Expanding the SnoRNA Interaction Network

Conservation of Guiding Function in Vertebrates

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt

von Diplom-Informatikerin Stephanie Kehr
geboren am 04. August 1983 in Warburg (Westfalen)

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Peter F. Stadler, Universität Leipzig

2. Professor Dr. Daniel Gautheret Université Paris-Sud

Die Verleihung des akademischen Grades erfolgt mit Bestehen der
Verteidigung am 12.12.2016 mit dem Gesamtprädikat *summa cum laude*.

Contents

Abstract

Small nucleolar RNAs (snoRNAs) are one of the most abundant and evolutionary ancient group of small non-coding RNAs. Their main function is to target chemical modifications of ribosomal RNAs (rRNAs) and small nuclear (snRNAs). They fall into two classes, box C/D snoRNAs and box H/ACA snoRNAs, which are clearly distinguished by conserved sequence motifs and the type of modification that they govern. The box H/ACA snoRNAs are responsible for targeting pseudouridylation sites and the box C/D snoRNAs for directing 2'-O-methylation of ribonucleotides. A subclass that localize to the Cajal bodies, termed scaRNAs, are responsible for methylation and pseudouridylation of snRNAs. In addition an amazing diversity of non-canonical functions of individual snoRNAs arose. The modification patterns in rRNAs and snRNAs are retained during evolution making it even possible to project them from yeast onto human. The stringent conservation of modification sites and the slow evolution of rRNAs and snRNAs contradicts the rapid evolution of snoRNA sequences.

Recent studies that incorporate high-throughput sequencing experiments still identify undetected snoRNAs even in well studied organisms as human. The snoRNAbase, which has been the standard database for human snoRNAs has not been updated since 2006 and misses these new data. Along with the lack of a centralized data collection across species, which incorporates also snoRNA class specific characteristics the need to integrate distributed data from literature and databases into a comprehensive snoRNA set arose. Although several snoRNA studies included *pro forma* target predictions in individual species and more and more studies focus on non-canonical functions of subclasses a systematic survey on the guiding function and especially functional homologies of snoRNAs was not available.

To establish a sound set of snoRNAs a computational snoRNA annotation pipeline, named `snoStrip` that identifies homologous snoRNAs in related species was employed. For large scale investigation of the snoRNA function, state-of-the-art target predictions

were performed with our software `RNAsnoop` and `PLEXY`. Further, a new measure the Interaction Conservation Index (ICI) was developed to evaluate the conservation of snoRNA function.

The `snoStrip` pipeline was applied to vertebrate species, where the genome sequence has been available. In addition, it was used in several ncRNA annotation studies (48 avian, spotted gar) of newly assembled genomes to contribute the snoRNA genes. Detailed target analysis of the new vertebrate snoRNA set revealed that in general functions of homologous snoRNAs are evolutionarily stable, thus, members of the same snoRNA family guide equivalent modifications. The conservation of snoRNA sequences is high at target binding regions while the remaining sequence varies significantly. In addition to elucidating principles of correlated evolution it was possible, with the help of the ICI measure, to assign functions to previously orphan snoRNAs and to associate snoRNAs as partners to known but so far unexplained chemical modifications. As further pattern redundant guiding became apparent. For many modification sites more than one snoRNA encodes the appropriate antisense element (ASE), which could ensure constant modification through snoRNAs that have different expression patterns. Furthermore, predictions of snoRNA functions in conjunction with sequence conservation could identify distant homologies. Due to the high overall entropy of snoRNA sequences, such relationships are hard to detect by means of sequence homology search methods alone.

The snoRNA interaction network was further expanded through novel snoRNAs that were detected in data from high-throughput experiments in human and mouse. Through subsequent target analysis the new snoRNAs could immediately explain known modifications that had no appropriate snoRNA guide assigned before. In a further study a full catalog of expressed snoRNAs in human was provided. Beside canonical snoRNAs also recent findings like AluACAs, sno-lncRNAs and extraordinary short SNORD-like transcripts were taken into account. Again the target analysis workflow identified undetected connections between snoRNA guides and modifications. Especially some species/clade specific interactions of SNORD-like genes emerged that seem to act as *bona fide* snoRNA guides for rRNA and snRNA modifications. For all high confident new snoRNA genes identified during this work official gene names were requested from the HUGO Gene Nomenclature Committee (HGNC)[1] avoiding further naming confusion.

[1] `http://www.genenames.org/`

Acknowledgment

I want to thank Peter for the scientific and personal support, for offering the opportunities, and for creating the good atmosphere at the institute.
I want to thank Jana! Beside many scientific stuff I learned how to start making sense out of data, and get structure into texts. However, what was equally important and far from natural is the constant support you provided, especially in the more difficult times. And 'therefore' I intend to say much more than a usual thanks.
I want to thank my colleagues for the collaborations and good times. Especially, I want to thank Sebastian for sharing all the good ideas, and like Axel and Jan, for the 'sofa-sessions'.
I want to thank Frau Generalsekretärin Petra, for answering all questions one could ever have concerning bureaucracy in an unusual kind way.
I want to thank Jens for 'fixing the computers'.
I want thank my friends and family: Christian for proof reading many lunches and everything, my family for believing in me and supporting me, the best Mädels on earth for growing older, growing together and Liebe, Nilz for the fun and calmness, and all my friends for being my friends.

Abbreviations

A	Adenine
ASE	Antisense Element
bp	base pair
CB	Cajal Body
C	Cytosine
CM	Covariance Model
DNA	Desoxyribonucleic Acid
ETS	External Transcribed Spacer
FDR	False Discovery Rate
GFP	Green Fluorescent Protein
G	Guanine
HGNC	HUGO Gene Nomenclature Committee
HTP	High-throughput Sequencing
ICI	Interaction Conservation Index
iCLIP	Individual-nucleotide Resolution UV-Cross Linking and Immunoprecipitation
IL3	Interleukin 3
ITS	Internal Transcribed Spacer
lncRNA	long non-coding RNA
LSU	Large Subunit
mc	multi-copy
miRNA	micro RNA
mRNA	messenger RNA
N	any nucleotide
ncRNA	non-coding RNA
NGS	Next Generation Sequencing

NMD	Nonsense Mediated Decay
nt	nucleotide
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
ψ	Pseudouridine
Pus	Pseudouridine synthase enzyme
PWM	Position Weight Matrices
PWS	Prader-Willi Syndrome
RBP	RNA Binding Protein
RISC	RNA Induces Silencing Complex
RNA	Ribonucleic Acid
RNP	Ribonucleoprotein Particle
R	Purine
rRNA	ribosomal RNA
sdRNA	small derived snoRNA
SNHG	snoRNA Host Gene
snoRNA	small nucleolar RNA
SNP	Single Nucleotide Polymorphism
snRNA	small nuclear RNA
SSU	Small Subunit
SVM	Support Vector Machine
TERC	Telomerase RNA component
TF	Transcription Factor
TGD	Teleost Genome Duplication
tRNA	transfer RNA
T	Thymine
UTR	Untranslated Region
U	Uracil
WT	Wild Type
Y	Pyrimidine

CHAPTER 0

The Puzzle

The thesis starts sharing the fascination of genome research in general and the motivation to investigate small nucleolar RNAs (snoRNAs) in particular. An overview about the structure of this work is provided, introducing all respective publications.

The Fascination of Genome Research... We are still at the beginning of understanding why each cell is running the right genetic program in the specific tissue and developmental state and according to heterogeneous environmental conditions, although each and every nucleus holds the exact same copy of DNA. Yet, it is clear that massive regulation is needed to ensure this phenomenon. Many regulatory non-coding RNAs (ncRNAs) have been discovered in genomic regions that have been overseen as 'junk' only 30 years ago. The ncRNAs affect all thinkable stages and mechanisms of gene expression, encompassing transcription, translation, splicing, secondary structure formation, chromatin packaging, and gene degradation, and many more. The emerging complexity with ongoing new and often unexpected discoveries, permanently presents new riddles. These riddles can be seen as parts of a huge puzzle. By solving parts of it, together the huge picture will become recognizable.

...and Small Nucleolar RNAs. This work focuses on small nucleolar RNA (snoRNAs), which are an abundant and evolutionarily ancient class among the mentioned regulatory non-coding RNAs. It is known that their main function is the modification of ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs) at multiple sites through base pairing interactions with specific regions. As such they impact two basic cellular processes. The rRNAs are part of the ribosomes, the site where the genetic code of a messenger RNA (mRNA) is translated into the aminoacid chain of a protein. SnRNAs are the constituents of the splicing machinery and thus responsible for correct (alternative) splicing of the pre-mRNA into mature mRNA. In the last years a considerable diversity of structure, genomic organization and function of snoRNAs has been discovered. They have been found to be involved in a multitude of other cellular processes, including such diverse mechanisms as stress response or chromatin remodeling.

Although snoRNA sequences have been annotated in several model organisms next generation sequencing data still unearth new and more variant snoRNA sequences. The question of how many snoRNA genes exist is still unanswered. Moreover a systematic survey, that annotates homologous snoRNAs considering also the functional homologies was still missing. On top the fact of more or less unconnected surveys in different organisms, led to confusion in snoRNA naming and obscure sequence homologies. Currently available snoRNA resources have different drawbacks. Some species specific databases

are outdated, like snoRNA-LBME-db[1] a snoRNA database for human snoRNAs and UMASS yeast snoRNAdb[2] which have both last been updated in 2006. Also information beyond nucleotide sequences are not designated for computational querying. Exemplary, details of the snoRNA-target RNA interactions in snoRNA-LBME-db are only provided as .gif-formatted images. A database under development capturing orthologic snoRNAs in metazoan is snoRNA Orthological Gene Database snOPY[3]. Unfortunately, the data is not yet complete and needs further curation. Also generic RNA databases like RFAM[4] (RNA families database) or Ensembl contain snoRNA data. However, the database scheme is not intended to capture specific snoRNA characteristics. For example box annotation or target information is not included there. Many of the contained sequences in the latter two databases have been added by scanning the genome for sequence (BLAST) and structure (Infernal) similarity alone, which involves the risk of providing false positives and non-functional pseudogenes, as in case of snoRNAs the presence of defined sequence motifs is obligatory.

In summary a comprehensive set of snoRNA families in vertebrates and other species was still missing and urgently needed. Although several studies included *pro forma* target predictions in individual species a systematic survey on the guiding function and especially functional homology of snoRNAs was not available. Obviously, only on a solid knowledge base the 'common' characteristics and interactions of snoRNAs can be studies. Herein, considering the class specific features of snoRNAs and focusing also on their evolution and the evolution of their function on a large scale. The integration of all these data into a snoRNA interaction network will help to draw sound conclusions about their contributions in diseases or understand their observed diversity and specialization.

To contribute a further piece to the complex puzzle of gene expression a computational snoRNA annotation pipeline, software for snoRNA target prediction, and a measure to evaluate the conservation of function were developed. The methods were applied to sets of vertebrate snoRNAs to gain insight into snoRNA evolution with respect to their targets.

[1] https://www-snorna.biotoul.fr/
[2] http://people.biochem.umass.edu/sfournier/fournierlab/snornadb/main.php
[3] http://snoopy.med.miyazaki-u.ac.jp/
[4] http://rfam.xfam.org

Structure of this Work

The thesis starts with introducing the relevant technical and biological concepts in the first two chapters. In the first chapter (Technical Introduction) basic bioinformatic methods that are used in the context of the work are introduced. In Chapter 2 (Introduction to Small nucleolar RNAs) gene expression is briefly summarized and detailed background information to small nucleolar RNAs (snoRNA) is provided.

Chapter 3 (Innovative Analysis of SnoRNA-Target Interactions) starts with providing the materials and methods used for the analysis. First, it is described how non-coding RNA annotation is performed in general than focusing on snoRNA annotation in particular. With our `snoStrip` pipeline and the associated `snoBoard` database (Section 3.1) homology based snoRNA annotation considering class specific features is automated. The snoRNA sequences that serve as start query set in vertebrates are specified. Further, the sources and processing steps for the target RNAs and according modifications are described. Also the methodologically contributions made to the field of snoRNA research are presented. To investigate the functions of snoRNAs, target predictions programs which minimize free energy were developed (Section 3.2). Due to different binding patterns, `RNAsnoop` solves this problem for box H/ACA snoRNAs and `PLEXY` for box C/D snoRNAs. To incorporate conservation information to support the short snoRNA-target RNA interaction predictions a score called Interaction Conservation Index (ICI) (Section 3.2.3) was developed. It also enables to study the conservation of snoRNA function. At the end of the chapter a workflow incorporating all tools to systematically analyze the evolution of snoRNAs and snoRNA function is described. There also the different conclusions, that expand the snoRNA interaction network and that can be drawn from this innovative way of analyzing the snoRNA functions are previewed.

The subsequent chapters present results of the different research projects where the developed methods were applied to gain new insights into snoRNAs, snoRNA functions and the conservation of both. Contribution to several ncRNA annotation projects (50 avian genomes and the spotted gar) was made by screening for snoRNAs with the use of the `snoStrip` pipeline (Chapter 4 (SnoRNA Screens)). Observed exceptional snoRNAs and scaRNAs were investigated in more detail. The study includes comparative genomics and the results are presented in Chapter 5 (Exceptional SnoRNAs). Then, the systematic study on the co-evolution of snoRNAs and their targets in verte-

brates are presented in Chapter 6 (Matching of Soulmates: Co-evolution of SnoRNAs and their Targets). There the general pattern of conservation of snoRNA function is treated. Afterwards, the project in which cross-linking experiments with coilin revealed novel snoRNAs in human and mouse is handled in Chapter 7 (SnoRNA Interactions with Coilin in Cajal Bodies). Therein also the exact binding between coilin and the snoRNAs was inspected. The comprehensive collection of human snoRNAs, including canonical and non-canonical, known and novel sequences is then described in Chapter 8 (The Human SnoRNAome). For all identified human snoRNAs, expression was inspected using ENCODE small RNA data. Furthermore, a detailed analysis of human snoRNA functionality is encompassed.
Finally, in the work is discussed (Chapter 9 (Discussion and Outlook)), emphasizing the novelties contributed to the research field and giving perspectives for future projects in an outlook section.

List of Publications

[1] S Bartschat, **S Kehr**, H Tafer, P F Stadler, and J Hertel. snoStrip: A snoRNA annotation pipeline. *Bioinformatics*, 30(1):115–6, Jan 2014. doi: 10.093/bioinformatics/btt604.

[2] H Tafer, **S Kehr**, J Hertel, I L Hofacker, and P F Stadler. RNAsnoop: efficient target prediction for H/ACA snoRNAs. *Bioinformatics*, 26(5):610–6, Mar 2010. doi: 10.1093/bioinformatics/btp680.

[3] **S Kehr**, S Bartschat, P F Stadler, and H Tafer. PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics*, 27(2):279–80, Jan 2011. doi: 10.1093/bioinformatics/btq642.

[4] Y Huang, Y Li, D W Burt, H Chen, Y Zhang, W Qian, H Kim, S Gan, Y Zhao, J Li, K Yi, H Feng, P Zhu, B Li, Q Liu, S Fairley, K E Magor, Z Du, X Hu, L Goodman, H Tafer, A Vignal, T Lee, K. W Kim, Z Sheng, Y An, S Searle, J Herrero, M A Groenen, R P Crooijmans, T Faraut, Q Cai, R G Webster, J R Aldridge, W C Warren, S Bartschat, **S Kehr**, M Marz, P F Stadler, J Smith, R H Kraus, Y Zhao, L Ren, J Fei, M Morisson, P Kaiser, D K Griffin, M Rao, F Pitel, J Wang, and N Li. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat. Genet.*, 45(7):776–83, Jul 2013.

[5] C Anthon, H Tafer, J H Havgaard, B Thomsen, J Hedegaard, S E Seemann, S Pundhir, **S Kehr**, S Bartschat, M Nielsen, R O Nielsen, M Fredholm, P F Stadler, and J Gorodkin. Structured RNAs and synteny regions in the pig genome. *BMC Genomics*, 15:459, 2014. doi: 10.1186/1471-2164-15-459.

[6] P P Gardner, M Fasold, S W Burge, M Ninova, J Hertel, **S Kehr**, T E Steeves, S Griffiths-Jones, and P F Stadler. Conservation and losses of non-coding RNAs in avian genomes. *PLoS One*, 10(3):e0121797, 2015. doi: 10.1371/journal.pone.0121797.

[7] I Braasch, A R Gehrke, J J Smith, K Kawasaki, T Manousaki, J Pasquier, A Amores, T Desvignes, P Batzel, J Catchen, A M Berlin, M S Campbell, D Barrell, K J Martin, J F Mulley, V Ravi, A P Lee, T Nakamura, D Chalopin, S Fan, D Wcisel, C Caestro, J Sydes, F E Beaudry, Y Sun, J Hertel, M J Beam, M Fasold, M Ishiyama, J Johnson, **S Kehr**, M Lara, J H Letaw, G W Litman, R T Litman, M Mikami, T Ota, N R Saha, L Williams, P F Stadler, H Wang, J S Taylor, Q Fontenot, A Ferrara, S M Searle, B Aken, M Yandell, I Schneider, J A Yoder, J N Volff, A Meyer, C T Amemiya, B Venkatesh, P W Holland, Y Guiguen, J Bobe, N H Shubin, F Di Palma, J Alföldi, K Lindblad-Toh, and J H Postlethwait. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet*, 48(4):427–37, Apr 2016. doi: 10.1038/ng.3526.

[8] P D Bhattacharya, S Canzler, **S Kehr**, J Hertel, I Grosse, and P F Stadler. Phlogenetic distibution of plant snoRNA families. submitted to *BMC Genomics*, 2016.

[9] M Marz, A R Gruber, C Höner zu Siederdissen, F Amman, S Badelt, S Bartschat, S H Bernhart, W Beyer, **S Kehr**, R Lorenz, A Tanzer, D Yusuf, H Tafer, I L Hofacker, and P F Stadler. Animal snoRNAs and scaRNAs with exceptional structures. *RNA Biol*, 8(6):938–46, 2011. doi: 10.4161/rna.8.6.16603.

[10] **S Kehr**, S Bartschat, H Tafer, P F Stadler, and J Hertel. Matching of Soulmates: Co-evolution of SnoRNAs and Their Targets. *Mol. Biol. Evol.*, 31(2):455–467, Feb 2014. doi: 10.1093/molbev/mst209.

[11] M Machyna, **S Kehr**, K Straube, D Kappei, F Buchholz, F Butter, J Ule, J Hertel, P F Stadler, and K M Neugebauer. The Coilin Interactome Identifies Hundereds of

Small Noncoding RNAs that Traffic through Cajal Bodies. *Molecular Cell*, 56(3): 389–399, Nov 2014. doi: 10.1016/j.molcel.2014.10.004.

[12] H Jorjani*, **S Kehr***, D J Jedlinski, R Gumienny, J Hertel, P F Stadler, M Zavolan, and A R Gruber. An updated human snoRNAome. *Nucleic Acids Res*, 44 (11):5068–82, Jun 2016. doi: 10.1093/nar/gkw386.

CHAPTER 1

Technical Introduction

The chapter is a technical introduction to relevant RNA bioinformatic approaches. Here the tools, which were frequently used in this work are explained. The concept of dynamic programming is presented. It can be used to compute alignments and RNA secondary structures. As next generation sequencing data have been included into several of our studies, important methods and according analysis steps are summarized, Last an introduction is given to databases that were used to retrieve different kind of genomic data.

1.1 Dynamic Programming and Recursion

A fundamental approach to solve complex optimization problems is dynamic programming. The approach encompasses a forward and a backward recursion. In the first phase the problem is recursively subdivided into smaller subproblems until a trivial solution exists. Solutions of the subproblems are captured in matrices allowing reuse of already computed subproblems. In the backward recursion, an optimal solution is assembled by backtracking through this matrix, i.e. re-finding that path, that led to the optimal outcome.

1.1.1 Alignments

Alignments are used to compare two strings. In bioinformatics these strings are sequences of amino acids or nucleotides. In the context of his work several different sequence alignment programs are used to compare RNA sequences. The problem is typically solved by means of dynamic programming.
The basic recursion for the matrix M to solve a global alignment between sequences $X = x_1x_2 \dots x_n$ and $Y = y_1y_2 \dots y_n$ is provided below. The matrix entry for the alignment at position (i,j) between $x_i \in X$ and $y_j \in Y$ is obtained by:

$$M_{ij} = min/max \begin{cases} M_{i-1,j-1} \pm \delta(x_i, y_j) & (mis)match \\ M_{i-1,j} \pm \sigma(x_i, -) & deletion \\ M_{i,j-1} \pm \sigma(-, y_j) & insertion \end{cases} \tag{1.1}$$

There are two ways to optimize an alignment of two sequences, either the distance is minimized or the similarity maximized.
Usually this model is parameterized as follows:

$$\begin{aligned} \delta(x_i, y_j) &= \begin{cases} \alpha, x_i == y_i \\ \beta, x_i \neq y_i \end{cases} \\ \sigma(x_i, -) &= \sigma(-, y_j) = \gamma \\ \alpha, \beta, \gamma &\in \mathbb{N} \end{aligned} \tag{1.2}$$

Then, α, β are the (mis)match costs or similarity bonus and deletion and insertion, i.e. the inclusion of a gap character to one of the sequences, are defined by costs γ.

Table 1.1: Overview of the alignment tools used in this work.

program	type	application	reference
`BLAST`	local pairwise alignment, heuristic	homology search	(Altschul et al., 1997)
`Infernal`	sequence structure alignment, probabilistic	homology search	(Nawrocki and Eddy, 2013)
`muscle`	multiple alignment	RNA alignments	(Edgar, 2004)

There are several variants of the algorithm. Inclusion of a 4th case into the recursion, which assigns 0 values for start and end gaps and/or when the score of the recursion falls below 0 (in max case), leads to local and semi-global alignments, allowing also to search for matching sub-sequences or small nucleotide sequences in a longer one.
Besides dynamic programming, especially when searching in large databases for short sequences (k-mers) heuristic methods are preferred.
Local alignments are also used for homology searches, i.e. identify sequences (genes) that have the same evolutionary origin in genomes. A smaller query sequence is aligned to a larger search sequence, typically a chromosome or scaffold. The widely used tool and probably best known bioinformatic software to solve this problem is `BLAST` (Altschul et al., 1997). For fast evolving RNAs, the sequence similarity alone is often not sufficient to uncover orthology of genes, a tool to incorporate also the conservation of structure is `Infernal` (Nawrocki and Eddy, 2013). It follows a probabilistic approach starting from covariance models of RNA families. An overview of different alignment methods used during this work is given in Table 1.1.

1.1.2 RNA Secondary Structure Prediction

An inherent property of RNA and also DNA is its potential to form base pairs. Base pairs can occur inter-molecular, i.e., an RNA molecule folding back on itself or intra-molecular between two different RNA molecules. The problem of predicting these base pairs is known as RNA secondary structure prediction or RNA folding and RNA-RNA interaction, respectively. To predict the secondary structure of an RNA sequence again the technique of dynamic programming is used. The problem is divided into small problems with known solutions.
For the computations three assumptions are made:

1. Each nucleotide can be part of at most one base pair, that means if (i, j) is a bp and (j, k) is a base pair $\implies i = k$.

2. There are no crossing base pairs, that means if (i, j) is a bp and (k, l) is a base pair $\implies i < k < l < j$.

3. Paired bases must have a minimal distance of three $i < j + 3$.

Nussinov et al. (1978) developed one of the first approaches to solve secondary structure prediction by simply maximizing the base pairs within an RNA sequence. The optimal structure of the whole sequence $S_{1,n} = s_1 \dots s_n$ is composed of optimal structures of its sub-sequences. There are two possibilities, for a base s_i in the sub-sequence $S_{i,j} = s_i \dots s_j$, $(1 \leq i < j \leq n)$ Figure 1.1:

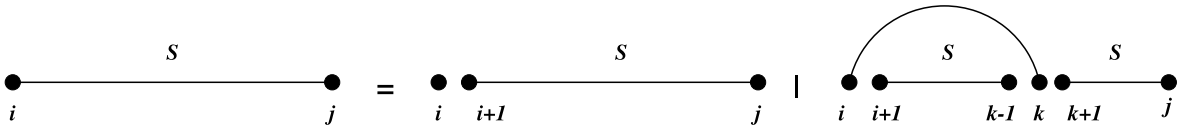

Figure 1.1: Schematic representation of maximizing the number of base pairs with the approach from Nussinov et al. (1978). Either a base s_i is unpaired or (s_i, s_k) form a base pair. Details see text. Figure adapted from Hertel et al. (2009).

1. s_i is unpaired, then the optimal solution is determined by the maximum number of base pairs in the shortened sequence $S_{i+1,j} = s_{i+1} \dots s_j$

2. s_i forms a base pair with s_k, then $S_{i,j}$ is divided into an enclosed sub-sequence $S_{i+1,k-1}$ with closing pair (i, k) and a sub-sequence $S_{k+1,j}$. The maximum number of base pairs (mnbp) in the sequence $S_{i,j}$ ($mnbp_{i,j}$) is then given by : $mnbp_{i,j} = max(mnbp_{i+1,j}, mnbp_{i+1,k-1} + mnbp_{k+1,j} + 1)$.

To predict the thermodynamically most stable structure of a sequence the problem can be interpreted as minimizing the energy of the system. Indeed not the hydrogen bond between the single paired nucleotides, but the dipole-dipole interaction between the aromatic rings of stacked base pairs are the main stabilizing structure components. Stretches of unpaired bases have a destabilizing effect. Combinations of stacked base pairs and unpaired bases shape different loops, which are the constituents of RNA secondary structures.

A decomposition of a structure into its constituting loops and according recursions are shown in Figure 1.2. The minimum free energy of the sequence F_{ij} is found by decomposing it into its loops. Therein the principal is similar to Nussinovs decomposition either a base is unpaired or paired. Obviously the resulting structures are more intricate, as enclosed structures $\mathcal{C}$ can be further decomposed into hairpin loops $\mathcal{H}$, interior loops $\mathcal{I}$ and multiloops of $\mathcal{M}$. In the latter case the number of sub-components is essential for the total multiloop energy. Therefore, introduction of substructure which forms a multiloop, always requires a 5'-component $\mathcal{M}$ and a 3'-component $\mathcal{M}^{\infty}$, which are then further decomposed.

$$F_{i,j} = \min \left\{ \begin{array}{l} F_{i+1,j}, \\ \min\limits_{i<k\leq j} C_{i,k} + S_{k+1,j} \end{array} \right\}$$

$$C_{i,j} = \min \left\{ \begin{array}{l} \mathcal{H}(i,j), \\ \min\limits_{i<k<l<j} C_{k,l} + \mathcal{I}(i,j;k,l), \\ \min\limits_{i<u<j} M_{i+1,u} + M^1_{u+1,j-1} + a \end{array} \right\}$$

$$M_{i,j} = \min \left\{ \begin{array}{l} \min\limits_{i<u<j} (u-i+1)c + C_{u+1,j} + b, \\ \min\limits_{i<u<j} M_{i,u} + C_{u+1,j} + b, \\ M_{i,j-1} + c \end{array} \right\}$$

$$M^1_{i,j} = \min \left\{ \begin{array}{l} M^1_{i,j-1} + c, \\ C_{i,j} + b \end{array} \right\}$$

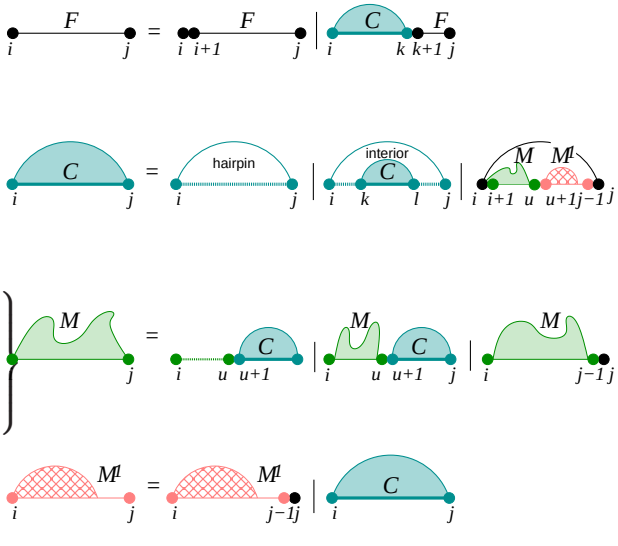

Figure 1.2: Loop decomposition to predict the RNA secondary structure. The picture and recursions are taken from Hofacker and Stadler (2007). The recursion have first been formulated by Zuker and Stiegler (1981). Arcs represent base pairs and dots unpaired regions. Details see text.

The minimum free energy of a structure is the sum over the constituting loop energies. The loop energies depend mainly on:

- the size of the loop,
- the closing base pair,
- the sequence of unpaired bases
- and for multiloops the number and type of its components.

The actual loop energies for hairpin-loops $\mathcal{H}$ and interior loops $\mathcal{I}$ are taken from experimental measurements (Mathews et al., 1999).
In nature each RNA is capable of folding into a large amount of different structures with similar stability. Within a cell, the differently folded RNAs are in equilibrium. Biological parameters like temperature or interacting proteins drive the equilibrium to a certain dominating structure. Thus state-of-the-art structure prediction programs do not only search for the one optimal structure but for all structures within a certain energy range.

RNA-RNA interaction In general RNA-RNA interaction is computed similar to RNA folding by free energy minimization based on thermodynamic modeling. The full problem is disassembled into smaller problems. Meaning the optimal and sub-optimal solution of the complete problem are gained by combination of small structural subunits with tabulated energy values by dynamic programming techniques.

1.2 Next Generation Sequencing

Complete sequencing of the total human genome took an international research team more than ten years (1990-2004) and about US$3 billion (International Human Genome Sequencing Consortium, 2004). Since then great strides in sequencing technologies have been made that massively increase speed and decrease costs. So called next generation sequencing (NGS) or also high-throughput sequencing (HTP) can now sequence human genomes in days and claimed to reach the milestone of cost of US$1000[1]. In the classical Sanger sequencing method (Sanger and Coulson, 1975), fluorescent labeled dideoxinuclosidetriphosphates (ddNTPs) cause a termination of the DNA chain during the polymerase chain reaction (PCR) Resulting DNA fragments of different length are separated through gel electrophoresis and the terminal base labels are measured. In contrast, for the new technologies sequencing libraries are prepared in a cell free system, up to millions of immobilized fragments are sequenced in parallel without the need for electrophoresis (van Dijk et al., 2014). Different technologies emerged that solve sample preparation and signal measurement through different approaches. The main technologies are: SOLID, Illumina, 454, each having specific advantages and drawbacks that need consideration in the subsequent analysis steps. Detailed reviews on next

[1] `https://www.veritasgenetics.com/documents/VG-launches-999-whole-genome.pdf`

generation sequencing can be found in Goldman and Domschke (2014); Goodwin et al. (2016); van Dijk et al. (2014).

The modern technologies allow to capture the RNA content directly from cells or tissues, which enables the identification of the actively transcribed parts of the DNA, revealing an astonishing complexity of genome architecture (Goodwin et al., 2016). Next generation sequencing has dramatically expanded the rate at which ncRNAs are discovered (Gardner et al., 2015). It is also possible to select certain subsets of genetic material, e.g. DNA associated with proteins through chromatin immunoprecipitation (ChIP-seq), RNA associated with proteins (CLIP) (see next Section), or RNA-DNA interactions (CHART, CHiRP), and others.

CLIP-seq

As experimental data from CLIP experiments (specifically iCLIP) form the basis for analysis in Chapter 7 it is more elaborated in following. To determine the RNAs that directly interact with proteins cross-linking and immunoprecipitation (CLIP) experiments have been developed (Ule et al., 2003)). To capture the actual interacting partners, a cell or tissue is UV irradiated thus covalently cross-linking RNAs to the proteins they contact. Afterwards the cell is lysed and the RNA binding protein (RNP) of interest is pulled out through immunoprecipitation, thus purifying the RNA-protein complexes. Through digestion of the protein only the RNA molecules remain. They are amplified and sequenced. In combination with next generation sequencing technologies this enables a genome wide search for RNA binding protein (RBP) interactors. Different CLIP technologies exist, the most prominent are HITS-CLIP, PAR-CLIP and iCLIP. Each version has its advantages, disadvantages and technical challenges, making them more or less suitable for different research scenarios. The iCLIP protocol is described in greater detail in following, for other methods it is referred to current literature (Wang et al., 2015).

iCLIP The method iCLIP (individual-nucleotide resolution UV-cross linking and immunoprecipitation) was introduced 2010 by König et al. (2010) (Schematic representation in Figure 1.3). The specialty of this CLIP protocol is that it takes advantage of the alleged problem that during reverse transcription the RNA polymerase can not cross the polypeptide rest, left at the RNA after digestion of the protein. In turn the synthesis exactly ends at the cross-linking position. A cleavable primer and the

barcode sequence are attached at the 3'-end and the transcript is circularized. Then the RNA circle is cleaved within the attached primer, leaving one part of the primer at the 3'-end and the other part with barcode at the 5'-end. After PCR, sequencing and adapter removing the first nucleotide of each transcripts is the position of cross-linking. That trick not only delivers the associated transcripts but also exactly defines the cross-linking site, which should correspond to the binding site between RNA and protein.

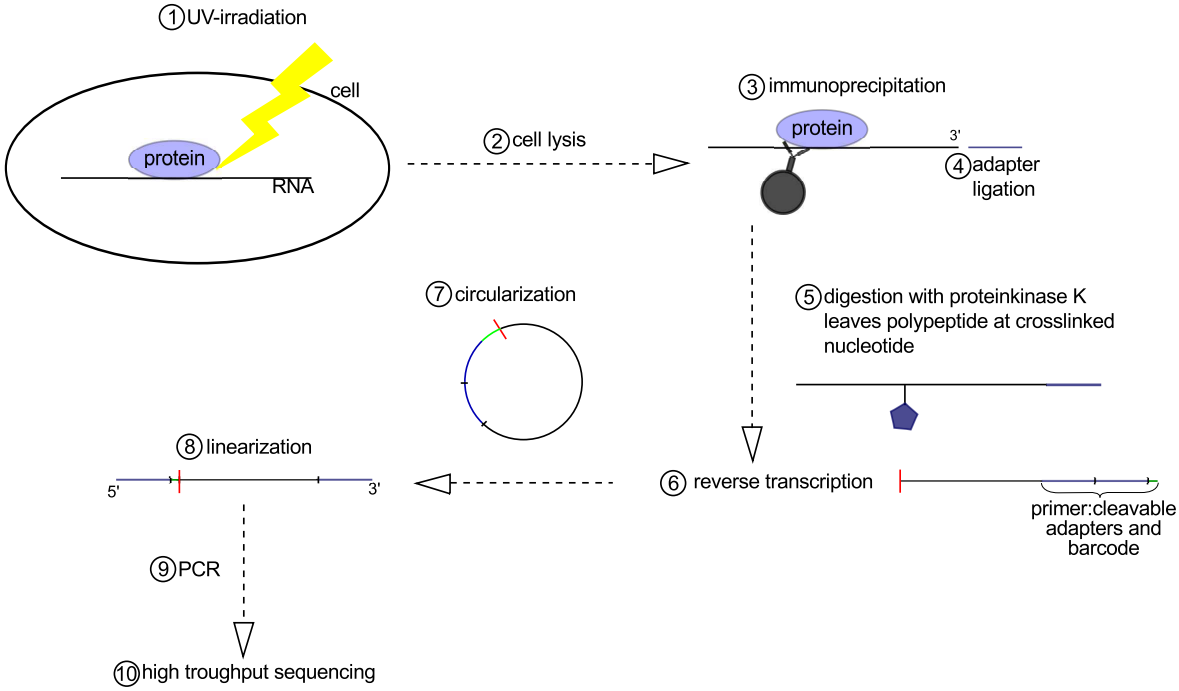

Figure 1.3: Schematic representation of the iCLIP workflow. 1. The cell is UV-radiated, establishing covalent bonds between proteins that contact RNAs. 2 Then the cell is lysed and 3. the protein of interest is selected through immunoprecipitation. 4. To the 3'-end of the RNA that is bound to the protein an adapter is ligated. 5. Then the protein is digested by protein kinase K. This step leaves the polypepdide rest bound to the RNA transcript. 6. During reverse transcription the RNA-polymerases can not pass this rest, with the effect that all transcripts end exactly at the cross-linked site. In the reverse transcription reaction a cleavable primer is added. This contains a adapter region, the cleavage site, another adapter region and the barcode. 7. Afterwards the amplified RNAs are first circularized and 8. then linearized. During linearization the cleavage is preformed at the designated position in the primer. The resulting RNA transcript has an adapter at both ends. At the 3'-end it is followed by the barcode, immediately adjacent to the cross-linked nucleotide. After 9. PCR the RNAs are 10. sequenced in high-throughput. Figure adapted from Huppertz et al. (2014); König et al. (2011)

Bioinformatic Analysis

The analysis of these huge masses of sequenced RNA snippets, called reads need efficient bioinformatic approaches. The main analyzing steps include:

1. **read preparation** - to avoid noise and bias several pre-processing steps are performed.
   - PCR artifacts are removed
   - adapters and barcodes are cut, and
   - reads of low quality and/or containing unrecognized nucleotides are eventually sorted out.

2. **read mapping** - the genomic origin of a read is determined by finding the best alignment to the genome (e.g. with `segemehl` (Otto et al., 2014; Hoffmann et al., 2009)). Main challenges of this step include:
   - immense data masses
   - mismatches through sequencing errors or genomic variance can obscure the origin,
   - multiple mappings can occur e.g. to genes that are encoded in high copy number,
   - split read mapping that results form reads that span intron-exon junctions or can result from non-canonical transcripts (Hoffmann et al., 2014).

3. **gene annotation** - identify the known and novel RNA transcripts.
   - intersection with available gene annotations, e.g. using BED-tools (Quinlan, 2014)
   - identification of conserved and structured regions with `RNAz` (Gruber et al., 2010)
   - identification of characteristic read patterns (Langenberger et al., 2012a)
   - application of customized filters for specific ncRNA classes e.g. snoRNAs (Kishore et al., 2013; Machyna et al., 2014; Jorjani et al., 2016).

1.3 Databases

The sequenced genomes, transcriptomes and according gene annotations, which are the ingredients of bioinformatic genome research are massive data collections. To be useful they need to be organized in consistent and comprehensive manner and should possibly have open access, so that researchers around the world can access them. Not far from seek, many database exist that provide biological data. They range from rather general data collections to focus on a specific issue and from optimized for automated data retrieval to merely displaying collected data. In the following a few databases (Table 1.2) that are especially important for this work are briefly introduced. Several other databases, which served as data source for a specific information are mentioned at the according passage in the text.

Table 1.2: Overview of main biological databases used for this work.

Name	content
ENSEMBL	Genome Assemblies, Gene Annotation, Genome Browser
UCSC	Genome Assemblies, Gene Annotation, Genome Browser
RFAM	RNA family alignments, covariance models
`snoRNA-LBME-db`	human snoRNAs
snOPY	orthologic snoRNAs

General and popular databases for genome data are ENSEMBL[2] and UCSC[3]. These large biological databases provide annotated genome assemblies for basically all sequenced organisms. As each genome assembly represents a separate database entity, `Ensembl` and `UCSC` are actually collections of databases. In `UCSC` gene annotations are provided as tracks which represent database tables. The tracks hold diverse information, beside gene annotations, which include also detailed gene structure (UTRs, intron, exon, ORFs), other tracks provide e.g. conservation, expression, repeats, SNPs. The data can be visualized in the genome browser, which is a central part of many genome assembly databases. This enables visual inspection of a gene of interest, its structure, its surrounding and even its conservation and expression. Furthermore, the web-interface provides basic analysis tools (like `BLAST` and `BLAT`), enabling basic genome research tasks. However, the complexity and bulk of genome data contained in the databases also leads to several uncertainties in the tracks. A few examples in

[2] `http://www.ensembl.org/index.html`
[3] `https://genome.ucsc.edu/`

human from the ncRNA field are small nuclear RNAs that are included in the Repeat track, which needs to be taken into consideration e.g. when all loci that overlap Repeats should be removed from a list, while snRNAs should be kept. Indeed, the RNAs are abundant in large copy numbers, yet the classification as repeat is not intuitive. In case of snoRNAs the annotations are scattered in several track, mostly the known genes and snoRNA track. Unfortunately, these are not congruent as both tracks include snoRNA genes that are not necessarily covered in the according other track (e.g. SNORD65). Another prevalent problem is ambiguous naming of genes. Nevertheless, both databases are indispensable, extensive sources for genome research.

A valuable resource for non-coding RNA families is RFAM[4] (Nawrocki et al., 2015). For each RNA family the database provides reliable alignments of representatives that include also the consensus secondary structure, and covariance models (CMs) that simultaneously models RNA sequence and structure. Elaborate analysis software and data formats have been developed in line with the database, which ensures optimal use of the data. The CMs can immediately be used for a probabilistic homology search with `Infernal`(Nawrocki and Eddy, 2013). The alignment format established by RFAM is called STOCKHOLM (.stk-files). Even an alignment editing extension for the emacs editor, called RALEE (RNA alignment editor in emacs) (Griffiths-Jones, 2005) is available. The great advantage of the stk-format is the possibility to add annotation lines to the alignment. Lines starting with #=GC mark column annotations. The classical example is annotation of the secondary structure consensus (#=GC SS_cons). Although this seems trivial, it is an valuable and outstanding possibility to add important information about sequence features to snoRNA families in a flexible manner.

A snoRNA specific database is the `snoRNA-LBME-db`[5], which provides snoRNA sequences, target interactions and gathered information about each snoRNA in human. Unfortunately, it is not maintained since 2006 and obviously does not contain the manifold findings of the recent years. It has also not been designed for computational querying, e.g. holding details of snoRNA-target interactions only recognizable for humans in .gif-files. For yeast a separate snoRNA database[6] exists. A recent approach to built a snoRNA orthological gene database (snOPY)[7] (Yoshihama et al., 2013) does not yet overcome the problem and remains incomplete and needs further curation.

[4] `www.rfam.xfam.org`
[5] `https://www-snorna.biotoul.fr/`
[6] `http://people.biochem.umass.edu/fournierlab/snornadb/main.php`
[7] `http://snoopy.med.miyazaki-u.ac.jp`

CHAPTER 2

Introduction to Small nucleolar RNAs

This chapter provides a basic biological background, first outlining non-coding RNAs (ncRNAs) in general, and then phasing deeper into the topic of small nucleolar RNAs (snoRNAs). A summary of the current knowledge of genomic organization and characteristics of snoRNAs prepares for ongoing chapters that handle certain aspects of snoRNAs in more detail. In particular details about the current state of knowledge on snoRNA function is outlined, including the interactions with ribosomal and spliceosomal RNAs but also the emerging involvement in alternative tasks.

2.1 Gene Expression and Non-coding RNAs

Gene Expression Every living cell comprises DNA. That is a macromolecule that holds the genetic information in nucleotide permutations. An information entity is called a gene. Genes can code for proteins that realize basic life supporting tasks in the cells, like metabolism, signaling, and transport. The Central Dogma that has had emerged from bacterial (so prokaryotic) genetic research in the 1950s and 60s said that a stretch of "DNA makes RNA makes protein". The actual situation turned out to be much more complex. A stretch of DNA can give rise to several and changing RNAs, via splicing and alternative splicing. These RNAs are not necessarily a mere intermediate on the way of getting a protein. Indeed already at RNA level, fundamental regulatory functions are performed by so called non-coding RNAs (Darnell, 2011, 137ff.).

In the cells of eukaryotes the double stranded DNA is spooled around histones which are organized in complexes, forming chromatin (Figure 2.1). Chromatin is further coiled to fibers of $10\mu m$ and $30\mu m$. The most condensed packaging is achieved during the metaphase when the chromatin is visible for electron microscopes in the typical chromosome shape. Inactive parts of the chromatin can have a similar high condensation state during interphase, called heterochromatin. Only fractions of the DNA, that are located in more loosely packed regions of the chromatin, called euchromatin, are accessible for transcription factors (TF) and thus are the regions of active gene expression (Cooper, 2000).

After binding of TFs to some regulatory elements (e.g. promoters, enhancers, silencer) on the DNA, the transcription machinery unwinds the DNA double strand and RNA polymerase molecules produce complementary RNA copies of the coding strand. This process is called transcription (Figure 2.1.

In eukaryotes the genetic information is usually not contiguously encoded on the DNA strand. Instead exons are surrounded by untranslated regions (UTRs) and interrupted by introns. In the progress of splicing introns are excised from the primary transcripts by RNA-protein complexes called spliceosomes.Two types of introns exist, the major and the minor introns. The according splicing machinery, respectively major and the minor, comprises five snRNAs each. While U5 is shared by both the other components are distinct: U1, U2, U4, and U6; or U11, U12, U4atac, and U6atac (for details on snRNA and spliceosomes see Reviews e.g. (Papasaikas and Valcárcel, 2016; Wahl et al., 2009; Chen and Moore, 2014)). The products of the splicing process are an exon-only

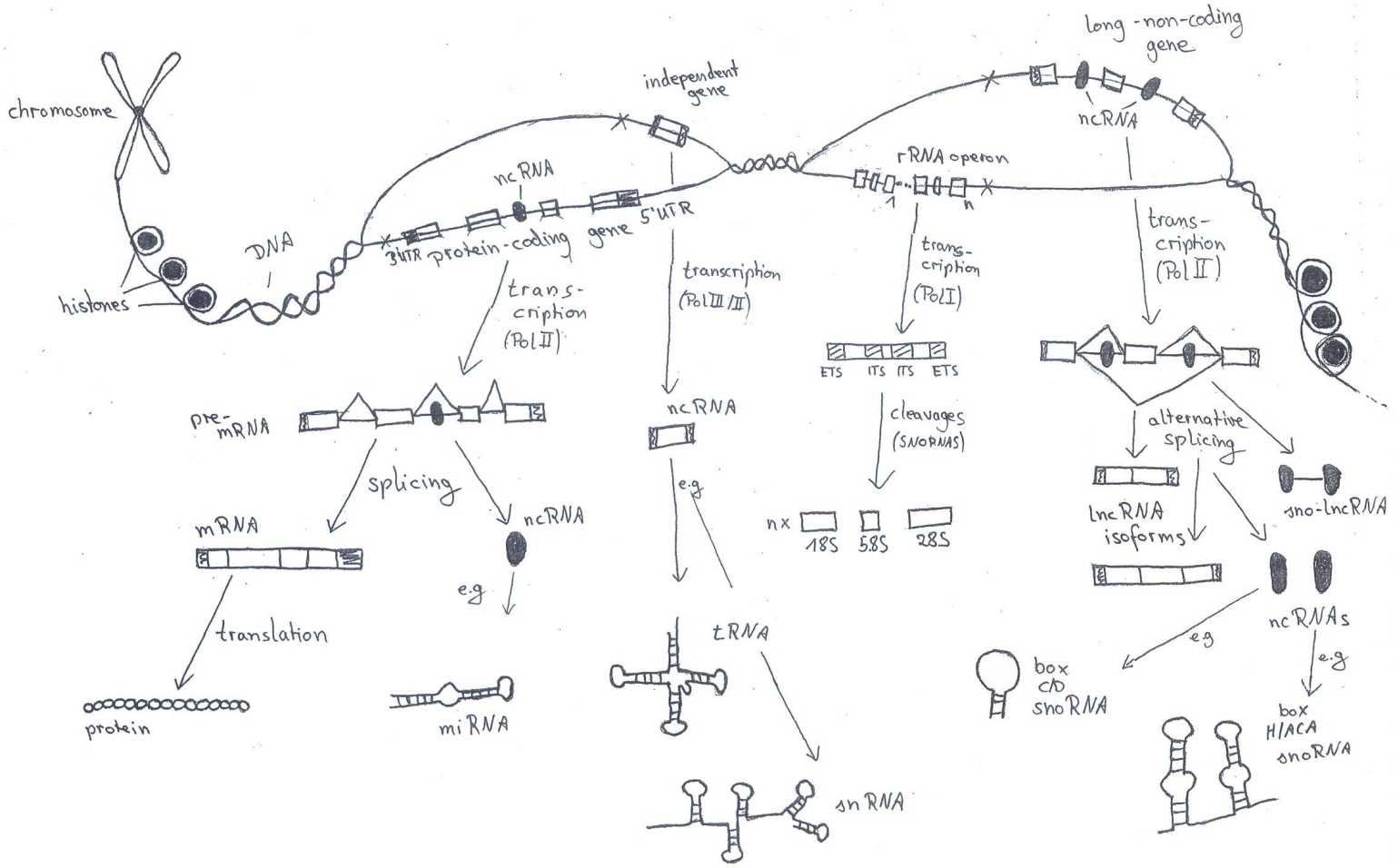

Figure 2.1: Schematic representation of gene expression. DNA is condensed in chromosomes, and wound around histones. Genes can be protein-coding or non-protein coding, reside in introns of other genes, or form independent transcription units. Transcription is catalyzed by different types of RNA polymerase (PolI/II/II) and produces immature RNAs. Depending on the type these are cleaved, capped and/or spliced. During splicing introns are excised, and exons merged. Thus, giving rise to several classes of small ncRNAs, rRNAs, mRNA or lncRNAs. The latter two might also be alternatively spliced revealing different isoforms. Protein coding mRNAs are translated into amino acid chains at ribosomes under involvement of tRNAs. The elements of the figure do not scale. The figure is inspired by (Morris and Mattick, 2014) and (Lindemeyer, 2006).

RNA transcript and several debranched intronic RNA stretches. The exonic RNAs can be coding for a protein, but also long non-coding RNA (lncRNA) transcripts, with defined regulatory functions exist. Many of the spliced introns are simply degraded, but several also give birth to functional small non-coding RNAs (ncRNAs).

Those RNAs, whose exon portion code for proteins, the messenger RNAs (mRNAs), accumulate at ribosomes. Like spliceosomes, these are large RNA-protein complexes. At the ribosomes the RNA sequence is translated into an amino acid chain. Each base triplet of the RNA template encodes for one amino acid. The codons are recognized by transfer RNAs (tRNAs), that supply the according amino acids, which is bound to the emerging protein (Figure 2.1).

The initial assumption of protein-coding regions being the only informative parts of the genome and all the rest being unused junk turned out to be fundamentally wrong in the last three decades (Hubé and Francastel, 2015). The current model assumes that only 2% of the genome encodes for proteins, while up to 80% is indeed functional at RNA level (ENCODE Project Consortium, 2012).

Regulatory non-coding RNAs In contrast to protein-coding genes no subsequent translation into an amino acid chain is needed for ncRNA genes. As such, they are energetically cheaper, faster adapting and more flexible (Eddy, 2001). Based on their length non-coding RNAs are classified as small or long non-coding RNAs. The RNA (and DNA) inherent mechanism of base pairing is fundamental and adequate for secondary structure formation of the molecule, and also constitutes the underlying mechanism to interact with other RNAs. The ncRNAs are functional in ribonucleoprotein complexes (RNPs), in which mostly the assembled proteins catalyze chemical reactions at sites which are defined by the ncRNA mostly through base pairing with target RNAs (Cech and Steitz, 2014). A famous examples in this sense are micro RNAs (miRNAs) that are part of an RNA induced silencing complex (RISC). This complex binds to target sequences in mRNAs (mostly in the 3'UTR) and promotes degradation of the mRNA and/or inhibition of translation (Bartel, 2004). Also small nucleolar RNAs (snoRNAs) base pair with target RNAs, here the binding defines single nucleotides, that are then chemically modified by the snoRNA associated proteins (Bachellerie et al., 2002). The induced modifications in turn have impact on the structure and stability of these RNAs (Decatur and Fournier, 2002). These few examples already show that ncRNAs represent a layer of regulation, and might be the main source of our complex characteristics and genetic variations, both within and between species (Mattick and Makunin, 2006).

Genomic organization of ncRNAs encompasses individually transcribed ncRNAs comprising their own promoters, while others are co-transcribed with their so-called host genes. In the latter scenario they reside in introns from which they are processed after splicing. Also some ncRNA genes have an intron-exon structure. Often small ncRNAs are encoded in introns of these so called long non-coding RNAs (lncRNA).
New classes of ncRNAs, new functions for known ncRNA classes, and new genomic organizations are still detected.

2.2 Introduction to SnoRNAs

SnoRNAs are an ancient class of small non-coding RNAs that are present in five of the six kingdoms of life: *Archaea*, *Plants*, *Fungi*, *Protists* and *Animalis*. Missing in *Bacteria* their invention dates back 2-3 billion years, previous to the split of archaea and eukaryotes.
The name **small nucleolar RNA** reflects their main nucleolar localization in the cells. The nucleolus is a sub-compartment of the nucleus, where ribosome assembly takes place (Matera et al., 2007).
During evolution the genomic encoding of snoRNAs changed from independent transcripts to a predominantly intronic localization (Tycowski et al., 2004). In *C. elegans* there is still evidence for promoter elements in the vicinity of many snoRNA genes (Deng et al., 2006). In plants the predominant encoding are independently transcribed clusters (Brown et al., 2008). Apart from some prominent exceptions like SNORD3 and SNORD118, the snoRNAs in *Metazoa* are encoded in introns of longer host genes and lack their own promoter elements (Tyc and Steitz, 1989; Dieci et al., 2009). Instead, they are co-transcribed with their host genes and released from the debranched introns by exonucleolytic cleavages (Bachellerie et al., 2002). In human, the largest fraction of snoRNAs is encoded within genes that code for proteins involved in ribosome biogenesis proteins (Filipowicz and Pogacić, 2002). Overall, the snoRNA host genes (SNHG) are very diverse, ranging from house keeping genes (among them the ribosomal proteins) to long non-coding genes of unknown function (Terns and Terns, 2002). For some of the lncRNA host genes it is even speculated that their only function is the hosting of snoRNAs (e.g. SNHG1) (Bortolin and Kiss, 1998). Several snoRNAs encoded in imprinted loci and in addition are exclusively expressed in neurons (and putatively restricted to brain)(Cavaillé et al., 2000). Two such snoRNA families have a very special

genomic organization . SNORD115 and SNORD116, are repeatedly encoded in arrays of 48 and 35 highly similar copies in the imprinted SNURF-SNRPN locus. Correct expression of both is suspected to have substantial impact in Prader-Willi syndrome (PWS) (Bortolin-Cavaillé and Cavaillé, 2012).

The main function of snoRNAs is to guide protein complexes to specific positions in ribosomal and spliceosomal RNA molecules. Those snoRNPs introduce chemical modifications of single residues in the target RNAs (Terns and Terns, 2002). This is described in detail in Section 2.3.

There are two main types of snoRNAs: box H/ACA and box C/D snoRNAs. They have different secondary structures, characteristic sequence elements, and catalyze different types of modifications in targeted RNAs. While the first guide pseudouridylation, the latter direct 2'-O-methylation of any kind of nucleotide. Both modifications are introduced concurrently or immediately after transcription of the rRNA operon, prior to cleavage of the 45S rRNA (primary transcription product of the rRNA operons). These modifications are essential for maturation of the rRNAs. The special snoRNAs SNORD3, SNORD118, SNORD14, SNORA73A/B, and SNORD22 direct cleavage steps of the 45S rRNA rather than chemical modifications (Atzorn et al., 2004).

Small Cajal body-specific RNAs (scaRNAs) constitute a third type of snoRNA. This subset accumulates in the Cajal Bodies (CBs), sub-organelles that are associated to the nucleolus and guide the modification of small nuclear RNAs (snRNAs) in the Cajal body of eukaryotic cells (Darzacq et al., 2002). Some scaRNAs combine features of both snoRNA classes, see Marz et al. (2011) for more details on these atypical RNAs.

Box C/D snoRNAs

The name of box C/D snoRNAs is derived from the characteristic sequence motifs within their sequence. Close to the 5'-, and 3'-end the C box (RTGATGA) and D box (CTGA) are encoded (Figure 2.2). In the medial sequence stretch, another pair of boxes with the same consensus sequence is found. Yet, the D' box and the C' box have higher sequence variation. The 3' and 5' end of the box C/D snoRNAs form a short terminal stem, the inner fraction of the sequence remains unpaired. The bases of the C- and D-box are the main constituents of a kink-turn that is obligatory for core protein assembly. Recognition of the target is achieved through antisense elements

(ASE) in the unpaired region upstream of the D and the D'-box. Here the snoRNA and the target RNA can form duplexes of 7-20nt length that have no bulges, and may only be interrupted by a few mismatches (Chen et al., 2007). The nucleotide base-paired to the 5th base upstream of the box is modified. For box C/D snoRNAs the modification is 2'-O-methylation i.e. an extra methyl group is added to the ribose ring of the nucleotide.

C/D snoRNPs The mature box C/D snoRNA ribonucleoprotein complex (C/D snoRNP) contains four core proteins, which assemble co-transcriptionally (Reichow et al., 2007). In vertebrates the protein 15.5kDA assembles with the C- and D-boxes and chaperons the k-turn formation. Subsequently, Nop56 and Nop58 stabilize this structure and recruit fibrillarin. The latter is the methylase, the catalytic component, accomplishing 2'-O-methylation in the targeted RNA molecule (Matera et al., 2007). Correct C/D snoRNP assembly involves several other proteins and protein complexes (Rothé et al., 2014). It is currently discussed, whether the functional RNP is monomeric, or rather a dimer combining two box C/D snoRNAs and two sets of core proteins in a single complex (Figure 2.2D) (Lapinaite et al., 2013). The latter case has been observed in archaea (Lapinaite et al., 2013; Lui and Lowe, 2013). In either case the assembly process starts in the nucleolus, and proceeds in the Cajal Body. Afterwards the complex has to be forwarded to their sites of function (Matera et al., 2007).

