Deciphering the landscape of snoRNA-mediated RNA modifications with high-throughput sequencing approaches

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
DOMINIK JAN JEDLINSKI
aus Muri bei Bern

Basel, 2017

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Mihaela Zavolan
Fakultätsverantwortliche und Dissertationsleiterin

Prof. Dr. Helge Grosshans
Korreferent

Basel, den 24.05.2016

Prof. Dr. Jörg Schibler
Dekan

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Mihaela Zavolan for the continuous support during my PhD studies, for her patience, and motivation. Mihaela’s guidance was of immense help in all the time of research and writing of this thesis. At the same time, Mihaela always gave me much freedom to pursue projects of my interest in- and outside of the laboratory. I am very grateful to have worked with such an inspiring scientist.

Besides my supervisor, I would like to thank the rest of my thesis committee: Prof. Helge Grosshans and Prof. David Gatfield for participating in my thesis committee, for their critical comments, and the pleasant scientific discussions.

Then I would like to thank my present and past colleagues from the lab, who where a pivotal reason for why I have truly enjoyed working in our research group. From the first day in the lab, my colleagues helped me to settle in and over the years many friendships have formed. I thank my fellow lab mates for the stimulating discussions and for all the fun we have had in the last four years. In particular, I would like to thank Georges Martin, Shivendra Kishore, Nitish Mittal, Afzal Syed, Arnau Viña Vilaseca, Aaron Grandy, Souvik Ghosh, Yoana Dimitrova, Beatrice Dimitriades, and Alexandra Gnann. I am also really thankful for the great teamwork with the people “from the other side” of the street, the bioinformaticians: Rafal Gumienny, Alexander Kanitz, Andrzej Rzepiela, Andreas R Gruber and Andreas J Gruber, Joao Guimaraes, Foivos Gypas, Ralf Schmidt, Jérémie Breda, and Andrea Riba. My sincere thanks go to Yvonne Steger. Thanks to Yvonne I never had to worry about any administrative hurdles. I am grateful to Harald Witte, who was always helpful and supplied me with mouse samples that were crucial for my research. I would like to thank Aaron Grandy and Pascal Engi for their critical comments on this written work.

Lastly, I would like to thank my family: my parents and my sister who have always supported me throughout my studies and my life in general.

Abstract

Recent years have witnessed a burst of studies in the rapidly developing field of “epitranscriptomics”, which encompasses post-transcriptional changes of transcripts that have a functional relevance. Several new experimental approaches coupled with high-throughput-sequencing enabled the transcriptome-wide mapping of various RNA modifications, including those that are guided by the well-characterized small nucleolar RNAs (snoRNAs). In the projects presented in this thesis, we have taken advantage of these new tools to comprehensively examine snoRNA functions in various cellular systems as well as in a health/disease context.

The first question that we set to answer is how complete is the catalog of human snoRNAs and snoRNA processing products, since it is known that a variety of small RNAs derive from other RNAs with well-known functions such as tRNAs and snoRNAs. To answer this question we sequenced long and short RNAs, RNA fragments obtained in photoreactive nucleotide-enhanced cross-linking and immunoprecipitation (PAR-CLIP) of core snoRNA-associated proteins and small RNAs that co-precipitate with the Argonaute 2 (Ago2) protein. A striking outcome of this study was that virtually all C/D box snoRNAs are specifically processed inside the regions of terminal complementarity, retaining in the mature form only 4-5 nucleotides upstream of the C box and 2-5 nucleotides downstream of the D box. Further we found several new non-coding RNA targets that were repeatedly identified as bound by the core snoRNPs and that were validated as carrying 2’-O-methyl sites and/or pseudouridines. Analysis of the total and Ago2-associated populations of small RNAs in human cells revealed that despite their cellular abundance, snoRNA-derived small RNAs are not efficiently incorporated into the Ago2 protein. We therefore concluded that a miRNA-like function for these products in human is unlikely.

Identification of the targets for the many newly discovered regulatory RNAs remains a challenge. To address this problem, in a second project, we combined two powerful experimental high-throughput methods (CLIP-seq and RiboMeth-seq) with computational modelling to map 2’-O-methylation sites in human rRNA and to comprehensively associate C/D box guide snoRNAs with targets. We thereby determined that many “orphan” snoRNAs appear to guide 2’-O-ribose methylation at sites that are targeted by other snoRNAs. Moreover, we found that snoRNAs can be reliably captured in interaction with many mRNAs, yet a subsequent 2’-O-methylation of these mRNAs cannot be detected. Our study provides a reliable approach to the comprehensive characterization of snoRNA-target interactions in species beyond those in which these interactions have been traditionally studied and contributes to the rapidly developing field of “epitranscriptomics”.

Finally, we applied the same approach to study a particular group of orphan snoRNAs that have been implicated in a rare neurodevelopmental disorder called Prader-Willi syndrome (PWS). PWS is characterized by excessive appetite, morbid obesity, mental and growth retardation, which are due to the loss of paternal expression of the maternally imprinted SNORD116 C/D box snoRNAs. snoRNP-CLIPs in mouse and human cell lines as well as mouse primary neurons revealed that SNORD116 snoRNAs associate with snoRNP proteins, yet the RiboMeth-seq indicates that they do not have a primary snoRNP guide function. Nevertheless, the 2’-O-methylation landscape of wild-type mouse differs from that of a mouse model that does not express Snord116, and the identified candidate target sites are now subject to validation by mass spectrometry.

Publications

Work discussed in this PhD thesis has previously appeared in the following publications:

1. Shivendra Kishore†, Andreas R. Gruber†, Dominik J. Jedlinski, Afzal P. Syed, Hadi Jorjani, Mihaela Zavolan. **Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing**. Genome Biology, 2013. **14**(5): p. R45. doi: 10.1186/gb-2013-14-5-r45.

2. Rafal Gumienny†, Dominik J. Jedlinski†, Georges Martin, Arnau Vina-Vilaseca, Mihaela Zavolan. **High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq.** Nucleic Acids Research, 2016. pii: gkw1321. doi: 10.1093/nar/gkw1321.

3. Hadi Jorjani†, Stephanie Kehr†, Dominik J. Jedlinski, Rafal Gumienny, Jana Hertel, Peter F. Stadler, Mihaela Zavolan, Andreas R. Gruber. **An updated human snoRNAome.** Nucleic Acids Research, 2016. 44(11):5068-82. doi: 10.1093/nar/gkw386.

4. Dominik J. Jedlinski, Rafal Gumienny, Harald Witte, Foivos Gypas, Boris Skryabin, Mihaela Zavolan. **Evaluation of a canonical snoRNA function of Prader-Willi syndrome-associated SNORD116**. *Manuscript in preparation.*

† contributed equally

Contents

Acknowledgements .... **i**

Abstract .... **ii**

Publications .... **iii**

List of Figures .... **vii**

List of Supplementary Figures .... **viii**

List of Tables .... **viii**

Chapter 1 Introduction **1**

1.1 Regulation of gene expression 2

1.1.1 Chromatin accessibility controls gene expression 3

1.1.2 Transcriptional regulation of gene expression 3

1.1.3 Post-transcriptional regulation of gene expression 3

1.1.3.1 Splicing and alternative splicing 4

1.1.3.2 5’ Capping and polyadenylation 4

1.1.3.3 MicroRNA regulation 5

1.1.4 The epitranscriptome 6

1.1.4.1 Small nucleolar RNAs guide RNA nucleotide modifications: Pseudouridylation and 2’-O-ribose methylation 7

1.1.4.2 Function of 2’-O-methylation and pseudouridylation 8

1.1.4.3 Biogenesis of snoRNAs 9

1.1.4.4 Alternative roles of snoRNAs 9

1.1.4.5 SnoRNA target identification 10

1.1.4.6 SnoRNA and disease 10

Chapter 2 Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing **11**

2.1 Abstract 12

2.2 Background 13

2.3 Results 14

2.3.1 PAR-CLIP of C/D box and H/ACA box snoRNP core proteins identifies their RNA binding partners 14

2.3.2 Binding patterns of core proteins on snoRNAs 16

2.3.3 Identification of novel snoRNA genes from PAR-CLIP and small RNA sequencing 18

2.3.4 Target prediction for newly identified snoRNA genes 19

2.3.5 Non-canonical RNA partners of core snoRNA proteins 20

2.3.6 snoRNA processing patterns 21

2.4 Discussion 24

2.5 Materials and methods 26

2.5.1 PAR-CLIP experiments 26

2.5.2 Small RNA sequencing 26

2.5.3 Identification of novel C/D snoRNAs and H/ACA snoRNAs from PAR-CLIP and small RNA sequencing data 27

2.5.4 Detection of 2'-O-ribose-methylated and pseudouridylated residues 27

2.5.5 Ago2 immunoprecipitation sequencing of asynchronous and mitotic cells 28

2.6 Abbreviations 28

2.7 Acknowledgements 29

2.8 Authors' contributions 29

2.9 Electronic supplementary material 29

Chapter 3 High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq 31

3.1 Author’s Contributions 45

3.2 Electronic Supplementary Information 45

Chapter 4 An updated human snoRNAome 47

4.1 Abstract 48

4.2 Introduction 48

4.3 RiM-seq 49

4.4 Methods 50

4.4.1 RiM-seq library preparation and sequencing protocol 50

4.4.2 Read mapping and 2’-O-methylation sites extraction 52

4.5 Discussion of the RiM-seq method 54

Chapter 5 Evaluation of a canonical snoRNA function of Prader-Willi syndrome-associated SNORD116 55

5.1 Introduction 56

5.2 Results 59

5.2.1 PWScr deletion does not affect alternative splicing of Rbfox2-regulated targets in mouse brain 59

5.2.2 Snord116 expression in neuronal samples 61

5.2.3 RiboMeth-seq in wt mice and PWScr$^{m+/p-}$ mice 64

5.2.4 Nop58-HITS-CLIP in SH-SY5Y cells 65

5.2.5 Fibrillarin-HITS/PAR-CLIPs in N1E-115 cells 66

5.2.6 Fibrillarin-HITS-CLIP in mouse primary neurons 67

5.2.7 Combined analysis of RiboMeth-seq and CLIP-seq data 69

5.2.8 Nop58-CLIP in N2A cells 71

5.3 Discussion 72

5.4 Methods 73

5.4.1 Splicing Analysis 73

5.4.2 HITS-CLIP experiments 74

5.4.3 Northern Blots 74

5.4.4 RiboMeth-seq 75

5.4.5 Cell culture and differentiation 75

5.4.6 Preparation of cerebellar granule cells for CLIP-experiments 76

5.4.6.1 Tissue dissection 76

5.4.6.2 Tissue dissociation 76

5.5 Author’s Contributions 77

5.6 Supplementary Figures 77

Chapter 6 Conclusion 79

References 83

List of Figures

Figure 1.1 **Schematic depiction of the two most abundant nucleotide modifications in rRNA and snRNA**..........7

Figure 1.2 **Schematic representation of C/D box snoRNPs (left) and H/ACA box snoRNPS (right).**......................8

Figure 2.1 **Summary of PAR-CLIP data of snoRNP core proteins**. 17

Figure 2.2 **Small RNA-seq and PAR-CLIP reads mapping to mini-snoRNAs**. 19

Figure 2.3 **Predicted structure of hybrids between novel snoRNAs and target RNAs**.......................................20

Figure 2.4 **Terminal processing of C/D box snoRNAs**..23

Figure 4.1 **RiM-seq protocol**. 49

Figure 4.2 **RiM-seq readout**. 50

Figure 4.3 **Plot of the $\log_2$ G4-to-G1 stoppage ratios**..52

Figure 4.4 **Plot of the G4-to-G1 $\log_2$ stoppage ratio** 53

Figure 4.5 **Matthews correlation coefficient plot**..53

Figure 4.6 **ROC curve and PR curve**..53

Figure 5.1 **Growth Differences among PWScr$^{m+/p-}$ and PWScr$^{m+/p+}$ siblings.** 57

Figure 5.2 **Rbfox binding in the Snord116 locus.**...59

Figure 5.3 **Distribution of expression changes of HITS-CLIP-determined Rbfox targets**...................................60

Figure 5.4 (A) **Exon inclusion/exclusion rates in wt and PWScr$^{m+/p-}$ mice.** 60

Figure 5.5 **Scatterplot of PSI scores of all possible Skipped Exons (SE)** 61

Figure 5.6 **Northern blots depicting the expression of Snord 116 (left) and Snord 115 (right)** 62

Figure 5.7 **Northern blots depicting expression of Snord116/SNORD116**..63

Figure 5.8 **Northern blot depicting Snord116 expression in brain tissue from wt and PWScr$^{m+/p-}$ mice (ko).**....64

Figure 5.9 **Correlation between angle score and score C on known methylation sites.**...................................65

Figure 5.10 **Snapshot from the CLIPZ genome browser.**...68

Figure 5.11 **Comparison of snoRNA expression levels calculated from Fibrillarin-CLIP** 69

Figure 5.12 **2'-O-methylation sites and CLIP interactions identified in 18S rRNA.**..70

Figure 5.13 **2'-O-methylation sites and CLIP interactions found in 28S rRNA.**..70

Figure 5.14 **Regions of 18S and 28S rRNAs showing known as well as novel positions of 2'-O-methylation** 71

Figure 5.15 **PR curve (A) and Matthews coefficient plot (B) of the angle score.** 75

List of Supplementary Figures

Supplementary Figure 5.1 **Murine neuroblastoma N1E-115 cells** 77

List of Tables

Table 2.1 **Summary of CLIPZ mapping statistics and annotation categories for PAR-CLIP samples.** 15

Table 2.2 **Annotation summary of the top 200 clusters inferred from PAR-CLIP**..16

Table 2.3 **Functional annotation of sequencing reads obtained in sRNA sequencing and HeLa Ago2 IP sequencing.** 22

Table 4.1 **Reverse transcription conditions used in RiM-seq**. 51

Chapter 1 Introduction

With the start of the human genome project (HGP) in 1990 research in biology entered a new and very exciting era. The project was the world's largest collaborative biological project [1, 2] having the goal of determining the sequence of base pairs that make up the *Homo sapiens* genome. The long-term purpose of this undertaking was to identify all the genes and their location in the genome, and to characterize their function in health and disease at an unprecedented level of detail. The HGP was successfully completed in early 2003 [1], providing the public with a high-quality version of the human genome sequence. This landmark event led to a paradigm shift in biomedical research and propelled it into the genomic and further into the post-genomic era, where we are today. Combining whole genome sequencing of multiple individuals with other information such as their medical history, the task of identifying genes responsible for diseases, once requiring large research teams, many years of work, and immense financial spending can be accomplished today within a few weeks by a single graduate student with access to deoxyribonucleic acid samples and associated phenotypes, an internet connection to the public genome databases, a thermal cycler, and a sequencing machine.

The implications of deciphering human and non-human genomes for biological and medical science are vast. Since all living organisms are related through evolution and store and process genetic information using the same molecules - deoxy ribonucleic acid (DNA) and ribonucleic acid (RNA) – powerful comparative genomics approaches [3] enabled rapid functional elucidation of many newly identified genome-encoded elements and a better understanding of how genetic networks and protein pathways contribute to cellular and organismal phenotypes. Analyzing the genetic makeup of increasing numbers of volunteers and patients progressively makes it possible to develop a detailed understanding of the heritable variation in the human genome [4-6]. Furthermore, genome studies particularly of model organisms such as yeast, worm, and mouse revealed many fundamental processes that are common to all living organisms, and at the same time opened the way for research directions that could translate into health benefits. Genes and pathways with a role in health and disease and their interactions with environmental factors can be identified more efficiently and studied in detail. The advent and progress of sequencing technologies facilitated the development of diagnostic methods for the prediction of susceptibility to disease, the prediction of drug response, the early detection of illness, and the accurate molecular classification of disease.

Paradoxically, the more data has been generated, the more seems to be necessary to be able to interpret the results of large-scale experiments. It is easy to underestimate the challenge of understanding the behavior of molecular networks with thousands of components considering that the number of conditions in which the operation of these networks was observed was orders of magnitude smaller. Furthermore, although sequencing technologies have become more efficient and affordable every year, the development of tools to analyze all of these data has lagged behind. It became more evident, if that were necessary, how important bioinformatics and computational biology are if one wants to take full advantage of these pioneering technologies and the richness of the generated data. On the other hand, the anticipation that all diseases would be fundamentally understood and would become more readily treatable once the genome was deciphered was curbed. Numerous human diseases are not monogenic but rather caused by the joint contribution of a number of independently acting or interacting genes and/or other non-genetic factors [7].

In addition, the contribution of individual genes to a particular phenotype may be small or context-dependent, making it very challenging to assign multifactorial diseases to a genetic locus in the DNA. Nonetheless, more than a decade after the announcement of the successful completion of the HGP, the hard work of biologists, computational scientists, and clinicians has laid a solid foundation that is embodied in many knowledge bases that enable researchers to explore many systems of their interest, putting new results in the perspective of the vast amount of data that has been collected so far by researchers worldwide.

With the current tools at hand, this is a truly exciting time for researchers to immerse themselves into the world of genomics to further pursue open questions in various fields ranging from fundamental biological processes to mechanisms that contribute to health or disease in humans.

Working together with interdisciplinary scientists at the interface between experimental biology and computation provided me with an excellent environment to carry out research along these lines. With the work presented here I hope to contribute my grain of knowledge to the scientific community.

1.1 Regulation of gene expression

All living organisms store and process their hereditary information using the same molecules, DNA and RNA [8]. The chemical building blocks of DNA and RNA are the nucleotides, each consisting of a sugar (deoxyribose in DNA and ribose in RNA), a phosphate group, and a nitrogenous base. There are four different DNA nucleotides, each with a specific base: adenine (A), thymine (T), guanine (G), and cytosine (C). The chemical structure of the nucleotides is such that they pair with each other through hydrogen bonds, A pairing with T, and G with C. A and T, and G and C are called complementary nucleotides. When the nucleotides are strung together in a chain through phosphodiester bonds, they form a structure known as polynucleotide [8], which has a directionality, one end of the strand being called the 5'end, the other the 3'end. Complementary DNA polynucleotides coil around each other to form a double helix. A specific sequence of nucleotides can make up a gene, the physical and functional unit of heredity that carries the information required for constructing RNA polymers in a process called transcription [9]. These RNA polymers can either serve as a template for the translation of proteins (in this case they are referred to as messenger RNA (mRNA)), or the RNA itself can be a functional entity itself, as is the case for ribosomal RNA (rRNA), spliceosomal RNA (snRNA), transfer RNA (tRNA), microRNAs (miRNAs), or small nucleolar RNAs (snoRNAs). The process of generating a functional effector molecule from a gene is referred to as gene expression. It is intriguing that all the cells of a eukaryotic multicellular organism such as the human carry the same genetic information, yet their phenotypes and functions can be highly distinct (e.g. neuron vs. muscle cell). This functional diversity is achieved through different patterns of gene expression in different cell types. The different expression patterns in turn are brought about by a complex regulation of gene expression, which occurs at various steps of the process (DNA/chromatin level, transcription level, and post-transcriptional level) through various mechanisms [8].

In the subsequent sections I will very briefly introduce the different layers of gene expression regulation and then particularly elaborate on post-transcriptional regulation with focus on the "epitranscriptome" and "snoRNA-mediated 2'-O-methylation", since the work presented in this thesis relates to these mechanisms.

1.1.1 Chromatin accessibility controls gene expression

In eukaryotic cells, the DNA is tightly folded and wrapped around histone proteins. One consequence of this packaging is that under normal circumstances, most of the DNA is not readily accessible to the RNA polymerase and Transcription Factors (TFs) [8]. Tightly folded and inaccessible DNA is referred to as heterochromatin, whereas accessible DNA is referred to as euchromatin. Thus, by selectively changing the accessibility of certain segments of the DNA to the transcription apparatus at specific times, eukaryotic cells can control gene expression simply by making DNA sequences sterically available to RNA polymerase binding [8]. The transition from "active" euchromatin to "silent" heterochromatin is regulated by histone modification. Among numerous histone modifications, the methylation and acetylation of specific lysine residues on the N-terminal histone tails are the best studied and are fundamental for the formation of euchromatin and heterochromatin [10]. For example, the acquisition of active chromatin marks, such as the acetylation of H3K9 and the addition of two or three methyl groups to H3K4 (H3K4Me2/3), is associated with chromatin decondensation which distinguishes actively transcribed genes from other genes [11]. Repressed genes are characterized by marks such as H3K27Me3 [12], H3K9Me2/3 [13, 14], H4K20Me3 [15]. The DNA itself can be subject to modification that impacts the chromatin state, for instance CpG islands are often found to be methylated on the fifth residue of the cytosine base [16]. DNA methylation is essential for mammalian development and DNA methylations are particularly frequent in the genome. The methylation state of CpG islands in promoters can impact the transcriptional activity of a gene [16]. A low level of CpG island methylation in promoters is associated with active transcription, whereas high methylation is associated with silenced genes. Histone and DNA modifications are brought about by various chromatin-interacting proteins that can reversibly shape the transcriptional status of a gene locus, depending on the need of the cell [10, 16].

1.1.2 Transcriptional regulation of gene expression

Another means by which cell fates and complex body plans are established is through TF-dependent, cell-type-specific transcription regulation. Again in eukaryotes, the transcriptional control of gene expression is very complex and involves numerous TFs. Initiation of transcription includes the binding of RNA Polymerase II and general TFs to the core promoter, a region of approximately 40 base pairs upstream and downstream of the transcription start site [17]. Transcription initiation is modulated by various *cis*-regulatory modules such as enhancers, which can be located up to 1'000'000 base pairs away from the transcription start site [17], and are bound by proteins to activate transcription. Repressor elements, also located upstream of transcription start sites, that bind repressor proteins to silence gene expression are also known [18]. It is estimated that the genome contains hundreds of thousands of such regulatory elements [19] that govern the gene expression. The activity of these regulatory elements can be restricted to a particular tissue or cell type, a time point in life, or to specific physiological, pathological or environmental conditions. This is accomplished through a variety of mechanisms [17, 20].

1.1.3 Post-transcriptional regulation of gene expression

Once the gene product is produced through transcription, its stability, subcellular traffic and localization, as well as its interactions with other cellular components are influenced through numerous processes that are referred to as "post-transcriptional regulation". They are briefly summarized in the following sections.

1.1.3.1 Splicing and alternative splicing

Splicing of mRNA precursors (pre-mRNAs), the removal of introns and joining of the exons, is a crucial step in the expression of most genes in higher eukaryotes. Although often referred to as a post-transcriptional mechanism, splicing also occurs co-transcriptionally [21], being carried out by the spliceosome, a large structure consisting of five small nuclear ribonucleoprotein particles (snRNPs) and a large number of proteins that cooperate to accurately recognize a splice site and to catalyze the splicing reaction[22]. The sequence of exons that that is spliced together may differ between cell types or conditions, leading to alternative splicing. Alternative splicing is an important mechanism for transcript and protein diversification in higher eukaryotes. The resulting proteins differ in their peptide sequence and hence in their chemical and biological activities [23]. Through alternative splicing many more proteins can be synthesized from the genome than would be expected from 20’000 protein-coding genes [24]. Interestingly, splicing requires the presence of five uridyl-rich snRNAs, namely U1, U2, U4, U5, and U6. snRNAs are closely associated with 6-10 proteins each, and they base pair with the pre-mRNA in the spliceosome complex that contains approximately 170 proteins [25].

1.1.3.2 5’ Capping and polyadenylation

All eukaryotic pre-mRNAs are modified at their two ends in the process of generating a mature mRNA. When an RNA is transcribed, the 5’ end of the nascent RNA chain that emerges from the surface of RNA polymerase is immediately targeted by several enzymes that together synthesize the 5’cap, a 7-methylguanylate that is connected to the terminal nucleotide of the RNA [25]. The cap protects the mRNA from enzymatic degradation through 5’-exoribonucleases and is important for its export to the cytoplasm.
In eukaryotes, all mRNAs except the histone mRNAs, have a 3’ poly(A) tail, which is added through a complex mechanism that starts with the 3’-end cleavage. Essential for the reaction is a sequence called poly(A) signal, which is usually AAUAAA and is located 10-35 nucleotides upstream of the cleavage site (also called poly(A) site), where cleavage takes place. A multitude of proteins associating in the 3’-end processing complex are involved in this process, of which the most generally involved are [25, 26]: the cleavage and polyadenylation specificity factor (CPSF), which first binds to and forms an unstable complex with the upstream poly(A) signal, the cleavage stimulatory factor (CStF), which interacts with a G/U-rich sequence typically located downstream of the cleavage site, and the cleavage factors I and II (CFI, CFII). Finally, poly(A) polymerase (PAP) binds to the complex before the cleavage can occur, in order that the free 3’ end generated after cleavage is rapidly polyadenylated and no essential information is lost to exonuclease degradation of an unprotected 3’ end. As soon as the synthesis of the poly(A) tail starts, poly(A)-binding protein (PABP) binds to the short A tail initially added by PAP, stimulating the further addition of A nucleotides. Once 200-250 nucleotides are added, the poly(A) tail allows the mature mRNA to be exported from the nucleus.

As discussed above, mature mRNAs must have their ends protected to avoid being degraded by nuclear exonucleases. Interestingly, these mechanisms can also be fine-tuned and impact the gene expression. Most genes typically have several poly(A) sites, leading to different isoforms of the gene product in a process called alternative polyadenylation. The different isoforms can e.g. have 3’ untranslated regions (3’-UTR) that differ in their length, where the shorter 3’-UTRs can be devoid of miRNA- or other RNP-binding sites and thus are associated with an altered protein output [26].

1.1.3.3 MicroRNA regulation

Once an mRNA is transported to the cytoplasm it is subject to several mechanisms that can control its stability and the efficiency of its translation to proteins. MiRNA-dependent regulation of translation is one of these mechanisms. It is a widespread post-transcriptional mechanism that can be found in all multicellular plants and animals. I deliberately provide only a short description of miRNAs, since miRNA-regulation is not the major focus of the work presented here. However, there are some parallels between miRNAs and other guide RNAs that I would like to emphasize.

MiRNAs are short, non-coding RNAs, approximately 22 nucleotides long that regulate the expression of target mRNAs. Since their discovery, it has become clear that miRNAs are involved in numerous biological processes and they are essential for organism development [27]; e.g. several miRNAs have been demonstrated to be crucial for development in *Caenorhabditis elegans* [28, 29] and in *Danio rerio* [30]. Because miRNAs are involved in the normal functioning of eukaryotic cells, it is not surprising that deregulation of miRNAs can result in disease. MiRNAs have been implicated in various diseases such as cancer and heart disease [31]. MiRNAs are transcribed from the genome, their primary transcripts folding into stem-loops that contain individual miRNAs. After passing through the miRNA biogenesis cascade (reviewed in [32]), one of the two strands of the resulting RNA duplex is loaded into a mature RNA-induced silencing complex (RISC). The single-stranded miRNA is bound by the multidomain Argonaute (Ago) protein which in many organisms has multiple homologues [25]. The miRNA-RISC complex then associates with target mRNAs by base pairing between the Ago-bound mature miRNA and complementary regions that are located predominantly in the 3'-UTR of the mRNA. This leads to the repression of target expression. There has been much debate regarding the repression of the mRNA and its precise mechanism. It is now generally accepted that binding of the miRNA-RISC complex to its targets, at least in animals, leads to an initial translational inhibition, later followed by mRNA destabilization (reviewed in [33]).

1.1.3.3.1 MicroRNA target identification/prediction

According to most recent estimations there are more than 6'000 miRNAs in the human genome, many of them only expressed in specific cell types [34]. Determining the function/targets of all these miRNAs has been a highly active area of research. The straightforward identification of miRNA targets has been hampered by the fact, that in human and animals the degree of complementarity between mRNA and miRNAs is only partial, involving about 7-8 nucleotides at the 5' end of the miRNA [35].Based on a developed biophysical model that takes into account additional interactions (other than the 2-7 seed) between miRNA and mRNA, our group has developed a tool to predict targets to improve and aid the identification of miRNA targets from cross-linking and immunoprecipitation (CLIP) data sets [36, 37], complementing other miRNA-target prediction tools [38-40].

CLIP experiments have been the state-of the art assay to capture both a guide RNA and its target from a ribonucleprotein complex [41, 42]. Because in humans and mice the main RISC effector is Arognaute 2 (Ago2), numerous Ago2-CLIPs have been performed with the intent to identify miRNA and corresponding targets [43-46]. However, until very recently [47], these approaches did not simultaneously capture the miRNA and the target. The task of identifying the guide miRNA for a specific Ago2-CLIP site was solved computationally [36-40] until it has been noticed that Ago2-CLIP data sets lead to the capture of miRNA-target interactions in the form of chimeric reads [47]. These are thought to form due to cellular enzymes ligating the guide RNA to its target RNA during CLIP [47-49]. The chimeras that form between miRNA and target in Ago2-CLIP experiments were first used for the systematic discovery of unambiguous miRNA-target interactions *in vivo* [47].

1.1.4 The epitranscriptome

In addition to the well-characterized post-transcriptional modifications that I outlined above, there are over 100 distinct chemical modifications that can be catalyzed on RNA nucleotides post-synthesis [50], potentially serving as yet another regulatory layer of gene expression.

Long before the first genome was sequenced, various nucleotide modifications of DNA had already been described, such as 5-methylcytosine [51] and 5-hydroxyl-methylcytosine [52]. Presently, numerous DNA modifications have been reported [53] and together with histone modifications they constitute important regulatory mechanisms for controlling gene expression and function. The sum of all DNA and histone modifications is often referred to as the "epigenome". Characterizing DNA modifications has become relatively easy, since approaches like bisulfite sequencing have significantly contributed to decipher the epigenetic landscape, and large-scale projects such as the NIH Roadmap Epigenomics Mapping Consortium [54] or the BLUEPRINT Consortium [55] are well underway to produce a rich resource of human epigenomic data from various human tissues and organs.

Likewise post-translational modifications of proteins, sometimes referred to as "epiproteome"[56], are well-recognized mechanisms necessary for the regulation of protein activity. Post-translational modification can occur on the amino acid side chains or at the protein's termini. Phosphorylation for instance, is the most common post-translational modification and is essential for regulating the activity of enzymes [57]. It is estimated that there are over 200 post-translational modifications in human adding to the complexity of the proteome [58].

The regulatory layer that lies between DNA and proteins, called the "epitranscriptome", is far less understood. Only in recent years, the development of high throughput methods enabled the transcriptome-wide study of various nucleotide modifications. One of the most studied nucleotide modification is the methyl-6-adenosine (m6A) which is found in thousands of mammalian genes [59, 60]. m6A was shown to be enriched in specific regions of mRNA, namely near the beginning of the 3'-UTR. Despite these advances in the mapping of m6A, the purpose and molecular function of m6A is still unknown. Several hypotheses were made (reviewed in [61]). The proposed functional implications of m6A are in protein recruitment, conformational change of RNA, effects on mRNA splicing, regulation of mRNA translation, and effects on mRNA expression and degradation. Initial mapping approaches localized m6A residues to transcript regions 100-200 nucleotides-long and could not identify the precise m6A positions, which made it challenging to answer questions regarding the precise molecular mechanism of m6A. However very recent work demonstrated transcriptome-wide single-nucleotide-resolution mapping of m6A [62]. This advancement is a bold example for the dynamics and the rapid technology development found in this highly active research area. Several other RNA modifications have been mapped in high throughput fashion and they include 5 methylcytosine [63], pseudouridine [64], and 2'-O-ribose methylation [65]. We took advantage of these emergent technologies, taking a recently developed high throughput assay called RiboMeth-seq [65], used to map snoRNA guided 2'-O-ribose methylation (2'O-Me), to improve it (CHAPTER 3), and apply it to study 2'-O-Me's and snoRNAs in a health and disease context (CHAPTER 5).

1.1.4.1 Small nucleolar RNAs guide RNA nucleotide modifications: Pseudouridylation and 2’-O-ribose methylation

SnoRNAs belong to a large and abundant family of small non-coding RNAs crucial for ribosome biogenesis and snRNA function. SnoRNAs can be found in all eukaryotes as well as archaea and they form well-characterized ribonucleoprotein complexes referred to as snoRNPs [66]. There are two main classes of snoRNAs, the box C/D and the box H/ACA snoRNAs, which differ in terms of their characteristic motifs, structure and in their protein binding preferences. C/D box snoRNPs (each snoRNA is complexed with four proteins; Fibrillarin, NOP56, NOP58, and 15.5K) guide and catalyze site-specific 2’-O-methylation of the RNA ribose (FIGURE 1.1 from [67]). H/ACA box snoRNPs (each snoRNA is complexed with Dyskerin, NHP2, GAR1, and NOP10) direct site-specific isomerization of the nucleoside uridine in a process called pseudouridylation (FIGURE 1.1) [68].

