The Fungi snoRNAome

Sebastian Canzler[1], Peter F. Stadler[1,1,1,1,1,1,1], Jana Hertel[1,]

[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/16-XXX`.

Keywords: 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 the cell 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. However, several snoRNAs have been shown to target residues in other (small) RNAs [**?**]. 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. Like other small ncRNAs, snoRNAs are functional due to their secondary structure *and* characteristic (short) regions in their primary structure. 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`[**?**] is therefore not feasible. Our recently introduced computational annotation pipeline `snoStrip`[**?**] 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 related organism. In this study we analysed a large set of fungal species with genomes that are available in decent quality for their snoRNA abundance. We started with experimentally veri-

[*]Corresponding author

fied snoRNAs in five fungi. Further, we studied the evolutionary conservation of those snoRNAs and the co-evolution of snoRNAs and their targets. As a result 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 target switches), family assignment and evolutionary history. All data can be downloaded from our electronic supplement.

2. Materials and Methods

2.1. Genome and snoRNA Data

Genome sequences from 147 fungal species were downloaded from various sources, e.g. NCBI[?], ENSEMBL[?], USCS[?], etc. (others? references?) An NCBI-based taxonomic tree displaying the relationship, genome source, and genome version of all fungal organisms in the evolutionary survey is shown in the supplement, see Figures S?? and S??. A simplified taxonomic tree that shows the important fungal lineages is provided in Figure ??.

For 63 out of the 147 species, snoRNA sequences were readily available at the snoStrip webservere[1]. This data is based on a previous snoStrip study[?] that started with experimentally detected snoRNAs extracted from five surveys for *Neurospora crassa* [?], *Aspergillus fumigatus* [?], *Candida albicans* [?], *Saccharomyces cerevisiae* [?], and *Schizosaccharomyces pombe* [?]. An overview of the retrieved snoRNAs and the corresponding publications can be seen in supplementary Table ?? A complete mapping of the species-specific names as they were used in the respective publication and the internal snoStrip names is shown in supplementary Tables ?? and ??. . Note that annotated *A.fumigatus* box C/D snoRNA AM921943 is treated as a box H/ACA snoRNA. It shows two separated, perfect hairpins and comprises convincing box motifs while it clearly lacks characteristics of box C/D snoRNA. In addition, *A.fumigatus* box C/D snoRNAs AM921919 and AM921934 are treated as a single snoRNA. Both sequences map to the same genomic location. Though, AM921934 comprises three point-mutations compared to AM921919. All but budding yeast snoRNAs have been taken from the corresponding publication. Budding yeast sequences have been downloaded from the UMass-database[2]. In total, this starting set of snoRNAs comprises more than 3500 snoRNA sequences assigned to 123 snoRNA families in the 63 species. (Wrong(!) Table: Table 4.1: 68+50+1=119 != 120??) This is the final result of snoRNA families..., The 123 families are in the starting set...

Figure 1: Simplified NCBI-derived taxonomic tree displaying major fungal lineages.

2.2. Homology search

The snoStrip pipeline was applied to the set of collected snoRNAs and the 147 fungal species in an iterative manner. Starting with Pezizomycotina, followed by Saccharomycotina, down to the other lineages towards the root of the phylogenetic tree. Each time new homologous snoRNAs were detected, the procedure was repeated to decrease the number of false negatives until no novel homologs were found.

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 with start positions that are not conserved were corrected. Too large or too short sequences were 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[?].

[1] http://snostrip.bioinf.uni-leipzig.de/

[2] http://people.biochem.umass.edu/fournierlab/snornadb/main.php

2.5. Phylogenetic analysis

To follow the evolution of the snoRNA families along the phylogenetic tree we used the software `ePoPE`[**?**]. It implements an efficient algorithm to predict the last common ancestor of a gene family and a potential number of the observed number of paralogs at any inner node of the tree. 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` script 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 [**?**]. Briefly, the conservation of the modification and the conservation in a specific snoRNA family are calculated as follows:

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

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

Where $\varepsilon(t,s,k) = \min_{x \in X(t,s,k)} E_{\text{mfe}}[x, y_{t,k}]$ 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

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

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

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

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 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:

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

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

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

3. Results

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 shown in 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 written in parentheses.

3.1. The snoRNA complement in fungi

A set of 68 box C/D snoRNA and 50 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 [**?**]. All snoStrip retrieved snoRNA sequences were carefully cross-checked in all species to reduce the number of false negatives and exclude potential wrong associations. This resulted in a total amount of 5595 and 2331 box C/D snoRNA and box H/ACA snoRNA sequences. Hence, we expanded the fungi snoRNA complement by XXX% and immensely increased its density.

