amount of cases, it might be that these families exhibit additional lineage specific target binding capabilities on their 'non-functional' ASE. Such a functionality might have evolved at a specific time point during evolution, and because of a potential benefit, is retained in all of today's organisms descending from this ancestor.

Interesting box C/D snoRNA families with a previously annotated functional D' targets and lineage specific D targets can be seen in Figure 4. Detailed information about all snoRNA families with an additional, lineage specific target are found in supplement Table S11.

Family snR87 (CD_10), for example, with its experimentally verified target 18S-479 (18S-436; D' target), 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 amongst 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. 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* [15] 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 [26, 27]. 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 marginal 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), it is thoroughly possible that this snoRNA is capable of modifying both targets. However, 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*, 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.

A highly interesting modification site is 25S-3941 (*S.cerevisiae* 25S-2724) whose actual methylation and the guidance by snR67 (CD_26) was experimentally shown [28]. 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 except for 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). 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). In a third family, snR54 (CD_49), the modification at 25S-3941 is predicted in Sordariales (ICI: 1.38, mean mfe: -14.14, D target).

In similarity to the box C/D snoRNA class, several box H/ACA snoRNAs comprise 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 supplement, see Table S16. Some of these functions might already been annotated, especially in snoRNA sequences of the budding yeast, see families snR189 (HACA_4) and snR191 (HACA_42) which are in fact officially denoted as double guides in *S.cerevisiae*. 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 con-