Typical C/D box snoRNAs are between 60 and 90 nucleotides in length and have characteristic conserved boxes C (consensus sequence RUGAUGA) and D (consensus sequence CUGA) near their 5’ and 3’ ends, respectively (FIGURE 1.2) [69-71]. The C and D boxes align and fold into the so-called kink-turn motif, and the ends of the snoRNA form a double-stranded stem structure [72]. This structure at the end of the snoRNAs is essential for biogenesis and proper localization, serving as a binding site for core for box C/D snoRNP proteins [69, 72]. C/D box snoRNAs carry additional motifs, the boxes C’ and D’ which have the same consensus sequences as the boxes C and D, respectively, and are found in the central region of the molecule, but typically these motifs are less well conserved and often degenerate. The snoRNA guide regions with complementarity to the targets are located directly upstream from the boxes D’ and/or D. The target nucleotide that pairs with the fifth nucleotide of the snoRNA anti-sense sequence acquires the 2’-O-Me mark [69, 72]. As previously mentioned, 2’-O-methylation requires the C/D box snoRNA to associate with a set of four evolutionary proteins, Fibrillarin, NOP56, NOP58, and 15.5K. The enzymatic activity is mediated by the highly conserved RNA methyltransferase fibrillarin [73].

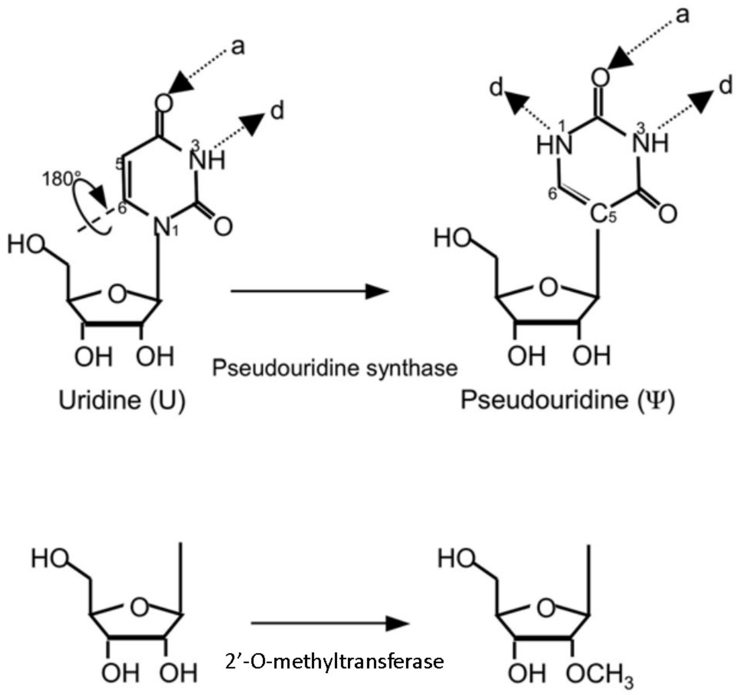


Figure 1.1 **Schematic depiction of the two most abundant nucleotide modifications in rRNA and snRNA**. Top: Pseudouridine is a rotational isomer of uridine, with one additional hydrogen bond donor (d), while the number of hydrogen bond acceptors (a) is unchanged. Bottom: Schematic representation of a 2’-O-methylated ribose.

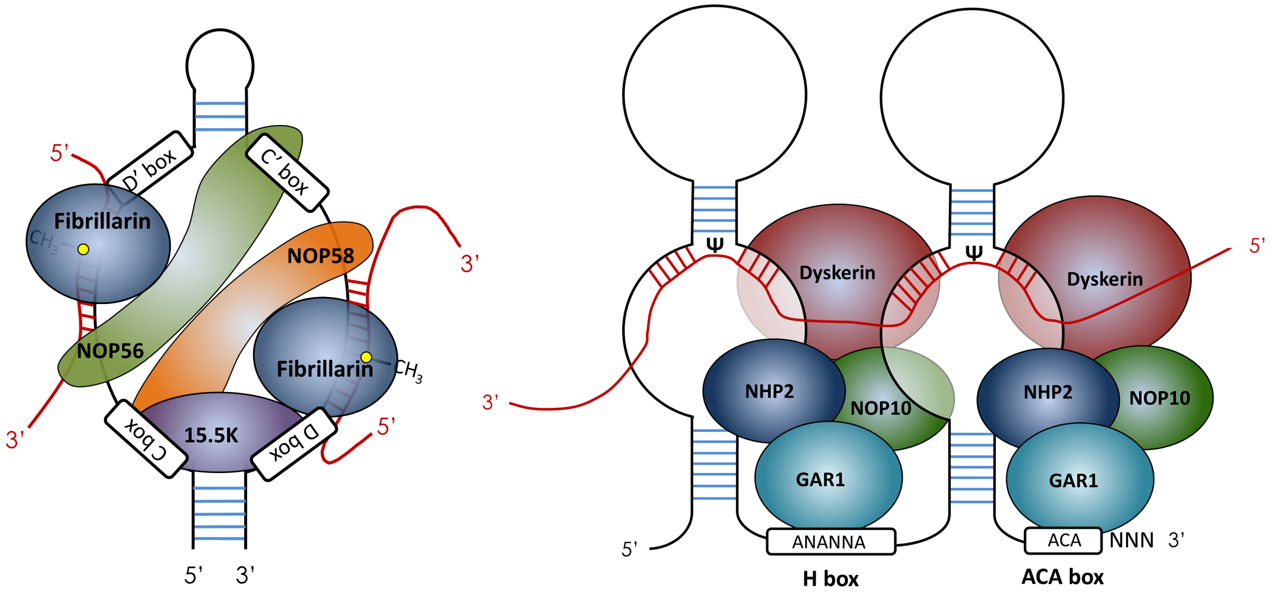


Figure 1.2 **Schematic representation of C/D box snoRNPs (left) and H/ACA box snoRNPS (right).** Target RNA is depicted in red, modified nucleotides are indicated by CH3 and Ψ for 2'-O-methylation and pseudouridylation, respectively.

The longer H/ACA box snoRNAs, which range from 120 to 140 nucleotides in length, display a characteristic secondary structure consisting of two hairpins (FIGURE 1.2). The two hairpins are connected by a hinge region which is formed of the H box (ANANNA where N can be any nucleotide). Upstream of the 3'-end of the molecule and immediately downstream of the second hairpin there is a highly conserved ACA box [74]. The guide regions of the H/ACA box snoRNAs can be found in the middle of the hairpins, specifying by complementarity the exact position to be pseudouridylated in the target. The pseudouridine site in the target RNA is typically located 14-15 nucleotides upstream from the H or ACA box [72, 74]. The enzymatic activity of H/ACA snoRNPs is conferred by the pseudouridine synthase Dyskerin [68]. Both C/D box snoRNAs and H/ACA box snoRNAs can recognize up to two different substrates [74].

1.1.4.2 Function of 2'-O-methylation and pseudouridylation

rRNA and spliceosomal snRNAs are the canonical targets of snoRNAs and carry numerous 2'-O-methylations and pseudouridines [75, 76]. Over the last few decades, much effort has been directed at studying the mechanisms by which these modifications are introduced as well as their molecular functions. To date, progress has been most significant in the area regarding the introduction of these modifications, and recent advances including the efforts of our group (CHAPTER 3) have now enabled the transcriptome wide study of pseudouridines [64] and 2'-O-Me's [65] and their H/ACA box and C/D box snoRNA guides, respectively. However, knowledge regarding the function of posttranscriptional modifications including 2'-O-Me's and pseudouridines has lagged behind.

Although snRNAs had been known to be extensively post-transcriptionally modified since their discovery, only in the 1990s their function became clear. Particularly studies performed on U2 and U6 snRNA shed light on the importance these nucleotide modifications. To this end, reconstitution systems were developed that involved the specific depletion of one of the endogenous spliceosomal snRNAs followed by supplementation of that respective snRNA synthesized *in vitro* [77]. As in *in vitro* synthesized snRNAs lack modifications, the ability or the lack thereof of the RNA to reconstitute pre-mRNA splicing would indicate whether the modifications were required for pre-mRNA splicing. Summarizing, several studies using reconstitution systems, performed in yeast [77-79], *Xenopus* [80, 81], and HeLa cells [82], led to the conclusion that post-transcriptional modifications of snRNAs are essential for proper pre-mRNA splicing. Particularly snRNP assembly is impaired when snRNAs are void of modifications [80].

rRNA is an integral component of the ribosome, and similar to snRNAs, it is subjected to extensive post-transcriptional modifications, with 2'-O-methylation and pseudouridylation being the most common modifications. Several lines of evidence suggest rRNA modifications are important for ribosome function. For instance, the analysis of a three-dimensional map obtained from *Escherichia coli* and *S. cerevisiae* ribosomes indicates the clustering of modifications in functionally important regions of the ribosome [83]. Even though it is difficult to elucidate the functions of individual modifications, since most deletions of any snoRNA alone result only in a minor phenotype [84], global deletions of 2'-O-methylation and pseudouridylation through mutations in Nop1 [85] and

Cbf5p (yeast homolog of Dyskerin) [86], respectively, resulted in significant growth defects and defects in ribosome assembly in *S. cerevisiae.*

The data are clear that post-transcriptional modifications within the rRNA and snRNA are important for pre-mRNA splicing and protein synthesis. The mechanisms behind how these modifications exert the effect are still not well understood. However, it is well-known that 2'-O-methylation and pseudouridylation differ in their chemical properties from their unmodified counterparts. These modifications can potentially impact various aspects of the RNA, including structure, thermal stability, and biochemical interactions.

For 2'-O-Me's several biophysical contributions to RNA were suggested: they may increase the stability of RNA conformations, alter the ability of the ribose to engage in hydrogen bonding, and may play a role in protecting the RNA from hydrolysis [87, 88]. Similarly, pseudouridylation seems to make the RNA more stable, alter/stabilize RNA conformation, and the base presents an extra hydrogen bond donor at the non-Watson-Crick edge that may potentially alter the pairing of pseudouridine with other bases [89].

1.1.4.3 Biogenesis of snoRNAs

SnoRNAs are typically generated from the introns of other, host genes. Once the intron is spliced out from the pre-mRNA and disbranched, the snoRNAs are processed by exonucleolytic trimming. Since vertebrate introns are rapidly degraded immediately upon co-transcriptional removal, there are mechanisms in place to protect the snoRNA. This is achieved by binding of the box C/D and H/ACA snoRNP proteins to the intron-embedded snoRNA sequences. This requires the snoRNP proteins to be actively recruited to the nascent intronic snoRNAs during the synthesis or before splicing of the host pre-mRNA [90]. Occasionally, the spliced exons of the pre-mRNA are devoid of open reading frames, indicating that the sole function of the transcript may be the expression of the snoRNA [91]. As previously mentioned, the 5' and 3' ends of the snoRNA form a specific structure called the kink-turn motif, which confers stability to the snoRNA and is important for proper snoRNA biogenesis.

To study snoRNA processing in more detail, we performed CLIP experiments on snoRNP components and extensively analyzed the 5' end 3' ends of the snoRNAs. Our findings are summarized in CHAPTER 2.

1.1.4.4 Alternative roles of snoRNAs

Other than exerting their canonical functions that are 2'-O-methylation and pseudouridylation, snoRNAs are suspected to have alternative roles. For instance the snoRNAs of the SNORD115 cluster have been implicated in the modulation of alternative splicing [92] and RNA editing [93]. Another interesting observation is the frequent generation of small, processed snoRNA fragments that, in isolated cases, have been demonstrated to exhibit miRNA-like features by being actively loaded into RISC and down-regulating their specific targets (reviewed in [94]). In **Chapter 2** we also explore the possibility of a miRNA-like behavior of snoRNAs in human cells. To this end we have sequenced and analyzed Ago2-associated populations of small RNAs in HeLa cells.

1.1.4.5 SnoRNA target identification

Assigning snoRNAs to their target site has always presented a challenge. This is reflected in the snoRNA database [95], where several known rRNA modifications remain without any guide snoRNA. Conversely, many so-called orphan snoRNAs remain without a known target. As previously mentioned, guide/target identification has recently been facilitated through the observation that chimeras form between a guide RNA and its target in CLIP experiments. At the same time, it has become possible to map 2'-O-methylations[65] and pseudouridines[64] in high throughput. In **Chapter 3** we demonstrate an integrated approach that combines two powerful experimental methods (snoRNP CLIP-seq and RiboMeth-seq) with computational modelling to map 2'-O-methylation sites in human rRNA and to assign them the C/D box guide snoRNAs.

In **Chapter 4** we present a new high-throughput variant of the classical reverse-transcriptase-based method for identifying individual 2'-O-methyl modifications in RNAs that we termed RiM-seq. RiM-seq presents an additional high-throughput method to validate 2'-O-Me's.

1.1.4.6 SnoRNA and disease

SnoRNAs have been implicated in several diseases. It has been known that defects in ribosome maturation and function can cause the disruption of vital processes and lead to diseases and transformation of healthy cells into cancer cells [96]. Therefore, it is plausible that snoRNA expression levels can affect the physiological conditions of cells and tissues, since they are involved in the regulation of post-transcriptional modification of rRNAs. Thus, snoRNA expression levels may be changed in disease or the change of snoRNA expression levels itself may influence emergence and progression of disease. Several snoRNAs have been shown to be increased or decreased in various cancers, suggesting that snoRNAs may exhibit oncogenic or tumor suppressor properties (reviewed in [97]). Further, the expression of snoRNAs seems to be perturbed in several other conditions such as in human cells during the antiviral response or in mammalian cells subjected to stress or drugs, although their role in these responses has not been clearly established yet [97].

A very well-described disorder where snoRNAs are thought to play an important role is the neurodevelopmental disease Prader Willi Syndrome (PWS). The lack of paternal expression of maternally imprinted C/D box snoRNAs is believed to be the main cause of this disorder [98, 99]. In **Chapter 5** we attempted to study these snoRNAs in more detail using a knockout mouse model void of these snoRNAs, mimicking the situation found in human PWS patients.

Chapter 2 Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing

Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing (Published in Genome Biology)

Shivendra Kishore[1†], Andreas R. Gruber[1,2,†], Dominik J. Jedlinski[1], Afzal P. Syed[1], Hadi Jorjani[1,2], and Mihaela Zavolan[1,2,*]
Computational and Systems Biology, Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

[1]Computational and Systems Biology, Biozentrum, University of Basel
[2]Swiss Institute of Bioinformatics

† Contributed equally
* To whom correspondence should be addressed. Tel: +41 61 267 1577; Fax: +41 61 267 1584; Email: mihaela.zavolan@unibas.ch

2.1 Abstract

Background
In recent years, a variety of small RNAs derived from other RNAs with well-known functions such as tRNAs and snoRNAs, have been identified. The functional relevance of these RNAs is largely unknown. To gain insight into the complexity of snoRNA processing and the functional relevance of snoRNA-derived small RNAs, we sequence long and short RNAs, small RNAs that co-precipitate with the Argonaute 2 protein and RNA fragments obtained in photoreactive nucleotide-enhanced crosslinking and immunoprecipitation (PAR-CLIP) of core snoRNA-associated proteins.

Results
Analysis of these data sets reveals that many loci in the human genome reproducibly give rise to C/D box-like snoRNAs, whose expression and evolutionary conservation are typically less pronounced relative to the snoRNAs that are currently cataloged. We further find that virtually all C/D box snoRNAs are specifically processed inside the regions of terminal complementarity, retaining in the mature form only 4-5 nucleotides upstream of the C box and 2-5 nucleotides downstream of the D box. Sequencing of the total and Argonaute 2-associated populations of small RNAs reveals that despite their cellular abundance, C/D box-derived small RNAs are not efficiently incorporated into the Ago2 protein.

Conclusions
We conclude that the human genome encodes a large number of snoRNAs that are processed along the canonical pathway and expressed at relatively low levels. Generation of snoRNA-derived processing products with alternative, particularly miRNA-like, functions appears to be uncommon.

2.2 Background

Small nucleolar RNAs (snoRNAs) are a specific class of small non-protein coding RNAs that are best known for their function as guides of modifications (2'-O-methylation and pseudouridylation) of other non-protein coding RNAs such as ribosomal, small nuclear and transfer RNAs (rRNAs, snRNAs and tRNAs, respectively) [83, 100, 101]. Based on sequence and structural features, snoRNAs are divided into two classes. C/D box snoRNAs share the consensus C (RUGAUGA, R = A or G) and D (CUGA) box motifs, which are brought into close proximity by short regions of complementarity between the snoRNA 5' and 3' ends [102, 103] and are bound by the four core proteins of the small ribonucleoprotein complex (snoRNP), namely 15.5K, NOP56, NOP58 and Fibrillarin (FBL) [74, 104, 105] during snoRNA maturation. Fibrillarin is the methyltransferase that catalyzes the 2'-O-methylation of the ribose in target RNAs [85]. Most C/D box snoRNAs also contain additional conserved C' and D' motifs located in the central region of the snoRNA. The other class of snoRNAs is defined by a double-hairpin structure with two single-stranded H (ANANNA, N = A, C, G or U) and ACA box domains [106], and are therefore called H/ACA box snoRNAs. They associate with four conserved proteins, Dyskerin (DKC1), Nhp2, Nop10 and Gar1, to form snoRNPs that are functionally active in pseudouridylation. Although all four H/ACA proteins are necessary for efficient pseudouridylation [106], it is Dyskerin that provides the pseudouridine synthase activity [107]. While H/ACA and C/D box snoRNAs accumulate in the nucleolus, some snoRNAs reside in the nucleoplasmic Cajal bodies (CBs) where they guide modifications of snRNAs [100] and are called small Cajal body-specific RNAs (scaRNAs). In addition to the typical H/ACA snoRNA features, vertebrate H/ACA box scaRNAs carry a CB localization signal called CAB box (UGAG) in the loop of their 5' and/or 3' hairpins [108].

Immediately upstream of the D box and/or the D' box, C/D box snoRNAs contain 10 to 21 nucleotide-long antisense elements that are complementary to sites in their target RNAs [109-111]. The nucleotide in the target RNA that is complementary to the fifth nucleotide upstream from the D/D' box of the snoRNA is targeted for 2'-O-methylation by the snoRNP [110, 111]. H/ACA box snoRNAs contain two antisense elements termed pseudouridylation pockets, located in the 5' and 3' hairpin domains of the snoRNA [112, 113]. Substrate uridines are selected through base-pairing interactions between the pseudouridylation pocket and target RNA sequences that flank the targeted uridine.

Deep-sequencing studies revealed a surprising diversity of small RNAs derived from non-coding RNAs (ncRNAs) known as small derived RNAs (sdRNAs) with well-established functions such as tRNAs [114, 115], Y RNAs [116], vault RNAs [117], ribosomal RNAs [118], spliceosomal RNAs [119] and snoRNAs [120-122]. In fact, the profile of sequenced reads observed for some of these small RNA species are very characteristic and have even been used for ncRNA gene finding based on sequencing data [123, 124]. The majority of C/D box and H/ACA snoRNAs seems to be extensively processed, producing stable small RNAs from the termini of the mature snoRNA [125] and the processing pattern is conserved across cell types [126]. Thus, it appears that snoRNAs are versatile molecules that give rise to snoRNA-derived miRNAs [120, 127], other small RNAs [121, 125] or longer processing fragments [128].

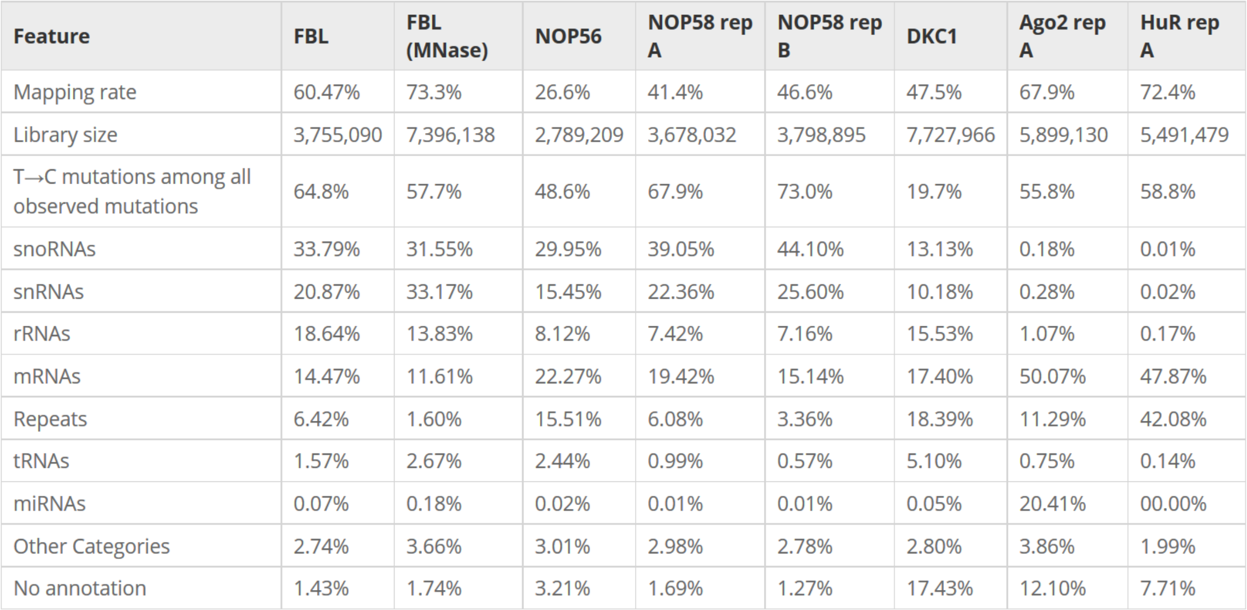
To gain insight into the complexity of snoRNA processing and the functional relevance of the derived sdRNAs, we undertook a comprehensive characterization of products generated from snoRNA loci, combining high-throughput sequencing of long and short RNA fragments with photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) of core snoRNA-associated proteins and with data from Argonaute 2 (Ago2) immunoprecipitation sequencing (IP-seq) experiments. We found that many loci in the human genome can give rise to C/D box-like snoRNAs. Among the novel snoRNAs that we identified are very short sequences, extending little beyond the C and D boxes, which are essential for the binding of core snoRNA proteins. Compared to the snoRNAs that are already known, the novel snoRNA candidates exhibit a lower level of evolutionary conservation and a lower expression level. These findings indicate that the C/D box snoRNA structure evolves relatively easily and that C/D box snoRNA-like molecules are produced from many more genomic loci than are currently annotated. We further found that C/D box snoRNAs are very specifically processed inside the regions of terminal complementarity, retaining in the mature form only four to five nucleotides upstream of the C box and two to five nucleotides downstream of the D box. Sequencing of the small RNA population as well as of the small RNAs isolated after Ago2 immunoprecipitation revealed that despite their cellular abundance, C/D box-derived small RNAs are not efficiently incorporated into the Ago2 protein. Our extensive data thus indicate that, contrary to previous suggestions [121, 129], snoRNA-derived small RNAs that carry out non-canonical, particularly miRNA-like, functions are rare.

2.3 Results

2.3.1 PAR-CLIP of C/D box and H/ACA box snoRNP core proteins identifies their RNA binding partners

To investigate the RNA population comprehensively that associates with both C/D box and H/ACA box small nucleolar ribonucleoproteins we performed PAR-CLIP as previously described [41] with antibodies against the endogenous Fibrillarin (FBL), NOP58 and Dyskerin (DKC1) proteins, in HEK293 cells (for details see Materials and methods). For NOP56 we used a stable cell line expressing FLAG-tagged NOP56 and anti-FLAG antibodies. Because we recently found that the choice of the ribonuclease and reaction conditions influences the set of binding sites obtained through cross-linking and immunoprecipitation (CLIP) [130, 131], we also generated a Fibrillarin PAR-CLIP library employing partial digestion with micrococcal nuclease (MNase) instead of RNase T1. PAR-CLIP libraries were sequenced on Illumina sequencers, mapped and annotated through the CLIPZ web server [131]. The obtained libraries were comparable to those from previous PAR-CLIP studies in terms of size, rates of mapping to genome and proportion of cross-link-indicative T→C mutations (TABLE 2.1). The DKC1 PAR-CLIP library shows a lower frequency of T→C mutations compared to all other libraries, but T→C mutations were still the most frequent in this library as well (data not shown).

Table 2.1 **Summary of CLIPZ mapping statistics and annotation categories for PAR-CLIP samples.**

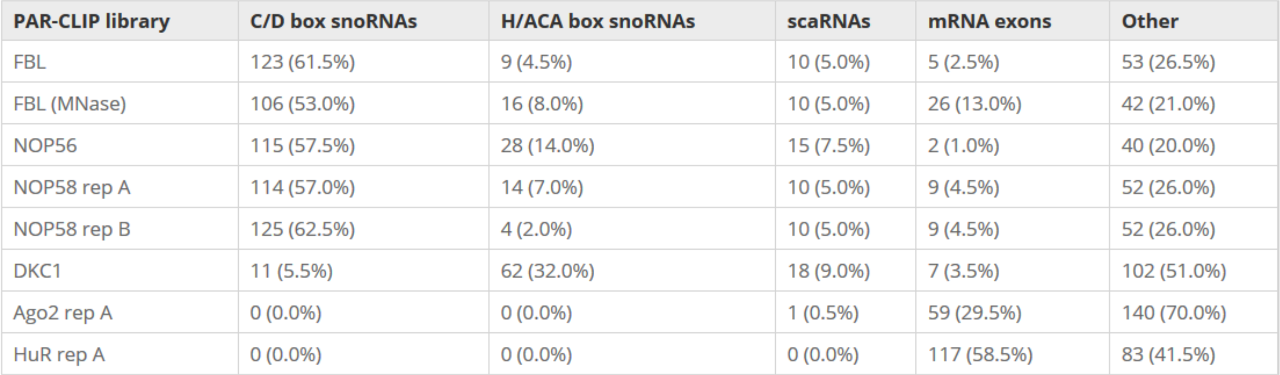


Feature	FBL	FBL (MNase)	NOP56	NOP58 rep A	NOP58 rep B	DKC1	Ago2 rep A	HuR rep A
Mapping rate	60.47%	73.3%	26.6%	41.4%	46.6%	47.5%	67.9%	72.4%
Library size	3,755,090	7,396,138	2,789,209	3,678,032	3,798,895	7,727,966	5,899,130	5,491,479
T→C mutations among all observed mutations	64.8%	57.7%	48.6%	67.9%	73.0%	19.7%	55.8%	58.8%
snoRNAs	33.79%	31.55%	29.95%	39.05%	44.10%	13.13%	0.18%	0.01%
snRNAs	20.87%	33.17%	15.45%	22.36%	25.60%	10.18%	0.28%	0.02%
rRNAs	18.64%	13.83%	8.12%	7.42%	7.16%	15.53%	1.07%	0.17%
mRNAs	14.47%	11.61%	22.27%	19.42%	15.14%	17.40%	50.07%	47.87%
Repeats	6.42%	1.60%	15.51%	6.08%	3.36%	18.39%	11.29%	42.08%
tRNAs	1.57%	2.67%	2.44%	0.99%	0.57%	5.10%	0.75%	0.14%
miRNAs	0.07%	0.18%	0.02%	0.01%	0.01%	0.05%	20.41%	00.00%
Other Categories	2.74%	3.66%	3.01%	2.98%	2.78%	2.80%	3.86%	1.99%
No annotation	1.43%	1.74%	3.21%	1.69%	1.27%	17.43%	12.10%	7.71%

Ago2: Argonaute 2; DKC1: Dyskerin; FBL: Fibrillarin; miRNA: micro RNA; MNase: micrococcal nuclease; PAR-CLIP: photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; rRNA: ribosomal RNA; snoRNA: small nucleolar RNA; snRNA: small nuclear RNA; tRNA: transfer RNA.

Compared to the libraries that we previously generated for HuR and Ago2 [130], two proteins whose primary targets are mRNAs, we found that snoRNAs, rRNAs and snRNAs were strongly enriched in PAR-CLIP libraries generated for the snoRNP core proteins (Table 2.1). The fact that not only snoRNAs but also the primary targets of snoRNAs, namely ribosomal RNAs and small nuclear RNAs, are enriched in these samples suggests that like Ago2 cross-linking, which captures both miRNAs and their targets [41, 130], cross-linking of core snoRNPs efficiently captures both snoRNAs and targets. To quantify the specificity of our PAR-CLIP libraries, we intersected the 200 clusters with the highest read density per nucleotide from each library with curated snoRNA gene annotations based on snoRNA-LBME-db [95] (Table 2.2). Currently, snoRNA-LBME-db lists about 153 human C/D box snoRNA loci and 108 human H/ACA box snoRNA loci that are known to be ubiquitously expressed. For each of the C/D box specific PAR-CLIP libraries, more than 100 of the top 200 clusters could be assigned to C/D box snoRNAs indicating the specificity of our CLIP experiments and the broad coverage of the snoRNA genes by the sequencing reads obtained from HEK293 cells. The Dyskerin PAR-CLIP data set showed a weaker enrichment in snoRNAs compared to the data sets for the core C/D box-specific proteins, with 57% of all known H/ACA box snoRNAs being represented among the 200 top-ranking clusters. scaRNAs were detected in both H/ACA box and C/D box specific libraries, as expected because many scaRNAs have both C/D box and H/ACA box elements. Finally, minor fractions of H/ACA box snoRNAs were also found in PAR-CLIP libraries of the C/D box-specific proteins, and *vice versa*. This could be caused by the close spatial arrangement of snoRNPs on the target molecule, or could indicate that H/ACA box snoRNAs and C/D box snoRNAs guide modifications on each other.

Table 2.2 **Annotation summary of the top 200 clusters inferred from PAR-CLIP experiments with snoRNA core proteins.**



PAR-CLIP library	C/D box snoRNAs	H/ACA box snoRNAs	scaRNAs	mRNA exons	Other
FBL	123 (61.5%)	9 (4.5%)	10 (5.0%)	5 (2.5%)	53 (26.5%)
FBL (MNase)	106 (53.0%)	16 (8.0%)	10 (5.0%)	26 (13.0%)	42 (21.0%)
NOP56	115 (57.5%)	28 (14.0%)	15 (7.5%)	2 (1.0%)	40 (20.0%)
NOP58 rep A	114 (57.0%)	14 (7.0%)	10 (5.0%)	9 (4.5%)	52 (26.0%)
NOP58 rep B	125 (62.5%)	4 (2.0%)	10 (5.0%)	9 (4.5%)	52 (26.0%)
DKC1	11 (5.5%)	62 (32.0%)	18 (9.0%)	7 (3.5%)	102 (51.0%)
Ago2 rep A	0 (0.0%)	0 (0.0%)	1 (0.5%)	59 (29.5%)	140 (70.0%)
HuR rep A	0 (0.0%)	0 (0.0%)	0 (0.0%)	117 (58.5%)	83 (41.5%)

Ago2: Argonaute 2; DKC1: Dyskerin; FBL: Fibrillarin; MNase: micrococcal nuclease; PAR-CLIP: photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; scaRNA: small Cajal body-specific RNA; snoRNA: small nucleolar RNA.

2.3.2 Binding patterns of core proteins on snoRNAs

As mentioned in the introduction, both C/D box and H/ACA box snoRNAs carry very specific functional sequence and structure elements, which are recognized by the snoRNP core proteins. We thus asked whether different C/D box core proteins have distinct preferences in binding different regions of the C/D box snoRNAs. FIGURE 2.1Adepicts PAR-CLIP read profiles along selected snoRNA genes (profiles for all scaRNA and snoRNA genes are in Additional file 1). Both C/D box core proteins as well as the H/ACA box specific Dyskerin bind to SCARNA6, which has a hybrid structure composed of both C/D box and H/ACA box elements. However, while the CLIP reads from the Fibrillarin, NOP56 and NOP58 samples cover the C and D box motifs, Dyskerin was preferentially cross-linked to the H-box motif and to the 5' end of the first H/ACA box stem. For the C/D box snoRNAs, different snoRNA core proteins gave very similar cross-linking patterns (FIGURE 2.1B), which we quantified through the correlation coefficient between read densities obtained along individual snoRNAs in pairs of samples. Comparing NOP58 to Fibrillarin and NOP56 we found that 109 (78%) and 111 (80%) snoRNA genes had a correlation coefficient of at least 0.9. To put this in perspective, between biological replicates of NOP58, 130 out of 139 snoRNAs investigated have a correlation coefficient of at least 0.9. This indicates that Fibrillarin, NOP56 and NOP58 form a tight complex that contacts the snoRNA. As noticed before, however [130], the nuclease treatment has a strong influence on the relative number of tags obtained from different positions along a snoRNA (FIGURE 2.1C). Only 19 snoRNA genes (14%) show a correlation ≥ 0.90 in their tag profiles obtained with RNase T1- and MNase-treated Fibrillarin PAR-CLIP samples, reflecting the fact that T1 nuclease is more efficient and generates a more biased position-dependent distribution of reads than MNase (FIGURE 2.1A). FIGURE 2.1D and FIGURE 2.1E summarize these results, showing that nucleotides in D' boxes are most frequently cross-linked, followed by nucleotides in the C' and C boxes, and then by nucleotides in the D box and in the rest of the snoRNA. MNase treatment in particular results in very poor coverage of the D box. On the other hand, we observed gene-specific differences in the binding of the core proteins. For example, SNORD20 only shows a peak of CLIP reads at the D box, SNORD30 only at the C box, while SNORD76 has peaks at both C and D boxes (FIGURE 2.1A).