A straightforward downstream analysis of this dataset is the evolutionary analysis on family and class level. The ePoPE approach was used to identify the last common ancestor of each individual snoRNA family and to trace the evolution of gain and loss events from the LCA down to the leafs – that represent our species – of a phylogenetic tree.

3.2. Fungi snoRNA characteristics

Box motifs. Sequenc motifs were created from all snoStrip-annotated and can be found in the supplement. In general these motifs follow the propagated rules for canonical snoRNA box motifs.

Box C (RTGATGA) and D (CTGA) match the consensus sequence motifs almost perfectly. Box C, for example, shows a purine (R) in 92% of all cases. The 5' GA dinucleotide is present in all detected snoRNA candidates. In case of box D, the 5' nucleotide shows small mutations (4.2%) mostly towards an adenine but the remaining positions are highly conserved (≥99.7%).

This is dramatically different in prime box motifs. For box C', merely the 5' TG dinucleotide and, to a lesser extent, the trailing GA dinucleotide show high conservation which might point at a role in the binding of snoRNP associated proteins. In case of box D', variations of the canonical nucleotides occur quite frequently (between 15% and 45% in each position).

In case of box H/ACA snoRNAs, the sequence of box ACA is highly conserved with slight variations in its middle position. Box H (ANANNA) on the other side comprises highly conserved adenine residues in its 1st and 3rd position while the trailing adenine is more variable. The 2nd position of this motif is a guanine in nearly 80%, whereas the other "N" associated positions (4th and 5th position) do not show a favored nucleotide.

Sequence. In accordance with the published box C/D snoRNA length, 90% of the novel snoStrip-annotated are found to be 80nts to 135nts in length. The median length is 93nts (data not shown). Family CD_53 (originally Nc_CD_53 in [**?**]) is the only exception since its members share sequences with lengths between 200 and 300nts. 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. In case of box C/D' distances, the minimal gap is found to be 11nts while the median space is 24nts long. The gap between box C' and box D is at least 9nts long while the median is 22nts. The D'-C' distance is not known to be of significant relevance although at least 2nts are required to form another kink-turn motif with the aid of snoRNP associated proteins. Larger distances do not pose a problem. Within the fungi snoRNAs, the shortest distance is 3nts while 80% of all prime box annotated sequences possess distance between 6 to 31 nucleotides.

Box H/ACA snoRNAs are reasonably longer. Their median sequence length is 188nt. The shortest sequence being annotated by snoStrip is 115nts while 90% of all sequences are between 148 and 266nts long. When comparing both hairpins, no significant difference can be observed. Both share similar median values of 85nts and 79nts for hairpin 1 (HP1) and hairpin 2 (HP2), respectively. Extraordinary long snoRNAs can be found in families HACA_36 (snR86) and HACA_41 (snR84) with lengths of ~1000nt and ~600nt, respectively. Family HACA_12 (snR30), which is ~600nt long, provides an exceptional secondary structure with extensively enlarged 5' hairpins and hinge regions, where the latter one is also able to form a so-called internal hairpin [**?**].

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 [**?**]. Owed to these rather short ends, only a small subset of snoRNA sequences were found to be capable of folding a short stem (1208 of 5595). When the trailing ends are enlarged to 10nts instead, a stem was detected in nearly 60% (3317). These observations and the rather large fraction of snoRNAs that are still unable to fold into a characteristic hairpin indicate that a specific naturally occurring secondary structure is probably not needed for box C/D snoRNAs to 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 in order to function appropriately. Only 15% (395 of 2269) of all snoRNAs were not found to develop a hairpin like structure in both hairpins. A constraint folding, i.e., the 14th position either upstream of box H or ACA is forced to be unpaired, should enable the folding algorithm to ”open” the pseudouridylation pocket by predicting an interior loop at this position. Apperently, when comparing the natural folding with the constraint one, both mfe values are nearly identical. This clearly indicates that the *in vivo* folding already supplies sufficient access to the bipartite anti sense element and hence, the snoRNP associated proteins have a rather stabilizing role than to mould the sequence into a specific shape.

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

3.3. The evolution of snoRNAs

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

Commonly, fungal snoRNA families encompass exactly one snoRNA sequence per organism. Exceptions of this rule are given by families CD_5 and CD_19 whose coverage number mainly lies between two and three. This is rather a technical result of how `snoStrip` works than a detected biological event where snoRNA families contain mutliple copies of the exact same snoRNA gene. Due to detected target switches and target duplications `snoStrip` was promted to automatically merge previously separate snoRNA familes and hence, virtually increases the copy numbers. 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 ore 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 CD_41 or Sordariomycetes in CD_28.

