The Fungal 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

Small nucleolar RNAs (snoRNAs) are essential players in the rRNA biogenesis through their involvement in the nucleolytic processing of the precursor and the subsequent guidance of nucleoside modifications. Within the kingdom of fungi, several species-specific surveys explore 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. Here, we report a comprehensive inventory of fungal snoRNAs with 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, shows apparently that the shape of the landscape is subject to consistent re-arrangements and adaptations, e.g., through lineage-specific targets and redundant guiding functions.

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 chemical 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 to form the small nucleolar ribonucleoprotein (snoRNP) complex [1]. 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) [2–4], although some snoRNAs have been shown to target residues in other RNA molecules such as transfer RNAs [5, 6], spliced leader RNAs [7], or brain-specific messenger RNAs [8, 9]. Furthermore, snoRNAs are known to be involved in the nucleolytic processing of rRNA precursors, the synthesis of telomeric DNA, genomic imprinting, and alternative splicing [10–13].

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 that they guide [11, 14]. 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

[*] 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)

snoRNAs are longer, fold into a thermodynamically more stable double stem-loop structure, and guide pseudouridylation of uracil residues in the target RNA. In addition to these two classes, there are chimeric snoRNAs that share features of both classes, are much longer and/or are described to have different functions [3]. Similar to other small ncRNAs, snoRNAs require both specific secondary structures and characteristic sequence motifs to perform their function. These features are therefore preserved by evolution and are clearly recognizable by comparative methods [11, 15]. While sequence motifs involved in protein binding are common to all members of each of the the two classes, the ASEs are conserved only among members of snoRNA families with the same target. Overall, therefore, snoRNA sequences evolve rapidly, making them hard to identify by purely sequence-base methods such as `blast` [16].

To overcome this limitation, we introduced a computational annotation pipeline `snoStrip` [17] that is specifically designed to track all specific characteristics of snoRNAs. Here we use this 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. 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. Manually curated snoRNA family alignments will be submitted to `Rfam` [18].

2. Materials and Methods

2.1. Genome and snoRNA Data

Genome sequences from 147 fungal species were downloaded from Ensembl Genomes [19], JGI [20], Broad Institute (Fungal Genome Initiative) , and Candida Genome Database [21]. 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, most snoRNA sequences were already retrieved in a previous study, mostly as test of `snoStrip` [17]. In this earlier work, we started with experimentally detected snoRNAs extracted from five surveys for *Neurospora crassa* [22], *Aspergillus fumigatus* [23], *Candida albicans* [24], *Saccharomyces cerevisiae* [25], and *Schizosaccharomyces pombe* [26]. An overview of the experimentally verified snoRNAs and the corresponding publications is compiled in Supplementary Table S2. The nomenclature of snoRNAs is consistent across different species. A dictionary relating the species-specific, traditional snoRNA names as they were used in the original publications and the internal snoRNA family designations used by `snoStrip` is included as Supplementary Table S3. Here we use the results of [17] 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.

2.2. Homology search

The `snoStrip` pipeline [17] 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 candidate snoRNAs identified by `snoStrip` were curated regarding the automatically identified box motifs, plausible lengths, and the overall fit of each snoRNA sequence in its respective family. To identify incorrectly annotated box motifs, the conservation of all predicted boxes was checked by a comparison of the start positions within the snoRNA family alignment. Motifs that start at unconserved positions are most probably false annotations and were re-adjusted to fit the snoRNA family specific box pattern and and box position. Sequences, where re-adjusted C- or D-boxes did not agree with canonical box motif pattern, were removed from further analysis.

Too large or too short candidate sequences were mainly caused by 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. Subsequently, sequences with re-adjusted box motifs were automatically trimmed or enlarged.

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[27].

2.5. Phylogenetic analysis

To follow the evolution of the snoRNA families along the phylogenetic tree we used the software `ePoPE`[28]. 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 [29, 30]. 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 [31]. In brief, the conservation of the modification and the conservation in a specific snoRNA family are calculated as follows:

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

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

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

are obtained by averaging over all all predictions of target t in species k or all targets t of snoRNA s in species k, respectively. There normalized parameters are then summed over all species $k \in O(t,s)$ in which a prediction of target t is found for snoRNA family s 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 represented 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 $V_\tau = \{\, v \mid \forall(v,u)\,,\ u \in V_\tau \ :\ LCA_T(v,u) \in V_\tau \,\}$ 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 τ can be calculated as follows:

$$ICI_{mod,\tau}(t,s) = \frac{1}{|O_\tau(s)|} * \left(\sum_{k \in O_\tau(t,s)} \frac{\varepsilon(t,s,k)}{\bar{\varepsilon}(t,k)} \right)$$
$$ICI_{sno,\tau}(t,s) = \frac{1}{|O_\tau(s)|} * \left(\sum_{k \in O_\tau(t,s)} \frac{\varepsilon(t,s,k)}{\hat{\varepsilon}(s,k)} \right) \quad (3)$$

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 τ and share at least one snoRNA of family s. v_k is the leaf that denotes organism k.