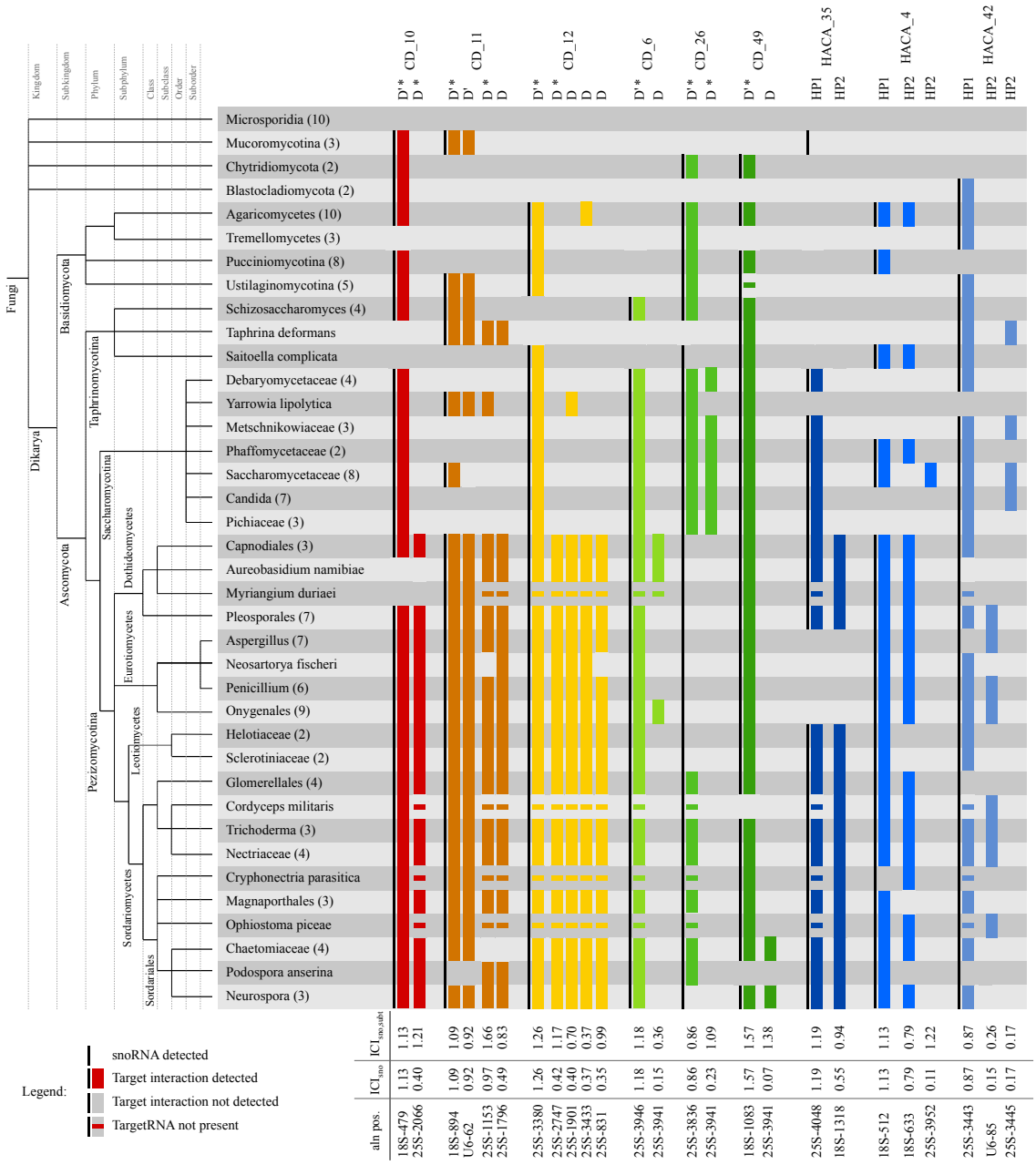


Figure 4: The conservation of predicted target interactions is shown for interesting single guide box C/D snoRNA families that exhibit an additional functional target at their 'non-functional' D box. Each family is depicted in a different color. The black bar in front of each family shows the presence 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 '*'.

spicuous. 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 amount of Pezizomycotina species ($ICI_{sub}$ = 0.73). This particular modification is not present in budding yeast 5.8S molecules which undoubtedly explains the missing predic-

Table 3: Interaction properties of four LSU modifications of CD_5 are shown. Properties for three SSU and two LSU methylations are given for clan CD_19.

| | modification | $ICI_{sno}$ | ∅ 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% |

tions 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 biological 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 present 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 sequence, upstream of box D/D or in HP1/HP2, is mostly preserved but it happens seldomly that this position is also shifted. Two highly complex rearrangements have been autmatically detected by `snoStrip`. Each of these two 'snoRNA clans' comprise 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 exlained and illustrated in the supplement section S8.

**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). The methylations at position 25S-1806, 25S-1898, and 25S-3915 map to known and verified modifications in human large subunit ribosomal RNAs and hence, are supposed to be ancient, which in consequence suggest the real existence of both the methylations and the guiding snoRNAs at the root of fungi. However, through individual target switches in the cause of fungal evolution, the history of these sequences became connected. A taxonomic tree displaying a potential evolutionary history involving snoRNAs that are predicted to guide the above mentioned modifications is shown in Figure 5. Therein, the putative ancient state is described to be constituted of two individual snoRNA sequences guiding the three ancient methylations. Parsimonious deletion and innovation events of target interactions 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.

Besides such ordinary processes of gain and loss events, it happened several times during fungal evolution that target interactions of the mentioned four modifications switched between different snoRNAs. It is a noteworthy fact that the actual target site within the snoRNA (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) are incorporated into one snoRNA sequence after the original guidance of 25S-1898 (D target) went missing.