Box H/ACA snoRNAs

The name box H/ACA snoRNA also reflects the sequence motifs that are encoded in the RNA sequence. The whole sequence folds into a typical hairpin-hinge-hairpin-tail structure. Therein the H-box (ANANNA) is encoded in the unpaired hinge region between the two hairpins, and the ACA-box is located in the single stranded tail sequence stretch at the 3'-end of the molecule. Each hairpin comprises a nearly symmetric interior loop, called the pseudouridylation pocket (Figure 2.3A). Within the interior loops a bipartite ASE is located 3' and 5' of the upper stem (Figure 2.3). There sno- and target- RNA form two small duplexes, with a total length of 8-20 nts. An uridine of the target RNA is anchored, unpaired, underneath the upper stem, central to the two parts of the ASE (Ganot et al., 1997; Ni et al., 1997). This U is isomerized to pseudouridine (Ψ) (Bachellerie et al., 2002). The downstream distance between the anchored U,

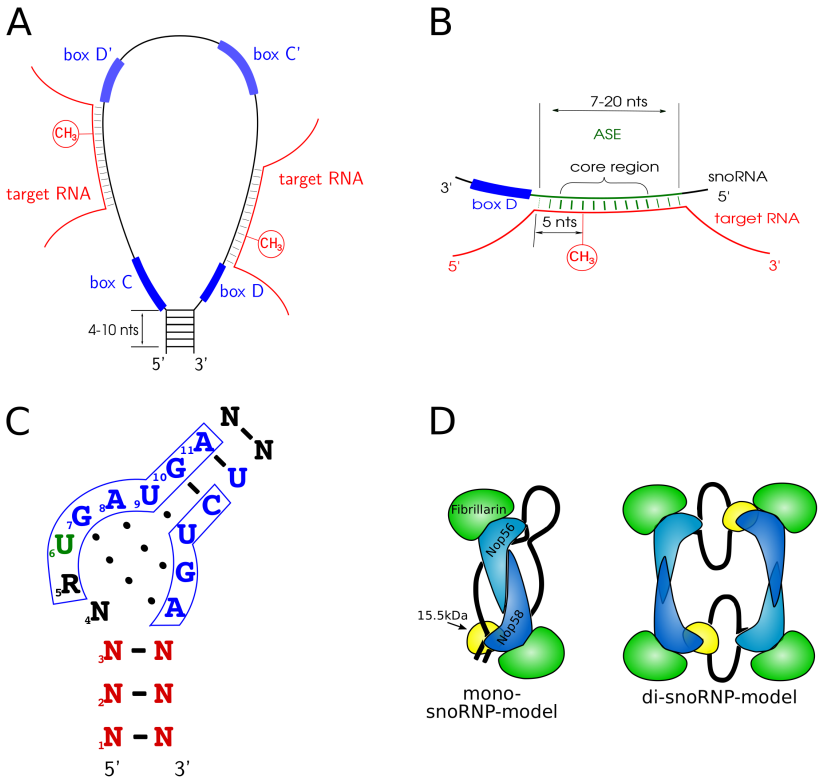

*Figure 2.2: Box C/D snoRNAs. A: In box C/D snoRNAs a short terminal helix encloses a large unstructured loop. Box motifs C and D and variant copies C' and D' are located within the loop. B: Target RNAs form small duplexes with the ASEs adjacent to D- and D'-box, resp. The 'core region' of the interaction ranges from 3rd to 11th position. The methyl-group is added to the nucleotide bound five nt upstream of the D/D'-box. C: The kink-turn in box C/D snoRNAs involves nucleotides of, and adjacent to the C- and D-box (in blue boxes). The respective frequencies for nucleotides and base pairs in the motif are: 1.Watson-Crick-pair(**WC**):82.8, 2.**WC**:81.53, 3.**WC**:85.14, 4.Pyrimidine(**Y**):68.85 & Purine(**R**):31.15, 5.**R**:90.98, 6.**U**:97.54, 7.**GA**:100, 8.**GA**:100, 9.**UU**:96.72, 10.**WC**:100, 11.**WC**:68.85 (adapted from Bartschat et al. (2014)). D: The core proteins of the eukaryotic box CD snoRNPs are 15.5kDA, 2xFibrillarin, Nop56, and Nop58. It is still unclear whether the complexes consist of one or two snoRNAs and accordingly four or eight core proteins (Figure D adapted from Lui and Lowe (2013)).*

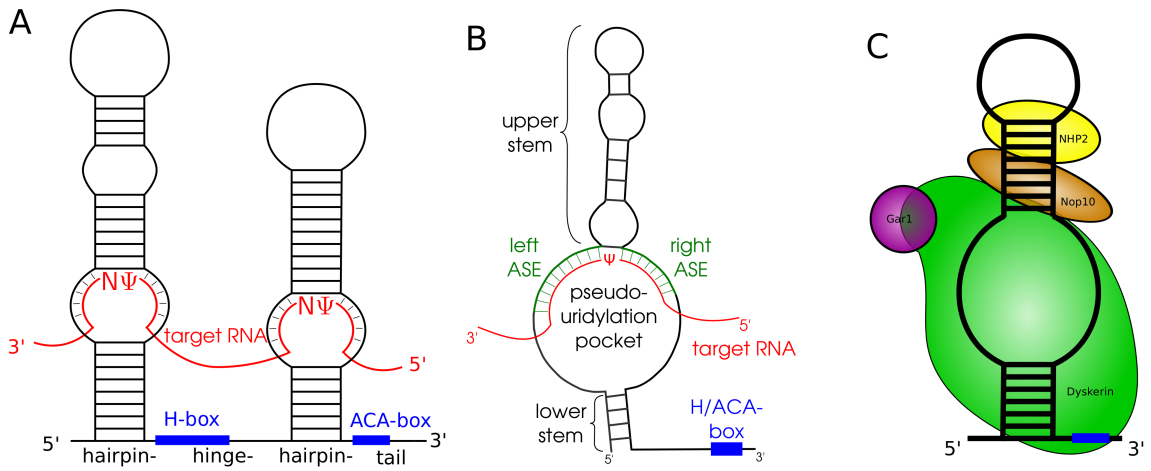

Figure 2.3: A: Box H/ACA snoRNAs have a hairpin-hinge-hairpin-tail architecture, with box H located in the hinge region, and box ACA in the tail. Target RNAs bind into the pseudouridylation pockets forming small duplexes with a bipartite ASE. B. The right and left ASE base pair with the target RNA, allowing no bulges and only few mismatches. The Uridine that is pseudouridylated is anchored unpaired underneath the upper stem. The distance between the introduced Ψ and the H-/ACA- box motif (blue box) is 14-16 nts. C: Each of the hairpin forms an H/ACA-snoRNP with proteins NHP2, Nop10, Gar1 and the pseudouridine synthase dyskerin (Figure C adapted from Lui and Lowe (2013)) .

which is equivalent to the closing pair to the upper stem to the following box motif (H or ACA resp.) encompasses 14-16 nucleotides. In archaea this distance restriction has been shown to result from the necessity of a certain conformation flexibility between the pseudouridylation pocket and the lower stem (Toffano-Nioche et al., 2015). Like in the case of the box C/D RNA duplexes, only a few mismatches but no bulges are tolerated.

H/ACA snoRNPs Box H/ACA snoRNAs also assemble with four core proteins (Figure 2.3C). Like in the case of box C/D snoRNAs the proteins bind the nascent transcript previous to splicing (Reichow et al., 2007; Richard et al., 2003). NHP2 directly binds the snoRNA in the apical loop, inducing a major bend in the RNA. NHP2 interacts with Nop10, which in turn interacts with the pseudouridine synthase dyskerin. Dyskerin is structurally similar to TruB, the bacterial pseudouridine synthase (Lui and Lowe, 2013). Gar1 is in contact with dyskerin and is involved in binding and release of the target RNA (Matera et al., 2007). An independent set of proteins seems to interact with each of the hairpins separately (Matera et al., 2007).

Small Cajal Body RNAs - ScaRNAs

Small Cajal Body RNA sequences are structured like ordinary box C/D snoRNAs or box H/ACA snoRNAs (with additional localization motifs) or combine structural elements of both. The latter results in tandem box C/D snoRNAs, tandem box H/ACA snoRNAs or hybrids of both (Figure 2.4). In the hybrid case, a box H/ACA domain is incorporated into the loop region of a surrounding box C/D domain. Some of these unconventional structures are more thoroughly surveyed in Chapter 5. Their defining and common feature is the Cajal Body localization, achieved by additional sequence motifs. For box H/ACA snoRNAs a CAB-box (ugAG) has been detected in the apical loop of one or both hairpins (Richard et al., 2003). A GT-repeat region cause accumulation of box C/D snoRNAs to Cajal Bodies (Kishore et al., 2013). ScaRNAs function as guide RNAs just like ordinary snoRNAs (Section 2.3). However, their main targets are pol-II (U1, U2, U4, U5, U11, U12, U4atac) and pol-III (U6, U6atac) transcribed small nuclear RNAs (snRNAs) (Darzacq et al., 2002).

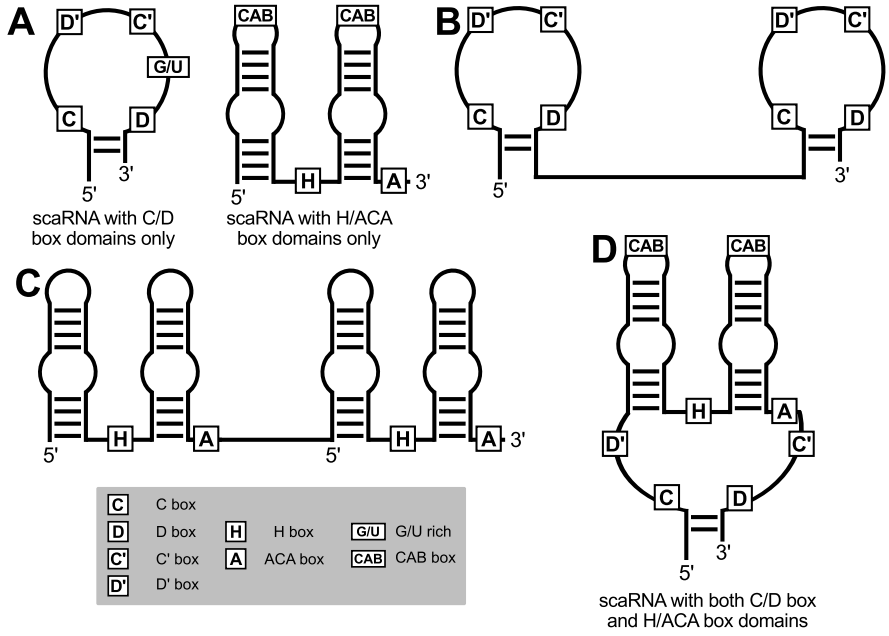

Figure 2.4: ScaRNAs occur in different shapes. A: Common box C/D snoRNA or box H/ACA snoRNA structure with additional Cajal Body localization motifs. B: Tandem box C/D scaRNA, combining two box C/D snoRNA domains. C: Tandem box H/ACA scaRNA, combining two box H/ACA snoRNA domains. D: Hybrid scaRNA combining a box H/ACA domain and a box C/D domain in a single transcript. Figure from Jorjani et al. (2016)

2.3 Common SnoRNA Function

The predominant function of snoRNAs can be summarized as maturation of ribosomal RNA (rRNA) and small nuclear RNAs (snRNAs).

Ribosomal RNAs The ribosomes, are the macromolecules in the cells where the proteins are synthesized from messenger RNA (mRNA) templates. Beside many involved protein components, four RNAs are part of the two ribosome subunits. One, 5S ribosomal RNA (rRNA) is transcribed by RNA-PolIII, while the others are synthesized from a single rRNA operon by RNA-PolI. 18S, 5.8S, and 28S are separated through so called internal transcribed spacers (ITS) and framed by two external transcribed spacers (ETS) (Figure 2.5). A cascade of several endo- and exonucleolytic cleavage steps involving many trans-acting protein and snoRNA factors removes these transcribed spacers to produce the rRNAs. The interaction with snoRNAs SNORD3, SNORD14, SNORD22, SNORA63, SNORA73, and SNORD118 at specific sites is obligatory for correct cleavage steps (Henras et al., 2015).

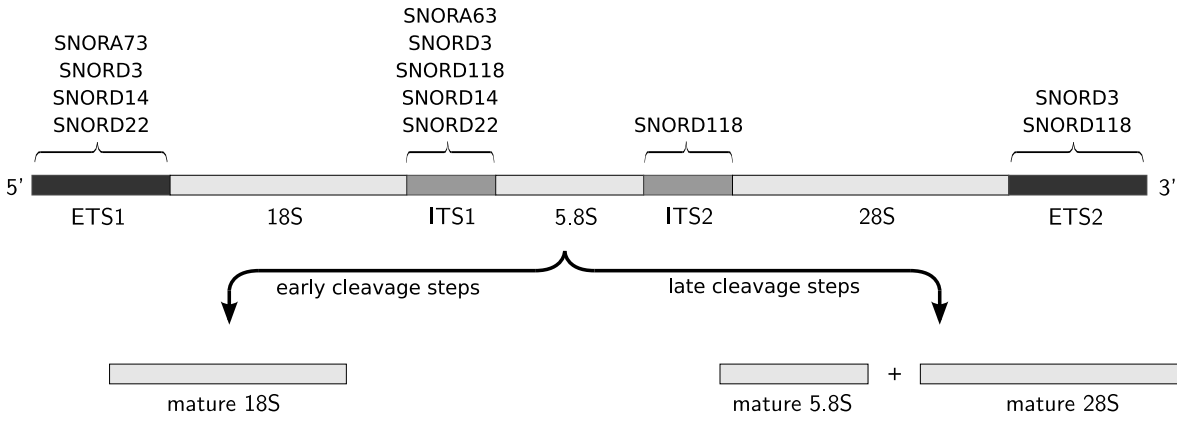

Figure 2.5: Cleavage steps of the primary ribosomal RNA transcript and associated snoRNAs. Early maturation steps involve removal of ETS1, and cleavage of ITS1 to produce mature 18S rRNA, the RNA component of SSU. This requires snoRNAs SNORD3, SNORD14, SNORD22, SNORA73. The later cleavage steps produce mature 5.8S and 28S rRNA, the RNA components of LSU. Cleavages in ITS1, ITS2, and ETS2 require snoRNAs SNORD3, and SNORD118, SNORA63. Figure adapted from Bartschat (2011), Henras et al. (2015), and Maxwell and Fournier (1995).

Single Nucleotide Modifications More in the foreground of snoRNA research are the single nucleotide modifications within the rRNA sequences. They are introduced

in the nucleolus by the snoRNA machinery. For a proper function of ribosomes numerous 2'-O-methylations (currently 104 sites identified) and pseudouridylations (currently 96 sites identified) of single nucleotides in the RNA components of the ribosome are needed. While Pseudouridine (Ψ) is introduced by several pseudouridine synthase enzymes (Pus) in bacteria, the task has been taken over by snoRNPs in archaea, protists, fungi, plants, and animals. Still Pus homologous that introduce Ψs in e.g. tRNAs (Spenkuch et al., 2014) are present in all organisms.

Pseudouridines The immediate effect of pseudouridines in the RNA sequence is an increased backbone rigidity through an additional hydrogen bond and improved base stacking (Spenkuch et al., 2014). Thus the modified bases stabilize secondary structures and enables a certain degree of flexible conformation changes. Still only four Ψs during zebra fish development have been detected to have an individual distinct and finally lethal effect through brain malformation, organ, and/or body maldevelopment (Higa-Nakamine et al., 2012). Apart from that, prevention of a single pseudouridylation in rRNAs has no measurable effect alone and only a disruption of several pseudouridylations lead to decreased cell growth (King et al., 2003). As such the modified residues fine-tune and adapt the molecule structures also to dynamic environmental and developmental demands.

2'-O-methylation The effect of 2'-O-methylation of ribose is even less understood. On molecular level 2'-O-methylation is known to hinder alkaline degradation. Methylations at the 2-OH favors a special conformation of the ribose, block sugar edge interactions and change the hydration sphere around the oxygen (Helm, 2006). Thus, on a more general level it is thought to have impact on RNA structure. Recent high-throughput methods to identify methylation sites help to annotate methylations and subsequently understand their consequences (Birkedal et al., 2015; Krogh et al., 2016). Interestingly, the mentioned publication detected interdependence between distant modifications. Furthermore, analysis of modification kinetics discovered that in LSU of yeast a subset of late post-transcriptional modifications can be distinguished from earlier co-transcriptional ones. All 2'-O-methylations in the SSU are of the latter class.

Specialized Ribosomes We assume that snoRNAs play an important role in specialization of the targeted RNA molecules to changing environmental conditions. In this scenario the modified residues adapt the ribosomes to the current cell challenges. Observed flexible snoRNA expression profiles, showing tissue and developmental dependent differential expression support this idea on the snoRNA side. On the modification

side evidence for specialized ribosomes can be drawn from some mTOR pathway induced Ψ residues that exist in the 28S rRNA of hamster ovary cells (Courtes et al., 2014). Also the possible between observed early and late 2'-O-methylations in the LSU, could reflect basic and adaptive rRNA modifications (Birkedal et al., 2015; Sloan et al., 2015). The concept of specialized ribosomes is not new, although the focus has rather been drawn on the interacting protein components so far (Xue and Barna, 2012). A specialized ribosome would be capable to favor the translation of a certain group of proteins.

Evolution of Modifications Interestingly, the general modification patterns of rRNAs (Appendix Figure A.1 and A.2) is retained during evolution making it even possible to identify distantly homologous snoRNAs based on their function (e.g. between yeast and human) (Ofengand and Bakin, 1997). In the rRNA molecules modification hotspots can be found in the regions that are also deeply conserved on sequence/structure level. The vast majority is located in the peptidyl transferase center and the intersubunit bridges, both highly functional and highly conserved regions of the ribosomes (King et al., 2003). Nevertheless, apart from the evolutionarily old and conserved modifications, there are also hints for species specific ones. One documented example is the methylation of 28S-G3524 and 28S-C4004 (GENBANK X02995) detected in the rRNA of frog (*X. tropicalis*). The guiding snoRNAs NET1 and NET3 (non-eutherian specific) are specifically found in Xenopus and some other vertebrate animals but neither in human nor any other placental mammals. This is in accordance with the detection of an unmethylated nucleotide at the analogue 28S rRNA site in human (Makarova and Kramerov, 2009), indicating a loss of function and the related gene.

Small Nuclear RNAs However, snoRNAs do not only target rRNAs, but were also discovered to guide modifications in small nuclear RNAs (snRNAs). The modified snRNAs constitute the RNA components of the spliceosomal machinery, that excises introns from primary RNAs transcripts (Section 2.1). The snRNAs (major and minor) are modified at a variety of sites. Responsible is a the Cajal body specific subset of snoRNAs: the scaRNAs. Accordingly, with the Cajal Body accumulation of snRNAs, also the scaRNA molecules amass in these cell organelles. Surprisingly, it became evident that CB are not the site where snRNA modification is taking place, but snRNAs are also modified in the nucleolus. Yu et al. (2001) studied U2, the most extensively modified snRNA (ca. 10% of the nts undergo 2'-O-methylation or pseudouridylation) (Massenet et al., 1998) in *Xenopus* oocytes. Mutation of the U2s' Sm-binding site

inhibits accumulation to the nucleolus, which correlates with unmodified U2 snRNA. Introduction of nucleolar localization signals (C-box and D-box) into the Sm-mutant U2 RNA re-establishes the modifications (Yu et al., 2001). As in rRNAs the modifications within the snRNA sequences are conserved between species and are enriched in functional important regions, especially those of RNA-RNA contact (Karijolich and Yu, 2010) (Appendix Figure A.3). Interestingly, not only RNA-dependent, but also RNA-independent modifications occur. An example of such a site is Ψ at position U2-35 in yeast and U2-34 in human. This site is exclusively catalyzed by Pus7p in yeast, while it is also established by the SCARNA8-snoRNP in human (Karijolich and Yu, 2010). In the snRNAs of the minor spliceosome fewer modifications, but at equivalent positions have been detected (Massenet and Branlant, 1999). It is suggested that the lesser extend of modification is correlated to an overall higher conversation of the minor introns (Karijolich and Yu, 2010).

Other Target RNAs Recent studies detected snoRNA guided pseudouridylation and 2'-O-methylation in further RNAs. SNORA70 and SNORA31 comprise antisense elements (ASEs) for experimentally verified Ψs in RN7SK and RN7SL. Furthermore, modified bases have been detected in snoRNAs themselves (Kishore et al., 2013) and in mRNAs (Schwartz et al., 2014; Carlile et al., 2014). Newly developed high-throughput protocols are capable of identifying base modifications in RNAs pulled from the cells. As snoRNAs and in consequence their editings are found differential expressed a broader variety of investigated samples from different tissues and under different environmental influences will probably pinpoint more RNAs with snoRNA guided modifications, meaning that the interaction network is still increasing.

2.4 Diversity of SnoRNAs and SnoRNA Function

Besides the canonical nucleotide modifications an amazing variety of additional snoRNA tasks have been discovered during the last years. See a recent broad review on this is: (Dupuis-Sandoval et al., 2015).

Small Derived RNAs Next generation sequencing experiments revealed that snoRNAs are frequently processed into smaller RNAs, so-called small derived RNAs (sdRNAs). They occur as stable transcripts in the cell (Taft et al., 2009; Ender et al., 2008). In fact, many box H/ACA snoRNAs, but not box C/D snoRNAs, are substrates for *Dicer* (Langenberger et al., 2012b). For some of them, there is also evidence to

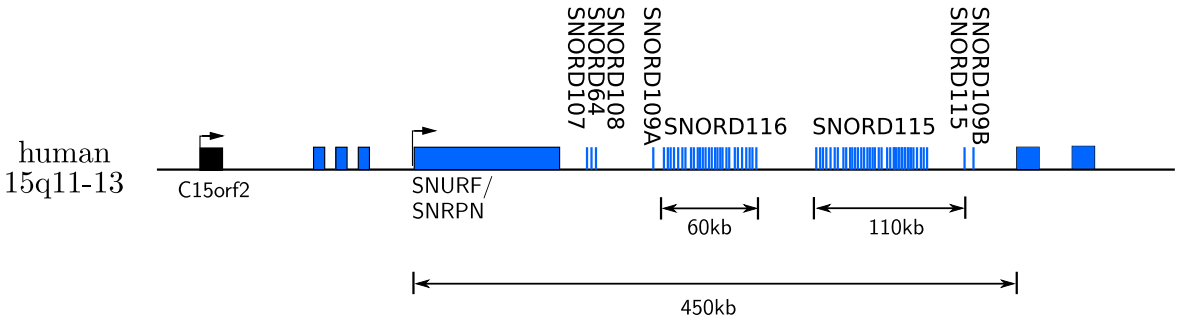

Figure 2.6: The SNURF-SNRPN region is imprinted in human. It codes for members of seven snoRNA families. SNORD116 and SNORD115 are encoded in arrays of 35 and 48 copies. Also sno-lncRNA with SNORD116 paralogous at each end occur. Defect in this locus can cause the Prader-Willi syndrome. Figure adapted from Bachellerie et al. (2002)

associate to the RISC-complex and act, like miRNAs, in regulation of mRNA translation. This is also a hint for an ancient relationship between snoRNAs and miRNAs (Scott and Ono, 2011). Still the majority of sdRNAs does not function like canonical miRNAs, in fact their cellular tasks remain unclear. Nevertheless, strong and conserved profiles and their accumulation in different tissues instead of being scattered in the cell, indicate an unknown cellular function and underline that they are more than transient degradation products (Scott et al., 2012).

SNURF-SNRPN snoRNAs Further prominent and yet not fully understood examples of non-canonical snoRNAs are encoded in the imprinted, paternally expressed SNURF-SNRPN region, a locus associated with the Prader-Willi syndrome (Figure 2.6). Among these snoRNAs are the members of the brain specific box C/D snoRNA family SNORD115 (HBII-52) that have been reported to influence alternative splicing of the serotonin receptor and several other mRNAs in human and mouse (Kishore and Stamm, 2006; Doe et al., 2009; Kishore et al., 2010; Soneo et al., 2010). Like SNORD115, also SNORD116 is encoded in the SNURF-SNRPN region. Both are arranged in arrays containing multiple (48 and 35), highly similar copies. Remarkably, alternative splicing of the SNORD116 locus also gives rise to long RNA transcripts spanning from snoRNA to snoRNA, with the region between the two 'common' snoRNAs being retained in these transcripts. Thus the whole region is expressed as a stable long non-coding RNA, termed sno-lncRNA (Zhang et al., 2014). Interestingly, these transcripts do neither accumulate to nucleoli nor to Cajal Bodies. There is strong evidence that the SNORD116 sno-lncRNAs act as sinks for Fox2, which regulates alternative splicing.

Sno-lncRNAs Sno-lncRNAs have been observed not only for SNORD116, but seem to be a more general phenomenon, including also lncRNAs with box H/ACA snoRNAs at the transcript ends (Yin et al., 2012). The prerequisite for these transcripts is to have a pair of snoRNAs within a single intron. Coinciding, these sno-lncRNAs have been observed within alternatively spliced host-genes, which in consequence dynamically give birth to snoRNAs or sno-lncRNAs (Yin et al., 2012). Considering the way snoRNAs are usually processed they seem to be a natural consequence of the snoRNA processing pathway. Where alternative splicing results in a pre-mRNA carrying two snoRNAs in one intron the core proteins bind the characteristic snoRNA boxes immediately after transcription. In consequence after linearization of the intron the exonucleases starts digesting their ends. However, the bound proteins are a protection against degradation, meaning that the snoRNAs and the sequence part in between is protected from degradation, leaving the lncRNA with snoRNA ends.

Chromatin Associated RNAs A further unexpected observation was made, when snoRNAs were found to make up a large fraction of chromatin associated RNAs (caRNAs). CaRNAs are responsible for proper decondensation of chromatin and as such regulate gene expression (Mondal et al., 2010). Mainly low abundant box H/ACA snoRNAs are found to be enriched in caRNAs. Those snoRNAs are not associated to the canonical core proteins. Instead, they are bound to Df31, a protein essential for decondensation of chromatin (Schubert et al., 2012). The exact role or mode of action has not been disclosed, yet.

SnoRNAs in Stress Response Another observed coherence exists between snoRNAs and several stress responses. As such SNORD32A, SNORD33, SNORD35A all encoded in introns of RPL13A turned out to be general mediators of oxidative stress, e.g. protecting against lipoxisity and water intoxication (Michel et al., 2011). Further snoRNAs are involved in the regulation of endoplasmatic reticulum (ER) function (SNORD3A) (Cohen et al., 2013) or the trafficking of cholesterol to the ER (SNORD60) (Brandis et al., 2013). The stress response modulation is independent from the guided 2'-O-methylation as down-regulation of the according box C/D snoRNAs does not result in reduced methylation levels in rRNAs. Thus, the mediatory effect must be caused either by 2'-O-methylation of additional target RNAs or by another still unknown mode of action. Also the occurrence of several differentially expressed snoRNAs in distinct types of cancer could be connected to the involvement in stress regulation. The expression profiles of snoRNAs can be that specific that they serve as bio-markers for

certain cancer types (e.g. SNORD50 (Tanaka et al., 2000) and SNORA42 (Mei et al., 2012)) (Williams and Farzaneh, 2012). Another explanation though is that the interplay actually is on immune response level. An indication for this is that SNORD33, involved in oxidative stress response and SCARNA22 are both frequently found overexpressed in cancer and also both have been found bound with Interleukin 3 (IL3) through immunoprecipitation (Dupuis-Sandoval et al., 2015).

Telomerase RNA A long known example of a special snoRNA-like domain is the telomerase RNA component (TERC). The 3'-end of animal telomerase RNA (hTR) is constituted by a box H/ACA scaRNA domain (SCARNA19) including a CAB-box (Mitchell et al., 1999; Zhang et al., 2011; Li et al., 2013). The RNA molecule is responsible for maintenance of chromosome telomere ends in the progress of DNA synthesis. It is discussed whether the snoRNA domains obligation merely is to localize hTR to the Cajal Body (Zhang et al., 2011) or whether the domain, which is even bound to dyskerin has a further function. Notably, mutation of either component of the telomerase enzyme complex, be it hTR, dyskerin, or human telomerase reverse transcriptase (hTERT) are the only cause of dyskeratosis congenita (Cohen et al., 2007).

CHAPTER 3

Innovative Analysis of SnoRNA-Target Interactions

The current chapter introduces the necessary new and improved tools that were developed to systematically study snoRNAs in respect to their evolution, their features and their function. These tools form the basis for the studies described in the following chapters. Additionally, the startset of experimentally verified snoRNAs, the target RNA sequences and reported modifications are provided. As no comprehensive and up-to date snoRNA data set in vertebrates exists, first, a broad collection of snoRNAs was needed. The `snoStrip` pipeline (Bartschat et al., 2014) enables easy genome-wide searches for homologous snoRNAs and comprehensively analyzes their features. To easily access the information, all snoRNA sequences and characteristics extracted during a `snoStrip` run are automatically stored in an attached `mySQl` database, named `snoBoard`. To study the guiding function of snoRNAs, the snoRNA-targetRNA interactions need to be predicted in thermodynamically correct manner. Due to the fundamental differences between box H/ACA snoRNAs and box C/D snoRNAs, different approaches are implemented in the tools `RNAsnoop` and `PLEXY`. The last step is then to combine the target predictions of homologous snoRNA sequences to study conservation of the interactions. All collected data can be integrated to find the general pattern of functional homologies during snoRNA evolution. To do this on a formal basis a score to measure conservation of the interactions, termed Interaction Conservation Index (ICI) was developed and is introduced in the chapter. Subsequently, an innovative analysis workflow combining all tools is outlined. The `snoStrip` pipeline, `RNAsnoop` and `PLEXY` are published in *Bioinformatics* (Bartschat et al., 2014; Tafer et al., 2010; Kehr et al., 2011). The ICI score is published in *Molecular Biology and Evolution* (Kehr et al., 2014).

3.1 Materials

In following the data collection that forms the basis for the studies described in this thesis are described. Starting with presenting an general approach to annotate ncRNA the focus is brought to the extraction of vertebrate snoRNA sequences through homology search with our `snoStrip` pipeline (Bartschat et al., 2014). Additionally, to examine the snoRNA guiding function the target RNAs, ribosomal RNA, and snRNA sequences, were collected from multiple resources. On top modifications that have been reported in the target RNAs were gathered. Details about the materials that were used for all parts for the work are provided in following. Further materials, that are only relevant for certain aspects are added at the appropriate passage in the sections.

SnoRNA Annotation

Today comprehensive annotation of non-coding RNAs is part of every relevant genome project. Nevertheless, it is a complex task and a standard pipeline does not exist. In contrast to protein coding genes ncRNAs lack strong statistical signals like open reading frames, high G+C content and codon usage bias (Rivas and Eddy, 2000). In following first a general ncRNA annotation workflow is explained to gain an overview of the necessary steps and putative data sources. This approach is applicable to each type of non-coding RNA family. The main drawback for fast evolving RNAs however is that sequence similarity alone, are often not sufficient to decide whether a candidate sequence is indeed a functional homolog, or is more likely a deviant nonfunctional pseudogene. Therefore, such general approaches, are often combined with type specific methods. For many ncRNA classes such annotation tools already exist, e.g. tRNAs or rRNAs. For snoRNAs no such pipeline was available so far. To fill this gap the `snoStrip` pipeline was developed.

General NcRNA Annotation Approach

A workflow of (ncRNA) gene annotation is shown in Figure 3.1. In general, the first approach is a homology based search using known RNA genes from related organisms as queries. The RNAs are highly structured and have high mutation rates, thus the search method of choice should capture sequence and structure similarities (e.g. `Infernal` (Nawrocki and Eddy, 2013), `GotohScan` (Hertel et al., 2009)). As runtime of these

programs is high, often a BLAST search is preposed to efficiently capture homologies that are already identifiable on sequence level only. Afterwards, the time consuming search needs to be done only for those queries without findings with BLAST.

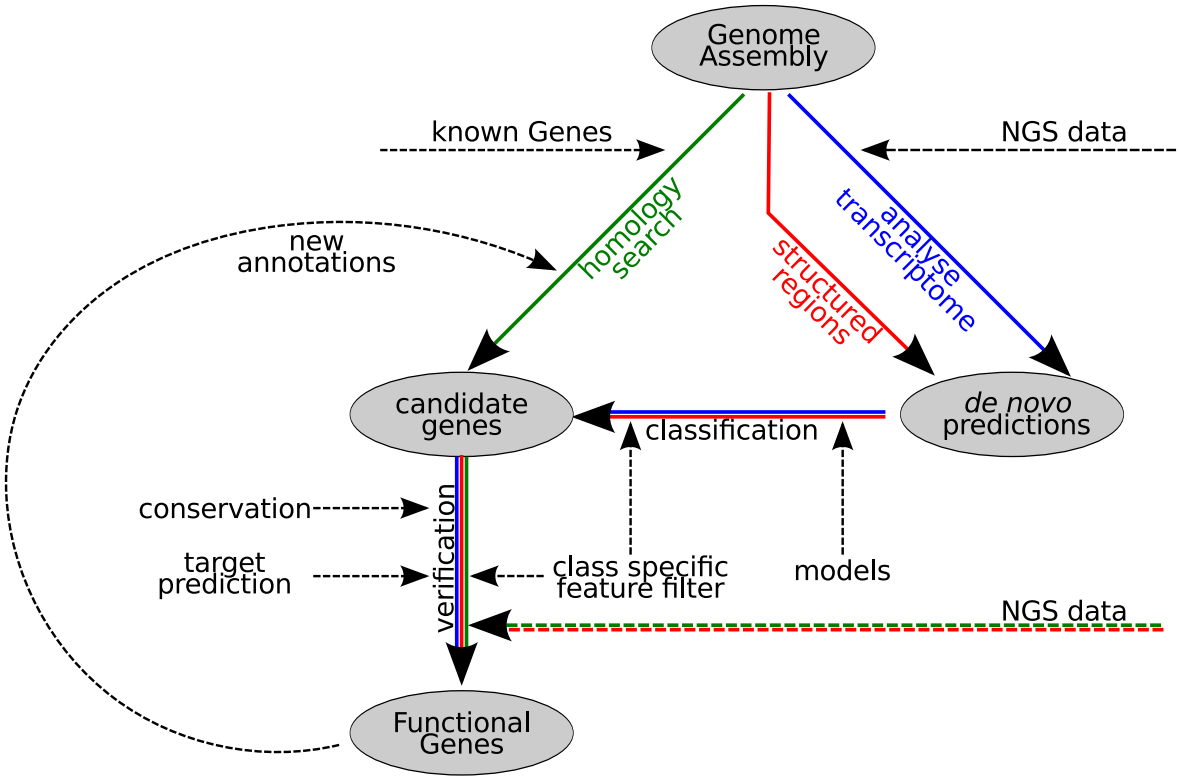

Figure 3.1: General workflow of ncRNA annotation. Starting from a genome assembly three approaches are followed. The comparative approach (green): based on known genes from related species, homology search reveals candidate genes. The genome approach (red): putative genes are extracted from the genome by identification of highly structured regions with ***RNAz****. The transcriptome approach (blue): if data from next generation sequencing (NGS) experiments are available, expressed genomic loci can be used to gain de novo gene predictions. Predictions from the latter two methods can be classified with machine learning techniques according using gene models or class specific feature filters. The sum of all candidate genes can still include nonfunctional pseudogenes and has to be further validated. Conservation analysis, ncRNA-class specific features like structure or sequence motifs, prediction of putative targets for functionality check and expression analysis can help to identify the set of functional genes in the studies genome. (Further details in the text.)*

To recognize *de novo* ncRNAs for which no homologous sequences are available, the genome is scanned for highly structured and conserved loci with RNAz (Gruber et al., 2010). These regions are putatively functional. The method has a high rate of false positive *de novo* predictions and needs further processing. Additionally, if available

high-throughput sequencing data is used to identify expressed ncRNAs that have significant transcription signals.
Overlapping RNA loci from the latter two *de novo* prediction approaches need categorization to known ncRNA classes. The classification is based on class specific features. Here known structural elements like kink-turns, pseudoknots, sequence motifs or other characteristics should be considered. Often machine learning techniques are applied. Therefore, models derived from known representatives of the ncRNA classes have to be available for positive training sets, as well as distinguishing negative training sets.
All identified candidate RNAs from the comparative, the genomic, and the transcriptomic method are then integrated and further validated. Validation can be based on conservation analysis and again fulfillment of class specific constraints, (e.g. obligatory box motifs and for structure). The challenge in this step is to find a reasonable balance between specificity and sensitivity to withdraw non-functional pseudogenes but keep sequences with species specific specialties. Additionally, target predictions can help to decide whether an RNA gene is more likely functional or a pseudogene. If available, candidates originating from the homology based and the genome based approaches should be validated by expression signals from next generation sequencing (NGS) data. The validated functional genes can be used as queries in future homology searches.

`SnoStrip` and `SnoBoard`: Homology Search Pipeline for SnoRNAs

For snoRNA homology search a more elaborate pipeline, termed `snoStrip` was developed in collaboration with Mr. Canzler (Bartschat et al., 2014; Canzler, 2016). The difficulty in detecting snoRNAs, like other RNAs, is their fast evolution. On top snoRNA genes are versatile elements that frequently duplicate and relocate. These attributes make decisions on sequence similarity alone, an precarious issue. To distinguish between functional orthologs and paralogs and non-functional pseudogenes, especially, the short snoRNA sequence motifs are crucial. In their absence the core proteins are not able to recognize the snoRNA sequence. For the same reason, folding into the specific secondary structure has to be possible. In addition the ability for analogous guiding function for the members of the same snoRNA families should be considered.
Given a snoRNA family, which can consist of only one snoRNA sequence in a single species or a set of homologous sequences in several organisms, the pipeline searches for

new homologous snoRNAs in the desired genomes. The pipeline embraces four main parts:

1. A basic homology search identifies potential snoRNA candidates in the analyzed genome. First, `blastn` with relaxed parameters is employed. As query sequences all snoRNAs already contained in the snoRNA family are used. If no candidate is returned, a covariance model (CM) is generated from the snoRNA family, which is subsequently used for a genome scan with `infernal` version 1.1 (Nawrocki and Eddy, 2013).

2. The identified snoRNA candidates are then filtered to ensure that only functional snoRNAs are kept. Specifically, the presencs of the characteristic box motifs is validated. These ensure functionality of the snoRNAs mainly through enabling binding of the core proteins (Section 2.2). For box C/D snoRNAs additionally only sequences that can form the obligatory kink-turn are retained. A further filtered warrants the conservation of at least one of the two putative target sites, as homologous snoRNA sequences should be able to execute the same function .

3. Additionally, characteristic features of the accepted snoRNA sequences are analyzed. An important feature of snoRNAs is their type-specific secondary structure. The structure is influenced by the bound proteins, due to this the folding method of choice is `RNAsubopt` (Wuchty et al., 1999). Thus from the set of high scoring putative secondary structures the best which conforms the characteristic snoRNAs structure is selected. During folding the fulfillment of type-specific folding constraints is enforced. These are:

- box C/D snoRNAs are required to contain an internal loop delimited by the boxes C and D,

```
.....CCCCCCC.....................................DDDD......  box-annotations
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......  folding constraint
(((((.....................................................))))))  ideal structure
```

- box H/ACA snoRNAs are prohibited to form base pairs in their hinge and tail regions.

```
...........................ANANNA.................................ACA...  box-annotations
...........................XXXXXX.................................XXXXXX folding constraint
.(((((....(((....)))....)))))............(((((....(((.....)))..... ideal structure
```

SnoRNA host genes are retrieved using the `Ensembl` API or by `mySQL` requests to UCSC. For all surrounding transcripts the number of the intron that encodes the snoRNA and the transcript ID is captured. As actual host gene the longest surrounding transcript is selected. Also the function of the added snoRNA is considered. Target predictions for the snoRNA sequences are either performed by `RNAsnoop` (box H/ACA snoRNAs) (Tafer et al., 2010) or by `PLEXY` (box C/D snoRNAs) (Kehr et al., 2011) (These programs are described in Section 3.2).

4. Finally, an alignment of the whole snoRNA family is computed with `muscle`.

The `snoStrip` pipeline is closely interwoven with the `snoBoard` database. First, each novel snoRNA identified during a `snoStrip` run and its corresponding derived information are stored in the linked `snoBoard` database. Secondly, during the homology search procedure all snoRNA sequences already assigned to the current family are considered as query sequences and the derived snoRNA family features are used as accurate filters for the identified snoRNA candidates.

The internal database structure reflects both functions in two main tables. In the *snoRNA* table each record keep detailed information about the snoRNA sequence, including nucleotide sequence, corresponding length, genomic coordinates, putative structure with predicted minimum free energy, host gene, alternative surrounding transcripts and corresponding number of the encoding intron. Additionally logged are the successful query sequence, corresponding search scores, date of record, and genome_source. The boxes and targets are stored in related tables. These tables can easily be joined based on snoRNA identifiers, but also independent requests on snoRNA motifs or targets are straightforward. The second main table *homology* holds the snoRNA family information. Each entry corresponds to one snoRNA family and lists all members in the investigated species. Each table column represents one species, so also conservation gaps, can easily be derived.

Chapter 4 presents the application of `snoStrip` to a plethora of species.

With the `snoStrip` pipeline a convenient and efficient way to annotate homologous snoRNAs in available genomes was provided. Conservation, of single snoRNA genes can be evolutionarily traced across a widespread of species. Therein the snoRNA annotation goes far beyond the confident identification of a snoRNA sequence. Each run already delivers box annotations, secondary structures, synteny information, alignments and target predictions for each detected snoRNA sequence. These are efficiently captured in the `snoBoard` database, where they can be comfortably accessed for further

analysis. SnoBoard is a valuable resource to derive and improve existing snoRNA models, derive fundamental principles, and also to identify specialties of certain snoRNA families or within certain clades. All in all shedding light on the evolution, function and potential functions of this fascinating class of non-coding RNAs.

SnoRNA Startset The snoRNAs in 47 vertebrate species (Appendix Table A.1 and Figure A.4) were annotated. As queries for the snoStrip homology search runs experimentally verified sequences from snoRNA-LBME-db[1] (Lestrade and Weber, 2006), chicken snoRNAs reported by Shao et al. (2009), and platypus snoRNAs from Schmitz et al. (2008) were combined. Highly similar sequences in terms of base identity, especially at antisense elements were merged into one family. The sequences are provided in fasta-format with a specific fasta-header:

>type_ID_species_(chromosome)_start,end_strand_[C' box_start_D' box_start]_C|H box_start_D|ACA box_start

>CD_95_H.sapiens_(chr5)_180602916,180602983_-_GTGCTGA_36_CTGA_26_GTGATGA_5_CTGA_59

GGCGGTGATGACCCCAACATGCCATCTGAGTGTCGGTGCTGAAATCCAGAGGCTGTTTCTGAGCTGCC

To extract the genome coordinates for the header the sequences were blated against the genome assemblies. If available, box-annotations were taken from literature. Otherwise, a PWM based approach was used to identify the box motifs in the sequences. The nucleotide frequencies in the matrices are derived from reported sequence motifs. Also the knowledge about the approximate position within the snoRNA sequences can help to identify the motifs. In box H/ACA snoRNAs the typical hairpin-hinge-hairpin-tail structure defines the location of the boxes in the hinge and the tail region. In box C/D snoRNAs a pair of boxes is situated near the sequence ends and another in its central region. However, especially in the case of variant prime boxes the correct motifs can occasionally be identified only after alignment of several members of the snoRNA family. The increased conservation of the motif itself and the adjacent ASE enables sound annotation of the according boxes.

To mention explicitly, to ensure a high confident query set, we do not incorporate snoRNA gene predictions from RFAM[2], ENCODE[3] and snOPY[4] here. The methods that have been used for their automated snoRNA annotations do not filter the sequences for presence of obligatory sequence or kink-turn motifs. Due to that, although valuable

[1] https://www-snorna.biotoul.fr/
[2] http://rfam.xfam.org/
[3] https://www.encodeproject.org/
[4] http://snoopy.med.miyazaki-u.ac.jp/snorna_db.cgi

resources, they naturally contain several pseudogenes and false positive predictions.

Target Set of rRNAs and snRNAs The sequences of all parts of the rRNA operon were collected for the investigated vertebrate species (Appendix A.1) as putative targets. In particular, available sequences for 18S, 5.8S and 28S rRNA were retrieved from the `SILVA rRNA database`[5] (Pruesse et al., 2007), `Ensembl`[6] and `NCBI`[7] databases. Using the `RNAmmer` (Lagesen et al., 2007) approach some of the sequences not annotated in the databases were successfully identified. For human and chicken we have used the same rRNA sequences as Lestrade and Weber (2006) and Shao et al. (2009) for verification reasons.
Unfortunately, rRNA operons are often excluded from genome assemblies. At the same time, in particular LSU rRNAs have been rarely cloned and sequenced in independent studies e.g. for phylogenetic purposes. As a consequence, we have a poor coverage of 28S rRNAs in many of the recently sequences vertebrates. While the majority of SSU RNAs (44/47) were collected, for LSU only 17/47 nearly full-length sequences could be used for comparative analyses of modification sites. A suitable 5.8S rRNA sequence was found in 41 species. Nevertheless, the phylogenetic range of the sequences in all three alignments (18S, 28S, 5.8S) spans the vertebrates from lamprey to human.
Small nucleolar RNA sequences (U1, U2, U4, U5, U6, U11, U12, U4atac, and U6atac) were taken from Marz et al. (2008).

Known SnoRNA-guided Modification Experimentally verified positions of chemical modifications within rRNAs and snRNAs were collected from Maden (1986, 1996); Ofengand and Bakin (1997), `The SSU rRNA Modification Database`[8] (McCloskey and Rozenski, 2005), `The RNA Modification Database`[9] (Cantara et al., 2011) and `snoRNA-LBME-db`[10] (Lestrade and Weber, 2006). Modifications in U2 snRNA were additionally taken from Dönmez et al. (2004) and Yu et al. (1998). Further modifications in other snRNAs are provided in Deryusheva et al. (2012) and Karijolich and Yu (2010). Predicted and verified interactions between snoRNAs and their targets were collected from `snoRNA-LBME-db` (Lestrade and Weber, 2006) and from the literature (Xiao et al.,

[5] `https://www.arb-silva.de/`
[6] `http://www.ensembl.org/index.html`
[7] `http://www.ncbi.nlm.nih.gov/`
[8] `http://rna.rega.kuleuven.be/ssu/`
[9] `http://mods.rna.albany.edu/`
[10] `https://www-snorna.biotoul.fr/`

2009; Higa-Nakamine et al., 2012; Badis et al., 2003). Very recently developed high-throughput protocols are able to map also 2'-O-methylation and Ψ in RNA transcripts and added further modifications sites to the set (Birkedal et al., 2015; Krogh et al., 2016; Zaringhalam and Papavasiliou, 2016; Jorjani et al., 2016).

Comparative Genomics of Target RNAs To determine homologous positions of known or predicted modification sites between different species high quality alignments of the vertebrate target RNAs are required. The rRNA sequences contain both regions of high and of low conservation. An appropriate alignment method is `RNAsalsa` (Stocsits et al., 2009), which was specifically designed to compute sequence structure alignments for ribosomal RNA sequences. The initial input alignments for `RNAsalsa` were computed with `muscle`. Secondary structure information of human 18S and 28S rRNA sequences (Cannone et al., 2002) were used as initial structural constraints. The initial structure constraint for human 5.8S rRNA was predicted using `RNAfold` (Hofacker et al., 1994).
Manually curated alignments of spliceosomal RNAs (U1, U2, U4, U5, U6, U11, U12, U4atac, and U6atac) are provided in Marz et al. (2008).
All computed target RNA alignments can be accessed at the supplement page to Kehr et al. (2014) www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/12-022 (S6.1, S6.2, and S6.3).

Mapping of Modifications Mapping between positions in rRNA and snRNA sequences in a certain species and their corresponding positions in the alignments has been realized with the `BioPerl` packages *AlignIO* and *SimpleAlign* (Stajich et al., 2002). A table of all modified positions in rRNAs and snRNAs is provided at www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/12-022/mapping.html.

3.2 Target Prediction for SnoRNAs

SnoRNAs guide the chemical modification of rRNAs and snRNAs at approximately 250 nucleotides. The specific site is defined through base pairing between the snoRNA ASE and the region surrounding the introduced modification.
In general RNA-RNA interaction is computed, similar to RNA folding, by free energy minimization based on thermodynamic modeling. The full problem is disassembled

into smaller problems. Meaning the optimal and suboptimal solution of the complete problem is gained by combination of small structural subunits with tabulated energy values by dynamic programming techniques (Section 1.1.2)

To efficiently identify the targets of snoRNAs `RNAsnoop` and `PLEXY` have been developed. Different algorithms are needed as the interaction modes differ substantially between box H/ACA snoRNAs and box C/D snoRNAs (Section 2.2). The programs have been published previously and have meanwhile become state-of-the-art in snoRNA target prediction (Bratkovič and Rogelj, 2014).

3.2.1 `RNAsnoop`: Target prediction for box H/ACA snoRNAs

`RNAsnoop` is a specialized co-folding algorithm to predict the intricate interactions between box H/ACA snoRNAs and their target RNAs (Tafer et al., 2010). As input, `RNAsnoop` requires the sequence of a single hairpin of a snoRNA and one or more sequences of putative target RNAs.

The hybrid structure formed by the snoRNA hairpin and a target RNA is a pseudo-knot, since the target RNA binds inside an interior loop of the snoRNA. Furthermore, the antisense element in the snoRNA sequence is disrupted such that the target RNA binds in-contiguously to two binding sides flanking a helical stem region (Section 2.2).

The common complexity to compute such a pseudo-knotted structure is bounded by $O(n^3 \cdot m^3)$, where n is the length of the snoRNA and m is the length of the target RNA. This would hamper a genome wide target search. Though the knowledge of the hybridization structure of snoRNA and target RNA makes it possible to reduce the runtime to $O(n^2 \cdot m)$ and thus enable a genome-wide screen, e.g., for possible mRNA targets influencing alternative splicing.

The hybridization between snoRNAs and their corresponding target RNA can be divided into three parts. First, the exact shape of the upper stem-loop in the snoRNA structure is predicted independently from the target RNA. Matrix M includes an unbranched structure prediction for the snoRNA (with sequence y) in absence of the interaction partner. In this way, possible locations of the pseudouridylation pocket, the interior loop in which the target RNA can bind, are determined (Figure 3.2A). The energies of the optimal substructures satisfy the recursion:

$$M_{p,q} = \min \begin{cases} \mathcal{H}(y[p,q]) \\ \min\limits_{k,l} M_{p-k,q+l} + \mathcal{I}(y[p-k,p]; y[q,q+l]) \end{cases}$$

where $\mathcal{H}(y[p,q])$ denotes the energy parameters (Lu et al., 2006; Mathews et al., 1999) for a hairpin loop formed by the subsequence $y[p,q] = y_p y_{p+1} \dots y_q$ including the closing pair (y_p, y_q). Analogously, $\mathcal{I}(y[p-k,p]; y[q,q+l])$ is the energy of an interior loop composed of the sequences $y[u,p]$ and $y[q,v]$, including the closing pairs (y_p, y_q) and (y_u, y_v).

Second, the hybridization between the target RNA and the 5'-part of the interior loop is calculated and stored in matrix L. For simplicity, this region will be referred to as left binding site in the following. In the duplex, stacked pairs and mismatches up to length two are allowed. However, bulges are absent in the interaction. Hence,there are three possibilities for nucleotide j from the snoRNA sequence y and nucleotide i from the target sequence x:
(1) they form a base pair (i,j), stacked to a pair $(j+1, i-1)$ (Figure 3.2B top),
(2) they form a base pair (j,i), although $(j+1, i-1)$ is a mismatch, but $(j+2, i-2)$ forms a base pair (Figure 3.2B middle), or
(3) they form a base pair (j,i), although $(j+1, i-1)$, and $(j+2, i-2)$ are mismatches, but $(j+3, i-3)$ pairs (Figure 3.2B bottom)

The recursion is:

$$L_{i,j} = \min_{k=1,2,3} \Big\{ L_{i-k,j+l} + \mathcal{I}(y[i-k,i]; y[j,j+l])$$

The last step consists of the computation of the duplex between the target RNA and the 3'-part of the interior loop (analogously called right duplex in the following). The right duplex is captured in matrix R. The start of the interaction is formed by the pair (j,i), which has to fulfill some special conditions (Figure 3.2C):
(1) matrix M has to contain the pair $(j+1, j+k)$, which is the closing pair of the upper stem,
(2) bases $(j+k+1, i-3)$ have to form the closing pair of the right interaction in matrix R.

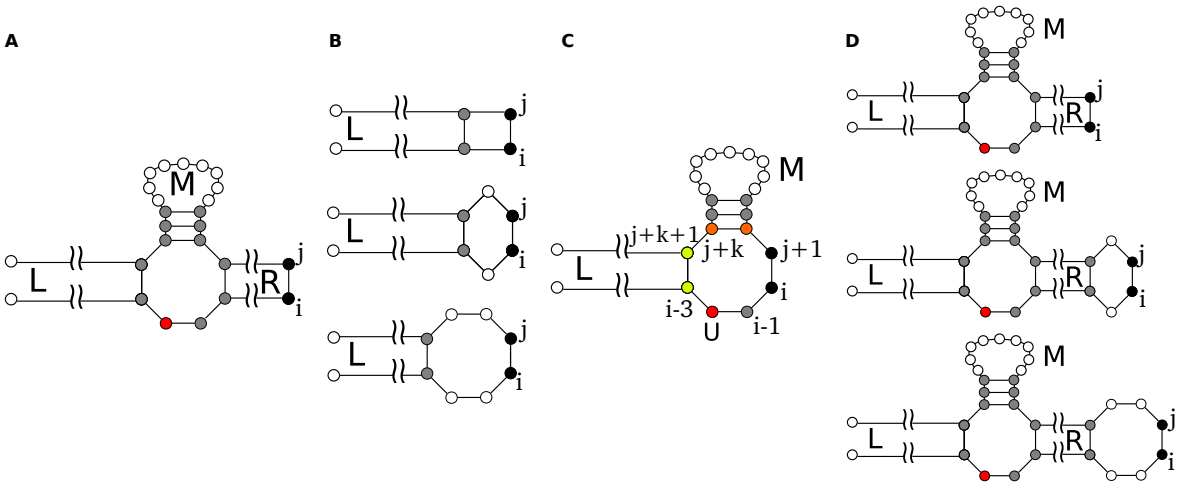

Figure 3.2: ***A*** *The upper stem loop structure of the snoRNA is computed and stored in matrix* M*.* ***B*** *The the left duplex is extended and contained in matrix* L*.* ***C*** *The start positions of the right duplex have to fulfill some additional constraints.* ***D*** *Then the right is computed analogously to the left and stored in matrix* R*. Details are described in the text.*

(3) the nucleotide $i-1$ must be unpaired, and (4) $i-2$ has to be an uracil that is to be pseudouridilated.

The extension of the right duplex (Figure 3.2D) is analogous to the extension of the left duplex, allowing only stacked pairs and symmetric mismatches up to length two, satisfying the recusion:

$$R_{i,j} \quad = \quad \min_{k=1,2,3} \begin{cases} \min\limits_{k,l\leq 2} R_{i-k,j+l} + \mathcal{I}(y[i-k,i];y[j,j+l]) \\ \min\limits_{l\in[3,|y|-j} L_{i-3,j+k+1} + M_{j+1,j+k}, \text{ if } x[i-2]=U \end{cases}$$

Finally, in the backtracking step the (sub)optimal solutions are traced by combining all three matrices, starting with the lowest energy value in matrix R.

The internal structure of the target RNA can also be relevant, because a putative structure containing the targeted nucleotides as base pairs needs to be opened previously to the hybridization with the snoRNA. The secondary structure of the RNAs can be provided in form of accessibility profiles, which contain the probability that a certain sequence interval is unpaired, and hence provide the energy necessary to make the interaction site available for binding.

If accessibility profiles are provided, `RNAsnoop` considers the energy needed to open

the internal structure of the target RNA. It is simply added as penalty to the overall computed minimum free energy of the interaction.

3.2.2 PLEXY: Target Prediction for box C/D snoRNAs

The interaction between a box C/D snoRNA and a target RNA is quite simple. A limited region of the snoRNA sequence, the $7 - 20$ nucleotides long antisense element (ASE), has the ability to form small duplexes with the target RNA. These duplexes comprise only few mismatches and no bulges. For further details on the interactions see Section 2.2 and Figure 2.2B. Nevertheless, no sufficient tool for the target prediction of box C/D snoRNAs was available. The only published tool is `snoTarget`, which uses a perl regular expression to search for a complementary stretch of RNA (Bazeley et al., 2008). The tool `PLEXY` was developed, to predict the interactions based on thermodynamic modeling (Kehr et al., 2011).
The following steps are performed for D-box and D'-box, consecutively. First, `PLEXY` uses the information about the box positions, to extract the antisense element immediately upstream of the D-box/D'-box. Then, `RNAplex` is used to compute all stable duplexes between the interacting region of the snoRNA and each putative target RNA of the relevant organism. `RNAplex` (Tafer and Hofacker, 2008; Tafer et al., 2011) implements a RNA folding algorithm, which is optimized for scanning speed. The energy model for long interior loops is simplified. As these do not occur in the short duplexes anyhow, we do not expect loss of accuracy, here. Parameters passed to `RNAplex` are a threshold on the minimum free energy of at most -7.00 [kcal/mol] and a limitation of duplex length to 20 base pairs. The short predicted duplexes returned from `RNAplex` are filtered with respect to additional criteria derived from (Chen et al., 2007).

- No bulges on either side of the interaction are allowed.
- The interaction should be at least 7 nucleotides long.
- The 3rd to 11th position of the interaction forms the 'core region'. In this interval, at most one mismatch is allowed (Figure 2.2B).
- The methylated residue is the nucleotide that base pairs with the fifth nucleotide upstream of the D-/D'-box in the snoRNA sequence. The formed base pair at this position has to be a real Watson-Crick pair.

Finally, the putative target sites are ranked by the computed duplex energies.
Analogously to `RNAsnoop`, `PLEXY` has the option to consider the opening energies for putative internal target RNAs, which have to be provided in form of accessibility profiles. Interestingly, the penalization leads to reduced snoRNA target RNA interaction recovery in yeast. However, this is in accordance with the observation that unlike in the case of pseudouridines, the accessibility around methylated nucleotides are not that different from accessibility of unmodified nucleotides (data not shown).