3. Results

There is at present no generally accepted nomenclature of snoRNA families across different fungal species. 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 *S. cerevisiae*, *N. crassa*, *A. fumigatus*, *C. albicans*, and *S. 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 Supplementary 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 Supplementary Table S4.

3.1. Expanded fungi snoRNA complement

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 considered separately due to its special function and characteristics and published elsewhere [32]. 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 5595 box C/D snoRNA and 2331 box H/ACA snoRNA sequences, expanding the collection of annotated fungal snoRNAs by more than 120%. The data substantially increase both the phylogenetic scope and the resolution of the snoRNA annotation.

3.2. Characteristics of fungal snoRNAs

Box motifs. Sequence motifs were extracted from all `snoStrip`-annotated snoRNAs. The complete collection is available for download from Supplement section S5. In general, these motifs are consistent with the published rules [33–36] 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. The 5' nucleotide (`C`) of box D is substituted in 4.2% of the cases, usually by `A`. The remaining positions are nearly perfectly conserved (≥99.7%). As expected from yeast and other animal snoRNAs the situation is different for the prime box motifs [35, 37]. In box C', merely the first `UG` dinucleotide and, to a lesser extent, the trailing `GA` dinucleotide 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 (between 15% and 45%) in each position.

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 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 4^{th} and 5^{th} `N` position do not show a significantly over-represented nucleotide. Again, these results are in accordance with previously published motif constraints [38].

Sequence length. Consistent 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 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 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 S6.2.

Box H/ACA snoRNAs are usually longer than box C/D snoRNAs. 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 S6.

Secondary 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 away from the C and D boxes, respectively [39]. Because of 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). That fact that more than 40% of the box C/D snoRNAs does not form a terminal stem even if precursor nucleotides are included strongly suggests that the terminal helix is not required for their function and snoRNP-associated proteins may be in 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 box H/ACA 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 [31].

3.3. Phylogenetic distribution of fungi snoRNAs

Phylogenetic distribution of snoRNAs. A heatmap depicting the distribution of fungal box C/D snoRNA families is shown in figure 1. Higher resolution heatmaps of both C/D and H/ACA snoRNA families are included in supplement section S7.

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, which typically have two or three members per species. This is explained due to several target switches and major rearrangements between three different snoRNA families which prompted `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. In addition, several families appear to be lineage-specific, e.g., seven in Saccharomycotina (see box 'A' in Figure 1), nine in Pezizomycotina (box 'B'), and six in Sordariomycetes (box 'C'). In all cases the originally reported representative of the family maps to these clades.

In contrast to lineage-specific families, lineage-specific losses of snoRNAs are also detectable. Basidiomycota, for example, do not seem 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 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 shows substantially more lineage specific innovation and deletion events than observed in box C/D snoRNAs, see Supplementary Figure S7. In total, 22 out of the 50 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, 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 Supplementary Figure S7.2). This stands in sharp contrast to C/D snoRNA sequence, where *P.tritici-*