Nearly 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 **??**), 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 is also detectable. Basidiomycota, for example, are not found to contain orthologs of families CD_8, CD_16, or CD_37, while in Saccharomycotina, no trace is found of snoRNAs of family CD_41. Members of CD_40 are not detected in Eurotiomycetes, while Sordariomycetes are attested to miss homologs of families CD_47 and 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.

In total, 22 out of 50 H/ACA families are merely

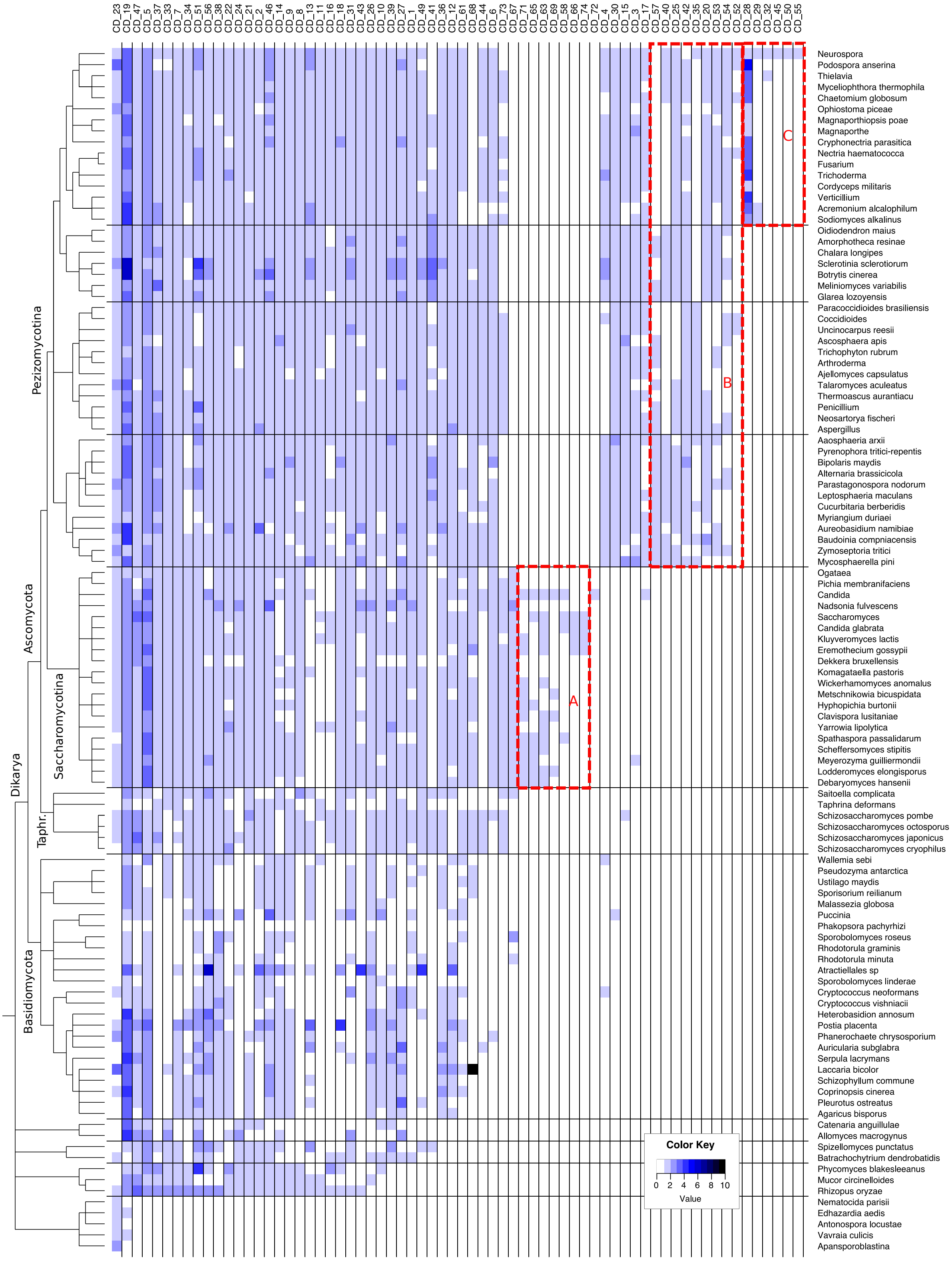

Figure 2: A heatmap of `snoStrip`-detected box C/D snoRNAs is shown on the previous site. Each column represents a specific snoRNA family, while each row either represents a certain species or genus. A taxonomic classification is shown on the left hand side. The amount of snoRNAs detected in a specific species and snoRNA family is encoded in a blue color scheme. Lineage specific families are boxed (A: Saccharomycotina, B: Pezizomycotina, C: Sordariomycetes).