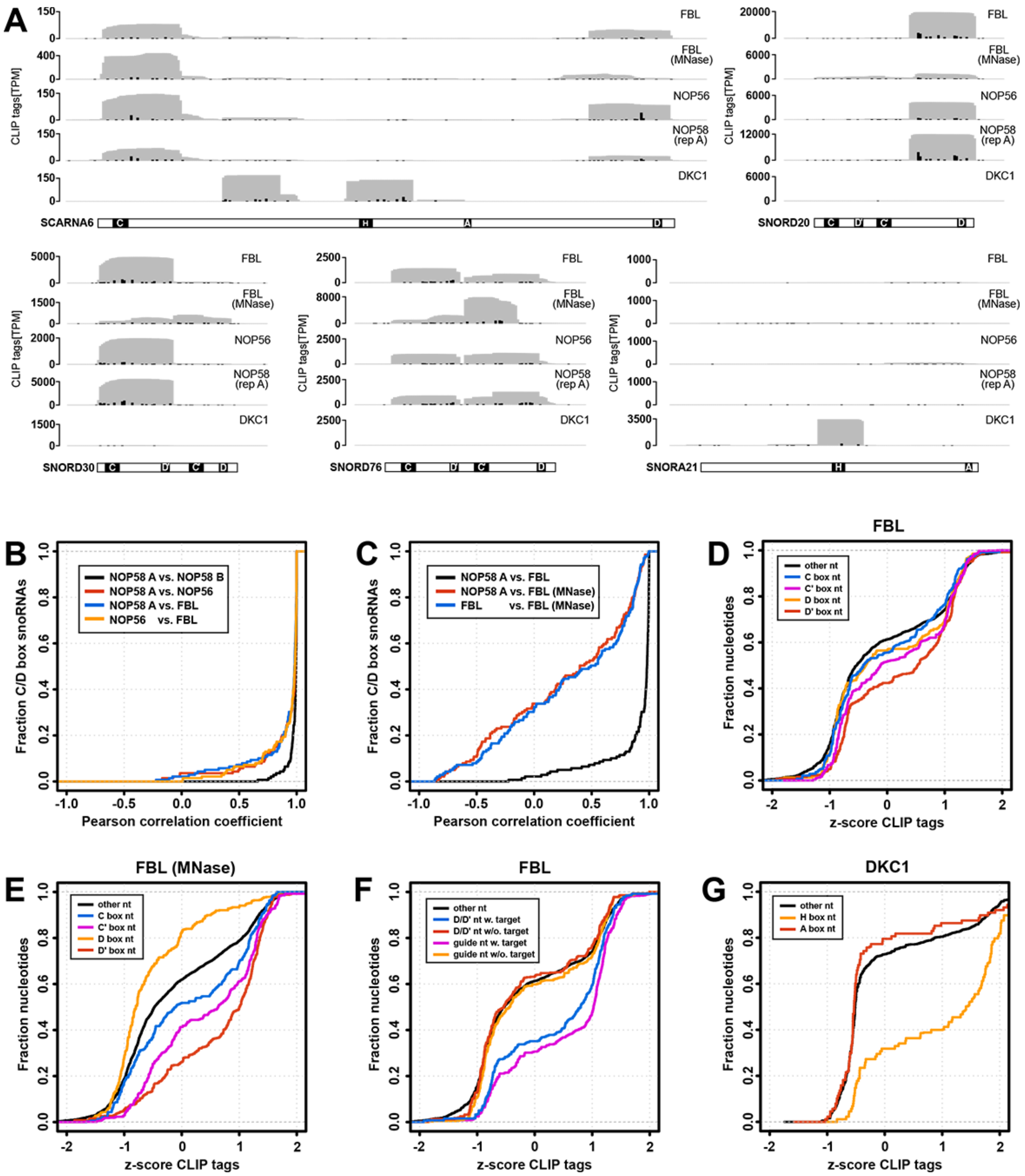


Figure 2.1 **Summary of PAR-CLIP data of snoRNP core proteins**. (A) Profiles of sequencing reads obtained from PAR-CLIP experiments for selected snoRNAs. Black bars in the profiles indicate the number of T→C mutations observed in PAR-CLIP reads at a particular nucleotide. (B) Similarity of binding profiles of core proteins that associate with C/D box snoRNAs. (C) Comparison of protein binding profiles as inferred from RNase T1-treated and MNase-treated PAR-CLIP samples. (D, E) Preferential binding of Fibrillarin to box elements as inferred from PAR-CLIP samples prepared with T1 (D) and MNase ribonucleases (E). (F) Comparison of binding preferences at D'/D box elements and guide regions for snoRNAs with and without a known target. (G) Analysis of binding preferences of Dyskerin for H/ACA box snoRNA-specific elements. D, E, F and G show the cumulative distributions of CLIP read coverage z-scores for nucleotides located in various regions of the snoRNA relative to the overall coverage of the snoRNA. CLIP: cross-linking and immunoprecipitation; MNase: micrococcal nuclease; PAR-CLIP: photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; snoRNA: small nucleolar RNA; snoRNP: small nucleolar ribonucleoprotein.

We further asked whether the binding pattern of Fibrillarin reflected in the abundance of CLIP reads differs between guide regions of the snoRNAs that have a target annotated in snoRNA-LBME-db and orphan guide regions. For guide regions, we took the nine nucleotides upstream of the D and D' boxes and as a reference we compared the coverage of the D and D' boxes themselves (FIGURE 2.1F). We found that guide regions with a known target and their associated D/D' boxes generally have a higher coverage compared to those that are orphan (70% compared to 40% positive *z*-scores of the average coverage per position in the guide region relative to the entire snoRNA, FIGURE 2.1G). This could indicate that the binding to the target renders the snoRNA-core protein complex more accessible to cross-linking.

For H/ACA box snoRNAs we found that Dyskerin strongly prefers the H box nucleotides (FIGURE 2.1G), which in 70% of the snoRNAs have a positive *z*-score for coverage compared to the entire snoRNA. This is expected because these snoRNAs are highly structured, with most nucleotides being engaged in base pairs in the two hairpin stems and a few nucleotides are free to interact with the proteins.

2.3.3 Identification of novel snoRNA genes from PAR-CLIP and small RNA sequencing

We screened the top 500 clusters from each PAR-CLIP library that did not overlap with known ncRNAs, mRNAs or repeat elements for potentially novel snoRNA genes. To identify H/ACA box genes we employed the SnoReport program [132], while for C/D box snoRNA detection we applied a custom approach searching for a C box motif (RUGAUGA, R = A or G; allowing one mismatch) at the 5' end and a D box motif (MUGA, M = A or C) at the 3' end, requiring that a terminal stem of at least four canonical base pairs can be formed by the nucleotides flanking the C and D boxes. We combined these computational screens with isolation and sequencing of the 20 to 200 nucleotide RNA fraction from HEK293 cells, which provides evidence for expression of the predicted snoRNAs. Requiring a minimal average coverage per nucleotide of at least 1 tag per million (TPM) in least one type-specific CLIP library as well as in the small RNA-seq library, we identified 77 and 20 putative C/D and H/ACA box snoRNAs, respectively (Additional files 2 and 3). We additionally screened 14 distinct small RNA sequence libraries from the recently released ENCODE data [133] and found that more than 75% of our putative C/D box snoRNAs were detected in at least one cell type other than HEK293 (see Additional file 4). We further tested the expression of the 20 most abundantly sequenced candidate snoRNAs by Northern blotting (see Additional file 5). Nine of the twenty candidates were also detectable in this assay, while an additional nine C/D box snoRNAs are supported by the ENCODE data (see Additional file 4).

To determine whether the candidates we identified as described are entirely novel snoRNA genes or so far undescribed homologs of known snoRNAs, we performed a BLAST search against the snoRNA genes from snoRNA-LBME-db (requiring an *E*-value ≤ 10^{-3}). We further compared the loci of the putative snoRNAs with the snoRNA annotation available in ENSEMBL release 65 [134], which is based on automatic annotation with sequence/structure models available in the Rfam database [135]. Out of the 20 H/ACA box snoRNA candidates, 18 show sequence or structural homology to known snoRNAs, while candidates ZL4 (annotated as nc053 in [136], but not classified as a snoRNA by the authors) and ZL36 appear to be novel H/ACA box snoRNAs without a known homolog. The homology search additionally revealed that ZL4 is conserved until *Xenopus tropicalis*.

Of the 77 C/D box snoRNAs, only seven showed sequence homology to known C/D box snoRNA genes, but in one case (ZL1) the homology consisted solely of a long GU-rich region. The evolutionary conservation of the guide regions of five of these snoRNAs (ZL11, ZL109, ZL126, ZL127 and ZL132) suggests that they target the same nucleotides on ribosomal RNA as their homologs. A sixth snoRNA, ZL142, appears to be a human homolog of the GGN68 snoRNA of chickens [137, 138]. An additional comparison with the results of another large snoRNA analysis [139], revealed that ZL2 and ZL107 have been previously described as SNORD41B and Z39, respectively. In order to further characterize the 69 potentially novel C/D box snoRNAs (including ZL1, which only had homology with a known snoRNA in a GU-rich region), we first asked whether their C and D boxes are evolutionarily conserved (Additional file 1). To this end, we computed their average position-wise phastCons scores [140], which we obtained from the UCSC genome browser. Five candidates including ZL1 showed an average phastCons score per nucleotide higher than 0.25 for C and D box nucleotides. A comprehensive homology search of vertebrate genomes allowed us to trace the evolutionary origin of these snoRNAs and to annotate C' and D' boxes as well as putative guide regions based on sequence conservation. ZL1 is highly conserved in vertebrates including *Petromyzon marinus*, while for ZL5, ZL6, ZL8 and ZL24 we were not able to retrieve any homologs outside of mammals.

The remaining C/D box snoRNAs show overall weak conservation in mammals and in primates (Additional file 1). The C' and D' box elements of these snoRNAs, which are typically more variable in sequence, were particularly difficult to annotate without supporting evidence from evolutionary conservation. Because it is not clear that these snoRNAs have a C-D'-C'-D box architecture, we refer to them as C/D box-like. The small RNA sequence data indicates that these C/D box-like snoRNAs are only weakly expressed (Additional file 6). Interestingly, while the shortest C/D box snoRNA that has been characterized so far is SNORD49B, which has 48 nucleotides, 23 of our C/D box-like snoRNAs are even shorter. FIGURE 2.2 depicts PAR-CLIP tags and small RNA-seq reads for four of these snoRNAs which we called mini-snoRNAs. ZL77 is among the shortest, with 27 nucleotides in length, and only 7 nucleotides available as a potential guide region between the C and D boxes, while ZL49 and ZL103 are slightly longer (14 and 15 nucleotides between the C and D boxes). Another mini-snoRNA, ZL63, generated a considerable number of reads in all the CLIP libraries as well as in the RNA sequence data.

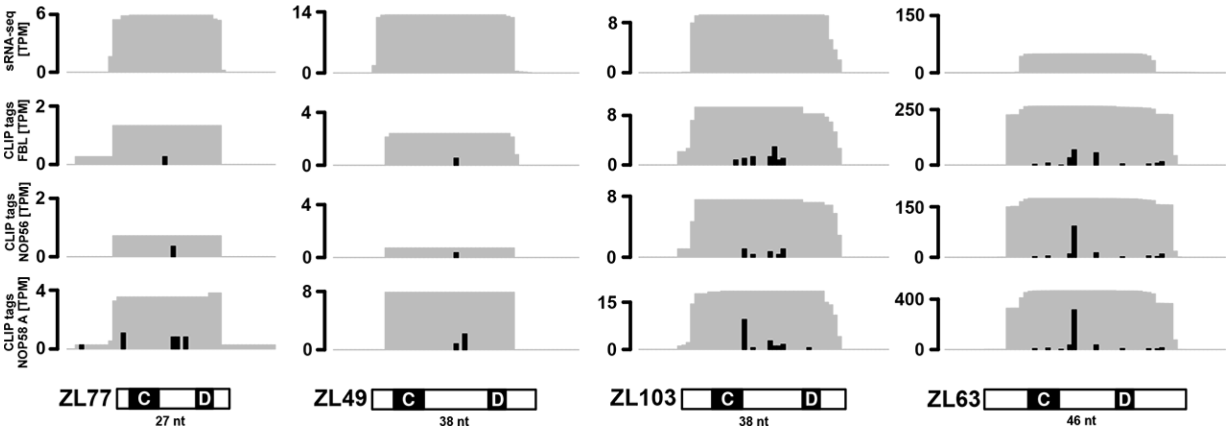


Figure 2.2 **Small RNA-seq and PAR-CLIP reads mapping to mini-snoRNAs**. Mini-snoRNAs ZL77, ZL49, ZL103 and ZL63 are shown. Black bars in the panels corresponding to PAR-CLIP libraries indicate the number of T→C mutations observed at individual nucleotides. CLIP: cross-linking and immunoprecipitation; PAR-CLIP: photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; snoRNA: small nucleolar RNA.

Our screen could further identify a snoRNA with mixed C/D box and H/ACA box structure. SCARNA21, a computationally predicted H/ACA box snoRNA [141], is surrounded by conserved C and D box elements enclosed by a terminal stem structure (Additional file 7). Northern blot analysis revealed that the prevalent form in the cells is the one that contains the C/D box elements and not the short form, which would be the single H/ACA box snoRNA.

2.3.4 Target prediction for newly identified snoRNA genes

To gain insight into the function of the novel snoRNAs that we identified, we sought to determine whether they have canonical targets. We employed the programs RNAsnoop and PLEXY to predict targets of H/ACA box and C/D box snoRNAs, respectively [142, 143]. As potential target sequences we considered ribosomal and spliceosomal RNAs obtained from snoRNA-LBME-db. Indeed, for the highly conserved C/D box snoRNAs ZL1, ZL5 and ZL6 (which share the guide region), as well as for the H/ACA box snoRNA ZL4, we could identify canonical targets (FIGURE 2.3). ZL1 and ZL4 are both predicted to guide modifications on the U2 snRNA, 2'-O-methylation of U47 and pseudouridylation of U15, respectively. The pseudouridylation of U2 snRNA at U15 has already been described, but the guiding snoRNA was not known [67]. With primer extension assays we could further validate the U47 modification (see Additional File 8). SnRNA modifications are known to occur in Cajal bodies. Consistent with ZL4 H/ACA box snoRNA being a scaRNA that is recruited to Cajal bodies, is the presence of the CAB box motif (UGAG), known to mediate this transport [108], in the hairpin loops. For the C/D box snoRNA ZL1 targeting U2 snRNA we could not identify an H/ACA box-like structural domain with a CAB box. Interestingly, however, this snoRNA candidate contains a long GU repeat, a feature shared by SCARNA9, the only Cajal body-associated snoRNA that lacks H/ACA and CAB boxes. This suggests that the GU element serves as an import signal into Cajal bodies. For ZL5/6, the predicted modification site on the 28S rRNA is in fact a known modification site for which the guide was so far unknown. We could not predict a target for the newly identified C/D box domain of SCARNA21.

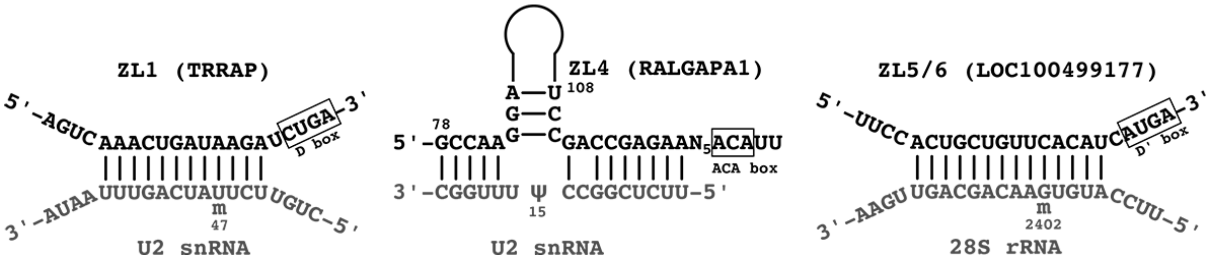

ZL1 (TRRAP)
5'-AGUC AAACUGAUAAGAU CUGA -3'
D box
3'-AUAA UUUGACUAUUCU UGUC-5'
m
47
U2 snRNA
ZL4 (RALGAPA1)
A—U 108
G—C
78
5'-GCCAA G—C GACCGAGAAN5 ACA UU
ACA box
3'-CGGUUU Ψ CCGGCUCUU-5'
15
U2 snRNA
ZL5/6 (LOC100499177)
5'-UUCC ACUGCUGUUCACAUC AUGA -3'
D' box
3'-AAGU UGACGACAAGUGUA CCUU-5'
m
2402
28S rRNA

Figure 2.3 **Predicted structure of hybrids between novel snoRNAs and target RNAs**. The snoRNAs are given at the top of each panel together with the symbol of the host gene in which the snoRNA resides (in parentheses). The targets are indicated at the bottom of the panels. rRNA: ribosomal RNA; snoRNA: small nucleolar RNA; snRNA: small nuclear RNA

We were especially interested to find out whether the non-conserved C/D box-like snoRNAs and in particular the mini-snoRNAs, could guide 2'-O-methylations. To this end, we took a simple approach searching for 8-mer Watson-Crick complementarity between the putative guide regions upstream of the D boxes to ribosomal and spliceosomal RNAs. We did indeed identify seven putative interaction sites, but none of these are known modification sites (Additional file 2). Thus, the targets of these C/D box-like snoRNAs remain to be identified.

2.3.5 Non-canonical RNA partners of core snoRNA proteins

Although snoRNAs are best known for guiding modifications of rRNAs, snRNAs and tRNAs [83, 100, 101], some evidence has emerged for the involvement of full-length mature snoRNAs also in other biological processes such as alternative splicing [92]. To investigate this possibility, we searched our PAR-CLIP data sets for RNAs that were abundantly cross-linked, yet not known to associate with the core snoRNA proteins. In contrast to the HuR PAR-CLIP that we performed before [130], the PAR-CLIP experiments conducted with C/D box snoRNP core proteins repeatedly identified several non-coding RNAs including vault RNA 1-2, 7SK RNA and 7SL RNA as well as H/ACA box snoRNAs. Similarly, in the Dyskerin PAR-CLIP we observed cross-linking of several C/D box snoRNAs.

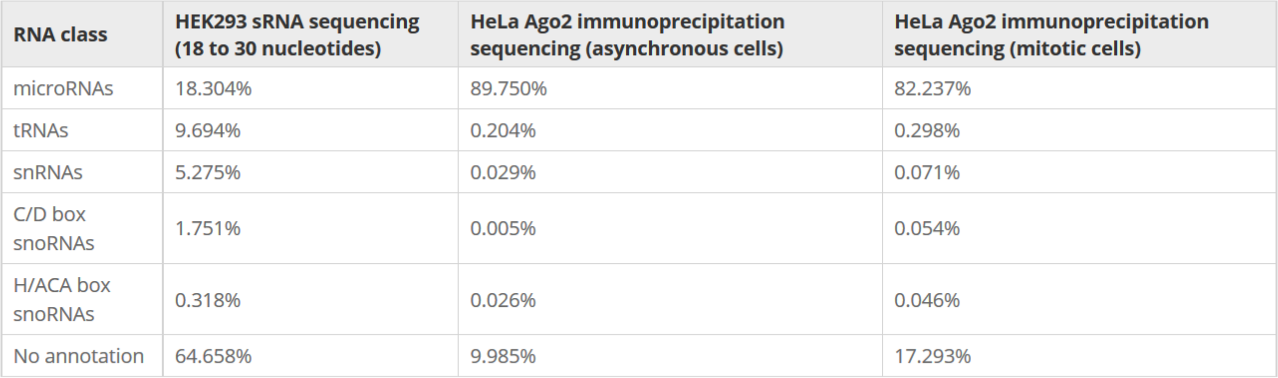
We performed primer extension experiments to determine potential sites for 2'-O-methyl and pseudouridine modification in prominent ncRNAs such as 7SK RNA, 7SL RNA and vault RNA 1-2 (see Additional file 9 for primer extension assays and Additional file 10 for a catalog of identified modifications sites and target predictions). Indeed, we found that all three of these RNA species carry modifications. Vault RNA 1-2 contains four 2'-O-methyl sites, 7SK RNA carries at least six 2'-O-methyl sites and one pseudouridylation site, and 7SL RNA contains several sites of pseudouridylation. Additionally, we sought to determine whether C/D box and H/ACA box snoRNAs guide modifications on each other. We thus performed 2'-O-methylation primer extension assays on SNORA61 and pseudouridylation assays on SNORD16 and SNORD35A. We found that SNORA61 potentially carries one 2'-O-methylation, while SNORD16 and SNORD35A carry two and six pseudouridylated residues, respectively. To identify C/D box snoRNAs that could guide the observed 2'-O-methylations, we searched for 8-mer complementarity upstream of D and D' boxes of C/D box and C/D box-like snoRNAs, but we did not find sequences complementary to the modification sites. To predict guiding H/ACA box snoRNAs we employed the program RNAsnoop using stringent filtering criteria. We identified potential guiding H/ACA box snoRNAs for 7SK RNA residue Ψ250 and 7SL RNA residue Ψ226.

Previous studies reported that snoRNAs may function in alternative splicing [92, 128] and we also repeatedly observed cross-linking of C/D box core proteins to regions that are annotated as exons of protein coding genes. To determine whether these mRNA regions are targeted by snoRNAs, we selected, from the top 1,000 clusters located in mRNA exons in NOP58 libraries, the 157 that were present in both NOP58 replicates and a third CLIP library with at least 10 TPM per nucleotide (Additional file 11). We identified complementarities to the 8-mer guide regions of snoRNAs in 79 of these clusters. In contrast, in shuffled CLIPed regions we only found 60 complementarities to snoRNA guide regions (average of 100 simulations on shuffled sequences). Thus, the mRNA sequences that we isolated in the CLIP experiments are consistent with the possibility that snoRNAs act as guides in some steps of mRNA processing.

2.3.6 snoRNA processing patterns

It has become apparent that many ncRNAs such as tRNAs, snRNAs, rRNAs and snoRNAs are extensively processed into small, stable RNA fragments originating mainly from the termini of the mature ncRNA [125], which in some cases are incorporated in the Argonaute proteins to function as microRNAs [120]. To identify snoRNA-derived small RNAs that could potentially act as miRNAs comprehensively, we isolated and sequenced the RNA fraction of 18 to 30 nucleotides from HEK293 cells. Small RNAs derived from C/D box snoRNAs constitute about 1.7% of the small RNA pool in this size range in HEK293 cells (TABLE 2.3). Consistent with the results of Li and colleagues [125], we found that most of the 513,339 reads overlapping with C/D box snoRNA genes originate from the 5' or 3' ends (38.7% and 46.0%, respectively). Visual inspection of the alignment of these reads to the snoRNAs revealed, however, that start and end positions of the reads do not generally coincide with the annotated snoRNA termini, which were inferred based on the characteristic C/D box snoRNA terminal stem (FIGURE 2.4A). Instead, the reads that we obtained indicate specific trimming that generates sharp 5' ends for 5'-end-derived reads and sharp 3' ends for 3'-end-derived reads. To determine whether this trimming may occur in the process of generating small RNAs from mature C/D box snoRNAs, we isolated small RNAs of length 20 to 200 nucleotides that presumably included the full-length, mature snoRNAs (average C/D box snoRNA length is 70 to 90 nucleotides) and performed a 150-cycle sequencing run. FIGURE 2.4A depicts the alignment of reads obtained in the small RNA fraction and the reads obtained in the 150-cycle sequencing run for three selected C/D box snoRNAs. Strikingly, the sharp ends of C/D box snoRNA-derived small RNAs coincide with the 5' and 3' ends of the mature form. More generally, we found that for 84% and 70% of the top 50 expressed C/D box snoRNAs, the most prominent start and end positions, respectively, obtained from long sequencing reads coincided with the most prominent start and end positions obtained from small RNA sequencing. This suggests that the observed trimming of the terminal closing stem occurs during the excision of the snoRNA from the intron and is not specific to the processing of the mature snoRNA form into smaller fragments. Furthermore, we found that it is the distance to the C or D boxes that seems to determine the observed ends of the snoRNAs rather than the length of the terminal closing stem (FIGURE 2.4B). The 5' end is sharply defined four to five nucleotides upstream of the C box, while the 3' end is more variably located two to five nucleotides downstream of the D box. In most cases this will leave mature C/D box snoRNAs with a terminal 5' overhang compared to the 3' end. This suggests that, similar to other small RNAs [127, 144, 145], snoRNAs are trimmed presumably by exonucleases, to boundaries that are determined by the proteins with which these small RNAs are complexed.

Table 2.3 **Functional annotation of sequencing reads obtained in sRNA sequencing and HeLa Ago2 IP sequencing.**



RNA class	HEK293 sRNA sequencing (18 to 30 nucleotides)	HeLa Ago2 immunoprecipitation sequencing (asynchronous cells)	HeLa Ago2 immunoprecipitation sequencing (mitotic cells)
microRNAs	18.304%	89.750%	82.237%
tRNAs	9.694%	0.204%	0.298%
snRNAs	5.275%	0.029%	0.071%
C/D box snoRNAs	1.751%	0.005%	0.054%
H/ACA box snoRNAs	0.318%	0.026%	0.046%
No annotation	64.658%	9.985%	17.293%

Ago2: Argonaute 2; IP: immunoprecipitation; snRNA: small nuclear RNA; snoRNA: small nucleolar RNA; sRNA: small RNA; tRNA: transfer RNA.

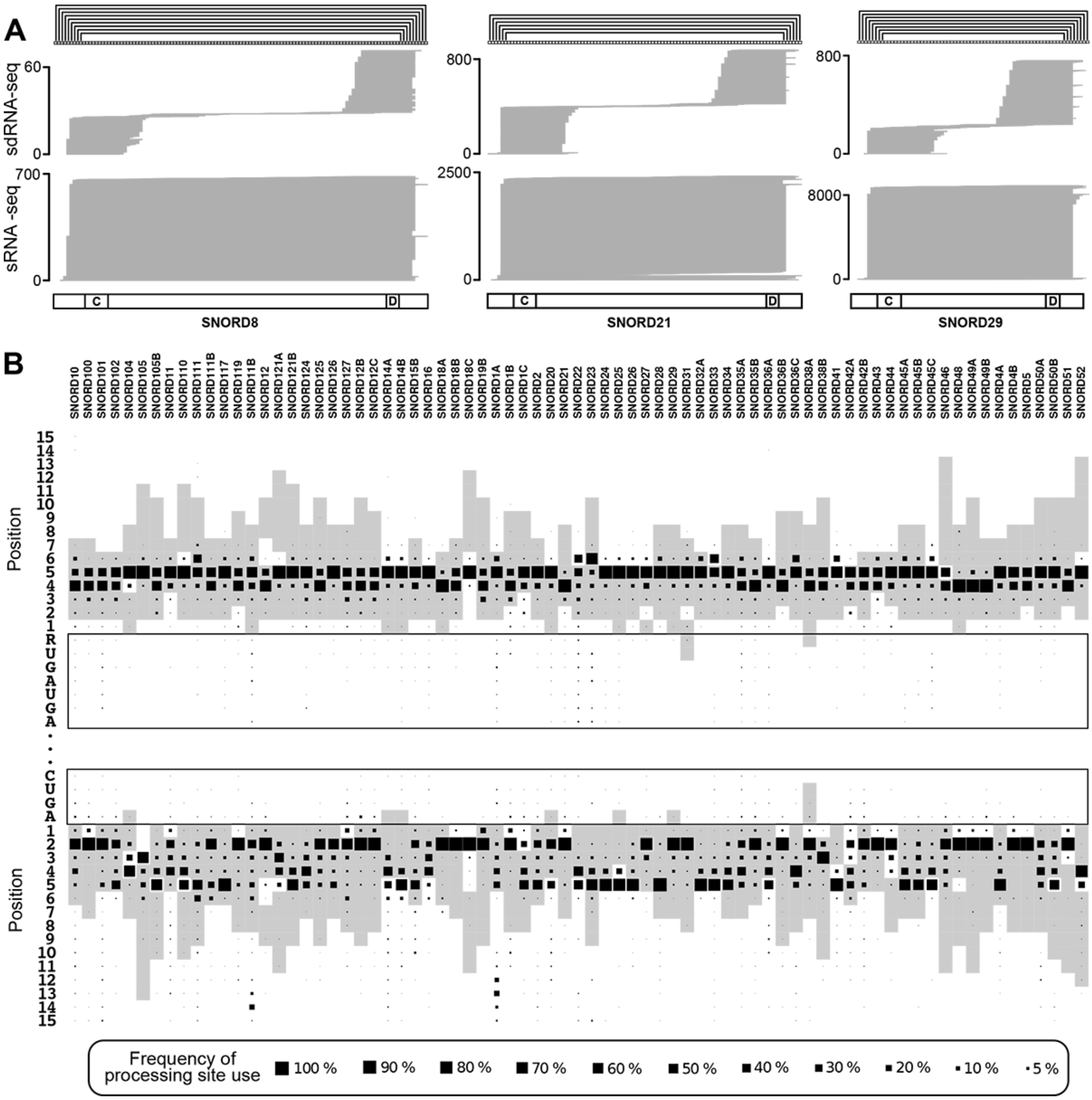


Figure 2.4 **Terminal processing of C/D box snoRNAs**. (A) Profiles of sequencing reads obtained from two small RNA seq libraries for three selected C/D box snoRNAs (SNORD8, SNORD21 and SNORD29). Upper: sdRNA sequencing, 18 to 30 nucleotides. Lower: sRNA sequencing, 20 to 200 nucleotides. Secondary structure annotation of the terminal closing stem is given on the top of the figure, while the locations of C and D motifs are shown on the bottom. (B) Detailed analysis of terminal stem processing for C/D box snoRNA expressed in HEK293 cells. The y-axis indicates individual nucleotides, with their specific identity for the nucleotides in C/D boxes and position relative to the boxes for the flanking nucleotides. Each column corresponds to a snoRNA, whose identity is shown at the top of the panel. Grey boxes indicate nucleotides that are predicted to be paired in the terminal stem. The size of black boxes is proportional to the number of sRNA sequencing reads that start (5' end) or end (3' end) at a particular nucleotide. See Additional File 16 for analysis of all C/D box snoRNAs expressed in HEK293 cells. sdRNA: small derived RNA; snoRNA: small nucleolar RNA; sRNA: small RNA.

Small RNAs derived from C/D box snoRNA termini appear to be abundant in the cells, and can be incorporated into Argonaute proteins to act as miRNAs [127]. To determine the relative participation of various small RNA classes in the Argonaute-dependent gene silencing, we immunopurified Ago2 from HeLa cells and sequenced the associated small RNA fraction. We found that, as expected, miRNAs constitute the most abundant RNA class that associates with Ago2 (approximately 90%), while C/D box snoRNAs account only for 0.005% of the IP-seq reads (TABLE 2.3). Assuming that overall proportions of small RNAs derived from tRNAs and snoRNAs are fairly constant across cell types, we can estimate the efficiency with which small RNAs (from the total small RNA pool) are incorporated in the Argonaute proteins. We found, for example, that although small RNAs derived from tRNAs are 5.6 times more abundant than C/D box derived snoRNAs, tRNA fragments are 40 times more abundant in the Ago2-associated fraction. Thus, tRNA-derived small RNAs appear to be more efficiently incorporated in Ago2 than C/D box snoRNA fragments. This is consistent with observations that tRNAs are cleaved by nucleases such as Angiogenin and even Dicer to generate processing fragments that are active in translation regulation [146, 147]. Similarly, small RNAs derived from H/ACA box snoRNAs are 5.5 times less abundant than small RNAs derived from C/D box snoRNAs in the total RNA fraction, but are 5.2 times more efficiently picked up by Ago2. The H/ACA box snoRNA SCARNA15, which has been shown to be processed into smaller fragments that act as microRNAs [120], is represented in this library with 3,636 reads, 29% of all reads mapped to H/ACA box snoRNA loci (see Additional file 12 for a full listing of all snoRNAs). The C/D box snoRNA with the highest number of reads in the Ago2 IP library is SNORD1A with 1,140 reads, but the majority of C/D box snoRNAs are represented by less than 50 reads.