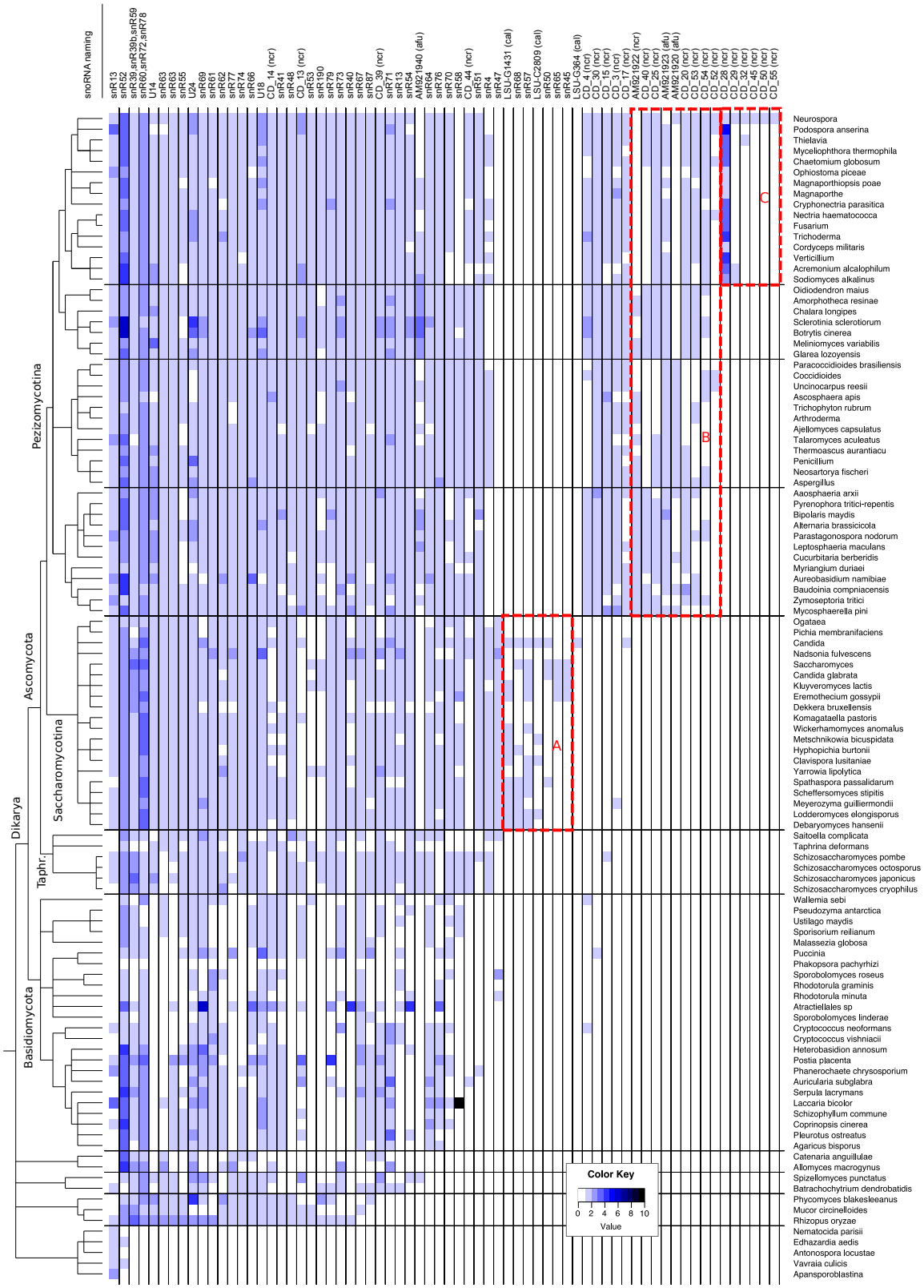

Figure 1: A heatmap of `snoStrip`-detected box C/D snoRNAs is shown on the previous site. Each column represents a specific snoRNA family, while each row either represents a certain species or genus. A taxonomic classification is shown on the left hand side. The amount of snoRNAs detected in a specific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

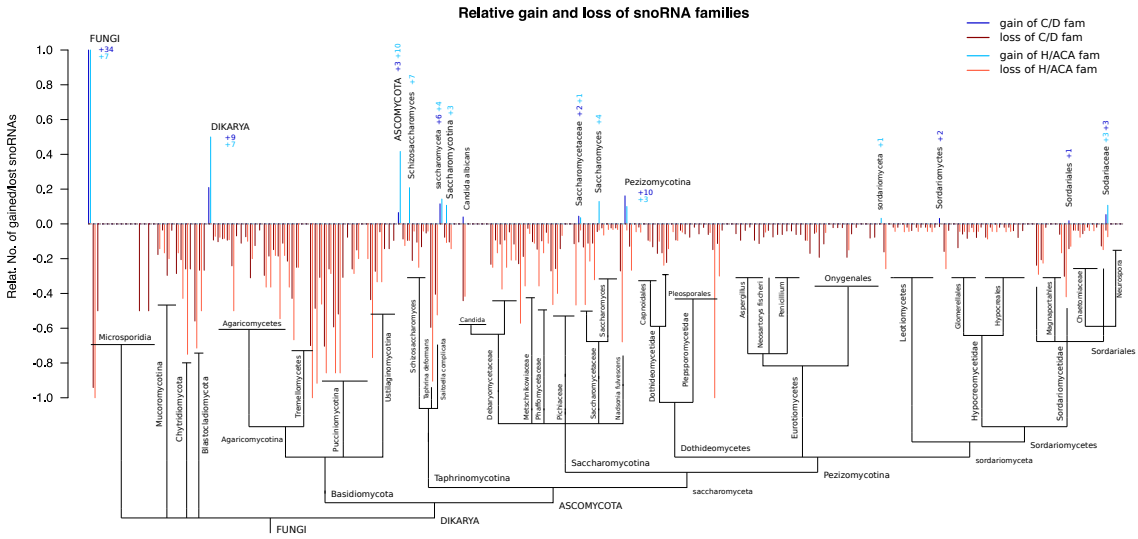

Figure 2: Relative number of gains and losses of entire snoRNA families during fungal evolution. The relative gain is the number of gained snoRNA families compared to the observed number of snoRNA families. The relative loss describes the number of lost snoRNA families compared to the number of snoRNA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