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 HACA_33, HACA_56, and 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 events in snoRNA history. Relative innovation and deletion events mapped to the preordered nodes of the NCBI-derived taxonomic tree up to species level are shown in Figure **??**. Absolute events that are traceable from the root of major fungal lineages up to families and orders are shown in the supplement. In both images, it is clearly visible that a large amount of snoRNA families has evolved at each major branch point along the backbone of the taxonomic tree. In case of box C/D snoRNAs, 34 families are already present at 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 are already present at the root of fungi and additional 7, 10, 4, and 3 families are gained at the root of Dikarya, Ascomycota, saccharomyceta, and Pezizomycotina, respectively. Due to the homology-based search procedure in `snoStrip` which is based on a experimentally verified set of snoRNAs, it is quite logical that innovation events are exclusively detectable at nodes leading to the leaves that represent these five species.

Major snoRNA losses in both, an absolute and relative point of view, can be seen in Microsporidia which are found to harbour only two distinct box C/D snoRNA families while all remaining C/D and H/ACA families are not detectable. Gardner *et. al* formerly mentioned the remarkable absence of snoRNA genes in this clade, although all components of the snoRNA machinery are clearly present [**?**]. They argued with a lack of experimental investigations and only insufficient bioinformatic methods. However, the more sophisticated `snoStrip` approach was also not able to detect a great variety of snoRNA genes which might point at a rather diversified snoRNA repertoire compared to other fungal lineages.

When focusing on species level, it is frequently observed that single organisms seem to have lost a large amount of their snoRNA repertoire. In particular, species in the Basidiomycota lineage miss a fairly high portion of their snoRNAs. Especially *W.sebi* and several Pucciniomycota seem to have lost nearly their entire set of box H/ACA snoRNAs (*W.sebi*: 0.92, *R.minuta*: 0.86, or *S.linderae*: 0,86). 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*: 0.15 C/D and 0.13 H/ACA; *O.piceae*: 0.30 C/D and 0.42 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 [**?**]. One of these sequences is found to share a homologous target binding region with a known *N.crassa* snoRNA (CD_39). Families CD_69 (LSU-C2809 in [**?**]) and CD_71 (LSU-G1431) 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 CD_72 (LSU-G364) 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 [**?**] 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`, HACA_46 (AJ632008 in [**?**]) and HACA_47 (AJ632011) have been detected to be functional homologs to HACA_36 (snR86) and HACA_27 (snR5), 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

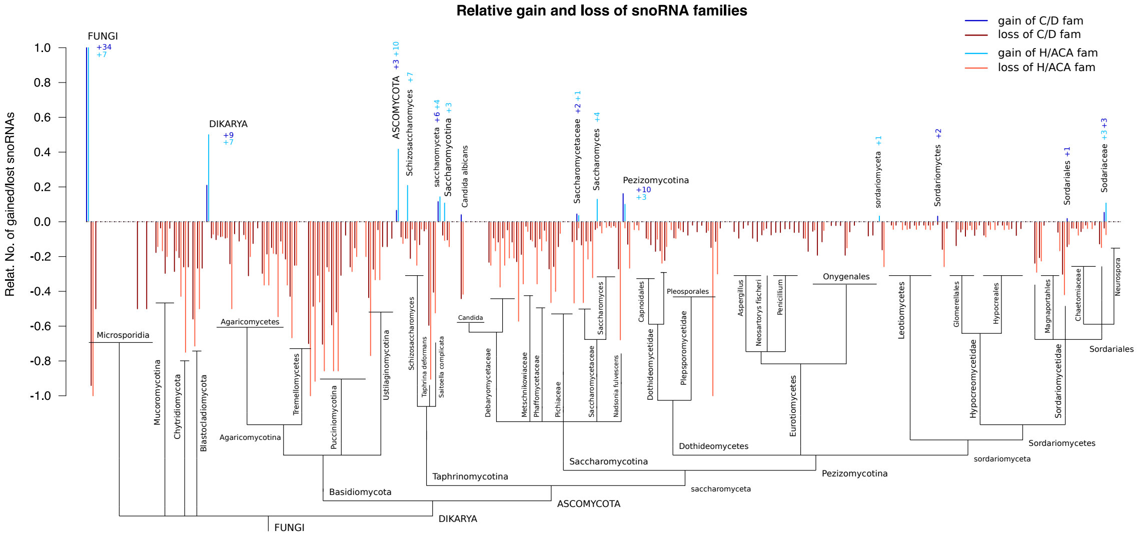

Figure 3: Relative number of gains and losses of entire snoRNA families during fungal evolution. The relative gain is the number of gained snoRNA families compared to the observed number of snoRNA families. The relative loss describes the number of lost snoRNA families compared to the number of snoRNA families in the parent node of the phylogenetic tree.