Of all categories of small RNAs, C/D box snoRNA fragments are those that show the strongest nuclear retention, and are found in the cytoplasm with only low frequency [148]. Thus, this physical separation could account for the low frequency of association between C/D box snoRNA-derived RNAs and Ago2. We therefore wondered whether the association of this abundant class of RNA fragments with Ago2 increases in the mitotic phase of the cell cycle, when the nuclear membrane is dissolved. We collected HeLa cells that were in the mitotic phase through mitotic shake off, immunopurified Ago2 and again sequenced the Ago2-associated small RNA fraction. We found that, indeed, the relative abundance of C/D box-derived fragments in Argonaute increased in this condition (TABLE 2.3), to 0.054% relative to 0.005%. Nonetheless, these results indicate that C/D box snoRNAs do not generally carry out miRNA-like functions, and that the number of H/ACA box snoRNAs with a dual function is very limited.

2.4 Discussion

To gain insight into the processing of snoRNAs and the functions of snoRNA-derived small RNAs, we performed PAR-CLIP experiments with snoRNP core proteins. Analysis of PAR-CLIP reads showed that C/D box core proteins Fibrillarin, NOP56 and NOP58 have a very similar binding pattern, overlapping with the box elements. Excluding snoRNA families SNORD113 to SNORD116, which are multi-copy families and do not have guide complementarity to rRNAs or snRNAs, snoRNA-LBME-db currently lists 153 C/D box snoRNAs, of which 40 and 78 have a guide region targeting a known modification at the D box and D' box, respectively. Evolutionary conservation profiles of the remaining putative guide regions suggest that most of them are not functional. In support of this concept, our analysis revealed that C/D box core proteins cross-linked more effectively to guide regions that are known to have a target compared to orphan guide regions.

Combining computational prediction with data from small RNA sequencing and PAR-CLIP we identified novel C/D and H/ACA box snoRNAs, and assigned guiding snoRNAs to several modifications on rRNAs and snRNAs that were previously described as orphans. In addition to these *bona fide* snoRNAs, we uncovered a group of C/D box-like snoRNAs that only have a C and a D box as opposed to the common C-D'-C'-D architecture. These C/D box-like snoRNAs are only weakly conserved and most of them are expressed at low levels. The unusual architecture and the weak evolutionary conservation are likely reasons why these RNA species have not been uncovered by computational ncRNA gene finders [149]. Some of the identified C/D box-like snoRNAs are extremely short, one being only 27 nucleotides in length, leaving hardly enough space for a guide region. The requirements for C/D box snoRNA biogenesis appear to be simply the presence of C and D boxes and a short region of complementarity flanking these boxes, leading probably to the production of many snoRNA-like molecules as the C/D box core proteins scan intronic regions of pre-mRNAs. An interesting lead to follow in further investigating the potential function of the C/D box-like snoRNAs originating in the introns of many genes comes from a recent study conducted in *Drosophila*, in which Schubert and colleagues showed that snoRNAs are required for maintenance of higher-order structures of chromatin accessibility [150].

In our PAR-CLIP experiments we also repeatedly cross-linked ncRNAs that are not usual snoRNA targets. We observed H/ACA box snoRNAs in PAR-CLIP experiments targeting the C/D box core proteins. *Vice versa*, we found C/D box snoRNAs in the PAR-CLIP targeting Dyskerin, which is an essential component of H/ACA box snoRNPs. Primer extension assays indicated that these snoRNAs carry modifications that would be expected from the protein complexes to which they were cross-linked, but we were, in general, not able to identify snoRNAs that could guide these modifications. One drawback may be that in the case of the 2'-O-methyl primer extension assays we cannot be sure that it was indeed a 2'-O-methyl modification as opposed to any other nucleoside modification that caused the stoppage of the reverse transcriptase. However, we can be fairly certain that we identified *bona fide* pseudouridylation sites. Particularly, in the case of SNORD35A we were able to identify five putative pseudouridylated residues but no convincing guiding sequence in a known H/ACA box snoRNA. This suggests either that even more snoRNAs remain to be identified or that these pseudouridylations are caused by a protein-only mechanism not requiring guidance by H/ACA box snoRNAs.

The processing patterns of snoRNAs have raised substantial interest and some controversy in recent years [94, 126, 128]. We strikingly found that snoRNA excision out of the intron follows a well-defined pattern leaving mature snoRNAs with four to five nucleotides upstream of the C box, and two to five nucleotides downstream of the D box, irrespective of the length of the terminal closing stem. Our data support the observations of Darzacq and Kiss [103] that the terminal stem serves to bring the C and D box elements into close proximity so as to be more easily recognized by snoRNP proteins, which then protect the snoRNA from further trimming by the exosome, but may not be needed for the functional, mature snoRNA. This implies that the core proteins actively protect and stabilize the maturing snoRNA.

We further quantified the abundance of snoRNA-derived small RNAs in HEK293 cells, and consistent with other studies [125], we found that small RNAs derived from the ends of C/D box snoRNAs are indeed abundant. However, we did not find evidence that these sdRNAs efficiently associate with Ago2 to act as microRNAs, even in conditions when the accessibility of these sdRNAs to Ago2 should be higher, such as in mitotic cells. We thus conclude that a microRNA-like function of snoRNA-derived small RNAs is an exception rather than a rule. Most of the sdRNAs from C/D box snoRNAs originate from the termini of mature snoRNAs, and hence carry C and D box motifs. It might be that snoRNA core proteins are still attached to these fragments, protect them from total degradation, sequester them in the nucleus and prevent these sdRNAs from being loaded into Ago2.

Deep-sequencing-based studies revealed a very complex landscape of transcription and processing of RNAs. The non-canonical products identified initially in such studies raises the question of additional, yet unknown, functions of molecules that have been studied for many years. What has become apparent more recently, however, is that deep sequencing allows us to construct a very detailed picture of the kinetics of processing various classes of RNAs and of their interactions with proteins that protect them from degradation. Intersection of many data sets such as those generated in our study will eventually reveal kinetic and regulatory aspects of cellular processes at a fine level of detail.

2.5 Materials and methods

2.5.1 PAR-CLIP experiments

PAR-CLIP was performed with HEK293 Flp-In cells (Invitrogen). Cells were grown in thirty 15-cm cell culture plates per experiment to approximately 80% confluency. At 12 h before harvest, 4-thiouridine (Sigma) was added to the cells to a final concentration of 100 µM. PAR-CLIP was carried out as described previously [41]. For immunoprecipitation, antibodies were coupled to protein-A or protein-G Dynabeads (Invitrogen). Antibodies used against endogenous proteins were α-NOP58 (sc-23705 from Santa Cruz Biotechnology), α-Dyskerin H-300 (sc-48794, Santa Cruz Biotechnology), α-Dyskerin C-15 (sc-26982, Santa Cruz Biotechnology) and α-Fibrillarin AFB01 monoclonal antibody line 72B9, lot 011 (from Cytoskeleton, Inc, AFB01). The α-Ago2 (11A9) monoclonal antibody was a gift from Gunter Meister. For PAR-CLIP with NOP56 we used a HEK293 cell line with a stably integrated FLAG-NOP56 fusion gene and IP was done with monoclonal α-FLAG antibody M2 from Sigma. For one Fibrillarin targeted PAR-CLIP the immunoprecipitated complexes were treated with micrococcal nuclease (MNase, from New England Biolabs) for 5 min at 37°C [130]. After SDS-PAGE, gels were blotted onto nitrocellulose membranes to reduce the background from free RNAs [151]. The PAR-CLIP libraries were prepared as described in Additional file 13 and submitted to deep sequencing on an Illumina HiSeq 2000.

The reads obtained from PAR-CLIP experiments were mapped to the human genome (hg19 assembly from UCSC, February 2009) and annotated with the CLIPZ server [131]. Reads marked with the CLIPZ annotation categories 'fungal', 'bacterial,' or 'vector' were discarded and only reads that mapped uniquely to the genome were used in the analyses. The library size was scaled to 1,000,000 for all samples to obtain a normalized expression value (tags per million).

2.5.2 Small RNA sequencing

Small RNA sequencing libraries were prepared from size-selected RNAs of 18 to 30 nucleotides (sdRNA sequencing) and 20 to 200 nucleotides (sRNA sequencing). HEK293 total RNA was extracted and treated with DNase. Next, 20 units of T4 polynucleotide kinase and 2 µl of [γ-32P] ATP (10 µCi/µl) were used to radiolabel 10 µg of RNA at the 5'-ends. The RNA was separated together with a radiolabeled 20-nucleotide ladder on a 12% polyacrylamide gel, the bands corresponding to 18 to 30 nucleotides (for sdRNA sequencing libraries) or 20 to 200 nucleotides (for sRNA sequencing libraries) were excised, the RNA was extracted overnight in a 0.4-M NaCl solution and finally precipitated with ethanol. Small RNA libraries were prepared according to a published protocol [152] and sequenced on an Illumina HiSeq 2000 instrument, for 36 (sdRNA sequencing) and 150 cycles (sRNA sequenicng library). Adaptor removal was done with the CLIPZ server, and the mapping to the human genome was then done with the Segemehl software (v. 0.1.3) with parameters '-D 1 -A 90' [153]. The Gene Expression Omnibus (GEO) accession number for the PAR-CLIP and sRNA-seq data is GSE43666.

2.5.3 Identification of novel C/D snoRNAs and H/ACA snoRNAs from PAR-CLIP and small RNA sequencing data

For each PAR-CLIP library we inferred binding regions of the proteins of interest by clustering reads whose corresponding loci were at most 25 nucleotides apart. To annotate known snoRNA and scaRNA genes we first retrieved sequences from the snoRNA-LBME-db [95], mapped them to the human genome (a list of motif and secondary structure annotated snoRNAs is available in Additional file 13). The 500 binding regions that accumulated the highest number of reads in each individual CLIP library, but did not overlap with known snoRNA or scaRNA genes, ncRNA genes or repeat elements, were screened for novel snoRNA candidates. We used SnoReport [132] to detect H/ACA box snoRNAs, while for detection of C/D box snoRNAs we searched for protein-binding regions that contained motifs corresponding to the C box (RTGATGA; allowing one mismatch) and to the two most common D box motifs (CTGA and ATGA). Sequences that contained both a C box and a D box motif were extended by ten nucleotides in order to search for a terminal closing stem. If a compact closing stem composed of at least four canonical base pairs with at least two G-C/C-G base pairs was found, the sequence was considered a snoRNA candidate. To evaluate the specificity of our C/D box snoRNA gene finding approach, we applied the same procedure to two types of clusters of PAR-CLIP reads from the NOP58 rep A sample both extended by 25 nucleotides on each side. First were the top 100 clusters (defined in terms of the number of reads associated with the cluster) that overlapped with C/D box snoRNA annotation, which served as a positive control. In this set, our program reported 80 sequences as putative snoRNAs. The second type of cluster contained the top 100 clusters that overlap with mRNA exon annotation. These should not contain snoRNAs, and indeed, we only obtained five putative C/D box snoRNAs candidates. Similarly low numbers of snoRNA candidates were obtained from randomized sequences (not shown). Altogether, these tests indicated that our method has very good specificity. In contrast, the number of predictions we obtained from CLIPed clusters without a known annotation was 11 for the top 100 such clusters.

Candidates that showed expression of at least 1 TPM per nucleotide in the 20 to 200 nucleotides small RNA sequencing run (only uniquely mapped reads that covered at least 50% of the candidate snoRNA sequence were considered), and had at least 1 TPM per nucleotide in at least one of the type-specific CLIP libraries were considered putative snoRNAs. They were consecutively numbered, and named as 'ZL#'. To further validate the newly found snoRNAs, we searched for evidence of expression in recently published small RNA-seq libraries from the ENCODE project [133, 154]. Files with the genome coordinates of mapped reads (BAM files) were obtained from the ENCODE data coordination center at UCSC [154] and uniquely mapping reads were used for the analysis. In addition, we selected the 20 candidate C/D box snoRNAs with the highest read count in our data for validation by Northern blotting (see Additional file 13 for details on the experiment). To evaluate the evolutionary conservation of the putative snoRNAs, we carried out a homology search against the vertebrate genomes available in the UCSC genome browser. Once an initial set of homologs was identified, we built sequence/structure models and continued to search for more distant homologs with the Infernal software [155].

2.5.4 Detection of 2'-O-ribose-methylated and pseudouridylated residues

To identify 2'-O-methylated residues we used a reverse transcriptase-based method coupled with polyacrylamide gel analysis as described in [156]. The method is based on the observation that cDNA synthesis is noticeably impaired in the presence of a 2'-O-methyl when deoxynucleotide triphosphate fragments (dNTPs) are limiting [156, 157], giving rise to a characteristic pattern of gel banding immediately preceding the 2'-O-methyls, with strong bands at low dNTP concentrations (0.004 mM) [157], becoming weaker with increasing concentrations of dNTPs.

To map pseudouridines in candidate RNAs we used a method that relies on chemical modification of RNA bases with N-cyclohexyl-N'-β (4-methyl morpholinium) -ethylcarbodiimide (CMC) [158]. The method involves carbodiimide adduct formation with U, G and pseudouridine followed by mild alkali treatment, which removes the adduct from U and G but not from the N-3 of pseudouridine. This modification results in the blockage of reverse transcription one residue 3' of the pseudouridine on the sequencing gel. For a detailed description of assays used to map 2'-O-methyls and pseudouridines see Additional file 13. As a proof of principle, we first applied these assays to the spliceosomal RNA U6, which is known to carry 2'-O-methylated and pseudouridylidated residues. In addition to the well-documented sites, we also observed novel 2'-O-methyl sites that have not been previously reported so far (Additional file 14).

To predict C/D box snoRNAs that could guide 2'-O-methylation, we searched for 8-mer complementarity (only canonical base pairs allowed) to regions immediately or one nucleotide upstream of the D and D' boxes of C/D box and C/D box-like snoRNAs. To predict H/ACA box snoRNAs that could guide pseudouridylations, we used the program RNAsnoop [142]. We first determined for each H/ACA snoRNA stem an energy cutoff value by running simulations on 1,000 random sequences of length 100. Only if an RNAsnoop prediction had an energy value lower than 90% of the random sequences, and at least three canonical base pairs on each side of the binding pocket, did we consider it as a hit.

2.5.5 Ago2 immunoprecipitation sequencing of asynchronous and mitotic cells

Mitotic cells were collected using mitotic shake-off [159, 160], a technique based on the observation that cells become rounded and more easily detachable from the culture vessel as they progress into metaphase during mitosis [161]. Details of the experimental setup are given in Additional file 13. To be able to confirm microscopically that we collected mitotic cells we used HeLa cells with the human histone H2B gene fused to green fluorescent protein (see Additional file 15).

Ago2 was immunoprecipitated from mitotic and asynchronous cells; the Ago2-associated RNAs were extracted and used to prepare cDNA libraries as described above [152], which were then submitted to deep sequencing. Adaptor removal was with the CLIPZ server, and reads were then mapped with Segemehl as described above. In the analysis of small RNA libraries (Ago2-IP and HEK293 sdRNA sequencing (18 to 30 nucleotides)), we considered both uniquely and multi-mapping reads that were annotated based on their mapping to genes in one of the following categories: tRNAs (from the UCSC Table Browser), microRNAs (from mirBase) and snRNAs (from ENSEMBL release 59), C/D box snoRNAs and H/ACA box snoRNAs (curated data set from this work).

2.6 Abbreviations

Ago2: Argonaute 2, *CB:* Cajal body, *CLIP:* cross-linking and immunoprecipitation, *DKC1:* Dyskerin, *dNTP:* deoxynucleotide triphosphate, *FBL:* Fibrillarin, *IP:* immunoprecipitation, *IP-seq:* immunoprecipitation sequencing *miRNA:* micro RNA, *MNase:* micrococcal nuclease, *ncRNA:* non-coding RNA, *PAR-CLIP:* photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation, *rRNA:* ribosomal RNA, *scaRNA:* small Cajal body-specific RNA, *sdRNA:* small derived RNA, *snoRNA:* small nucleolar RNA, *snoRNP:* small nucleolar ribonucleoprotein, *snRNA:* small nuclear RNA, *sRNA:* small RNA, *TPM:* tags per million, *tRNA:* transfer RNA.

2.7 Acknowledgements

The authors thank Erich Nigg (Biozentrum) for providing the HeLa cells in which the histone H2B gene is fused to green fluorescent protein and for advice on the isolation of mitotic cells. We are grateful to Gunter Meister for the anti-Ago2 antibody and Ina Nissen and Christian Beisel from the Quantitative Genomics Facility of the D-BSSE of ETH Zurich, for help with deep sequencing. All computations were carried out on the [BC]2 HPC infrastructure at the University of Basel. Work in the Zavolan lab is supported by the University of Basel and the Swiss National Science Foundation (grant #31003A 127307). SK acknowledges the support of the Gebert Rüf Foundation Rare Diseases Program (grant GRS-046/11).

2.8 Authors' contributions

SK and MZ conceived the project. SK performed the experiments, with help from DJJ (Ago2 IP, primer extensions, and novel snoRNA validation) and APS (novel snoRNA validation). ARG performed the computational analysis of the sequencing data, with help from HJ (computational prediction of snoRNA targets). ARG, DJJ, SK and MZ wrote the manuscript. All authors read and approved the final manuscript.

2.9 Electronic supplementary material

To access the additional files that are mentioned in the manuscript please consult the electronic supplementary material PMID.

Chapter 3 High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq

Nucleic Acids Research Advance Access published December 27, 2016

doi: 10.1093/nar/gkw1321

High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq

Rafal Gumienny[1,2,†], Dominik J. Jedlinski[1,†], Alexander Schmidt[3], Foivos Gypas[1,2], Georges Martin[1], Arnau Vina-Vilaseca[1] and Mihaela Zavolan[1,2,*]

[1]Computational and Systems Biology, Biozentrum, University of Basel, Switzerland, [2]Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Switzerland and [3]Proteomics Core Facility, Biozentrum, University of Basel, Switzerland

Received September 02, 2016; Revised December 08, 2016; Editorial Decision December 15, 2016; Accepted December 19, 2016

ABSTRACT

High-throughput sequencing has greatly facilitated the discovery of long and short non-coding RNAs (ncRNAs), which frequently guide ribonucleoprotein complexes to RNA targets, to modulate their metabolism and expression. However, for many ncRNAs, the targets remain to be discovered. In this study, we developed computational methods to map C/D box snoRNA target sites using data from core small nucleolar ribonucleoprotein crosslinking and immunoprecipitation and from transcriptome-wide mapping of 2′-*O*-ribose methylation sites. We thereby assigned the snoRNA guide to a known methylation site in the 18S rRNA, we uncovered a novel partially methylated site in the 28S ribosomal RNA, and we captured a site in the 28S rRNA in interaction with multiple snoRNAs. Although we also captured mRNAs in interaction with snoRNAs, we did not detect 2′-*O*-methylation of these targets. Our study provides an integrated approach to the comprehensive characterization of 2′-*O*-methylation targets of snoRNAs in species beyond those in which these interactions have been traditionally studied and contributes to the rapidly developing field of ‘epitranscriptomics’.

INTRODUCTION

RNAs are extensively modified in all living organisms (1). Recently, high-throughput approaches have been developed to map 2′-*O*-methylated riboses (2′-*O*-Me, (2)) and nucleobases carrying the most frequent modifications, including N6-methyladenosine (m6A, (3)), pseudouridine (ψ, (4)) and 5-methylcytosine (m5C, (5)), transcriptome-wide. These studies have catalyzed the birth of ‘epitranscriptomics’ (6) and have rekindled the interest in the functions of RNA modifications and their relevance for human diseases (7,8). Whereas 2′-*O*-ribose methylation has long been implicated in the stability and structure of ribosomal RNAs (reviewed in (9)) and m6A appears to modulate the rate of mRNA translation (10–13), the role of most RNA modifications remains to be characterized.

The 2′-*O*-methylation of riboses in ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs) and in Archaea, transfer RNAs (tRNAs) (14–16), is catalyzed by the protein fibrillarin. Fibrillarin is part of a larger ribonucleoprotein (snoRNP) complex whose protein components in mammals and yeast are: FBL (fibrillarin)/Nop1 (17), SNU13/Snu13 (18), NOP56/Nop56 and NOP58/Nop58 (19). As summarized in (20), it is generally accepted that the snoRNP complex assembles sequentially. SNU13/Snu13 initially binds the guide RNA, leading to the folding of the K-turn motif, and the subsequent binding of the NOP56/Nop56:NOP58/Nop58 heterodimer. This complex helps position the guide RNA in its active conformation and is completed by the binding of FBL/Nop1, the snoRNP component responsible for the 2′-*O*-methylation enzymatic activity. As we here focus on human snoRNA, to simplify reading we use hereafter the corresponding nomenclature. The guiding C/D-box small nucleolar RNAs (snoRNAs) (in Archaea small RNAs) take their names from conserved C/C’ (RUGAUGA, R = A or G) and D/D’ (CUGA) boxes. Molecules with more complex structure, which can include additional H/ACA boxes and signals that direct their localization to Cajal bodies (therefore called small Cajal body-associated RNAs or scaRNAs (21)) have also been identified and are essential for the modification and proper functioning of snRNAs. The C/C’ and D/D’ boxes are important for snoRNA biogenesis and for the interaction with RNA binding proteins (22). ‘Anti-sense’ elements located upstream of the D and/or D’ boxes, base-pair with the targets. The target nucleotide that pairs with the fifth nucleotide of the anti-sense element acquires the 2′-*O*-Me mark. Base-pairing adjacent to the target site can further enhance 2′-*O*-methylation (23).

[*]To whom correspondence should be addressed. Tel: +41 61 267 1577; Fax: +41 61 267 1584; Email: mihaela.zavolan@unibas.ch
[†] These authors contributed equally to this work as first authors.

Many studies have investigated snoRNA-guided modifications, particularly in yeast (24–27). As a result, features that define snoRNA target sites have been identified and incorporated into computational methods for snoRNA target prediction (28,29). They include a high complementarity to the 3′ end of the anti-sense box, with no more than one mismatch over at least seven nucleotides, and no bulges (29). A few snoRNAs including U3, U8, U13 have been found to be essential for the processing of rRNA precursors in multiple species, whereas U14 functions in both guiding 2′-*O*-methylation as well as rRNA precursor processing (30–33).

Until the introduction of the crosslinking, ligation and sequencing of hybrids (CLASH) (34), experimental characterization of snoRNA target sites was laborious and addressed only a few sites at a time (35). Progress on method development was further driven by the need to generalize target identification approaches to other guide RNAs, such as the miRNAs (36). Interestingly, miRNA–target hybrids are produced by the action of endogenous ligases and can be obtained through crosslinking and immunoprecipitation (CLIP) of Argonaute proteins, without a specific ligation step (37). MiRNA targets inferred from the chimeric reads obtained with CLIP seem to behave more as canonical miRNA targets, responding more strongly to miRNA transfection, than CLASH-determined targets (38). Whether snoRNA–target chimeras can also be obtained from the CLIP of core snoRNPs has not been investigated.

In parallel with the capture of snoRNA–target interactions, efforts were undertaken to map 2′-*O*-methylated riboses in ribosomal RNAs, also in high-throughput (2). Taking advantage of the resistance of 2′-*O*-methylated riboses to alkaline hydrolysis, the RiboMeth-seq method was used to map 54 annotated and 1 predicted 2′-*O*-methylated site in *Saccharomyces cerevisiae* and is now applied to the profiling of rRNA modifications in human cells as well (39).

Studies from various groups have recently expanded the set of human snoRNAs, beyond those that are catalogued in snoRNAbase (https://www-snorna.biotoul.fr/ (40)) (41–44). Taking advantage of the processing pattern that most C/D-box snoRNAs seem to follow (42) and of the small RNA sequencing data sets generated by the ENCODE consortium, we have recently constructed an updated catalog of human snoRNAs (44). Interestingly, in data sets from both small RNA sequencing and from core snoRNP CLIP we reproducibly identified snoRNA-like sequences which contained only a subset of the C/D box snoRNA-specific sequence elements. For most snoRNA-like molecules we could not predict target sites.

Given the surge in data sets pertaining to snoRNA interactions, we here sought to provide relevant computational analysis methods. First, we developed a model to identify chimeric sequences, composed of a C/D box-containing RNA and a corresponding target part, among the reads obtained by CLIP of core C/D-box snoRNPs. To further enable the functional characterization of the chimera-documented interactions, we developed a model to identify sites of 2′-*O*-Me from RiboMeth-seq data (2). Our data supports the concept that some rRNA sites are only partially methylated (39) and indicates that some of the snoRNAs which are not known to guide 2′-*O*-methylation interact with sites whose methylation is guided by other snoRNAs. Interactions with strong chimeric read support outside of the canonical snoRNA targets, do not seem to lead to 2′-*O*-ribose methylation that can be detected with RiboMeth-seq. This suggests that the sensitivity of RiboMeth-seq is low or that C/D box snoRNA interaction with non-canonical targets may serve yet uncharacterized functions.

MATERIALS AND METHODS

CLIP of snoRNP core proteins

To identify chimeric snoRNA–target reads, we analyzed five CLIP data sets that were published before (42): two NOP58-CLIP (Gene Expression Omnibus (GEO) accession numbers GSM1067861 and GSM1067862), 1 NOP56-CLIP (GEO accession # GSM1067863) and 2 FBL-CLIP (GEO accession # GSM1067864 and GSM1067865). We also generated an additional FBL-CLIP data set with the protocol described in (45) (GEO accession GSE77027).

Identification of snoRNA–target chimera

SnoRNA and target sets. We obtained the most comprehensive annotation of human snoRNA sequences, genome coordinates and known or predicted targets from the human snoRNA atlas that was recently published (44). We downloaded the sequences of known snoRNA targets (rRNA and snRNA) from the snoRNA database (40) and we further obtained tRNA sequences from GtRNAdb (46). We added one tRNA sequence per codon to the set of putative snoRNA targets. The database of putative snoRNA targets thus consisted of the GRCh37 version of the human genome assembly, augmented with rRNA, snRNA and tRNA sequences.

Computational analysis of chimeric reads

Analogous to a previous study that developed a strategy to uncover chimeric miRNA–target reads from Argonaute-CLIP data (37), we here developed a method that uses snoRNA-specific information to identify snoRNA–target chimera in core snoRNP CLIP data sets. The challenge is that the very low frequency of chimeric reads in CLIP data sets and the short length of the snoRNA and target parts in the typically short reads obtained from CLIP can lead to a high rate of false positive chimeras, making it necessary to use additional information, such as the specific pattern of hybridization of the guide RNA to the target.

Read selection. We carried out an initial annotation of CLIP data sets with the CLIPZ web server (47), which provides as output genome-mapped reads with their respective annotations, as well as the unmapped reads. Because we look for snoRNA–target interactions that take place within the snoRNP complex, we expect that target sites are also captured on their own in the core snoRNP CLIP, just as miRNA targets are captured in Argonaute-CLIP (48). Thus, to reduce the search space, we used clusters of at least two overlapping genome-mapped reads as putative target regions. To have sufficiently long snoRNA and target parts in the chimeric reads, we only used unmapped reads longer than 24 nucleotides.

Detection of snoRNA subsequences in unmapped reads. To speed up the identification of snoRNA subsequences within unmapped reads we first generated all possible subsequences of 12 nucleotides in length ('anchors') from all snoRNAs. We then searched the unmapped reads for exact matches to any of these anchors and, when a match was found, we carried out the local alignment of the respective snoRNA to the unmapped read with the swalign python package (https://pypi.python.org/pypi/swalign) (parameters for a match = 2, mismatch = −5, gap opening = −6, gap extension = −4). For each chimeric read, we retained only the snoRNA(s) with the best local alignment score. To evaluate the significance of the alignment scores, we applied the same procedure to shuffled reads. For most of the reads, the score of the alignment with the snoRNA presumed to be contained in the read was much higher compared to the score of aligning the snoRNA to a shuffled version of the read (Supplementary Figure S1A). Thus, as it appears that many unmapped reads indeed contain snoRNA subsequences, we split chimeric reads into the part that could be aligned to a snoRNA (the 'snoRNA fragment') and the rest of the read ('putative target fragment'). All reads with a putative target fragment of at least 15 nucleotides were considered candidate chimeras which we analyzed further as described below.

Annotation of putative target fragments extracted from chimeric reads

The search space for putative target fragments consisted of CLIPed sites as well as rRNA, snRNA and tRNA sequences, which we explicitly included because the reference genome assembly may not contain all of the repetitive loci of these RNAs. As the PAR-CLIP protocol yields reads in which C nucleotides are incorporated at the sites of crosslinked U's, before carrying out the mapping of the putative target fragments we generated single-point variants of the reads, with one C nucleotide changed to a U (37). For the mapping, we used Bowtie2 (49) in the local alignment mode with the following command line parameters: -f -D100 -L 13 -i C, 1 –score-min C, 30 –local -k 10. For reads that mapped to multiple genomic loci, we checked whether at least one of these loci corresponded to a canonical snoRNA target, rRNA or snRNA. If so, we kept only the canonical locus. Otherwise, we kept all putative target loci. The statistics for each experimental data set can be viewed in Supplementary Table S1.

Training a model of snoRNA–target interaction

To better distinguishing *bona fide* snoRNA–target interactions captured in chimeras from false positives, we developed an additional model as follows. We extracted putative target sites that were captured in multiple chimeras with the same snoRNA and had a PLEXY-predicted energy of interaction (28) lower than -12 kcal/mol. From the combined CLIP experiments we identified 362 such sites in the 28S and 18S ribosomal rRNAs. 67 of these are known to undergo 2′-*O*-ribose methylation (we called these 'positives'), whereas for the remaining 295 sites a modification is not so far known to occur ('negatives'). For each site, we calculated the features described below and trained a model to predict the class ('positive' or 'negative') of sites in the 28S rRNA. We evaluated the performance of the model using the the known modification sites on the 18S rRNA as true positives and all other sites in the 18S rRNA as true negatives. As the performance was high, we combined the two data sets and retrained a model for the comprehensive identification of snoRNA–target interactions.

Feature definition and computation

Predicted energy of snoRNA–target interaction. PLEXY is a tool for the transcriptome-wide prediction of C/D box snoRNA targets. It uses nearest-neighbor energy parameters to compute thermodynamically stable C/D-box snoRNA - target RNA interactions (28,50), but applies additional rules to further reduce the false positive rate. For each putative target fragment that mapped to the database of putative targets (see section SnoRNA and target sets) we extracted a 50 nucleotides long sequence centered on the target part of the chimeric read, and calculated its interaction energy with the snoRNA also identified from the chimeric read. PLEXY also assigns the position of the snoRNA-induced modification and we kept this information for further analyses. To assess the value of the PLEXY score in identifying *bona fide* interactions, we shuffled the snoRNA associated with each target part in a chimeric read and repeated the calculation.

Target site accessibility. Known snoRNA–target site interactions involve perfect base-pairing of the nucleotides at the 3′ end of the anti-sense box, which is anchored at the D box. This interaction region defines the 5′ end of the target site. Therefore, we defined the accessibility of the target region as the probability that the 5′-anchored 21 nts-long region in the target is in single stranded conformation within an extended region of 30 nucleotides upstream and 37 nucleotides downstream of 5′ end of the putative site. We computed this value with CONTRAfold (51).

Nucleotide content of flanking regions. We defined the 'Flanks A content' as the proportion of adenines within the 67 nts-long region defined above. We similarly computed frequencies of other nucleotides. Because the frequency of adenines was most predictive of true interaction sites (Supplementary Figure S2) we only used this feature in the model.

Model training

The histograms constructed separately for the positive and negative sites in the 28S and 18S rRNAs indicated that the features described above are informative for distinguishing positive from negative sites (Figure 1) and we therefore trained a generalized linear model (GLM) with the logit link function (logistic regression) using these features, with the Statsmodels python library (52). We built the model based on all 18S rRNA and 28S rRNA sites. The code that we used to extract putative snoRNA–target interactions from CLIP data can be obtained from the github (https://github.com/guma44/snoRNAHybridSearchPipeline) and additional information is available on the accompanying web site (http://www.clipz.unibas.ch/snoRNAchimeras).