repentis orthologs are found in nearly all families that are present in the *P.tritici-repentis*-containing Dothideomycetes lineage.

3.4. 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 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 supplementary Figure S8.1 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, Ascomycota, 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.* already mentioned the remarkable absence of snoRNA genes in this clade, although all components of the snoRNA machinery are clearly present [40]. 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 is frequently observe that individual organisms seem to have lost a large amount of their snoRNAs, i.e. in the Basidiomycota lineage. In particular, *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 (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 of 40 snoRNA genes that showed no high sequence similarity towards already annotated budding yeast sequences [24]. 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 [24] (`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 [26] 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 [26] (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)).

3.5. 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.

IS THIS CLEARER NOW??: Among the 68 box C/D snoRNA families, the majority (40) are *true* single guides (28 families share a functional D' target; 12 a conserved D target). Another 14 box C/D snoRNA families are *predominantly* single guides, i.e., these families share exactly one overly 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. Another eight families harbor two functional target binding regions that are conserved throughout all lineages where these families are detected. The remaining 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, 15 families in hairpin 2). Six families exhibit a lineage specific HP2 target in addition to the globally conserved target in HP1. The opposite situation can be seen in 3 box H/ACA snoRNA families. 11 families are double guides, and 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 (sections S10-S21).

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 '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

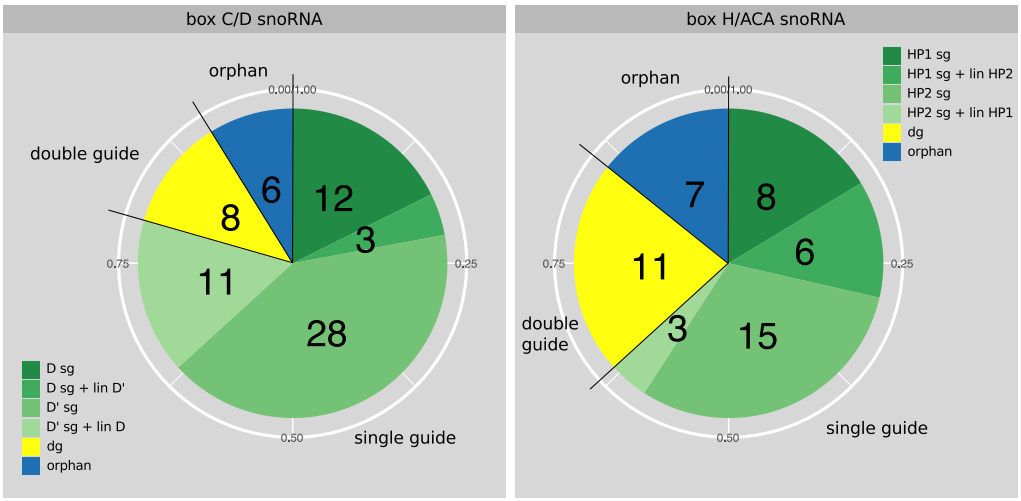

Figure 3: Pie chart of both major snoRNA classes. A snoRNA family is classified based on its conserved target prediction either as single guide (sg), single guide with a lineage specific target in its non-conserved target region (lin), double guide (dg), or orphan.

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 known [41] 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 [22], 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 S14. Highly conserved target interaction that are predicted by `snoStrip` are shown in Table 1.

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

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

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*

based on conservation aspects.

In 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 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 [26]. 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) [22]. 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 S20.

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 ~ 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 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 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 Supplementary Table S12.

Family snR87 (CD_10), for example, with its experimentally verified target 18S-479 (18S-436; D' target) [42], 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 [43]. 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* [22] 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 [44, 45]. 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 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*, 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.