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

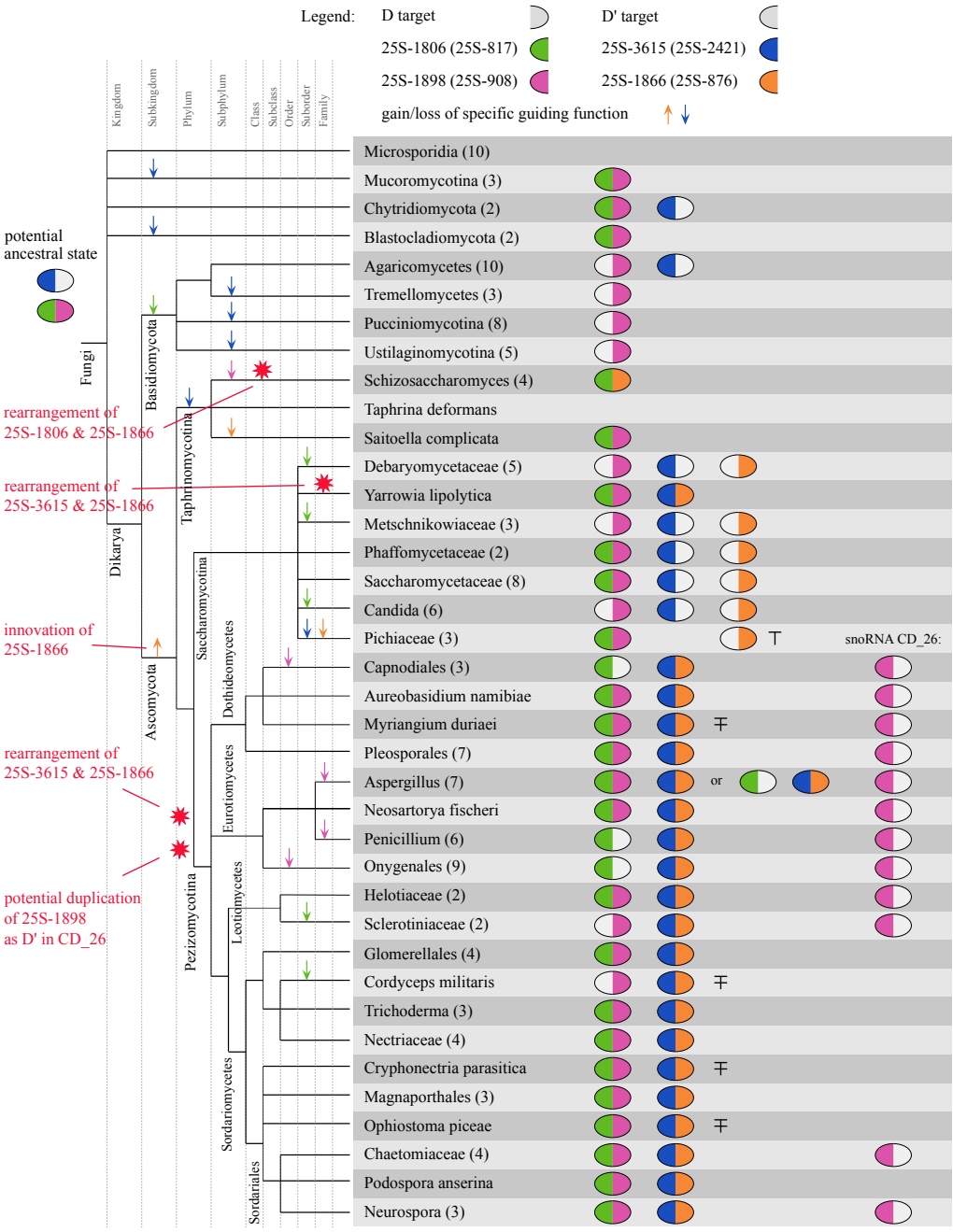


Figure 5: Potential evolutionary history of snoRNA clan CD_5 involving four modification sites on the LSU rRNA. Gain/loss events are displayed with arrows, while potential rearrangements are shown with red stars. ⊤ 25S-1866 is solely found in Pichia. ∓ Putative since LSU sequences are missing; snoRNAs show convincing ASE conservation.

all Ascomycota.