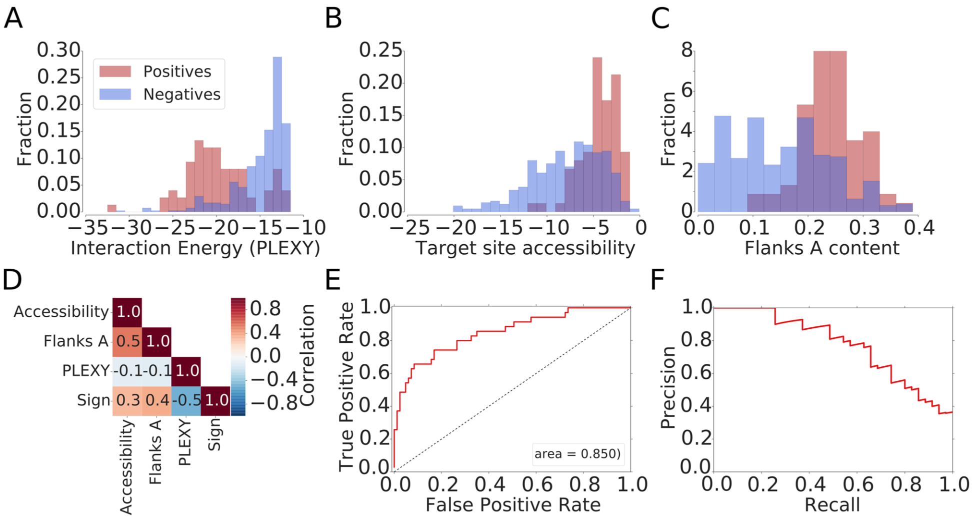


Figure 1. Features that are relevant for the identification snoRNA–target interactions based on chimeric reads. Distributions of (**A**) the interaction energy calculated with PLEXY (28), (**B**) the target site accessibility calculated with CONTRAfold (51) and (**C**) the A nucleotide composition of the neighborhood of positive (known) and negative (captured in chimeras but unknown) snoRNA interaction sites. (**D**) Correlation between features used for model training and the indicator function, taking the value of −1 for negative and 1 for positive sites. (**E**) Receiver operating characteristic (ROC) curve and (**F**) Precision-Recall (PR) curve constructed based on snoRNA target predictions in 18S rRNA with the model trained on 28S rRNA target sites.

Annotation of modification sites

We annotated the biotypes of the targets in which predicted modification sites resided based on the ENSEMBL version 75 (53) and the RMSK table from University of California Santa Cruz genome browser (54), for the repeat elements. From the known interactions that we retrieved with our model from chimeric reads, we separately extracted those that involve the anti-sense elements at the D and D' boxes and constructed profiles of coverage of the corresponding snoRNAs by fragments from chimeric reads, relative to the position of the D box. As shown in Supplementary Figure S1B and C, the appropriate anti-sense elements were captured preferentially in chimeric reads, although other parts of the snoRNAs have also been ligated with some frequency to the targets.

RiboMeth-seq

Preparation and sequencing of RiboMeth-seq libraries. The principle behind RiboMeth-seq is that nucleotides with a 2′-*O*-Me ribose are resistant to alkaline hydrolysis. Thus, products of partial alkaline hydrolysis should not start or end at 2′-*O*-Me sites, leading to an underrepresentation of these positions among read starts and ends. The read starts and ends thus provide a negative image of the methylation landscape (2). We carried out RiboMeth-seq experiments in HEK 293 cells, using either total RNA or poly(A)-enriched RNA from either the nucleus or cytoplasmic fractions. We also carried out the alkaline hydrolysis for different time intervals of 8, 14 or 20 min. The samples that we prepared were as follows:

RiboMethSeq_HEK_totalRNA_8min
RiboMethSeq_HEK_totalRNA_14min
RiboMethSeq_HEK_totalRNA_20min
RiboMethSeq_HEK_polyARNA_8min
RibomethSeq_HEK_cytoplasmic1_14min
RibomethSeq_HEK_cytoplasmic2_14min
RibomethSeq_HEK_nuclear1_14min
RibomethSeq_HEK_nuclear2_14min

We extracted total RNA with TRI Reagent (Sigma) and prepared the mRNA with the Dynabeads mRNA DIRECT Kit (Life Technologies), from HEK293 cells according to the manufacturer's instructions. For mapping of 2′-*O*-methyl sites in rRNA we used 1 μg of total RNA as starting material. To explore the existence of 2′-*O*-methyl sites in mRNAs, we used poly(A)-selected RNA (200 ng). In both protocols, the RNA was partially degraded under alkaline conditions in a sodium carbonate/bicarbonate buffer at pH 9.2 for 14 min and then put on ice. Samples were separated parallel to a low molecular weight marker ladder (10–100 nt) on a 15% denaturing polyacrylamide gel for 1 h at 1400 V and 20 W. The gel was stained with GR Green nucleic acid stain (Excellgen) for 3 min and fragmented RNA ranging from 20 to 40 nt was cut out from the gel and extracted overnight in 0.4 M NaCl. The RNA was precipitated with 1 μl of co-precipitant (GlycoBlue) in 75% ethanol at −20°C for 2 h and then centrifuged at maximum speed for 10 min at 4°C. The RNA pellet was washed twice with 70% ethanol and air-dried. The pellet was dissolved in water, the RNA was dephosphorylated with FastAP alkaline phosphatase (Thermo Scientific) at 37°C for 30 min and the enzyme was heat-inactivated at 75°C for 10 min. Subsequently, the RNA was phosphorylated with polynucleotide kinase (Thermo Scientific) in the presence of 1 mM ATP at 37°C for 1 h and then extracted with phenol-chloroform and precipitated in 80% ethanol, washed with 70% ethanol twice and air-dried. The pellet was dissolved in 8 μl mix (4 μl H_2O, 1 μl 10× truncated T4 RNA Ligase 2 buffer, 1 μl 100 uM 3′ rApp-adapter (5′ adenylated 3′ adapter, 5′-App-TGGAATTCTCG GGTGCCAAGG-amino-3′), 2 μl 50% DMSO), denatured at 90°C for 30 s and chilled on ice. Next, RNasin Plus RNase inhibitor (Promega) and

T4 RNA Ligase 2 truncated were added to a final concentration of 2 U/μl and 30 U/μl, respectively, and the reaction was incubated at 4°C for 20 h over night. The next day, 1 μl of RT primer (100 μM; 5′-GCCTTGGCAC CCAGAGAATTCCA-3′) was added (for quenching of remaining 3′ adapter molecules, preventing adapter dimers ligation in the next step), the samples were heated at 90°C for 30 s, at 65°C for 5 min, then placed on ice. A 5′-adapter ligation mix was then directly added to the sample (1.5 μl 10 mM ATP, 1 μl 100 uM 5′ RNA Adapter RA5 (Illumina TruSeq RNA sample prep kit), 1 μl T4 RNA Ligase 1 (20 U/μl), 0.5 μl RNasin Plus RNase inhibitor (40 U/μl) and reactions were incubated at 20°C for 1 h and 37°C for 30 min. The RNA was then directly reverse transcribed in a 30 μl reaction by adding dNTPs to 0.5 mM, DTT to 5 mM, 1× SSIV buffer, RNAsin to 2 U/μl and 1 μl Superscript IV reverse transcriptase (Life Technologies). The sample was incubated at 50°C for 30 min and inactivated at 80°C for 10 min. 5 μl of the resulting cDNA was then used in a pilot polymerase chain reaction (PCR) reaction. To this end, aliquots were taken from reactions at every second cycle between 12 and 22 cycles and analyzed on a 2.5% agarose gel. The number of cycles causing a first visible amplification was chosen for a large scale PCR (10 μl cDNA in a 100 μl reaction). The PCR product was ethanol precipitated and run along a 20 bp marker on a 9% non-denaturing polyacrylamide gel in TBE for 1 h at 250 V, 20 W. The gel was dismantled and stained for 3 min with GR Green. PCR products between 125 and 175 bp were cut out, the gel piece was mashed and DNA was eluted overnight into 400 μl of H2O. The supernatant was separated from the gel particles in a SpinX filter column (Costar), NaCl was added to 0.4 M, DNA was ethanol precipitated, the pellet washed in 75% ethanol and dissolved in 20 μl H_2O. Libraries were sequenced on an Illumina HiSeq-2500 deep sequencer (GEO accession: GSE77024). Their summary can be found in Supplementary Table S2.

Mapping of RiboMeth-seq reads. We obtained ∼50 million reads for each of the RiboMeth-seq samples. We removed adaptors with Cutadapt (–minimum-length 15, other parameters left with default values) (55) and mapped the reads with STAR (parameters: –outFilterMultimapNmax 20 –outFilterMismatchNoverLmax 0.05 –scoreGenomicLengthLog2scale 0 –outSAMattributes All) (56) to a human GRCh37 assembly version-based transcriptome composed of rRNAs, snRNAs, tRNAs and snoRNAs (see SnoRNA and target sets section) as well as to lincRNAs, miscRNAs, and all unspliced protein coding genes (obtained from GRCh37 version of ENSEMBL, http://grch37.ensembl.org/index.html (53)).

Computation of the RiboMeth-seq score. For each target of interest such as the 18S rRNA, we calculated the log2 normalized (to a total library size of 10^6 reads) profile of cleavage positions. We used separately the 5′ and 3′ ends of the reads, as both ends are determined by alkaline hydrolysis. We then calculated the angle defined by the $\log_2$ coverage values at positions −1, 0 and +1 for each position along the RNA. An angle of 180° indicates that the frequency of cleavage at the three adjacent positions is identical, 0° indicates that the central position has very high coverage compared to the neighboring positions (and is therefore not protected from cleavage) and 360° indicates that the central position has no coverage (and therefore it is protected from cleavage) compared to the neighboring positions. As a RiboMeth-seq score we took the average angle computed based on 5′ and 3′ read ends. We used a score threshold of 290° for predicting sites in individual RiboMeth-seq experiments, favoring slightly recall over precision. Detailed statistics for individual experiments can be found in Supplementary Table S2. Finally, we used putative 2′-*O*-Me sites that had a score above the threshold in at least one experiment and calculated their average score across the seven experiments. To determine a threshold for this average score and then compute the PR curve and Matthews correlation coefficient, we included among the positives the 19 sites that were did not score above the threshold in any individual experiment, but are known to undergo methylation. This resulted in a set of 105 known sites in the 18S and 28S rRNAs.

Validation of 2′-*O*-methylation sites with RTL-P

Similar to the classic primer extension assays (57), the 'Reverse Transcription at Low deoxy-ribonucleoside triphosphate (dNTP) followed by polymerase chain reaction' method (RTL-P, (58)) takes advantage of the observation that cDNA synthesis through a 2′-*O*-Me nucleotide is impaired when dNTPs are limiting. However, RTL-P is simpler and more sensitive than primer extension assays. RTL-P consists of a site-specific primer extension by reverse transcriptase at a low dNTP concentration and a semi-quantitative PCR amplification step, followed by agarose gel electrophoresis to obtain ratios of PCR signal intensities. To increase sensitivity and reproducibility, we implemented a real-time PCR (qPCR) step to facilitate the analysis of the signal intensities (qPCR parameters and primer sequences are shown in Supplementary Table S3).

Validation of 2′-*O*-methylation at G2435 in 28S with mass spectrometry

The rRNA fragment isolation for mass spectrometry analysis (MS) was adapted from (59). The isolated fragment was treated with RNase T1 to yield a specific digestion pattern and dephosphorylated prior to LC–MS/MS analysis. As reference we used 11-nts long synthetic RNA oligonucleotides identical in sequence to the 28S rRNA around the G2435 site. 20 pmol of the unmodified synthetic UCCUGAGAGAU as well as the 2′-*O*-methylated synthetic variant UCCUG*AGAGAU (the methylated G is indicated by *) were subjected to RNase T1 digestion and dephosphorylation.

Samples were analyzed on a LTQ-Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) using a targeted LC-MS/MS workflow as described recently (60). UCCUG and UCCUG* specific MS assays were generated from the synthetic RNA oligonucleotides and applied to all samples. Data analysis was carried out using the Qual Browser tool of the Xcalibur software (version: 3.0.63). Full details of the sample preparation and LC-MS/MS experiment are described in Supplementary Figure S3.

RESULTS

Crosslinking and immunoprecipitation of core snoRNPs captures snoRNA–target site chimeras

Although miRNAs and snoRNAs differ entirely in their function, they share the ability to guide ribonucleoprotein complexes to target RNAs. Thus, by analogy with miRNAs (37), we hypothesized that chimeric molecules, composed of snoRNAs and their targets, are captured in CLIP experiments that target one of the core snoRNP proteins. Therefore, we designed a method to identify snoRNA–target chimeric reads from among the unmapped (to genome or transcriptome) reads obtained in six photoreactive nucleoside-enhanced (PAR)-CLIP experiments that targeted one of the NOP58, NOP56 and FBL proteins. We found that on average, ∼10% of the reads that were not mapped to the genome or transcriptome had at least a 12-nt match to a snoRNA. However, only for ∼half of these reads was the remaining, putative target part of the read, longer than 15 nucleotides. Because multi-family snoRNAs have very low expression in the HEK 293 cells, most of the putatively chimeric reads yielded a high-scoring alignment to a single snoRNA, and only ∼20% aligned to multiple snoRNAs. A summary of the data obtained in all of these experiments is shown in Supplementary Table S1. To determine whether the apparent snoRNA–target chimera do reflect real interactions, we randomized the snoRNA assigned to each target fragment in the chimeras and calculated the predicted energies of interaction of the real and randomized pairs of molecules with PLEXY (28). Although the interaction energy predicted for the presumed chimeras was significantly lower compared to randomized sequence pairs, the difference between the average PLEXY energies was relatively low (∼1.2 kcal/mol, Figure 2A). This indicated that that a more sophisticated approach is needed to reliably identify snoRNA–target interactions from these data, which likely contain a large number of false positives.

A model to identify high-confidence snoRNA–target chimeras

For training a model to predict snoRNA–target interactions, we selected presumed snoRNA–rRNA chimeras with low predicted energy of interaction (<−12 kcal/mol), separated them into those containing ‘positive’ target sites (known from previous studies) and those containing ‘negative’ target sites (not known to undergo snoRNA-guided methylation) and compared the distributions of features that have been found to play a role in other small RNA-guided interactions (61) between the two sets. The PLEXY interaction score (28) discriminated best these two data sets (as shown in Figure 1A and D). However, known snoRNA target sites also reside in structurally accessible regions (Figure 1B), rich in adenines (Figure 1C). We used chimeric reads involving the 28S rRNA to train a generalized linear model (GLM) based on these features and then tested the model on chimeric reads involving the 18S rRNA. The area under the receiver operating characteristic (ROC) curve was ∼85%, the model being able to recall 70% of the known interaction sites with 65% precision (Figure 1E and F). We then combined the sites in the 28S and 18S rRNAs, retrained the model, and found that at a score threshold of 0.15 we obtained good performance in predicting rRNA modification sites, with a Matthews correlation coefficient (MCC) of ∼0.75, precision of 0.75 and recall value of 0.74 (Figure 2B–D). Our predictions finally consisted of putative interactions that were supported by chimeric reads from at least two experiments and had a minimum score of 0.15. For completeness, we have also predicted interactions in individual data sets and show the overlap of sites obtained in pairs of experiments in Supplementary Figure S4.

Chimeric reads reveal novel C/D box snoRNA target sites within structural RNAs

We applied the derived model to the full chimeric read data and identified 980 putative interactions, involving 852 unique target sites. We focused on the snoRNA interactions with structural RNAs, including not only the rRNAs, but also snRNAs, tRNAs and the snoRNAs themselves. Only one of the 2′-*O*-Me sites in rRNAs that have been mapped so far is is ‘orphan’, meaning that its guide snoRNA is unknown. Our data indicates that this modification, located at position A1383 in the 18S rRNA (62), is guided by SNORD30 (Figure 3A), a snoRNA which was reported to guide the 2′-*O*-methylation at position A3804 in 28S rRNA (63). The chimeric reads also revealed 35 potentially novel 2′-*O*-Me sites in rRNAs (13 in 18S rRNA, 21 in 28S rRNA and 1 in 5.8S rRNA), some of which were found in interaction with multiple snoRNAs, thus corresponding to 40 novel interactions. Eleven of the 40 interactions involve snoRNAs that have been so far classified as ‘orphan’ (Supplementary Table S4). As an example, a snoRNA of unknown family (snoID_372) was found in three experiments in interaction with the 28S rRNA (predicted energy of interaction of −24.8 kcal/mol), in which it may guide the modification at position 4953 (Figure 3B). Similarly, in two experiments we found the recently uncovered snoID_0701 (family unknown) orphan snoRNA, which has low but broad expression across tissues (44), in a very stable (−28.2 kcal/mol) interaction with the 28S rRNA. This snoRNA is predicted to guide the 2′-*O*-methylation at position U2756 (Figure 3C).

SnRNAs are also known targets of scaRNA-guided 2′-*O*-methylation. Of the nine such sites that are known, we were able to recover four over our prediction threshold. Additionally, we identified chimeric reads of the SNORD23, a snoRNA that is currently considered orphan, with the U6 snRNA (Figure 3D). In previous work (42), we have studied the methylation pattern of this snRNA by primer extension. We found evidence of 2′-*O*-methylation at positions 60, 62 and 63 of U6, but not at position 64, which is predicted to be modified as a result of the interaction with SNORD23. Thus, the significance of this interaction, supported by 11 reads in our data, remains to be determined.

Additionally, we identified three apparent interactions of snoRNAs with other snoRNAs (SNORD5 with SNORD56, SNORD50 with SNORD57 and SNORD34 with SNORD38A), as well as an intra-molecular chimera of SNORD4B. The predictions are summarized in Supplementary Table S4 and all alignments of putative chimeric reads to putative target sites and snoRNAs can be viewed at http://www.clipz.unibas.ch/snoRNAchimeras/.

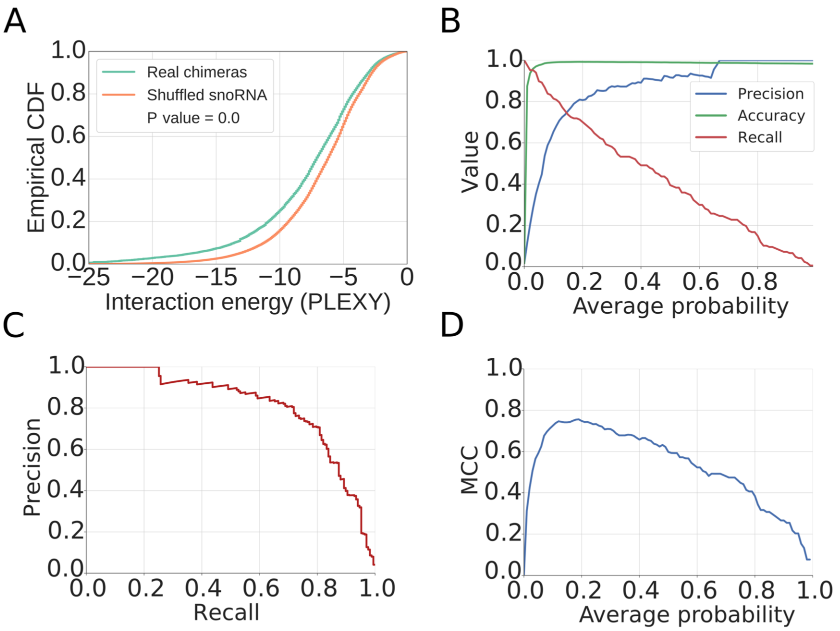


Figure 2. Characterization of the model for inferring snoRNA–target interactions from chimeric reads. (**A**) Empirical cumulative distribution function of the interaction energy estimated with PLEXY between target fragment and snoRNA found in the chimera ('Real chimeras') or between target fragment and a randomly assigned snoRNA ('Shuffled snoRNA'). *P*-value from the Mann–Whitney *U* test is also shown. (**B**) Metrics illustrating the performance of the method, as a function of the minimum average probability of the considered sites from the 18S and 28S rRNAs. (**C**) Precision-Recall curve for the method. (**D**) Matthews correlation coefficient (MCC) as a function of the minimum average probability of the considered sites.

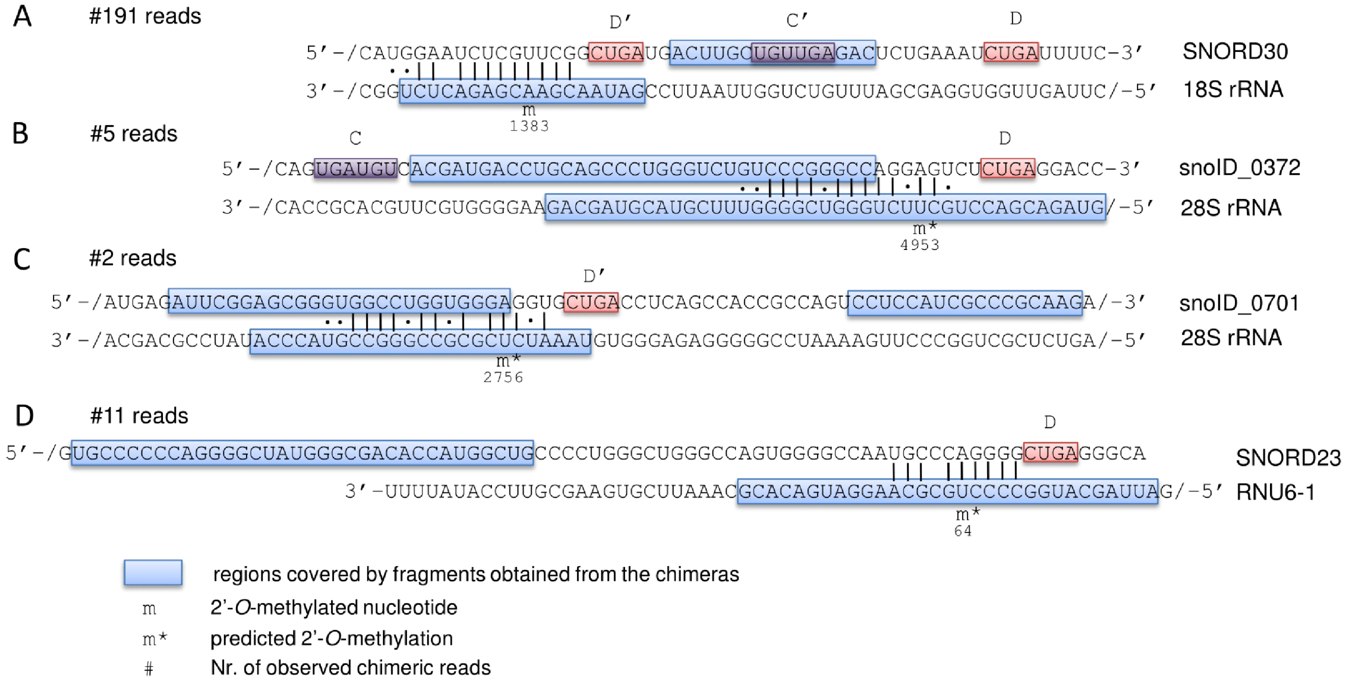

A
#191 reads
D′
C′
D
5′-/CAUGGAAUCUCGUUCGGCUGAUGACUUGCUGUUGAGACUCUGAAAUCUGAUUUUC-3′
3′-/CGGUCUCAGAGCAAGCAAUAGCCUUAAUUGGUCUGUUUAGCGAGGUGGUUGAUUC/-5′
m
1383
SNORD30
18S rRNA
B
#5 reads
C
D
5′-/CAGUGAUGUCACGAUGACCUGCAGCCCUGGGUCUGUCCCGGGCCAGGAGUCUCUGAGGACC-3′
3′-/CACCGCACGUUCGUGGGGAAGACGAUGCAUGCUUUGGGGCUGGGUCUUCGUCCAGCAGAUG/-5′
m*
4953
snoID_0372
28S rRNA
C
#2 reads
D′
5′-/AUGAGAUUCGGAGCGGGUGGCCUGGUGGGAGGUGCUGACCUCAGCCACCGCCAGUCCUCCAUCGCCCGCAAGA/-3′
3′-/ACGACGCCUAUACCCAUGCCGGGCCGCGCUCUAAAUGUGGGAGAGGGGGCCUAAAAGUUCCCGGUCGCUCUGA/-5′
m*
2756
snoID_0701
28S rRNA
D
#11 reads
D
5′-/GUGCCCCCCAGGGGCUAUGGGCGACACCAUGGCUGCCCCUGGGCUGGGCCAGUGGGGCCAAUGCCCAGGGGCUGAGGGCA
3′-UUUUAUACCUUGCGAAGUGCUUAAACGCACAGUAGGAACGCGUCCCCGGUACGAUUAG/-5′
m*
64
SNORD23
RNU6-1
regions covered by fragments obtained from the chimeras
m 2'-O-methylated nucleotide
m* predicted 2'-O-methylation
# Nr. of observed chimeric reads

Figure 3. Schematic representation of snoRNA–target interactions that are predicted based on chimeric reads from CLIP experiments. For each interaction, the snoRNA sequence is shown at the top and the target sequence at the bottom of the panel. '/' indicates that only part of the sequence is shown, for readability. Regions of both snoRNAs and targets that are represented in the chimeric reads are encompassed in blue boxes. Indicated are also the presumed C/C' and D/D' boxes as well as the number of chimeric reads supporting each of the interactions. PLEXY-predicted sites of 2′-*O*-methylation are marked by 'm*' and the previously mapped site is labeled with 'm'.

Redundant targeting of known sites of 2′-*O*-ribose methylation by multiple snoRNAs

One of the main open questions in the snoRNA field concerns the targets and functions of the 330 orphan snoRNAs, which belong to 219 families (44). As mentioned in the introduction, some of these snoRNAs are involved in pre-rRNA processing. Interestingly however, the chimeric read data shows that SNORD118, also known as U8, a snoRNA which is necessary for the proper maturation of 5.8S and 28S rRNAs (31), interacts with the region of the 28S rRNA where the SNORD80 is known to guide the modification of G1612. The evidence for this interaction is very strong, chimeras having been captured in six distinct experiments (Figure 4). Although the base-pairing between SNORD118 and the putative target site is not as extensive as that of the SNORD80 snoRNA, it still includes 10 consecutive base-pairs, two of which are G-U base pairs. This example suggests that some snoRNAs are capable of interacting with sites whose 2′-*O*-methylation is guided by other snoRNAs. We detected fragments from 66 orphan snoRNAs in chimeric reads.

Identification of snoRNA-guided 2′-*O*-Me sites with RiboMeth-seq

Surprisingly, ∼300 predicted interaction sites mapped to loci encoding protein-coding genes. To evaluate whether these sites could undergo 2′-*O*-methylation, we implemented a high-throughput version of the recently developed RiboMeth-seq method (2). To be able to capture non-canonical targets, we carried out seven experiments, six using total RNA, which contained both the canonical rRNAs targets as well as other RNA species, and one using poly(A)+ RNAs, which was thereby strongly enriched in mRNAs. Two of the total RNA samples were prepared from total cell lysate, two from the nuclear fraction and two from the cytoplasmic fraction.

2′-*O*-Me sites were previously identified from RiboMeth-seq by comparing the number of reads ending at a particular position in the target with the average number of reads ending at the flanking regions (‘score A’ in (2)). Reasoning that 2′-*O*-methylation of adjacent nucleotides is very rare and that 2′-*O*-Me positions should yield much fewer cleavage events compared to the unmethylated adjacent nucleotides, we here tested additionally another score. Specifically, at each position of a target of interest (e.g. 18S rRNA), we evaluated the shape (angle) of the trough defined by the $\log_2$ normalized read coverage at the specific position and the immediately adjacent positions (Figure 5A). We found that this score yields a higher precision compared to the ‘score A’ proposed before (2) (Figure 5B and C) and a very high Matthews correlation coefficient in classifying the sites (Figure 5D).

Applying this method to the combined RiboMeth-seq data, we identified 168 2′-*O*-Me sites, 80 of which were known. These included 32 out of the 45 known 2′-*O*-Me sites in 18S rRNA (71%), 44 out of the 60 in 28S rRNA (73%), the known site at position 75 in 5.8S rRNA, 2 sites in the U6 snRNA and one site in U1 snRNA. Figure 6 shows the location of previously known 2′-*O*-methylation sites in the 18S and 28S rRNAs, as well as the corresponding chimeric read and RiboMeth-seq evidence that we obtained here for these rRNAs. The 88 novel sites were mostly located in canonical snoRNA/scaRNA targets—snRNA, rRNAs and tRNAs—34 being located in other RNA species. Although both the chimeric read method and RiboMeth-seq identified the majority of known 2′-*O*-Me sites, with comparable sensitivity (∼70%), none of the 34 novel target sites in structural RNAs that were found in chimeric reads had a RiboMeth-seq score above the threshold.

Position G2435 in the 28S rRNA, captured in interaction with SNORD2, is partially methylated

To assess whether the limited sensitivity of RiboMeth-seq could be a reason for the limited validation of sites that are reproducibly captured in chimeric reads, we investigated in depth the predicted SNORD2-guided 2′-*O*-methylation of position G2435 in the 28S rRNA. This interaction was captured in four CLIP experiments (Figure 7A).

We applied the recently published method ‘Reverse Transcription at Low deoxy-ribonucleoside triphosphate concentrations followed by polymerase chain reaction’ (RTL-P) (58), which we then followed with qPCR, to improve quantification. After showing that the method yields the expected results on a positive (position A1031 in the human 18S rRNA) and a negative control (U1991 in 28S rRNA) (Supplementary Figure S5), we tested position G2435 in 28S rRNA. We found that the unanchored MeU-RT primer yielded significantly less cDNA and hence PCR product than the anchored MeA-RT primer at low dNTP concentrations (Figure 7B), indicating that the site indeed carries a 2′-*O*-Me modification.

To unambiguously show that the RT stoppage at G2435 is due to 2′-*O*-methylation, we applied targeted mass spectrometry (60). Figure 7C shows the extracted ion chromatograms of specific UCCUG* fragments that were measured in 28S rRNA as well as in a control sample. We manually checked the identities of the employed fragments using the control sample (Supplementary Figure S3A) and found that they matched those obtained from the HEK rRNA (Supplementary Figure S3B), confirming the presence of UCCUG* in the HEK sample. The LC-MS analysis also identified the unmodified fragment UCCUG from HEK rRNA (Supplementary Figure S3C), albeit at a lower level than UCCUG* (Supplementary Figure S3D). These results show that the G2435 28S rRNA site identified among the chimeric reads is predominantly 2′-*O*-methylated.

mRNAs captured in chimeras with snoRNAs do not show evidence of 2′-*O*-methylation

Finally, we wondered whether some of the chimera-supported interactions that did not reside in the typical snoRNA targets, particularly those annotated as being located in mRNAs, were also below the sensitivity of RiboMeth-seq. We therefore applied RTL-P to four mRNA-annotated sites, located in APP, CCDC93, DHFR and ZC3H12C transcripts, but did not find evidence of 2′-*O*-methylation (data not shown).

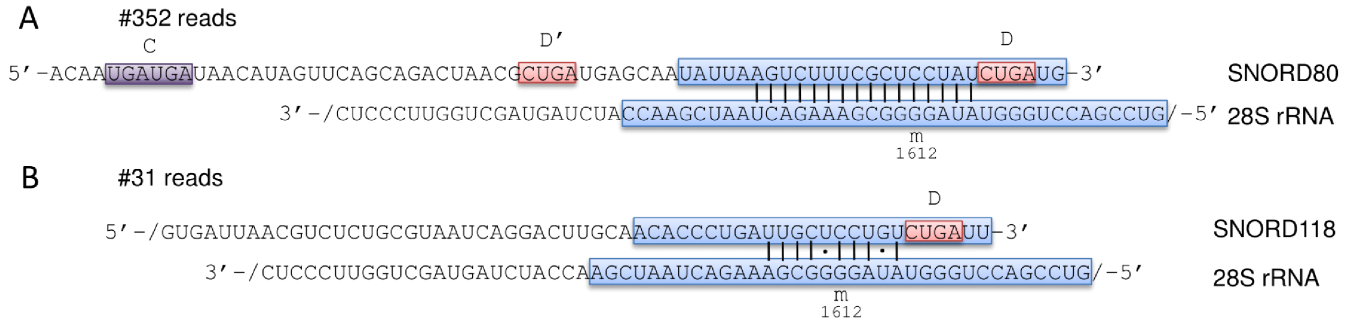

A
#352 reads
C
D′
D
5′-ACAAUGAUGAUAACAUAGUUCAGCAGACUAACGCUGAUGAGCAAUAUUAAGUCUUUCGCUCCUAUCUGAUG-3′
SNORD80
3′-/CUCCCUUGGUCGAUGAUCUACCAAGCUAAUCAGAAAGCGGGAUAUGGGUCCAGCCUG/-5′
28S rRNA
m
1612
B
#31 reads
D
5′-/GUGAUUAACGUCUCUGCGUAAUCAGGACUUGCAACACCCUGAUUGCUCCUGUCUGAUU-3′
SNORD118
3′-/CUCCCUUGGUCGAUGAUCUACCAAGCUAAUCAGAAAGCGGGAUAUGGGUCCAGCCUG/-5′
28S rRNA
m
1612

Figure 4. Similar to Figure 3, representation of the data supporting the interaction of both SNORD80 and SNORD118 with the 28S rRNA, around the known position of 2′-*O*-methylation at G1612.