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 [43]. 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 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

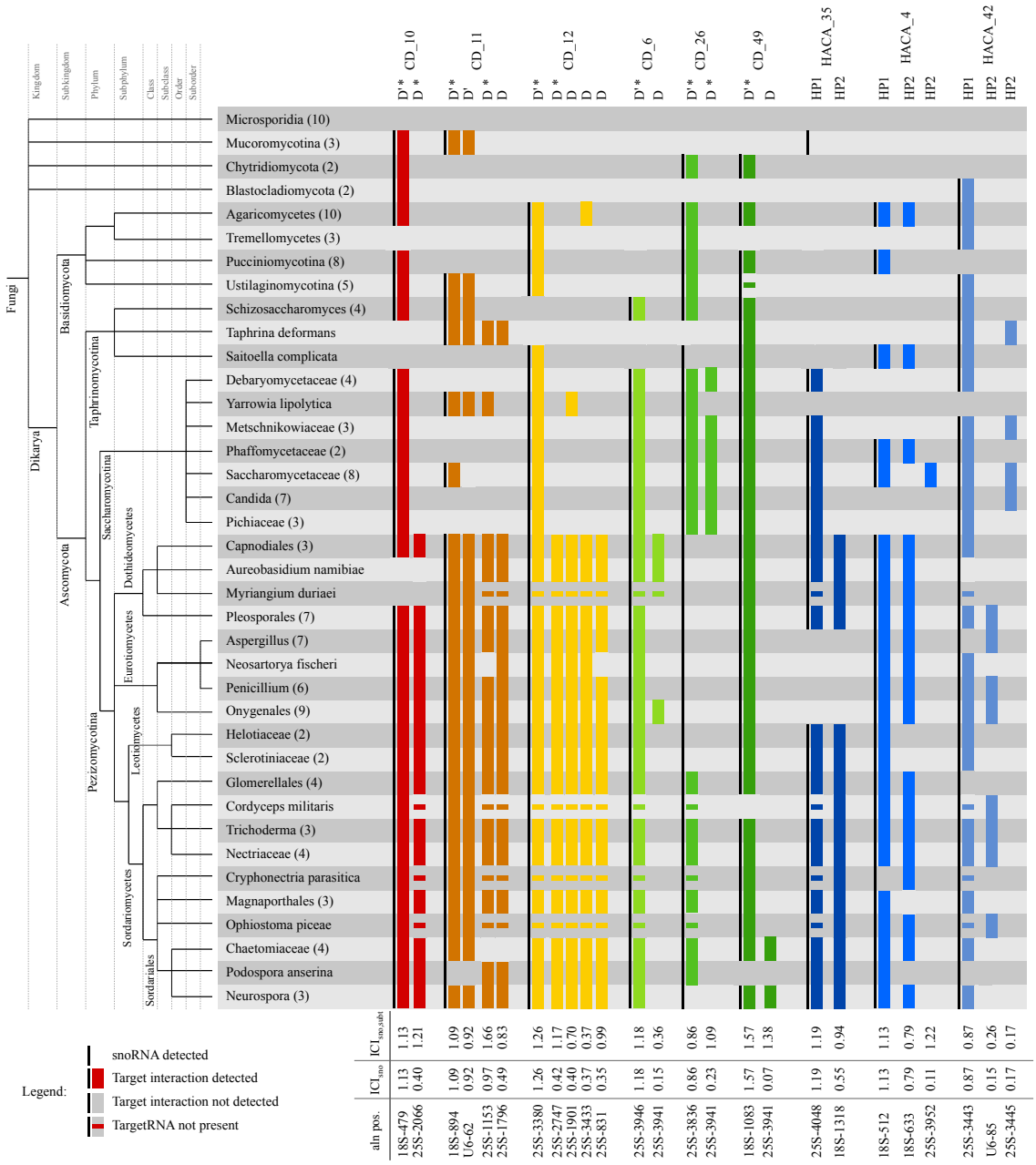

Figure 4: The conservation of predicted target interactions is shown for interesting single guide box C/D snoRNA families that exhibit an additional functional target at their 'non-functional' D box. Each family is depicted in a different color. The black bar in front of each family shows the presence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

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 supplement, see Table S18. 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* [41, 46]. 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 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 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 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.

3.6. 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 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 automatically 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).