3.2.3 Interaction Conservation Index

The interactions between the snoRNAs and and their target RNAs can be fairly short, ranging from 7-20 nts. Assuming a nucleotide sequence of length 7, there are $4^7 = 16384$ permutations. In the human genome with 3.2 billion base pairs, a rather simplified calculation, that assumes equal distribution of nucleotides in the genome, each such 7 nts long sequence stretch would occur $3,200,000,000 \div 16384 \sim 195310$ times by coincidence. In contrast an interaction of 20 nts is already significant and rare: $3,200,000,000 \div 4^{20} \sim 0.0029$. However, there can be hordes of putative interaction candidates for each ASE making it hard to select the 'true' interaction within a set of equally good scoring hybridization in terms of MFE. On the one hand the number of putative targets can be massively reduced by limiting the search space from the whole genome to the transcriptome or to known targeted RNAs. On the other hand, evolutionary preservation can give further support.
Moreover, the availability of individual target predictions of a snoRNA sequence in single species (from `RNAsnoop` and `PLEXY`) leads us to ask additional general questions on conservation and evolution of snoRNA guiding function. It has been observed, that modifications at equivalent sites are conserved over long evolutionary distances (Ofengand and Bakin, 1997). It has yet only been speculated, that they are introduced by homologous snoRNA guides (Chen et al., 2008).
To formally investigate the conservation of RNA-RNA interactions the Interaction Conservation Index (ICI) was developed. It combines the quality of an interaction in a single organism with the scope of the presence of this interaction in a set of other species. It serves as efficient measure to evaluate the conservation of the interaction between snoRNA and target RNA. In this section this important measurement that forms the basis of analyzing the conservation of the snoRNA guiding function is elaborated.

A target t specifies a particular column in the alignments of the target sequences. A snoRNA family s is defined by homology, i.e., it may contain more than one paralog. $X(t,s,k)$ denotes the set of all snoRNAs from family s in species k that are predicted to target t in species k. Furthermore, $S(t,k)$ is written for the set of snoRNA families predicted to target t in species k, i.e., $S(t,k) = \{s|X(t,s,k) \neq \emptyset\}$. Similarly $O(t,s) = \{k|X(t,s,k) \neq \emptyset\}$ denotes the set of species in which family s has a representative that targets t and $T(s,k) = \{t|X(t,s,k) \neq \emptyset\}$ is the set of targets of the snoRNA family s in species k.

The interaction is scored at the level of families

$$\varepsilon(t,s,k) = \min_{x \in X(t,s,k)} E_{\text{mfe}}[x, y_{t,k}] \tag{3.1}$$

where $E_{\text{mfe}}[x, y_{t,k}]$ is the energy of the interaction between the snoRNA x and the sequence $y_{t,k}$ around the target t in species k, computed as minimum free energy of the interaction given by `PLEXY` or `RNAsnoop`. Hence, if more than one paralog in the same family s exists, only the one with the best interaction energy is considered.

The average predicted interaction energy for target t in species k is $\bar{\varepsilon}(t,k) = \sum_{s \in S(t,k)} \varepsilon(t,s,k)/|S(t,k)|$, while the average interaction energy of snoRNA s with all its putative targets t in species k is $\hat{\varepsilon}(s,k) = \sum_{t \in T(s,k)} \varepsilon(t,s,k)/|T(s,k)|$. Averaging over all species in which the interaction is predicted, the two normalized parameters are introduced

$$\begin{aligned} \text{ici}_{\text{mod}}(t,s,k) &= \varepsilon(t,s,k)/\bar{\varepsilon}(t,k) \\ \text{ici}_{\text{sno}}(t,s,k) &= \varepsilon(t,s,k)/\hat{\varepsilon}(s,k) \end{aligned} \tag{3.2}$$

For $k \notin O(t,s)$ these energy-based scores are not defined since no member of family s interacts with target t in species k.

In order to summarize these data over all species, the *Interaction Conservation Indices* is defined

$$\begin{aligned} ICI_{mod}(t,s) &= 1/|O_s| \sum_{k \in O(t,s)} \varepsilon(t,s,k)/\bar{\varepsilon}(t,k) \\ ICI_{sno}(t,s) &= 1/|O_s| \sum_{k \in O(t,s)} \varepsilon(t,s,k)/\hat{\varepsilon}(s,k) \end{aligned} \tag{3.3}$$

for the modification (target) and the snoRNA, respectively. Both scores measure how

much better s fits the target t compared to the predicted alternatives for which an interaction is also feasible. Large values of $ICI(t,s) \geq 1$ suggest that t is consistently a target of snoRNA family s. The parameter $ICI_{mod}(t,s)$ emphasizes the conservation of the modification site, while $ICI_{sno}(t,s)$ emphasizes the conservation of the snoRNAs ASE. By design this score should be applicable to other RNA-RNA interactions, e.g to miRNA targets in mRNAs. Prerequisite however are reliable alignments of the ncRNAs and the target RNAs.

3.3 SnoRNA Analysis Workflow

A workflow follows from the consecutive, coordinated application of the tools to the data described in the previous sections. It is suitable, to evolutionarily track snoRNA genes, especially also considering the conservation of their guiding function on a large scale. This unique method of analyzing snoRNAs in such detail was developed in cooperation with Sebastian Canzler and Jana Hertel. It expands the snoRNA interaction network by combining five main parts (Figure 3.3).
i) Starting point are snoRNA sequences that were collected from databases and literature in one or several species (here human, chicken, and platypus; compare Section 3.1). The snoRNAs are provided in fasta-format. The header of each sequence has to contain the box-annotation in a format defined by the `snoStrip` pipeline. These serve as queries for homology search with `snoStrip` retrieving a set of snoRNA sequences in the species investigated (here vertebrates).
ii) For all species the target RNA set is collected. The individual target RNA sequences in fasta format are available from databases and literature (see Section 3.1). Associated accessibility profiles that represent the internal structure of the target RNA sequences are computed with `RNAup` (Mückstein et al., 2006). Further, reliable alignments of the target RNAs are important to identify equivalent modifications in different species. In case of ribosomal RNAs, that contain both, regions of high and others of low conservation, an adequate alignment program is `RNAsalsa` (Stocsits et al., 2009). Manually curated snRNA alignments are provided in Marz et al. (2008).
iii) For each single snoRNA sequence of a snoRNA family an independent target prediction is performed considering thermodynamics of hybridization. With `RNAsnoop` (Tafer et al., 2010) (from `Vienna RNA Package` 1.7) prediction of box H/ACA snoRNAs targets is realized (Section 3.2.1), here the internal structure of the target RNA is relevant.

Target prediction for box C/D snoRNAs is performed with PLEXY (Section 3.2.2) (Kehr et al., 2011). This is done without considering accessibility information of the target, since prediction accuracy of box C/D snoRNA targets declines otherwise. As result a set of individual target predictions in the investigated species is present.
iv) Additionally, available information on experimentally 2'-O-methylation and pseudouridylation sites in the target RNAs is collected from various databases and literature (see Section 3.1). All positions in the individual species are mapped to the according target RNA alignment columns.
v) Last, the set of conserved targets is predicted. Therefore, the individual target predictions are evaluated using both types of ICI scores (Section 3.2.3), as such combining the previous data. On the one hand the snoRNA families with the best fitting ASE to a known modified nucleotide is determined with the ICI_{mod}. On the other hand for each snoRNA family the best scoring interaction within the target RNAs is identified with the ICI_{sno}. The assignments between snoRNA families and modification sites consider both, the free energy of the interaction in the single organisms and the level of its conservation across the considered species.
Finally, the data gained in steps i)-v) is connected in several ways to add new pieces to the puzzle of snoRNA-target RNA interactions. More precisely:

1. Each snoRNA can be classified as double guide, single guide or orphan, depending on whether a conserved target exists for both, one or none of the ASE in the snoRNA sequence.

2. The information that two snoRNA families target the same modification can be used to unravel distant homology. These are not always identifiable on sequence similarity alone.

3. For orphan snoRNAs without assigned function high scoring conserved targets can be identified.

4. For reported modifications that lack a recognized snoRNA guide high scoring snoRNA families can be identified.

5. Matches between 3. and 4. reveal interactions between unknown modifications and orphan snoRNAs.

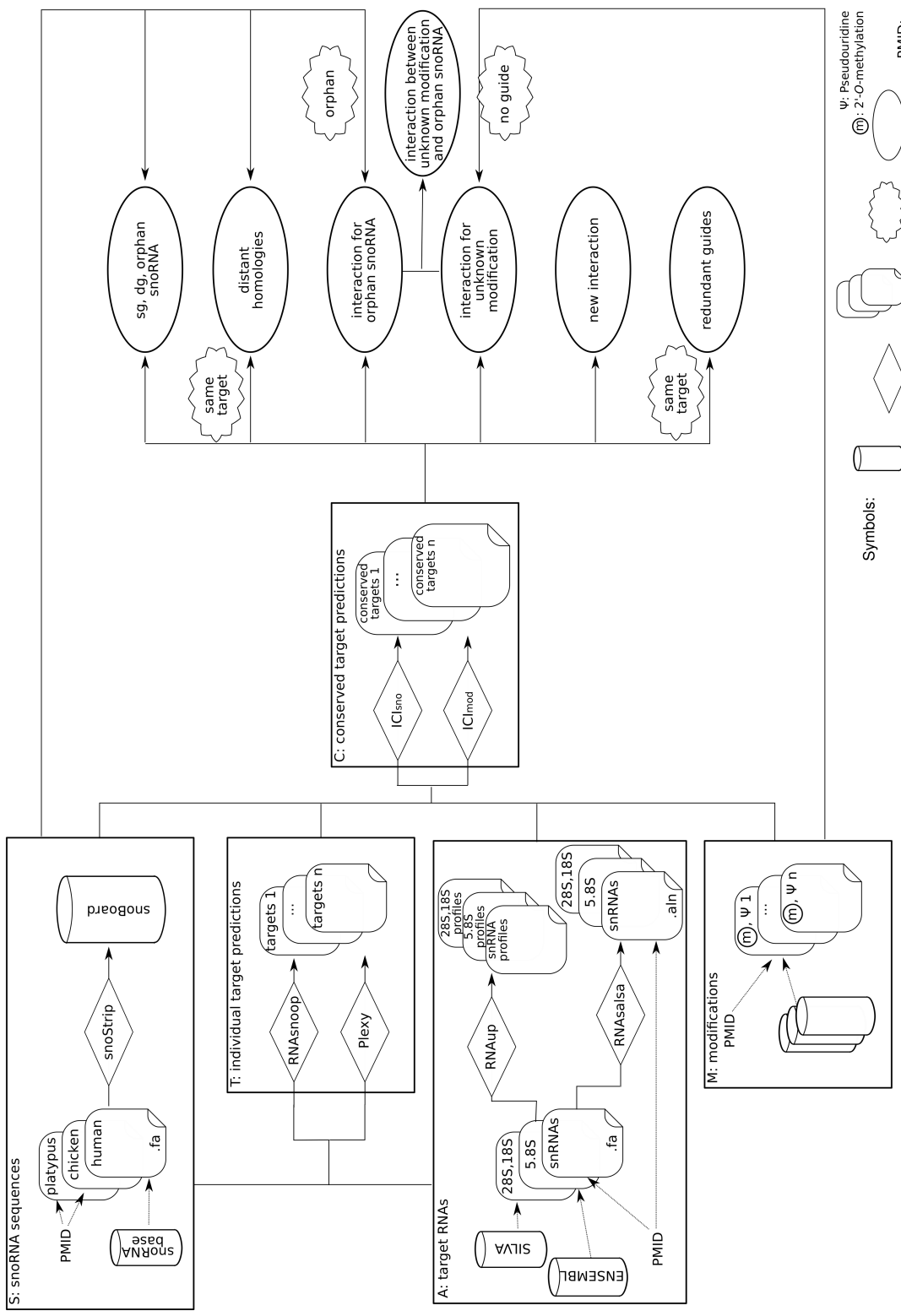

Figure 3.3: Schematic representation of the workflow to expand the snoRNA interaction network. Details are provided in the text.

6. From the set of conserved target predictions also complete new interactions between snoRNAs and target RNAs can be recognized. These might reflect modifications that have not yet been observed in experiments, e.g. because of tissue-, environmental condition-, or species- specific occurrence of the according modification.

7. Several Ψs and methylated nucleotides are redundantly guided, i.e. have more than one putative snoRNA that can guide the core proteins to the according position. These alternative guides can be derived from overlapping target predictions, that do not show signs of homology (compare item 2.).

Thus the workflow combines our target analysis tools in order to add new snoRNA families, merge homologous families and connect snoRNA guides and modification sites. No comparable approach to expand the snoRNA interaction network has been published. In this study the workflow was applied to vertebrates. However, it is not limited to certain species and can also be used to study the snoRNA-target RNA interactions also in completely different phyla. It was used by Sebastian Canzler to study the snoRNA interaction network in fungi (Canzler, 2016).

CHAPTER 4

SnoRNA Screens

In the last years the `snoStrip` pipeline was used to annotate snoRNAs in a plethora of species. To gain a comprehensive set of snoRNAs for further analysis it was used to annotate snoRNAs in 47 vertebrate species. As part of several ncRNA annotation projects it was applied to newly assembled genomes. The main contribution to these projects was to annotate the set of snoRNAs in the species of interest. These were pig, duck (both not shown here), the spotted gar and 48 avian genomes. Depending on the projects focus some extra analyses were done, e.g conservation analysis or host gene annotation. Additionally, `snoStrip` was used to annotate homologous sequences in 24 plant genomes, obviously starting with a plant snoRNA query set. In the following sections the studies are briefly introduced and summarized. The presented studies are published in different peer-reviewed journals: *PLOS ONE, Nature Genetics, and BMC Genomics* (Gardner et al., 2015; Braasch et al., 2016; Patra Bhattacharya et al., 2016)

4.1 Vertebrate snoRNA dataset

Although snoRNAs are a numerous and ancient class of ncRNAs a comprehensive set of snoRNAs across vertebrate species was still missing when the work was started. However, highly confident and partly experimentally verified snoRNA sequences were available chicken (Shao et al., 2009), platypus (Schmitz et al., 2008) and human (Lestrade and Weber, 2006). These were used as queryset for the initial `snoStrip` (Bartschat et al., 2014) run. For each snoRNA family homologous sequences were searched in related organisms following phylogenetic relationships. The snoRNA sequences newly detected in each species were added to the queryset for homology search in the next species. Any sequence that was identified could also be a missing link to identify a homolog in a related but already analyzed species. Thus, it is necessary to apply `snoStrip` in an iterative manner until no new sequence is annotated.

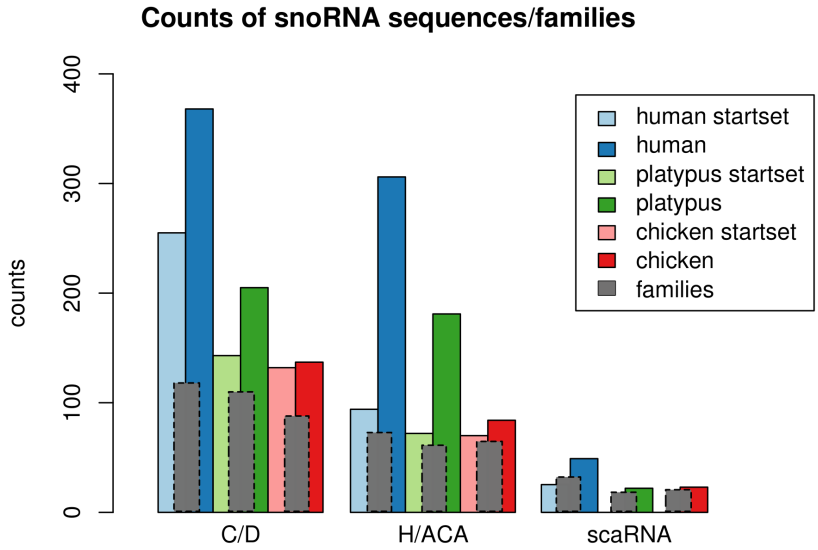

Figure 4.1: SnoRNA sequence counts before (lighter colors) and after (darker color) application of `snoStrip` to 47 vertebrate genomes. Resulting family counts are shown as gray bars.

Through application of `snoStrip` a dataset containing 259 snoRNA families in 47 vertebrate species, including a total of 22,061 single snoRNA sequences was compiled. This set includes 723 human snoRNA sequences (Figure 4.1). A list of investigated species and their phylogenetic tree is provided in Appendix Table A.1 and Figure A.4. As is generally true for ncRNA sequences also the set of snoRNA sequences is biased

towards model organisms (Hoeppner and Poole, 2012; Gardner et al., 2010). In this particular case these are the species that were used as initial dataset (human, chicken and platypus). It is beyond the scope of homology based approaches to detect putative novel species or group specific snoRNAs e.g. in carnivores.

The heatmaps in Figure 4.2 - Figure 4.3 show the numbers of paralogous per snoRNA sequences in each snoRNA family. Most snoRNAs homologous sequences can be identified in all amniotes. Several families are additionally conserved in Teleostomi, suggesting abundance of these snoRNA families in the common ancestor of vertebrates. A few more recent inventions of box C/D snoRNAs and scaRNAs seem to date back to common ancestor of mammals. Single white gaps (so undetected snoRNA sequences in single species) are probably due to incomplete genome assemblies and limitations in homology search rather than representing genuine gene losses. Some families that originate form the chicken startset and the platypus startset appear as avian or even platypus specific.

In the case of box C/D snoRNAs (Figure 4.2) the snoRNAs are usually abundant in moderate (mostly one, several two) copies. A few exceptions are the multi-copy families SNORD113-116 and SNORD13, which cluster at the bottom of the heatmap. These are also specific to eutherians. The pattern for scaRNAs (Figure 4.4) is similar.

Box box H/ACA snoRNAs generally have a higher amount of paralogs per family (Figure 4.3) than observed in box C/D snoRNAs. Considering the recently detected AluACAs (Jády et al., 2012), this type of snoRNAs seemingly share certain repetitive elements. The frequently increased copy numbers in *D. novemcinctus* and *C. hoffmanni* however probably represent false positive predictions, which might also be caused by poor assemblies.

The alignments that were computed with muscle in the last `snoStrip` step were manually curated in `RALEE` mode (Griffiths-Jones, 2005). With the .stockholm format (Section 1.3) it was taken advantage of the possibility to add customized annotation lines. The standard example is annotation of the secondary structure consensus (#=GC SS_cons). For snoRNAs additionally annotation of characteristic box motifs (#=GC Anno), the kink-turn (#=GC cons_kink), and occasionally target regions and predictions (#=GC target) were added. Based on the vertebrate snoRNA dataset a large scale study on conservation of snoRNA-target interactions was performed (Chapter 6). The set, including alignments and fasta-files is published as part of that study (Kehr et al., 2014): www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/12-022.

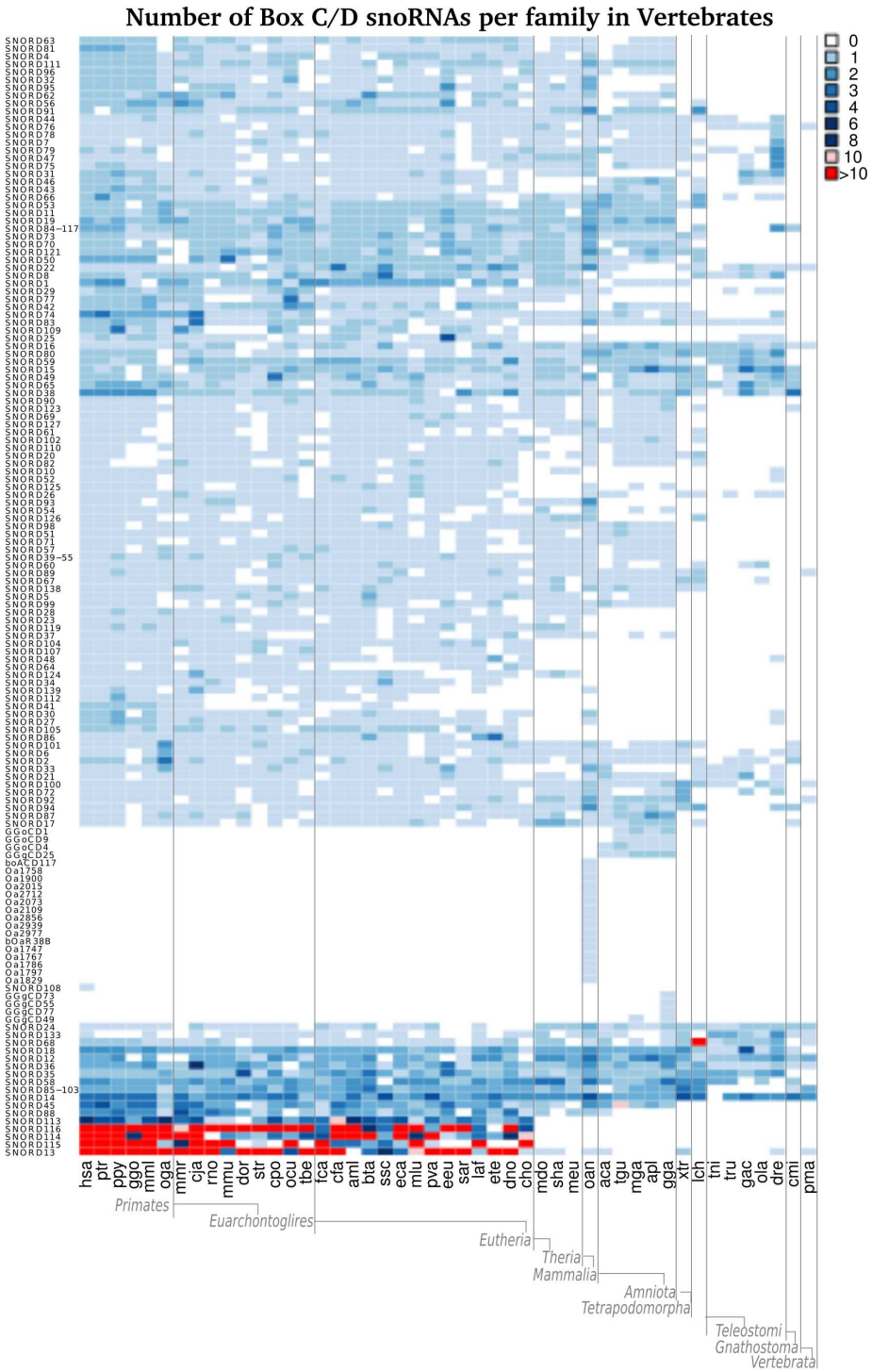

Figure 4.2: The heatmap represents the distribution of box C/D snoRNA families (rows) among vertebrate species (columns). (Species abbreviations are provided in Appendix Table A.1.) The color code reflects the number of paralogs found in the according family in the investigated vertebrate genomes. It ranges from light blue for one paralog to red for more than 10 paralogs. White cells mark species where no member of the snoRNA family has been identified.

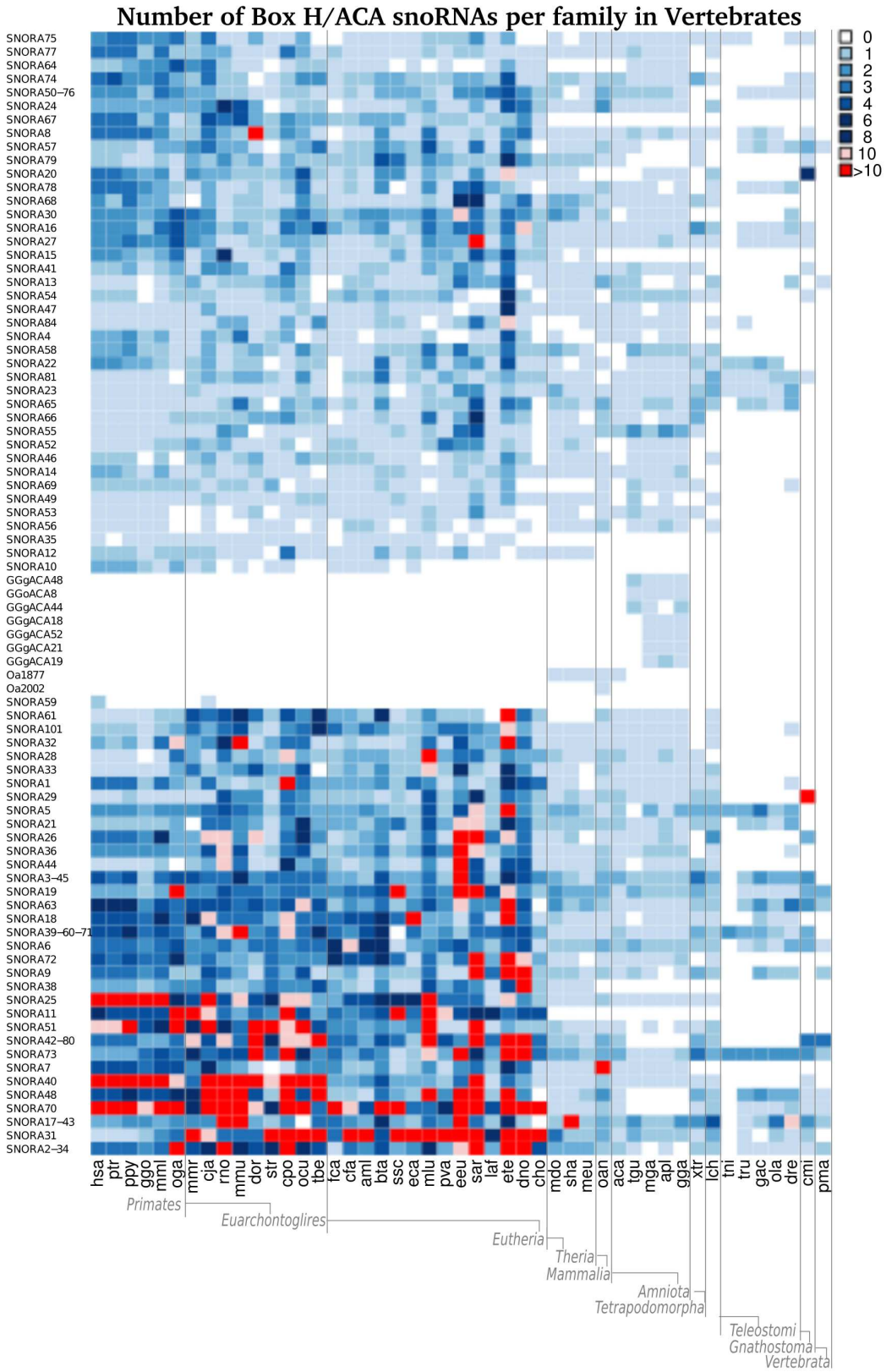

Figure 4.3: Heatmap representing paralog number of box H/ACA snoRNAs in investigated vertebrate genomes. Description equivalent to Figure 4.2

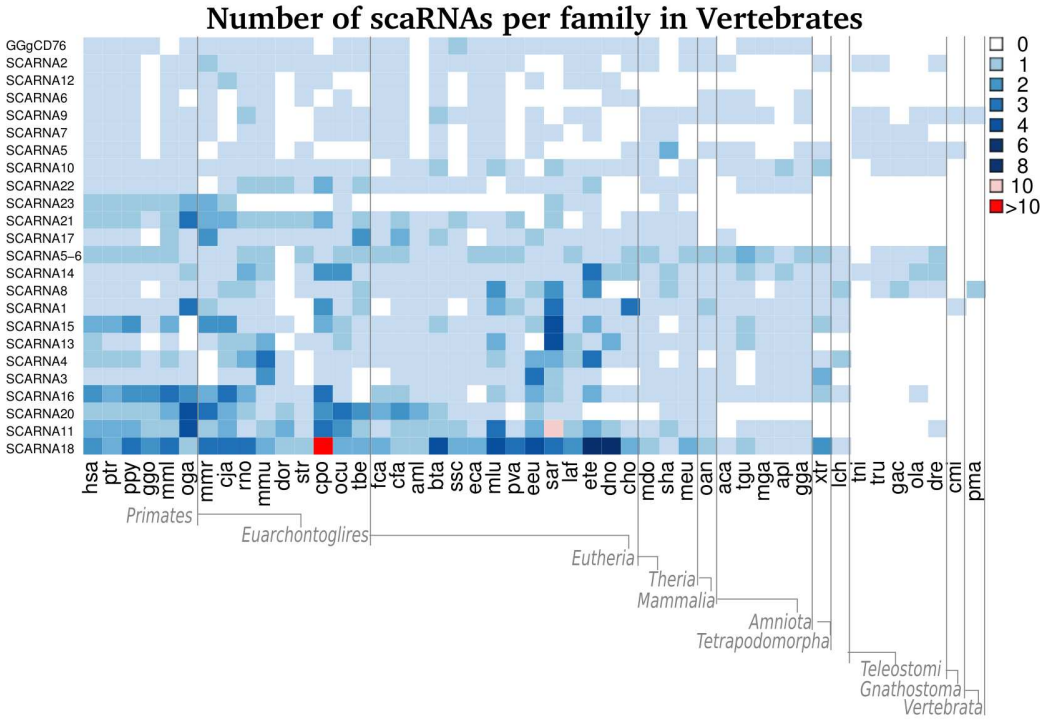

Figure 4.4: Heatmap representing paralog number of scaRNAs in investigated vertebrate genomes. Description equivalent to Figure 4.2

4.2 Conservation and Losses of non-coding RNAs in Avian Genomes

A collaborative project realized large-scale non-coding RNA annotation in 48 avian genomes (Gardner et al., 2015). The bioinformatic approach for homology based gene annotation combined probabilistic models of ncRNA families from the RFAM database for a infernal search, with class-specific models and search tools. Also in this projects `snoStrip` was used to infer the snoRNA sequences. Subsequently the gene annotations were validated by RNA-seq data. The 48 avian genomes (Appendix Figure A.5) contain only three model birds with chromosome assemblies (zebra finch (Warren et al., 2010), chicken (International Chicken Genome Sequencing Consortium, 2004) and turkey (Dalloul et al., 2010)), while the other 45 are recently published predominantly non-model avian species (Jarvis et al., 2014; Zhang et al., 2014; Huang et al., 2013; Zhan et al., 2013; Shapiro et al., 2013; Ganapathy et al., 2014). As outgroups the alligator and turtle are included.

Compared to mammals birds have an overall reduced genome size. Their karyotype is characterized by a large number of chromosomes (average $2n \approx 80$) generally consisting of approximately 5 larger macrochromosomes and many smaller microchromosomes (Griffin et al., 2007; Solinhac et al., 2010; Douaud et al., 2008). The presence of microchromosomes presents significant assembly challenges (International Chicken Genome Sequencing Consortium, 2004; Dalloul et al., 2010; Ellegren, 2005). Indeed, of the 48 published genomes, 20 of which are high-coverage ($> 50X$), only two had relatively complete chromosomal assemblies when this study was initiated (chicken, zebra finch; (Warren et al., 2010; Zhang et al., 2014)).

In total 626 different ncRNA families were annotated. The snoRNA findings are summarized in Table 4.1. Most of the snoRNAs are present with stable copy numbers across all bird genomes.

There are 59 snoRNA families that guide corresponding ribosomal modification sites, between human and yeast (Lestrade and Weber, 2006). Of these ancient snoRNA families 45 are conserved in the bird data set. Most of the apparent losses cluster on two loci of the ancestral vertebrate genome.

The first cluster corresponds to a human cluster at host gene SNHG1 which contains a total of eight C/D box snoRNAs: SNORD25, SNORD26, SNORD27, SNORD28,

Table 4.1: Summary of snoRNA sequences annotated in 48 avian genomes.

Human	median birds	chicken	confirmed	type
281	120.0	106	90 (84.9%)	box C/D snoRNAs
336	85.5	68	48 (70.6%)	box H/ACA snoRNAs
34	13.0	12	12 (100%)	ScaRNAs
7340	1080.0	1194	865 (72.4%)	Total

SNORD29, SNORD22, SNORD30 and SNORD31 (Tycowski et al., 1996). Each of which are also found in the alligator and turtle genomes within a 34 KB locus, yet these have partly been lost in birds. However, five of the eight snoRNAs are located in the tinamou genome. These are encoded on the same scaffold and are within 2 KB of each other. This implies that SNHG1 is conserved in the tinamou. Loci with four of the eight snoRNAs can be found in zebra finch, ground-finch, and bald eagle. Still, three of the eight are located in the ostrich, crow, and cuckoo genomes, again within 2 KB of each other on the same scaffolds. Also mousebird, duck and rifleman have three of the eight snoRNA genes, but they are observed in different loci. In the remaining bird genomes only two, one, or even none of the snoRNAs was detected. The most conserved snoRNA in the cluster is SNORD27. The complex pattern of loss could be attributed to several different models, e.g. multiple losses in birds, poor homology modeling or incomplete genome sequences.

The second cluster is encoded in an intron of ribosomal protein L13a in human. It contains two copies of SNORD33 and one SNORD34 sequence within a 1 KB genomic region. The turtle and alligator genomes retain the two copies of SNORD33 yet do not have an obvious SNORD34 gene on the same scaffold. Within the bird genomes, the crow and rifleman each retain a single SNORD33 and SNORD34 gene on the same scaffold. The ground-finch and bald eagle have a single SNORD33 and the zebra finch and seriema a single SNORD34 copy. Interestingly, a protein `BLAST` (version 2.2.18) searches for the RPL13A gene, recovered the protein conserved in the alligator and turtle genomes as well as in the bald eagle, crow, rifleman and zebra finch (data not shown). Therefore, the RPL13A gene and corresponding intronic snoRNAs show the same conservation. This supports a pattern of coherent loss of the RPL13A gene and the intronic snoRNAs that it hosts in the bird genomes.

There are three causes for the observed losses of the other ancient snoRNA families

(this is true also for all other observed losses of ncRNA genes):

1. A genuine gene loss in the avian lineage. Due to a smaller genome size of birds compared to mammals also a lower number of ncRNA genes is expected.

2. High divergence of ncRNA that results in undetectable homology due to significant sequence and structure alteration. In general when sequence similarity drops below 60%-50% alignment tools fail.

3. Unsequenced and/or unassembles microchromosomes might harbor a high fraction of undetected ncRNA genes.

Indeed a Fisher's exact test showed that significantly more missing ncRNAs are located on microchromosomes in chicken than on macrochromosomes, ($P < 10e - 16$). Thus, it is suggested that many of the ancient ncRNAs families are missing because they are predominantly found on microchromosomes and the vast majority of avian microchromosomes remain unsequenced (Zhang et al., 2014; Ellegren, 2005).
Subsequent validation of the ncRNA annotation with RNA-seq delivers expression signals above background levels for 865 (72.4%) genes (see Table 4.1). Particularly, 100%, 85%, and 70% of annotated scaRNAs, box C/D snoRNAs and box H/ACA snoRNAs are validated with RNA-seq in chicken. This percentage is a lower bound of the actually expressed RNAs, as only a fraction of the developmental stages and tissues of chicken have been characterized with RNA-seq. It is known for some ncRNAs that they are expressed in highly specific conditions (Mercer et al., 2008; Johnston and Hobert, 2003).

4.3 SnoRNAs in the Spotted Gar

The spotted gar (*L. oculatus*) diverged from Teleost fishes before the Teleost genome duplication (TGD). It can be viewed as evolutionary bridge between Teleosts and other vertebrates, including human. Within the project a new genome assembly was established and the genes were annotated. For ncRNAs a general `Infernal` search with RFAM family models was combined with ncRNA-class specific methods (Braasch et al., 2016). Again our `snoStrip` pipeline was used to annotate the snoRNA ensemble. The gar karyotype ($2n = 58$) contains both macro- and microchromosomes. Gar is the first ray-finned fish genome sequence not affected by the TGD. Strikingly, almost half of

the gar karyotype (14/29 chromosomes) showed a nearly one-to-one relationship in gar-chicken comparisons, including macro- and micro chromosomes with highly correlated chromosome assembly lengths. This similarity in chromosome size and gene content is a strong evidence that the karyotype of the common bony vertebrate ancestor of gar and chicken possessed both macro- and microchromosomes as Ohno et al. (1969) hypothesized. Of the 99 snoRNA families, for which 242 homologous sequences were annotated, 90% are also conserved in human, 75% in chicken, and 58% in zebra fish.

4.4 Phylogenetic Distribution of Plant snoRNA Families

Although there is good evidence for the conservation of many of the chemical modification sites on rRNAs and snRNAs between eukaryotic kingdoms (Lestrade and Weber, 2006), it has remained an open question to what extent individual snoRNA families are homologous at large phylogenetic distances. This is difficult to address since snoRNA sequences evolve rapidly apart from the conserved boxes and the antisense region. To tackle this question it is necessary to first understand in detail the evolutionary patterns of snoRNAs within each kingdom and to distinguish those snoRNA families that may have originated in the eukaryotic ancestor from those that are more recent innovations. In plants, snoRNA evolution has attracted comparably little attention. In a recent survey submitted to *BMC genomics* the evolutionary history of snoRNAs in the plant kingdom is now addressed (Patra Bhattacharya et al., 2016).

As `snoStrip` is not limited to a certain kingdom, even the screening for snoRNAs across the plant kingdom was realized through application of the pipeline. The initial query set of 554 snoRNA genes was collected from available (plant) snoRNA databases (Brown et al., 2003; Yoshihama et al., 2013). These sequences were assigned to 222 box C/D and 74 box H/ACA snoRNA families.

The search was applied to 24 sequenced plant species roughly but evenly covering the plant kingdom. The resulting plant snoRNA set comprises a total of 5670 sequences in 302 snoRNA families, Many of the plant snoRNA families comprise multiple paralogs. The sequence conservation of snoRNAs is sufficient to establish homologies between phyla. The signal tapers off, however, between land plants and algae. Also the snoRNA targets are found well-conserved for most snoRNA families. Plant snoRNAs are fre-

quently organized in highly conserved clusters.
The study provides the most comprehensive collection of snoRNAs in plants. It is a valuable resource for more detailed studies on snoRNAs and their evolution in the plant kingdom and to identify the ancient snoRNA core.

The frequent use of `snoStrip` in genome annotation projects (Huang et al., 2013; Anthon et al., 2014; Gardner et al., 2015; Braasch et al., 2016) prooved that this tool is useful. It enables to track hotspots of snoRNA inventions and losses during evolution. Beside pure snoRNA locations it also extracts snoRNA features, which can subsequently be used to derive general snoRNA models e.g. for automated snoRNA *de novo* searches with SVM based classification. It can now be regarded as state-of-the-art method to reliably annotate homologous snoRNA sequences. It is not limited to specific phyla. Besides, the generation of a comprehensive snoRNA annotation in vertebrates it has already been used to identified snoRNA genes in in fungi (Canzler, 2016) and plants (Patra Bhattacharya et al., 2016).

CHAPTER 5

Exceptional SnoRNAs

During the work with snoRNAs several sequences have been noticed that exhibit many of the characteristic snoRNA elements but feature unexpected deviations of the norm. These encompass exceptional length, deviant secondary structures, or hybrid architectures and are mainly but not exclusively assigned to the scaRNA subclass. Their careful investigation included conservation in metazoan animals, analysis of their structure, their synteny and *in silico* function analysis. This chapter presents the findings concerning the snRNAs that are composed of two snoRNA domains. These and exceptionally structured box C/D snoRNAs and box H/ACA snoRNAs have been published in *RNA Biology* (Marz et al., 2011). They are supplemented with uncommon scaRNA families, and further details that came up after the publication of the paper in studies with (Machyna et al., 2014) and (Jorjani et al., 2016).

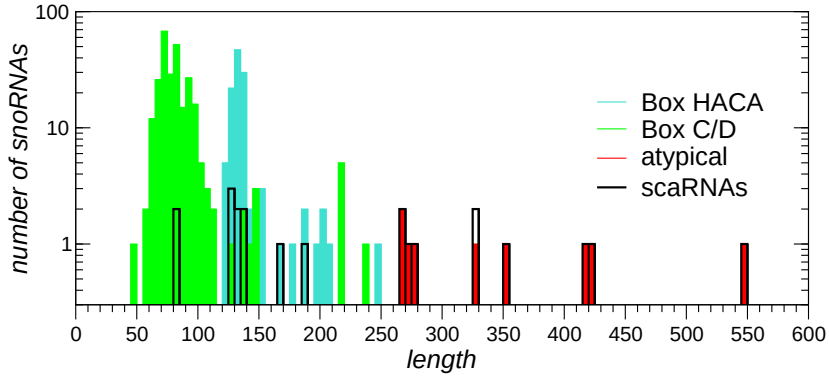

Figure 5.1: Length distribution of human snoRNAs taken from `snoRNA-LBME-db` *(Lestrade and Weber, 2006). Box C/D, box H/ACA, and snoRNAs with atypical architecture (e.g. those with both a C/D and a H/ACA domain) are shown by different colors. The scaRNAs, characterized by an additional localization signal, belong to either one of these three classes. They are marked by the black bars.*

5.1 Atypical Length and Composite Structures

While length distribution of box C/D snoRNAs respective box H/ACA snoRNAs is quite uniform the length of scaRNAs varies significantly. Figure 5.1 shows the length distribution of human snoRNAs taken from `snoRNA-LBME-db` (Lestrade and Weber, 2006). Typical box C/D snoRNAs have a length of 80 ± 11 nt. The average length of box H/ACA snoRNAs is 134 ± 7. In contrast to both the length of scaRNAs is more variant, ranging from 80 to 550. Besides scaRNAs that have regular box C/D structure, or regular box H/ACA structure, several have exceptional composite structures consisting of two box C/D domains, two box H/ACA domains, or combine a box C/D and a box H/ACA domain. A further special case is telomerase RNA (TERC) which is not considered here.

In following eight exceptional scaRNA families with composite structure are reviewed. The focus is drawn on their structural peculiarities and evolutionary conservation, including also inspection of their synteny. Exceptionally long box C/D and H/ACA structured snoRNAs are discussed in the paper (Marz et al., 2011) and in (Canzler, 2016).

The selected snoRNA sequences were retrieved from `snoRNA-LBME-db` (Lestrade and Weber, 2006), and `Rfam` (v.9.1 and v.10.0, seed sequences). This study predates the development of the `snoStrip` annotation pipeline. The comparative analysis, struc-

ture predictions, alignment construction and host gene annotation was performed as follows. First, iterative `Blast` searches (Altschul et al., 1997) in 105 downloadable animal genomes were conducted. After generating alignments (with `clustalw` (Thompson et al., 2002), `cmalign` (Nawrocki et al., 2009), and `locarna` (Will et al., 2007)) and predicting consensus secondary structures (with `RNAalifold` (Bernhart et al., 2008)), `infernal` (v.1.0) (Nawrocki et al., 2009) was used to construct and calibrate covariance models. These were used to search with a combined sequence-structure approach in those genomes for which the purely sequence based approaches have remained unsuccessful. Homologous snoRNA candidates were added to the alignments and evaluated. The final structure-annotated alignments were manually refined using the `emacs` editor in `RALEE` mode (Griffiths-Jones, 2005) and `RNAalifold` (Bernhart et al., 2008) for structure prediction. SnoRNA host gene sequences were retrieved from `ENSEMBL` genome browser (Spudich and Fernández-Suárez, 2010). The sequences were aligned with `clustalw` and possible homology of protein-coding host genes with different names was verified. In addition, we searched for the homologous proteins in genomic sequences to determine the relationships between host genes. Furthermore, the results of the protein search step on local genome versions were used to identify putative locations of more divergent snoRNAs. In order to assess distant homologies between snoRNA families the similarities between the covariance models of the families were scored with `cmcompare` (Höner zu Siederdissen and Hofacker, 2010). Since snoRNAs of the same class by definition have similar secondary structures, a randomized set of sequences using `RNAinverse` (Hofacker et al., 1994) was generated and the bit score distributions were compared.

5.2 Composite structured scaRNAs

Several scaRNAs are composed of two complete snoRNA domains. They either are of the same type or comprise both a box C/D and a box H/ACA domain. For details see Section 2.2 and Figure 2.4.

SCARNA9 (mgU2-19/30) This scaRNA is composed of two box C/D domains with predicted targets in the U2 snRNA (nucleotides G19 and A30) (Tycowski et al., 2004). The two domains are separated by a G/U-rich linker. The full-length molecule appears to localize to the Cajal body. The two component snoRNAs, designated mgU2-

Table 5.1: Phylogenetic distribution and host genes of six composite scaRNAs. Numbers of animals that contain the snoRNAs are listed. The type of parentheses indicates association with the host genes listed in the last row of the table. The symbol ♢ refers to a non-coding transcript located between or adjacent to gene(s) listed in the parentheses. Numbers without parentheses refer to species where no host gene was determined. Horizontal lines indicate the phylogenetic range of previously reported sequences in `snoRNA-LBME-db` *(Lestrade and Weber, 2006).*

	Dual scaRNA RNAs					
Organisms	scaRNA9	scaRNA12	scaRNA6	scaRNA5	scaRNA10	scaRNA13
Primates	[5],1	[7]	[5]	[5]	[8]	[5]
Euarchontoglires	[5]	[5]	[4]	[6]	[7]	[6]
Laurasiatheria	[5],(1),{1}	[6]	[5]	[7]	[8-1]	[5]
Afrotheria	[2]	[2]	[1]	[1]	[2]	[2]
Xenarthra	[1]	[1]	[1]	[2]	[2]	[1]
M. domestica	[1]	[1]	[1]	[1]	[1]	[1]
O. anatinus	[1]	–	[1]	[1]	–	[1]
A. carolinensis	1	–	[1]	[1]	[1]	[1]
Aves	[3]	–	[2]	[2]	[2]	[3-1]
X. tropicalis	–	–	[1]	[1]	[2-1]	–
Teleostei	[2],2	–	[2]	[6-1]	[5]	[5]
C. milii	1	–	[1]	[1]	–	–
P. marinus	1	–	–	–	[1]	–
B. floridae	–	–	–	–	[1]	–
Tunicata	–	–	–	–	–	–
S. purpuratus	–	–	–	–	–	–
S. kowalevskii	–	–	–	–	–	–
Drosophila	–	–	–	–	(12)	–
Panarthropoda	–	–	–	–	6	–
P. humanus	–	–	–	–	–	–
D. pulex	–	–	–	–	–	–
H. robusta	–	–	–	–	–	–
L. gigantea	–	–	–	–	–	–
C. capitata	–	–	–	–	1	–
A. californica	–	–	–	–	1	–
N. vectensis	–	–	–	–	–	–
R. spez	–	–	–	–	–	–
T. adhaerens	–	–	–	–	–	–
Host Genes	[KIAA1731] (PLRG1) {SETD1B}	[PHB2]	[APG16L1]	[APG16L1]	[NCAPD2] (CG1142)	♢ [GLRX5]

Primates: *H. sapiens*, *P. troglodytes*, *P. pygmaeus*, *M. mulatta*, *C. jacchus*, *O. garnettii*, *M. murinus*; **Euarchontoglires**: *M. musculus*, *R. norvegicus*, *S. tridecemlineatus*, *C. porcellus*, *O. princeps*, *O. cuniculus*, *T.belangeri*; **Laurasiatheria**: *F. catus*, *C. familiaris*, *B. taurus*, *S. scrofa*, *E. caballus*, *M. lucifugus*, *E. europaeus*, *S. araneus*; **Afrotheria**: *L. africana*, *E. telfairi*; **Xenarthra**: *D. novemcinctus*, *C. hoffmanni*; **Aves**: *T. guttata*, *G. gallus*, *M. gallopavo*; **Teleostei**: *D. rerio*, *T. nigroviridis*, *T. rubripes*, *G. aculeatus*, *O. latipes*; **Tunicata**: *O. dioica*, *C. intestinalis*, *C. savignyi*; **Drosophila**: *D. pseudoobscura*, *D. yakuba*, *D. melanogaster*, *D. erecta*, *D. simulans*, *D. sechellia*, *D. grimshawi*, *D. mojavensis*, *D. persimilis*, *D. virilis*, *D. ananassae*; **Panarthropoda**: *G. mositans*, *A. aegypti*, *A. gambiae*, *N. vitripennis*, *C. quinquefasciatus*, *T. castaneum*;

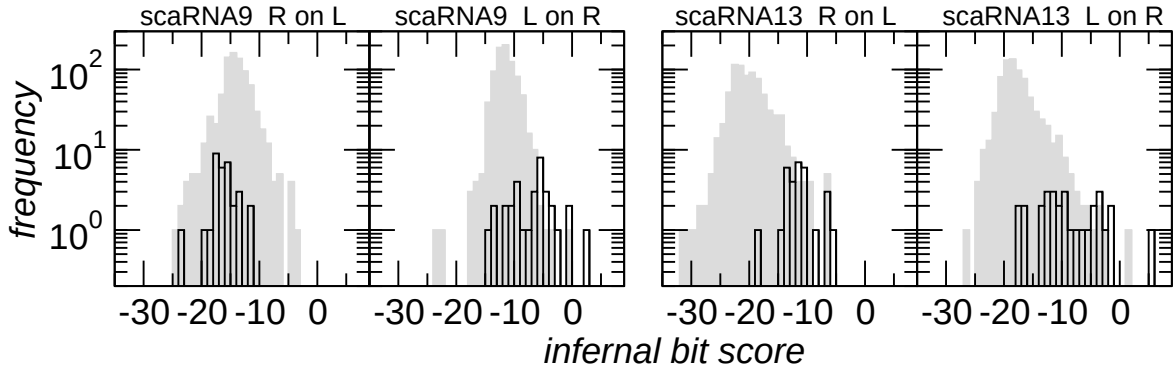

Figure 5.2: Distribution of `infernal` *bit scores of the sequence of the 5' component aligned to the covariance model of the 3' component (R on L) and* vice versa *for SCARNA9 and SCARNA13. The background distribution (randomized sequences) is shown in gray. While there is no indication that the 5' and 3' components are related for SCARNA9, the shift of the bit score distributions towards higher values for SCARNA13 shows that the sequences of the two parts (L and R) of this snoRNA are more similar than expected. Although this does not constitute an iron-clad proof, it serves at least as a strong indication that L and R are homologs and likely arose through a tandem duplication event.*

19 and mgU2-30, also exist as separate entities that localized to the nucleolus (Tycowski et al., 2004). A mouse homolog of mgU2-30 has been reported as Z32 snoRNA in GenBank entry ***AJ242789***. The SCARNA9 sequence is quite well conserved and can be found in all vertebrates. In tetrapods and some teleosts it is associated with KIAA1731, a conserved protein of unknown function. In zebra fish, GBAS serves as host gene for the SCARNA9 homolog. In *M. lucifugus* and *E. europaeus* SCARNA9 is associated with the proteins PLRG1 and SETD1B, respectively. A comparison of the two box C/D components shows no evidence that they might have arisen by tandem duplication, (Figure 5.2).

SCARNA13 (U93) SCARNA13 has been shown to co-localize with coilin in Cajal bodies and to guide the pseudouridylation of residue 54 in the U2 spliceosomal snRNA and of residue 53 in snRNA U5 (human coordinates) (Kiss et al., 2002; Schattner et al., 2006). The pseudouridylation of both positions was experimentally validated in human, mouse, and cow (Kiss et al., 2002). The scaRNA consists of two tandemly arranged, otherwise inconspicuous, box H/ACA domains. The sequence comprises a total of four hairpins. With a size of 252 nt (*Tetraodon*) to 274 nt (*Echinops*) it perfectly matches the expectation for a duplicated box H/ACA snoRNA. In Tetrapoda and teleosts it is located in a poorly characterized non-coding host gene SNHG10, downstream of the

highly conserved GLRX5 gene. In contrast to SCARNA9 the two H/ACA components show signs of distant homology (Fig. 5.2), suggesting that this dual snoRNA arose from a tandem duplication of a canonical H/ACA snoRNA.

SCARNA12 (U89) SCARNA12 is composed of a box H/ACA domain and a box C/D domain and has been shown to localize to the Cajal bodies (Darzacq et al., 2002). It is predicted to modify nucleotide U46 in U5 snRNA. The length ranges from 235 nt in *Echinops* to 283 nt in *Monodelphis*. Since it was not possible to annotate the gene beyond Theria, SCARNA12 is suspect to be a recent innovation. All identified SCARNA12 homologs are located in an intron of prohibitin 2.

SCARNA10 (U85) SCARNA10 has a similar architecture as SCARNA12. In contrast to the latter, however, it is among the best-conserved snoRNAs, which could be traced through most of the metazoan phyla. Originally detected in human (Jády and Kiss, 2001), the 5' end of its mouse homolog was reported as mouse MBI-52 box C/D snoRNA in (Hüttenhofer et al., 2001). The cow "microRNA" *mir-2424-1* originates from the 3' end of the bovine SCARNA10 gene. The *Drosophila* homolog, snoRNA:MeU5-C46, was also described in (Jády and Kiss, 2001) and compared in detail to the human sequence. Mutagenesis studies showed that SCARNA10 contains two functional copies of the CAB box (Richard et al., 2003).
In deuterostomes, SCARNA10 is consistently encoded in an intron of the NCAPD2 gene. Its *Drosophila* homolog is located in an intron of CG1142, a gene of unknown function. However, based on sequence alignment CG1142 can be identified as a homolog of NCAPD2. In *C. elegans*, the box H/ACA snoRNA ΨCeU5-48 (Huang et al., 2007) (also known as CeN105 (Deng et al., 2006)), is predicted to guide the modification of the homologous position in the U5 snRNA. In (Huang et al., 2007), an evolutionary relation between ΨCeU5-48 and SCARNA10 as well as SCARNA12 was suggested. The discovery of complete SCARNA10 homologs in several lophotrochozoan taxa suggests, however, that SCARNA10 has lost its box C/D domain in nematodes.
The similarity of the unusual architectures of the SCARNA10 and SCARNA12 families suggests that they are ancient paralogs. In order to test this hypothesis, the `infernal` bit scores were computed for aligning members of the one family against the covariance model of the other family (Figure 5.3). In each case, the scores, which average above 0 are significantly larger than the expected score from a randomized background con-

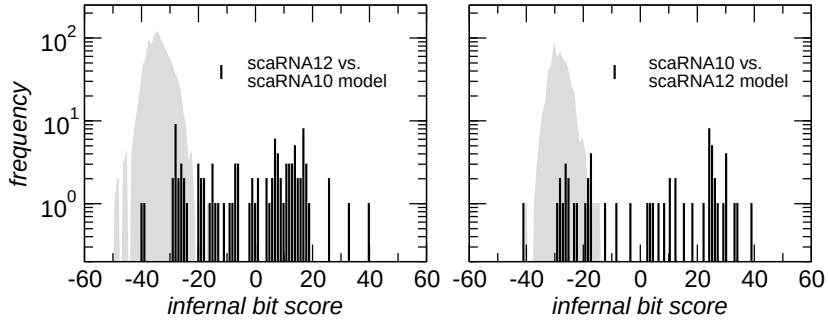

Figure 5.3: Distant homology of the SCARNA10 and SCARNA12 families. A comparison of `infernal` *bit scores for alignments of SCARNA12 against the SCARNA10 covariance model (l.h.s., black histogram), and* vice versa *(r.h.s., black histogram) shows that the sequences of one family fit much better to the model of the other family than random sequences fitting the same secondary structure (gray background).*

trol. This indicates that SCARNA10 and SCARNA12 are indeed paralogous snoRNA families. In contrast, there is no evidence for an evolutionary relation of the nematode CeN105 snoRNA and the H/ACA domain of either SCARNA10 or SCARNA12.

SCARNA5 (U87) and SCARNA6 (U88) These paralogous scaRNAs contain both a box C/D domain (targeting U5 snRNA, position U41 in human) and a H/ACA domain. Furthermore, the box C/D component of SCARNA6 guides methylation of snRNA U4 (human position A65). Both scaRNAs are encoded in distinct introns of the human ATG16L1 mRNA (Darzacq et al., 2002). The mouse RNA MBI-46 (Hüttenhofer et al., 2001) is the homolog of SCARNA6. In chicken, a 177 nt fragment of the SCARNA5 homolog (GGN31) was reported in (Zhang et al., 2009). The length of SCARNA5 ranges from 260 nt in *Takifugu* to 291 nt in *Monodelphis*, while SCARNA6 is slightly shorter, (225 nt *Canis* to 276 nt in *Monodelphis*). It is worth noting that SCARNA5 is frequently mis-annotated as SCARNA6 in the current release of `ENSEMBL`. An interesting peculiarity of both SCARNA5 and SCARNA6 is the absence of well-conserved target sequences. Possibly, the H/ACA domain, which – similar to SCARNA10 – contains conserved CAB boxes in its hairpin loops, only mediates transport to the Cajal body.

Figure 5.4 shows that SCARNA5 and SCARNA6 can be traced throughout gnathostome evolution. Although remaining in association with the host gene ATG16L1 (with the possible exception of *C. milii*), both RNAs have been relocated to different introns several times during vertebrate evolution.