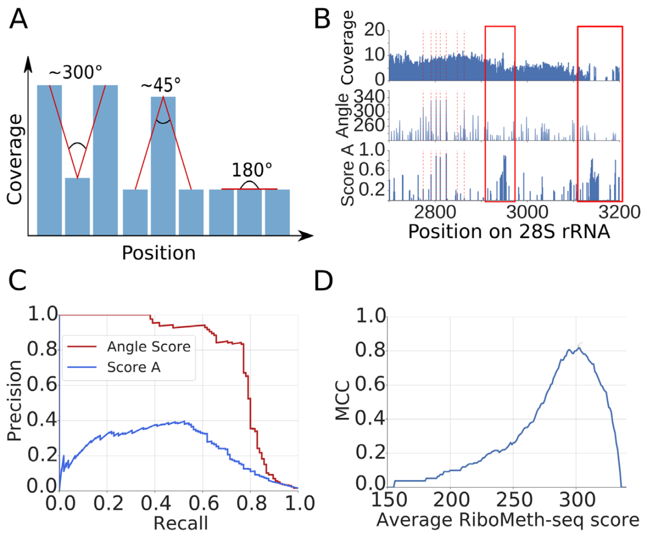


Figure 5. Analysis of RiboMeth-seq data. (**A**) Strategy for evaluating the RiboMeth-seq data. The score was calculated based on the normalized log2 coverage of a position and of its immediately adjacent neighbors by RiboMeth-seq reads. A large score indicates stronger depletion of the position by 3/5 ‘ ends of reads and thus resistance to alkaline hydrolysis. (**B**) Example of a normalized log2 coverage profile along 28S rRNA and calculated scores (Angle and Score A). With red dashed lines positions of known 2′-*O*-methylation sites are indicated. The red rectangles indicates regions where no 2′-*O*-methylation has been mapped, which is also predicted by the angle score but not by score A. (**C**) Example of Precision-Recall curves obtained for the two scoring methods applied to rRNAs from the RiboMethSeq_HEK_totalRNA_8min experiment. (**D**) Matthews correlation coefficient (MCC) plot of average RiboMeth-seq score indicating the optimal angle score.

DISCUSSION

High-throughput sequencing of samples prepared from cells that underwent various treatments have enabled the characterization of transcriptomes at ever increasing depth and resolution. This lead to the realization that the non-coding transcriptome is as large as the protein-coding fraction (64). New members of all classes of RNAs, including miRNAs and snoRNA have also been discovered (65, 66). The large number of novel molecular species increased the need for functional characterization methods, ideally in high-throughput. The aim of our study was to provide such methods for a specific class of non-coding RNAs, the C/D-box snoRNAs.

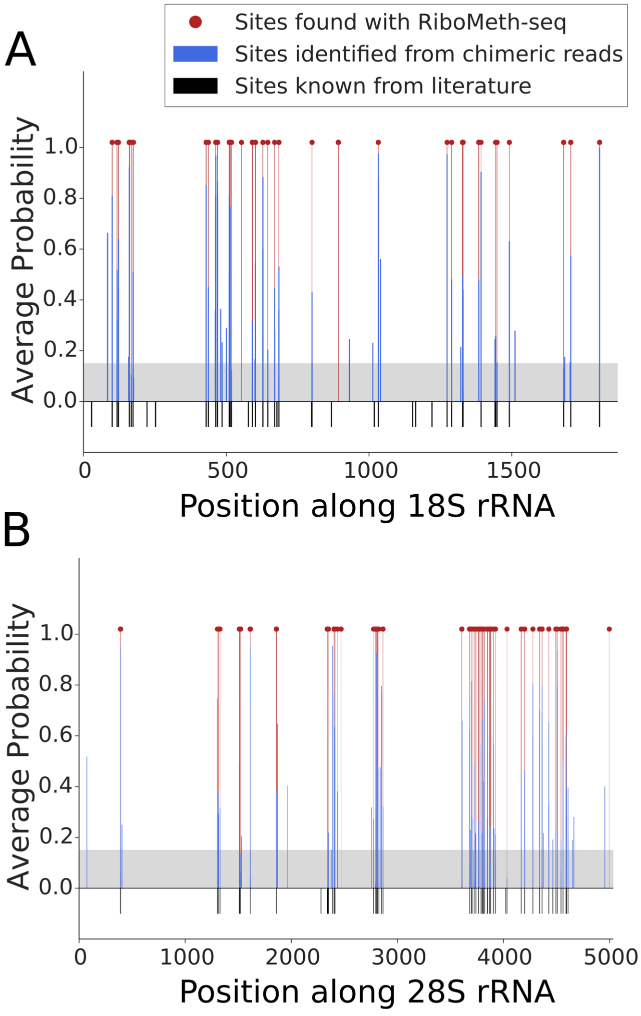


Figure 6. Location of snoRNA interaction sites and 2′-*O*-ribose methylation in the (**A**) 18S and (**B**) 28S ribosomal subunits. 2′-*O*-Me positions that are known from literature are shown as black bars. Interaction sites identified from chimeric reads are shown as blue bars, with their associated probabilities. The gray area indicates the score threshold that we used to extract the high-confidence sites from chimeric reads. The locations of 2′-*O*-Me sites identified with RiboMeth-seq are shown with red lines and dots.

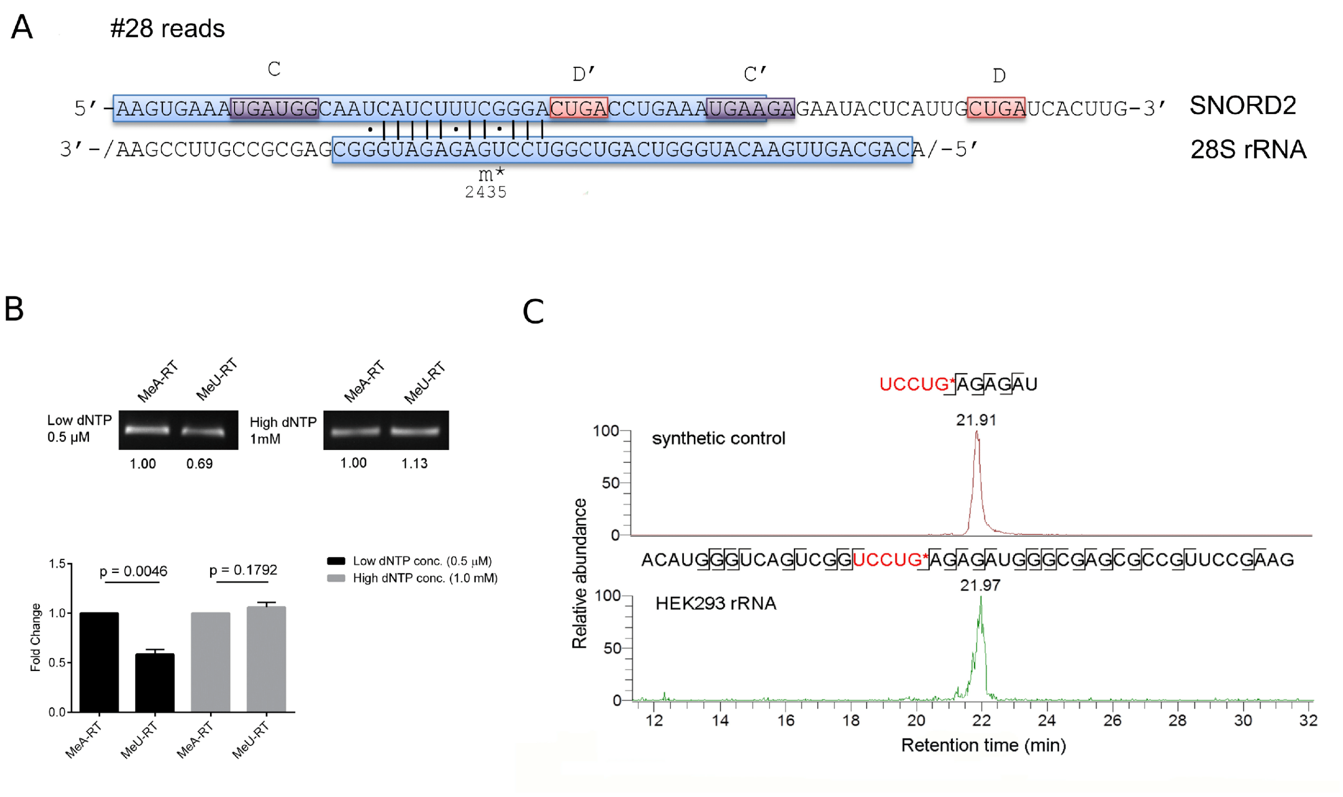


Figure 7. SNORD2-guided 2′-*O*-methylation of G2435 in the 28S rRNA (**A**) Schematic representation of the predicted interaction, which is supported by 28 chimeric reads (see also legend of Figure 3). (**B**) Confirmation of the G2435 2′-*O*-methylation by RTL-P followed by agarose gel analysis and followed by qPCR analysis. Error bars represent the standard deviation of the mean, and the *P* -value of the t-test computed over three replicate experiments, each with three technical replicates is indicated. (**C**) Targeted LC-MS/MS analysis of UCCUG*, confirming the 2′-*O*-methylation at G2435. A synthetic RNA oligonucleotide control (on top) and fragment A2416-G2461 from 28S rRNA (at the bottom) were digested with RNase T1 and specific transitions measured by targeted mass spectrometry.

We have combined two high-throughput approaches, the first aiming to identify direct interactions between snoRNAs and targets and the second to map sites of 2′-*O*-methylation transcriptome-wide. The first approach is based on the observation that chimeric reads, resulting from the ligation of a guide RNA to its target by endogenous ligases, are generated during CLIP (37). Whether CLIP of core snoRNP proteins can be used to identify snoRNA targets has not been investigated so far. Due to the low frequency of chimeric sequences (less than a percent of the reads (37)), the large 'background' of CLIP (48), and the short length of the snoRNA and target fragments that are captured, a snoRNA-centric analysis, taking into account the specific base-pairing pattern of snoRNAs with targets, is necessary. We found that a model that uses the predicted energy of interaction between the snoRNA and target, the accessibility of the target site and the A nucleotide context of the regions flanking the putative site, can identify over 70% of the known 2′-*O*-Me sites in rRNAs, with similar specificity. The model assigns SNORD30 as guide for the 'orphan' A1383 site in the 18S rRNA, and identifies an interaction between the SNORD118 snoRNA, so far known to be involved in pre-rRNA processing (31), with G1612 in the 28S rRNA, whose methylation is guided by SNORD80. The multi-copy nature of many of the 'orphan' snoRNAs, other homologies that they have in the genome, and the presence of crosslinking-induced mutations in the CLIP data pose substantial challenges to the identification of their targets and will benefit from an increase in the length of the reads generated with CLIP.

The model also predicted 40 novel interactions with rRNAs as well as many outside of structural RNAs. To evaluate 2′-*O*-methylation at these sites we implemented the RiboMeth-seq method (67). Although with this method we were able to recover the majority of known methylation sites, we did not find support for 2′-*O*-methylation of any novel sites in rRNAs. To determine whether these results are partly due to the limited sensitivity of RiboMeth-seq, we used low-throughput methods to evaluate 2′-*O*-methylation at position G2435 site in the 28S rRNA, which was supported by chimeric read data from four experiments. Both RTL-P and mass spectrometry provided evidence for 2′-*O*-methylation at this site. These data, as well as a closer inspection of the RiboMeth-seq scores of this site in individual experiments, indicate that the site is only partially methylated. The cause and consequences of partial methylation at rRNA sites will be fascinating topics for future studies, as the evidence for partial and cell type-specific methylation of rRNAs is mounting (39,44). Of note, the interaction of SNORD48 with C1868 in the 28S rRNA, presumed to lead to the observed partial methylation of this site (39) was also captured in our chimeric read data. Another possibility to consider is that the CLIP-derived chimera provide evidence for snoRNA-rRNA interactions that are relevant for rRNA processing but not 2′-*O*-methylation. Indeed, it has been proposed that the ancestral function of snoRNAs was in rRNA processing, a function that is still preserved in the U3, U8, U13, and U14 snoRNAs (26,30–32,68,69). Because the corresponding snoRNA-interacting sites may also need to be structurally accessible and have low-energy interaction with the snoRNAs, and because the D/D’ box sequences are short and not perfectly conserved in sequence, our method may misclassify these sites as 2′-*O*-methylation sites. Because PLEXY enforces the snoRNA interaction with the target to take place close to already annotated D boxes, we do not expect such cases in our final list of candidates. However, a careful inspection of the hybrids and chimeric read alignments that we provide on the accompanying web site should help identify these cases.

Although the chimeric read data suggested some interactions of snoRNAs with mRNAs, we were not able to validate these with RiboMeth-seq. This could be due to the much lower expression of the mRNAs compared to rRNAs, which makes the reliable detection of troughs in read coverage difficult. However, the RTL-P method also failed to provide evidence of 2′-*O*-methylation at mRNA sites (not shown). Thus, these sites may be the result of spurious ligation events. Alternatively, the snoRNA interaction with these sites may have other outcomes than 2′-*O*-methylation. Consistent with this hypothesis, a recent study that analyzed globally RNA–RNA interactions also found many interactions of snoRNAs with mRNAs and further demonstrated a function of SNORD83B in controlling the level of its target mRNAs (70).

Finally, RiboMeth-seq revealed a few high-confidence sites for which we did not find any corresponding chimeric reads. The low rate of capture of interactions in the chimeric reads may account for this observation. Alternatively, the RiboMeth-seq-documented sites may be resistant to alkaline hydrolysis for reasons other than 2′-*O*-Me. Supporting this latter hypothesis, these sites are generally located in rRNAs or snRNAs, molecules that are extensively modified and highly structured. In contrast to the known modification sites in rRNAs, which do not exhibit any nucleotide bias, the new sites recovered by RiboMeth-seq show a strong G-bias (not shown). This could again indicate that these sites are spurious or that modifications are introduced at these sites by specific enzymes such as the transfer RNA methyltransferase 7 protein (71). Interestingly, a recent study reported that G3771 in the 28S rRNA is 2′-*O*-methylated, guided by SNORD15A (39). Although we also find strong evidence for the methylation of this site in our RiboMeth-seq data, we did not find chimeric read evidence for SNORD15A acting as guide at this site. Rather, our chimeric read data supports a previous prediction (72) that SNORD15A guides the methylation at A3764 in the 28S rRNA.

Our study thereby provides computational methods that enable the mapping of snoRNA–target interactions in high-throughput. We believe that the application of these two complementary and high-throughput approaches, namely interaction capture via CLIP-seq and RiboMeth-seq will accelerate the accurate assignment of snoRNA guides to already mapped as well as newly discovered sites of 2′-*O*-methylation across cell types. This is especially relevant for studying the landscape of rRNA modification, which seems to be much more dynamic than anticipated, and for extending the study of snoRNA-guided methylation beyond species such as yeast and human.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

R.G. and D.J.J. would like to thank the members of the Zavolan lab for input on the project.

FUNDING

Marie Curie Initial Training Network RNPnet project [289007 to R.G.] from the European Commission; Marie Curie Initial Training Network RNAtrain project [607720 to F.G.] from the European Commission. Funding for open access charge: Core funding.

Conflict of interest statement. None declared.

REFERENCES

1. Li,S. and Mason,C.E. (2014) The pivotal regulatory landscape of RNA modifications. *Annu. Rev. Genomics Hum. Genet.*, **15**, 127–150.
2. Birkedal,U., Christensen-Dalsgaard,M., Krogh,N., Sabarinathan,R., Gorodkin,J. and Nielsen,H. (2014) Profiling of ribose methylations in RNA by high-throughput sequencing. *Angew. Chem. Int. Ed. Engl.*, 10.1002/anie.201408362.
3. Dominissini,D., Moshitch-Moshkovitz,S., Schwartz,S., Salmon-Divon,M., Ungar,L., Osenberg,S., Cesarkas,K., Jacob-Hirsch,J., Amariglio,N., Kupiec,M. *et al.* (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **485**, 201–206.
4. 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. *et al.* (2014) Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*, **159**, 148–162.
5. Schaefer,M. (2015) RNA 5-methylcytosine analysis by bisulfite sequencing. *Methods Enzymol.*, **560**, 297–329.
6. Saletore,Y., Meyer,K., Korlach,J., Vilfan,I.D., Jaffrey,S. and Mason,C.E. (2012) The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.*, **13**, 175.
7. Lee,M., Kim,B. and Kim,V.N. (2014) Emerging roles of RNA modification: m (6)A and U-tail. *Cell*, **158**, 980–987.
8. Sun,W.-J., Li,J.-H., Liu,S., Wu,J., Zhou,H., Qu,L.-H. and Yang,J.-H. (2015) RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res.*, 10.1093/nar/gkv1036.
9. Karijolich,J., Kantartzis,A. and Yu,Y.-T. (2010) RNA modifications: a mechanism that modulates gene expression. *Methods Mol. Biol.*, **629**, 1–19.
10. Heilman,K.L., Leach,R.A. and Tuck,M.T. (1996) Internal 6-methyladenine residues increase the in vitro translation efficiency of dihydrofolate reductase messenger RNA. *Int. J. Biochem. Cell Biol.*, **28**, 823–829.
11. Tuck,M.T., Wiehl,P.E. and Pan,T. (1999) Inhibition of 6-methyladenine formation decreases the translation efficiency of dihydrofolate reductase transcripts. *Int. J. Biochem. Cell Biol.*, **31**, 837–851.

12. Wang,X., Lu,Z., Gomez,A., Hon,G.C., Yue,Y., Han,D., Fu,Y., Parisien,M., Dai,Q., Jia,G. *et al.* (2014) N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, **505**, 117–120.
13. Wang,X., Zhao,B.S., Roundtree,I.A., Lu,Z., Han,D., Ma,H., Weng,X., Chen,K., Shi,H. and He,C. (2015) N (6)-methyladenosine modulates messenger RNA translation efficiency. *Cell*, **161**, 1388–1399.
14. Tycowski,K.T., You,Z.H., Graham,P.J. and Steitz,J.A. (1998) Modification of U6 spliceosomal RNA is guided by other small RNAs. *Mol. Cell*, **2**, 629–638.
15. Dennis,P.P., Omer,A. and Lowe,T. (2001) A guided tour: small RNA function in Archaea. *Mol. Microbiol.*, **40**, 509–519.
16. Zemann,A., op de Bekke,A., Kiefmann,M., Brosius,J. and Schmitz,J. (2006) Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res.*, **34**, 2676–2685.
17. Tollervey,D., Lehtonen,H., Jansen,R., Kern,H. and Hurt,E.C. (1993) Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly. *Cell*, **72**, 443–457.
18. Watkins,N.J., Véronique,S., Bruno,C., Stephanie,N., Patrizia,F., Angela,B., Matthias,W., Michael,R., Christiane,B. and Reinhard,L. (2000) A common core RNP structure shared between the small nucleoar box C/D RNPs and the Spliceosomal U4 snRNP. *Cell*, **103**, 457–466.
19. Gautier,T., Bergès,T., Tollervey,D. and Hurt,E. (1997) Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis. *Mol. Cell. Biol.*, **17**, 7088–7098.
20. Quinternet,M., Marc,Q., Marie-Eve,C., Benjamin,R., Decebal,T., Bruno,C. and Xavier,M. (2016) Structural features of the box C/D snoRNP Pre-assembly process are conserved through species. *Structure*, **24**, 1693–1706.
21. Jády,B.E. and Kiss,T. (2001) A small nucleolar guide RNA functions both in 2′-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. *EMBO J.*, **20**, 541–551.
22. Caffarelli,E., Fatica,A., Prislei,S., De Gregorio,E., Fragapane,P. and Bozzoni,I. (1996) Processing of the intron-encoded U16 and U18 snoRNAs: the conserved C and D boxes control both the processing reaction and the stability of the mature snoRNA. *EMBO J.*, **15**, 1121–1131.
23. van Nues,R.W., Sander,G., Grzegorz,K., Sloan,K.E., Matthew,C., David,T. and Watkins,N.J. (2011) Box C/D snoRNP catalysed methylation is aided by additional pre-rRNA base-pairing. *EMBO J.*, **30**, 2420–2430.
24. Tollervey,D., Lehtonen,H., Carmo-Fonseca,M. and Hurt,E.C. (1991) The small nucleolar RNP protein NOP1 (fibrillarin) is required for pre-rRNA processing in yeast. *EMBO J.*, **10**, 573–583.
25. Tollervey,D. and Kiss,T. (1997) Function and synthesis of small nucleolar RNAs. *Curr. Opin. Cell Biol.*, **9**, 337–342.
26. Lafontaine,D.L. and Tollervey,D. (1998) Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. *Trends Biochem. Sci.*, **23**, 383–388.
27. Fatica,A. and Tollervey,D. (2003) Insights into the structure and function of a guide RNP. *Nat. Struct. Biol.*, **10**, 237–239.
28. Kehr,S., Bartschat,S., Stadler,P.F. and Tafer,H. (2011) PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics.*, **27**, 279–280.
29. Chen,C.-L., Perasso,R., Qu,L.-H. and Amar,L. (2007) Exploration of pairing constraints identifies a 9 base-pair core within box C/D snoRNA–rRNA duplexes. *J. Mol. Biol.*, **369**, 771–783.
30. Kass,S., Tyc,K., Steitz,J.A. and Sollner-Webb,B. (1990) The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing. *Cell*, **60**, 897–908.
31. Peculis,B.A. and Steitz,J.A. (1993) Disruption of U8 nucleolar snRNA inhibits 5.8S and 28S rRNA processing in the Xenopus oocyte. *Cell*, **73**, 1233–1245.
32. Cavaillé,J., Hadjiolov,A.A. and Bachellerie,J.P. (1996) 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**, 206–213.
33. Li,H.D., Zagorski,J. and Fournier,M.J. (1990) Depletion of U14 small nuclear RNA (snR128) disrupts production of 18S rRNA in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **10**, 1145–1152.
34. Kudla,G., Granneman,S., Hahn,D., Beggs,J.D. and Tollervey,D. (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10010–10015.
35. Maden,B.E. (2001) Mapping 2′-O-methyl groups in ribosomal RNA. *Methods*, **25**, 374–382.
36. Helwak,A., Kudla,G., Dudnakova,T. and Tollervey,D. (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, **153**, 654–665.
37. Grosswendt,S., Filipchyk,A., Manzano,M., Klironomos,F., Schilling,M., Herzog,M., Gottwein,E. and Rajewsky,N. (2014) Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol. Cell.*
38. Breda,J., Rzepiela,A.J., Gumienny,R., van Nimwegen,E. and Zavolan,M. (2015) Quantifying the strength of miRNA–target interactions. *Methods*, **85**, 90–99.
39. Krogh,N., Jansson,M.D., Häfner,S.J., Tehler,D., Birkedal,U., Christensen-Dalsgaard,M., Lund,A.H. and Nielsen,H. (2016) Profiling of 2′-O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. *Nucleic Acids Res.*, 10.1093/nar/gkw482.
40. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.
41. Jády,B.E., Ketele,A. and Kiss,T. (2012) Human intron-encoded Alu RNAs are processed and packaged into Wdr79-associated nucleoplasmic box H/ACA RNPs. *Genes Dev.*, **26**, 1897–1910.
42. Kishore,S., Gruber,A.R., Jedlinski,D.J., Syed,A.P., Jorjani,H. and Zavolan,M. (2013) Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol.*, **14**, R45.
43. Machyna,M., Kehr,S., Straube,K., Kappei,D., Buchholz,F., Butter,F., Ule,J., Hertel,J., Stadler,P.F. and Neugebauer,K.M. (2014) The coilin interactome identifies hundreds of small noncoding RNAs that traffic through Cajal bodies. *Mol. Cell*, **56**, 389–399.
44. Jorjani,H., Kehr,S., Jedlinski,D.J., Gumienny,R., Hertel,J., Stadler,P.F., Zavolan,M. and Gruber,A.R. (2016) An updated human snoRNAome. *Nucleic Acids Res.*, 10.1093/nar/gkw386.
45. Martin,G., Gruber,A.R., Keller,W. and Zavolan,M. (2012) Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. *Cell Rep.*, **1**, 753–763.
46. Chan,P.P. and Lowe,T.M. (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, **37**, D93–D97.
47. Khorshid,M., Rodak,C. and Zavolan,M. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **39**, D245–D252.
48. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.-C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
49. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
50. Tafer,H. and Hofacker,I.L. (2008) RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics*.
51. Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
52. Seabold,S. and Perktold,J. (2010) Statsmodels: Econometric and statistical modeling with python. *Of the 9th Python in Science Conference*.
53. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
54. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
55. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10.

56. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
57. Maden,B.E., Corbett,M.E., Heeney,P.A., Pugh,K. and Ajuh,P.M. (1995) Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA. *Biochimie*, **77**, 22–29.
58. Dong,Z.-W., Shao,P., Diao,L.-T., Zhou,H., Yu,C.-H. and Qu,L.-H. (2012) RTL-P: a sensitive approach for detecting sites of 2′-O-methylation in RNA molecules. *Nucleic Acids Res.*, **40**, e157.
59. Andersen,T.E., Porse,B.T. and Kirpekar,F. (2004) A novel partial modification at C2501 in Escherichia coli 23S ribosomal RNA. *RNA*, **10**, 907–913.
60. Bauer,M., Ahrné,E., Baron,A.P., Glatter,T., Fava,L.L., Santamaria,A., Nigg,E.A. and Schmidt,A. (2014) Evaluation of data-dependent and -independent mass spectrometric workflows for sensitive quantification of proteins and phosphorylation sites. *J. Proteome Res.*, **13**, 5973–5988.
61. Gumienny,R. and Zavolan,M. (2015) Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Res.*, 10.1093/nar/gkv050.
62. Maden,B.E. (1986) Identification of the locations of the methyl groups in 18 S ribosomal RNA from Xenopus laevis and man. *J. Mol. Biol.*, **189**, 681–699.
63. Tycowski,K.T., Shu,M.D. and Steitz,J.A. (1996) A mammalian gene with introns instead of exons generating stable RNA products. *Nature*, **379**, 464–466.
64. Mattick,J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, **2**, 986–991.
65. Morin,R.D., O'Connor,M.D., Griffith,M., Kuchenbauer,F., Delaney,A., Prabhu,A.-L., Zhao,Y., McDonald,H., Zeng,T., Hirst,M. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.
66. Chen,H.-M. and Wu,S.-H. (2009) Mining small RNA sequencing data: a new approach to identify small nucleolar RNAs in Arabidopsis. *Nucleic Acids Res.*, **37**, e69.
67. Birkedal,U., Christensen-Dalsgaard,M., Krogh,N., Sabarinathan,R., Gorodkin,J. and Nielsen,H. (2015) Profiling of ribose methylations in RNA by high-throughput sequencing. *Angew. Chem. Int. Ed. Engl.*, **127**, 461–465.
68. Tollervey,D. (1987) A yeast small nuclear RNA is required for normal processing of pre-ribosomal RNA. *EMBO J.*, **6**, 4169–4175.
69. Enright,C.A., Maxwell,E.S., Eliceiri,G.L. and Sollner-Webb,B. (1996) 5′ETS rRNA processing facilitated by four small RNAs: U14, E3, U17, and U3. *RNA*, **2**, 1094–1099.
70. Sharma,E., Sterne-Weiler,T., O'Hanlon,D. and Blencowe,B.J. (2016) Global mapping of human RNA-RNA interactions. *Mol. Cell*, **62**, 618–626.
71. Guy,M.P., Shaw,M., Weiner,C.L., Hobson,L., Stark,Z., Rose,K., Kalscheuer,V.M., Gecz,J. and Phizicky,E.M. (2015) Defects in tRNA anticodon loop 2′-O-methylation are implicated in nonsyndromic X-linked intellectual disability due to mutations in FTSJ1. *Hum. Mutat.*, **36**, 1176–1187.
72. Kiss-László,Z., Henry,Y., Bachellerie,J.P., Caizergues-Ferrer,M. and Kiss,T. (1996) Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, **85**, 1077–1088.

3.1 Author’s Contributions

RG, DJJ, MZ conceived the project. DJJ, RG, and MZ co-wrote the manuscript. RG performed all bioinformatic analyses with help of FG. DJJ performed all experiments with help of AVV (RiboMeth-seq) and AS (Mass Spectrometry) and GM (CLIP-seq). All authors read and approved the final MS.

3.2 Electronic Supplementary Information

For Supplementary Figures and Tables please consult the online version of the manuscript PMID.

Chapter 4 An updated human snoRNAome

An updated human snoRNAome (Published in Nucleic Acids Research)

Hadi Jorjani[1,†], Stephanie Kehr[2,†], Dominik J. Jedlinski[1], Rafal Gumienny[1], Jana Hertel[2], Peter F. Stadler[2], Mihaela Zavolan[1], Andreas R. Gruber[1]

[1] Computational and Systems Biology, Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel CH-4056, Switzerland.
[2] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, D-04107 Leipzig, Germany.

[†]Contributed equally

Keywords: C/D box snoRNA, H/ACA box snoRNA, scaRNA, non-coding RNAs, snoRNA-derived fragments, gene annotation, target prediction

4.1 Abstract

Small nucleolar RNAs (snoRNAs) are a class of non-coding RNAs that guide the post-transcriptional processing of other non-coding RNAs (mostly ribosomal RNAs), but have also been implicated in processes ranging from microRNA-dependent gene silencing to alternative splicing. In order to construct an up-to-date catalog of human snoRNAs we have combined data from various databases, de novo prediction and extensive literature review. In total, we list more than 750 curated genomic loci that give rise to snoRNA and snoRNA-like genes. Utilizing small RNA-seq data from the ENCODE project, our study characterizes the plasticity of snoRNA expression identifying both constitutively as well as cell type specific expressed snoRNAs. Especially, the comparison of malignant to non-malignant tissues and cell types shows a dramatic perturbation of the snoRNA expression profile. Finally, we developed a high-throughput variant of the reverse-transcriptase-based method for identifying 2’-O-methyl modifications in RNAs termed RiM-seq. Using the data from this and other high-throughput protocols together with previously reported modification sites and state-of-the-art target prediction methods we re-estimate the snoRNA target RNA interaction network. Our current results assign a reliable modification site to 83% of the canonical snoRNAs, leaving only 76 snoRNA sequences as orphan.

4.2 Introduction

Currently there are two known high-throughput methods that can be used to study 2’-O-ribose-methylation transcriptome-wide: CLIP-seq and RiboMeth-seq. CLIP-seq captures direct snoRNA-target interactions and RiboMeth-seq enables the identification of actual 2’-O-methyl modifications. Both of these methods were extensively described in CHAPTER 3.

In this chapter of the thesis I would like to address only the part that I have contributed to the study “**An updated human snoRNAome**”, which is the new experimental method that we named RiM-Seq. We applied RiM-Seq to validate computationally predicted sites of 2’-O-methylation. We began the development of RiM-seq before we adapted RiboMeth-seq to validate 2’-O-methylation sites and CLIP-seq to identify chimeras. RiM-Seq can be viewed as another, supplemental method to confirm 2’-O-methylation sites transcriptome-wide because its principle is different from those of RiboMeth-seq and CLIP-seq. RiM-Seq is essentially a high-throughput variant of the low-throughput reverse-transcriptase-based protocol which we used in the form of primer extension assays to validate individual 2'-O-Me's in various targets in CHAPTER 2.