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%

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 Supplement 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) [43]. 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 phylogenetic 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

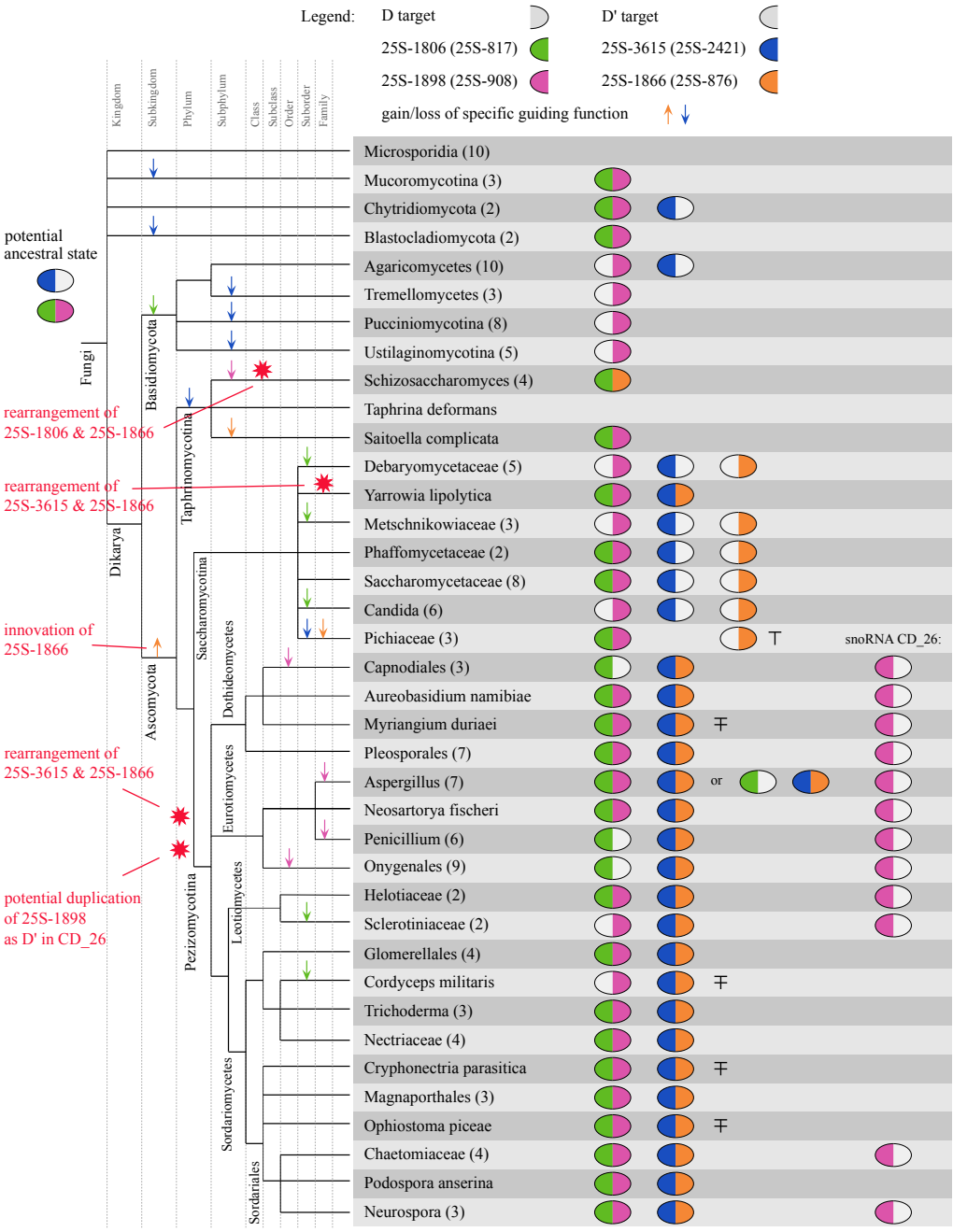

Figure 5: Potential evolutionary history of snoRNA clan CD_5 involving four modification sites on the LSU rRNA. Gain/loss events are displayed with arrows, while potential rearrangements are shown with red stars. ⊤ 25S-1866 is solely found in Pichia. ∓ Putative since LSU sequences are missing; snoRNAs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

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. The 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 in-