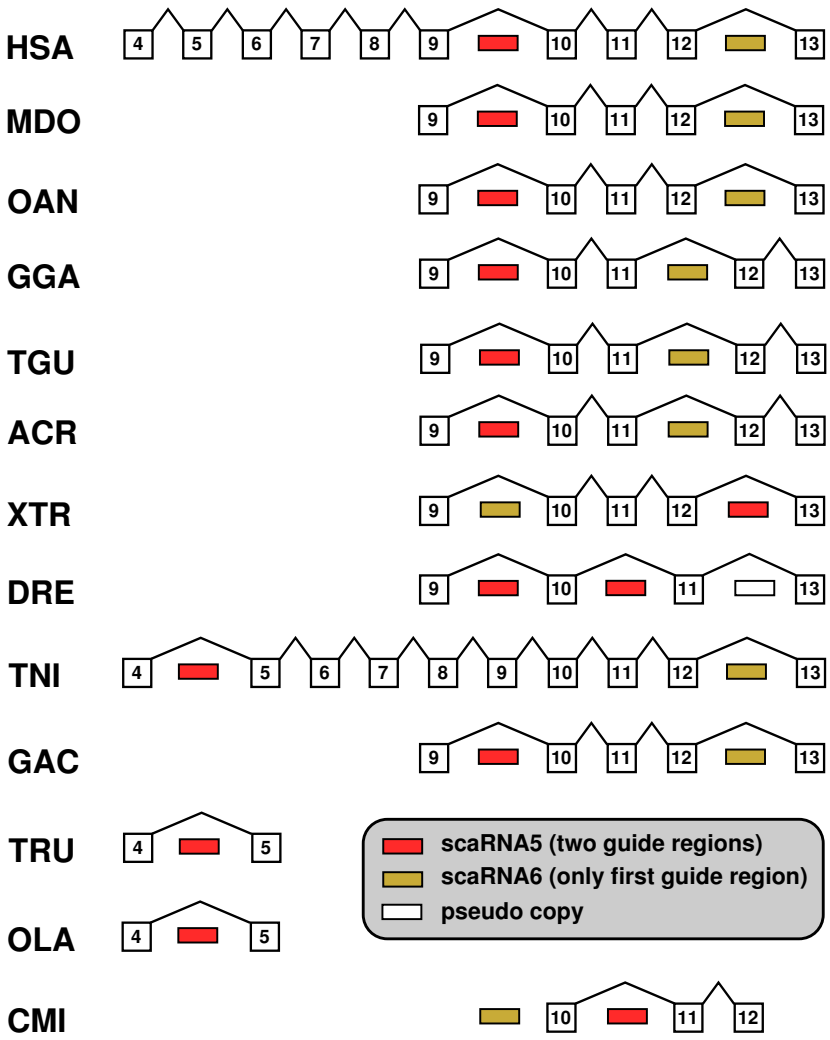

Figure 5.4: Location of the homologs of SCARNA5 and SCARNA6 in the ATG16L1 gene. Homology of introns was established by sequence alignments. The scaRNAs jumped to different positions several times during vertebrate evolution. Exons numbers correspond to the human gene. Species abbreviation are listed in (Appendix Table A.1)

SCARNA21 A further scaRNA with composite structure is SCARNA21. The scaRNA has two paralogs in mammals. A recent parCLIP study revealed that one of the previously unsuspicious 'common' H/ACA structured transcripts is indeed surrounded by a expressed box C/D domain. It is described in detail in Section 8.2.

SCARNA28 This scaRNA, was newly identified as GGgCD76 homolog in human by `snoStrip` reported as ZL1 in Kishore et al. (2013) and also identified with the coilin-iCLIP experiments. It has a box C/D structure with a GT-insert of differing length in vertebrates. It is described in more detail in Section 7.5.

In conclusion, a class of outstanding snoRNAs arose through fusion of two snoRNA genes. The four scaRNAs with C/D-H/ACA hybrid structures consist of a H/ACA component that is inserted into the loop of the C/D component. Two of these, SCARNA5 and SCARNA6 share the same host gene and are clearly paralogous. The other two examples, SCARNA10 and SCARNA12, come from different genomic locations. They still share enough sequence similarity to identify them as ancient paralogs. There is no evidence, on the other hand, that all four C/D-H/ACA hybrids share a common ancestor. SCARNA9, is a fusion of two box C/D snoRNAs that can be expressed as both a single and two separate molecules. No signs of homology between the two components could be observed. In contrast, SCARNA13, a fusion of two complete box H/ACA domains appears to be the product of a tandem duplication.
The expanded and manually curated alignments including consensus structures for all scaRNA families are a useful resource for genome annotation and further studies into snoRNA evolution alike.

CHAPTER 6

Matching of Soulmates: Co-evolution of SnoRNAs and their Targets

The modification patterns of ribosomal (rRNAs) and small nuclear (snRNAs) are retained during evolution making it even possible to project them from yeast onto human. The stringent conservation of modification sites and the slow evolution of rRNAs and snRNAs contradicts the rapid evolution of snoRNA sequences. To explain this discrepancy the co-evolution of snoRNAs and their targeted sites throughout vertebrates was investigated. With the vertebrate snoRNA data set, established with the `snoStrip` pipeline (Section 4.1), the collected target RNA set, including their modifications (Section 3.1) and the ICI score (Section 3.2.3) a systematic investigation was possible for the first time. In this study it was shown that functions of homologous snoRNAs in general are evolutionary stable, thus, members of the same snoRNA family guide equivalent modifications. The conservation of snoRNA sequences is high at target binding regions while the remaining sequence varies significantly. In addition to elucidating principles of correlated evolution it was possible, to assign functions to previously orphan snoRNAs and to associate snoRNAs as partners to known chemical modifications unassigned to a snoRNA guide before. Furthermore, the predictions of snoRNA functions in conjunction with sequence conservation were used to identify distant homologies. Due to the high overall entropy of snoRNA sequences, such relationships are hard to detect by means of sequence homology alone. The findings presented in this chapter have been published in *Molecular Biology and Evolution* (Kehr et al., 2014). The snoRNA alignments and further supplemental material is available at `http://www.bioinf.uni-leipzig.de/supplements/12-022`.

6.1 Conservation of Reported Guiding Function

Despite the emerging diversity of snoRNA function, here the attention was limited to the common modification tasks, pseudouridylation and 2'-O-methylation of ribosomal and small nuclear RNAs.

Starting from the vertebrate snoRNA dataset, first the targets were predicted for all single snoRNA sequences with `RNAsnoop` and `PLEXY` considering thermodynamic principles. The performance was evaluated by comparing recovery rates of the available human interactions. `RNAsnoop` returned 59 known human box H/ACA snoRNA interactions although it has been trained on yeast only. Deactivation of the yeast model and scoring predictions based on interaction energy only, yielded 86 of the 112 known interactions www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/12-022/mapping.html), including all those predicted with the yeast model. In contrast, the recovery of known human box C/D snoRNA targets performed better without the use of accessibility information. While considering the internal structure of the target RNA recovered 103 of the 115 known interactions, neglection of accessibility information recovered 111. This is in agreement with the observation that accessibility around methylated site does, in contrast to pseudouridylated residues, not significantly differ from the average accessibility of nucleotides in the ribosomal RNAs (see Section 3.2.2). All, predicted pseudouridylated and methylated positions were mapped to the corresponding columns of the target RNA alignments (Section 3.1 using `BioPerl`. For further details about the developed methods see Section 3.2.

Then, the conservation of the known human interactions (data retrieved from `snoRNA-LBME-db`) were investigated. Therefore, the individual target predictions from all investigated species were evaluated using the new ICI scores (Section 3.2.3 and Section 3.3).

Of the reported interactions 87% were recovered as conserved within vertebrate species with known snoRNA and target RNA sequences. In 18S rRNA, all 35 reported human interactions with box C/D snoRNAs are conserved at least in *Eutheria* (Figure 6.1). For 18S rRNA and box H/ACA snoRNAs, 31 of the 39 reported human interactions were found to be evolutionarily conserved at least in mammals. Five Ψs in 18S rRNA are reported to have two matching box H/ACA guides. Conserved function of both guides was predicted only in one case. For the remaining doubly guided modifications only a single conserved snoRNA guide (Figure 6.2) was identified.

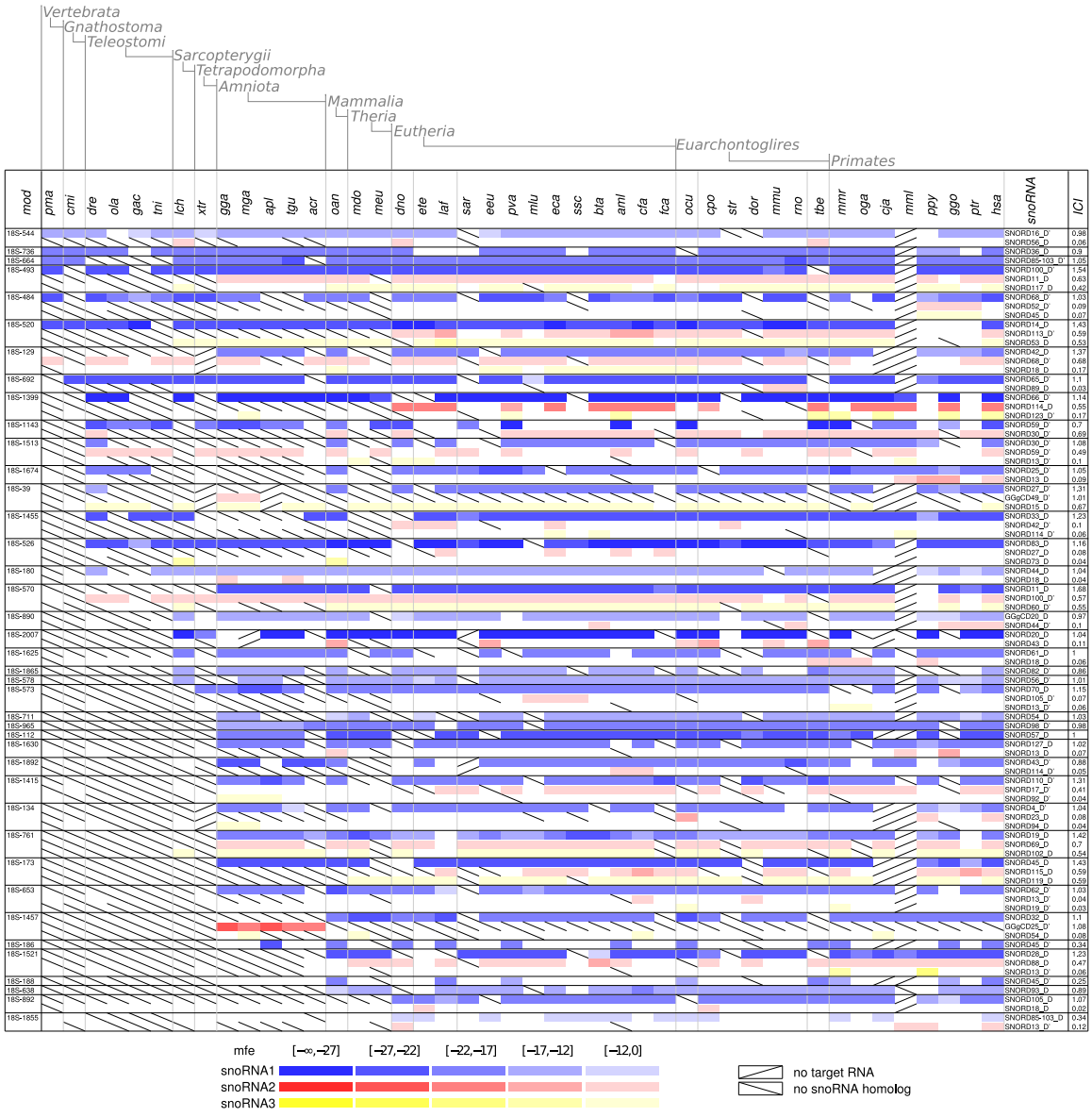

Figure 6.1: Interaction conservation of box C/D snoRNA and targets in 18S rRNA. For each target $t \leq 3$ snoRNA families are displayed in different colors. Interaction energies $\varepsilon(t, s, k)$ determine the saturation of the color. ASEs of both D and the D' box are considered. Crossed-out fields indicate a missing part of the rRNA (from upper left to lower right) and a missing snoRNA (from lower left to upper right). Empty (white) fields indicate that no interaction was predicted. Details see text.

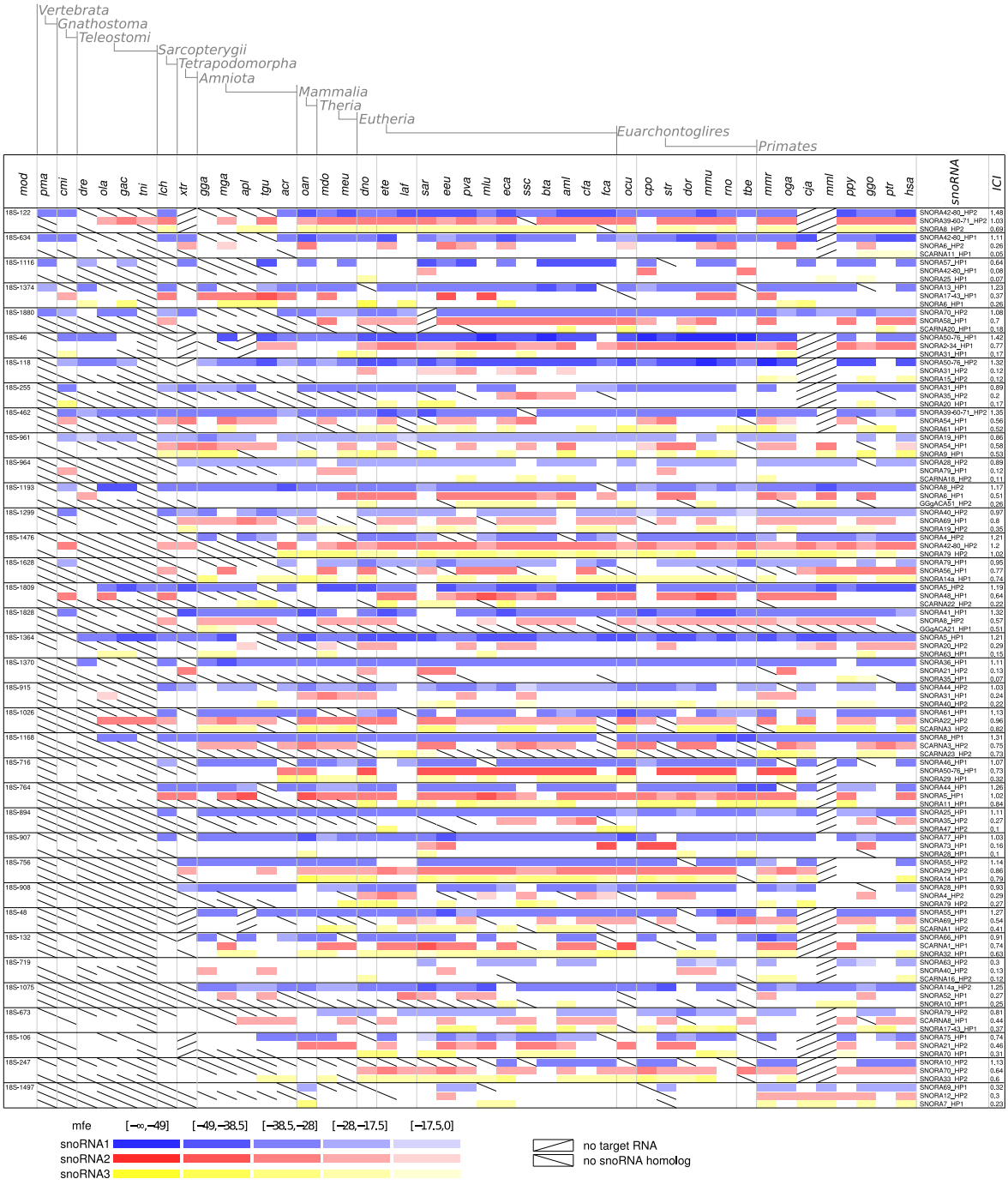

Figure 6.2: Interaction conservation of box H/ACA snoRNA and targets in 18S rRNA. Description of the figure analogously to Figure 6.1.

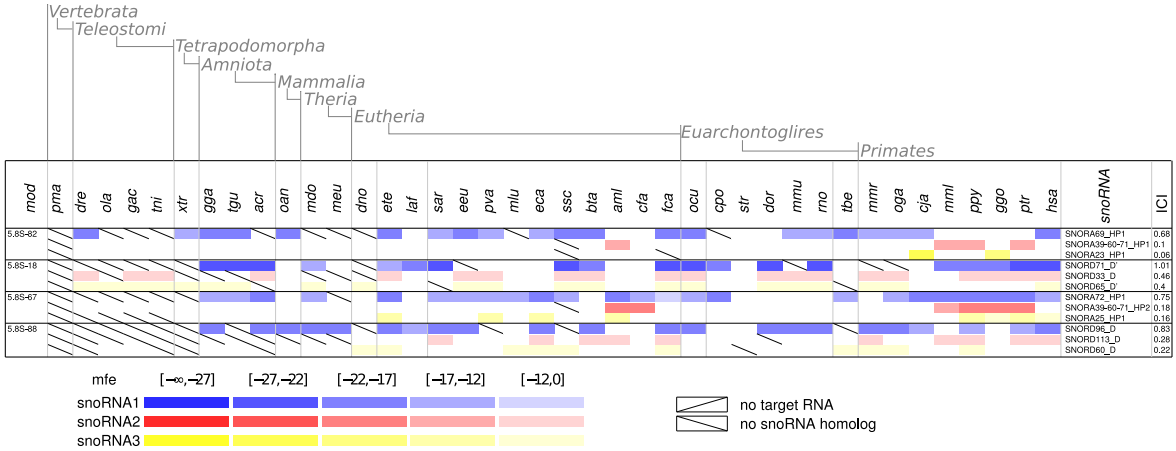

Figure 6.3: Interaction conservation of box C/D and box H/ACA snoRNA and targets in 5.8S rRNA. Description of the figure analogously to Figure 6.1

Figure 6.1 shows the conservation pattern of interactions between box C/D snoRNA guides and methylated sites in the 18S rRNA. The header row lists the abbreviated species names of investigated vertebrates (Appendix Table A.1). The subsequent rows provide detailed information about the interactions of a modification site (first column) and certain snoRNAs (2nd last column). For each target site ≤ 3 snoRNAs are shown. These are ordered by their ICI_{mod} scores (last column). The color intensity of each field is correlated to the predicted minimum free energy of the individual interaction. A field is crossed out, if the snoRNA (form lower left to upper right) or rRNA sequence (from upper left to lower right) is not available for the species. For empty white fields no interaction was predicted. The rows are ordered according to the range of conservation, i.e., interactions with stable partners in all vertebrates appear in higher rows than interactions conserved only within mammals. In general it is observed that once a snoRNA family has occurred its function is conserved. This results in three main groups. The first group of snoRNAs emerged at the root of vertebrates and accordingly its function is conserved in all *Vertebrata*, the second group of interactions appeared within *Teleostomi*, and the third main group arose in *Amniota*. Figures 6.2 - 6.3 and Appendix Figures A.6 - A.8 are analogously.

For LSU (28S and 5.8S rRNAs) 61 of 62 human box C/D snoRNA interactions (Figures 6.3 and Appendix Figure A.6) and 47 of 54 box H/ACA snoRNA interactions were found conserved in vertebrates (Figures 6.3 and Appendix Figure A.7). In LSU for one of two doubly guided 2'-O-methylations and three of four Ψs only a single guide

was found to be conserved. One position (alignment site:28S-4754, corresponding to human 28S-3797) is reported as methylated and pseudouridylated. Interestingly, high ICI scores agree with both interactions.
In snRNAs 16 of 18 box C/D snoRNA interactions and 17 of the 20 box H/ACA snoRNA interactions were recovered (Figure A.8). Two Ψs and two methylations are reported to have two interacting scaRNAs. Both guides were recovered for the methylated residues but only one guiding scaRNA for both pseudouridylations.

ICI-scores For the interactions reported in human the ICI-scores were inspected. The distributions are given in Figure 6.4 (left half). The median ICI_{mod} values for box C/D snoRNAs interactions with the SSU, LSU and snRNAs are 1.04, 0.81, and 0.85, respectively. In analogy median ICI_{sno} values are 1.64, 1.14, and 1.13. Box H/ACA snoRNAs interacting with SSU, LSU and snRNAs yield median ICI_{mod} scores of 1.07, 0.73, and 0.75, respectively. Here, median ICI_{sno} values are 1.08, 0.69, and 0.89 (Table 6.1).

Table 6.1: Median ICI values for the interactions listed in `snoRNA-LBME-db`

score	rRNA	all	C/D	H/ACA
ICI_{mod}	18S	1.05	1.04	1.07
	28S	0.76	0.81	0.73
	snRNAs	0.81	0.85	0.75
ICI_{sno}	18S	1.23	1.64	1.08
	28S	0.90	1.14	0.69
	snRNAs	0.91	1.13	0.89

Only one of 98 interactions between box C/D snoRNAs and ribosomal RNAs was missed and two of 18 with small nuclear RNAs. Due to a more complicated interaction structure of box H/ACA snoRNAs and target RNAs `RNAsnoop` is less sensitive. Missed interactions explain the zero peaks in box H/ACA snoRNA concerning ICI curves.
For all interactions with targets on 18S rRNA the scores are > 1, displaying high conservation of the interactions. As explained in Section 3.1 only for 17 of the 47 investigated vertebrate species full length 28S sequences were available. Due to a resulting lower alignment quality and therefore putative unaligned modification sites lower average ICI scores were obtained for 28S rRNA interactions. This emphasizes

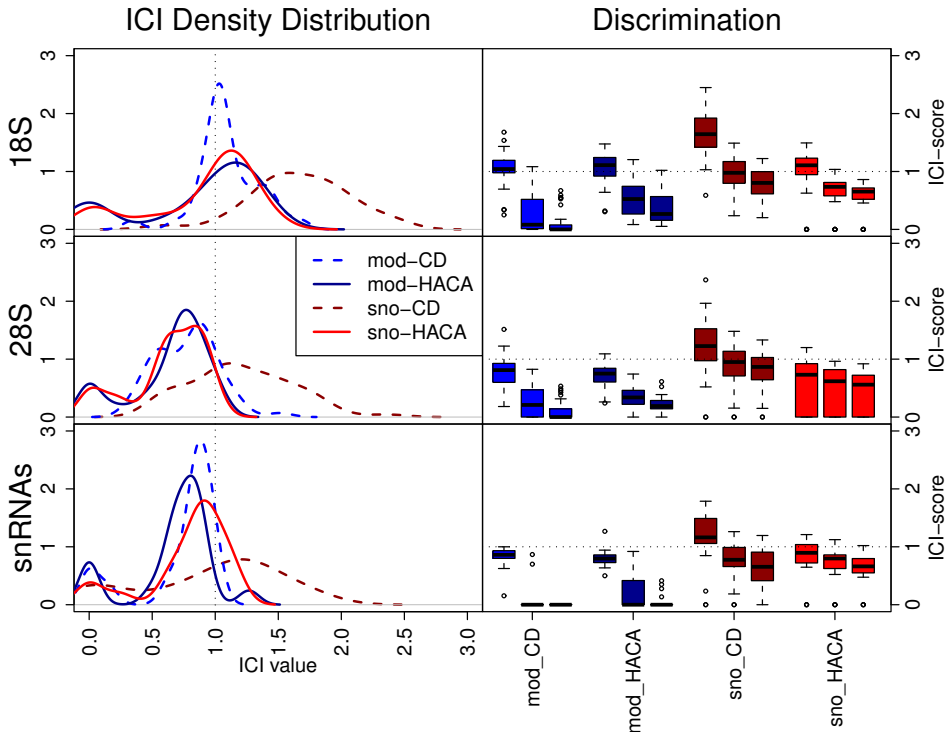

Figure 6.4: ICI_{mod} and ICI_{sno} scores. The plots separately show the values of both types of ICI score (sno & mod) and both types of snoRNA (C/D & H/ACA) on the putative targets 18S rRNA, 28S rRNA and snRNAs. The left part shows the density distribution of the ICI scores. The right part shows a comparison of the best three snoRNA families according to ICI scores for each target site.

the importance of high quality alignments for reliable statements on targets and target conservation.

For many target sites more than one snoRNA family was predicted to have an appropriate ASE to interact. Comparing the ICI_{mod} scores for these families show that there is almost always a single dominating family whose ICI_{mod} scores can clearly be discriminated from those of alternative predictions, see Figure 6.4 right part. The ICI_{sno} scores show similar behavior.

SNORD68 and 18S-484 is a good example for a conserved interaction. The region containing the methylated uridine at 18S-484 (corresponding alignment position to 18S-428 in the human sequence) and the D'-ASE of SNORD68 comprises 11 nts complementarity. The two interacting sequence segments are almost completely conserved from human to sea lamprey Figure 6.5, and Figure 6.1. Two alignment columns show

mutations on the rRNA side, but these mutated nucleotides can still form G-U pairs with the corresponding snoRNA nucleotides. The snoRNA regions not involved in the interaction display lower sequence conservation. The ICI reflects the good conservation of the interaction: $ICI_{mod}(18S-484, SNORD68\_D') = 1.03$. The average individual mfe $\varepsilon(t,s,k)$ of the interaction is $-21.6[kcal/mol]]$. In macaques no interaction could be predicted because the rRNA sequence of the crucial segment is missing. For elephant shark, duck, zebra finch, wallaby, shrew, pig, kangaroo rat, and tree shrew no SNORD68 homolog has been detected with `snoStrip`. In bush baby no interaction with 18S-484 was predicted with `PLEXY` although both the rRNA sequence and the snoRNA homolog are present. Though the ASE of the snoRNA comprises four mutations, two of which do not disrupt base-pairing at the respective positions but result in wobble pairs. Nevertheless, the other two mutations lead to mismatches within the 'core region' of the interaction, wherein only one mismatch is tolerated according to Chen et al. (2007). Two other snoRNA ASEs show a certain amount of complementarity to the 18S-484 target region in chimp, gorilla and orangutan. Low ICI values 0.09 and 0.07, and low average $\varepsilon(t,s,k)$ of -9.76 and -10.2 imply low stability and weak conservation of these interactions, although the snoRNA families are present in *Eutheria* and *Amniota*, respectively. Furthermore, SNORD68 is a double guiding box C/D snoRNA. The ASE upstream of the D-box is complementary to alignment site 28S-3267. This interaction is well conserved in vertebrates. An ICI value of 0.94 and an average mfe of -20.01 indicate that this interaction is very stable.

Special Cases Interestingly there are several modifications for which at some point in evolution a second snoRNA guide occurs which is then retained. A feasible explanation may be differential snoRNA expression so that a back up of the modification under different cellular conditions is necessary. Several examples support this hypothesis. Pseudouridylation of site 28S-5501 (corresponding to human 28S-4491) has been reported to be guided by SNORA10. With the `snoStrip` pipeline homologs were identified in supraprimates, carnivors, cow and horse. For four of five species, where snoRNA and LSU sequences are available, the interaction was predicted with $ICI_{mod} = 0.73$. As additional guide matching 28S-5501 SNORA63 was found. The interaction is conserved throughout vertebrates with $ICI_{mod} = 0.81$. The two snoRNAs are encoded in introns (sense direction) of RPS2 (ribosomal protein 2) and EIF4A2 (eukaryotic translation initiation factor), respectively. The expression of these proteins in human is antagonis-

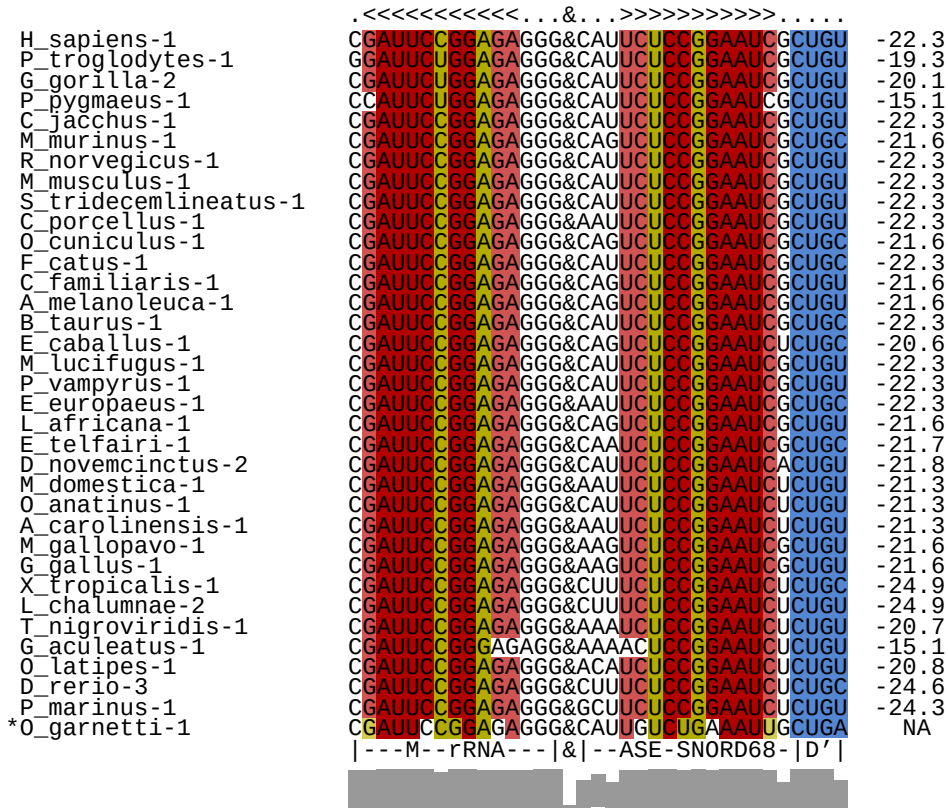

Figure 6.5: Conservation of the interaction between the region upstream of D'-box of snoRNA family SNORD68 (right side) and the region around the 2'-O-methylated uridine at alignment column 484 (left side). Target RNA segment and ASE are separated by &. The methylated residue is marked with M in the last row. The first row provides information about the structure of the RNA-RNA interaction. Intermolecular basepairs are indicated by arrow brackets. The blue box marks the D'-box. Red and green columns highlight conservation of the RNA-RNA interaction. Completely conserved basepairs are shown in red. Green columns mark basepairs with compensatory mutations. Lighter colors indicate loss of basepairs in individual species. The gray bars at the bottom correspond to the degree of sequence conservation computed by RNAalifold (Bernhart et al., 2008). MFEs for the interactions ($\varepsilon(t,s,k)$) as predicted by PLEXY are provided in the last column. In O. garnettii no interaction can occur due to additional mutations in the 'core region', therefore it is marked with asterisk.

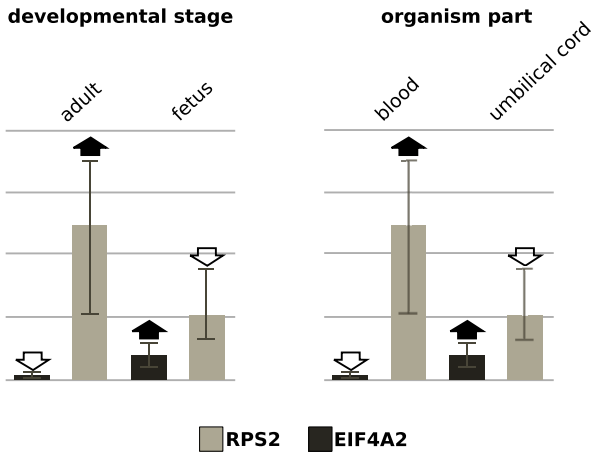

Figure 6.6: Expression of SNORA10 and SNORA63 hostgenes EIF4A2 and RPS2 is shown at different developmental stages (left) and in different organism parts (right). Arrows above the bars indicate down and up regulation, respectively. Figures are taken from the Gene Expression Atlas.

tically up and down regulated in fetus and adult state as well as in blood and umbilical cord tissue (according to data from the Gene Expression Atlas E-GEOD-6236[1]) (see Figure 6.6). As a consequence, the expression patterns of these hostgenes ensure that at least one of the two snoRNAs is present in the different tissues and for different developmental stages.

Another example of redundant guiding is the modification of site 18S-761 (corresponding to human 18S-683) by SNORD19 and SNORD69. Both interactions are conserved in amniotes with ICI values of 1.41 and 0.7, respectively. Both snoRNAs reside in introns of GNL3 (guanine nucleotide binding protein-like 3). This protein comprises different isoforms. All transcripts include the intron encoding SNORD19 but at least one isoform ends in front of the intron hosting SNORD69. Thus, SNORD69 expression is regulated by alternative splicing of GNL3. Additional data are compiled in Supplementary Table S2 on www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/12-022. A comprehensive analysis of hostgene expression patterns, however, has not to be conducted.

Another exception from the general pattern of canonical snoRNA functionality is a change-over of the snoRNA guide for a single modification, e.g. the methylation of 18S-1457 (corresponding to sequence position 1328 in human and 1286 in chicken). The site matches the 3'-ASE of the mammalian-specific SNORD32 family. In aves, on

[1] `http://http://www.ebi.ac.uk/gxa`

the other hand, the modification is addressed by the 5'-ASE of bird specific snoRNA family GGgCD25. This interesting behavior was detected by high scoring ICIs of 1.1 and 1.08 for the same modification. Zemann et al. (2006) reported similar observations in nematodes.

Case Studies

The ICI score was used to investigate conservation of the few published experimentally verified interactions between a snoRNA and its guided modification.

Human experimentally verified interactions. Xiao et al. (2009) tested 16 predicted interactions of box H/ACA snoRNAs and Ψs in human rRNAs. While 12 have been verified, four predictions have been rejected in their study. The conservation throughout vertebrates of all 16 interactions was measured using the ICI score (Table 6.2). High ICI values agree with the experimentally verified interactions. Two of the four negative results are not conserved at all, so that no ICI could be computed. Of the remaining two, one cannot be conclusively resolved by our method, leaving a single case where our predictions disagree with the experimental results.

Table 6.2: Comparison of ICI_{mod} values with experimentally tested interactions (+: interaction verified, − interaction rejected) between box H/ACA snoRNAs and ribosomal RNAs. The SNORA50 and SNORA76 as well as SNORA80 and SNORA42 are paralogs and hence members of the same family in this survey.

alignment	human	guiding families	verified	ICI_{mod}
18S-46	18S-34	SNORA50,SNORA76	+,+	1.47
18S-118	18S-105	SNORA50,SNORA76	+,+	1.33
18S-122	18S-109	SNORA80,SNORA42	+,+	1.47
18S-634	18S-572	SNORA80,SNORA42	+,-	1.11
18S-673	18S-609	SNORA24	-	-
18S-908	18S-815	SNORA28	+	0.93
18S-961	18S-863	SNORA24,SNORA19	+,-	0.38,0.86
18S-964	18S-866	SNORA28,SNORA19	+,-	0.89,-
28S-4573	28S-3618	SNORA19	+	0.76
28S-4665	28S-3709	SNORA19	+	0.58

Due to high sequence similarity snoRNA families SNORA50 and SNORA76 as well as SNORA42 and SNORA80, respectively, have been merged into one family each in the snoRNA dataset compiled with `snoStrip`. They are denoted SNORA50-76 and SNORA42-80 in following (Alignments can be downloaded at `http://www.bioinf.uni-leipzig.de/supplements/12-022` S3.1). In terms of guiding potential, snoRNA families were treated as entities and it was not distinguished between guiding potential of single paralogous sequences. This is not problematic for family SNORA50-76, where both sequences have been verified to guide 18S-46 (human 18S-34) and 18S-118 (human 18S-105) in human by Xiao et al. (2009). In the second case, however, SNORA42 and SNORA80 both have been verified to guide 18S-122 (human 18S-109), but only SNORA80 interacts with 18S-634 (human 18S-572) while SNORA42 does not. The resolution of our analysis correctly predicts that family SNORA42-80 targets 18S-572. The confirmed interaction between SNORA24 and modification corresponding to 18S-961 (human 18S-863) yielded a very low value of $ICI_{mod} = 0.38$. Nevertheless, SNORA24 homologs are predicted to interact with the verified target in *L. africana*, *M. domestica*, *D. novemcinctus*, *T. guttata*, *A. carolinensis*, *D. ordii*, *R. norvegicus*, *C. jacchus*, *M. gallopavo*, *G. gallus*, *M. eugenii*, *L. chalumnae* with an average interaction energy of -26.96. SNORA19, however, is the best scoring interaction for the modification at 18S-961 (18S-863 in publication). Although, this interaction seems to be conserved throughout vertebrates it has been rejected by Xiao et al. (2009). In the remaining two cases, (18S-673 (human 18S-609) & SNORA24 and 18S-964 (human 18S-866) & SNORA19), our analysis agrees with the published negative experimental results.

Zebra fish snoRNAs that are essential during embryonal development have recently been identified by Higa-Nakamine et al. (2012). Three methylated residues in the rRNAs are guided by snoRNA families SNORD44, SNORD78, and SNORD26, respectively. Our method identified all three interactions as conserved within vertebrates. The interaction between SNORD44 and site 18S-180 (18S-163 in publication) yielded a high ICI_{mod} score of 1.04. SNORD78 guides 2'-`O`-methylation of the according guanine 28S-5615 (28S-3745 in publication) in vertebrates with $ICI_{mod} = 0.85$. SNORD26 was recovered as conserved guide for modification of adenosine at alignment position 28S-939 (28S-398 in publication) with $ICI_{mod} = 0.62$. The lower values can be explained by the lower quality of this alignment.

The two most conserved Ψs are modified by SNORA74 (U19) at least in vertebrates. The modifications at alignment positions 28S-4697 and 28S-4699 (corresponding to human residues 28S-3741 and 28S-3743) are conserved even in bacteria. There, these modifications are produced by the specialized pseudouridine synthase RluD (Ofengand, 2002; Ofengand and Bakin, 1997). Both Ψs are located in the decoding center, a central region of the ribosome contacting the SSU and the passing tRNAs.
SNORA74 has an exceptional three hairpin structure conserved from yeast to human (Badis et al., 2003). The computed ICI_{mod} values of 0.95 and 0.92 confirm the complementarity of corresponding alignment sites and ASEs in the 5'- and 3'-hairpins in all vertebrates. (See Figure A.7 and alignment on supplementary page)

6.2 New SnoRNAs for Known Human Modifications

Although most 2'-O-methylations and pseudouridylations in human rRNAs have already been assigned to snoRNA guides, some cases remain not matching with a given snoRNA. SnoRNA families with above average ICI_{mod} values are most likely the unrecognized conserved snoRNA guides for these modifications. Thus, the appropriate snoRNA guides were identified for eight of 21 rRNA sites and two of ten accounted for U2 snRNA modifications. The results are summarized in Table 6.3 and illustrated in Figures 6.7, A.9, A.10. For the rest of the modifications, still unassigned to known snoRNA guides, it is suggested that the vertebrate genomes still harbour a few undiscovered snoRNA families containing matching antisense elements.

SSU rRNA For three methylations and two Ψs in 18S with previously unassigned snoRNAs appropriate guides were predicted (Figure 6.7). The 3'-hairpin of SNORA55 putatively guides pseudouridylation of 18S-756 (human 18S-681) in tetrapods and the 5'- hairpin of SNORA61 pseudouridylation of 18S-1026 (human 18S-918) in tetrapods and coelacanth. The interactions have ICI scores of 1.14 and 1.13, respectively. Apparently, these hairpins have a second guiding function, as both hairpins have alternative targets listed in snoRNA-LBME-db. These interactions are also widely conserved in amniotes and in tetrapods and coelacanth with ICI_{mod} values of 1.24 and 1.05, respec-

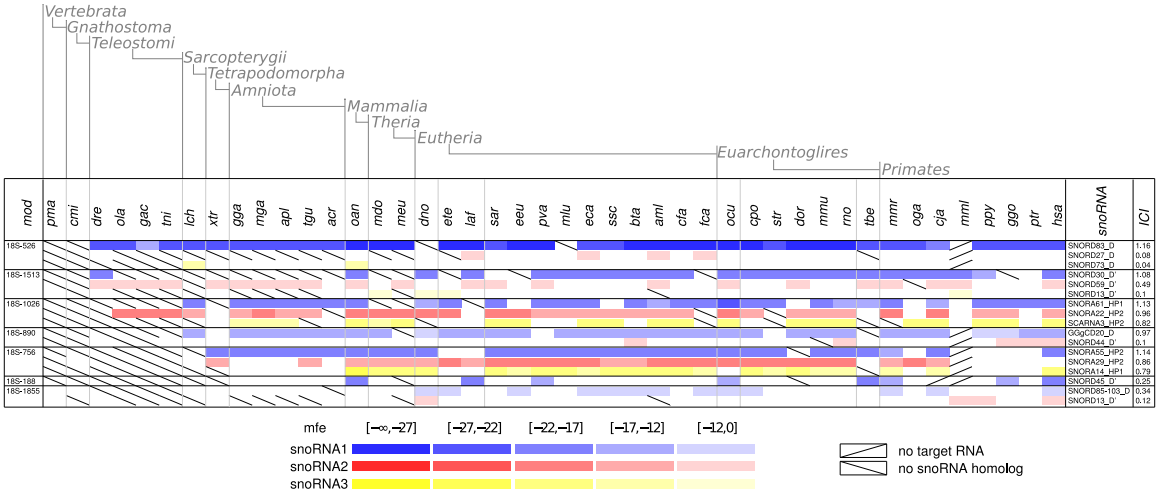

Figure 6.7: SnoRNA guides to known modifications in 18S rRNA

tively). A second guiding function of a target binding region is not only observed for box H/ACA snoRNAs but also within the box C/D snoRNA family SNORD30. The reported target to the ASE adjacent to the D'-box is 28S-4761 (28S-3804 in human). This interaction is supported by conservation and an additional conserved guiding potential to the unassigned methylation of the cytosine corresponding to alignment column 18S-1513 (18S-1383 in human) is suggested. For orphan methylation of the cytosine at 18S-890 (18S-797 in human), conserved complementarity ($ICI_{mod} = 0.98$) was detected from coelacanth throughout tetrapods to chicken snoRNA GGgCD20. For this snoRNA family no vertebrate homologs were known previously to the `snoStrip` search. At last, an interaction between residue 18S-526 (18S-468 in human) and the orphan snoRNA family SNORD83 was identified. This interaction is conserved in *Teleostomi* and has not been reported before.

LSU rRNA Three box H/ACA snoRNAs with ICI_{mod} value above average were found as putative guides for modified nucleotides with unknown guide in 28S (Figure A.9). The first hairpin of SNORA78 was predicted to interact with the pseudouridylated residue at 28S-5263 (28S-4266 in human), while the other hairpin of this snoRNA is known to target position 28S-5339 (28S-4331 in human). The fact that box H/ACA snoRNAs often guide nearby pseudouridylation enhances the prediction. The two other box H/ACA snoRNAs interacting with the previously unknown guided Ψs 28S-2722 (28S-1849 in human) and 28S-4824 (28S-3863 in human) are SNORA51 and SNORA84,

Table 6.3: Predictions for known modifications without assigned snoRNA guides.

alignment	human	type	snoRNA	ASE	ICI_{mod}
18S-526	18S-468	Am	SNORD83	D	1.16
18S-756	18S-681	Ψ	SNORA55	HP2	1.14
18S-890	18S-797	Cm	GGgCD20	D	0.98
18S-1026	18S-918	Ψ	SNORA61	HP1	1.13
18S-1513	18S-1383	Am	SNORD30	D'	1.1
28S-2722	28S-1849	Ψ	SNORA51	HP1	1.09
28S-4824	28S-3863	Ψ	SNORA84	HP2	0.86
28S-5263	28S-4266	Ψ	SNORA78	HP1	0.83
U2-23	U2-15	Ψ	GGoACA7	HP2	0.83
U2-79	U2-47	Um	GGgCD76	D	0.9

respectively. Both families have been reported as orphan snoRNAs, so far. SNORA51 was identified as homolog of orphan chicken snoRNA GGoACA9 during `snoStrip` run and the predicted interaction turned out to be also conserved in chicken.

U2 snRNA is pseudouridylated at position 23 (U2-15 in human). Here, an interaction of this orphan Ψ with snoRNAs that belong to chicken annotated orphan GGoACA7 was predicted. The interaction is conserved throughout vertebrates with $ICI_{mod} = 0.83$ (Figure A.10). A high scoring snoRNA family for unassigned 2'-O-methylation at site 79 (corresponding to human 47) is chicken snoRNA GGgCD76. The chicken interaction has already been predicted by (Shao et al., 2009) without respect to any conservation issues. With $ICI_{mod} = 0.9$ this interaction is conserved in vertebrates since coelacanth.

6.3 Functions for Orphan snoRNAs

The function of 41 human snoRNAs is still unknown. The ICI_{sno} score was used to identify conserved complementarity of these orphan snoRNAs to rRNAs or snRNAs. Table 6.4 summarizes all predictions where at least one of the ICI scores is above treshold, as well as those where the predicted target site is known to be modified. The interactions are illustrated in Figures A.11 and A.12.

Orphan guides for modifications without matching snoRNA Four orphan snoRNAs were identified as conserved guides for known modifications with previously

Table 6.4: Predictions for orphan snoRNA families. Known modifications are marked by asterisk.

snoRNA	ASE	alignment	human	type	ICI_{mod}	ICI_{sno}
GGgCD20	D	18S-890	18S-797	Cm *	0.97	1.49
SNORD109	D'	28S-5424	28S-4414	m^5C	0.9	0.93
SNORD116	D'	18S-1286	18S-1162	C	1.09	1.21
SNORD125	D	18S-1623	18S-1440	C	0.88	1.77
SNORD83	D	18S-526	18S-468	Am *	1.16	1.86
SNORD86	D	18S-1345	18S-1219	C	1.1	1.64
SNORA49	HP1	28S-3725	28S-2826	U	0.75	0.71
	HP2	28S-3729	28S-2830	U	0.9	0.87
SNORA51	HP1	28S-2722	28S-1849	Ψ *	1.1	1.07
SNORA84	HP2	28S-4824	28S-3863	Ψ *	0.86	0.72

unknown guides (GGgCD20 & 18S-890, SNORD83 & 18S-468, SNORA51 & 28S-2722, SNORA84 & 28S-4824). These interactions were already described in the previous section.

Orphan snoRNA targets m^5C Orphan SNORD109 has two paralogs in human encoded in introns of the paternally expressed SNURF-SNRNP locus. SNORD109 is expressed in brain and kidney, and, at lower levels in lung and muscle (Runte et al., 2001). Surprisingly, the analysis revealed ten nucleotides conserved complementarity of the D'-ASE to target site 28S-5424 (corresponding to human 28S-4414) with ICI_{mod} and ICI_{sno} values of 0.9 and 0.93, respectively (Figure 6.8. This nucleotide is reported as 5-methylcytidine (m^5C) in `3D Ribosomal Modification Maps Database` (Appendix Figure A.2[2]). The methylation of the nucleobase instead of the associated ribose is a chemical modification not associated with snoRNAs. It should be kept in mind, however, that also other SNURF-SNRNP-snoRNA complexes, such as SNORD115, have been shown to exhibit non-canonical behavior.

Orphan snoRNAs and intricate structures Furthermore, complementarities were predicted between stretches of ribosomal RNA without reported modifications and orphan snoRNAs. A careful examination of how differential expression of snoRNAs affects

[2] `http://people.biochem.umass.edu/fournierlab/3dmodmap/humlsu2dframes.php`

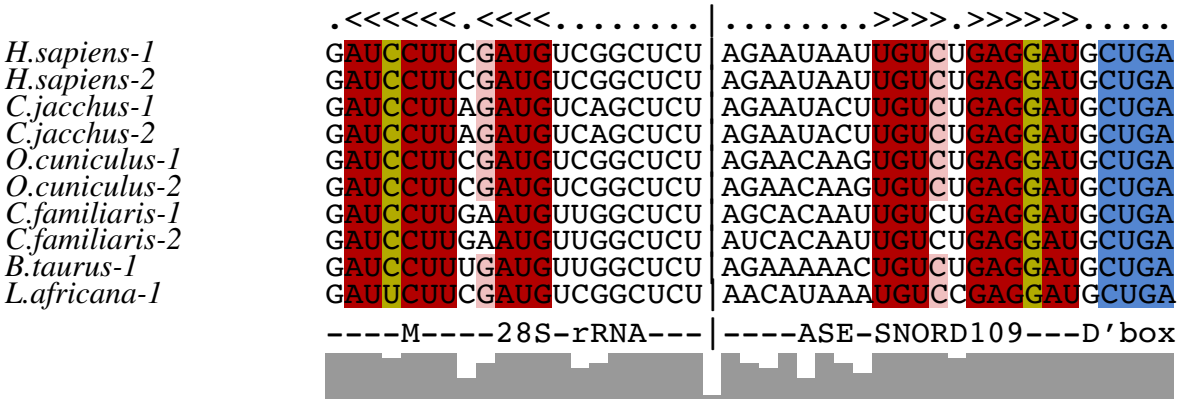

Figure 6.8: Interaction between SNORD109 and 28S-5424. Explanation equivalent to Figure 6.5.

rRNA modification has not been carried out so far (Xue and Barna, 2012). Hence, the possibility of undetected modifications occurring only under certain conditions should not be excluded.

High scoring interactions were predicted for SNORD116 and 18S-1286 (human position 18S-1162) and SNORD86 and 18S-1345 (human 18S-1219). The putatively targeted nucleotides are located next to pseudoknotted RNA stretches (according to `3D Ribosomal Modification Maps Database` (Appendix Figure A.1[3]). Since, modifications are capable of stabilizing intricate structures the predictions are not implausible.

Orphan snoRNA and extensively modified regions High scores of $ICI_{mod} = 0.88$ and $ICI_{sno} = 1.77$ are computed for the ASE upstream of the D-box of SNORD125 as guide for modification of 18S-1623 (human 18S-1440). It is located in a functional region of the SSU, close to tRNA binding sites where other reported modifications are proximal.

For orphan SNORA49, targets were predicted for both hairpins. The putative modifications are 28S-3725 and 28S-3729 (human 28S-2826 and 28S-2830). They are located in a small helical structure with multiple reported modifications (Appendix Figure A.2[4]).

A complete list of predictions of rRNA and snRNA targets for the orphan snoRNAs is provided at www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/12-022 Table S4.2.1. The remaining snoRNAs that lack complementarity to rRNAs and snRNAs

[3] `http://people.biochem.umass.edu/fournierlab/3dmodmap/humssu2dframes.php`
[4] `http://people.biochem.umass.edu/fournierlab/3dmodmap/humlsu2dframes.php`

might have alternative functions, e.g., cleavage of the primary rRNA transcript, targeting mRNA and altering their alternative splicing, or translational control in a miRNA-like fashion after being processed into smaller fragment (sdRNAs).

6.4 Identification of Distant Homologs

Early studies into snoRNAs frequently used homologous target sites as an argument for the homology of the snoRNAs themselves. The chicken snoRNAs GGgCD3, GGgCD4, GGgCD14, GGgCD24, GGgCD29, GGgCD63, GGgCD64, and GGgCD66 reported by Shao et al. (2009), all members of the box C/D class, may serve as a good example for the validity of this approach. According to the `BLAST`-based homology search procedure implemented in `snoStrip` they are specific to the avian lineage. Target prediction with `PLEXY` and their ICI_{mod}-scores identified them as conserved guides for methylations of the target alignment positions 18S-129, 18S-134, 18S-653, 18S-1415, 18S-1892, 28S-5348, 28S-5371, and 28S-5474 (the coordinates refer to the homologous nucleotides in the human RNAs). These positions are targeted by human snoRNA families SNORD42, SNORD4, SNORD62, SNORD110, SNORD43, SNORD60, SNORD1, and SNORD69, respectively. Homologs of these families were readily identified by `snoStrip` in other mammals but not in sauropsids. The families could be combined into common alignments and a detailed inspection showed that they are indeed homologs. (Alignments are available on the Supplementary page www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/12-022 S3.2.)

Although most platypus snoRNA sequences from the study of Schmitz et al. (2008) could be merged to mammalian families during the `snoStrip` search procedure already, some appeared as species specific. Analogously to the avian snoRNA families, platypus snoRNA sequences Oa1759, Oa2916, and Oa2126 could be identified as homologs of the vertebrate snoRNA families SNORD110, SNORD96, and SNORD4 respectively.

In case of box H/ACA snoRNAs the inference of a evolutionary origin from functional homology was possible only for SNORA64 and GGgACA47, both guiding pseudouridylation of 28S-6029 (28S-4975 in human).

Particularly divergent sequences were observed for the family containing human SNORD62, chicken GGgCD14, and platypus bOaCD62i. The alignment reveals a large deletion in aves lineage in the 3'-part of the snoRNA. The selective pressure is focused on the ASE downstream of the D' box which retains complementarity to

Table 6.5: Based on conserved targets distant homologies could be identified between chicken snoRNAs thought to be avian specific, platypus sequences thought to be species specific and mammalian snoRNA families. Sequence [1)] present in RFAM alignment [2)] homology identifiable by RFAM search [3)] homology confirmed by Makarova and Kramerov (2011). The third column provides a mapping of the alignment column (aln) and modified positions in the human (hsa), chicken (gga), and platypus (oan) rRNA sequences.

RFAM	Human	Chicken	Platypus	rRNA	modified position in			
					aln	hsa	gga	oan
RF00150	SNORD42	GGgCD3[2)]	Oa1817[1),2)] Oa2691[1),2)]	18S	129	116	116	95
RF00266 CL00053	SNORD4	GGgCD4[2),3)]	Oa2132[1),2)] Oa2126[1),2)]	18S	134	121	121	100
RF00153 CL00068	SNORD62	GGgCD14[3)]	bOaCD62i[1),2)] Oa2054[1),2)]	18S	653	590	551	594
RF00610 CL00076	SNORD110	GGgCD24	Oa1759	18S	1415	1288	1246	1293
RF00221 CL00059	SNORD43	GGgCD29[2),3)]		18S	1892	1705	1658	1706
RF00055 CL00072	SNORD96		Oa2961	5.8S	88	75	74	74
RF00271 CL00066	SNORD60	GGgCD63		28S	5348	4340	3825	NA
RF00213	SNORD1	GGgCD64[2),3)]	Oa1765[1),2)]	28S	5371	4362	3847	NA
RF00574	SNORD69	GGgCD66[2)]	Oa2470[1),2)]	28S	5474	4464	3949	NA
RF00264 CL00041	SNORA64	GGgACA47[2),3)]		28S	6029	5285	4282	NA

the target in the SSU. The aves lineage has 12 nucleotides perfect complementary to the region around the modified adenine 653 (18S-590 in human), while in mammals the interaction is with 14 nucleotides longer but comprises a mismatch at the 8th position (Alignment and figure of interaction is provided on the www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/12-022 S3.2).

Alignments of avian, platypus and mammalian families can readily be combined by manual inspection and editing according to the list of correspondences in Table 6.5, demonstrating that these families are indeed homologs.

Vertebrate rRNAs have a highly conserved modification pattern and a high level of sequence conservation in the vicinity of the modified nucleotides. In contrast, snoRNAs are famous for their overall high sequence entropy. Nevertheless, this study on the conservation of snoRNA-target RNA interactions confirmed that in general the same modifications are guided by the same snoRNA families (Hoeppner and Poole, 2012). In fact, together with the functional sequence boxes the ASEs form the regions of strongest sequence conservation within the snoRNAs, disclosing co-evolution of snoRNAs and their targets. Furthermore, the interactions are maintained by compensatory mutations within the snoRNA sequences preserving the base pairing.

Evaluation of all known human interactions with the ICI score recovered $\sim$ 87% (snRNAs), 83% (box H/ACA snoRNAs-rRNAs) and nearly 100% (box C/D snoRNAs-rRNAs) of the interactions as conserved in eutherians or further. This and consistent high ICI values for experimentally verified interactions also support the usefulness of the ICI measure.

Besides the observed stable partnerships between snoRNAs and their associated target sites, redundant guides and a few changeover of guides could be detected using our evaluation method. In individual cases we find that redundant guides are processed from host genes with anti-correlated expression profiles, explaining how such redundant snoRNAs are evolutionarily maintained.

The target analysis workflow (Section 3.3) that was applied to all snoRNA sequences (ICI_{sno}) and all putative modifications (ICI_{mod}) added new edges to the network of snoRNAs and the network of interactions. SnoRNA families, reported in distantly related organisms, could be merged based on sequence similarity and equivalent function. Further, snoRNA guides could be assigned to ten of the 31 so-far unexplained modifications in rRNA and U2 snRNA, and putative functions were designated for nine of 41 orphan snoRNAs. Thus, the ICI proved to be a very useful measure to reunite so far lonesome soulmates.

CHAPTER 7

SnoRNA Interactions with Coilin in Cajal Bodies

In this chapter an iCLIP experiment with the Cajal Body marker protein coilin is described. The work was a collaborative project with Martin Manycha and Karla Neugebauer, who did all the wet lab work and parts of the analysis. The experiments that were performed in human and mouse identified a surprisingly high diversity of ncRNAs populating the Cajal Bodies. Among them also the majority of the known snoRNAs could be detected. It was even possible to annotate additional novel snoRNAs in the portion of expressed transcripts that had no annotations assigned previously. For these the conservation and function in vertebrates was investigated and official HGNC names were allocated. On top the coilin binding context within the snoRNA sequences was examined. The study is published in *Molecular Cell* (Machyna et al., 2014).

7.1 iCLIP, Coilin and Cajal Bodies

Performed iCLIP (individual-nucleotide resolution UV-cross linking and immunopreciptitation) experiments (Section 1.2) with the Cajal Body (CB) marker protein coilin in human and mouse revealed an unexpected diversity of RNA interaction partners. The aim of cross-linking experiments (CLIP) (developed by laboratory of Robert Darnell in 2003 (Ule et al., 2003)) is to find RNA interaction partners of proteins *in vivo*. A cell is UV-irradiated, which establishes covalent bonds between RNAs and proteins. Afterwards, the protein of interest (with permanently linked RNAs) is selected and isolated from the cell by immunoprecipitation. After digestion of the protein the remaining RNAs are amplified for high-throughput sequencing (Milek et al., 2012).

Cajal Bodies and Coilin The Cajal Body (CB), named for his discoverer Santiago Ramon y Cajal is a sub-cellular compartment in the nucleus. Cajal Bodies have been known to be involved in biogenesis of ribonucleoprotein particles (RNPs), containing mainly snRNAs, and also SNORD3, SNORD118, and telomerase RNA (Machyna et al., 2013; Matera et al., 2009). However, how the assembly of CB itself (and also other cellular compartments without lipid bilayer membrane) is organized *in vivo* is still unsolved.

The coilin protein is required for CB formation and maintenance in many species and cell types (Liu et al., 2009; Machyna et al., 2013; Strzelecka et al., 2010; Tucker et al., 2001). It is an intrinsically disordered protein with a coiled structure and has no domains with defined function (Tucker et al., 2000). Through coilin-coilin self interaction The CB is supposed to hold together (Hebert and Matera, 2000).