4.3 RiM-seq

To validate predicted 2'-O-methylated residues transcriptome-wide we developed a high-throughput sequencing-based variant of the well-established, low-throughput reverse transcriptase-based protocol [156], which is usually coupled with polyacrylamide gel analysis. The method is based on the observation that cDNA synthesis is noticeably impaired in the presence of a 2'-O-methylation when deoxynucleoside triphosphate fragments (dNTPs) are limiting [156, 157] (FIGURE 4.1). This gives rise to a characteristic pattern of gel banding immediately preceding the 2'-O-Me, with strong bands at low dNTP concentrations (0.004mM) [157] becoming weaker with increasing concentrations of dNTPs (CHAPTER 2). These stoppages, which correspond to 2'-O-methylation sites, will generate read ends when RNA fragments are reverse-transcribed under different dNTP concentrations, ligated to adapters and sequenced. 2'-O-methyl positions can be subsequently identified by calculating the "stoppage ratio" of reads starting at a given position (5' ends) to the reads that cover that position (readthrough reads + 5' ends) and comparing this ratio between samples generated with low and normal, high dNTP concentrations.

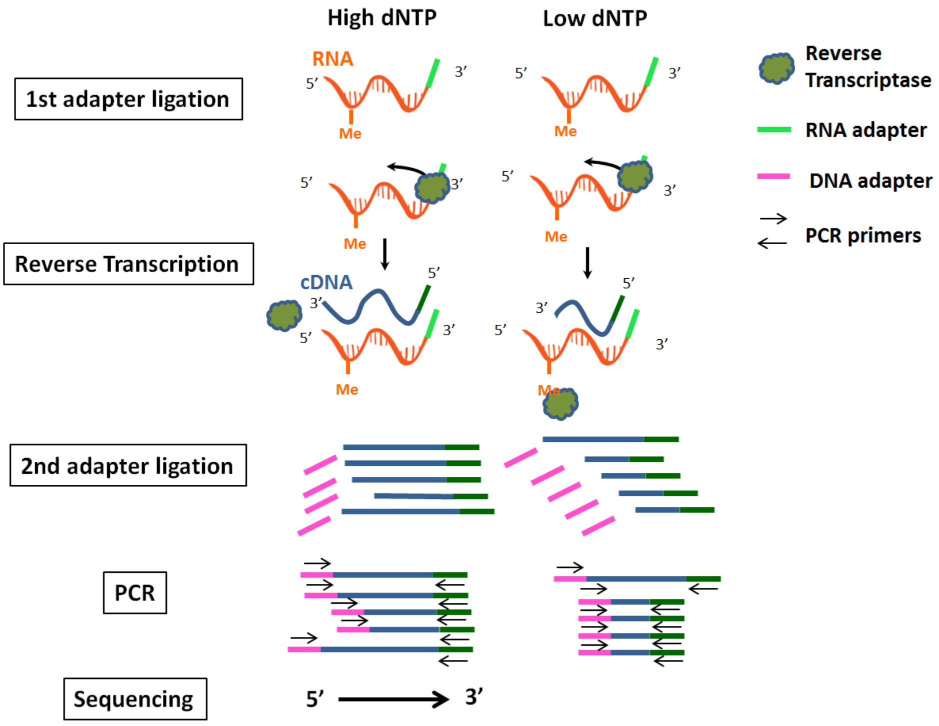


Figure 4.1 **RiM-seq protocol**. RNA is reverse transcribed under high (left) or low (right) dNTP concentration. When the dNTP concentration is low, reverse transcription in impaired in the presence of a 2'-O-methylation site.

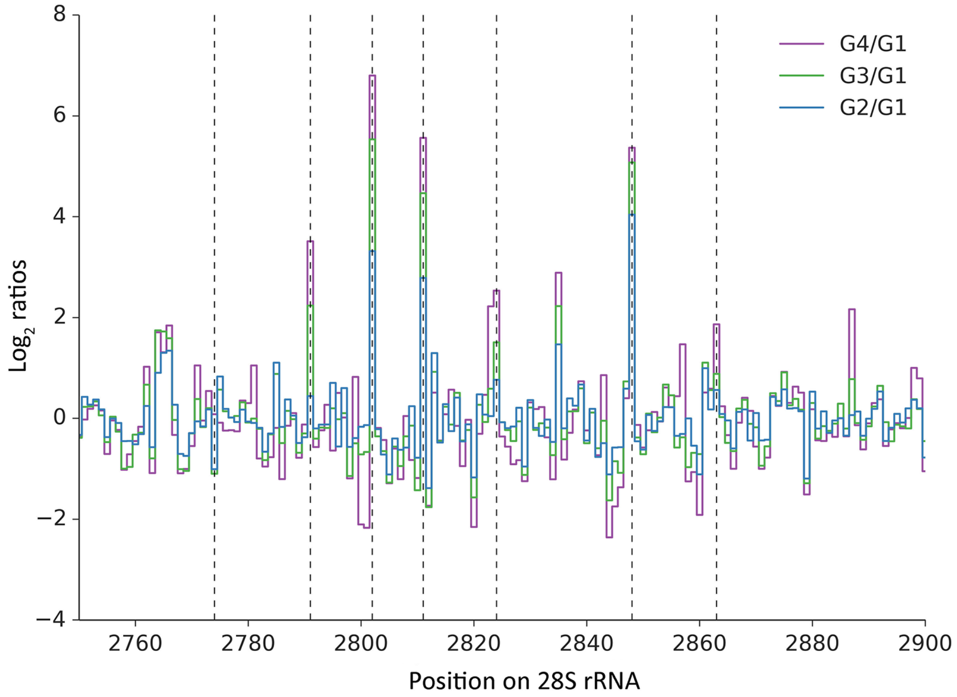


Figure 4.2 **RiM-seq readout**. Plot of three log$_2$ stoppage ratios along a fragment of 28S rRNA obtained from samples prepared with four different concentrations of nucleotides during primer extension (G1 being the highest dNTP concentration, G4 the lowest): concentration 4 to concentration 1 (G4/G1), concentration 3 to concentration 1 (G3G1) and concentration 2 to concentration 1 (G2G1). Dashed lines indicate known 2'-O-methylation sites from snoRNABase [95].

FIGURE 4.2 depicts a region of the 28S rRNA with sites of 2'-O methylation that were identified with RiM-Seq. Dashed lines indicate known 2'-O-methylation sites from the snoRNA-LBME-db [95]. G1 represents the stoppage ratio of the control sample that was reverse transcribed under normal dNTP concentrations (500µM), G4 represents the stoppage ratio of the sample that was reverse transcribed under the lowest dNTP concentrations (0.4 µM) (see TABLE 4.1). G2, G3, and G4 samples generated significantly more RT stoppages at 2'-O-methylation positions compared to the G1 control, which resulted in an increase of reads starting at this positions (5'ends). Also evident is that the stoppage ratio is correlated with the RT-dNTP concentrations as would be expected.

4.4 Methods

4.4.1 RiM-seq library preparation and sequencing protocol

Total RNA was extracted with TRI Reagent (Sigma) and mRNA was prepared with the Dynabeads mRNA DIRECT Kit (Life Technologies), both from HEK293 cells according to the manufacturer's instructions. For mapping of 2'-O-methyl sites in rRNA, 5µg of total RNA was used as starting material, whereas for mapping 2'-O-methyl sites in mRNA, 1µg of poly(A)-selected RNA was used. In both protocols, the RNA was degraded under alkaline conditions in a sodium carbonate/bicarbonate buffer at pH 9.2 for 5 minutes, put on ice and purified with RNeasy clean-up columns (QIAGEN). The sample was eluted in 12µl of water and dephosphorylated in a 14µl reaction volume using FastAP (life technologies) according to manufacturer's instructions (1.4µl FastAP buffer, 0.6µl RNasin Plus RNase Inhibitor (Promega), 1µl FastAP). 1µl of 40µM 3' RNA-adapter (5' 5'-p-UGGAAUUCUCGGGUGCCAAGG-amino-3) were added and the sample was denatured at 70°C for 2 minutes and then transferred on ice. A ligation mixture (5µl 10x NEB Ligase 1 buffer, 12 µl DMSO (50%), 16µl PEG 8000 (50%), 0.5µl RNasin Plus RNase inhibitor, 2µl T4 RNA Ligase 1 (30 U/µl)) was added and the sample was incubated 2 hours at 25°C. The sample was purified using RNeasy clean-up kit (QIAGEN) and eluted in 48µl H_2O. For reverse transcription, 1µl of 10µM RT primer (5'-GCCTTGGCAC CCAGAGAATTCCA-3') was added to 9µl of the RNA sample and the mixture was incubated at 72°C for 2 minutes and then cooled down on ice for 2 minutes. The reverse transcription reactions were set up as follows in TABLE 4.1:

Table 4.1 **Reverse transcription conditions used in RiM-seq**. G1-G4 (high dNTP conc-low NTP conc) are the different conditions used in RT.

Condition	**G1**	**G2**	**G3**	**G4**
dNTP Concentration	500µM	5µM	2µM	0.4µM
10mM dNTP Mix	**1µl**	-	-	-
0.1mM dNTP Mix	-	**1µl**	**0.4µl**	-
0.01mM dNTP Mix	-	-	-	**0.8µl**
Water	1µl	1µl	1.6ul	1.2µl
5x First Strand Buffer	4µl	4ul	4µl	4µl
100mM DTT	2µl	2µl	2µl	2µl
RNase Inhibitor	1µl	1µl	1µl	1µl
SuperScript III	1µl	1µl	1µl	1µl
SubTotal	10µl	10µl	10µl	10µl
Total	20µl	20µl	20µl	20µl

The reaction was incubated at 42°C for 20 min, 50°C for 20 min, and 55°C 20 min and then 4°C for at least one minute. Excess RT primers were digested by adding 3µl of ExoSAP-IT (Affymetrix) and incubated at 37°C for 12 minutes. The enzyme was inactivated by incubation at 80°C for 15 minutes. Subsequently the RNA was degraded by adding 2µl of RNase H (NEB) and incubating at 37°C for 15 minutes. The RNase H was inactivated by incubation at 65°C for 20 minutes and the sample placed on ice.

The cDNA (25µl) was cleaned up using 50µl of Agencourt AMPure XP beads (BECKMAN COULTER) with 25µl of isopropanol. The mixture was mixed well and incubated 5 minutes at room temperature. The beads were collected on a magnet and washed twice with 80% ethanol. The beads were dried on the magnet for 10 minutes and the cDNA was eluted in 20 µl of H_2O.

To each sample 1µl of 40µM 5'-DNA adapter (5'-p-GATCG TCGGA CTGTA GAACT CTGAA C-amino-3') was added and the samples were denatured at 75°C for 2 minutes and then transferred on ice. 39.5µl of the following ligation mix were added to every sample: 6µl 10X NEB Ligase 1 Buffer, 4.5µl DMSO (50%), 27µl PEG 8000 (50%), 2µl T4 RNA ligase 1 (30U/µl) and the mixture was vortexed and incubated at room temperature overnight. The cDNA (60µl) was again cleaned up adding 120µl of Agencourt AMPure beads with

30µl of isopropanol by incubating the mixture 5 minutes and collecting the beads on a magnet. The beads were washed twice with 80% ethanol and dried 10 minutes on the magnet. The cDNA was eluted in 60µl of H_2O.

5µl of the resulting cDNA were then used in a pilot PCR reaction. To this end, aliquots were taken from reactions at every second cycle between 12 and 22 cycles and analyzed on a 2.5 % agarose gel. The number of cycles causing a first visible amplification was chosen for a large scale PCR (10µl cDNA in a 100µl reaction). The PCR product was then cleaned up with the QIAquick PCR Purification Kit (QIAGEN) followed by a clean-up with Agencourt AMPure XP beads according to manufacturer's instructions. The PCR product was eluted in 25µl of H_2O. Libraries were sequenced on an Illumina HiSeq-2500 deep sequencer.

4.4.2 Read mapping and 2'-O-methylation sites extraction

We obtained ~40 million reads for each of the RiM-seq samples G1-G4 (four different dNTP concentrations). The adapter was cut with Cutadapt (--minimum-length 15, and other parameters set to default) [186] and we mapped the reads with STAR [187] (additional parameters: --outFilterMultimapNmax 20 --outFilterMismatchNoverLmax 0.05 --scoreGenomicLengthLog2scale 0 --outSAMattributes All) to the RNA transcriptome composed of rRNAs and snRNAs. Mapped reads were used to calculate coverage and the 5' ends of the reads were used to identify locations where the reverse transcription reaction had stopped. Putative 2'-O-methylation sites were detected with the following procedure: For each sample a stoppage ratio was calculated as the ratio between the number of reads starting at given position (reads with 5' end at that position) and the number of reads covering a specific position (reads with 5' end at that position + readthrough reads). To avoid division by 0 a pseudocount was added to both coverages. For each condition (G1-G4) a fold change was calculated as $\log_2$ of the ratio between the stoppage ratio at that concentration and the stoppage ratio in the control sample (G1). This gave three $\log_2$ ratios corresponding to decreasing concentrations of nucleotides: concentration 4 to concentration 1 (G4/G1), concentration 3 to concentration 1 (G3G1) and concentration 2 to

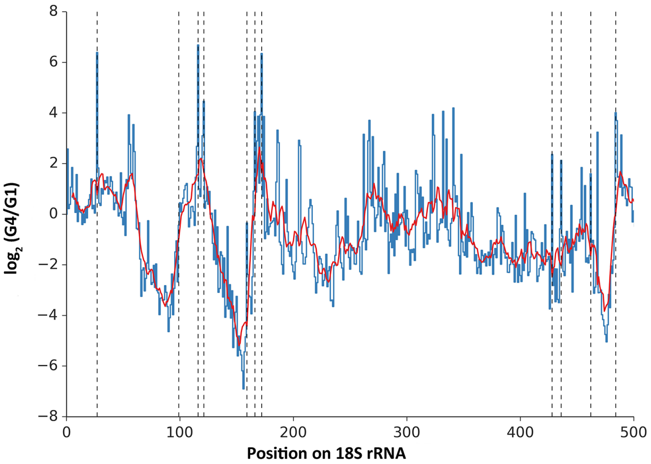


Figure 4.3 **Plot of the $\log_2$ G4-to-G1 stoppage ratios** (see the text for details) along the 18S rRNA (blue line) with local background calculated (red line). Dashed lines indicate 2'-O-methylation sites from snoRNABase.

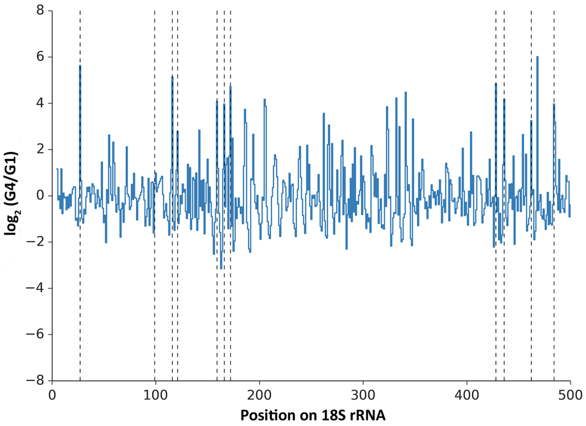


Figure 4.4 **Plot of the G4-to-G1 $\log_2$ stoppage ratio** along the 18S rRNA with local background subtracted. Dashed lines indicate 2'-O-methylation sites from snoRNABase.

concentration 1 (G2G1). Because the stoppage ratio seems to be strongly region-specific, we subtracted the local background defined as the mean stoppage ratio in the neighborhood of a given position (+/- 5 nucleotides, excluding the central position) see FIGURE 4.3 and FIGURE 4.4. According to the reverse transcription protocol there should be an increase in sample to control ratio with decreasing nucleotide concentration and this pattern was used as a first filtering step. As there is no indication that the ratio should increase linearly, we required that there is increase in $\log_2$ ratio going from G2/G1 to G3/G1 and from G3/G1 to G4/G1 i.e. G2/G1 < G3/G1 < G4/G1. Thus, only positions with increased ratios were kept for further analysis (FIGURE 4.2). We used as a score the $\log_2$ ratio between lowest and control concentration (G4/G1). The threshold for the G4/G1 ratio was deduced from the Matthews coefficient plot and is 3.63 (FIGURE 4.5).

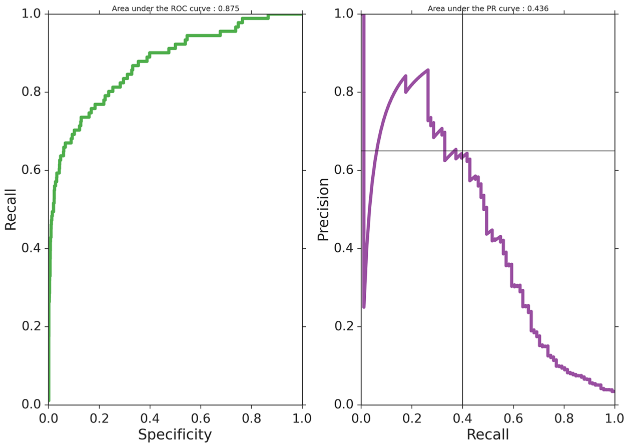


Figure 4.6 **ROC curve and PR curve** calculated for G4/G1 ratio as a score on 18S and 28S rRNAs.

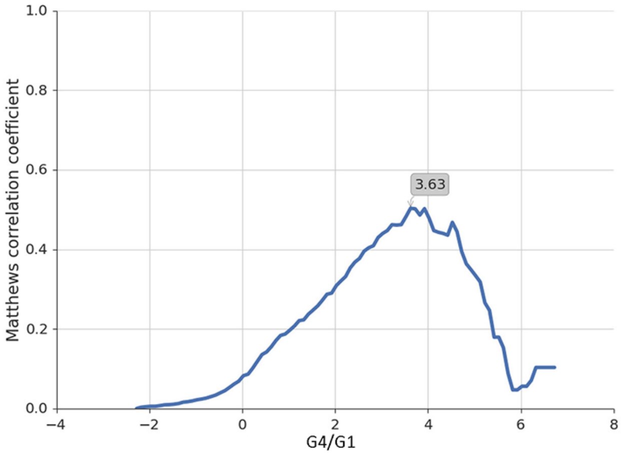


Figure 4.5 **Matthews correlation coefficient plot** for the G4/G1 ratio calculated on 18S and 28S rRNAs.

However, at this threshold only a small number of putative target sites we found and the recall of known rRNA sites was only 0.47 (precision 0.56, FIGURE 4.6). Because this experiment is used to validate computational predictions rather than discover stoppage sites de novo, we lowered the threshold to 2.0, where the precision is low but recall is ~70%.

4.5 Discussion of the RiM-seq method

The RiM-seq approach to detect 2’-O-methylation sites is based on a well-known procedure often used in primer extension assays. As such, it preserves all the advantages and drawbacks of this method. The main advantage of the method is its simplicity: it is easy to conduct and fairly easy to analyze. However, false positives can be generated by pausing induced by other types of modifications (although with much lower rate), by degradation of RNA at sensitive sites and by the secondary structure of the RNA [201]. One way to deal with this problem when performing gel analysis is to compare the results to those obtained with unmodified RNA [201]. This, however, is not possible with the high-throughput approach. There is also great variability in the $\log_2$ ratio of the stops - some of them very clear and some not visible at all. Additionally, there is a contribution of so called shading to the false negative rate. This corresponds to the situation when two sites are in close proximity and one of them cannot be reached due to strong pausing at the other site. The ROC curve calculated for rRNAs indicates that the approach allows one to distinguish between methylated and non-methylated sites. However, the precision-recall (PR) curve shows that this method does not have both high precision and recall (FIGURE 4.6). Nonetheless, we present this method as an alternative high-throughput method for the validation of computationally predicted sites of 2’-O-methylation, complementing the known approaches CLIP-seq and RiboMeth-seq (described in CHAPTER 3).

Chapter 5 Evaluation of a canonical snoRNA function of Prader-Willi syndrome-associated SNORD116

Evaluation of a canonical snoRNA function of Prader-Willi syndrome-associated SNORD116 (Manuscript in preparation)

Dominik J Jedlinski[1], Rafal Gumienny[1,2], Harald Witte[3], Foivos Gypas[1,2], Boris Skryabin[4], Mihaela Zavolan[1,2]
[1]Computational and Systems Biology, Biozentrum, University of Basel, Switzerland
[2]Swiss Institute of Bioinformatics
[3] Biozentrum, University of Basel, Switzerland
[4]Institute of Experimental Pathology, University of Münster, Germany

5.1 Introduction

Small nucleolar RNAs (snoRNAs) are well-characterized non-coding RNAs that localize in eukaryotic and archeal nucleoli where they primarily guide chemical modifications of target RNAs. These are mainly ribosomal RNA (rRNA) and small nuclear RNA (snRNA) [166], but other RNA species can also undergo snoRNA-guided modification [177], and the spectrum of snoRNA targets is not entirely characterized. Based on sequence and structural features, snoRNAs can be divided into the C/D box snoRNAs, which guide 2’-O-methylation, and H/ACA box snoRNAs, which guide pseudouridylation of their target RNA. During their biogenesis, which involves processing from excised and debranched introns by exonucleolytic trimming [75, 202], snoRNAs associate with evolutionary conserved proteins to form the enzymatically active small ribonucleoprotein complex (snoRNP). The core proteins of the C/D box snoRNP are 15.5K, NOP56, NOP58, and Fibrillarin [74, 104, 105], the latter being the enzyme that catalyzes the site-specific transfer of a methyl group from S-adenosylmethionine onto the 2’-hydroxy group of the ribose. The target site specificity is defined by a 6-20 nucleotides-long region known as the anti-sense box and located upstream of the D or D’ box of the snoRNA, which pairs with the target site. The target nucleotide that pairs with the 5th nucleotide upstream of the D box undergoes methylation [109-111]. Computational prediction of snoRNA targets also makes use of these properties. The number of target sites that can be predicted based on these principles in the entire genome vastly exceeds the number of sites where modifications have been described. However, additional determinants seem difficult to uncover [73, 102, 107, 173]. Furthermore, a considerable number of C/D box snoRNAs that carry the conserved C/C’ (RUGAUGA, R = A or G) and D/D’ (CUGA) boxes remain “orphan”, without a known RNA target [203, 204]. Recent studies have described atypical characteristics and expression patterns of orphan snoRNAs, as well as unexpected associations with RNA-binding proteins, which suggest that snoRNAs may have a broader set of functions than currently known [94, 205].

Among the orphan snoRNAs, a cluster located on chromosome 15q11-13 has been extensively studied due to their association with the Prader-Willi syndrome (PWS), a congenital disease marked by excessive appetite and morbid obesity as well as mental and growth retardation. The molecular basis of PWS is unknown, but genome analysis of a number of patients as well as studies in animal models implicate the loss of paternal expression of maternally-imprinted chromosome 15q11-13 genes - caused by non-inherited deletions, uniparental disomy, or imprinting defects - as the cause of PWS [206]. The chromosome 15q11-13 locus comprises multi-copy snoRNAs and several protein-coding genes, but the latter do not appear to contribute to the development of the syndrome [206, 207]. The snoRNAs SNORD64, SNORD107, SNORD108, SNORD109A, SNORD115, and SNORD116 are processed from the introns of a non-protein coding transcript (U-UBE3A-ATS) that is transcribed from this locus [208]. SNORD115 and SNORD116 are multicopy (47, and 29, respectively) snoRNAs, well-conserved in mammals [209, 210], highly expressed in the brain [211]. Initially, SNORD115 seemed a good candidate PWS gene, because it has a striking complementarity to the serotonin receptor 2C pre-mRNA, whose splicing it may regulate, potentially contributing to the development of obesity [92, 212]. However, other studies in human [213] and mouse [214] essentially ruled out SNORD115 as a main contributor to PWS. In particular, Runte and colleagues [213] demonstrated that family members of Angelman syndrome patients who carried paternal deletions of the whole SNORD115 cluster did not exhibit an obvious clinical phenotype. Genetic studies have described several microdeletions in the 15q11-13 locus in PWS patients and the data are notably in favor of the interpretation that paternal deficiency of SNORD116 alone is sufficient to cause the key manifestations of the PWS phenotype [98, 99, 215, 216]. The question whether other genes in the region, particularly the lncRNA IPW116 that hosts SNORD116, may aggravate the PWS phenotype, remains open [217, 218].

In mouse, the PWS locus is located on chromosome 7. Based on the findings that implicated SNORD116 in the pathogeny of PWS in humans, two groups generated mouse models of the disease by deleting the PWS critical region (PWS*cr*) that spans the Snord116 cluster and IPW exons A1/A2, B and C [219, 220]. Previously mouse models of PWS had mostly larger deletions of several paternally-expressed, imprinted genes [221-224]. Mice that inherited the deleted allele maternally (PWScr$^{m-/p+}$) showed no phenotypic abnormalities and were indistinguishable from wild-type (wt) littermates. However, mice that inherited the allele paternally (PWScr$^{m+/p-}$) displayed postnatal growth retardation, delayed sexual maturation, increased anxiety, motor learning deficit, hyperphagia, and hyperghrelinemia, thus recapitulating a subset of the PWS symptoms [219, 220] (FIGURE 5.1). Features of human PWS that were not observed in the mouse models are hypotonia, obesity, and decreased sensitivity to pain. The authors of the respective studies argued that the PWS may have species-specific components [220].

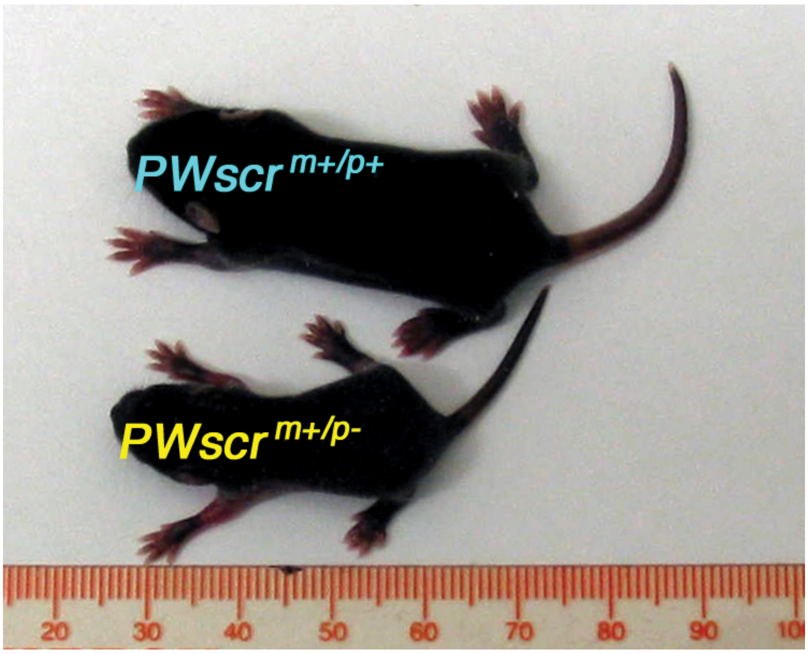


Figure 5.1 **Growth Differences among PWScr$^{m+/p-}$ and PWScr$^{m+/p+}$ siblings.** Representative pair of mice from the same litter at postnatal day 10 (129SV x C57BL/6 genetic crosses). Figure from [219].

Although a general consensus that PWS is caused primarily by the lack of SNORD116 expression has seemingly been reached, the molecular function of SNORD116 has remained elusive. Mouse Snord115 and Snord116 yield a variety of processed snoRNAs (psnoRNAs) and some studies suggested that they modulate alternative splicing of multiple transcripts [122, 128]. MiRNA-like functions have also been attributed to snoRNA-derived fragments (reviewed in [94]). The question of which type of molecule is predominantly generated from the Snord115 and Snord116 loci has been addressed in a few studies, with the conclusion that it is a classical C/D box-containing snoRNA [225] that associates with core C/D box snoRNP proteins [93]. Furthermore, incorporation of psnoRNA-type products into Argonaute 2 proteins in human cells was very limited [177], arguing against a miRNA-like role for psnoRNAs in human.

Yin and colleagues also described a long non-coding RNA (lncRNA), sno-lncRNAs, derived from the SNORD116 locus and demonstrated its involvement in PWS pathogenesis [205]. Interestingly, the ends of this sno-lncRNA correspond precisely to two SNORD116 that are embedded in the same intron. Generation of the sno-lncRNA likely involves the snoRNP biogenesis machinery, which would explain why these transcripts are found complexed to snoRNP proteins. It has been proposed that these interactions stabilize the sno-lncRNAs by inhibiting exonucleolytic degradation [205]. Interestingly, five sno-lncRNAs are generated from the PWS locus, each containing multiple binding sites for the splicing regulator RBFOX2. It has been suggested that the sno-lncRNAs act as molecular sinks, titrating RBFOX proteins in specific cells and thereby inducing subtle alterations in the splicing of RBFOX-regulated exons. Given that the expression of sno-lncRNAs was found to be highest in pluripotent stem cells [225], RBFOX-dependent splicing alterations are expected in the early embryonic development of PWS patients.

In this study we use a broad set of assays to evaluate possible functions for SNORD116 snoRNAs in the brain, the tissue in which they are most highly expressed. We first considered that Snord116 or sno-lncRNAs from the PWS region may alter patterns of splicing of Rbfox2-regulated targets in brain, as they appear to do in stem cells and ovarian carcinoma PA1 cells [205]. Analyzing the impact of Snord116-deletion (in PWScr$^{m+/p-}$ compared to wt mice) on alternative splicing in the brain with mRNA-seq, we found neither substantial changes in splicing nor an enrichment in splicing changes in Rbfox2 targets compared to other genes. Next we re-evaluated the possibility that these snoRNAs have a canonical function, comprehensively mapping 2'-O-methylated riboses in wt mice and PWScr$^{m+/p-}$ mice that do not express Snord116 with the RiboMeth-Seq method [65]. We identified several promising targets sites that appear to be differentially methylated in the two conditions. To determine whether these could be direct targets of Snord116 snoRNAs, we then performed snoRNP-Crosslinking and Immunoprecipitation (CLIP) experiments in several neuronal cell lines as well as in mouse primary neurons (cerebellar granule cells). Although we recapitulated previous findings that Snord116 snoRNAs associate with the snoRNP machinery, especially in primary neurons, we were not able to identify guide RNA-target hybrids in these data sets, even though hybrids of other, canonical snoRNAs, could be identified. At this point, the data that we obtained indicates that in neurons, Snord116 snoRNAs do associate with core snRNPs and generate primarily typical snoRNA transcripts. However, a guide function of these snoRNAs in the brain remains elusive. The 2'-O-methylation patterns in the brain tissue of PWScr$^{m+/p-}$ mice are similar to those in wt mice with some notable differences, which will be validated with mass spectrometry. The mechanistic basis of these differences in methylation will have to be determined.

5.2 Results

5.2.1 PWScr deletion does not affect alternative splicing of Rbfox2-regulated targets in mouse brain

To learn about the Rbfox2 association with the snoRNAs, we reanalyzed the data from [226] where binding sites of Rbfox1/2/3 in mouse brain were globally mapped with HITS-CLIP. We reproduced the observation that Rbfox proteins bind extensively to RNAs derived from the snord116 locus [205], specifically to binding sites that are located between the snoRNAs in the sno-lncRNA (FIGURE 5.2).

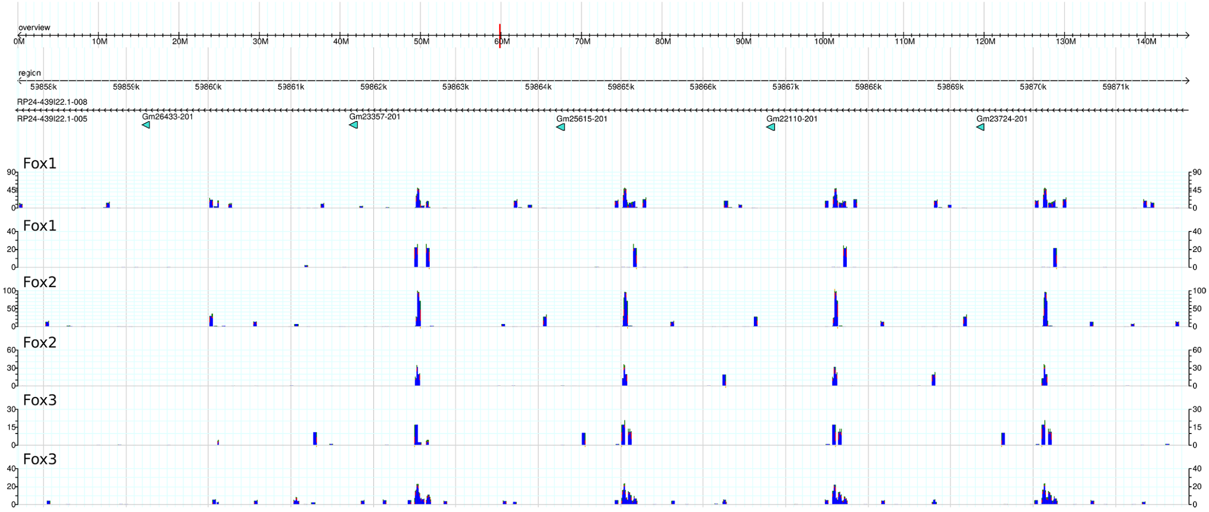


Figure 5.2 **Rbfox binding in the Snord116 locus.** The Snord116 copies are indicated with light blue triangles at the top of the panel. Densities of reads from the Rbfox1-3-HITS-CLIP experiments mapped to the negative strand of the chromosome are shown with dark blue.