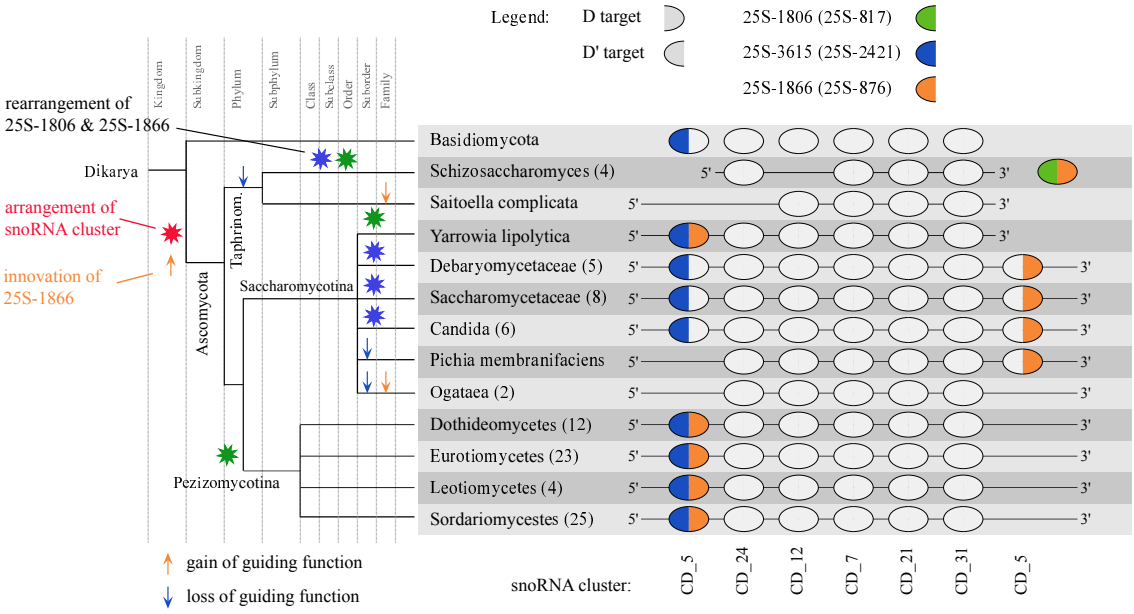

Figure 6: Sequences of the CD_5 snoRNA family are incorporate into a polycistronic transcript that harbors up to seven snoRNA genes. This cluster with its highly conserved structure and size occurred at the root of Ascomycota, but most of its genes arose at least at the root of Dikarya. There are different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

corporated into one snoRNA sequence after the original guidance of 25S-1898 (D target) was lost in this family.

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 approximately 1.5kb is 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 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 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 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 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

Table 4: Summary of 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

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 kcal/mol). Neurospora species are also predicted to guide this methylation with its Nc_CD_26 (CD_26) snoRNA[22]. 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).

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 given by box C/D snoRNA family snR40 (CD_43). It is predicted and experimentally validated [43] 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*, nor human, this prediction, albeit its overly convincing nature, remains hypothetical.

An even more impressive example is the D' ASE. of the snR70 (CD_61) family. No less than 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 [43], whereas the other predicted positions so far have not been shown to be chemically modified.

4. Discussion

We provide here a comprehensive inventory of snoRNAs in fungi together with a detailed analysis of the evolution of snoRNA families and their target specificities. The investigation of 147 different taxa provides a detailed history of gain, loss, and duplication events for 68 families of box C/D snoRNAs and 50 families of box H/ACA snoRNAs involving more than 7,700 individual snoRNA sequences. The processing of this amount 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 [17]. 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 tracked with high resolution based on the functional characteristics of the snoRNAs as determined by `snoStrip` together with a quantitative assessment of predicted RNA-RNA interaction based on the the Interaction Conservation Index (ICI) [31].

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 timescales. This effect may limit the scope of the homology search procedure. The short interacting regions

also make these molecules more vulnerable to mutations that disrupting the snoRNA-targetRNA 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 scoring 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 reported from 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 [41]. Both interactions can be tracked across the 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 phases, 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 [47]. In the budding yeast, it was already reported that stress-induced conditional pseudouridylations indeed exist in U2 snRNA [48]. Small nucleolar RNA snR81, which is also responsible for guidance of a constitutive U2 pseudouridylation, was therein shown to guide one of the novel modifications through imperfect and redundant base pairing abilities. The authors speculate that conditionally induced modifications in RNA in general are quite more frequent than previously thought.

On the other hand, 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 for the latter situation is the predicted pseudouridine at position 5.8S-18. Although there is not 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 supplement 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 recently was reported to be strongly regulated in development and between tissues or cell lines [49, 50]. 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 by means of paralog or through the redundant target binding capability of another snoRNA family.

In summary we observe that the landscape of snoRNAs keeps changing over the the evolutionary time-scale of the kingdom Fungi. We observe both the extinction of entire snoRNA families and the innovation of new ones. The function of snoRNA families itself also 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` [44, 51]) 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").

On the other hand, there are many similarities between the fungal and the metazoan snoRNAome. 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 animals, as reported by the human snoRNA atlas [50]: Over 70% of the human box C/D carrying snoRNAs are found to be single guided (75% in Fungi). In both human and fungi the remainder is about equally split between 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.

evtl noch 1-2 salbungsvolle schlusssaetze ...