as homologous sequences. Families HACA_9 (AJ632018), HACA_48 (AJ632010), HACA_53 (AJ632016), and HACA_54 (AJ632012) 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 (HACA_50 (AJ632009), HACA_51 (AJ632017) and HACA_55 (AJ632013)) or exclusively found in *S.pombe* (HACA_45 (AJ632015), HACA_49 (AJ632019), and HACA_56 (AJ632014)).

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 **??**. 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.

Solely a minority of box C/D snoRNAs is found

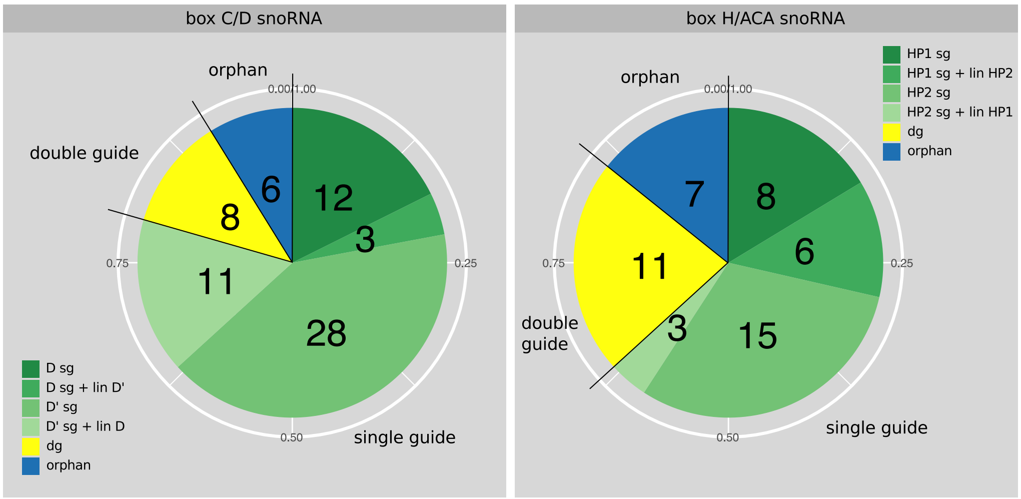

Figure 4: Pie chart of both major snoRNA classes. A snoRNA family is classified based on its conserved target prediction either as single guide (sg), single guide with a lineage specific target in its non-conserved target region (lin), double guide (dg), or orphan.

to contain two overly conserved target regions upstream of box D and D'. However, except for CD_5 and CD_19, none of the remaining six families is traceable amongst all major fungal lineages. Two families, CD_17 and CD_35, are found in Pezizomycotina while CD_67 is exclusively found in Saccharomycotina. The remaining families are either found in Sordariales (CD_32), a subgroup of Sordariomycetes, or in Glomerellales and Neurospora (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 (HACA_2, HACA_3, HACA_6, HACA_37). Two more families are traced to the root of Ascomycota (HACA_27, HACA_29), whereas the remaining five families are lineage (two found in Saccharomycotina, HACA_31, HACA_39) or genus specific (two found in Saccharomyces, HACA_26, HACA_30; one found in Schizosaccharomyces, HACA_46).

Family HACA_3 is published to guide three targets in both the budding yeast and fission yeast (annotated as snR3 in *S.cerevisiae*, AJ632000 in *S.pombe*);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, 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), *N.crassa* (2 sequences), 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 the supplement. Highly conserved target interaction that are predicted by `snoStrip` are shown in Table **??**.

Unfortunately, potential targets for both orphan

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*

N.crassa snoRNAs are not unambiguously discovered by snoStrip. The best prediction yields an ICI$_{sno}$ score of 0.71 for family CD_53 and is loosely found in several Pezizomycotina species (25S-3500, mean mfe: -11.56). The second family (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 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 **??**) 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 [**?**]. Ten of these targets were confirmed by conservation 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 (CD_10, D' target, ICI: 1.13; CD_26, D' target, ICI: 0.86), see Table **??**.