It might have happened that a snoRNA of clan CD_5 guiding methylation at 25S-3615 is already present at the 5' end of this cluster when it emerged. However, there are several possibilities how the snoRNA cluster evolved after the innovation of guiding function for 25S-1866. One hypothesis (Blue stars in Figure 6) is the initial incorporation of 25S-1866 into the snoRNA that already guides 25S-3615, creating a double guide snoRNAat 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 the 25S-1806

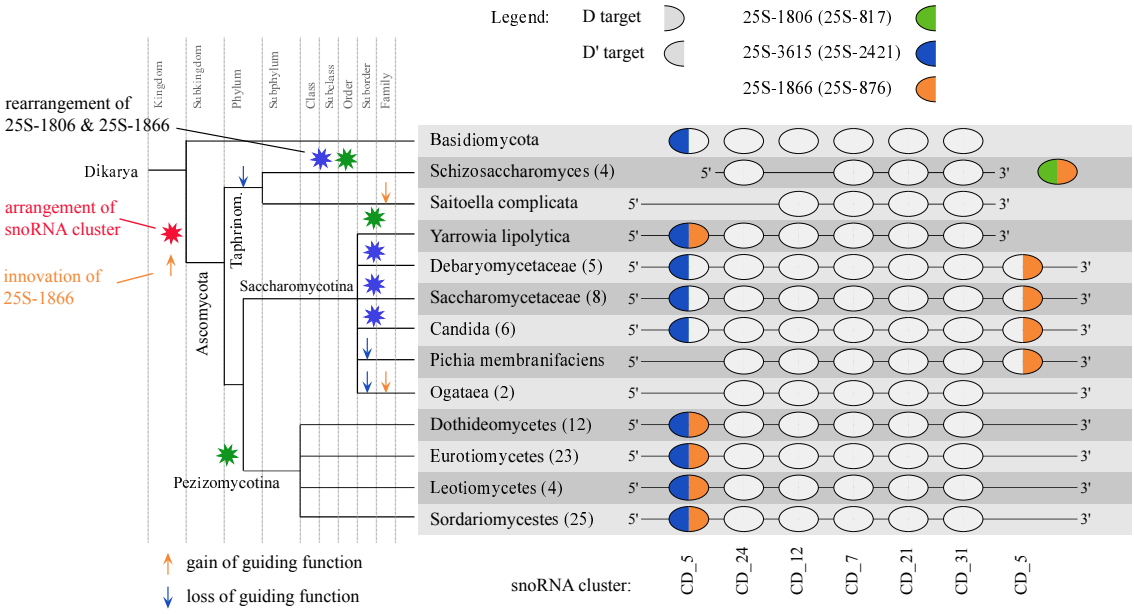


Figure 6: Sequences of the CD_5 snoRNA family are incorporate into a polycistronic transcript that harbors up to seven snoRNA genes. This cluster with its highly conserved structure and size occurred at the root of Ascomycota, but most of its genes arose at least at the root of Dikarya. There are different 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 hypothesis A and B, respectively.

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). The original formation is solely conserved in *Yarrowia lipolytica*. In another hypothesis, evolution might have taken the other way round (green stars in Figure 6). Assuming the innovation of 25S-1866 led to a novel single guide snoRNA that is 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. In result, 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 happened at the root of Pezizomycotina, where the first snoRNA of the cluster is found to guide modifications at position 25S-3615 (D') and 25S-1866 (D). Proteins that are located up- and downstream of 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). Neurospora species are also predicted to guide this methylation with its Nc_CD_26 (CD_26) snoRNA[15]. In reverse the original D' target of snR67, 25S-3836, was abolished in these organisms and is not found to be restored in any other snoRNA family. Please also confer Figure 4 for more detailed information of family CD_26. 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 organisms, or 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). Think about this: Interestingly, in lineages where the original target 25S-3836 of CD_26 is lost, there is no functional connection to the orginal snoRNA family anymore. This means organisms of such lineages

Table 4: Summary on multiple target predictions of families snR40 (CD_43) and snR70 (CD_61) that are guided with the same ASE.

| | pos | $ICI_{sno}$ | ∅ 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 |

have no functional target sitem either D or D' in common with organisms of other lineages in this particular family. They show clear sequence homology, meaning that they probably descend from one another. evolution of a new snoRNA family due to target shift??

***Multiple Target Interactions***. It might happen, that 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 given by box C/D snoRNA family snR40 (CD_43) which is predicted and experimentally validated to guide methylation at position 18S-1400 (18S-1271) with its D' target binding region. 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. 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 a vast majority of cases even with a better binding energy. But since the genuine modification is neither reported in *S.cerevisiae*, *N.crassa*, or human, this prediction, albeit its overly convincing nature, remains hypothetical.