7.2 An Unexpectedly Complex Set of RNA Interactors

It has been previously suggested that coilin binds RNA, and *in vitro* data had shown that purified coilin is capable of associating directly with RNA homopolymers and selected snRNAs (Broome and Hebert, 2013; Makarov et al., 2013). These findings raised the intriguing possibility that coilin might interact directly with specific RNA *in vivo*. To test this hypothesis, UV-cross-linking immunoprecipitation (iCLIP) (König et al.,

2010) of coilin-GFP from human and mouse cells was performed in the Neugebauer lab. GFP stand for green fluorescent protein, for simplicity the fluorescent labeled coilin, will be referred to as coilin only. CLIP is preferable to RNA immunoprecipitation without cross-linking (RIP), which in the case of coilin would be expected to yield CB localized RNAs whether or not interactions were direct (Broome and Hebert, 2013; Riley and Steitz, 2013). After pre-processing of the sequenced reads, they were mapped to the human and mouse reference genomes (hg19 and mm9) with `bowtie` (Langmead and Salzberg, 2012). CLIP-tags within a range of ±15 nucleotides were merged and the false discovery rate (FDR) of the CLIP-tag positions was computed against a set where their position was 100× randomized. Cross-Linking sites with $FDR < 0.5$ were considered significant (König et al., 2010; Sugimoto et al., 2012). Indeed, coilin is cross-linked to RNA, and independent biological replicates were highly reproducible. Coilin CLIP-tags mapped to defined peaks in U3, the Pol II-driven spliceosomal RNAs (U1, U2, U4, and U5) as well as histone mRNAs and U7 snRNA (Figure 7.1A). All of these transcripts have also been detected in a previously performed coilin Chip-seq (data not shown here).

Coilin Interacts with Intron-encoded SnoRNAs

The highest enrichment of significant coilin CLIP-tags was detected in non-coding RNAs (ncRNAs) (Figure 7.1B). Surprisingly, the majority of these tags (30%) were not from snRNAs but rather intron-encoded snoRNAs (Table 7.1, Figure 7.1B). To date, only SNORD3, SNORD118 and SNORD14 box C/D snoRNAs have been detected co-localizing with CBs (Boulon et al., 2004; Narayanan et al., 1999). Moreover, intronic snoRNA genes have not been detected by coilin Chip-seq (data not shown). Alternatively, introns may be released from chromatin before coilin can accumulate at these chromosomal sites, owing to the rapidity of co-transcriptional splicing (Brugiolo et al., 2013).

To determine whether coilin interacts with snoRNAs before or after processing from introns, CLIP-tag densities mapping to the snoRNAs, surrounding intron sequence, and spanning the snoRNA-intron boundaries were analyzed. Coilin CLIP-tags mapped exclusively inside the snoRNA 5'- and 3'-end boundaries, suggesting that coilin binds snoRNAs after processing (Figure 7.2A). To address the specificity of coilin interactions with snoRNAs, the distribution of CLIP-tags within regions of each class of snoRNA

Table 7.1: Transcripts enriched with significant coilin CLIP-tags in mouse (P19) and human (HeLa) (ENSEMBL gene annotation v59) cells. Percentage represents fraction of total CLIP-tags.

		Mouse			Human		
	Name	Genes	Clip-tags	(%)	Genes	Clip-tags	(%)
snRNA	U1	8	531	0.1	11	380	0.1
	U2	18	28,767	5.4	25	31,462	7.1
	U4	2	3,435	0.6	6	8,565	1.9
	U5	8	18,825	3.5	8	16,245	3.7
	U6	66	2,030	0.4	45	2,589	0.6
	U7	1	7,055	1.3	2	2,287	0.5
	U11	2	1,473	0.3	1	1,041	0.2
	U12	3	3,309	0.6	1	6,464	1.5
	U4atac	1	4,446	0.8	3	3,842	0.9
	U6atac	1	1,253	0.2	5	2,158	0.5
snoRNA	SNORD3	9	10,436	2.0	3	1,534	0.3
	SNORD118	1	12,420	2.3	2	9,086	2.1
	SNORD13	3	5,358	1.0	1	2,591	0.6
	TERC	1	3,520	0.7	1	4,857	1.1
	H/ACA	132	59,148	11.1	104	44,889	10.2
	C/D	144	95,603	17.9	134	74,725	17.0
	scaRNA	29	102,886	19.3	28	66,508	15.1
Other	Y-RNA	2	819	0.2	2	140	0.0
	7SK	2	2,396	0.4	2	1,647	0.4
	RNAse MRP	1	907	0.2	1	1,449	0.3
	MALAT1	0	0	0.0	1	33	0.0
	NEAT1	0	0	0.0	1	145	0.0
	Histone RNAs	44	949	0.2	22	287	0.1
	tRNA	221	7668	1.4	262	5882	1.3
	7SL	0	0	0.0	1	75	0.0
	RNase P	1	139	0.0	1	263	0.1
	miRNA	7	2849	0.5	1	33	0.0

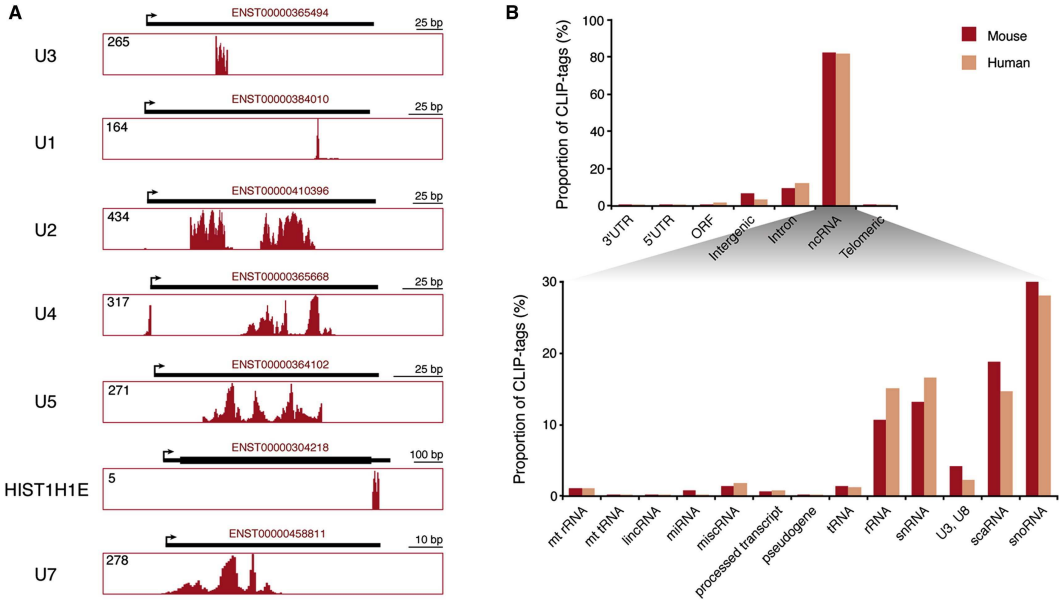

Figure 7.1: Coilin CLIP Identified Hundreds of Small Non-Coding RNAs as Targets in Mouse and Human Cells. ***A****: UCSC genome browser view of selected human transcripts with a high number of coilin CLIP-tags (red).* ***B****: Significant CLIP tags were assigned to specific groups based on biotype of their respective transcript. Protein-coding transcripts were further subdivided into 3'- and 5'- untranslated regions (UTR), open reading frames (ORF), and intronic regions. Bars represent relative CLIP-tag abundance. Note that snoRNAs comprise the most abundant class of ncRNAs bound to coilin in both mouse and human cells.*

was examined. Each of the box H/ACA and box C/D snoRNAs were divided into three segments based on their predicted secondary structures (Figure 7.2B-C). A comparison of CLIP-tag counts between different segments revealed that coilin preferentially binds to the 5'-half of the loop in box C/D snoRNAs and the 3' stem loop structure of box H/ACA snoRNAs (Figure 7.2B-C). To investigate coilin binding position at higher resolution, all snoRNA genes were scaled to the same length and divided into 50 equal bins; CLIP-tags were calculated and summed for each bin. The CLIP-tag distribution in box H/ACA snoRNAs (Figure 7.2B) reveals multiple peaks in the 3'- region of RNAs. These roughly correspond to the small terminal loop and both sides of the pseudouridylation pocket, the internal loop carrying the guide element(s). In contrast, a broad peak of coilin CLIP-tags was detected within the central region of box C/D snoRNAs, where base pairing is not predicted (Figure 7.2C). These patterns of coilin binding to snoRNAs with correlations based on structure suggest specificity. To identify putative binding motifs in the snoRNA sequences extensive motif search on

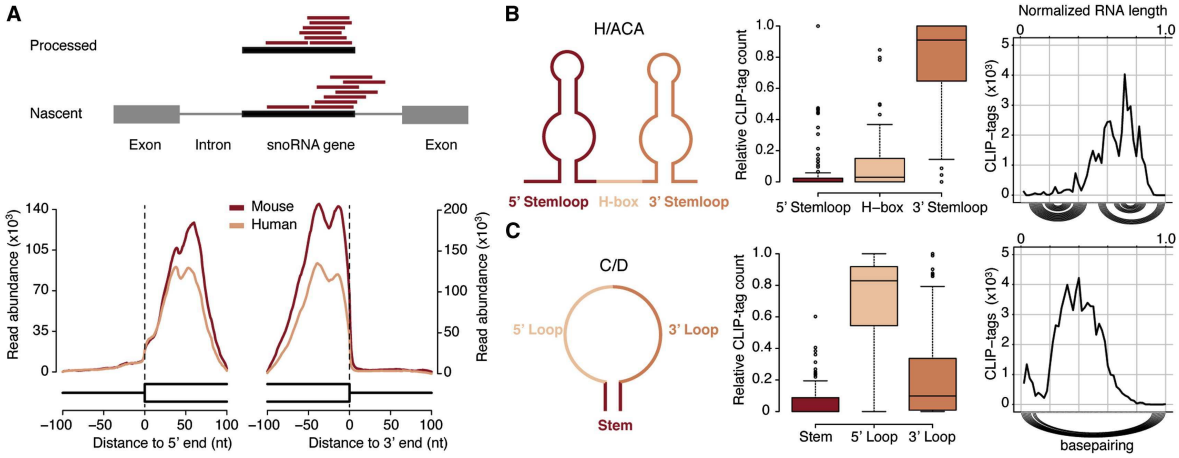

Figure 7.2: Coilin contacts specific regions within box H/ACA snoRNAs and box C/D snoRNAs. ***A****: The upper panel schematizes the possibility that coilin CLIP-tags could be associated with processed or nascent snoRNAs. Lower panel: meta-analysis for all detected mouse (red) and human (salmon) coilin CLIP-tags shows binding within snoRNA genes and not within surrounding intron sequence. Line plots represent total read coverage at each nt independently derived for the 5' and 3' ends of all snoRNAs.* ***B*** *and* ***C****: All human snoRNAs containing at least one significant CLIP-tag were folded with* `RNAsubopt` *and divided into three separate regions: 3' stem loop. H box and 5' stem loop for box H/ACA snoRNAs; stem 5' and/or 3' half of the loop for box C/D snoRNAs (schematic left panel). Boxplots show CLIP-tags abundance in each region relative ti the total CLIP-tag count in each snoRNA (middle panel). SnoRNA sequence lengths were normalized, such that 0 represents the 5' end and 1 the 3' end. Relative CLIP-tag count (y-axis) represents the sum of significant CLIP-tags for each of the normalized nucleotide positions. Traces below the plot indicate typical base-pairing patterns predicted within each snoRNA class (right panel).*

extended CLIP-tags in snoRNAs using two independent computational approaches was performed.

`MEME` In the extended CLIP-tags the characteristic snoRNA sequence motifs were masked. To avoid bias caused by a single snoRNA family with high copy number, also paralogous sequences from multi-copy families were removed. Then a `MEME` (Bailey et al., 2009) search was performed on the masked sequences for human and mouse separately. The search was performed on one set containing all types of snoRNAs and on sets in which box H/ACA snoRNAs and box C/D snoRNAs were separated. A variety of different `MEME` parameter settings were tested on each of the sets:

1. a maximum of 1 motif per sequence
2. exactly 1 motif per sequence

3. up to 10 motifs in the set

4. a motif length between 4 and 10 nt

Short but conserved sequence motifs (CTG) with low E-values of $2.6e-129$ (human) and $2.5e-119$ (mouse) were reported by the program (Figure 7.3). To clarify, this CTG motif can not originate from the D/D' boxes (CTGA), because these have been masked in the set.
For further validation the data was analyzed with a very new method specifically designed for CLIP data: `GraphProt` (Maticzka et al., 2014). Of the extended cross-link sites, the centering 12 nucleotides are considered as putatively interacting, while the surrounding nucleotides are used to identify the surrounding secondary structure. As `GraphProt` computes its binding model based on machine learning a negative training set of unbound sites was obtained by randomly picking a site from the surrounding host gene. Independent models where computed for box H/ACA and box C/D snoRNAs as well as for human and mouse.
This independent approach lends additional credibility to the structural influence of the binding site (Figure 7.3). On sequence level apart from a single cytosine residue, nothing can be observed.
In conclusion although no significant sequence motif for coiling binding was determined, there is, however, a preference for RNA binding within a short hairpin loop featuring a C at its central position.

7.3 Novel SnoRNA and ScaRNA Genes

Preliminary analysis of the CLIP data identified pre-mRNAs of protein coding genes (Figure 7.1B). However, the vast majority of this binding was specifically positioned in the introns, grouped in dense clusters over regions with elevated sequence conservation, which is very reminiscent of CLIP signal found in snoRNAs. It was hypothesized that these represent unannotated snoRNAs (Figure 7.4).
Coilin CLIP-tags clustered less than 5nt apart were combined and candidate intronic regions selected. These sequences were first screened for presence of conserved snoRNA box-motifs using position weight matrices. Secondly, fulfillment of snoRNA type specific structural constraints was checked by folding the sequence with `RNAsubopt` (Wuchty et al., 1999). The according filter criteria are illustrated in Figure 7.5.Position

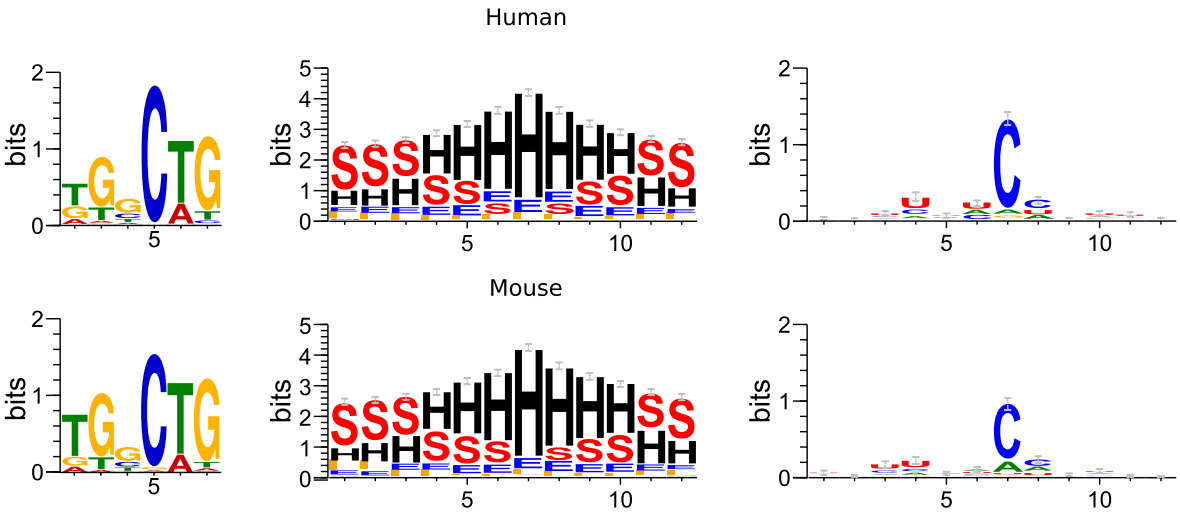

Figure 7.3: WebLogos of motifs found by `MEME` *in the vicinity of coilin CLIP-tags in (left panel) snoRNAs. Although conserved, the calculated E-values are not significant (see text).* `GraphProt` *motifs showing sequence motifs (middle panel) and structure motifs (right panel) in human and mouse for CLIP-tags. The motifs are conserved between box H/ACA snoRNAs and box C/D snoRNAs, and thus merged. Structure motifs are annotated as stems(S), external region (E), hairpin loops (H), internal loops (I), multiloops (M) and bulges (B). An over-represented C is present at the cross link site followed by C or U, surrounded by U nucleotides, located in a single hairpin loop enclosed by at least 2 base pairs.*

Weight Matrices (PWM) derived from vertebrate snoRNAs were used to score the sequences in a sliding window approach. In sequences considered as box H/ACA snoRNA candidates the variant H box (ANANNA) had to score 0.3 and the highly conserved ACA box 0.9 of the maximal score of the PWM. Additionally, the presence of two hairpins with a length of 40 nt is obligatory. In each of them at least eight base pairs have to be possible. box C/D snoRNA candidates were accepted if the C box motif (RTGATGA) scored 0.4 and the D box motif 0.8, relative to the respective maximal score of the PWM. The distance between both boxes need to be at least 40 nts, and at least three bp need to form the terminal stem.

The remaining candidates, i.e. human: 18 box C/D and 53 box H/ACA snoRNAs; mouse 14 box C/D and 11 box H/ACA snoRNAs, built the initial query-set for a conservation analysis in 47 vertebrate species with the `snoStrip` pipeline (Bartschat et al., 2014). It turned out that 13 box C/D snoRNAs, 26 box H/ACA snoRNAsand 3 scaRNAs (human), and 3 box C/D snoRNAs, 6 box H/ACA snoRNAs, and 3 scaRNAs

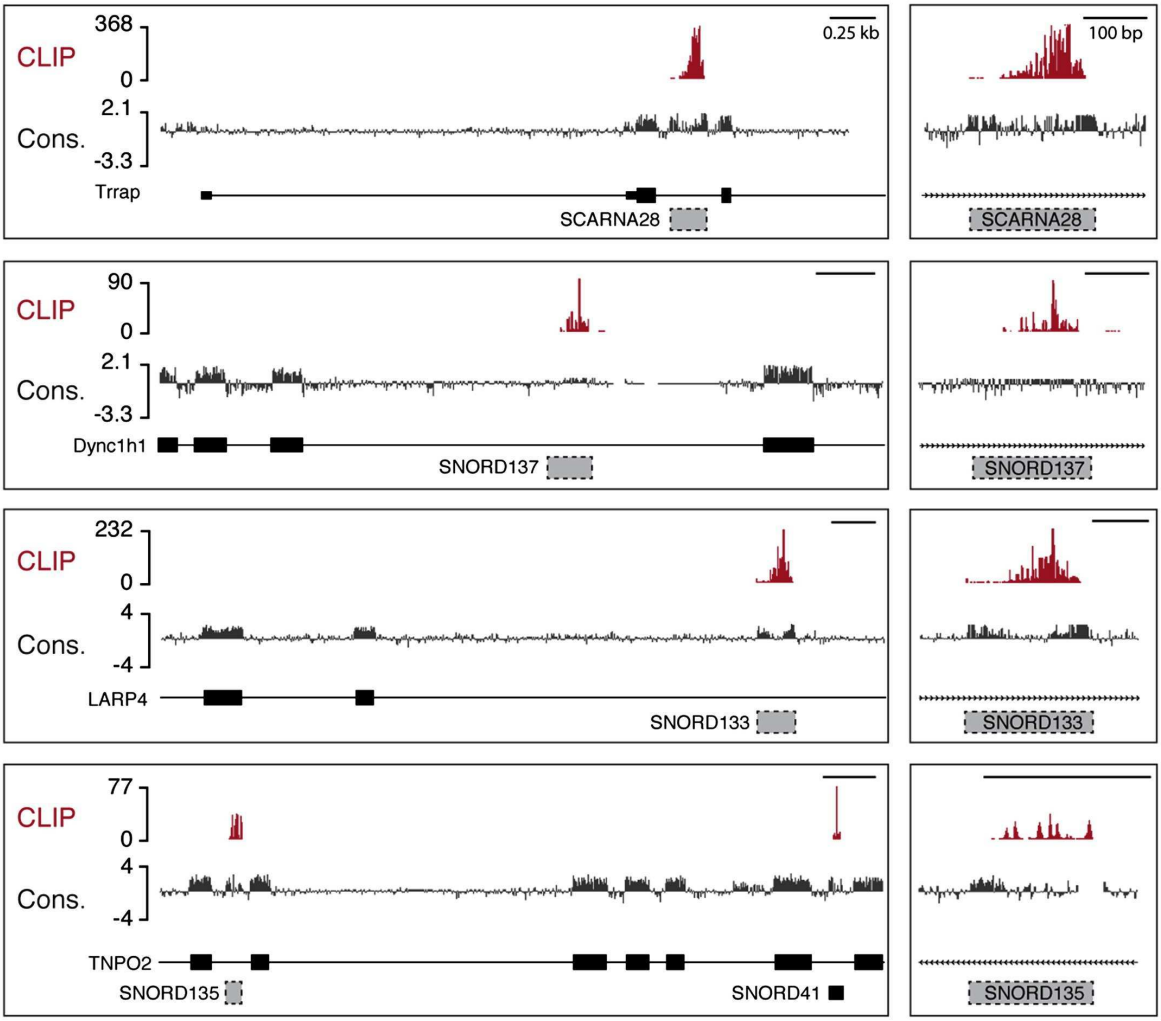

Figure 7.4: Coilin CLIP signals reveals mouse and human snoRNA genes. UCSC genome browser view of selected coilin CLIP signals assigned to introns of protein-coding gene, with location of novel snoRNA gene predictions (gray boxes). CLIP-tags (red) were often found positioned within highly conserved intronic patches.Mammalian conservation [Cons] in dark gray.

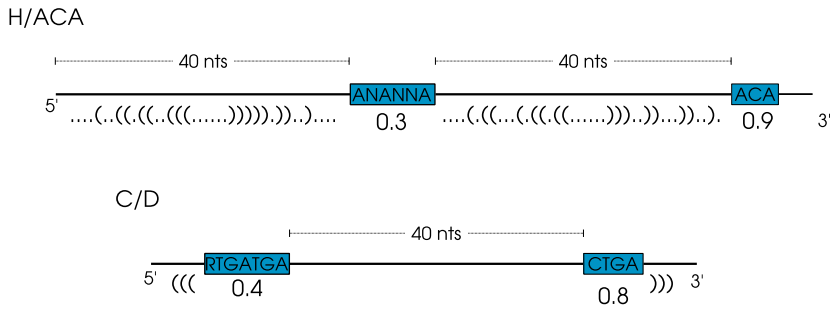

Figure 7.5: To identify novel snoRNAs in significant CLIP-tags within introns, the extended sequences were filtered for characteristic box motifs and structure. A: H box (ANANNA) was excepted with a score of 0.3, ACA box with 0.9 of the maximal score of the PWM. Additionally, the presence of two hairpins with a length of 40 nt and at least 8bps is obligatory. B: A C box (RTGATGA) had to score 0.4 and the D box motif 0.8, relative to the respective maximal score of the PWM. The distance between both boxes had at least 40 nts, and at least three bp need to form the terminal stem.

(mouse) show conservation in vertebrate species (Appendix Table A.2 and Table A.3). This further supports these snoRNA predictions. Official names for the high confident predictions have been assigned by the HUGO Gene Nomenclature Committee (Gray et al., 2013). To validate expression of these snoRNAs and compare their expression to previously annotated snoRNAs, RT-PCR and RNA-seq on poly(A)- RNA was performed (data not shown). The newly identified snoRNAs are expressed at lower but still significant levels, potentially explaining how they have escaped detection so far.

Afterwards the remaining reads that do not show snoRNA characteristics were analyzed further. Using the RepeatMasker track from UCSC led to exclusion of 1438 repeat regions in human (680 in mouse) and with GeneScan from UCSC further 59 predictions were discarded in human (628 in mouse). This leaves 315 unannotated reads in human and 87 reads in mouse, respectively. These were cross-checked with a recent study on lncRNAs that are processed in a snoRNA-like manner (Yin et al., 2012). No overlap with any of these sno-lncRNAs was found.

So the remaining reads were evaluated for their potential of being a non-coding RNA at all. First homologs of the reads were searched in five more vertebrates (human, mouse, chicken, dog and cow). Then the respective alignments were scored with RNAz (Gruber et al., 2010). Based on machine learning this approach distinguishes ncRNAs from other RNAs with respect to conserved secondary structure and sequence. In human 24 reads were conserved. Although all those reads show mean pair wise identity

larger than 80% RNAz did not predict a single ncRNA in this data. In mouse only a single read is conserved in sequence, however, its RNAz classification is 'other' with probability 0. There is no structural conservation in the according alignment.
This additional analysis shows that the additional CLIP-tags belong mostly to repetitive, often low-complexity sequences. There are no further sequences/RNAs that are well conserved or would have been found consistently between human and mouse. Hence, we can rule out that there is a coherent and evolutionarily conserved class of non-snoRNA coilin partners.

SnoRNAs Traffick Through Cajal Bodies

To further validate that classical snoRNAs with nucleolar functions pass the CBs, snoRNAs were fluorescent labeled and microinjected into HeLa cell nuclei by Martin Machyna in Karla Neugebauers lab. As expected, SNORD3 and SNORD118, and the scaRNAs were concentrated in CBs and nucleoli validating the approach. Remarkably, also the other tested coilin snoRNA interactors showed similar localization pattern including the novel identified. The scaRNAs, are retained in CBs by CAB-box elements that interact with the CB protein WDR79/WRAP53 (Darzacq et al., 2002; Deryusheva et al., 2012; Marnef et al., 2014; Tycowski et al., 2009) and appear only in CBs while prominent nucleolar localization was observable for sno- but not scaRNAs after ~ 60 minutes. The observations confirm that all snoRNAs traffick to CBs but scaRNAs are uniquely retained there.

7.4 New Target predictions for Novel snoRNAs

All novel snoRNAs were also tested for conserved complementarity to rRNAs and snRNAs following the approach described in Section 3.2.
For human and mouse, high scoring interactions (ICI values, higher than average for known snoRNA-target RNA interactions and good average minimum free energies (MFEs) among the species) are reported. For box C/D snoRNAs these values are: $ICI > 0.81$ and $MFE < -9.80[kcal/mol]$ and for box H/ACA snoRNAs: $ICI > 0.69$ and $MFE < -13.9[kcal/mol]$. For further support, we checked whether the appropriate modifications have been observed at the predicted positions and also whether another snoRNA guide was already assigned as complementary to this region.

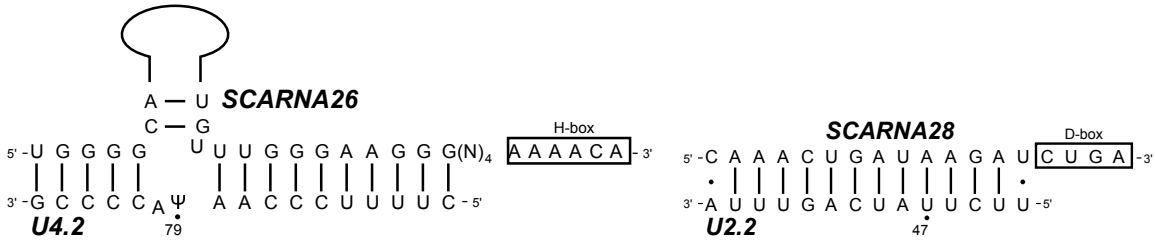

Figure 7.6: Interaction between SCARNA26A/B and U4-79 (left) and SCARNA28 and U2-47 (right). Both scaRNAs are newly detected within the iCLIP experiments in human and mouse. The Ψ at position 79 in snRNA U4 and the 2'-O-methylation of residue 47 have been reported by Karijolich and Yu (2010) and Dönmez et al. (2004)

Among the high scoring predicted modifications are U4-79 and U2-47. These sites have been reported as modified nucleotides in (Dönmez et al., 2004; Karijolich and Yu, 2010) but no guide had been assigned to them. The predicted guide for pseudouridylation of U4-79 is novel SCARNA26 (Figure 7.6). The interaction between SCARNA28 and U2-47 is described in the following section.

7.5 SCARNA28

A scaRNA with exceptional structure was captured in the iCLIP coilin experiments in human and mouse. It is a homolog of known chicken snoRNA GGgCD76 (Shao et al., 2009). It was found in Amniotes using the `snoStrip` pipeline and has also been reported as ZL1 in human by Kishore et al. (2013). With a length of 147-220 nucleotides (197 in human) it clearly exceeds the usual length of box C/D snoRNAs. It has a GT-repeat insert of differing length as has been observed for other scaRNAs with box C/D domains (Marz et al., 2011). Such GT-repeats have been shown to be the Cajal Body localization signal in scaRNAs with box C/D snoRNA structure (Kishore et al., 2013; Marnef et al., 2014; Machyna et al., 2014) equivalent to the CAB box in scaRNA box H/ACA domains. (Richard et al., 2003). In agreement deletion of the GT-repeat of novel SCARNA28 revealed that wild-type (WT) SCARNA28 is present only in CBs, while the mutant that lacks the CB localization signal additionally appears in nucleoli. An highly conserved guiding region adjacent to the D box is complementary to the a known methylated site at U2-47 (Dönmez et al., 2004) (Figure 7.6 right). The interaction has been simultaneously identified by (Kishore et al., 2013) and (Kehr

et al., 2014).
The iCLIP experiments with the Cajal Body specific protein coilin revealed direct interaction of coilin with numerous non-coding RNA. These include the expected snRNAs, histone mRNAs and scaRNAs, but also most known snoRNAs and several other small and long ncRNAs, which suggests a broader function of the CB in RNA maturation and RNP assembly. Different approach to identify the exact binding pattern have revealed that coilin contacts a small hairpin loop within the snoRNA sequences. In the portion of intronic CLIP-tags without annotated genes, several (42 human, 12 mouse) new snoRNAs belonging to the scaRNA, H/ACA and C/D class could be identified. Fluorescent labeling could verify these novel genes as stable transcripts that take the snoRNA typical routes in the cell. Intensive analysis of their function using the ICI score could assign some of them to experimentally detected modifications in rRNA and even to modifications in snRNAs for which no fitting snoRNA ASE was recognizable before. As such the snoRNA interaction network was further expanded, by finding new snoRNA members that could immediately add missing edges to unexplained modifications. On top the exceptional SCARNA28 was detected (in parallel to (Kishore et al., 2013; Kehr et al., 2014)). Like other scaRNAs it comprises a low complexity insert of differing length in vertebrates. It is the predicted guide for 2'-O-methylation of Uridine at position 47 in U2 snRNA.

CHAPTER 8

The Human SnoRNAome

The human snoRNA data resources that used to be standard in the field have either ceased to exist or to be updated. Additionally, recent studies (Kishore et al., 2013; Machyna et al., 2014) have demonstrated that our catalog of human snoRNA loci is far from complete. Especially, many non-standard snoRNA transcripts seem to have been overseen. The focus of the research community has moved towards characterization of snoRNA genes in species other than human. In order to construct an up-to-date catalog of human snoRNAs data from the existing resources was combined with *de novo* predictions. Therein the recent findings of snoRNA-like transcripts that share some but not all snoRNA characteristics were considered. Besides experimental validation cross-checking with small RNA-seq data from the ENCODE project, also the plasticity of snoRNA expression could be characterized. Consecutively as well as cell type specific expressed snoRNAs were characterized. Finally, the snoRNA target RNA interaction network was re-estimated. A newly developed high-throughput variant of the reverse-transcriptase-based method for identifying 2'-O-methylation in RNAs termed RimSeq was combined with previously reported modification sites and state-of-the-art target prediction methods. The study was a collaborative project with the Zavolan lab and has been published in *Nucleic Acids Research* (Jorjani et al., 2016). Official gene symbols for high confident novel snoRNAs were assigned by the HGNC. The new comprehensive data collection on human snoRNAs is provided in a basic database. The snoRNA Atlas can be accessed via `http://www.bioinf.uni-leipzig.de/publications/supplements/15-065`.

8.1 An updated catalog of human snoRNAs

To update the human snoRNA catalog, data from several sources were integrated. Specifically, snoRNAs were collected from RFAM-based predictions generated by the GENCODE consortium (Derrien et al., 2012), from deepBase (Yang et al., 2010), and from a `snoStrip` generated dataset in vertebrates (Bartschat et al., 2014; Kehr et al., 2014). Additionally, recently published literature was screened for further snoRNAs and sequences from Jády et al. (2012); Kishore et al. (2013); Zhang et al. (2014); Machyna et al. (2014) were added. It was checked for all snoRNAs if official gene names are available at the HGNC website[1] (Gray et al., 2015).

Table 8.1: Overview of known and novel snoRNAs analyzed in this study. Known snoRNAs are either listed in the HGNC collection or are extracted from recently published literature (Jády et al., 2012; Kishore et al., 2013; Zhang et al., 2014; Machyna et al., 2014; Kehr et al., 2014), and/or the public databases GENCODE and deepBase. Novel snoRNAs are those genes that do not overlap any of the known ones, naturally. In brackets, we provide counts of snoRNA-like processed transcripts, that do not fall into the 'common' snoRNA classes and should therefore gain a new HGNC prefix (pending). Other snoRNA candidates only partially fulfill the criteria for applying at the HGNC for gene names (details see text) and need further validation before HGNC gene names are requested.

	Known snoRNAs				Novel snoRNAs		
	Total	HGNC	HGNC Requests	Other	Total	HGNC Requests	Other
box H/ACA	179	136	+39/-3	4	11	9	2
AluACA	348	0	0	348	6	0	6
box C/D	376	295	48 (+4)	29	41	14 (+21)	6
SNORD-like	18	0	0 (+8)	10	98	0 (+98)	0
scaRNAs	29	27	2	0	0	0	0
TERC	1	1	0	0	0	0	0
sno-lncRNA	11	0	0	11	0	0	0

Moreover, a genome-wide screen for *de novo* snoRNAs was established, considering canonical and also non-canonical snoRNA transcripts. Especially, a recent high-throughput study identified very short snoRNA-like transcripts (Kishore et al., 2013). Due to the high computational demand of gene finding programs, first genomic regions

[1] `www.genenames.org`

that show expression in the sRNA-seq data set generated by the ENCODE consortium (Djebali et al., 2012) were selected. The extracted genomic regions were screened for potential snoRNA genes with `snoReport` (Hertel et al., 2008) and `snoSeeker` (Yang et al., 2006). Additionally, a search algorithm was implemented that screens for potential SNORD-like snoRNA genes (Kishore et al., 2013). The expressed sequences were screened for D and C boxes, allowing a few mutations. The distance between the box motifs could vary between 10 and 90 nucleotides. Due to the vast number of snoRNA candidates obtained from these sources, the candidates were further filtered. First, those that overlapped with repeat-annotated genomic regions with more than 25% of their sequence were excluded. This might have led to rejection of some true box H/ACA snoRNAs. However, subsequent validation through conservation analysis is not feasible for these candidates. Then a set of rules was established to accept those snoRNA candidates whose expression as mature forms is confidentially supported by the sRNA-seq data. In case of box C/D snoRNAs reads were considered as supportive if their 5' end is located 4-5 nts upstream of the C box and their 3' end 2-5 nts downstream of the D box. If the snoRNA exceeds the maximal read length of 100 nts, at least 75% of the snoRNA gene or 90 nts have to overlap the gene. A supportive read for a box H/ACA snoRNA ends diverge from the snoRNA locus by maximal 5 nts and the read covers 75% or 90 nts of the gene. In total, the collection of known and novel snoRNAs yielded 156 novel canonical and non-canonical human snoRNA sequences and 87 elsewhere reported snoRNAs that were not contained in the HUGO Gene Nomenclature Committee (HGNC) annotation (Table 8.1).

For all gathered snoRNA sequences conservation analysis was performed with `snoStrip`. As such the level of conservation throughout vertebrates, structure, box annotation, host genes and alignments are available for the whole set. In order to distinguish candidates which have relatively close homologs among the already known snoRNAs, each snoRNA was mapped against the RFAM-families. Sequences were assigned as family members when the $p-value < 10^{-5}$. (Griffiths-Jones et al., 2003; Nawrocki and Eddy, 2013). The comprehensive data including all features that were annotated, as well as read profiles, and target predictions is provided in a small searchable database at www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/15-065. Also gff-files, fasta-files and alignments can be downloaded there.

Remarkably, for most of the novel snoRNAs, homologs were only recovered in primates (93 box C/D snoRNA and 8 box H/ACA snoRNA). No homologs at all were retrieved

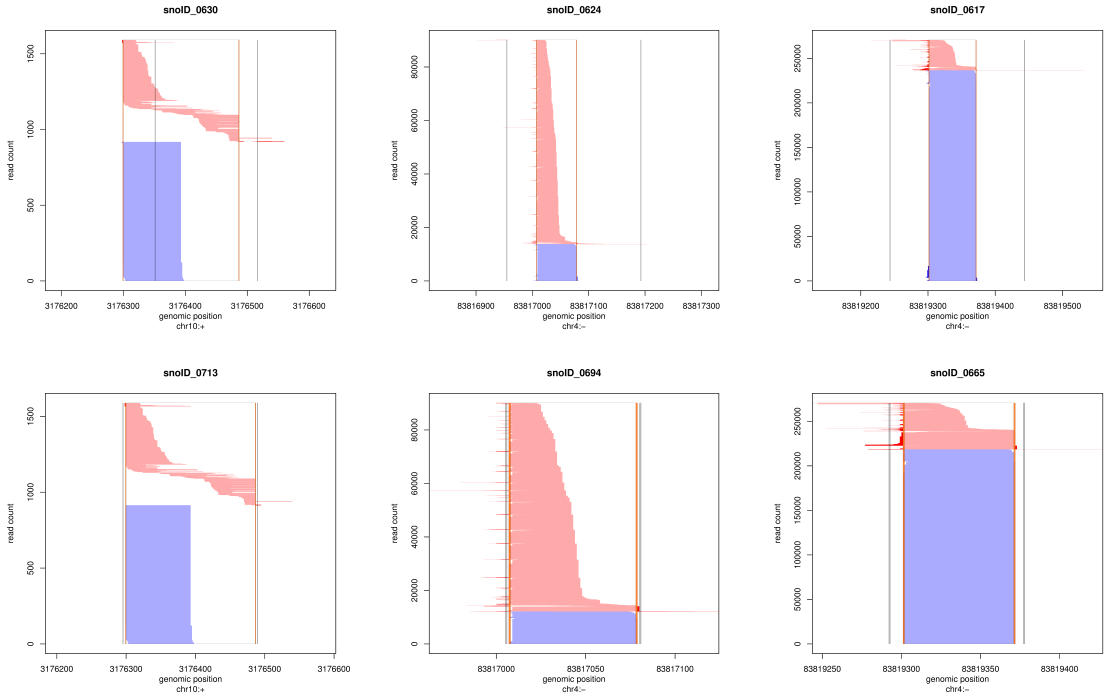

Figure 8.1: The read profiles visualize how the ENCODE small RNA-seq reads(Djebali et al., 2012) span the snoRNA loci. Blue reads are considered as supportive. Other overlapping reads are displayed in red. Grey vertical lines mark the gene borders, while orange vertical lines mark the expression start and end. The RNA-seq data uncover the actual ends of the snoRNAs. The box H/ACA snoRNAs SNORA85, SNORA96, and SNORA97 (from left to right in upper panel) were revised as box C/D snoRNAs SNORD142, SNORD143, SNORD144 (from left to right in lower panel).

for 13 box C/D snoRNAs and 7 box H/ACA snoRNAs. Reliably determining if these snoRNAs are indeed evolutionary new inventions, specific to human and primates, is beyond the current methodology. However, a lineage specific pattern is documented for miRNA (Stark et al., 2007; Ladewig et al., 2012; Ruby et al., 2006) and has already been suggested for snoRNAs by Zhang et al. (2010). Though, the previous article has sharply been criticized by Makarova and Kramerov (2011).

The read profiles composed in the study were used to refine snoRNA genes. For several snoRNAs the gene ends were updated. Additionally, three snoRNAs that have been annotated as box H/ACA snoRNAs (SNORA85, SNORA96, SNORA97) in the coilin cross-linking study (Chapter 7) had to be revised as box C/D snoRNAs (SNORD142, SNORD143, SNORD144). The according sequences comprise both types of box pairs.

Only the actual borders of the transcript, which became apparent through the established read profiles enabled to correctly categorize them (Figure 8.1). The overall relatively low expression of the snoRNAs identified in the coilin study can be explained by a putative more special expression of these snoRNAs genes in the cell, which might not be captured by the ENCODE data. Low expression and abnormal read profiles were also revealed for several members of the multi-copy families SNORD113, SNORD114, SNORD115, and SNORD116, which are encoded in imprinted or close to imprinted genomic regions in human and a few other loci.
Official HGNC gene symbols were allocated for novel snoRNAs that show evolutionary conservation in hominids and beyond, contain all expected sequence motifs, are found expressed as full-length snoRNAs in human and fold into a canonical structure (H box, ACA box and hairpin-hinge-hairpin-tail structure for box H/ACA snoRNAs, and C box, D box, the typical kink-turn formed by these boxes, and a terminal stem of at least 2 bps for box C/D snoRNAs).

8.2 An updated catalog of human snoRNA targets

The primary function of snoRNAs is to guide the modification of specific sites in ribosomal and spliceosomal RNAs. To provide an up-to-date annotation of the targets in the human snoRNA catalog, state-of-the-art computational methods (Bratkovič and Rogelj, 2014) were combined with experimental data on snoRNA-guided RNA modifications.
The extensive computational target analysis follows the workflow described in Section 3.3. First, homologs of the gathered human snoRNA sequences are annotated with `snoStrip`. To predict the snoRNA targets, `RNAsnoop` (Tafer et al., 2010) and `PLEXY` (Kehr et al., 2011) (Section 3.2) were applied to the set of snoRNA sequences and of target RNAs (Section 3.1). Hereby, considering primary sequence features, secondary structure of the snoRNA, the accessibility of the target region, and the predicted minimum free energy of the snoRNA-target duplex. Then evolutionary conservation of the predicted interaction within vertebrates was evaluated using the Interaction Conservation Index (ICI) (Kehr et al., 2014) (Section 3.2.3). As reminder, the ICI combines stability of an interaction between snoRNA and target RNA within a single species with the range of conservation of an equivalent interaction among species in which a snoRNA homolog was identified. Roughly, an ICI score > 1 can be interpreted as

the specific interaction being better than alternative predictions in all species where a snoRNA homolog is present. Further a coarse-grained encoding of the conservation that indicates the depth of conservation in the phylogenetic tree of eukaryotes was considered. Based on these sets in-depth analysis combining the predicted targets with known information on true modifications were possible.
Information about target sites was gathered with respect to three categories for each snoRNA antisense element. First, any previously reported target site (r). Second, the best scoring human target prediction (h_1) within the set of human target predictions considering the minimum free energy of the snoRNA-target RNA interaction duplex. And third, the best scoring conserved target prediction (c_1) within the set of conserved interactions evaluated by the Interaction Conservation Index. The final assignment of a snoRNA antisense to a target site was based on following rules:

1. h_1, if the best scoring conserved target is best scoring human target ($c_1 = h_1$)

2. r, if the reported target is best scoring human target ($r = h_1$)

3. c_1, if the reported target is not the best scoring human target ($r \neq h_1$); and a human target prediction (h_i) exists within the best scoring conserved target predictions ($h_i = c_1$)

4. h_1, if no human target prediction exists within the best scoring conserved target predictions ($h_i \neq c_1$).

Selected interactions were accepted, if the interaction is well conserved in deuterostomes with an $ICI > 1.0$ for box C/D snoRNAs and an $ICI > 0.8$ for box H/ACA snoRNAs (compare Chapter 6 for information on these thresholds). A predicted interaction was classified as highly confident if the resulting modification overlaps a confirmed modified position, that has been identified by a high-throughput approach, or has been reported in literature.
Finally, all snoRNA sequences were classified as orphan, single guides or double guides. Orphan snoRNAs lack identifiable guiding function, single guides have an interaction assigned at one ASE and double guides at both ASE.
In the target analysis box C/D snoRNAs, box H/ACA snoRNAs, and small cajal body snoRNAs were separately treated. For box C/D snoRNAs it was further distinguished between canonical box C/D snoRNAs, non-canonical SNORD-like genes and

snoRNAs that belong to the multi-copy families SNORD113-SNORD116. The numbers of snoRNAs in the different datasets are listed in Table 8.2 (column Total). Besides also the numbers of identified antisense elements (ASE) harbored in the data sets are listed.

Table 8.2: For each category we provide total (full), known (upper) and novel (lower) sequence counts. SnoRNAs can comprise two ASEs, all box H/ACA snoRNAs, and box C/D snoRNA sequences for which D and D' box were identified, or one ASE, box C/D snoRNAs where the D' box is too variant to recognize and only the D box was annotated. Note that for simplicity the SNORD3 (13 members) and SNORD13 (11 members) families are not listed.

	SNORD-like	Multi-copy	Canonical box C/D	ALUACA	Canonical box H/ACA
Total known	116 18	118 118	275 234	354 348	190 179
Total novel	98	0	41	6	11
2 ASE identified known	25 0	113 113	216 201		190 179
2 ASE identified novel	25	0	15		11
1 ASE identified known	91 18	5 5	59 33		
1 ASE identified novel	73	0	26		

Box C/D snoRNAs For box C/D snoRNAs target prediction the SNORD3 and SNORD13 snoRNA families that have established non-canonical functions in pre-rRNA cleavage (Cavaillé et al., 1996; Kass et al., 1990) were not considered. Hence, we obtained a total of 393 snoRNA sequences, of which 275 are canonical box C/D snoRNAs another 118 are members of the multi-copy (mc) gene families SNORD113, SNORD114, SNORD115, and SNORD116. The regions adjacent to the D and D' box define the ASEs. For the majority (83%) of the canonical box C/D snoRNA sequences both box motifs could be annotated (Table 8.2). Thus both ASEs were considered for target prediction. In contrast, only a few (21%) of the SNORD-like snoRNAs appear to possess both D and D' boxes. In many cases the D' box could not be reliably annotated either due to the short length of these snoRNA like genes or the lack of evolutionary conservation hinders reliable identification of the variant sequence motif in the alignments. In total, we applied target prediction to $863 (= (25 + 113 + 216) * 2 + (91 + 5 + 59))$ antisense elements (ASEs) covering all cataloged box C/D snoRNAs and all SNORD-like snoRNAs. The snoRNA target prediction results are also listed in www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/15-065.

With our method we were able to increase the percentage of functional ASE in the 234 known box C/D snoRNAs from previously reported 46% to 56%. On the level of whole snoRNA sequences almost 20% (34) of the snoRNAs had no identified function, afterwards only 5% (9) of the sequences remain orphan. The majority of reported human interactions 76% have high ICI values, meaning that the interaction is conserved in other vertebrate species. For all newly accepted target predictions, conservation is a prerequisite.

Counting also the 41 novel canonical box C/D snoRNAs, which obviously had no targets assigned before, 82% (227/275) have a predicted target in rRNA or snRNA (Figure 8.2A). Although most of the known and novel box C/D snoRNAs have a D and D' box, only a minority of those 17% (38) indeed interacts with targets at both antisense elements. Most sequences with reported function 83% (189) are single guides with their target-RNA complementary either at their D- or at their D'-box. Still 48 box C/D snoRNAs with canonical features remain without a predicted or known target in rRNA or snRNA. Of these, SNORD97 is reported as enriched in chromatin-associated RNAs (caRNAs) (Dupuis-Sandoval et al., 2015).

Multi-copy box C/D snoRNAs Because a detailed analysis of the mc snoRNA families did not reveal convincing target predictions, these families are separately treated. The four box C/D snoRNA families: SNORD113, SNORD114, SNORD115, and SNORD116 have 118 members. These families are famous for their unusual genomic organization in imprinted regions in multi-copy manner. Among them are SNORD115 and SNORD116 from the SNURF-SNRPN locus associated to the Prader-Willi syndrome. The sequences show neuron specific expression. None of the families has reported targets in ribosomal or spliceosomal RNAs. However, SNORD115 and SNORD116 have been shown to change expression of multiple genes (Falaleeva et al., 2015) and SNORD115 has been shown to be involved in alternative splicing of the serotonin receptor mRNA (Kishore and Stamm, 2006) and several other mRNAs (Kishore et al., 2010).

The majority of the sequences remain orphan after target prediction. All SNORD115 sequences lack complementarity to considered target RNAs. Some members of the other families show complementarity to target RNA regions in one or even both (SNORD114) ASEs. In general the picture remains fuzzy as the predictions are not consistent within the families, some paralogs deviate and there are over-proportionally many complementary regions in the target RNAs. Thus no convincing targets in rRNA and snRNA

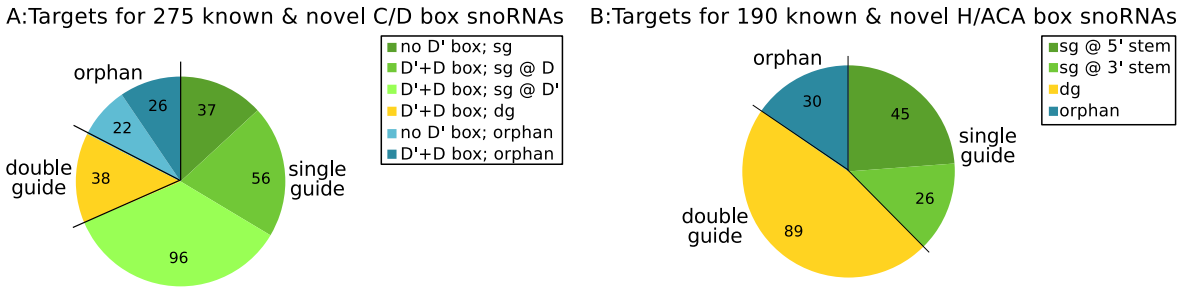

Figure 8.2: Distribution of orphan, single guide (sg), and double guide (dg) among known and novel snoRNAs based on our target predictions. A: Of the 275 canonical box C/D snoRNAs, 48 are orphan, 38 are double guide and 189 are single guide. Of the latter, 93 (56 D'+D box, +37 no D' box) have a functional ASE adjacent to the D-box and 96 adjacent to the D'-box. B: Of 190 canonical box H/ACA snoRNA sequences 30 remain orphan (of which SNORA73A/B have a non-canonical role in 18S rRNA maturation (Fayet-Lebaron et al., 2009)), 89 are double guide and 71 are single guide. Of these 45 have a functional ASE in the 5' stem, and 26 in the 3' stem.

sequences could be identified.

SNORD-like snoRNAs More than half (73) of the 116 box SNORD-like snoRNA sequences are shorter than 50 nucleotides. Such short snoRNA genes have first been identified by Kishore et al. (2013). Due to their shortness they do not comprise the prime box pair, and thus harbor only one putative antisense element each. Against expectation for 77 of the non-canonical snoRNAs high scoring interactions with ribosomal or spliceosomal RNAs were predicted. Of these 47 involve the short SNORD-like genes. The interactions, as the snoRNAs themselves, are mostly identified in primates only. Strikingly, a share of $\sim$ 30% of these predictions derive 2'-O-methylation in snRNA, of which the majority (80%) are located in the snRNAs of the minor spliceosome. In all newly predicted interactions for *bona fide* snoRNAs the share of snRNA targets is only $\sim$ 13% (31/237) of which not even half are located in the snRNAs of the minor spliceosome. (According target predictions and interaction Figures are provided at www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/15-065.)

One such example is SNORD-like snoID_0373 (37 nts). It has an eight nts long ASE that is predicted to guide the modification machinery to the reported methylation of U12-7 in human and chimp (Figure 8.4 left). The extra methyl-group at the uridine has been identified by Dönmez et al. (2004), but so far no matching snoRNA guide could explain the modification.

Interestingly, a methylation is also predicted for an observed pseudouridylated residue.

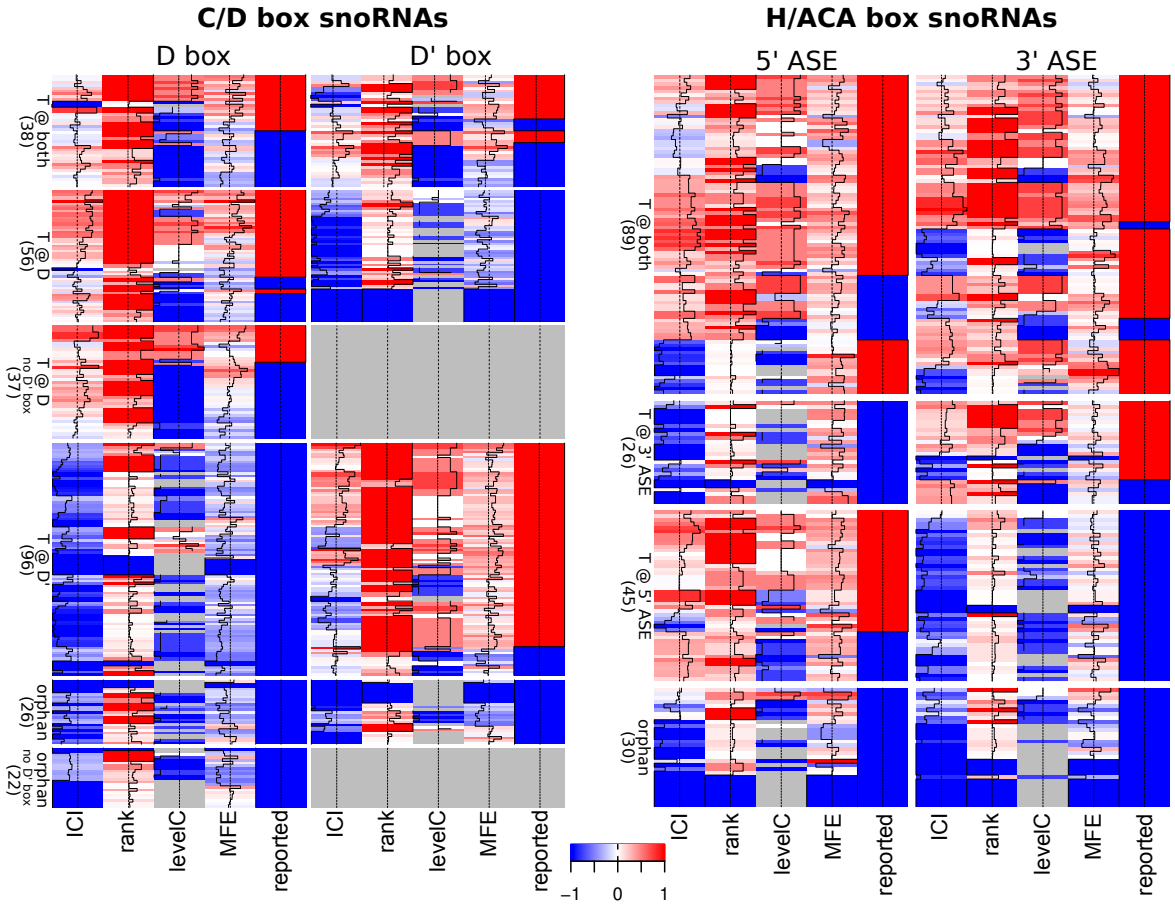

Figure 8.3: Visualization of predicted snoRNA-target RNA interactions. Each row displays binding properties of one snoRNA sequence. The columns represent: ICI: interaction conservation index of selected target (scaled to [-1,1]); rank: rank of selected interaction within set of predictions for the snoRNA ASE in the human genome (1:respective interaction is best for this ASE, else:1/(log of rank) (scaled)); levelC: level of conservation of the interaction among deuterostomes (1:primates, 2:eutherians, 3:therians, 4:mammals, 5:amniotes, 6:tetrapodes, 7:tetrapodes and teleosts, 8:vertebrates, 9:deuterostomes, scaled to [-1,1], gray denotes human-specific); MFE: minimum free energy (MFE) of the interaction (scaled to [-1,1]); reported: target reported in the literature (1:yes, -1:no). For each cell, the value is also illustrated by the position of the vertical black line relative to the 0-value line, located in the middle of the cells. Appendix Figures A.13 - A.22 provide the heatmaps in higher resolution with snoRNA names next to the rows. A: box C/D snoRNAs (not including multi-copy and C/D-like snoRNAs). The left set of columns refer to the ASE upstream of D box; the right set of columns the ASE adjacent to D' box (a gray line if the D' box and the particular ASE could not be annotated). B: box H/ACA snoRNAs (not including the ALUACA class). The left set of columns refer to the 5' stem-associated ASE, whereas the right set of columns refer to the 3' stem-associated ASE.

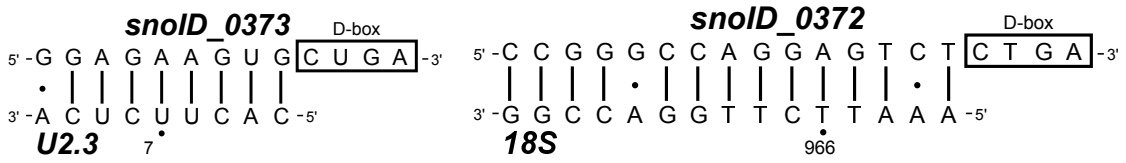

Figure 8.4: Predicted interaction between SNORD-like snoRNA snoID_0373 and region around methylated thymine at U2-7 (left) and SNORD-like snoID_0372 and region around Ψ at 18S-966 (right).

The snoRNA snoID_0372 is predicted to guide methylation of the Ψ at 18S-966 in primates (Figure 8.4 right). The pseudouridylation is reported to be guided by SNORA14. Methylation of a Ψ has previously been reported at 28S-3797.

As in our previous studies redundant guides were observed for known methylated residues that already have snoRNA guides assigned: snoID_0356 is predicted to interact with the methylated uridine at 18S-121 (also targeted by SNORD4); snoID_0427 putatively guides methylation of 28S-2338 (also targeted by SNORD24), snoID_0417 is complementary to 18S-1442 (SNORD61) and ID_0337 can hybridize with 18S-1326 (targeted by SNORD33).

