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 [**?**]. 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) [**? ? ?**], although some snoRNAs have been shown to target residues in other RNA molecules such as transfer RNAs [**? ?**], spliced leader RNAs [**?**], or brain-specific messenger RNAs [**? ?**]. Furthermore, snoRNAs are known to be involved in the nucleolytic processing of rRNA precursors, the synthesis of telomeric DNA, genomic imprinting, and alternative splicing [**? ? ? ?**].

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 [**? ?**]. Box C/D snoRNAs form a stem-loop structure with a rather long loop which is sta-

[*] 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)

bilized by the associated proteins and guide the 2'-O-methylation of ribose groups. Box H/ACA snoRNAs are longer, fold into a thermodynamically more stable double stem-loop structure, and guide 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 [**?**]. 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 [**? ?**]. 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` [**?**].

To overcome this limitation, we introduced a computational annotation pipeline `snoStrip` [**?**] 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` [**?**].

2. Materials and Methods

2.1. Genome and snoRNA Data

Genome sequences from 147 fungal species were downloaded from Ensembl Genomes [**?**], JGI [**?**], Broad Institute (Fungal Genome Initiative) , and Candida Genome Database [**?**]. 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` [**?**]. In this earlier work, we 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 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 [**?**] 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 [**?**] 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[**?**].

2.5. Phylogenetic analysis

To follow the evolution of the snoRNA families along the phylogenetic tree we used the software `ePoPE`[**?**]. It implements a variant of Sankoffs parsimony algorithm using the Dollo variant that excludes the loss and re-gain of a gene family along the same lineage during evolution. Innovation and deletion/loss/divergence events are deduced and mapped to the branches of the phylogenetic tree. The `ePoPE` results are combined for *all* snoRNA families using the `ePoPE_summarize.pl` tool that comes with the `ePoPE` distribution.

2.6. Target prediction and analysis

Target prediction is part of the `snoStrip` pipeline. There, the computational tools `PLEXY` and `RNAsnoop` are employed to predict targets for box C/D snoRNAs and box H/ACA snoRNAs, respectively [**?** **?**]. SnoRNAs are investigated for single or double guide potential based on these predictions and/or confirmed target interactions. SnoRNAs that remain without target association are considered orphan. SnoRNAs that are assigned to the same family but show variance in their associated target are investigated manually for a potential target switch.

2.7. Lineage specific conservation of target interactions

To study the conservation of interactions, the targets for each individual snoRNA sequence are initially predicted and subsequently their conservation in other species is evaluated. To formally investigate the conservation, the Interaction Conservation Index (ICI) was developed by [**?**]. 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 [**?**]. 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 [**? ? ? ?**] 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 [**? ?**]. 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 [**?**].

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