An even more vital example is provided by family snR70 (CD_61) at its D' ASE. Not less than five potential targets are predicted with an ICI score above 1.0, a mean mfe below -11.30 and more than 80 single sequence predictions. Details are shown in Table 4. This time, the most persuasive prediction is experimentally confirmed, whereas the other predicted positions were not shown to be modified yet.

## 4. Discussion

Within this study, the `snoStrip` pipeline was applied to a small set of experimentally verified snoRNAs with the aim to merge non-identified homologous families and uncover the snoRNAome in a wide range of fungal species. The detected snoRNA genes and families helped to trace evolutionary events such as innovations and losses and the functional analysis of potential target interactions added a new layer of information. Based on the functional characteristics of the snoRNAs and the Interaction Conservation Index (ICI), the coevolution of snoRNAs and their targets can be measured. This measure combines the evolutionary conservation of the precise RNA-RNA interaction with its thermodynamic stability and hence serves as an extraordinary marker for highly conserved modification sites and interactions.

The starting point of this study includes five different sets of mostly experimentally verified snoRNAs. These were subsequently merged and used for querying 147 fungal organisms. By means of `snoStrip`, a total set of over 5500 box C/D snoRNAs (68 families) and 2200 box H/ACA snoRNAs (50 families) was assembled. The automated annotation of snoRNAs and their characteristics and the highly efficient target prediction in combination with the ICI scores were key-factors to sort and rearrange the landscape of fungal snoRNAs.

Similar to Metazoa, it is apparent that fungal box H/ACA snoRNAs show a higher loss-ratio compared box C/D snoRNAs. This might have 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 timescales, both on sequence level and a functional point of view. But due to its short interacting regions, these molecules are more vulnerable for target site disrupting mutations and, in consequence, for a presumable loss of functionality which might in fact lead to a higher rate of losses.

In general, fungal snoRNAs are found to stably preserve their 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 or even species specific targets. Due to the novel ICI score and its adaptation to work on subtrees, this scenario is evidently measurable from a com-

putational point of view. To what extent this still holds *in vivo* remains unclear, since target predictions and the measurement of conservation of certain interactions in a small set of organisms is only of limited value and highly restricted without experimental evidence.

The aspect of additional target interactions that are predicted at the highly conserved ASE of a snoRNA family is still mainly unexplored, but the possibility that a single snoRNA target site comprises two distinct guiding functions has at least been reported for budding yeast box H/ACA snoRNAs. Distinct in that sense means target sites that are not directly adjacent. The budding yeast snoRNA family snR3 (HACA_3), for example, is verified to target two modification sites in its second hairpin. Both interactions are furthermore traceable across Dikarya. Despite this special case where both targets are experimentally validated, most detected 'double' target sites require experimental verification. In some cases of box H/ACA snoRNAs, these additional targets gain better ICI scores than the annotated modification site. Such highly convincing predictions might not be regarded as junk although they lack experimental evidence on both the interaction level and the validation of the genuine modification itself. Based on the specialized ribosome hypothesis, the possibility of distinct ribosomal conformations in different developmental stages and stress levels might also affect the modification level of ribosomal RNAs and hence, might lead to still hidden modifications and interactions [29]. Convincing examples of remarkably conserved multiple interactions are given by box C/D snoRNA families snR40 (CD_43) and snR70 (CD_61) that 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 anti sense element. This might be dependent on developmental phases, or more complex mechanisms that might be triggered by probability rates with respect to the actual binding energy. In a potential scenario, interactions with extraordinary low binding energies are preferentially executed while additional guiding functions might be performed less often or even on demand.

On the other hand, we also find 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; and second, this includes single interactions that are split up over different snoRNA families depending on the taxonomic lineage. A perfect example of the latter situation is given by the predicted pseudouridine at position 5.8S-18. This particular position is not known to be modified yet, but several highly convincing predictions in distinct families have been made by `RNAsnoop` (see supplement material). 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. When thinking about tissue specificity or developmental stage specificity of snoRNA families, it might happen that certain families are underexpressed or even completely silenced under particular conditions which might lead to an insufficient rate of pseudouridines or methylations. Therefore, the necessity of a precise modification might have let to a shift or duplication of the target binding capability to another snoRNA family.