The details of snoRNA-target RNA interactions were depicted as heatmaps in Figure 8.3 (See Appendix Figures A.13 - A.18 for a high resolution version with snoRNA ID next to each row.). Blue and red colors indicate low and high evidence for the interaction, respectively. It is apparent that after our analysis only a small fraction of snoRNAs remains orphan, which is indicated by the blue color in the column reported and by a low value of the interaction conservation index (ICI) for both ASEs. Several interactions, mainly for the newly identified snoRNAs, seem to be primate specific (column levelC: blue and column ICI: white/red). Interestingly, box C/D snoRNAs seem to have a single-guide tendency (column ICI is white/red for either D or D' box, but relatively rarely for both). For the 59 snoRNAs for which we could not identify a D' box, the classification as single-, double-guide or orphan snoRNA remains preliminary (gray cells on D' box side). Although the majority of box C/D snoRNAs encode both a D and a D' box and have associated ASEs, for only 17% high-scoring interactions were predicted for both ASEs. Among single-guide box C/D snoRNAs, the predicted interaction preferentially involves the D' box-associated ASE (96 cases vs. 56 with guiding at the D box-associated ASE). This is in strong contrast to the pattern of evolutionary conservation, since the D box shows generally stronger conservation.

Box H/ACA snoRNAs For box H/ACA snoRNA target prediction, only canonical genes were considered. Sequences encoded within Alu repeats (AluACAs) were excluded. For these evolutionary conservation information cannot be reliably obtained. In the canonical box H/ACA snoRNA structure each of the two stem-loops possesses a pseudouridylation pocket with respective bipartite ASE element. This made a total of 380 ASEs (Table 8.2) that were considered for target prediction. In total, counting known and novel canonical box H/ACA snoRNAs the analysis associated 85% of the 190 box H/ACA snoRNA sequences with at least one target Uracil in an rRNA or snRNA (Figure 8.2B and Figure 8.3). In contrast to box C/D snoRNAs, box H/ACA snoRNAs predominantly (56%) double guides. For those with one guiding ASE, the ASE is preferentially located in the 5' stem (45 of cases compared to 26 that have the single guiding ASE in the 3' stem). Among the 30 snoRNAs for which no canonical targets were reported or predicted is also SNORA73A/B. This is in agreement with the reported non-canonical interaction of the yeast homolog snR30 with the 18S RNA (Fayet-Lebaron et al., 2009). The conserved potential for base-pairing of these molecules suggests that the mechanism is well conserved to vertebrates. Furthermore, there is evidence that SNORA73A functions as a putative regulator of chromatin function (Dupuis-Sandoval et al., 2015). The higher amount of double guides, in contrast to box C/D snoRNAs, seems explainable by the overall higher evolutionary pressure on the binding pockets through structure constraints for the hairpins that comprise the interior loops.

ScaRNAs The human scaRNAs can be grouped into tandem box C/D (4), tandem H/ACA box (1), hybrids of box C/D and box H/ACA domains (5), canonical box C/D (2) and canonical box H/ACA (17). Thus, the pool of scaRNAs can potentially interact with target RNAs at $78 = ((4 + 1 + 5) * 4 + (2 + 17) * 2)$ sites. Due to their intricate structure, we could not reliably annotate all potential ASEs for six scaRNAs, leaving a total of 71 ASEs that were subjected to further analysis. An evolutionarily conserved target could be recovered for 43 cases including seven sites that are newly predicted.

SCARNA21

SCARNA21 has previously been annotated as scaRNA with 'common' H/ACA structure. Two paralogs are present in the human genome, one on chr17 and one on chr1. Kishore et al. (2013) have detected that the sequence on chr17 is embedded in a conserved box C/D part. With `snoStrip` homologous sequences of the long and the short version were found to be present in eutherian species.

Computation of the consensus structure revealed a well conserved structure comprising all snoRNA characteristic elements. A perfect kink-turn motif between the C/D part as well as two stable hairpins in the H/ACA part (Figure 8.5). All structural elements are either fully conserved or supported by compatible or co-varying mutations. Even a prime-box pair is clearly detectable within the embedding C/D part. The elongated isoform of SCARNA21, was found to harbor three additional functional ASEs (Figure 8.5).

The guiding regions (guide2l and guide2r in Figure 8.5) in the interior loop of the 5'-hairpin have already been reported to bind to spliceosomal RNA U12 in human. Hybridization of both RNAs anchors the Uracil at position 18 available for the pseudouridine synthase of the snoRNP complex. The interaction is well conserved in 24 eutherian species ($ICI_{sno} = 1.16$). Interestingly, the newly identified guiding region adjacent to the D'-box (guide1), is complementary to the same region of U12. The resulting 2'-O-methylation of Guanine at position 17 has been detected by Deryusheva et al. (2012). This interaction is conserved within 17 eutherian species ($ICI_{sno} = 1.11$). For the 3'-hairpin no widely conserved target is predicted. Nevertheless, in human, (guide3l and guide3r) is complementary to the region around U6atac-83. This has been reported as the only experimentally verified Ψ in human U6atac (Deryusheva et al., 2012). In spite of high variability of the scaRNA-ASEs that was observed in the alignment this interaction is also predicted for galago, mouse lemur and horse. For the ASE adjacent to the D'-box (guide4) a moderately conserved interaction with 28S rRNA is predicted ($ICI_{sno} = 0.71$). The known guide for the according 2'-O-methylation of Cytosine at position 4426 is SNORD49. Putatively contributing three modifications in RNAs of the minor spliceosome (U12-17, U12-18 and U6atac-83), we concluded that the elongated form of SCARNA21 is involved in maturation and therefore proper function of the minor spliceosomal machinery.

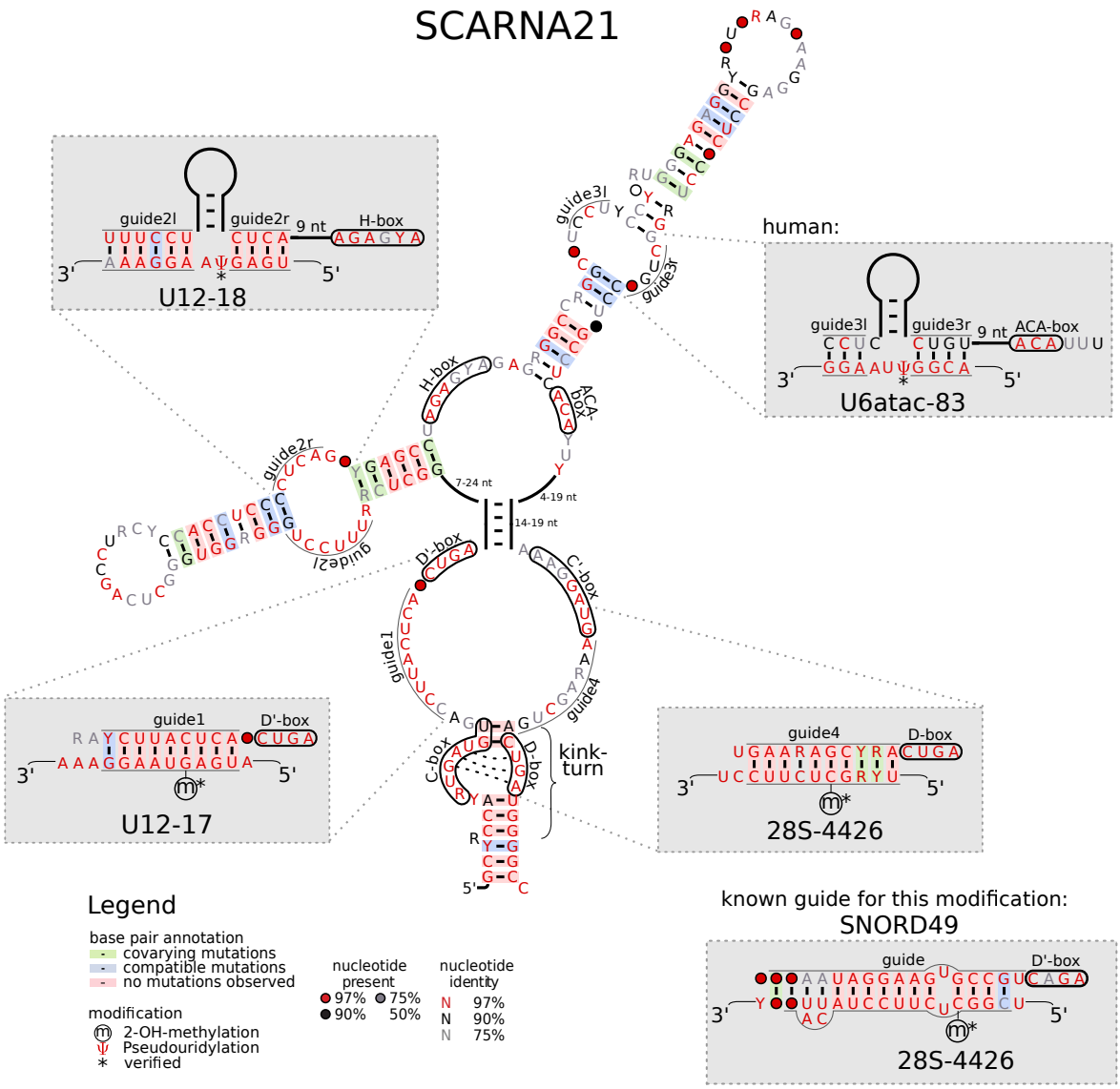

Figure 8.5: Structure of elongated SCARNA21. Characteristic sequence motifs are circled in black. The box C/D sequence parts form the characteristic terminal stem and the obligatory kink-turn motif. The H/ACA part folds into the typical double-hairpin structure. Putative functions for the ASEs are displayed in the gray boxes. From 5' to 3' the predicted functions are: guide1: U12-17, guide2r&l: U12-18, guide3: U6atac-83, guide4: 28S-4426. See text for details about these interactions. The figure is produced by R2R.*(Weinberg and Breaker, 2011)*

Using the computational predictions and the data obtained from high-throughput experiments and modifications reported in literature ten novel high confidence interactions between canonical snoRNAs and target molecules were identified. For two target sites whose methylation has been reported to be guided by a known snoRNA an additional guiding snoRNA was predicted: the D'-box ASE of SNORD136 for 18S-683, and snoID_0337 for 18S-1326. Additionally, the methylations that were experimentally identified at 18S-1606 and 18S-1410 could be assigned to previously considered orphan snoRNAs SNORD73A/B and to novel snoRNA snoID_0340, respectively. Guiding H/ACA snoRNAs can putatively guide the two previously mapped pseudouridylation sites: 18S-681 and 28S-4266. Concerning the pseudouridylation sites that emerged from high-throughput data, a guiding snoRNAs could be predicted in three (18S-1046, 18S-1232, and 28S-2619) out of the four cases. We could not identify a guiding snoRNA for the pseudouridine at position 1177 in human 18S rRNA reported by (Carlile et al., 2014). Details of this analysis are summarized in Table 8.3.

Here, a careful collection of snoRNA sequences in human was established. Known snoRNAs from several sources were combined and supplemented by further *de novo* predicted sequences that show reliable expression in the ENCODE small RNA data. With the *de novo* search it was possible to detect also non-canonical snoRNA transcripts that occurred in recent studies. Application of the `snoStrip` pipeline also annotated the characteristic features of all collected snoRNAs. Thus a comprehensive up-to-date catalog of human snoRNAs, including expression profiles, evolutionary conservation, host genes, boxes, alignments, structure, and target prediction is provided on www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/15-065 under the term human snoRNA Atlas. Summarizing results obtained from target prediction and reported interactions, it is possible to associate more than two thirds (70%) of the box C/D snoRNAs and 85% of the box H/ACA snoRNAs with a specific rRNA or snRNA target. However, 118 box C/D snoRNA and 30 box H/ACA snoRNA genes remain classified as orphan. Only 48 box C/D snoRNAs remain orphan, after exclusion of the multi-copy family members. For some of these a special function in affecting alternative splicing of mRNA has already been shown. It is quite possible that their special encoding and expression profiles also reflect a special function not related to chemical modifications in rRNAs and snRNAs. Counting on ASE level also the single guides

Table 8.3: List of predicted interactions between guide snoRNAs and nucleotides whose modification has been confirmed experimentally. The modification data originated either from `snoRNA-LBME-db`[2]*, in which case guide snoRNAs were sometimes already assigned, or from the high-throughput (HTP) approaches, in which case the guiding snoRNAs were not known so far. We further provide the location of the ASE which is predicted to take part in the interaction, the Interaction Conservation Index (ICI) of the interaction and the conservation level of the predicted snoRNA guide.*

Modification	Assigned Guide	HTP	Predicted Guide	ASE	ICI	Conservation Level
18S-683	SNORD19	-	SNORD136	D'	1.22	Eutherians
18S-1326	SNORD33	+	snoID_0337	D'	1.84	Primates
18S-1410	NA	+	snoID_0340	D	1.33	Primates
18S-1606	NA	+	SNORD73A/B	D'	1.17	Tetrapodes and Teleostes
18S-681	unknown	+	SNORA14A/B	5' stem	0.84	Amniotes
18S-681	unknown	+	SNORA55	3' stem	1.2	Tetrapodes
28S-4266	unknown	-	SNORA78	5' stem	0.92	Tetrapodes and Teleostes
18S-1046	NA	+	SNORA57	3' stem	0.9	Deuterostomes
18S-1232	NA	+	SNORA70A/B/E SNORA70-11/14	5' stem	1.18	Vertebrates
28S-2619	NA	+	SNORA38A/B	5'-stem	0.84	Therians

that have both boxes annotated, might have a second undetected function at their free ASE. Here, the question remains, if these snoRNAs interact with their target RNAs in a way that fails to be recognized by our computational target prediction methods or if these snoRNAs execute biologically different function than guiding modifications.
Interestingly, among the modifications predicted for the non-canonical SNORD-like sequences the snRNAs of the minor spliceosome (U4atac, U6atac, U11, U12) have an extraordinary high share. Additionally, these snoRNAs and according function are often observed in primate only. Most intriguingly, the results suggest an important role of SCARNA21 in the maturation of snRNAs of the minor spliceosome.
With the study again the network of snoRNA target RNA interactions in human was extended. It was possible to suggest functions for many of the novel snoRNAs as well as assign snoRNA guides to three previously reported 'orphan' modifications and five modifications identified by high-throughput methods during this study.

CHAPTER 9

Discussion and Outlook

Prior to this work collections of snoRNAs existed only in individual species. The most valuable data source for human snoRNAs is `snoRNA-LBME-db` (Lestrade and Weber, 2006). Unfortunately, information about characteristic features of the sequences although contained in the database is not intended for automated querying. Also the database ceased to be updated and is missing out the technological advances of next generation sequencing. SnoRNA sets in species other than human had to be collected from literature or from general databases that do not capture snoRNA specific features. Automated snoRNA annotations do generally not extract and filter for characteristic features, consequently containing several non-functional pseudogenes and false positive predictions. On top different naming conventions used in the studies additionally obscure homologies. Apart from human (Lestrade and Weber, 2006) and chicken (Shao et al., 2009), only few studies focused on interactions between modification sites in target RNAs. Although modifications of equivalent sites in species as divergent as yeast and human have been observed, no systematic analysis of the conservation of the guiding relation between snoRNAs and their targets was performed.

Prediction of the targeted modifications is a challenging task as the interaction typically involves only a short region, making it necessary to take additional signs of evidence such as evolutionary conservation into consideration. There have been assumptions that snoRNAs frequently change targets (Shao et al., 2009) and have lineage specific functions (Zemann et al., 2006), while other studies assume conserved functions for snoRNA families (Hoeppner and Poole, 2012). For a number of reasons previous studies were unable to investigate snoRNA function over time in such detail. They were based on a more limited set of species, and/or snoRNA sequences, and had no formal method

to evaluate conservation of targets.

This work contributes tools for extensive analysis of snoRNAs. These include a sophisticated homology search pipeline for snoRNAs, called `snoStrip` (Bartschat et al., 2014). Due to particularly considering obligatory snoRNA features it is able to differentiate between functional and non-functional genes. The features that were extracted are stored in a comprehensive database, termed `snoBoard` for automated processing. Further, based on thermodynamic modeling of the interactions between snoRNAs and their targets in single sequences (with our programs `RNAsnoop` (Tafer et al., 2010) and `PLEXY` (Kehr et al., 2011)) a measurement to score the conservation of these interactions among a set of species was conceived and named Interaction Conservation Index (ICI) (Kehr et al., 2014). On the one hand supporting relatively short target predictions in single species through conservation signals and on the other hand enabling to study the general pattern of co-evolution between snoRNA guides and their targets. Integration of all tools into an innovative workflow allows to study and expand the snoRNA interaction network.

`SnoStrip` can now be regarded as state-of-the-art snoRNA homology annotation pipeline. With its application we contributed the snoRNA genes to several RNA annotation projects for newly assembled 48 avian species (Gardner et al., 2015) and spotted gar (Braasch et al., 2016)) reflecting the usability and the benefits of our methods. Meanwhile, it was used to build comprehensive sets of snoRNA sequences among vertebrate species (Kehr et al., 2014), plant genomes (Patra Bhattacharya et al., 2016) and fungi (Canzler, 2016). The effective storage of all derived snoRNA related information further improves the wealth of these sets and helps to answer more in-depth questions on snoRNAs, their features, and their evolution. In this sense this work focused on snoRNA function and its evolutionary aspects.

Some snoRNAs, that deviate from the norm in terms of sequence length and exceptional structure composition were studied in more detail (Marz et al., 2011). Especially, the evolutionary origin of these unusual transcripts was focused. In case of four hybrid scaRNAs that comprise a box H/ACA domain embedded into a box C/D domain two orthologous pairs were identified. However, no evidence was found for a common ancestor for all of them. Another type of deviations are fusions of snoRNAs into a tandem transcript. These seem to result both from a fusion of separate molecules and of single molecules undergoing a tandem duplication concurrently with fusion.

The Interaction Conservation Index (ICI) measures snoRNA and target co-evolution.

The score combines thermodynamic stability of the RNA-RNA duplex with its evolutionary preservation. Evaluation of this measure on all known human interactions recovered $\sim$ 87% (scaRNAs-snRNAs), $\sim$ 83% (box H/ACA snoRNAs-rRNAs) and nearly 100% (box C/D snoRNAs-rRNAs) as evolutionary conserved (at least) in *Eutheria*. The correctness of the ICI measure is further supported by consistently high values for experimentally verified interactions. The lack of published negative experimental results makes the estimation of false positive rates not feasible, since functionality of high scoring interactions cannot be excluded.

Using the ICI, the evolutionary history of all snoRNA families was traced and stable partnerships between snoRNAs and their associated target sites were observed throughout vertebrates. This is at odds with the stringent conservation of modification sites and the slow evolution of rRNAs and snRNAs on the one hand while snoRNAs show high variation in their sequences on the other. Closer inspection of the snoRNA sequences revealed high selective pressure on the antisense elements preserving the complementarity to the target. Application of the ICI to all putative modifications and all snoRNAs could add new edges to the network of snoRNAs and the network of interactions. Based on sequence similarity and homologous function several snoRNA families reported in distantly related organisms were merged. Further for 10 of the 31 modifications with so-far unknown snoRNAs in rRNA and U2 snRNA an appropriate snoRNA guide was found. *Vice verca* for nine of 41 orphan snoRNAs a plausible function could be assigned. The ICI proved to be a very useful measure to match so far lonesome soulmates. Additionally, redundant guides and less common changeovers of guides could be resolved using the innovative evaluation method. In individual cases it was observed that redundant guides might be processed from host genes with anti-correlated expression profiles, suggesting how such snoRNAs could evolutionarily be maintained. This is in agreement with the hypothesis of (Hoeppner and Poole, 2012), who suggest constrained drift during the evolution of snoRNAs: an ongoing mobility of snoRNA genes with occupation of genomic locations that maintain the adequate expression pattern. The fact that at least 20% of the observed modifications in rRNAs show complementary to two or more snoRNAs (Dupuis-Sandoval et al., 2015) underlines the importance of their existence. For these cases the absence of individual snoRNA does not automatically alter the modification pattern, because the according modifications are redundantly guided (Kehr et al., 2014). In addition one can hypothesize that differentially expressed snoRNAs can contribute to the task

of specializing ribosomes. On protein level such adaptations to changing cellular conditions have already been observed (Xue and Barna, 2012). The capability of snoRNA directed modifications to alter structural conformations of the ribosome proposes them as further fine-tuners. New sequencing protocols are able to detect modifications in high-throughput (Birkedal et al., 2015; Krogh et al., 2016; Zaringhalam and Papavasiliou, 2016; Jorjani et al., 2016). They already observed two phases of modification in LSU, which might reflect early obligatory modifications and later adaptive ones (Birkedal et al., 2015). It will be interesting to study if and how modification patterns differ in different tissues, cell stages and environmental conditions.

A study using iCLIP experiments to identify the RNA interacting partners of the Cajal Body specific protein coilin discovered that it binds virtually all classes of small non-coding RNAs in the cell (Machyna et al., 2014). Especially, including hundreds of intron-encoded snoRNAs that traverse CBs. A scan for snoRNA features within unannotated significant reads identified several snoRNA candidates. Subsequent validation through conservation analysis using the `snoStrip` homology search pipeline identified 42 novel snoRNA sequences in human and 12 in mouse. Beside the new snoRNA nodes also two edges assigning novel snoRNAs as guides for previously unexplained modifications in snRNAs could be added to the interaction network. Among them SCARNA28 is the fitting guide for modification of position U47 in snRNA U2. A peculiarity of that sequence is a low complexity GT-insert like observed in a few other scaRNAs (Marz et al., 2011). This GT-repeats has been shown to function as Cajal-Body localization signal for box C/D snoRNAs (Kishore et al., 2013; Marnef et al., 2014), similar to the CAB-box in box H/ACA snoRNAs (Richard et al., 2003). For snoRNA that were crosslinked to coilin detailed analysis of the binding pattern between coilin and snoRNAs revealed that a small stem loops structure is found in the vicinity of the RNA-protein contact.

Broader availability of next generation sequencing data discovered additional and especially more variant snoRNA transcripts in the recent years even in human (Kishore et al., 2013; Deschamps-Francoeur et al., 2014; Machyna et al., 2014)). In combination with the fact that `snoRNA-LBME-db` (Lestrade and Weber, 2006) that used to be the human standard resource for snoRNAs is no longer maintained, the urgent need arose for a central, comprehensive, and up-to-date collection of human snoRNA. To provide a snoRNA set as complete as possible, data from available databases and literature were combined (Jorjani et al., 2016). Additionally, a screen for further *de novo* snoRNA

was performed. The study compiled a collection of 1118 human snoRNA genes including canonical box C/D snoRNAs, box H/ACA snoRNAs, and scaRNAs, but also sno-lncRNAs, AluACAs, and SNORD-like transcripts.

The workflow that was established to analyze snoRNA function (Section 3.3) again proved suitable to extend the snoRNA interaction network in human. It was possible to suggest functions for many of the novel snoRNAs as well as assign snoRNA guides to three previously reported 'orphan' modifications and another five modifications identified by high-throughput methods during this study. Interestingly, a significant enrichment in predicted targets in snRNA and especially snRNA from the minor spliceosome was observed for the newly detected non-canonical SNORD-like sequences. Most intriguingly, the snRNA U12 residue targeted by the 5' ASE of the H/ACA domain of the SCARNA21 is directly adjacent to a newly predicted target at the D box. The associated ASE is located in the elongated version of the gene. Additionally the 3' stem in the H/ACA part shows conserved complementarity to U6atac snRNA. Both snRNAs are part of the minor spliceosome. Thus, the results suggest an important role of SCARNA21 in the maturation of snRNAs of the minor spliceosome. Yet experimental validation of these predicted interactions is pending.

Furthermore, the study revealed that box C/D snoRNAs having a predicted or reported guide for both ASEs are only a minority constituting about 15% of all cataloged SNORD-like and box C/D sequences. Among the snoRNAs with a single target, the D' box ASE is surprisingly preferred over the ASE located at the generally more conserved D box. The underlying reason for this observation is not clear. It might be that the D box ASE is catalytically more active or that the region at the D box is more often involved into another cellular function performed by the snoRNP. In contrast to box C/D snoRNAs most box H/ACA snoRNAs function as double guides. This is in accordance with higher constraints on the sequences through the need of structure formation which results in higher overall evolutionary conservation of the sequences.

In total, it was possible to reduce the percentage of reported orphan snoRNAs in human from 40% to 20% compared to data currently listed in `snoRNA-LBME-db`. Among canonical, evolutionarily conserved snoRNAs (not including multi-copy families) currently still 76 have no assigned rRNA and snRNA target, of which six have a reported alternative function (Dupuis-Sandoval et al., 2015; Atzorn et al., 2004). How many of the orphan snoRNAs and the ASEs without detected function are to execute non-canonical functions remains difficult to answer and will in most cases require ex-

periments for each snoRNA in question. On modification site, still for 30 no fitting ASE within the snoRNAs could be predicted, which probably means that still a few snoRNA genes are hidden in the human genome. Another explanation could be that these modifications and/or snoRNAs are restricted to human and are consequently not detectable by our method since conservation is a requirement here. In the human snoRNAome, a considerable amount of human and primate specific snoRNAs was recognized, suggesting also a primate specific modification pattern. This is in agreement with evolutionary younger snoRNAs with brain specific expression found in eutherian species only. Leading to the assumption that snoRNAs with their high evolution rates and the according guided modifications in rRNAs and snRNAs contribute to the peculiarities of primates and humans.
Many newly predicted interactions target nucleotides that were not detected to be chemically modified. Further experiments that involve also new next-generation sequencing will have to resolve if these modifications can be found e.g under specific cellular conditions. It will be interesting to study the relation between differentially expressed snoRNAs, redundant guiding and adapting modification patterns.
For all human snoRNAs that were subject to the studies and that had no official gene symbol available, these were assigned by the HGNC. This is true for elsewhere reported sequences lacking a name so far and high confident novel ones.

Outlook

With the snoRNA sets in plants (Patra Bhattacharya et al., 2016), fungi (Bartschat et al., 2014) and vertebrates (Kehr et al., 2014) the distant homologies between snoRNAs of these kingdoms can be addressed. Sequence similarities, structure conservation and the information derived about snoRNA-target RNA co-evolution will help to solve this problem. Thus clarifying which snoRNA families are of ancient origin and which are recent innovations. It will be of interest if certain features can be identified that distinguish ancient from recent snoRNAs.
Further the synteny of snoRNAs should be focused in more detail. It is still an open question if snoRNAs are more often than not retained in the same introns of the same host genes. On the one hand stable associations between the snoRNA and the retaining introns exist, on the other hand intragenomic mobility has been observed (Weber, 2006; Shao et al., 2009; Hoeppner et al., 2009; Hoeppner and Poole, 2012).

However, snoRNAs are often encoded in clusters, i.e. in the introns of the same host genes. These clusters are also retained during evolution. Beside a large fraction of house keeping protein-coding genes (Dieci et al., 2009) several long non-coding RNAs encode snoRNAs in their introns. First underestimated as merely serving as vehicles for the snoRNAs they emerge to be involved in multiple malignancies, including many types of cancer (Zfas1, GAS5) (Krell et al., 2014; Williams and Farzaneh, 2012; Askarian-Amiri et al., 2011).

These lncRNA host genes are often subject to alternative splicing and differential expression. SnoRNAs are released from introns through exonucleolytic trimmings in which the snoRNA ends are protected through the bound core proteins (Reichow et al., 2007). Hence if an isoform harbors two snoRNAs in a contiguous excised sequence it seems natural that a transcript with snoRNAs at each end is retrieved. Exactly this pattern that generate so called sno-lncRNAs has been found at the SNURF-SNRPN locus and other loci (Yin et al., 2012; Zhang et al., 2014). This also correlated with the controversy about snoRNP monomers or dimers (Lapinaite et al., 2013). The first would originate from an intron that contains a single snoRNA, the latter from an isoform that encodes two snoRNAs in one intron and is retained as sno-lncRNA transcript with each snoRNA domain assembling a set of core proteins. The abundance and function of such transcripts could be estimated by studying isoforms of the host genes and could be further validated through transcriptome data.

The relationship between snoRNAs and their host genes could also be linked with the interesting findings concerning nonsense mediated decay (NMD) from Karijolich and Yu (2010) and Lykke-Andersen et al. (2014). On the one hand they identified the possibility to change a premature termination codon (PTM) inside a mRNA sequence into an active codon through box H/ACA snoRNA guided pseudouridylation. Thus, a degradation signal is turned into an codon, which determines the incorporation of a specific amino acid into the protein sequence. On the other hand the enrichment of PTMs within snoRNA host genes was observed. The authors suggest that this can uncouple the expression of host genes and that of snoRNAs. The combination of these observations can also lead to the suspicion that the hosted snoRNAs gain control over the translation of its host gene by turning on and off degradation and/or altering the amino acid chain of the encoded proteins. For an initial test of this hypothesis the mRNAs of snoRNA host genes could be scanned for complementarity to the ASE elements of the box H/ACA snoRNAs retained in their introns.

List of Figures

List of Tables

APPENDIX A

Appendix

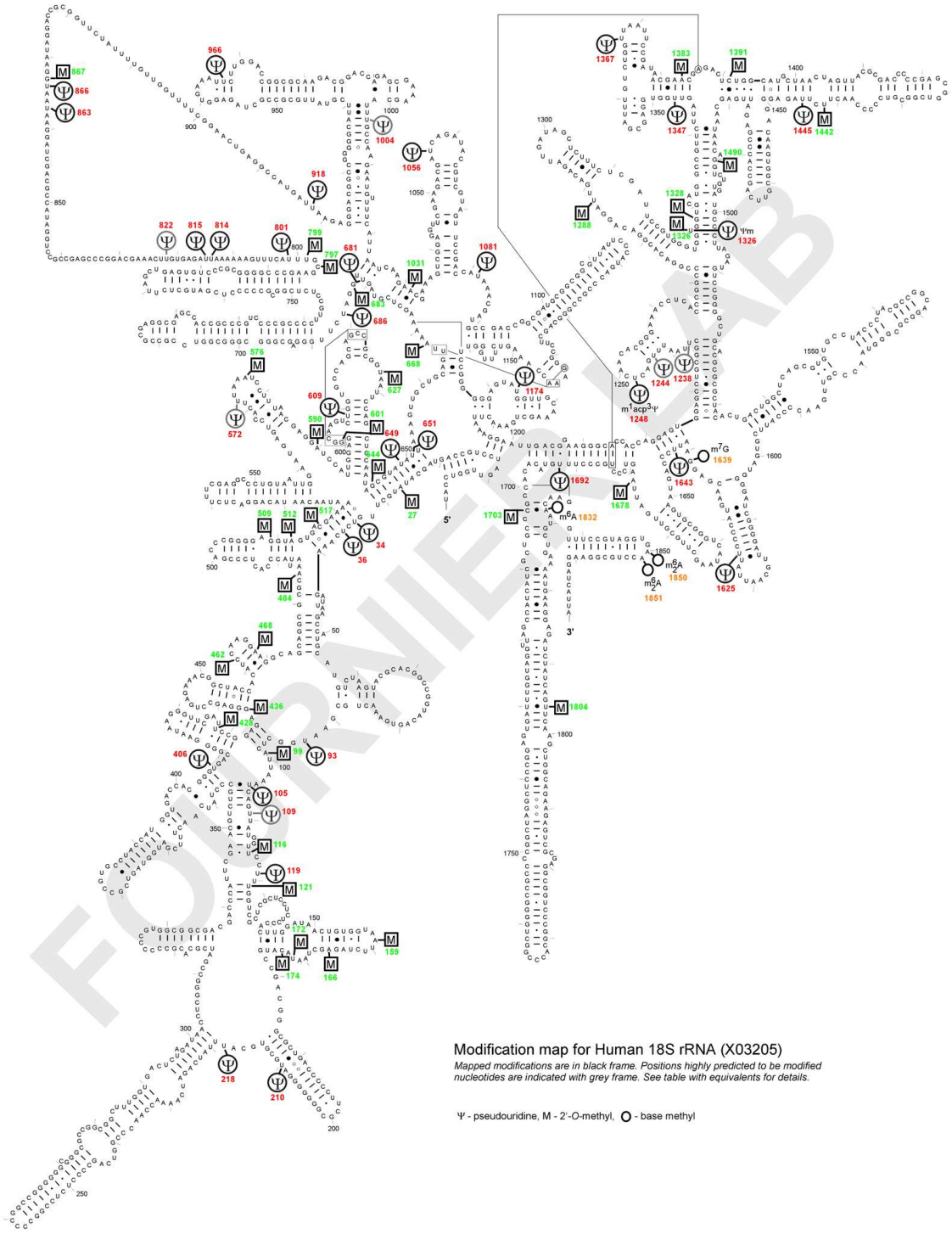

Figure A.1: Modification Map of Human SSU from `http://people.biochem.umass.edu/sfournier/fournierlab/3dmodmap/hum2dssu.htm`

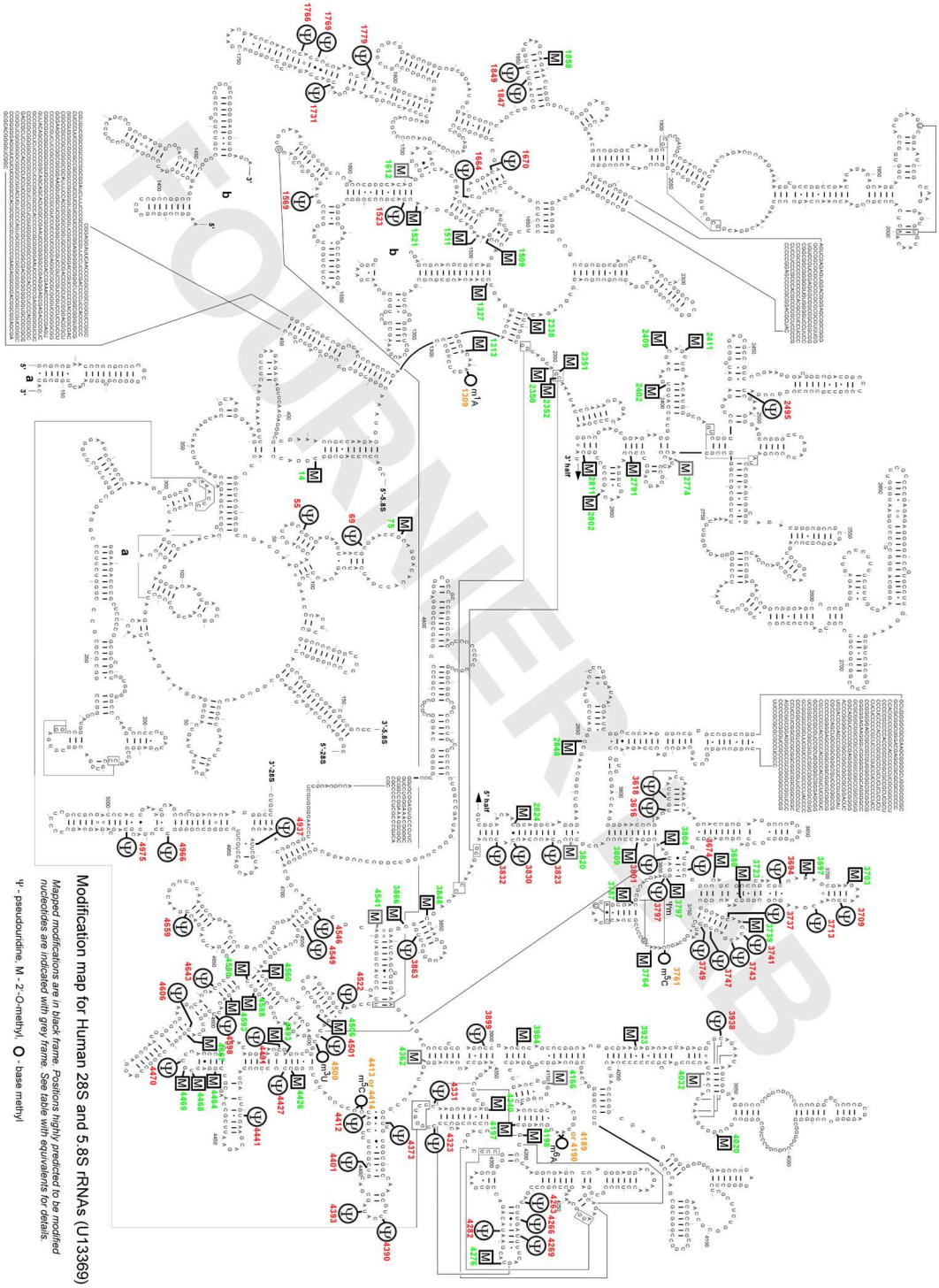

Figure A.2: Modification Map of Human LSU from `http://people.biochem.umass.edu/sfournier/fournierlab/3dmodmap/hum2dlsu.htm`

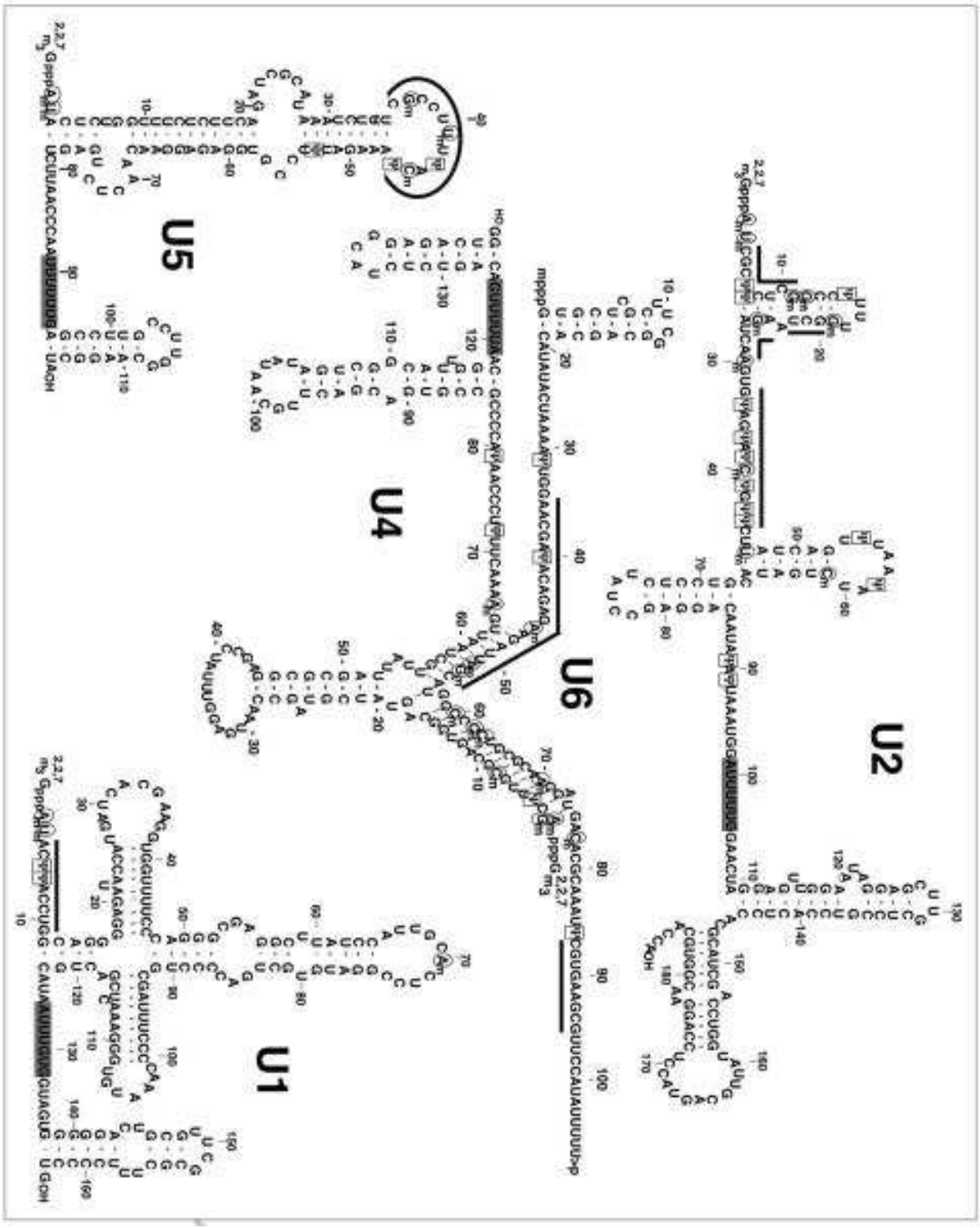

Figure A.3: Modification Map in Human snRNAs from (Karijolich and Yu, 2010)

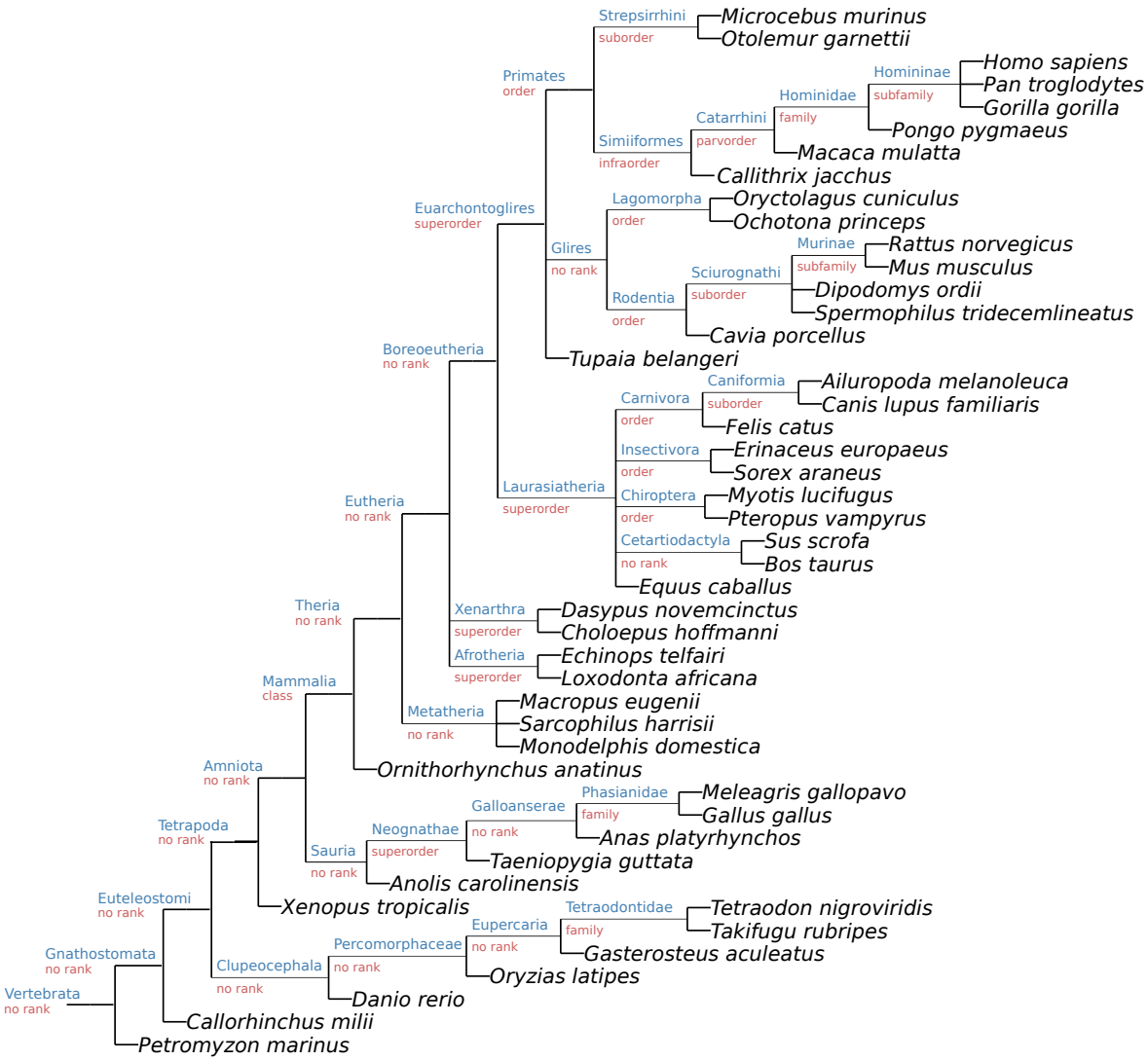

Figure A.4: Phylogenetic tree of 47 investigated vertebrate species. The tree is generated with help of the ETE Toolkit (Huerta-Cepas et al., 2016)

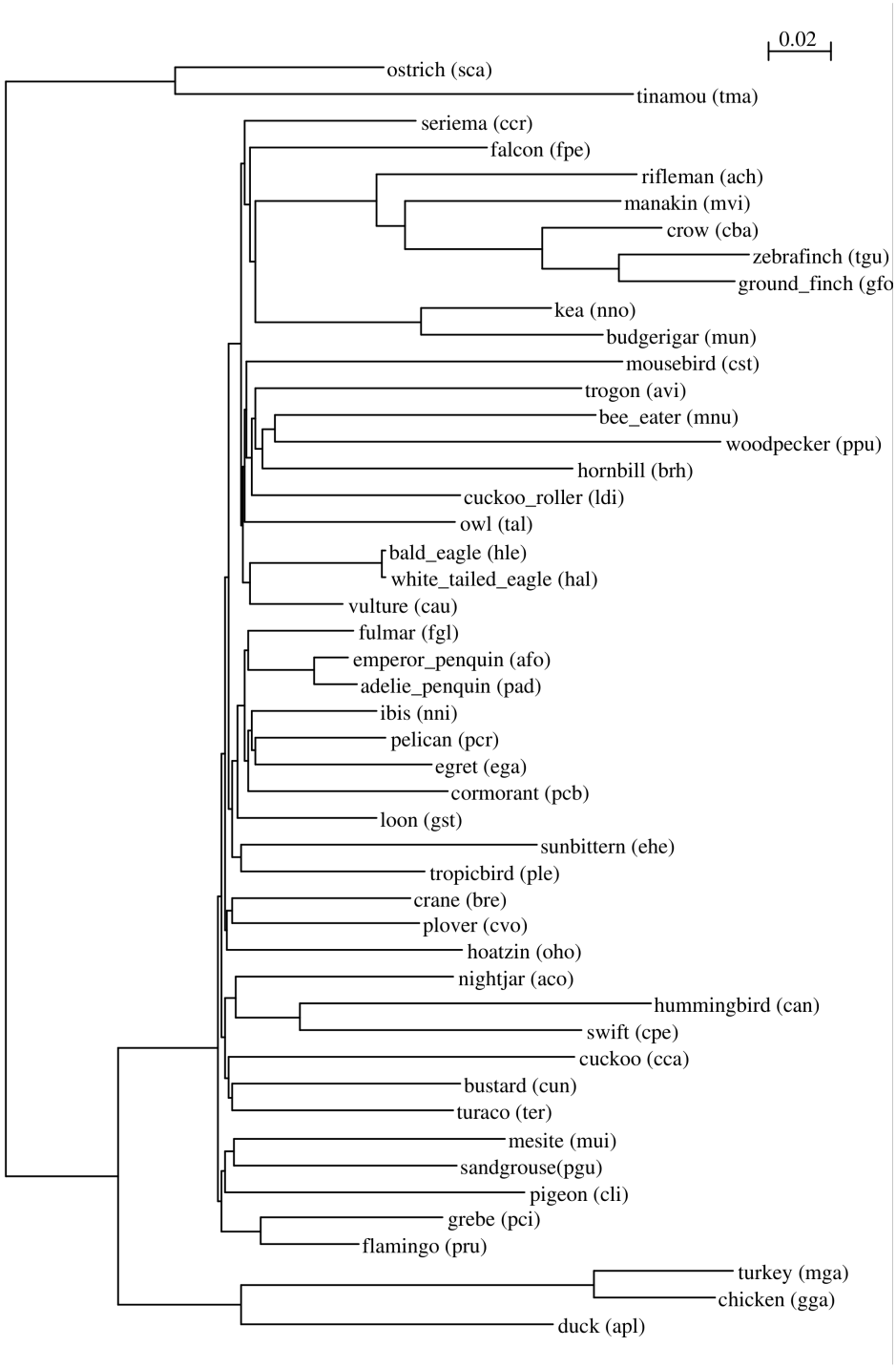

Figure A.5: Phylogenetic tree of 48 investigated avian species.

Table A.1: Considered 47 vertebrate species. Providing common name, species name, abbreviation, database (ENS, UCSC) and assembly/database identifier

Name	Species	Abbreviation	source	assembly
human	Homo sapiens	hsa	UCSC	hg19
chimp	Pan troglodytes	ptr	ENS	CHIMP2.1
orangutan	Pongo pygmaeus	ppy	ENS	PPYG2
gorilla	Gorilla gorilla	ggo	ENS	gorGor1
rhesus	Macaca mulatta	mml	UCSC	rheMac2
bushbaby	Otolemur garnettii	oga	ENS	BUSHBABY1
lemur	Microcebus murinus	mmr	ENS	micMur1
marmoset	Callithrix jacchus	cjc	ENS	calJac3
rat	Rattus norvegicus	rno	ENS	RGSC3.4
mouse	Mus musculus	mmu	UCSC	mm9
kangaroo rat	Dipodomys ordii	dor	ENS	dipOrd1
squirrel	Spermophilus tridecemli.	str	ENS	speTri1
guinea pig	Cavia porcellus	cpo	ENS	cavPor3
rabbit	Oryctolagus cuniculus	ocu	ENS	RABBIT
treeshrew	Tupaia belangeri	tbe	ENS	TREESHREW
cat	Felis catus	fca	ENS	CAT
dog	Canis familiaris	cfa	ENS	canFam2
panda	Ailuropoda melanoleuca	aml	UCSC	ailMel1
cow	Bos taurus	bta	ENS	Btau_4.0
pig	Sus scrofa	ssc	ENS	Sscrofa10.2
horse	Equus caballus	eca	ENS	EquCab2
mirobat	Myotis lucifugus	mlu	ENS	MICROBAT1
flying fox	Pteropus vampyrus	pva	ENS	pteVam1
hedgehog	Erinaceus europaeus	eeu	ENS	HEDGEHOG
shrew	Sorex araneus	sar	ENS	COMMON_SHREW1
elephant	Loxodonta africana	laf	ENS	loxAfr2
tenrec	Echinops telfairi	ete	ENS	TENREC
armadillo	Dasypus novemcinctus	dno	ENS	dasNov2
sloth	Choloepus hoffmanni	cho	ENS	choHof1
opossum	Monodelphis domestica	mdo	UCSC	monDom4
tasmanian devil	Sarcophilus harrisii	sha	ENS	DEVIL7.0
wallaby	Macropus eugenii	meu	ENS	Meug_1.0
platypus	Ornithorhynchus anatinus	oan	ENS	OANA5
lizard	Anolis carolinensis	acr	ENS	AnoCar2.0
zebra finch	Taeniopygia guttata	tgu	UCSC	taeGut1
turkey	Meleagris gallopavo	mga	OTHER	beta
duck	Anas platyrhynchos	apl	OTHER	beta
chicken	Gallus gallus	gga	ENS	galGal4
crawled frog	Xenopus tropicalis	xtr	ENS	xenTro3
coelacanth	Latimeria chalumnae	lch	OTHER	
green pufferfish	Tetraodon nigroviridis	tni	UCSC	tetNig2
japanese pufferfish	Takifugu rubripes	tru	ENS	fr3
stickleback	Gasterosteus aculeatus	gac	UCSC	gasAcu1
medaka	Oryzias latipes	ola	ENS	MEDAKA1
zebra fish	Danio rerio	dre	ENS	Zv9
elephant shark	Callorhinchus milii	cmi	OTHER	1.4x
sea lamprey	Petromyzon marinus	pma	UCSC	petMar2

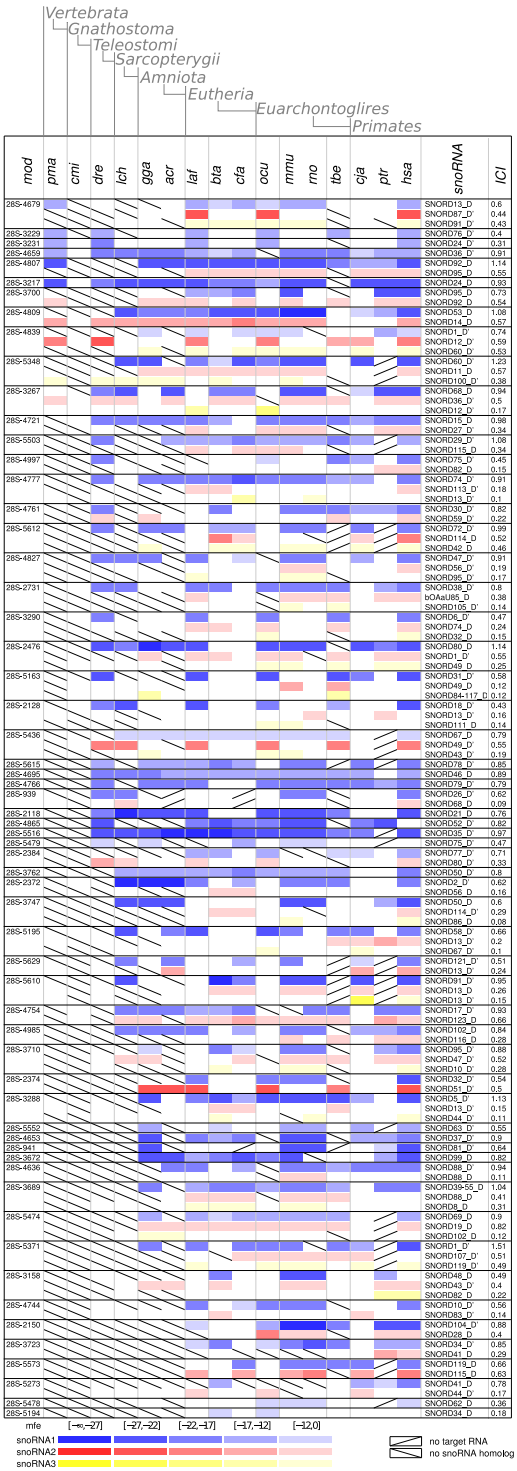

Figure A.6: Interaction conservation of box C/D snoRNA and targets in 28S rRNA. *Description of the figure analogously to Figure 6.1*

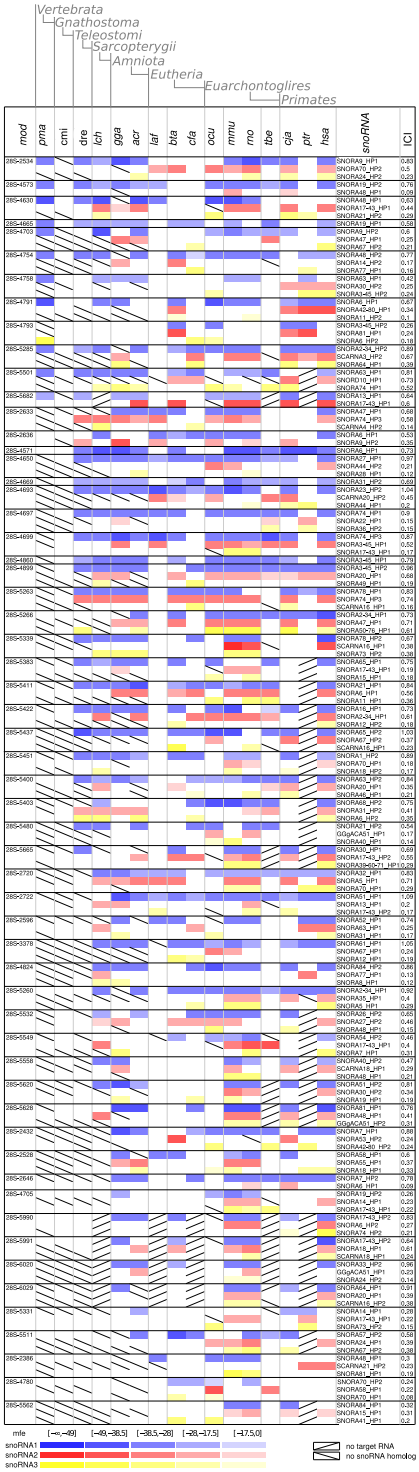

Figure A.7: Interaction conservation of box H/ACA snoRNA and targets in 28S rRNA. *Description of the figure analogously to Figure 6.1*