Family CD_54 was originally published to guide modification at 25S-1648 (*N.crassa* 25S-667; D target) [**?**]. By means of snoStrip, family 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 **??**. 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 sequence), and *S.pombe* (8 sequences). Again, a detailed summary of these sequence can be seen in the supplement.

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 (HACA_11 matches snR11, HACA_12 matches snR30, HACA_13 matches snR10, AM921943 matches snR32, and AJ632018 matches snR43). The three remaining snoRNA families comprise a conserved target in HP2, see Table **??**. Family HACA_7 is found to be a distant homolog to family 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 ; HACA_7 is ~ 180nt in length), `snoStrip` was unable to detect a potential common origin. Family 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 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 **??**. Detailed information about all snoRNA families with an additional, lineage specific target are found in the supplement.

Family CD_10, for example, with its experimentally verified target 18S-479 (*S.cerevisiae* snR87; 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. CD_11 was shown to guide the methylation at position 18S-894 (*S.cerevisiae* snR53; 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* [**?**] 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 **??**. Position 45 in U6 snRNA was not found to be modified in the budding yeast [**?** **?**]. 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 (*S.cerevisiae*) was experimentally shown [**?**]. The conserved interaction of this position is traceable in at least three different families, each in another fungal lineage. Family CD_26, which contains the budding yeast sequence, 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