In general, one can say that the snoRNA landscape is permanently changing, i.e., whole snoRNA sequences vanish and novel genes are introduced, guiding functions may be shifted from one snoRNA to another, they may be duplicated, or they get lost. That means, the creation, change, and loss of snoRNA genes is an on-going process, that also leads to a large number of lineage or even species specific snoRNAs, detectable target switches, and the loss of single families or even large fractions of the whole snoRNAome. Additionally, the amount of present snoRNA families is found to be considerably higher in Metazoa, for example, in human, than for lower eukaryotes such as yeasts. This is a direct consequence of the observation that higher eukaryotes contain more modifications in their rRNAs and snRNAs than Bacteria or lower eukaryotes.

Besides that, several aspects about the snoRNAome in Metazoa and Fungi are similar. A common feature is the detectable burst in the snoRNA diversity at each major branching point in the taxonomic tree of both kingdoms. In case of box C/D snoRNAs, the distribution of orphan, single guided, and double guided snoRNAs is quite similar compared between fungi and the human snoRNA atlas [30]. Therein, over 70% of box C/D carrying snoRNAs are found to be single guided (75% in Fungi), while the other fraction is to one part double guided and to the other part orphan (same in Fungi). In box H/ACA snoRNAs, the situation looks a little bit different, since human double guided snoRNAs comprise the largest

group (47%). In Fungi, solely 22% of box H/ACA snoRNA families is found to guide two distinct pseudouridines with both hairpins.

## Acknowledgments

This work was funded in part by the European Union FP-7 project QUANTOMICS (no. 222664). MMML-seq project of the International Cancer Genome Consortium (ICGC) funded by German Federal Ministry of Education and Research. LIFE – Leipzig Research Center for Civilization Diseases is funded by the State of Saxony and the European Union.

## References

[1] Clouet d'Orval B, Bortolin ML, Gaspin C, Bachellerie JP. Box c/d rna guides for the ribose methylation of archaeal trnas. the trnatrp intron guides the formation of two ribose-methylated nucleosides in the mature trnatrp. Nucleic Acids Res 29, 2001:4518–29.

[2] Dennis PP, Omer A, Lowe T. A guided tour: small rna function in archaea. Mol Microbiol 40, 2001:509–19.

[3] Uliel S, Liang XH, Unger R, Michaeli S. Small nucleolar rnas that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. Int J Parasitol 34, 2004:445–54. doi:10.1016/j.ijpara.2003.10.014.

[4] Cavaill J, et al. Identification of brain-specific and imprinted small nucleolar rna genes exhibiting an unusual genomic organization. Proc Natl Acad Sci U S A 97, 2000:14311–6. doi:10.1073/pnas.250426397.

[5] Kishore S, Stamm S. Regulation of alternative splicing by snornas. Cold Spring Harb Symp Quant Biol 71, 2006:329–34. doi:10.1101/sqb.2006.71.024.

[6] Kiss T. Small nucleolar rnas: an abundant group of noncoding rnas with diverse cellular functions. Cell 109, 2002:145–8.

[7] Matera AG, Terns RM, Terns MP. Non-coding rnas: lessons from the small nuclear and small nucleolar rnas. Nat Rev Mol Cell Biol 8, 2007:209–20. doi:10.1038/nrm2124.

[8] Bratkovi? T, Rogelj B. Biology and applications of small nucleolar rnas. Cell Mol Life Sci 68, 2011:3843–51. doi:10.1007/s00018-011-0762-y.

[9] Darzacq X, Jdy BE, Verheggen C, Kiss AM, Bertrand E, Kiss T. Cajal body-specific small nuclear rnas: a novel class of 2'-o-methylation and pseudouridylation guide rnas. EMBO J 21, 2002:2746–56. doi:10.1093/emboj/21.11.2746.