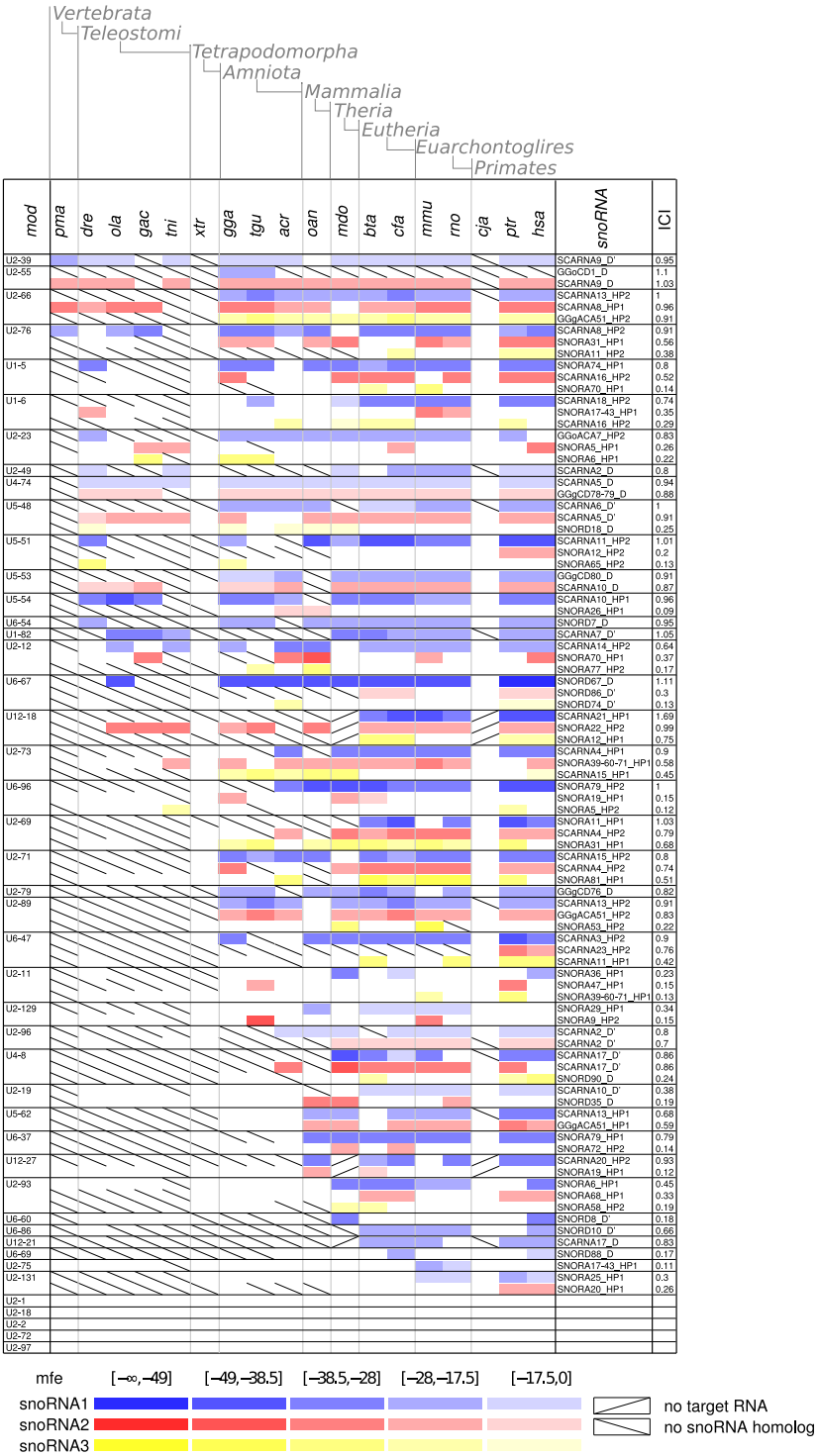

Figure A.8: Interaction conservation of box C/D and box H/ACA snoRNA and targets in snRNAs. Description of the figure analogously to Figure 6.1

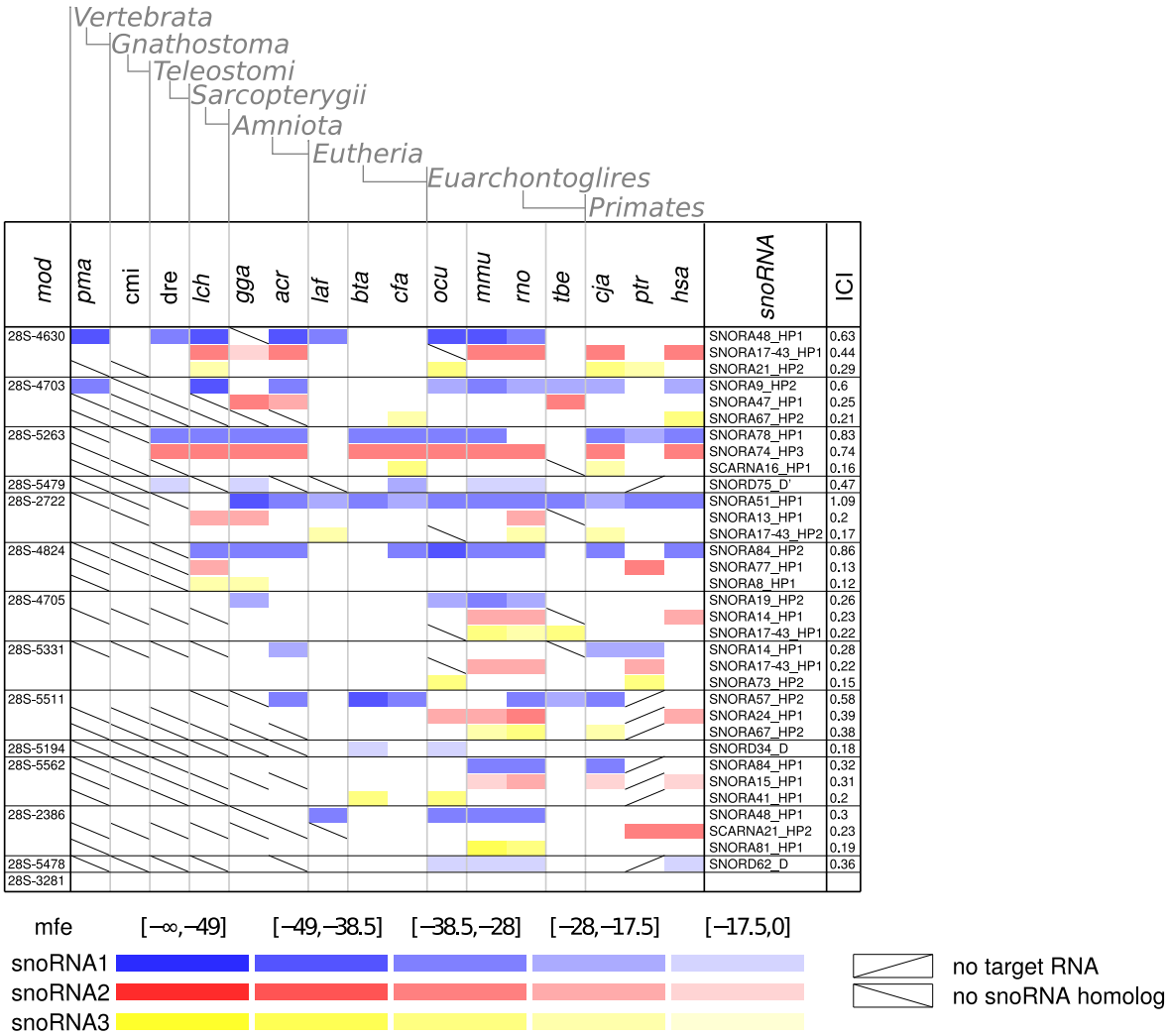

Figure A.9: SnoRNA guides to nown modifications in 28S rRNA

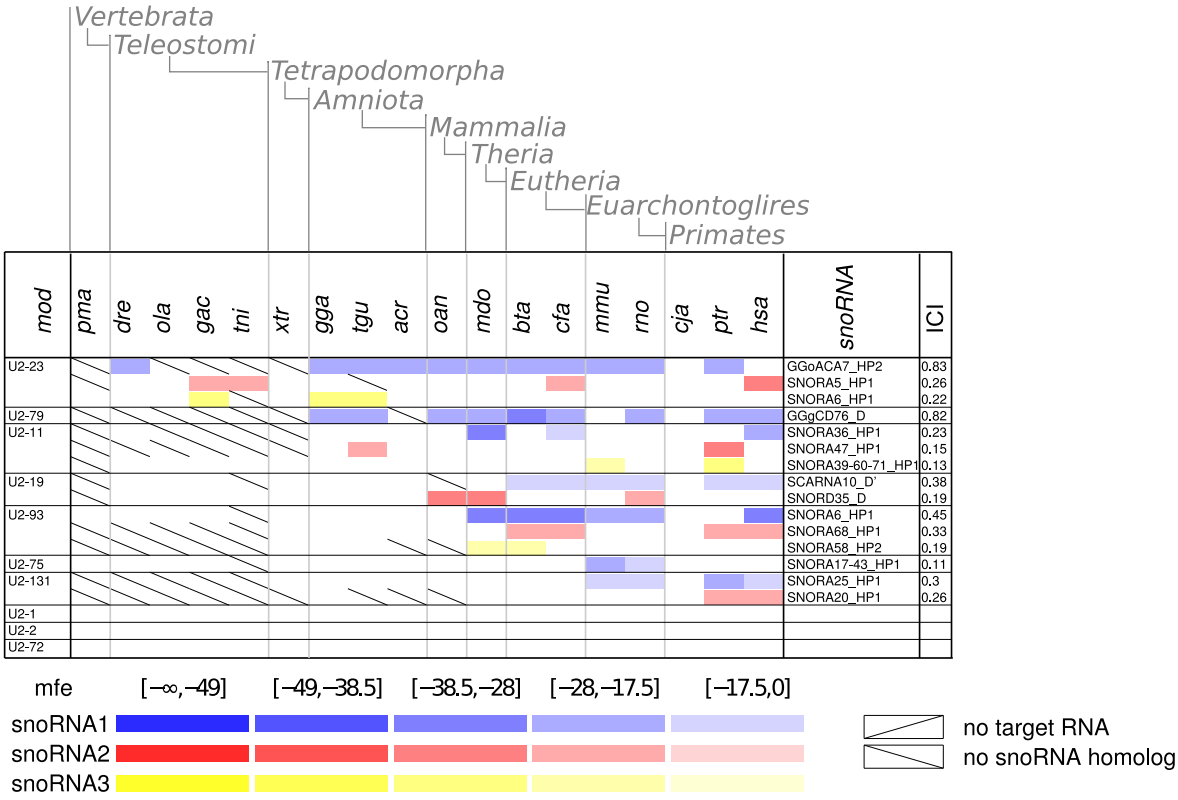

Figure A.10: SnoRNA guides to Known modifications in U2 snRNA

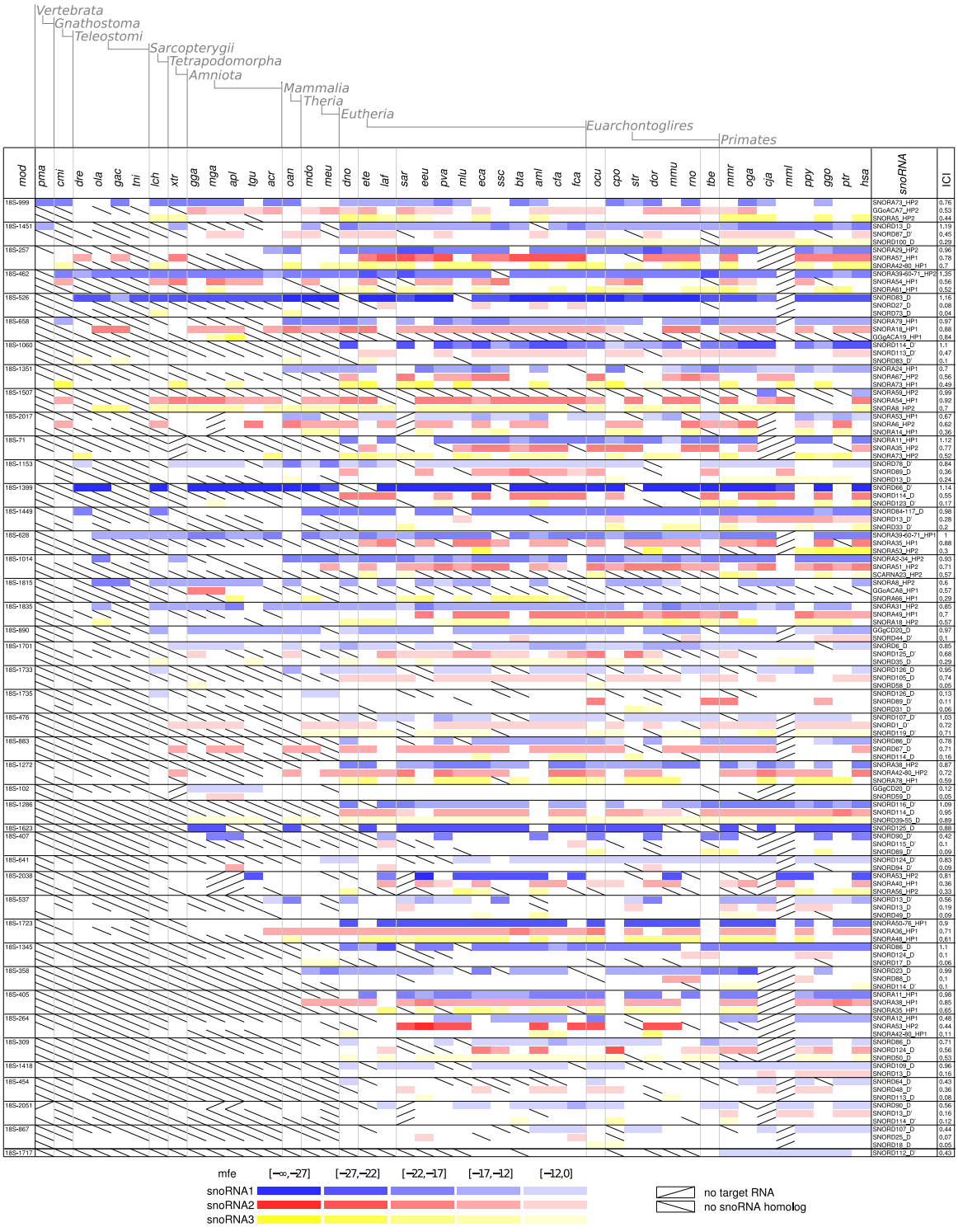

Figure A.11: Predicted Modification sites of orphan snoRNA in 18S rRNA

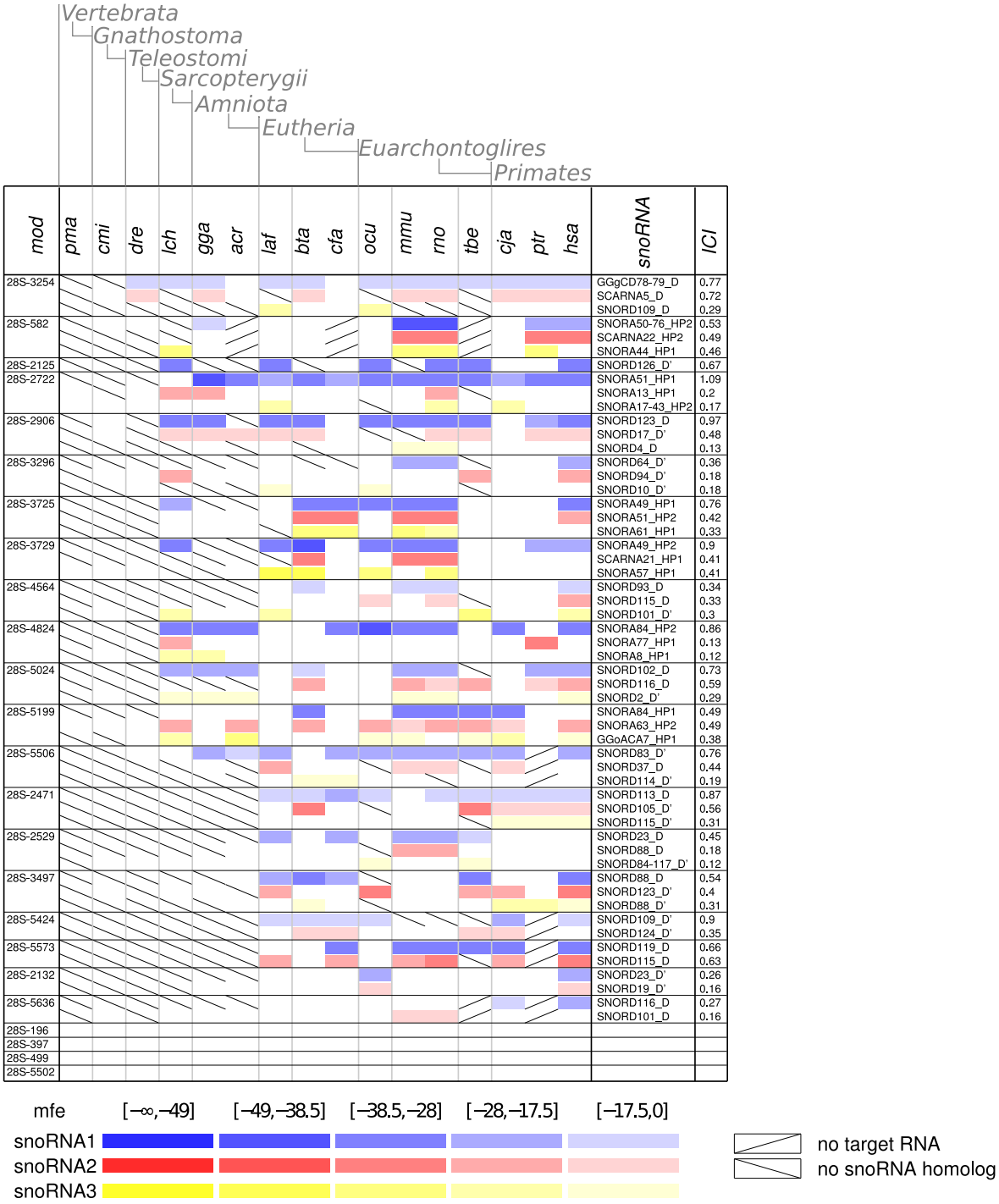

Figure A.12: Predicted Modification sites of orphan snoRNA in 28S rRNA

Table A.2: Novel CLIP SnoRNAs in Human. E:Eutheria, M:Marsupilia, A:Aves, P:Primates, S:Sauropsids, T:Teleost, O:Platypus, :modified, ̂:pseudogene, !:unknown guide

ID	Coordinates	(hg19)	Boxmotifs	CAB	Hostgene	Conservation	Target
SNORA85	chr10:3176352-3176517	(+)	AAATAA_75_ACA_161	TGAG_44	PFKP	P	U5-78
SNORA86	chr10:33190262-33190400	(-)	ACAAAA_90_ACA_134		ITGB1 (5'UTR)	E;M;A;O	
SNORA87	chr10:114805114-114805335	(+)	AGAGCA_118_ACA_217	GGAG_184	TCF7L2	P	18S-28&18S-109
SNORA88	chr11:32165288-32165458	(-)	ATAAGG_71_ACA_166		THEM7P	P	
SNORA89	chr14:34178143-34178319	(+)	AAAATA_94_ACA_172		NPAS3	E;M	
SNORA90	chr17:35601717-35601891	(+)	ACAGAA_84_ACA_170		ACACA	P	
SNORA91	chr21:43368214-43368391	(-)	AAAACA_127_ACA_173		C2CD2	E	
SNORA92	chr22:38620486-38620727	(-)	AGAGCA_92_ACA_237		TMEM184B	P	U1-78&18S-1509
SNORA93	chr3:13659881-13660048	(+)	AAAATA_73_ACA_163		FBLN2	P	18S-1155
SNORA94	chr3:48642353-48642583	(-)	ACAGGA_103_ACA_226		UQCRC1	P	U2-22
SNORA95	chr3:85111304-85111415	(-)	AAAAAA_60_ACA_107		CADM2	E	
SNORA96	chr4:83816955-83817193	(-)	AAATAA_107_ACA_234		SEC31A	P	
SNORA97	chr4:83819244-83819444	(-)	ACATAA_104_ACA_196	TGAG_150	SEC31A	E;M	
SNORA98	chr6:144726791-144726903	(+)	AAAACA_48_ACA_108		UTRN	E	
SNORA99	chr8:11562871-11563067	(+)	AAAATA_108_ACA_192		GATA4	P	18S-742
SNORA100	chr1:245017389-245017519	(-)	ACAACA_59_ACA_126		HNRNPU (5'UTR)	E	28S-5011
SNORA103	chr1:173835343-173835428	(-)	AAAATA_46_ACA_81		GAS5	P	
SNORA104	chr19:14733025-14733163	(+)	ACAAAA_90_ACA_134		EMR3	E;M;A;O	
SNORA105A	chr5:21884563-21884678	(+)	AGATAA_54_ACA_111		CDH12	E;M;S;O	
SNORA105B	chr2:198351512-198351627	(-)	AGATAA_54_ACA_111	HSPD1 (5'UTR)	E;M;S;O		
SNORA105C	chr12:56906649-56906764	(+)	AGATAA_54_AAA_111		intergenic	E;M;S;O	
SNORA107	chr13:85357789-85357913	(+)	ATAGGA_69_ACA_120		intergenic	E	
SNORA108	chr18:56267852-56267984	(+)	ATAGCA_55_ACA_128		ALPK2	P	
SNORA109	chrX:55209913-55210040	(-)	ACAATA_52_ACA_123		integenic	P	
SNORA110	chr1:44718016-44718231	(-)	ACAGCA_129_AGA_211	TGGG_180	ERI3	E;A	18S-1172
SNORA111	chr18:35046302-35046467	(-)	ATACCA_111_ACA_161		CELF4	E	18S-918
SNORD128	chr1:8554860-8554973	(-)	ATGAGAT_74_CTGT_39-GTGATGT_7_CTGA_106		RERE	P	
SNORD129	chr10:7228642-7228746	(-)	CTGATGT_5_CTGA_98		SFMBT2	E;M;O	
SNORD130	chr10:28362166-28362307	(-)	GTGATGC_9_CTGA_131		MPP7	P	18S-A221
SNORD131	chr11:1970561-1970706	(+)	TTAGTGA_97_CTGC_51-GTGAAGA_17_CTGA_131		MRPL23	P	28S-G2342
SNORD132	chr2:111415727-111415826	(-)	GTGATAG_8_CTGA_89		BUB1	P	U2-U44()
SNORD133	chr12:50850354-50850569	(+)	GTGATGA_8_CTGA_207		LARP4	E;M;A	18S-C676
SNORD134	chr17:80047836-80048016	(-)	GTGATGA_8_CTGA_173		FASN	P	18S-C251&18S-G1151
SNORD135	chr19:12814411-12814485	(-)	GTGATGG_8_CTGA_66		TNPO2	E	28S-U4276
SNORD136	chr3:52722903-52723049	(+)	ATAATGC_85_CTGA_60-ATGATGA_9_CTGA_137		GNL3	P	
SNORD137	chr9:12972554-12972632	(-)	ATGATGT_9_CTGA_69		intergenic	E	U2-A30
SNORD140	chr22:41469612-41469725	(-)	ATGATCA_3_CTGA_108		intergenic	P	
SNORD141A	chr9:135895817-135895921	(+)	AGGATGT_9_CTGA_95		EEF1A1P5̂	E	
SNORD141B	chr5:14652387-14652491	(-)	AGGATGT_9_CTGA_95		EEF1A1P13̂	E	
SCARNA26A	chr1:155648899-155649046	(-)	AAAACA_64_ACA_142	TGAG_100	YY1AP1	E;M;A	
SCARNA26B	chr1:155753475-155753623	(-)	AAAACA_64_ACA_145	TGAG_100	YY1AP1	E;M;A	U4-79!
SCARNA28	chr7:98479320-98479513	(+)	GTGATGA_8_CTGA_182		TRRAP	E;M;A;O	U2-U47!

Table A.3: Novel CLIP SnoRNAs in Mouse. No official HGNC gene symbols were requested for the novel mouse snoRNAs. Thus internal ID are listed in the table.

ID	Coordinates	(mm9)	Boxmotifs	CAB	Hostgene	Conservation	Target
SNORA97	chr5:100855847-100856043	(-)	ACATAC_100_ACA_192	TGAG_134	5430416N02Rik	E;M	
HACA_112-1	chr4:117317740-117317956	(+)	ACAGCA_129_ACA_212	TGGG_184	Eri3	E;A	18S-1173
HACA_119-1	chr9:26949884-26950025	(-)	AAATAA_75_ACA_137		Jam3	E	
HACA_39-1	chr18:25820640-25820803	(-)	ATAGCA_109_ACA_159		Celf4	E	
HACA_137-1	chr15:36345512-36345620	(-)	AAAAGA_45_ACA_104		intergenic	E (loss in P)	
SNORA100	chr1:180259122-180259247	(-)	ACAACA_55_ACA_121		Hnrnpu (5'UTR)	E	28S-4693
SNORD135	chr8:87577488-87577560	(+)	GTGATGG_8_CTGA_64		Tnpo2	E	28S-U3958
CD_109-1	chrX:5988955-5989084	(+)	CTGAGGA_8_CTGA_119		Shroom4	E	
CD_119-1	chr1:24619973-24620045	(-)	TTGATGA_6_CTGA_65		intergenic	E	
SCARNA26A	chr3:88603786-88603932	(+)	AAAGTA_64_ACA_142	TGAG_99	YY1AP1	E;M;A	U4-79
SCARNA26B	chr1:148785978-148786124	(-)	AAAGCA_64_ACA_142	TGAG_99	intergenic	E;M;A	U4-79
SCARNA28	chr5:145532200-145532397	(+)	ATGATGA_8_CTGA_189		Trrap	E;M;A;O	U2-U47

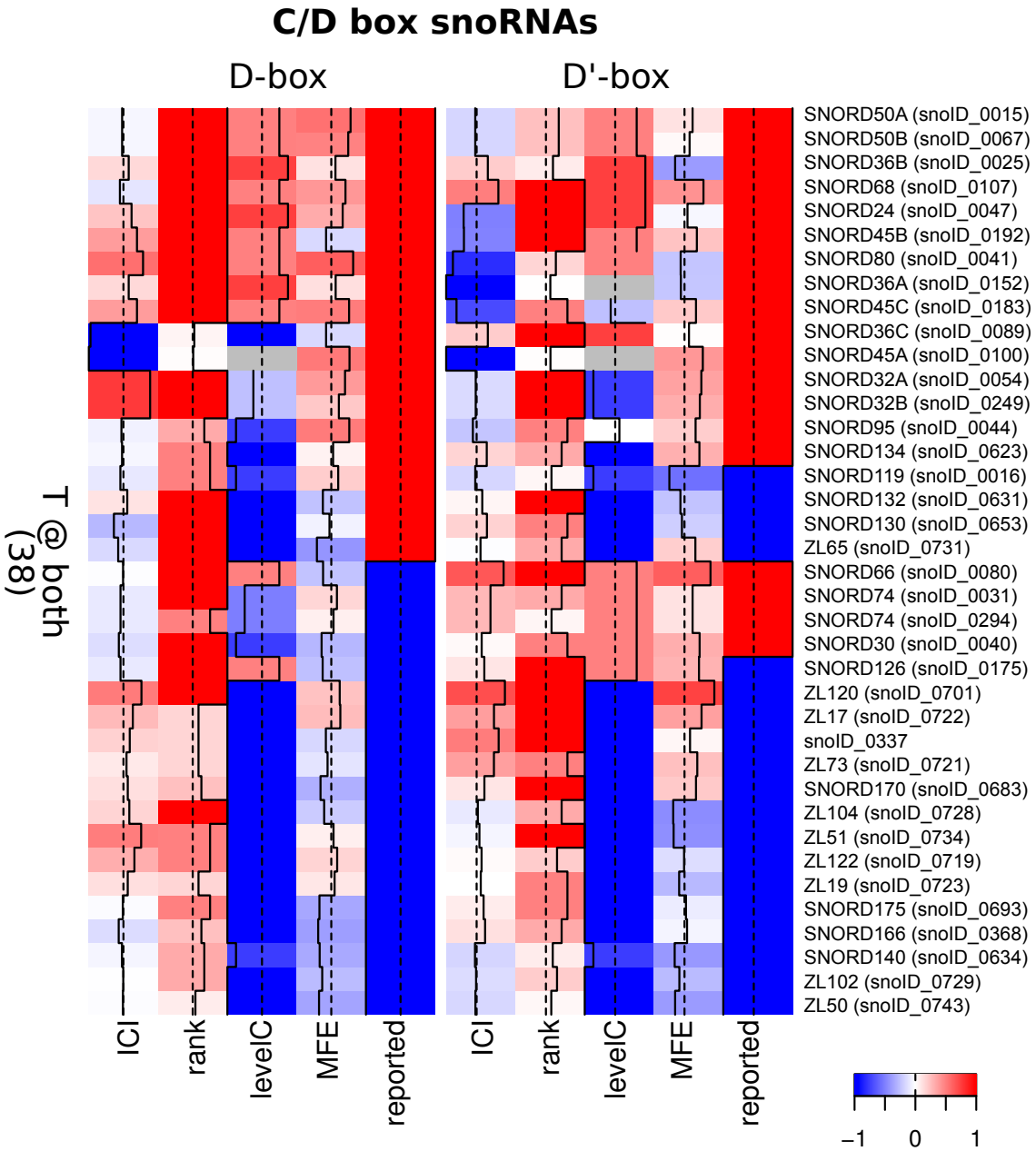

Figure A.13: High-resolution heatmaps of target binding characteristics in double guiding box C/D snoRNAs

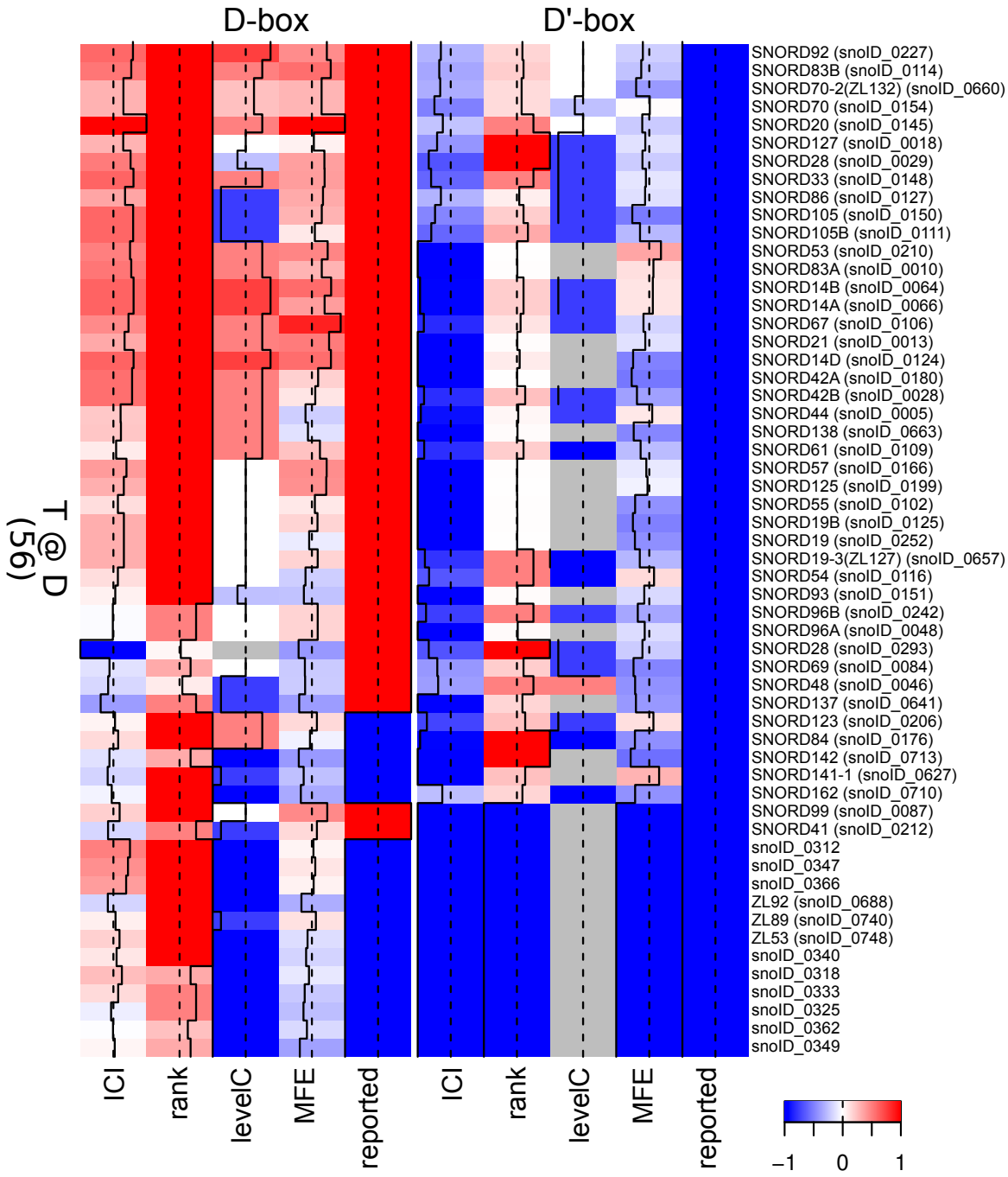

Figure A.14: High-resolution heatmaps of target binding characteristics in single guiding box C/D snoRNAs @D box

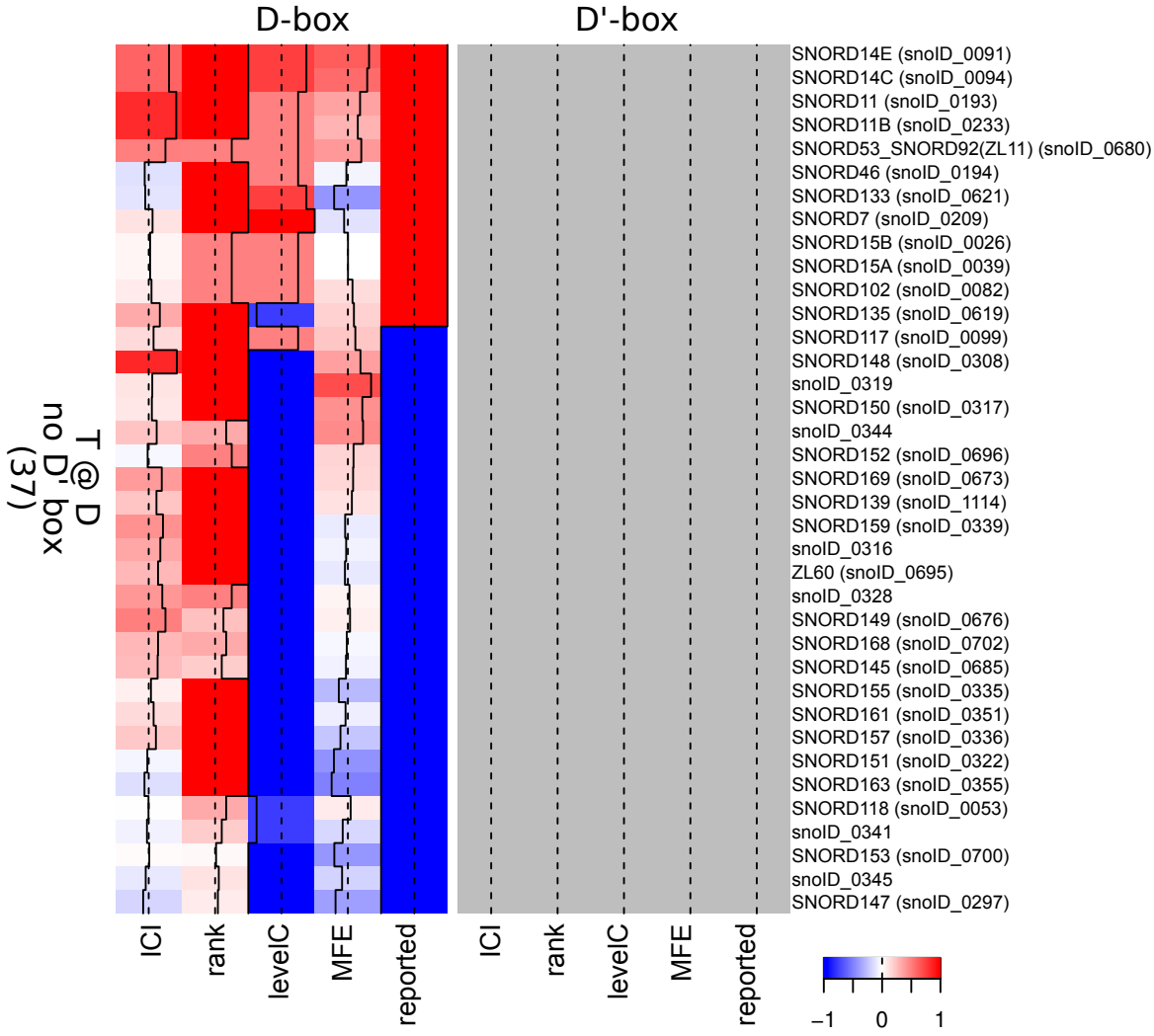

Figure A.15: High-resolution heatmaps of target binding characteristics in single guiding box C/D snoRNAs @D box without D' box

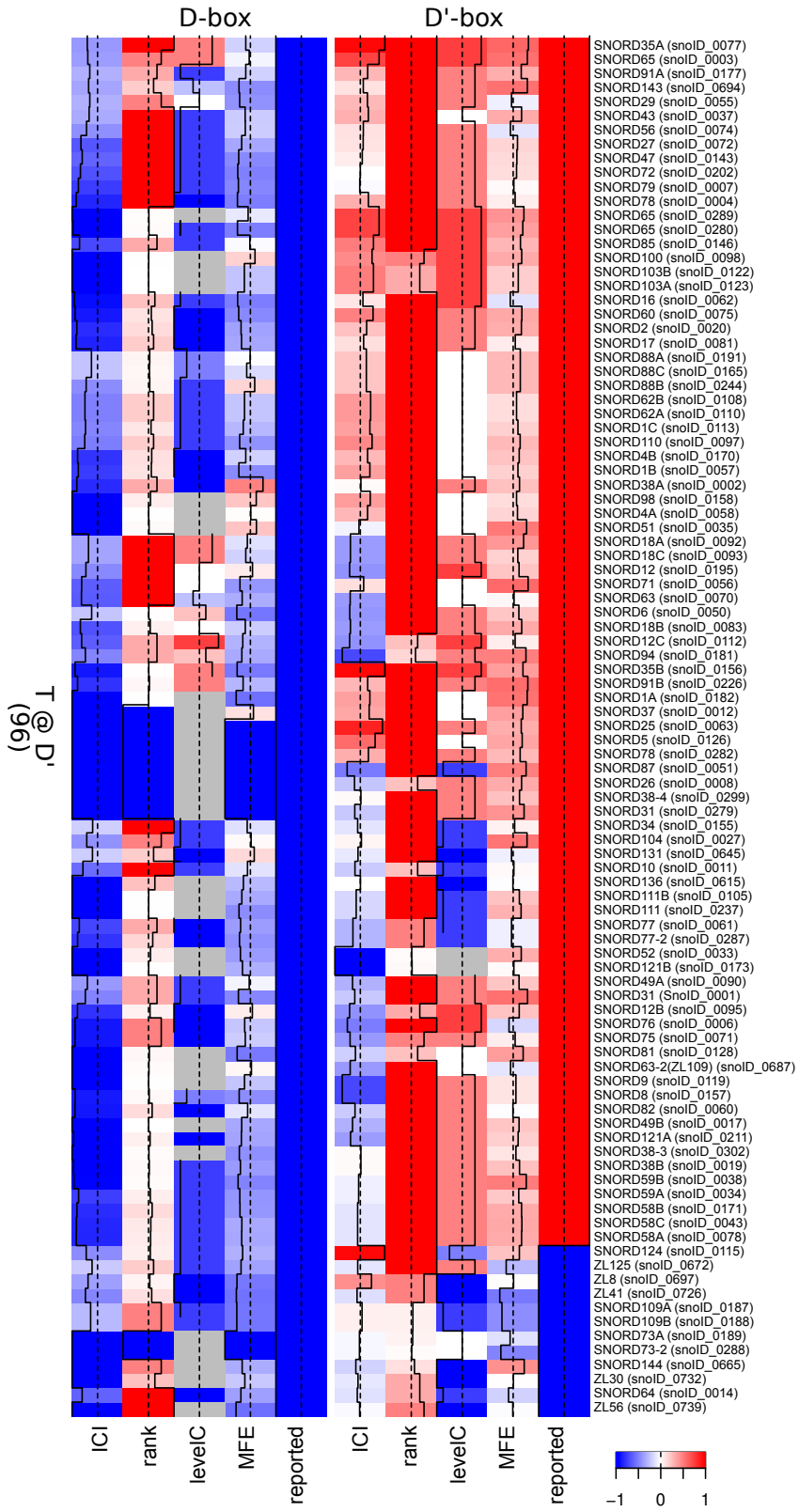

Figure A.16: High-resolution heatmaps of target binding characteristics in single guiding box C/D snoRNAs @D' box

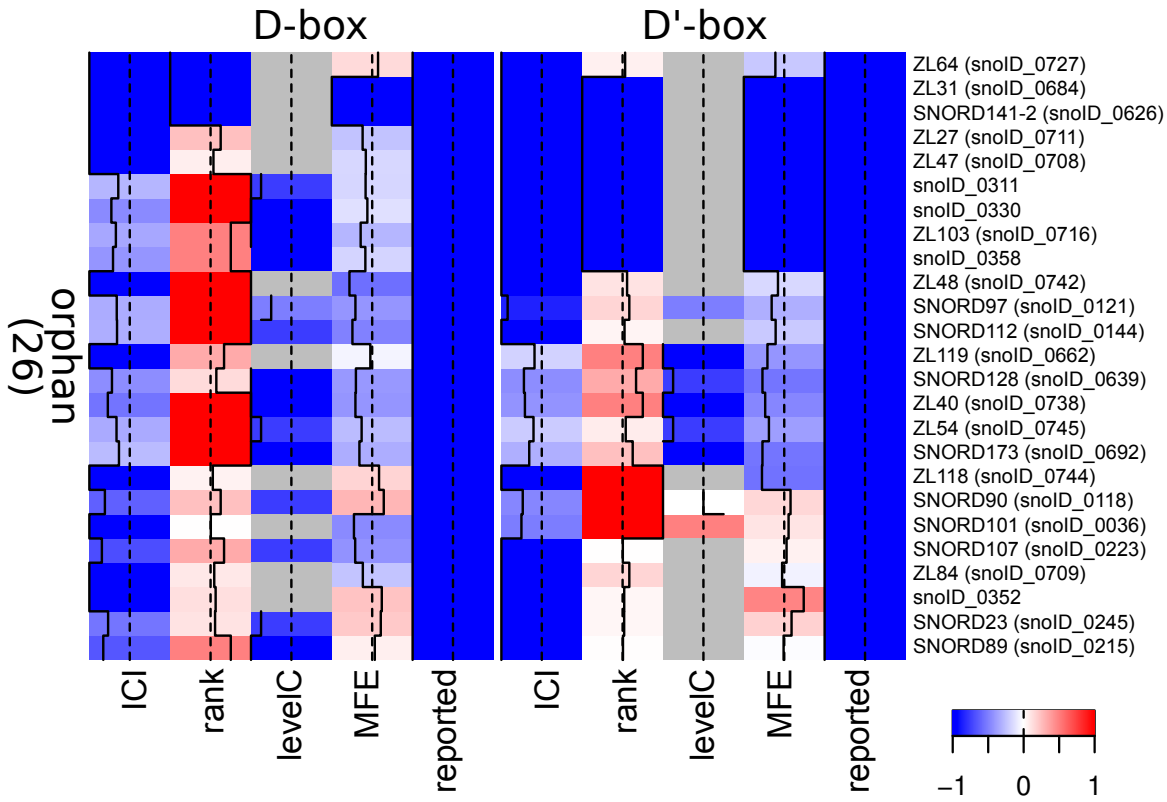

Figure A.17: High-resolution heatmaps of target binding characteristics in orphan box C/D snoRNAs with both boxes identified

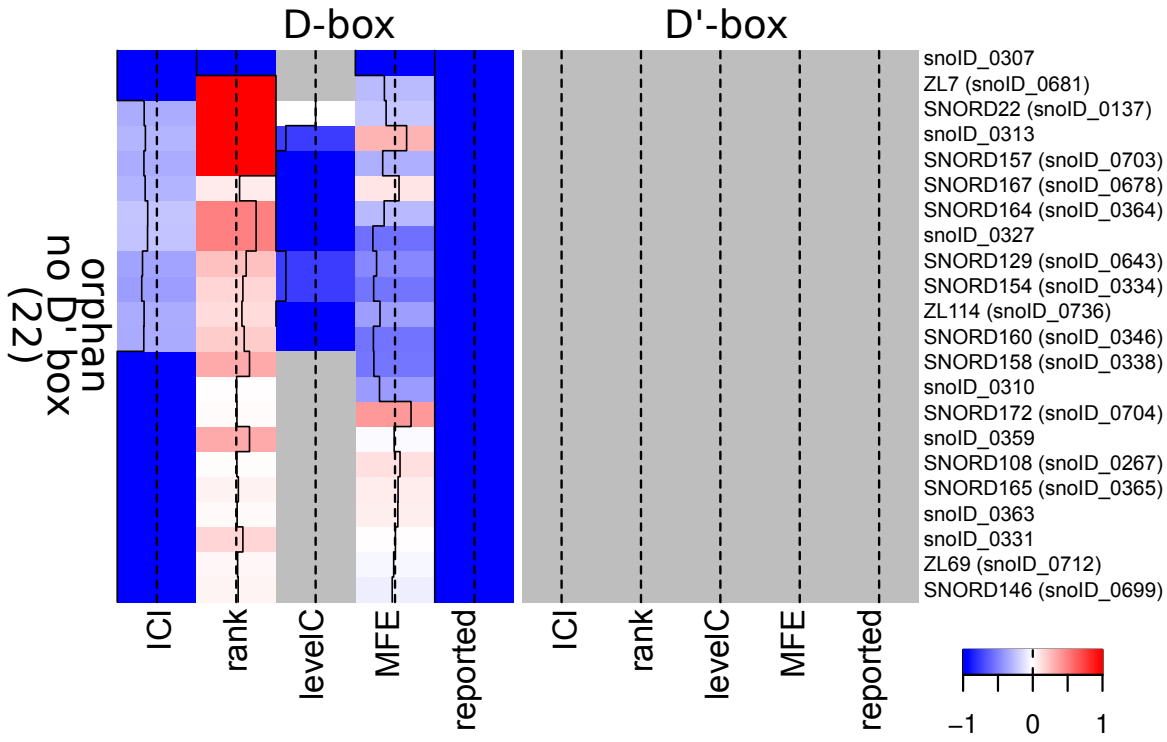

Figure A.18: High-resolution heatmaps of target binding characteristics in orphan box C/D snoRNAs without D' box

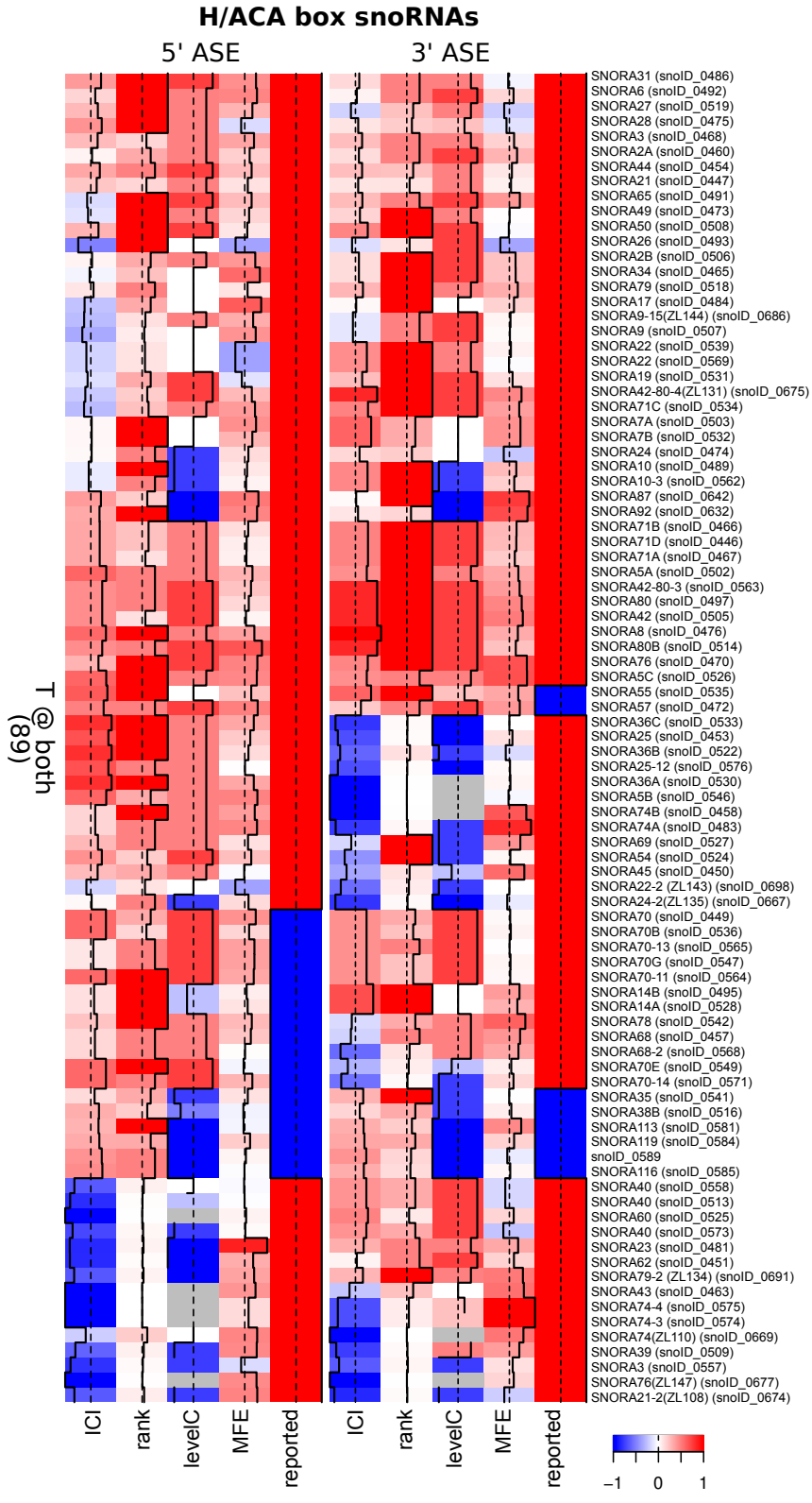

Figure A.19: High-resolution heatmaps of target binding characteristics of doube guiding box H/ACA snoRNAs

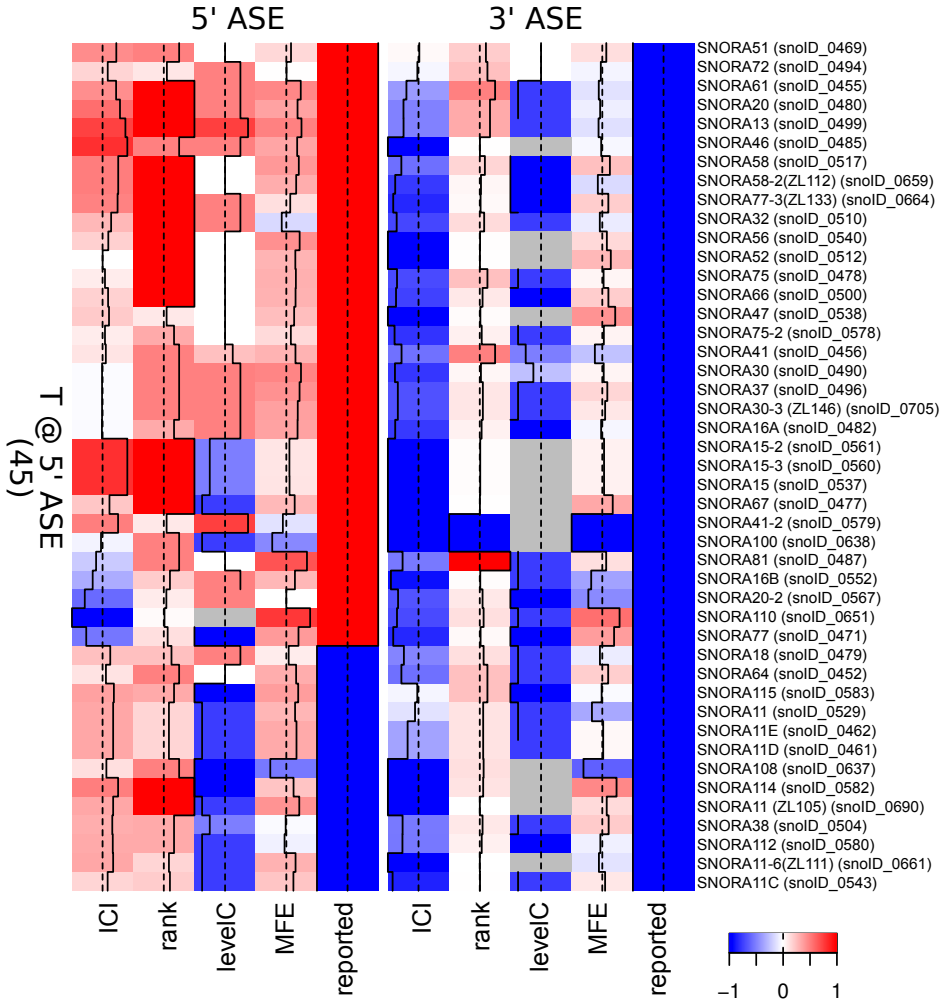

Figure A.20: High-resolution heatmaps of target binding characteristics of single guiding box H/ACA snoRNAs @5' hairpin

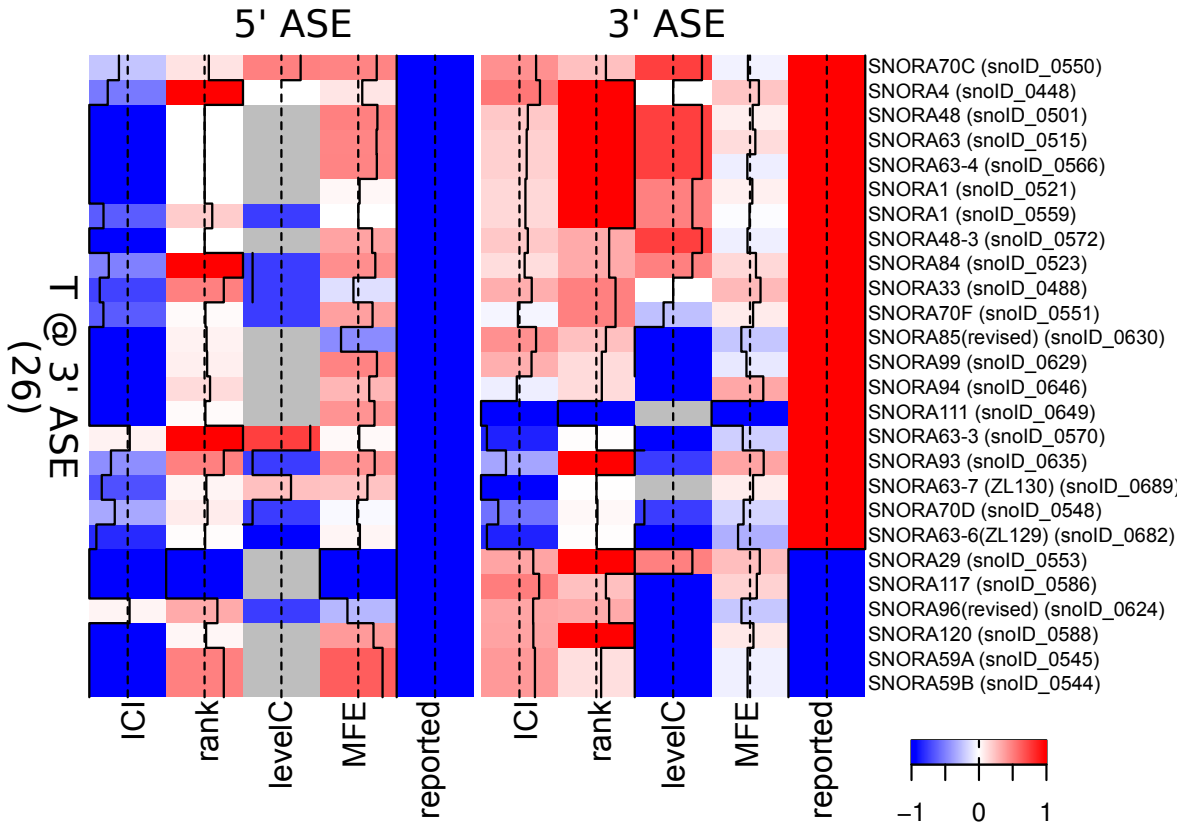

Figure A.21: High-resolution heatmaps of target binding characteristics of single guiding box H/ACA snoRNAs @ 3' haipin

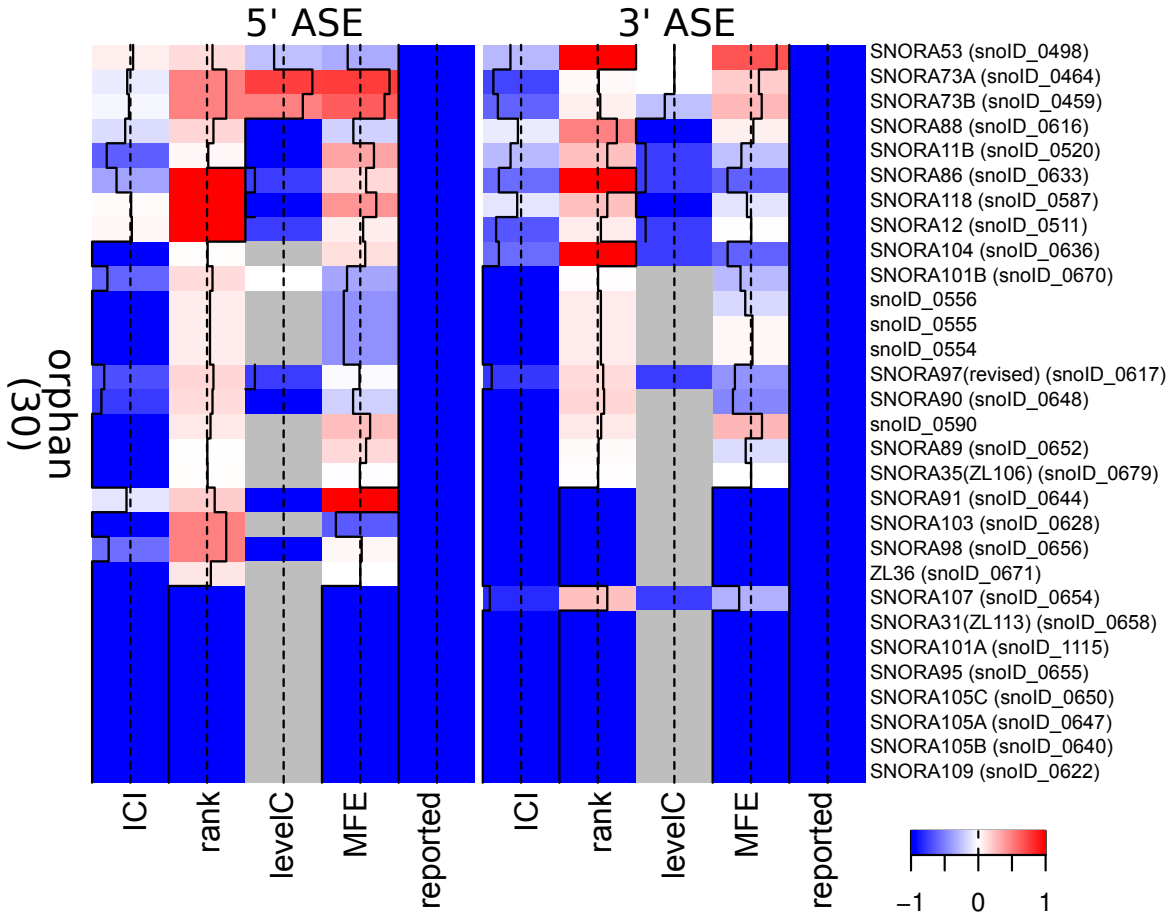

Figure A.22: High-resolution heatmaps of target binding characteristics of orphan box H/ACA snoRNAs

Bibliography

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, Sep 1997.