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 German Federal Ministry of Education and Research, the CRC 1052 "Obesity", and by the Deutsche Forschungsgemeinschaft (Project DFG STA 850/15-1). more??. LIFE – Leipzig Research Center for Civilization Diseases is funded by the State of Saxony and the European Union.

References

[1] Reichow SL, Hamma T, Ferr-D'Amar AR, Varani G. The structure and function of small nucleolar ribonucleoproteins. Nucleic Acids Res 35, 2007:1452–64. doi: 10.1093/nar/gkl1172.

[2] Decatur WA, Fournier MJ. rrna modifications and ribosome function. Trends Biochem Sci 27, 2002:344–51.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] Maxwell ES, Fournier MJ. The small nucleolar rnas. Annu Rev Biochem 64, 1995:897–934. doi: 10.1146/annurev.bi.64.070195.004341.

[11] Tollervey D, Kiss T. Function and synthesis of small nucleolar rnas. Curr Opin Cell Biol 9, 1997:337–42.

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

[13] 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.

[14] Balakin AG, Smith L, Fournier MJ. The rna world of the nucleolus: two major families of small rnas defined by different box elements with related functions. Cell 86, 1996:823–34.

[15] Ganot P, Caizergues-Ferrer M, Kiss T. 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, 1997:941–56.

[16] 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.

[17] 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.

[18] Nawrocki EP, et al. Rfam 12.0: updates to the rna families database. Nucleic Acids Res 43, 2015:D130–7. doi: 10.1093/nar/gku1063.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] Kehr S, Bartschat S, Stadler PF, Tafer H. Plexy: efficient target prediction for box c/d snornas. Bioinformatics 27, 2011:279–80. doi:10.1093/bioinformatics/btq642.

[30] Tafer H, Kehr S, Hertel J, Hofacker IL, Stadler PF. Rnasnoop: efficient target prediction for h/aca snornas. Bioinformatics 26, 2010:610–6. doi: 10.1093/bioinformatics/btp680.

[31] 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.

[32] 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.

[33] Xia L, Watkins NJ, Maxwell ES. Identification of specific nucleotide sequences and structural elements required for intronic u14 snorna processing. RNA 3, 1997:17–26.

[34] Watkins NJ, et al. A common core rnp structure shared between the small nucleoar box c/d rnps and the spliceosomal u4 snrnp. Cell 103, 2000:457–66.

[35] Cahill NM, et al. Site-specific cross-linking analyses reveal an asymmetric protein distribution for a box c/d snornp. EMBO J 21, 2002:3816–28. doi: 10.1093/emboj/cdf376.

[36] Watkins NJ, Dickmanns A, Lhrmann R. 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, 2002:8342–52.

[37] Kiss-Lszl Z, Henry Y, Kiss T. Sequence and structural elements of methylation guide snornas essential for site-specific ribose methylation of pre-rrna. EMBO J 17, 1998:797–807. doi:10.1093/emboj/17.3.797.

[38] Normand C, Capeyrou R, Quevillon-Cheruel S, Mougin A, Henry Y, Caizergues-Ferrer M. Analysis of the binding of the n-terminal conserved domain of yeast cbf5p to a box h/aca snorna. RNA 12, 2006:1868–82. doi: 10.1261/rna.141206.

[39] 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.

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

[41] Schattner P, Decatur WA, Davis CA, Ares M Jr, Fournier MJ, Lowe TM. Genome-wide searching for pseudouridylation guide snornas: analysis of the saccharomyces cerevisiae genome. Nucleic Acids Res 32, 2004:4281–96. doi:10.1093/nar/gkh768.

[42] Davis CA, Ares M Jr. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit rrp6p in saccharomyces cerevisiae. Proc Natl Acad Sci U S A 103, 2006:3262–7. doi: 10.1073/pnas.0507783103.

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

[44] 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.

[45] 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.

[46] Badis G, Fromont-Racine M, Jacquier A. A snorna that guides the two most conserved pseudouridine modifications within rrna confers a growth advantage in yeast. RNA 9, 2003:771–9.

[47] 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.

[48] Wu G, Xiao M, Yang C, Yu YT. U2 snrna is inducibly pseudouridylated at novel sites by pus7p and snr81 rnp. EMBO J 30, 2011:79–89. doi:10.1038/emboj.2010.316.

[49] Kapushesky M, et al. Gene expression atlas update–a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res 40, 2012:D1077–81. doi:10.1093/nar/gkr913.

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

[51] Cantara WA, et al. The rna modification database, rnamdb: 2011 update. Nucleic Acids Res 39, 2011:D195–201. doi:10.1093/nar/gkq1028.