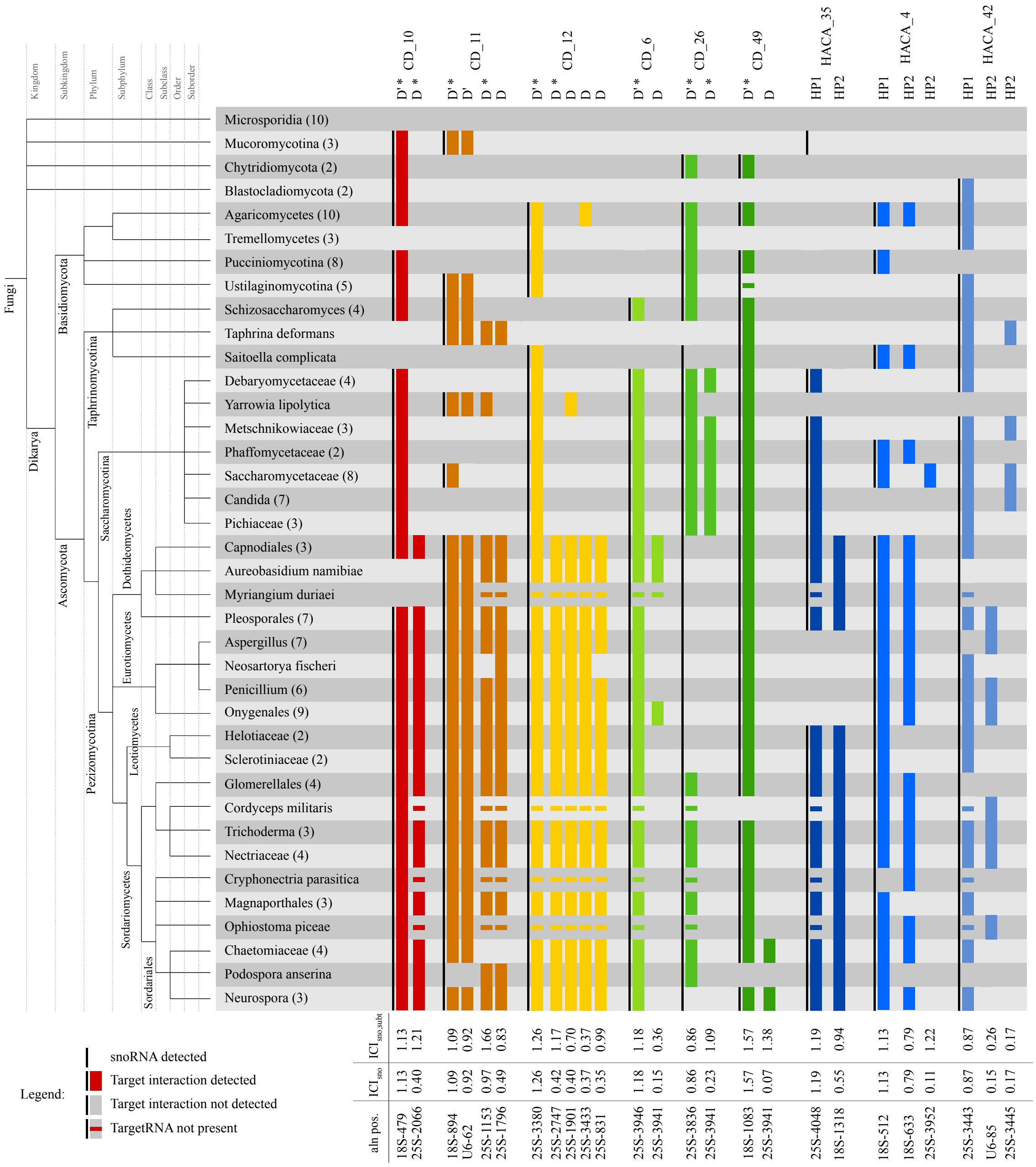

Figure 5: The conservation of predicted target interactions is shown for interesting single guide box C/D snoRNA families that exhibit an additional functional target at their 'non-functional' D box. Each family is depicted in a different color. The black bar in front of each family shows the presence 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 '*'.

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 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, CD_49, the modification at 25S-3941 is predicted in Sordariales (ICI: 1.38, mean mfe: -14.14).

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 **??**. Again, a detailed summary can be found in the supplement. Some of these functions might already been annotated, especially in snoRNA sequences of the budding yeast, see families HACA_4 and HACA_42 which are in fact officially denoted as double guides in *S.cerevisiae*. Both families show an interesting pattern in their second hairpins. HP1 is highly conserved in both cases 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. HACA_4, 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 HACA_42, 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 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.

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 **??** summarizes the target interactions that are convincingly predicted in the snoRNA clans CD_5 and CD_19.

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 exlained and illustrated in the supplement.

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 **??**. 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