Anthon, C., Tafer, H., Havgaard, J. H., Thomsen, B., Hedegaard, J., Seemann, S. E., Pundhir, S., Kehr, S., Bartschat, S., Nielsen, M., Nielsen, R. O., Fredholm, M., Stadler, P. F., and Gorodkin, J. Structured RNAs and synteny regions in the pig genome. *BMC Genomics*, 15:459, 2014. doi: 10.1186/1471-2164-15-459.

Askarian-Amiri, M. E., Crawford, J., French, J. D., Smart, C. E., Smith, M. A., Clark, M. B., Ru, K., Mercer, T. R., Thompson, E. R., Lakhani, S. R., Vargas, A. C., Campbell, I. G., Brown, M. A., Dinger, M. E., and Mattick, J. S. SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA*, 17(5):878–91, May 2011. doi: 10.1261/rna.2528811.

Atzorn, V., Fragapane, P., and Kiss, T. U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for ribosome assembly. *Cell*, 72:443–457, 2004.

Bachellerie, J. P., Cavaillé, J., and Hüttenhofer, A. The expanding snoRNA world. *Biochimie*, 84 (8):775–90, Aug 2002.

Badis, G., Fromont-Racine, M., and Jacquier, A. A snoRNA that guides the two most conserved pseudouridine modifications within rRNA confers a growth advantage in yeast. *RNA*, 9(7):771–9, Jul 2003.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, 37 (Web Server issue):W202–208, Jul 2009.

Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, Jan 2004.

Bartschat, S., Kehr, S., Tafer, H., Stadler, P. F., and Hertel, J. snoStrip: A snoRNA annotation pipeline. *Bioinformatics*, 30(1):115–116, Jan 2014.

Bartschat, S. *A Homology based annotation pipeline for small nucleolar RNAs in Deuterostomia*. PhD thesis, University of Leipzig, 2011.

Bazeley, P. S., Shepelev, V., Talebizadeh, Z., Butler, M. G., Fedorova, L., Filatov, V., and Fedorov, A. snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene*, 408(1-2):172–9, Jan 2008. doi: 10.1016/j.gene.2007.10.037.

Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008. doi: 10.1186/1471-2105-9-474.

Birkedal, U., Christensen-Dalsgaard, M., Krogh, N., Sabarinathan, R., Gorodkin, J., and Nielsen, H. Profiling of ribose methylations in RNA by high-throughput sequencing. *Angew Chem Int Ed Engl*, 54(2):451–5, Jan 2015. doi: 10.1002/anie.201408362.

Bortolin, M. L. and Kiss, T. Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding. *RNA*, 4(4):445–54, Apr 1998.

Bortolin-Cavaillé, M. L. and Cavaillé, J. The SNORD115 (H/MBII-52) and SNORD116 (H/MBII-85) gene clusters at the imprinted Prader-Willi locus generate canonical box C/D snoRNAs. *Nucleic Acids Res*, 40(14):6800–7, Aug 2012. doi: 10.1093/nar/gks321.

Boulon, S., Verheggen, C., Jády, B. E., Girard, C., Pescia, C., Paul, C., Ospina, J. K., Kiss, T., Matera, A. G., Bordonné, R., and Bertrand, E. PHAX and CRM1 are required sequentially to transport U3 snorna to nucleoli. *Mol Cell*, 16(5):777–87, Dec 2004. doi: 10.1016/j.molcel.2004.11.013.

Braasch, I., Gehrke, A. R., Smith, J. J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A. M., Campbell, M. S., Barrell, D., Martin, K. J., Mulley, J. F., Ravi, V., Lee, A. P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Caestro, C., Sydes, J., Beaudry, F. E., Sun, Y., Hertel, J., Beam, M. J., Fasold, M., Ishiyama, M., Johnson, J., Kehr, S., Lara, M., Letaw, J. H., Litman, G. W., Litman, R. T., Mikami, M., Ota, T., Saha, N. R., Williams, L., Stadler, P. F., Wang, H., Taylor, J. S., Fontenot, Q., Ferrara, A., Searle, S. M., Aken, B., Yandell, M., Schneider, I., Yoder, J. A., Volff, J. N., Meyer, A., Amemiya, C. T., Venkatesh, B., Holland, P. W., Guiguen, Y., Bobe, J., Shubin, N. H., Di Palma, F., Alföldi, J., Lindblad-Toh, K., and Postlethwait, J. H. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet*, 48(4):427–37, Apr 2016. doi: 10.1038/ng.3526.

Brandis, K. A., Gale, S., Jinn, S., Langmade, S. J., Dudley-Rucker, N., Jiang, H., Sidhu, R., Ren, A., Goldberg, A., Schaffer, J. E., and Ory, D. S. Box C/D small nucleolar RNA (snorna) U60 regulates intracellular cholesterol trafficking. *J Biol Chem*, 288(50):35703–13, Dec 2013. doi: 10.1074/jbc.M113.488577.

Bratkovič, T. and Rogelj, B. The many faces of small nucleolar RNAs. *Biochim Biophys Acta*, 1839 (6):438–43, Jun 2014. doi: 10.1016/j.bbagrm.2014.04.009.

Broome, H. J. and Hebert, M. D. Coilin displays differential affinity for specific RNAs in vivo and is linked to telomerase RNA biogenesis. *J Mol Biol*, 425(4):713–24, Feb 2013. doi: 10.1016/j.jmb.2012.12.014.

Brown, J. W., Echeverria, M., Qu, L. H., Lowe, T. M., Bachellerie, J. P., Httenhofer, A., Kastenmayer, J. P., Green, P. J., Shaw, P., and Marshall, D. F. Plant snoRNA database. *Nucleic Acids Res*, 31(1):432–5, Jan 2003.

Brown, J. W., Marshall, D. F., and Echeverria, M. Intronic noncoding RNAs and splicing. *Trends Plant Sci*, 13(7):335–42, Jul 2008. doi: 10.1016/j.tplants.2008.04.010.

Brugiolo, M., Herzel, L., and Neugebauer, K. M. Counting on co-transcriptional splicing. *F1000Prime Rep*, 5:9, 2013. doi: 10.12703/P5-9.

Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Müller, K. M., Pande, N., Shang, Z., Yu, N., and Gutell, R. R. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3:2, 2002.

Cantara, W. A., Crain, P. F., Rozenski, J., McCloskey, J. A., Harris, K. A., Zhang, X., Vendeix, F. A., Fabris, D., and Agris, P. F. The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res*, 39(Database issue):D195–201, Jan 2011. doi: 10.1093/nar/gkq1028.

Canzler, S. *Insights into the Evolution of small nucleolar RNAs*. PhD thesis, University of Leipzig, 2016.

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., and Gilbert, W. V. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, 515(7525):143–6, Nov 2014. doi: 10.1038/nature13802.

Cavaillé, J., Hadjiolov, A. A., and Bachellerie, J. P. Processing of mammalian rRNA precursors at the 3' end of 18S rRNA. identification of cis-acting signals suggests the involvement of U13 small nucleolar RNA. *Eur J Biochem*, 242(2):206–13, Dec 1996.

Cavaillé, J., Buiting, K., Kiefmann, M., Lalande, M., Brannan, C. I., Horsthemke, B., Bachellerie, J. P., Brosius, J., and Httenhofer, A. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc Natl Acad Sci U S A*, 97(26):14311–6, Dec 2000. doi: 10.1073/pnas.250426397.

Cech, T. R. and Steitz, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, 157(1):77–94, Mar 2014. doi: 10.1016/j.cell.2014.03.008.

Chen, C. L., Perasso, R., Qu, L. H., and Amar, L. Exploration of pairing constraints identifies a 9 base-pair core within box C/D snoRNA-rRNA duplexes. *J Mol Biol*, 369(3):771–83, Jun 2007. doi: 10.1016/j.jmb.2007.03.052.

Chen, C. L., Chen, C. J., Vallon, O., Huang, Z. P., Zhou, H., and Qu, L. H. Genomewide analysis of box C/D and box H/ACA snoRNAs in Chlamydomonas reinhardtii reveals an extensive organization into intronic gene clusters. *Genetics*, 179(1):21–30, May 2008. doi: 10.1534/genetics.107.086025.

Chen, W. and Moore, M. J. The spliceosome: disorder and dynamics defined. *Curr Opin Struct Biol*, 24:141–9, Feb 2014. doi: 10.1016/j.sbi.2014.01.009.

Cohen, E., Avrahami, D., Frid, K., Canello, T., Levy Lahad, E., Zeligson, S., Perlberg, S., Chapman, J., Cohen, O. S., Kahana, E., Lavon, I., and Gabizon, R. SNORD3A: a molecular marker and modulator of prion disease progression. *PLoS One*, 8(1):e54433, 2013. doi: 10.1371/journal.pone.0054433.

Cohen, S. B., Graham, M. E., Lovrecz, G. O., Bache, N., Robinson, P. J., and Reddel, R. R. Protein composition of catalytically active human telomerase from immortal cells. *Science*, 315(5820):1850–3, Mar 2007. doi: 10.1126/science.1138596.

Cooper, G. *The Cell: A Molecular Approach*. ASM Press, 2000. ISBN 9780878931026. URL `http://www.ncbi.nlm.nih.gov/books/NBK9863/`.

Courtes, F. C., Gu, C., Wong, N. S., Dedon, P. C., Yap, M. G., and Lee, D. Y. 28S rRNA is inducibly pseudouridylated by the mTOR pathway translational control in CHO cell cultures. *J Biotechnol*, 174:16–21, Mar 2014. doi: 10.1016/j.jbiotec.2014.01.024.

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Blomberg, L. A., Bouffard, P., Burt, D. W., Crasta, O., Crooijmans, R. P., Cooper, K., Coulombe, R. A., De, S., Delany, M. E., Dodgson, J. B., Dong, J. J., Evans, C., Frederickson, K. M., Flicek, P., Florea, L., Folkerts, O., Groenen, M. A., Harkins, T. T., Herrero, J., Hoffmann, S., Megens, H. J., Jiang, A., de Jong, P., Kaiser, P., Kim, H., Kim, K. W., Kim, S., Langenberger, D., Lee, M. K., Lee, T., Mane, S., Marcais, G., Marz, M., McElroy, A. P., Modise, T., Nefedov, M., Notredame, C., Paton, I. R., Payne, W. S., Pertea, G., Prickett, D., Puiu, D., Qioa, D., Raineri, E., Ruffier, M., Salzberg, S. L., Schatz, M. C., Scheuring, C., Schmidt, C. J., Schroeder, S., Searle, S. M., Smith, E. J., Smith, J., Sonstegard, T. S., Stadler, P. F., Tafer, H., Tu, Z. J., Van Tassell, C. P., Vilella, A. J., Williams, K. P., Yorke, J. A., Zhang, L., Zhang, H. B., Zhang, X., Zhang, Y., and Reed, K. M. Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol*, 8(9), 2010. doi: 10.1371/journal.pbio.1000475.

Darnell, J. *RNA: Life's Indispersable Molecule*. Cold Spring Harbor Laboratory Press, 2011. ISBN 9781936113194. URL `https://books.google.de/books?id=cLq4cQAACAAJ`.

Darzacq, X., Jády, B. E., Verheggen, C., Kiss, A. M., Bertrand, E., and Kiss, T. Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. *EMBO J*, 21(11):2746–56, Jun 2002. doi: 10.1093/emboj/21.11.2746.

Decatur, W. A. and Fournier, M. J. rRNA modifications and ribosome function. *Trends Biochem Sci*, 27(7):344–51, Jul 2002.

Deng, W., Zhu, X., Skogerbø, G., Zhao, Y., Fu, Z., Wang, Y., He, H., Cai, L., Sun, H., Liu, C., Li, B., Bai, B., Wang, J., Jia, D., Sun, S., He, H., Cui, Y., Wang, Y., Bu, D., and Chen, R. Organization of the Caenorhabditis elegans small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome Res*, 16(1):20–9, Jan 2006. doi: 10.1101/gr.4139206.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras,

T. R., Hubbard, T. J., Notredame, C., Harrow, J., and Guigó, R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, 22(9):1775–89, Sep 2012. doi: 10.1101/gr.132159.111.

Deryusheva, S., Choleza, M., Barbarossa, A., Gall, J. G., and Bordonné, R. Post-transcriptional modification of spliceosomal RNAs is normal in SMN-deficient cells. *RNA*, 18(1):31–6, Jan 2012. doi: 10.1261/rna.030106.111.

Deschamps-Francoeur, G., Garneau, D., Dupuis-Sandoval, F., Roy, A., Frappier, M., Catala, M., Couture, S., Barbe-Marcoux, M., Abou-Elela, S., and Scott, M. S. Identification of discrete classes of small nucleolar RNA featuring different ends and RNA binding protein dependency. *Nucleic Acids Res*, 42(15):10073–85, Sep 2014. doi: 10.1093/nar/gku664.

Dieci, G., Preti, M., and Montanini, B. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, 94(2):83–8, Aug 2009. doi: 10.1016/j.ygeno.2009.05.002.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Rder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guig, R., and Gingeras, T. R. Landscape of transcription in human cells. *Nature*, 489(7414):101–8, Sep 2012. doi: 10.1038/nature11233.

Doe, C. M., Relcovic, D., Garfield, A. S., Dalley, J. W., Theobald, D. E., Humby, T., Wilkinson, L. S., and Isles, A. R. Loss of imprinted snoRNA mbii-52 leads to increased 5htr2c pre-RNA editing and altered 5HT2CR-mediated behaviour. *Hum Mol Genet*, 18(12):2140–8, Jun 2009. doi: 10.1093/hmg/ddp137.

Dönmez, G., Hartmuth, K., and Lührmann, R. Modified nucleotides at the 5' end of human U2 snRNA are required for spliceosomal E-complex formation. *RNA*, 10(12):1925–33, Dec 2004. doi: 10.1261/rna.7186504.

Douaud, M., Fve, K., Gerus, M., Fillon, V., Bardes, S., Gourichon, D., Dawson, D. A., Hanotte, O., Burke, T., Vignoles, F., Morisson, M., Tixier-Boichard, M., Vignal, A., and Pitel, F. Addition of the microchromosome GGA25 to the chicken genome sequence assembly through radiation hybrid and genetic mapping. *BMC Genomics*, 9:129, 2008. doi: 10.1186/1471-2164-9-129.

Dupuis-Sandoval, F., Poirier, M., and Scott, M. S. The emerging landscape of small nucleolar RNAs in cell biology. *Wiley Interdiscip Rev RNA*, 6(4):381–97, 2015. doi: 10.1002/wrna.1284.

Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2(12):919–29, Dec 2001. doi: 10.1038/35103511.

Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, 2004.

Ellegren, H. The avian genome uncovered. *Trends Ecol Evol*, 20(4):180–6, Apr 2005. doi: 10.1016/j.tree.2005.01.015.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012. doi: 10.1038/nature11247.

Ender, C., Krek, A., Friedländer, M. R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., and Meister, G. A human snoRNA with microRNA-like functions. *Mol Cell*, 32(4):519–28, Nov 2008. doi: 10.1016/j.molcel.2008.10.017.

Falaleeva, M., Surface, J., Shen, M., de la Grange, P., and Stamm, S. SNORD116 and SNORD115 change expression of multiple genes and modify each other's activity. *Gene*, 572(2):266–73, Nov 2015. doi: 10.1016/j.gene.2015.07.023.

Fayet-Lebaron, E., Atzorn, V., Henry, Y., and Kiss, T. 18S rRNA processing requires base pairings of snR30 H/ACA snoRNA to eukaryote-specific 18S sequences. *EMBO J*, 28(9):1260–70, May 2009. doi: 10.1038/emboj.2009.79.

Filipowicz, W. and Pogacić, V. Biogenesis of small nucleolar ribonucleoproteins. *Curr Opin Cell Biol*, 14(3):319–27, Jun 2002.

Ganapathy, G., Howard, J. T., Ward, J. M., Li, J., Li, B., Li, Y., Xiong, Y., Zhang, Y., Zhou, S., Schwartz, D. C., Schatz, M., Aboukhalil, R., Fedrigo, O., Bukovnik, L., Wang, T., Wray, G., Rasolonjatovo, I., Winer, R., Knight, J. R., Koren, S., Warren, W. C., Zhang, G., Phillippy, A. M., and Jarvis, E. D. High-coverage sequencing and annotated assemblies of the budgerigar genome. *Gigascience*, 3:11, 2014. doi: 10.1186/2047-217X-3-11.

Ganot, P., Bortolin, M. L., and Kiss, T. Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, 89(5):799–809, May 1997.

Gardner, P. P., Bateman, A., and Poole, A. M. SnoPatrol: how many snoRNA genes are there? *J Biol*, 9(1):4, 2010. doi: 10.1186/jbiol211.

Gardner, P. P., Fasold, M., Burge, S. W., Ninova, M., Hertel, J., Kehr, S., Steeves, T. E., Griffiths-Jones, S., and Stadler, P. F. Conservation and Losses of Non-Coding RNAs in Avian Genomes. *PLoS One*, 10(3):e0121797, 2015. doi: 10.1371/journal.pone.0121797.

Goldman, D. and Domschke, K. Making sense of deep sequencing. *Int J Neuropsychopharmacol*, 17(10):1717–25, Oct 2014. doi: 10.1017/S1461145714000789.

Goodwin, S., McPherson, J. D., and McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6):333–51, May 2016. doi: 10.1038/nrg.2016.49.

Gray, K. A., Daugherty, L. C., Gordon, S. M., Seal, R. L., Wright, M. W., and Bruford, E. A. Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res*, 41(Database issue):D545–52, Jan 2013. doi: 10.1093/nar/gks1066.

Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., and Bruford, E. A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res*, 43(Database issue):D1079–85, Jan 2015. doi: 10.1093/nar/gku1071.

Griffin, D. K., Robertson, L. B., Tempest, H. G., and Skinner, B. M. The evolution of the avian genome as revealed by comparative molecular cytogenetics. *Cytogenet Genome Res*, 117(1-4):64–77, 2007. doi: 10.1159/000103166.

Griffiths-Jones, S. RALEE–RNA Alignment editor in Emacs. *Bioinformatics*, 21(2):257–9, Jan 2005. doi: 10.1093/bioinformatics/bth489.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res*, 31(1):439–41, Jan 2003.

Gruber, A. R., Findei, S., Washietl, S., Hofacker, I. L., and Stadler, P. F. RNAz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput*, pages 69–79, 2010.

Hebert, M. D. and Matera, A. G. Self-association of coilin reveals a common theme in nuclear body localization. *Mol Biol Cell*, 11(12):4159–71, Dec 2000.

Helm, M. Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic Acids Res*, 34(2):721–33, 2006. doi: 10.1093/nar/gkj471.

Henras, A. K., Plisson-Chastang, C., O'Donohue, M. F., Chakraborty, A., and Gleizes, P. E. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdiscip Rev RNA*, 6(2):225–42, 2015. doi: 10.1002/wrna.1269.

Hertel, J., Hofacker, I. L., and Stadler, P. F. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24(2):158–164, Jan 2008.

Hertel, J., de Jong, D., Marz, M., Rose, D., Tafer, H., Tanzer, A., Schierwater, B., and Stadler, P. F. Non-coding RNA annotation of the genome of Trichoplax adhaerens. *Nucleic Acids Res*, 37 (5):1602–15, Apr 2009. doi: 10.1093/nar/gkn1084.

Higa-Nakamine, S., Suzuki, T., Uechi, T., Chakraborty, A., Nakajima, Y., Nakamura, M., Hirano, N., Suzuki, T., and Kenmochi, N. Loss of ribosomal RNA modification causes developmental defects in zebrafish. *Nucleic Acids Res*, 40(1):391–8, Jan 2012. doi: 10.1093/nar/gkr700.

Hoeppner, M. P. and Poole, A. M. Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC Evol Biol*, 12:183, 2012.

Hoeppner, M. P., White, S., Jeffares, D. C., and Poole, A. M. Evolutionarily stable association of intronic snoRNAs and microRNAs with their host genes. *Genome Biol Evol*, 1:420–8, 2009.

Hofacker, I. L. and Stadler, P. F. *Bioinformatics: From Genomes to Therapies*, chapter RNA Secondary Structure, pages 439–489. T Lengauer(Ed.), Wiley-VCH, Weinheim, Germany, 2007.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. Fast Folding and Comparison of RNA Secondary Structures (The vienna RNA Package). *Monatshefte f. Chemie*, 2(125):167–188, 1994. doi: 10.1007/BF00818163.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F., and Hackermüller, J. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, Sep 2009. doi: 10.1371/journal.pcbi.1000502.

Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., Kunz, M., Holdt, L. M., Teupser, D., Hackermüller, J., and Stadler, P. F. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol*, 15(2):R34, 2014. doi: 10.1186/gb-2014-15-2-r34.

Höner zu Siederdissen, C. and Hofacker, I. L. Discriminatory power of RNA family models. *Bioinformatics*, 26(18):i453–9, Sep 2010. doi: 10.1093/bioinformatics/btq370.

Huang, Y., Li, Y., Burt, D. W., Chen, H., Zhang, Y., Qian, W., Kim, H., Gan, S., Zhao, Y., Li, J., Yi, K., Feng, H., Zhu, P., Li, B., Liu, Q., Fairley, S., Magor, K. E., Du, Z., Hu, X., Goodman, L., Tafer, H., Vignal, A., Lee, T., Kim, K. W., Sheng, Z., An, Y., Searle, S., Herrero, J., Groenen, M. A., Crooijmans, R. P., Faraut, T., Cai, Q., Webster, R. G., Aldridge, J. R., Warren, W. C., Bartschat, S., Kehr, S., Marz, M., Stadler, P. F., Smith, J., Kraus, R. H., Zhao, Y., Ren, L., Fei, J., Morisson, M., Kaiser, P., Griffin, D. K., Rao, M., Pitel, F., Wang, J., and Li, N. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat. Genet.*, 45(7): 776–783, Jul 2013.

Huang, Z. P., Chen, C. J., Zhou, H., Li, B. B., and Qu, L. H. A combined computational and experimental analysis of two families of snorna genes from Caenorhabditis elegans, revealing the expression and evolution pattern of snoRNAs in nematodes. *Genomics*, 89(4):490–501, Apr 2007. doi: 10.1016/j.ygeno.2006.12.002.

Hubé, F. and Francastel, C. Mammalian introns: when the junk generates molecular diversity. *Int J Mol Sci*, 16(3):4429–52, 2015. doi: 10.3390/ijms16034429.

Huerta-Cepas, J., Serra, F., and Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol*, 33(6):1635–8, Jun 2016. doi: 10.1093/molbev/msw046.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L. E., Sibley, C. R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. iCLIP: protein-RNA interactions at nucleotide resolution. *Methods*, 65(3): 274–87, Feb 2014. doi: 10.1016/j.ymeth.2013.10.011.

Hüttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J. P., and Brosius, J. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J*, 20(11):2943–53, Jun 2001. doi: 10.1093/emboj/20.11.2943.

International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, Dec 2004. doi: 10.1038/nature03154.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, Oct 2004. doi: 10.1038/nature03001.

Jády, B. E. and Kiss, T. A small nucleolar guide RNA functions both in 2'-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. *EMBO J*, 20(3):541–51, Feb 2001. doi: 10.1093/emboj/20.3.541.

Jády, B. E., Ketele, A., and Kiss, T. Human intron-encoded Alu RNAs are processed and packaged into Wdr79-associated nucleoplasmic box H/ACA RNPs. *Genes Dev*, 26(17):1897–910, Sep 2012. doi: 10.1101/gad.197467.112.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P., Prosdocimi, F., Samaniego, J. A., Vargas Velazquez, A. M., Alfaro-Nez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jnsson, K. A., Johnson, W., Koepfli, K. P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T., and Zhang, G. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–31, Dec 2014. doi: 10.1126/science.1253451.

Johnston, R. J. and Hobert, O. A microRNA controlling left/right neuronal asymmetry in caenorhabditis elegans. *Nature*, 426(6968):845–9, Dec 2003. doi: 10.1038/nature02255.

Jorjani, H., Kehr, S., Jedlinski, D. J., Gumienny, R., Hertel, J., Stadler, P. F., Zavolan, M., and Gruber, A. R. An updated human snoRNAome. *Nucleic Acids Res*, 44(11):5068–82, Jun 2016. doi: 10.1093/nar/gkw386.

Karijolich, J. and Yu, Y. Spliceosomal snRNA modifications and their function. *RNA Biol.*, 7(2): 192–204, March/April 2010. doi: 10.4161/rna.7.2.11207.

Kass, S., Tyc, K., Steitz, J. A., and Sollner-Webb, B. The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing. *Cell*, 60(6):897–908, Mar 1990.

Kehr, S., Bartschat, S., Stadler, P. F., and Tafer, H. PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics*, 27(2):279–80, Jan 2011. doi: 10.1093/bioinformatics/btq642.

Kehr, S., Bartschat, S., Tafer, H., Stadler, P. F., and Hertel, J. Matching of Soulmates: Co-evolution of SnoRNAs and Their Targets. *Mol. Biol. Evol.*, 31(2):455–467, Feb 2014. doi: 10.1093/molbev/mst209.

King, T. H., Liu, B., McCully, R. R., and Fournier, M. J. Ribosome structure and activity are altered in cells lacking snoRNPs that form pseudouridines in the peptidyl transferase center. *Mol Cell*, 11(2):425–35, Feb 2003.

Kishore, S. and Stamm, S. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, 311(5758):230–2, Jan 2006. doi: 10.1126/science.1118265.

Kishore, S., Khanna, A., Zhang, Z., Hui, J., Balwierz, P. J., Stefan, M., Beach, C., Nicholls, R. D., Zavolan, M., and Stamm, S. The snoRNA MBII-52 (SNORD115) is processed into smaller RNAs and regulates alternative splicing. *Hum Mol Genet*, 19(7):1153–64, Apr 2010. doi: 10.1093/hmg/ddp585.

Kishore, S., Gruber, A. R., Jedlinski, D. J., Syed, A. P., Jorjani, H., and Zavolan, M. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small rna sequencing. *Genome Biol*, 14(5):R45, 2013. doi: 10.1186/gb-2013-14-5-r45.

Kiss, A. M., Jády, B. E., Darzacq, X., Verheggen, C., Bertrand, E., and Kiss, T. A Cajal body-specific pseudouridylation guide RNA is composed of two box H/ACA snoRNA-like domains. *Nucleic Acids Res*, 30(21):4643–9, Nov 2002.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17(7):909–15, Jul 2010. doi: 10.1038/nsmb.1838.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. iCLIP–transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J Vis Exp*, -(50), 2011. doi: 10.3791/2638.

Krell, J., Frampton, A. E., Mirnezami, R., Harding, V., De Giorgio, A., Roca Alonso, L., Cohen, P., Ottaviani, S., Colombo, T., Jacob, J., Pellegrino, L., Buchanan, G., Stebbing, J., and Castellano, L. Growth arrest-specific transcript 5 associated snoRNA levels are related to p53 expression and DNA damage in colorectal cancer. *PLoS One*, 9(6):e98561, 2014. doi: 10.1371/journal.pone.0098561.

Krogh, N., Jansson, M. D., Häfner, S. J., Tehler, D., Birkedal, U., Christensen-Dalsgaard, M., Lund, A. H., and Nielsen, H. Profiling of 2'-O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. *Nucleic Acids Res*, Jun 2016. doi: 10.1093/nar/gkw482.

Ladewig, E., Okamura, K., Flynt, A. S., Westholm, J. O., and Lai, E. C. Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Res*, 22(9):1634–45, Sep 2012. doi: 10.1101/gr.133553.111.

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H. H., Rognes, T., and Ussery, D. W. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*, 35(9):3100–8, 2007. doi: 10.1093/nar/gkm160.

Langenberger, D., Pundhir, S., Ekstrm, C. T., Stadler, P. F., Hoffmann, S., and Gorodkin, J. deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns. *Bioinformatics*, 28(1): 17–24, Jan 2012a. doi: 10.1093/bioinformatics/btr598.

Langenberger, D., Çakir, M. V., Hoffmann, S., and Stadler, P. F. Dicer-Processed Small RNAs: Rules and Exceptions. *J. Exp. Zool: Mol. Dev. Evol.*, 320:35–46, 2012b.

Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4): 357–9, Apr 2012. doi: 10.1038/nmeth.1923.

Lapinaite, A., Simon, B., Skjaerven, L., Rakwalska-Bange, M., Gabel, F., and Carlomagno, T. The structure of the box C/D enzyme reveals regulation of RNA methylation. *Nature*, 502(7472):519–23, Oct 2013. doi: 10.1038/nature12581.

Lestrade, L. and Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*, 34(Database issue):D158–62, Jan 2006. doi: 10.1093/nar/gkj002.

Li, Y., Podlevsky, J. D., Marz, M., Qi, X., Hoffmann, S., Stadler, P. F., and Chen, J. J. Identification of purple sea urchin telomerase RNA using a next-generation sequencing based approach. *RNA*, 19 (6):852–60, Jun 2013. doi: 10.1261/rna.039131.113.

Lindemeyer, M. *Evolution of Spliceosomal RNAs in Metazoan Animals*. PhD thesis, University of Leipzig, 2006.

Liu, J. L., Wu, Z., Nizami, Z., Deryusheva, S., Rajendra, T. K., Beumer, K. J., Gao, H., Matera, A. G., Carroll, D., and Gall, J. G. Coilin is essential for Cajal body organization in Drosophila melanogaster. *Mol Biol Cell*, 20(6):1661–70, Mar 2009. doi: 10.1091/mbc.E08-05-0525.

Lu, Z. J., Turner, D. H., and Mathews, D. H. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res*, 34(17):4912–24, 2006. doi: 10.1093/nar/gkl472.

Lui, L. and Lowe, T. Small nucleolar RNAs and RNA-guided post-transcriptional modification. *Essays Biochem*, 54:53–77, 2013. doi: 10.1042/bse0540053.

Lykke-Andersen, S., Chen, Y., Ardal, B. R., Lilje, B., Waage, J., Sandelin, A., and Jensen, T. H. Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. *Genes Dev*, 28(22):2498–517, Nov 2014. doi: 10.1101/gad.246538.114.

Machyna, M., Heyn, P., and Neugebauer, K. M. Cajal bodies: where form meets function. *Wiley Interdiscip Rev RNA*, 4(1):17–34, 2013. doi: 10.1002/wrna.1139.

Machyna, M., Kehr, S., Straube, K., Kappei, D., Buchholz, F., Butter, F., Ule, J., Hertel, J., Stadler, P. F., and Neugebauer, K. M. The Coilin Interactome Identifies Hundereds of Small Noncoding RNAs that Traffic through Cajal Bodies. *Molecular Cell*, 56(3):389–399, Nov 2014. doi: 10.1016/j.molcel.2014.10.004.

Maden, B. E. Identification of the locations of the methyl groups in 18 s ribosomal RNA from Xenopus laevis and man. *J Mol Biol*, 189(4):681–99, Jun 1986.

Maden, T. Ribosomal RNA. Click here for methylation. *Nature*, 383(6602):675–6, Oct 1996. doi: 10.1038/383675a0.

Makarov, V., Rakitina, D., Protopopova, A., Yaminsky, I., Arutiunian, A., Love, A. J., Taliansky, M., and Kalinina, N. Plant coilin: structural characteristics and RNA-binding properties. *PLoS One*, 8(1):e53571, 2013. doi: 10.1371/journal.pone.0053571.

Makarova, J. A. and Kramerov, D. A. Analysis of C/D box snoRNA genes in vertebrates: The number of copies decreases in placental mammals. *Genomics*, 94(1):11–9, Jul 2009. doi: 10.1016/j.ygeno.2009.02.003.

Makarova, J. A. and Kramerov, D. A. SNOntology: Myriads of novel snoRNAs or just a mirage? *BMC Genomics*, 12:543, 2011. doi: 10.1186/1471-2164-12-543.

Marnef, A., Richard, P., Pinzón, N., and Kiss, T. Targeting vertebrate intron-encoded box C/D 2'-O-methylation guide RNAs into the Cajal body. *Nucleic Acids Res*, 42(10):6616–29, Jun 2014. doi: 10.1093/nar/gku287.

Marz, M., Kirsten, T., and Stadler, P. F. Evolution of spliceosomal snRNA genes in metazoan animals. *J Mol Evol*, 67(6):594–607, Dec 2008. doi: 10.1007/s00239-008-9149-6.

Marz, M., Gruber, A. R., Höner zu Siederdissen, C., Amman, F., Badelt, S., Bartschat, S., Bernhart, S. H., Beyer, W., Kehr, S., Lorenz, R., Tanzer, A., Yusuf, D., Tafer, H., Hofacker, I. L., and Stadler, P. F. Animal snoRNAs and scaRNAs with exceptional structures. *RNA Biol*, 8(6):938–46, 2011. doi: 10.4161/rna.8.6.16603.

Massenet, S. and Branlant, C. A limited number of pseudouridine residues in the human atac spliceosomal UsnRNAs as compared to human major spliceosomal UsnRNAs. *RNA*, 5(11):1495–503, Nov 1999.

Massenet, S., Mougin, A., and Branlant, C. Posttranscriptional Modifications in the U Small Nuclear RNAs. In *Modification and Editing of RNA*, pages 201–227. American Society of Microbiology, 1998. URL `http://www.asmscience.org/content/book/10.1128/9781555818296.chap11`.

Matera, A. G., Terns, R. M., and Terns, M. P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol*, 8(3):209–20, Mar 2007. doi: 10.1038/nrm2124.

Matera, A. G., Izaguire-Sierra, M., Praveen, K., and Rajendra, T. K. Nuclear bodies: random aggregates of sticky proteins or crucibles of macromolecular assembly? *Dev Cell*, 17(5):639–47, Nov 2009. doi: 10.1016/j.devcel.2009.10.017.

Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5):911–40, May 1999. doi: 10.1006/jmbi.1999.2700.

Maticzka, D., Lange, S. J., Costa, F., and Backofen, R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*, 15(1):R17, 2014. doi: 10.1186/gb-2014-15-1-r17.

Mattick, J. S. and Makunin, I. V. Non-coding RNA. *Hum Mol Genet*, 15 Spec No 1:R17–29, Apr 2006. doi: 10.1093/hmg/ddl046.

Maxwell, E. S. and Fournier, M. J. The small nucleolar RNAs. *Annu Rev Biochem*, 64:897–934, 1995. doi: 10.1146/annurev.bi.64.070195.004341.

McCloskey, J. A. and Rozenski, J. The Small Subunit rRNA Modification Database. *Nucleic Acids Res*, 33(Database issue):D135–8, Jan 2005. doi: 10.1093/nar/gki015.

Mei, Y. P., Liao, J. P., Shen, J., Yu, L., Liu, B. L., Liu, L., Li, R. Y., Ji, L., Dorsey, S. G., Jiang, Z. R., Katz, R. L., Wang, J. Y., and Jiang, F. Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene*, 31(22):2794–804, May 2012. doi: 10.1038/onc.2011.449.

Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F., and Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A*, 105(2):716–21, Jan 2008. doi: 10.1073/pnas.0706729105.

Michel, C. I., Holley, C. L., Scruggs, B. S., Sidhu, R., Brookheart, R. T., Listenberger, L. L., Behlke, M. A., Ory, D. S., and Schaffer, J. E. Small nucleolar RNAs U32a, U33, and U35a are critical mediators of metabolic stress. *Cell Metab*, 14(1):33–44, Jul 2011. doi: 10.1016/j.cmet.2011.04.009.

Milek, M., Wyler, E., and Landthaler, M. Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing. *Semin Cell Dev Biol*, 23(2):206–12, Apr 2012. doi: 10.1016/j.semcdb.2011.12.001.

Mitchell, J. R., Cheng, J., and Collins, K. A box H/ACA small nucleolar RNA-like domain at the human telomerase RNA 3' end. *Mol. Cell. Biol.*, 19:567–576, 1999.

Mondal, T., Rasmussen, M., Pandey, G. K., Isaksson, A., and Kanduri, C. Characterization of the RNA content of chromatin. *Genome Res*, 20(7):899–907, Jul 2010. doi: 10.1101/gr.103473.109.

Morris, K. V. and Mattick, J. S. The rise of regulatory RNA. *Nat Rev Genet*, 15(6):423–37, Jun 2014. doi: 10.1038/nrg3722.

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., and Hofacker, I. L. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–82, May 2006. doi: 10.1093/bioinformatics/btl024.

Narayanan, A., Speckmann, W., Terns, R., and Terns, M. P. Role of the box C/D motif in localization of small nucleolar RNAs to coiled bodies and nucleoli. *Mol Biol Cell*, 10(7):2131–47, Jul 1999.

Nawrocki, E. P. and Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–5, Nov 2013. doi: 10.1093/bioinformatics/btt509.

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–7, May 2009. doi: 10.1093/bioinformatics/btp157.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*, 43(Database issue):D130–7, Jan 2015. doi: 10.1093/nar/gku1063.

Ni, J., Tien, A. L., and Fournier, M. J. Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell*, 89(4):565–73, May 1997.

Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J. Algorithms for Loop Matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978. doi: 10.2307/2101031.

Ofengand, J. Ribosomal RNA pseudouridines and pseudouridine synthases. *FEBS Lett*, 514(1): 17–25, Mar 2002.

Ofengand, J. and Bakin, A. Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaebacteria, mitochondria and chloroplasts. *J Mol Biol*, 266(2):246–68, Feb 1997. doi: 10.1006/jmbi.1996.0737.

Ohno, S., Muramoto, J., Stenius, C., Christian, L., Kittrell, W. A., and Atkin, N. B. Microchromosomes in holocephalian, chondrostean and holostean fishes. *Chromosoma*, 26(1):35–40, 1969.

Otto, C., Stadler, P. F., and Hoffmann, S. Lacking alignments? the next-generation sequencing mapper segemehl revisited. *Bioinformatics*, 30(13):1837–43, Jul 2014. doi: 10.1093/bioinformatics/btu146.

Papasaikas, P. and Valcárcel, J. The Spliceosome: The Ultimate RNA Chaperone and Sculptor. *Trends Biochem Sci*, 41(1):33–45, Jan 2016. doi: 10.1016/j.tibs.2015.11.003.

Patra Bhattacharya, D., Canzler, S., Kehr, S., Hertel, J., Grosse, I., and Stadler, P. F. Phylogenetic distribution of plant snoRNA families. *BMC Genomics*, 17(1):969, Nov 2016. doi: 10.1186/s12864-016-3301-2.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glöckner, F. O. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, 35(21):7188–96, 2007. doi: 10.1093/nar/gkm864.

Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature analysis. *Curr Protoc Bioinformatics*, 47:11.12.1–11.12.34, 2014. doi: 10.1002/0471250953.bi1112s47.

Reichow, S. L., Hamma, T., Ferré-D'Amaré, A. R., and Varani, G. The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res*, 35(5):1452–64, 2007. doi: 10.1093/nar/gkl1172.

Richard, P., Darzacq, X., Bertrand, E., Jády, B. E., Verheggen, C., and Kiss, T. A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs. *EMBO J*, 22(16): 4283–93, Aug 2003. doi: 10.1093/emboj/cdg394.

Riley, K. J. and Steitz, J. A. The "Observer Effect" in genome-wide surveys of protein-RNA interactions. *Mol Cell*, 49(4):601–4, Feb 2013. doi: 10.1016/j.molcel.2013.01.030.

Rivas, E. and Eddy, S. R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583–605, Jul 2000.

Rothé, B., Back, R., Quinternet, M., Bizarro, J., Robert, M. C., Blaud, M., Romier, C., Manival, X., Charpentier, B., Bertrand, E., and Branlant, C. Characterization of the interaction between protein Snu13p/15.5K and the Rsa1p/NUFIP factor and demonstration of its functional importance for snoRNP assembly. *Nucleic Acids Res*, 42(3):2015–36, Feb 2014. doi: 10.1093/nar/gkt1091.

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D. P. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, 127(6):1193–207, Dec 2006. doi: 10.1016/j.cell.2006.10.040.

Runte, M., Hüttenhofer, A., Gross, S., Kiefmann, M., Horsthemke, B., and Buiting, K. The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. *Hum Mol Genet*, 10(23):2687–700, Nov 2001.

Sanger, F. and Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*, 94(3):441–8, May 1975.

Schattner, P., Barberan-Soler, S., and Lowe, T. M. A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA*, 12(1):15–25, Jan 2006. doi: 10.1261/rna.2210406.

Schmitz, J., Zemann, A., Churakov, G., Kuhl, H., Grützner, F., Reinhardt, R., and Brosius, J. Retroposed SNOfall–a mammalian-wide comparison of platypus snoRNAs. *Genome Res*, 18(6): 1005–10, Jun 2008. doi: 10.1101/gr.7177908.

Schubert, T., Pusch, M. C., Diermeier, S., Benes, V., Kremmer, E., Imhof, A., and Längst, G. Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. *Mol Cell*, 48(3): 434–44, Nov 2012. doi: 10.1016/j.molcel.2012.08.021.

Schwartz, S., Bernstein, D. A., Mumbach, M. R., Jovanovic, M., Herbst, R. H., León-Ricardo, B. X., Engreitz, J. M., Guttman, M., Satija, R., Lander, E. S., Fink, G., and Regev, A. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*, 159 (1):148–62, Sep 2014. doi: 10.1016/j.cell.2014.08.028.

Scott, M. S. and Ono, M. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. *Biochimie*, 93(11):1987–92, Nov 2011. doi: 10.1016/j.biochi.2011.05.026.

Scott, M. S., Ono, M., Yamada, K., Endo, A., Barton, G. J., and Lamond, A. I. Human box C/D snoRNA processing conservation across multiple cell types. *Nucleic Acids Res*, 40(8):3676–88, Apr 2012. doi: 10.1093/nar/gkr1233.

Shao, P., Yang, J. H., Zhou, H., Guan, D. G., and Qu, L. H. Genome-wide analysis of chicken snoRNAs provides unique implications for the evolution of vertebrate snoRNAs. *BMC Genomics*, 10:86, 2009. doi: 10.1186/1471-2164-10-86.

Shapiro, M. D., Kronenberg, Z., Li, C., Domyan, E. T., Pan, H., Campbell, M., Tan, H., Huff, C. D., Hu, H., Vickrey, A. I., Nielsen, S. C., Stringham, S. A., Hu, H., Willerslev, E., Gilbert, M. T., Yandell, M., Zhang, G., and Wang, J. Genomic diversity and evolution of the head crest in the rock pigeon. *Science*, 339(6123):1063–7, Mar 2013. doi: 10.1126/science.1230422.

Sloan, K. E., Leisegang, M. S., Doebele, C., Ramrez, A. S., Simm, S., Safferthal, C., Kretschmer, J., Schorge, T., Markoutsa, S., Haag, S., Karas, M., Ebersberger, I., Schleiff, E., Watkins, N. J., and Bohnsack, M. T. The association of late-acting snoRNPs with human pre-ribosomal complexes requires the RNA helicase DDX21. *Nucleic Acids Res*, 43(1):553–64, Jan 2015. doi: 10.1093/nar/gku1291.

Solinhac, R., Leroux, S., Galkina, S., Chazara, O., Feve, K., Vignoles, F., Morisson, M., Derjusheva, S., Bed'hom, B., Vignal, A., Fillon, V., and Pitel, F. Integrative mapping analysis of chicken microchromosome 16 organization. *BMC Genomics*, 11:616, 2010. doi: 10.1186/1471-2164-11-616.

Soneo, Y., Taya, Y., Stasyk, T., Huber, L. A., Aoba, T., and Hüttenhofer, A. Identification of novel ribonucleo-protein complexes from the brain-specific snoRNA MBII-52. *RNA*, 16(7):1293–300, Jul 2010. doi: 10.1261/rna.2109710.

Spenkuch, F., Motorin, Y., and Helm, M. Pseudouridine: still mysterious, but never a fake (uridine)! *RNA Biol*, 11(12):1540–54, 2014. doi: 10.4161/15476286.2014.992278.

Spudich, G. M. and Fernández-Suárez, X. M. Touring Ensembl: a practical guide to genome browsing. *BMC Genomics*, 11:295, 2010. doi: 10.1186/1471-2164-11-295.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8, Oct 2002. doi: 10.1101/gr.361602.

Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G. J., and Kellis, M. Systematic discovery and characterization of fly microRNAs using 12 drosophila genomes. *Genome Res*, 17(12):1865–79, Dec 2007. doi: 10.1101/gr.6593807.

Stocsits, R. R., Letsch, H., Hertel, J., Misof, B., and Stadler, P. F. Accurate and efficient reconstruction of deep phylogenies from structured RNAs. *Nucleic Acids Res*, 37(18):6184–93, Oct 2009. doi: 10.1093/nar/gkp600.

Strzelecka, M., Oates, A. C., and Neugebauer, K. M. Dynamic control of Cajal body number during zebrafish embryogenesis. *Nucleus*, 1(1):96–108, 2010. doi: 10.4161/nucl.1.1.10680.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., and Ule, J. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol*, 13 (8):R67, 2012. doi: 10.1186/gb-2012-13-8-r67.

Tafer, H. and Hofacker, I. L. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, 24(22):2657–63, Nov 2008. doi: 10.1093/bioinformatics/btn193.

Tafer, H., Kehr, S., Hertel, J., Hofacker, I. L., and Stadler, P. F. RNAsnoop: efficient target prediction for H/ACA snoRNAs. *Bioinformatics*, 26(5):610–6, Mar 2010. doi: 10.1093/bioinformatics/btp680.

Tafer, H., Amman, F., Eggenhofer, F., Stadler, P. F., and Hofacker, I. L. Fast accessibility-based prediction of RNA-RNA interactions. *Bioinformatics*, 27(14):1934–40, Jul 2011. doi: 10.1093/bioinformatics/btr281.

Taft, R. J., Glazov, E. A., Lassmann, T., Hayashizaki, Y., Carninci, P., and Mattick, J. S. Small RNAs derived from snoRNAs. *RNA*, 15(7):1233–40, Jul 2009. doi: 10.1261/rna.1528909.

Tanaka, R., Satoh, H., Moriyama, M., Satoh, K., Morishita, Y., Yoshida, S., Watanabe, T., Nakamura, Y., and Mori, S. Intronic U50 small-nucleolar-RNA (snoRNA) host gene of no protein-coding potential is mapped at the chromosome breakpoint t(3;6)(q27;q15) of human B-cell lymphoma. *Genes Cells*, 5(4):277–87, Apr 2000.

Terns, M. P. and Terns, R. M. Small nucleolar RNAs: versatile trans-acting molecules of ancient evolutionary origin. *Gene Expr*, 10(1-2):17–39, 2002.

Thompson, J. D., Gibson, T. J., and Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2.3, Aug 2002. doi: 10.1002/0471250953.bi0203s00.

Toffano-Nioche, C., Gautheret, D., and Leclerc, F. Revisiting the structure/function relationships of H/ACA(-like) RNAs: a unified model for Euryarchaea and Crenarchaea. *Nucleic Acids Res*, 43 (16):7744–61, Sep 2015. doi: 10.1093/nar/gkv756.

Tucker, K. E., Massello, L. K., Gao, L., Barber, T. J., Hebert, M. D., Chan, E. K., and Matera, A. G. Structure and characterization of the murine p80 coilin gene, Coil. *J Struct Biol*, 129(2-3): 269–77, Apr 2000. doi: 10.1006/jsbi.2000.4234.

Tucker, K. E., Berciano, M. T., Jacobs, E. Y., LePage, D. F., Shpargel, K. B., Rossire, J. J., Chan, E. K., Lafarga, M., Conlon, R. A., and Matera, A. G. Residual Cajal bodies in coilin knockout mice fail to recruit Sm snRNPs and SMN, the spinal muscular atrophy gene product. *J Cell Biol*, 154(2): 293–307, Jul 2001.

Tyc, K. and Steitz, J. A. U3, U8 and U13 comprise a new class of mammalian snRNPs localized in the cell nucleolus. *EMBO J*, 8(10):3113–9, Oct 1989.

Tycowski, K. T., Shu, M. D., and Steitz, J. A. A mammalian gene with introns instead of exons generating stable RNA products. *Nature*, 379(6564):464–6, Feb 1996. doi: 10.1038/379464a0.

Tycowski, K. T., Aab, A., and Steitz, J. A. Guide RNAs with 5' caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. *Curr Biol*, 14(22):1985–95, Nov 2004. doi: 10.1016/j.cub.2004.11.003.

Tycowski, K. T., Shu, M. D., Kukoyi, A., and Steitz, J. A. A conserved WD40 protein binds the Cajal body localization signal of scaRNP particles. *Mol Cell*, 34(1):47–57, Apr 2009. doi: 10.1016/j.molcel.2009.02.020.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–5, Nov 2003. doi: 10.1126/science.1090095.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet*, 30(9):418–26, Sep 2014. doi: 10.1016/j.tig.2014.07.001.

Wahl, M. C., Will, C. L., and Lührmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–18, Feb 2009. doi: 10.1016/j.cell.2009.02.009.

Wang, T., Xiao, G., Chu, Y., Zhang, M. Q., Corey, D. R., and Xie, Y. Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res*, 43(11):5263–74, Jun 2015. doi: 10.1093/nar/gkv439.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., Searle, S., White, S., Vilella, A. J., Fairley, S., Heger, A., Kong, L., Ponting, C. P., Jarvis, E. D., Mello, C. V., Minx, P., Lovell, P., Velho, T. A., Ferris, M., Balakrishnan, C. N., Sinha, S., Blatti, C., London, S. E., Li, Y., Lin, Y. C., George, J., Sweedler, J., Southey, B., Gunaratne, P., Watson, M., Nam, K., Backström, N., Smeds, L., Nabholz, B., Itoh, Y., Whitney, O., Pfenning, A. R., Howard, J., Völker, M., Skinner, B. M., Griffin, D. K., Ye, L., McLaren, W. M., Flicek, P., Quesada, V., Velasco, G., Lopez-Otin, C., Puente, X. S., Olender, T., Lancet, D., Smit, A. F., Hubley, R., Konkel, M. K., Walker, J. A., Batzer, M. A., Gu, W., Pollock, D. D., Chen, L., Cheng, Z., Eichler, E. E., Stapley, J., Slate, J., Ekblom, R., Birkhead, T., Burke, T., Burt, D., Scharff, C., Adam, I., Richard, H., Sultan, M., Soldatov, A., Lehrach, H., Edwards, S. V., Yang, S. P., Li, X., Graves, T., Fulton, L., Nelson, J., Chinwalla, A., Hou, S., Mardis, E. R., and Wilson, R. K. The genome of a songbird. *Nature*, 464 (7289):757–62, Apr 2010. doi: 10.1038/nature08819.

Weber, M. J. Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet*, 2(12): e205, dec 2006.

Weinberg, Z. and Breaker, R. R. R2R–software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, 12:3, 2011. doi: 10.1186/1471-2105-12-3.

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4): e65, Apr 2007. doi: 10.1371/journal.pcbi.0030065.

Williams, G. T. and Farzaneh, F. Are snoRNAs and snoRNA host genes new players in cancer? *Nat Rev Cancer*, 12(2):84–8, Feb 2012. doi: 10.1038/nrc3195.

Wuchty, S., Fontana, W., Hofacker, I. L., and Schuster, P. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165, Feb 1999.

Xiao, M., Yang, C., Schattner, P., and Yu, Y. T. Functionality and substrate specificity of human box H/ACA guide RNAs. *RNA*, 15(1):176–86, Jan 2009. doi: 10.1261/rna.1361509.

Xue, S. and Barna, M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat Rev Mol Cell Biol*, 13(6):355–69, Jun 2012. doi: 10.1038/nrm3359.

Yang, J. H., Zhang, X. C., Huang, Z. P., Zhou, H., Huang, M. B., Zhang, S., Chen, Y. Q., and Qu, L. H. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res*, 34(18):5112–23, 2006. doi: 10.1093/nar/gkl672.

Yang, J. H., Shao, P., Zhou, H., Chen, Y. Q., and Qu, L. H. deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res*, 38(Database issue):D123–30, Jan 2010. doi: 10.1093/nar/gkp943.

Yin, Q. F., Yang, L., Zhang, Y., Xiang, J. F., Wu, Y. W., Carmichael, G. G., and Chen, L. L. Long noncoding RNAs with snoRNA ends. *Mol Cell*, 48(2):219–30, Oct 2012. doi: 10.1016/j.molcel.2012.07.033.

Yoshihama, M., Nakao, A., and Kenmochi, N. snOPY: a small nucleolar RNA orthological gene database. *BMC Res Notes*, 6:426, 2013. doi: 10.1186/1756-0500-6-426.

Yu, Y. T., Shu, M. D., and Steitz, J. A. Modifications of U2 snRNA are required for snRNP assembly and pre-mRNA splicing. *EMBO J*, 17(19):5783–95, Oct 1998. doi: 10.1093/emboj/17.19.5783.

Yu, Y. T., Shu, M. D., Narayanan, A., Terns, R. M., Terns, M. P., and Steitz, J. A. Internal modification of U2 small nuclear (sn)RNA occurs in nucleoli of *Xenopus* oocytes. *J Cell Biol*, 152 (6):1279–88, Mar 2001.

Zaringhalam, M. and Papavasiliou, F. N. Pseudouridylation meets next-generation sequencing. *Methods*, Mar 2016. doi: 10.1016/j.ymeth.2016.03.001.

Zemann, A., op de Bekke, A., Kiefmann, M., Brosius, J., and Schmitz, J. Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Research*, 34(9):2676–85, 2006. doi: 10.1093|nar|gkl359.

Zhan, X., Pan, S., Wang, J., Dixon, A., He, J., Muller, M. G., Ni, P., Hu, L., Liu, Y., Hou, H., Chen, Y., Xia, J., Luo, Q., Xu, P., Chen, Y., Liao, S., Cao, C., Gao, S., Wang, Z., Yue, Z., Li, G., Yin, Y., Fox, N. C., Wang, J., and Bruford, M. W. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat Genet*, 45(5):563–6, May 2013. doi: 10.1038/ng.2588.

Zhang, Q., Kim, N. K., and Feigon, J. Architecture of human telomerase RNA. *Proc Natl Acad Sci U S A*, 108(51):20325–32, Dec 2011. doi: 10.1073/pnas.1100279108.

Zhang, X. O., Yin, Q. F., Wang, H. B., Zhang, Y., Chen, T., Zheng, P., Lu, X., Chen, L. L., and Yang, L. Species-specific alternative splicing leads to unique expression of sno-lncRNAs. *BMC Genomics*, 15:287, 2014. doi: 10.1186/1471-2164-15-287.

Zhang, Y., Wang, J., Huang, S., Zhu, X., Liu, J., Yang, N., Song, D., Wu, R., Deng, W., Skogerbø, G., Wang, X. J., Chen, R., and Zhu, D. Systematic identification and characterization of chicken (Gallus gallus) ncRNAs. *Nucleic Acids Res*, 37(19):6562–74, Oct 2009. doi: 10.1093/nar/gkp704.

Zhang, Y., Liu, J., Jia, C., Li, T., Wu, R., Wang, J., Chen, Y., Zou, X., Chen, R., Wang, X. J., and Zhu, D. Systematic identification and evolutionary features of rhesus monkey small nucleolar RNAs. *BMC Genomics*, 11:61, 2010. doi: 10.1186/1471-2164-11-61.

Zuker, M. and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–48, Jan 1981.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

.....
(Ort, Datum)

.....
(Unterschrift)

Curriculum Scientiae

Personal Information:

Name	Stephanie Kehr
Phone	0341/ 97 16695
Mail	steffi@bioinf.uni-leipzig.de
Birth	August 04 1983

Working Experience:

since 05/2014	Research Assistent Bioinformatics Group Universität Leipzig Project SFB 1052: Obesity Mechanism
02/2014–04/2014	Research Assistent Theoretical Biochemistry Group at Universität Wien Project: *In silicio* annotation of non-coding RNAs
09/2011–01/2014	Research Assistent Bioinformatics Group Universität Leipzig Project: Quantomics - From Sequence to Consequence: Tools for the Exploitation of Livestock Genomes
06/2008–09/2011	Student Assistent Bioinformatics Group Universität Leipzig
12/2008–04/2009	Internship Theoretical Biochemistry Group at Universität Wien

Education:

since 09/2011	PhD Student Bioinformatics Group Universität Leipzig
10/2007–09/2011	Diploma Student Computer Science with focus on Bioinformatics at Universität Leipzig Thesis: Functional Analysis of small nucleolar RNAs - Target Prediction, Target Conservation, and Hostgenes
10/2002–09/2003	Bachelor Student Bioinformatics at FH OOW Emden

Awards:

FTI Award 2012	Best Diploma Thesis in Computer Science in Germany in 2011 provided by Fakultätentag Informatik (FTI)

Languages:

German	native speaker
English	fluent
Spanish	basic

IT-Knowledge:

Operating System	Linux
Programming	Perl, Shell, PHP, R
Other	LaTeX, HTML, PostScript, mySQL, Emacs, Vim

14. Dezember 2016