[10] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 215, 1990:403–10. doi:10.1016/S0022-2836(05)80360-2.

[11] Bartschat S, Kehr S, Tafer H, Stadler PF, Hertel J. `snoStrip`: a snoRNA annotation pipeline. Bioinformatics 30, 2014:115–6. doi:10.1093/bioinformatics/btt604.

[12] Kersey PJ, et al. Ensembl genomes 2016: more genomes, more complexity. Nucleic Acids Res 44, 2016:D574–80. doi:10.1093/nar/gkv1209.

[13] Nordberg H, et al. The genome portal of the department of energy joint genome institute: 2014 updates. Nucleic Acids Res 42, 2014:D26–31. doi:10.1093/nar/gkt1069.

[14] Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The candida genome database (cgd): incorporation of assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res 45, 2017:D592–D596. doi:10.1093/nar/gkw924.

[15] Liu N, et al. SnoRNAs from the filamentous fungus *Neurospora crassa*: structural, functional and evolutionary insights. BMC Genomics 10, 2009:515. doi:10.1186/1471-2164-10-515.

[16] Jöchl C, et al. Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein synthesis. Nucleic Acids Res 36, 2008:2677–89. doi:10.1093/nar/gkn123.

[17] Mitrovich QM, Tuch BB, De La Vega FM, Guthrie C, Johnson AD. Evolution of yeast noncoding RNAs reveals an alternative mechanism for widespread intron loss. Science 330, 2010:838–41. doi:10.1126/science.1194554.

[18] Piekna-Przybylska D, Decatur WA, Fournier MJ. New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. RNA 13, 2007:305–12. doi:10.1261/rna.373107.

[19] Li SG, Zhou H, Luo YP, Zhang P, Qu LH. Identification and functional analysis of 20 Box H/ACA small nucleolar RNAs (snoRNAs) from *Schizosaccharomyces pombe*. J Biol Chem 280, 2005:16446–55. doi:10.1074/jbc.M500326200.

[20] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast Folding and Comparison of RNA Secondary Structures (The Vienna RNA Package). Monatshefte f Chemie 2, 1994:167–188. doi:10.1007/BF00818163.

[21] Hertel J, Stadler PF. The Expansion of Animal MicroRNA Families Revisited. Life (Basel) 5, 2015:905–20. doi:10.3390/life5010905.

[22] Kehr S, Bartschat S, Tafer H, Stadler PF, Hertel J. Matching of Soulmates: coevolution of snoRNAs and their targets. Mol Biol Evol 31, 2014:455–67. doi:10.1093/molbev/mst209.

[23] Canzler S, Stadler PF, Hertel J. Evolution of Fungal U3 snoRNAs: Structural Variation and Introns. Non-Coding RNA 3, 2017. ISSN 2311-553X. doi:10.3390/ncrna3010003.

[24] Kishore S, Gruber AR, Jedlinski DJ, Syed AP, Jorjani H, Zavolan M. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. Genome Biol 14, 2013:R45. doi:10.1186/gb-2013-14-5-r45.

[25] Gardner PP, Bateman A, Poole AM. SnoPatrol: how many snoRNA genes are there? J Biol 9, 2010:4. doi:10.1186/jbiol211.

[26] Machnicka MA, et al. `MODOMICS`: a database of RNA modification pathways–2013 update. Nucleic Acids Res 41, 2013:D262–7. doi:10.1093/nar/gks1007.

[27] Massenet S, Mougin A, Branlant C. Post-transcriptional Modifications in the U Small Nuclear RNAs. In H Grosjean, R Benne, editors, *Modification and Editing of RNA*. ASM Press, 1998.

[28] Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. Science 283, 1999:1168–71.

[29] Xue S, Barna M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. Nat Rev Mol Cell Biol 13, 2012:355–69. doi:10.1038/nrm3359.

[30] Jorjani H, et al. An updated human snoRNAome. Nucleic Acids Res 44, 2016:5068–82. doi:10.1093/nar/gkw386.