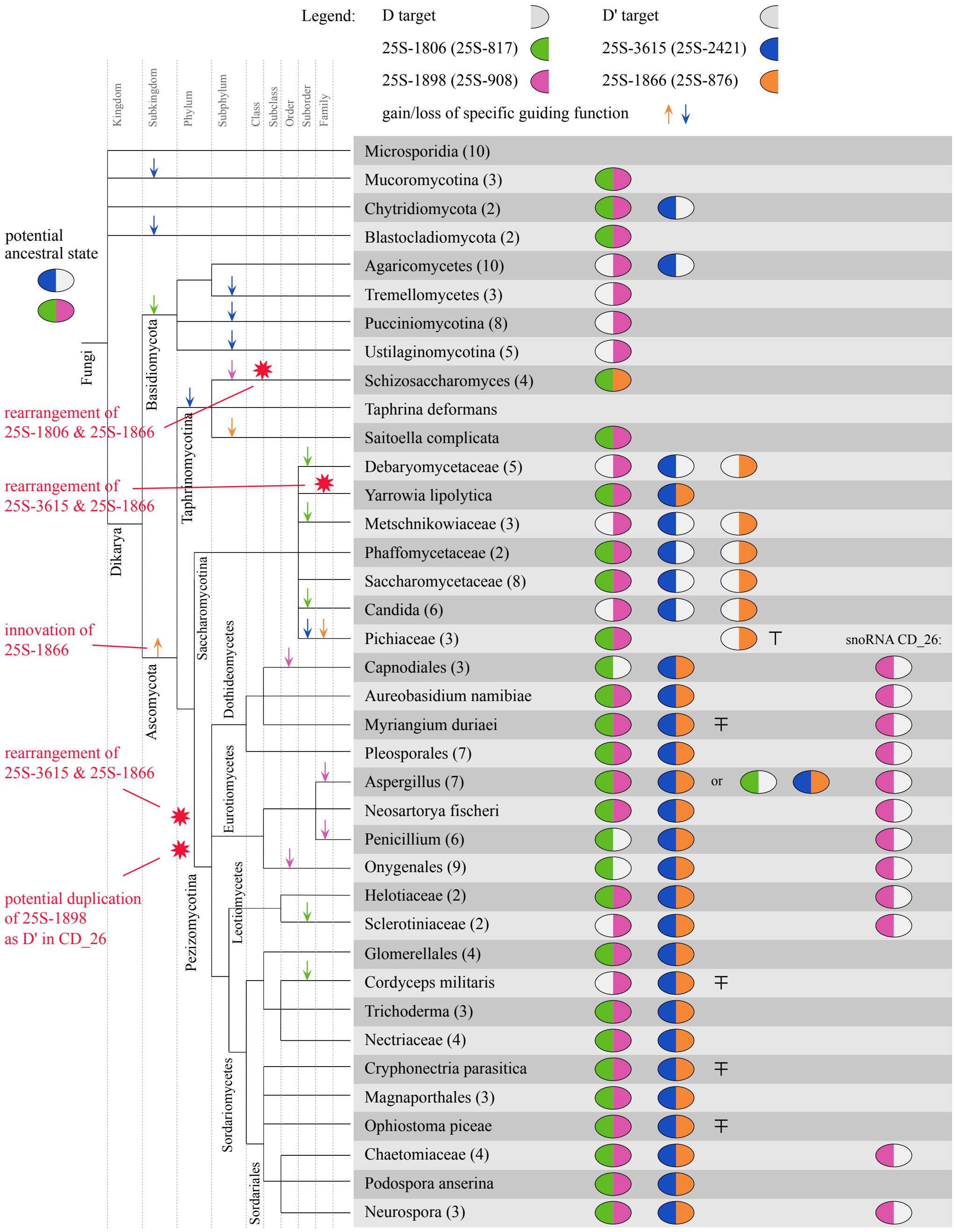

Figure 6: Potential evolutionary history of snoRNA clan CD_5 involving four modification sites on the LSU rRNA. Gain/loss events are displayed with arrows, while potential rearrangements are shown with red stars. ⊤ 25S-1866 is solely found in Pichia. ∓ Putative since LSU sequences are missing; snoRNAs show convincing ASE conservation.

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

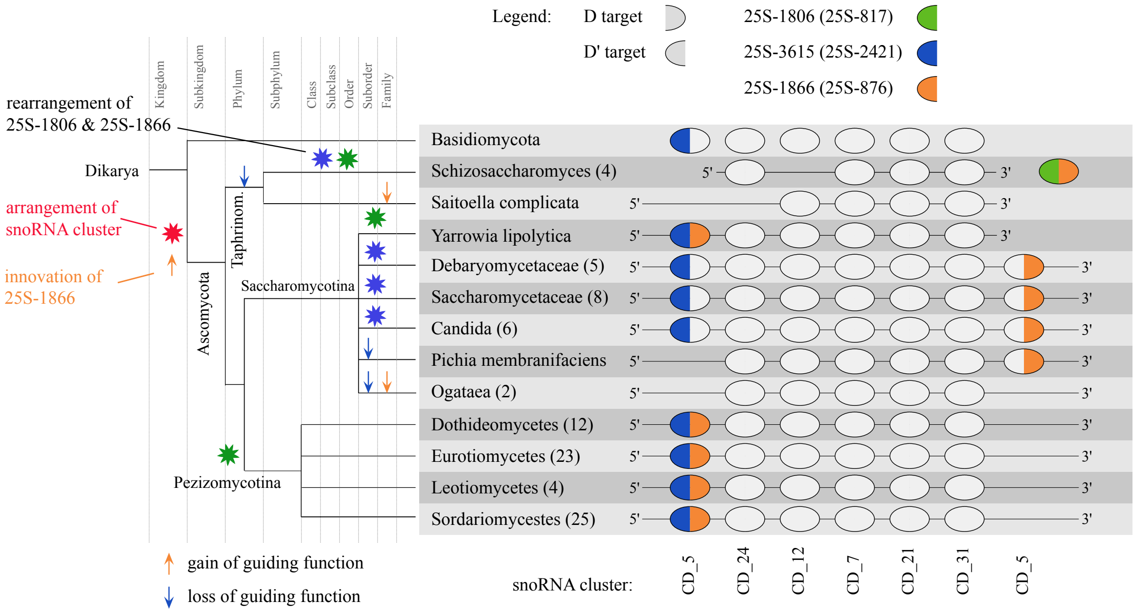

Figure 7: Sequences of the CD_5 snoRNA family are incorporate into a polycistronic transcript that harbors up to seven snoRNA genes. This cluster with its highly conserved structure and size occurred at the root of Ascomycota, but most of its genes arose at least at the root of Dikarya. There are different 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.

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 CD_24, CD_12, CD_7 CD_21, and CD_31 in 5'-3' direction, see Figure **??**. 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 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 **??**) 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 **??**). 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,

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

	pos	ICI$_{sno}$	∅ mfe	# ia
CD_43	18S-1400	0.95	-12.96	67/90
	18S-614	1.61	-21.96	71/90
CD_61	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

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 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 CD_26 snoRNA. In reverse the original D' target of CD_26, 25S-3836, was abolished in these organisms and is not found to be restored in any other snoRNA family. Please also confer Figure **??** 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. 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 CD_43 (*S.cerevisiae* snR40) 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 CD_61 (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 **??**. 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 keyfactors 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 computational 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 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 [**?**]. Convincing examples of remarkably conserved multiple interactions are given by box C/D snoRNA families CD_43 and 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 [**?**]. 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.