To determine whether targets of the Rbfox proteins respond to the loss of Snord116 expression, we first used the site extraction tool that is implemented in the CLIPZ web server (www.clipz.unibas.ch) to obtain Rbfox1-3 binding sites, which are enriched in reads compared what is expected based on the mRNA level expression of the genes [227]. Asking whether Rbfox targets (transcripts for which significantly enriched sites were identified in at least 6 of the 8 HITS-CLIP experiments) are differentially expressed in the knock-out mouse samples, we found this not to be the case (FIGURE 5.3).

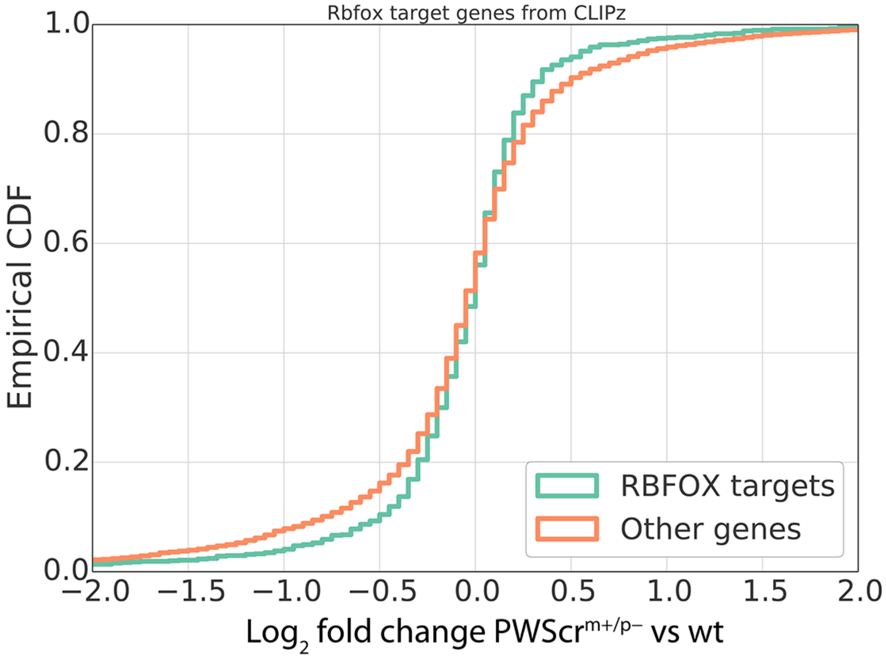


Figure 5.3 **Distribution of expression changes of HITS-CLIP-determined Rbfox targets** in PWScr$^{m+/p-}$ compared to wt mice.

We used the rMATs analysis tool [228] to determine whether the deletion of Snord116, which is predicted to prevent the "sponging" of the Rbfox proteins, leads to changes in the splicing of Rbfox targets. The distributions of percent spliced in (PSI) scores of individual exons/introns indicate that there is no global shift in splicing (FIGURE 5.4). Nevertheless, we examined the 24 high confidence events (p-value < 0.01) with |PSI difference| of at least 0.3 (FIGURE 5.5) and found that only one (Enah) of these, corresponding to the enabled homolog gene (ENAH) with role in cytoskeleton remodeling, is an Rbfox target. The exons of Rbfox targets that were identified in [205] changed very marginally in the PWScr$^{m+/p-}$ compared to wt mice, arguing against a Rbfox-sponging effect of the snoRNAs in the mouse brain (FIGURE 5.4). We also investigated all putative exon inclusion/exclusion events by sashimi plots (data not shown) but did not find any meaningful differences between PWScr$^{m+/p-}$ and wt mice.

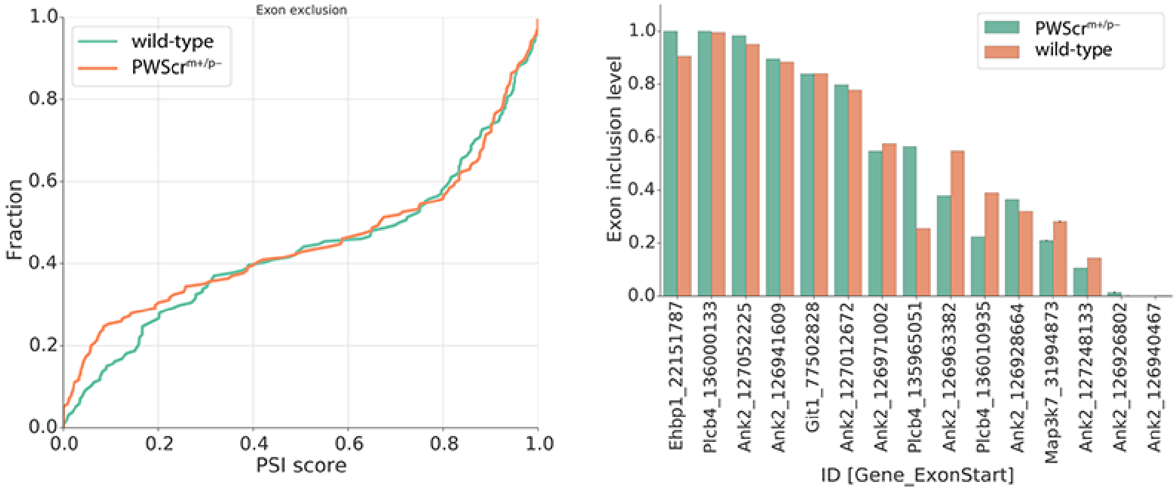


Figure 5.4 (A) **Exon inclusion/exclusion rates in wt and PWScr$^{m+/p-}$ mice.** (B) Exon inclusion/exclusion rates in genes identified in [205].

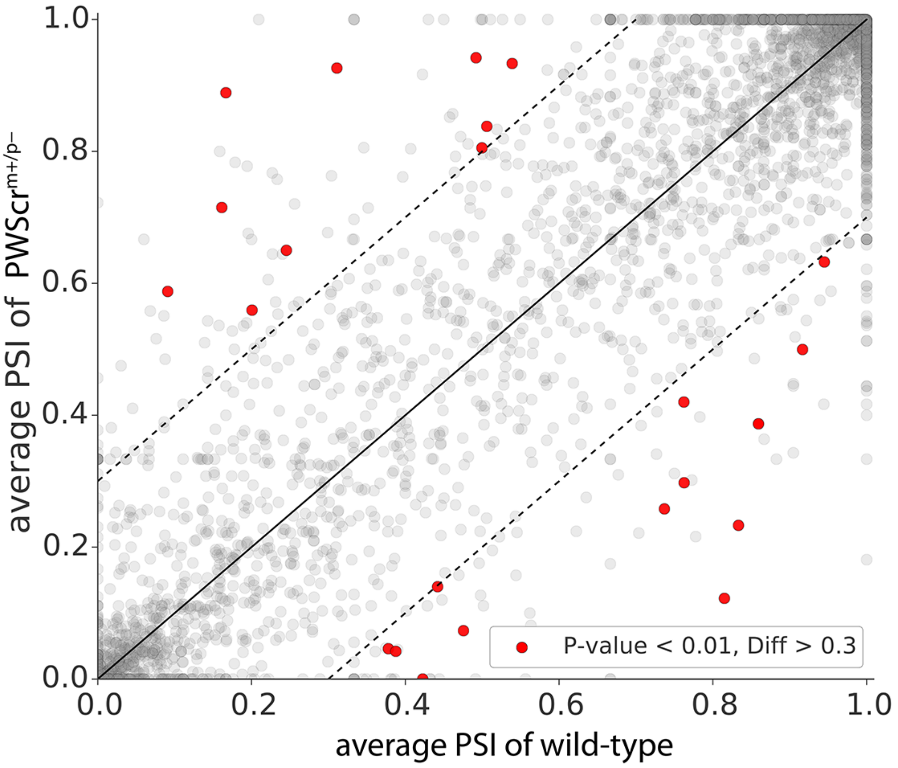


Figure 5.5 **Scatterplot of PSI scores of all possible Skipped Exons (SE)** in PWScr$^{m+/p-}$ and wt brain samples. Red dots indicate the significant SE events (p-value<0.01, in a likelihood-ratio test evaluating whether the difference of the mean PSI values between two sample groups exceeds a user defined threshold which was set of 0.0001) that have an absolute average inclusion difference > 0.3.

5.2.2 Snord116 expression in neuronal samples

It is well established that the imprinted snoRNAs from the PWS locus show highest expression in the brain in human as well as in mouse [211]. However, to identify a good cellular system for studying Snord116 function, particularly with CLIP, we first re-assessed the expression of these snoRNAs in several neuronal cell lines as well as brain tissue. We carried out northern blots in two mouse neuroblastoma cell lines N1E-115 and N2A and the human neuroblastoma cell line SH-SY5Y. Simultaneously we also probed for the expression of Snord116 in mouse brain (C57BL/6JRj).

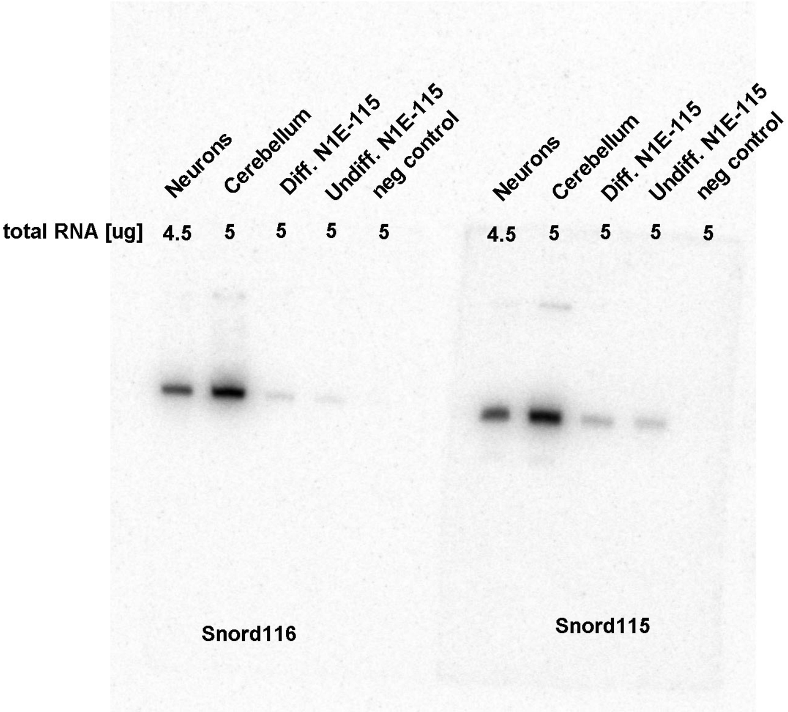


Figure 5.6 **Northern blots depicting the expression of Snord 116 (left) and Snord 115 (right)** in mouse neurons, total mouse cerebellum, differentiated and undifferentiated N1E-115 cells.

FIGURE 5.6 depicts the expression of Snord115 and Snord116 in mouse cerebellum, mouse cerebellar neurons, undifferentiated and differentiated N1E-115 cells (for differentiation see Methods) determined with northern blotting. We found that both snoRNAs have a higher expression in primary cells (total cerebellum and cerebellar neurons) compared with the N1E-115 cell line. Differentiation of N1E-115 cells, although successfully realized (SUPPLEMENTARY FIGURE 5.1), does not substantially increase the expression of these snoRNAs. Results for the fibroblast cell line used as a negative control are also shown. We also assessed the expression of Snord116 in the mouse N2A (FIGURE 5.7A and B) and human SH-SY5Y neuroblastoma cell lines (FIGURE 5.7C). Differentiation of both of these cell types into neuron-like cells by means of bovine serum albumin (BSA) and/or retinoic acid addition slightly increases the Snord116 expression.

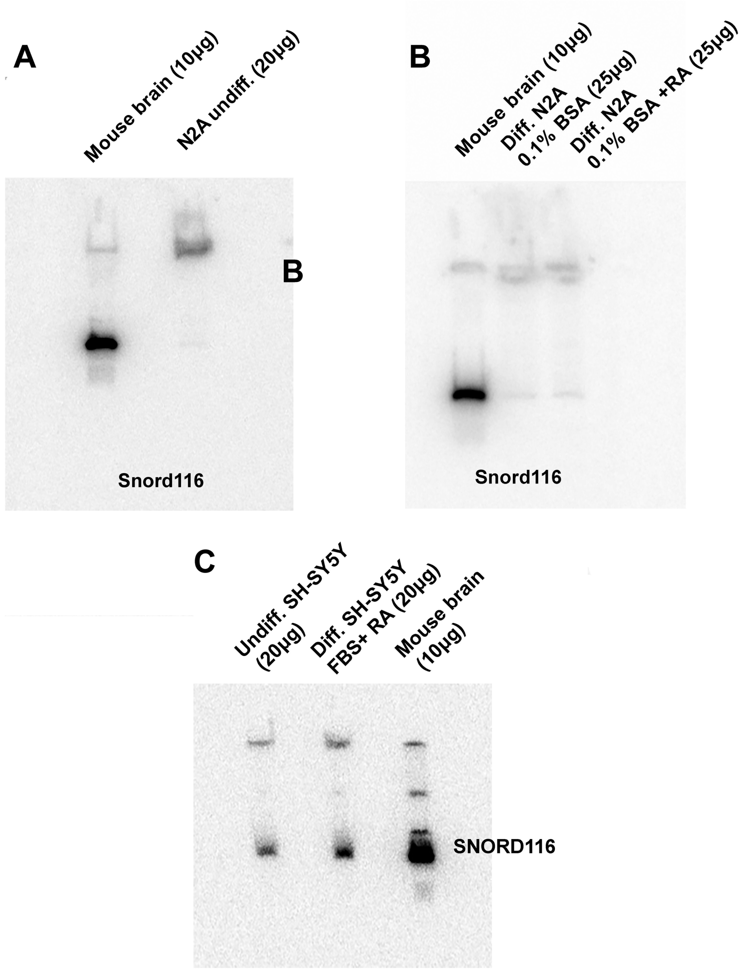


Figure 5.7 **Northern blots depicting expression of Snord116/SNORD116** in (A) undifferentiated N2A cells, (B) in differentiated N2A cells, and (C) in undifferentiated and differentiated SH-SY5Y cells. In every blot Snord116 expression was also probed in mouse brain, as a positive control

These results showed that Snord116 expression is much higher in primary neurons than in neuronal cell lines, including those in which differentiation towards neuron-like cells was performed. This indicates that it is unlikely to find a substitute model cellular system for studying Snord116 function, and primary neurons should be used.

5.2.3 RiboMeth-seq in wt mice and PWScr$^{m+/p-}$ mice

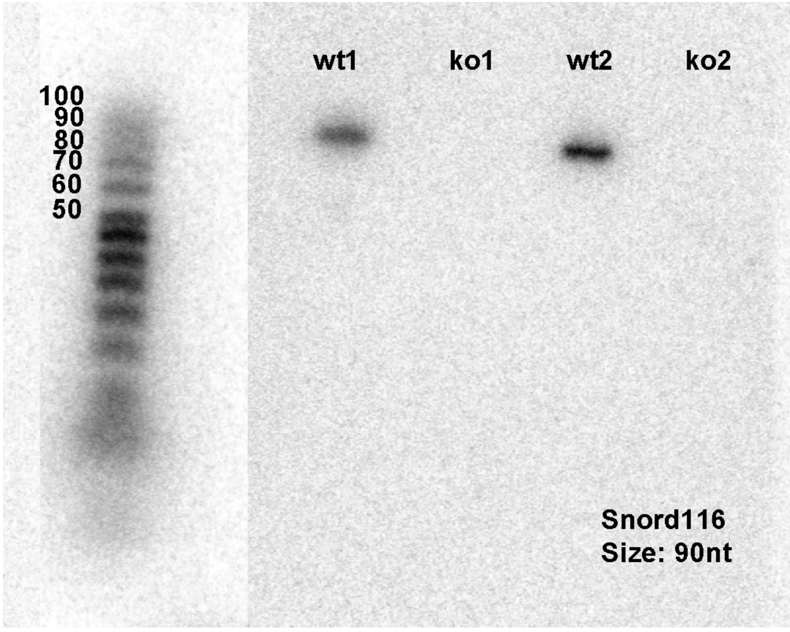


Figure 5.8 **Northern blot depicting Snord116 expression in brain tissue from wt and PWScr$^{m+/p-}$ mice (ko).**

SNORD116 being highly expressed in the brain, their role could be to sequester the core snoRNA proteins, thereby affecting the levels of rRNA 2'-O-methylation in the brain. To examine this possibility as well as the possibility that Snord116 still has a canonical function in guiding 2'-O-methylation, we generated RiboMeth-Seq [65] (CHAPTER 3) libraries from wt mice and from PWScr$^{m+/p-}$ mice, which lack Snord116 [219]. Before we prepared the RiboMeth-Seq libraries, we first assessed Snord116 expression in brains from two wt and two PWScr$^{m+/p-}$ (labeled as "ko") animals in FIGURE 5.8. As expected, we confirmed the presence and absence of Snord116 expression in wt and PWScr$^{m+/p-}$, respectively.

We analysed the RiboMeth-seq data as described previously (Gumienny and Jedlinski et al., CHAPTER 3). Additionally we have calculated a score (score 'C') proposed in [65] to reflect the fraction of methylated molecules in the cell. We observed high correlation between our score (angle) and score C (FIGURE 5.9).

For each putative 2'-O-methylation site a t-test on angle scores and C scores was performed to assess the level of difference between the sites in PWScr$^{m+/p-}$ and wt mice. We set a threshold of 0.01 for the p-value. Additionally we required that the change is in a specific direction i.e. that a site has to be present in wt and absent in PWScr$^{m+/p-}$. To call a site "present" in a given wt sample, we used specific score thresholds (280 for the angle score and 0.7 for the C score), which we set based on their power in accurately identifying known 2'-O-methylation sites in ribosomal RNAs.

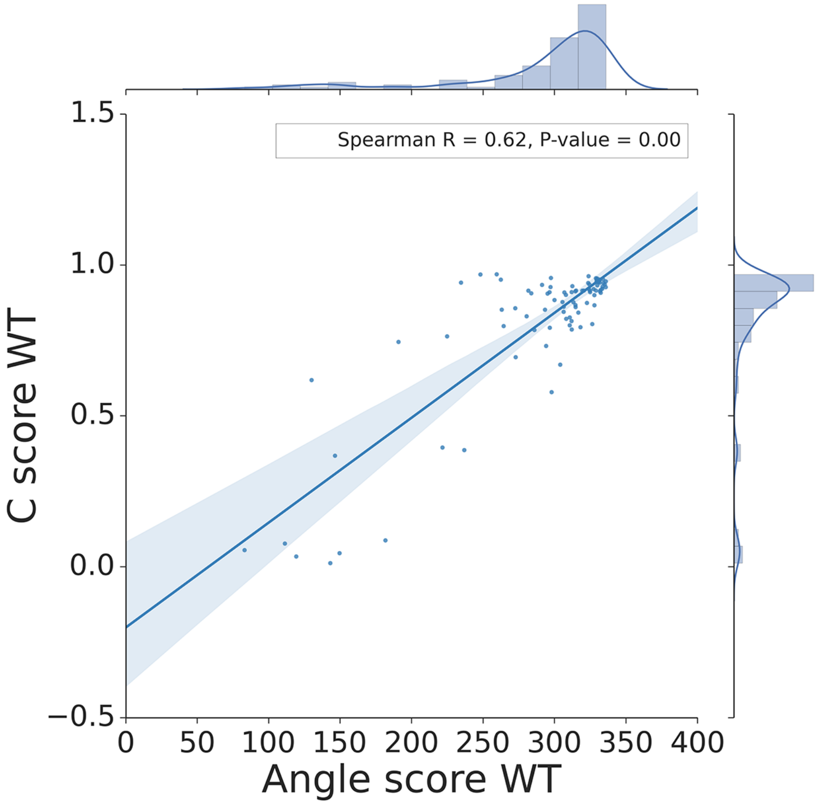


Figure 5.9 **Correlation between angle score and score C on known methylation sites.**

We have identified 173 and 144 differentially methylated sites for angle score and score C, respectively. To determine whether any of these sites could be a Snord116 site we applied the PLEXY algorithm [143] for predicting snoRNA targets. We found that PLEXY assigned only one site identified by angle score, located in the cholecystokinin gene (chr9:121494409 in mm10), to Snord116. A similar analysis applied to sites predicted by the score C identified two possible Snord116 targets: one in the adenylate cyclase (Adcy1, chr11:7174707 / mm10 genome assembly version from the UCSC genome server) and the other in the ring finger protein 14 (Rnf14, chr18:38300497 / mm10).

Finally, to assess the association of these snoRNAs with core snoRNP proteins, we performed HITS-CLIP experiments as previously described [130], with antibodies targeting the Nop58 (for N2A and SH-SY5Y cells) and Fibrillarin (for N1E-115 cells and mouse neurons) proteins. To this end, the N2A and SH-SY5Y cells were differentiated, whereas the N1E-115 cells were left untreated, as no increase of Snord116 could be observed in these cells in the northern blot. Neurons were extracted from mice, pooled, cross-linked with UV, and then also carried through the CLIP library preparation as described [130] (for details see METHODS).

5.2.4 Nop58-HITS-CLIP in SH-SY5Y cells

Analysis of the Nop58-CLIP from SH-SY5Y cell line was performed as previously described in (Gumienny and Jedlinski et al., CHAPTER 3). We have performed only one experiment in this cell line. In human there are 190 known snoRNA-target interactions of a total of 116 distinct 2’-O-methylation sites in rRNAs and small nuclear snRNAs. In this experiment we were able to recover 83 out of 190 interactions (44%) and 66 out of 116 2’-O-methylations (57%). A more stringent selection of the sites based on their probability (calculated as described in Gumienny and Jedlinski et al., CHAPTER 3) being greater than 0.5 left 62 out of the 190 interactions (33%) and47 out of 116 distinct 2’-O-methylation sites (40%). These relatively low numbers could reflect both the inefficiency of capture of snoRNA-target interactions as well as the tissue specificity of neurons/neuron-like cells in regard to 2’-O-methylation.

Applying the same stringent threshold to sites that involved mRNAs and other RNA species that are not considered canonical snoRNA targets, we have found 256 putative interactions. These included 7 novel interactions with rRNAs, 6 of which predicted to lead to the 2’O-methylation of sites that were previously known to be methylated (but the guide snoRNA was not known). 6 of these methylation sites did not emerge from RiboMeth-seq, which may reflect either the partial coverage of RiboMeth-seq, the partial methylation status of these sites, or false positive sites (albeit with previous evidence in the literature). One interaction, at position 462 of 18S rRNA that is known to be 2’-O-methylated, seems to involve the SNORD116 snoRNA and has very low predicted energy of interaction (-18 kcal/mol).

HITS-CLIP revealed 200 potential interaction sites that are novel and non-canonical. About half of these seem to involve orphan snoRNAs. Most of the target sites were located in repeats (46%) but a substantial fraction is located in protein coding genes (28%). In contrast to the well established snoRNAs with canonical targets, orphan snoRNAs were frequently found in interactions with miscRNAs (in fact with the many copies of Y RNA that is involved in DNA replication [229, 230]).

Among the HITS-CLIP-defined sites, 15 seem to interact with SNORD116. One of these mapped to the NOP58 CDS. Another corresponded to a known methylation site (nucleotide 463) in 18S rRNA, which however, also interacts with SNORD14. This may explain why no significant change in methylation level could be inferred for this site with RiboMeth-seq. Finally, one of the putative SNORD116 interactions was with an tRNA and some with Alu elements.

For the SNORD115 snoRNAs we only identified two putative targets, both of them in repeats (one in ribosomal protein L5 repeat and one in (A)n repeat), consistent with these snoRNAs being less incorporated into snoRNPs.

5.2.5 Fibrillarin-HITS/PAR-CLIPs in N1E-115 cells

In mouse there are 136 known snoRNA-target interactions of total 90 2’-O-methylation sites (in rRNAs, according to snOPY database [231]). In this experiments we were able to recover 95 out of 136 interactions (70%) and 66 out of 90 2’-O-methylations (73%). Setting again a more stringent threshold for site probability (determined based on the accuracy of rRNA site identification and set to 0.1), we retained 91 out of 136 interactions (67%) and 63 out of 90 2’-O-methylation sites (70%). In this case, combined analysis of more replicate experiments (four as opposed to one in SH-SY5Y) resulted in increased sensitivity and also precision.

In rRNA we have identified 30 new putative 2’-O-methylation sites stemming from 34 interactions. Overall, in mouse snoRNA-target interactions and ribosomal 2’-O-methylation are less characterized than in human or yeast. Thus, many of these sites are sites that are already known to be methylated in human, e.g. position 469 of mouse 18S rRNA which is equivalent to position 468 in human.

We have identified 881 interaction sites located in RNAs not considered to be canonical targets of snoRNAs. The functional categories of the corresponding targets are similar to those inferred from the SH-SY5Y cell line with the exception that in N1E-115 we do not observe many target sites in misc RNAs such as Y RNAs. We have identified five interactions sites guided by Snord116 and one interaction site guided by Snord115. This pattern is is in agreement with the data that Snord116 is more efficiently captured in the CLIP experiments (**Figure 5.10** and **Figure 5.11**)

5.2.6 Fibrillarin-HITS-CLIP in mouse primary neurons

We have performed several Fibrillarin-CLIP experiments in mouse primary neurons, aiming to capture snoRNA-target chimeric reads. Only one of these experiments yielded a number of chimeric reads that seemed to be meaningful and we present the analysis of these data here. In this experiment we were able to recover only 75 out of 136 interactions (55%) and 56 out of the 90 2’-O-methylation sites (62%). After application of a threshold of 0.5 (same as for the CLIP in the human cell line) the numbers decreased from 75 to 53 and 56 to 39 for interactions and 2’-O-methylation sites, respectively.

Here we have identified six novel interaction sites in rRNAs and almost 2000 putative interactions, coming mostly from the non-orphan snoRNAs, in other RNAs. However, in the brain tissue a substantial amount (34%) of guide-target interactions came from orphan snoRNAs. Out of these sites, 90 were predicted to be targeted by Snord116 family members and 24 by Snord115 family members. As before, there were more chimeric read-based predicted target sites for Snord116 than for Snord115.

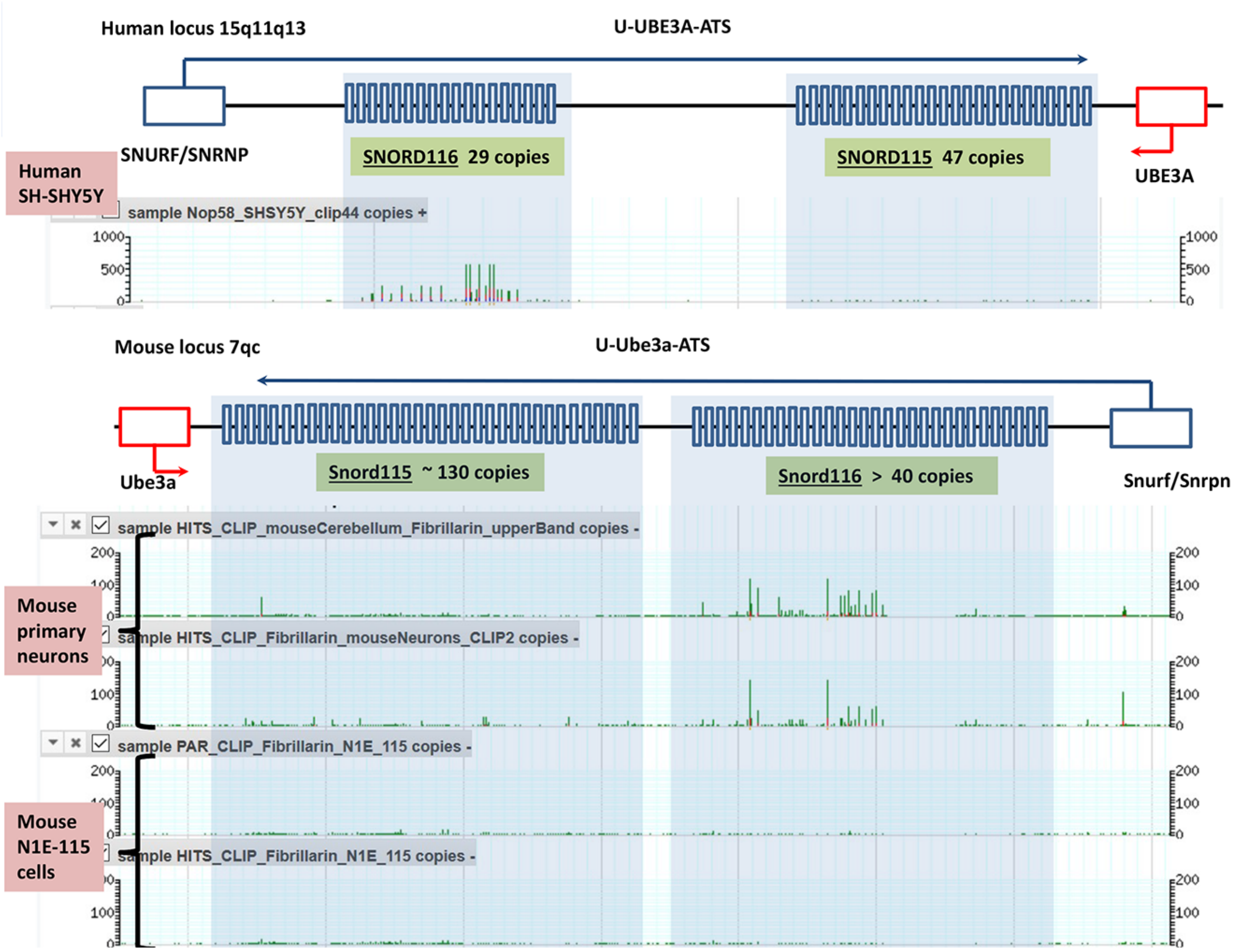

Human locus 15q11q13
U-UBE3A-ATS
SNURF/SNRNP
SNORD116 29 copies
SNORD115 47 copies
UBE3A
Human SH-SHY5Y
sample Nop58_SHSY5Y_clip44 copies +
1000
500
0
Mouse locus 7qc
U-Ube3a-ATS
Ube3a
Snord115 ~ 130 copies
Snord116 > 40 copies
Snurf/Snrpn
sample HITS_CLIP_mouseCerebellum_Fibrillarin_upperBand copies -
Mouse primary neurons
sample HITS_CLIP_Fibrillarin_mouseNeurons_CLIP2 copies -
sample PAR_CLIP_Fibrillarin_N1E_115 copies -
Mouse N1E-115 cells
sample HITS_CLIP_Fibrillarin_N1E_115 copies -
200
100
0

Figure 5.10 **Snapshot from the CLIPZ genome browser.** Overlaid is a scheme of the PWS loci in human (upper) and mouse (bottom) with SNORD115/Snord115 and SNORD116/Snord116 snoRNAs loci. Green vertical lines indicate reads that were obtained from the individual CLIP experiments and mapped with CLIPZ. The CLIP experiments are as follows (top to bottom): Fibrillarin-HITS-CLIP in SH-SY5Y, Fibrillarin-HITS-CLIP in mouse primary neurons (Replicate1), Fibrillarin-HITS-CLIP in mouse primary neurons (Replicate2), Fibrillarin-PAR-CLIP in N1E-115, Fibrillarin-HITS-CLIP in N1E-115.

FIGURE 5.10 illustrates the capture of SNORD115/Snord115 and SNORD116/Snord116 snoRNAs in human and mouse CLIP experiments. Overlaid is a schematic representation of the PWS locus in human and mouse with its multi-copy genomic arrangements of the PWS snoRNAs. The green vertical lines overlap with the snoRNA loci. SNORD116/Snord116 are more efficiently captured in the CLIP experiments than SNORD115/Snord115 in both human and mouse cells. Remarkably, in the CLIPs performed in the SH-SY5Y cells and mouse primary neurons SNORD116/Snord116 were more efficiently captured than in the CLIPs performed in N1E-115 cells, which largely correlates with the expression levels observed in these cell lines and in the primary neurons (FIGURE 5.6 and FIGURE 5.7)

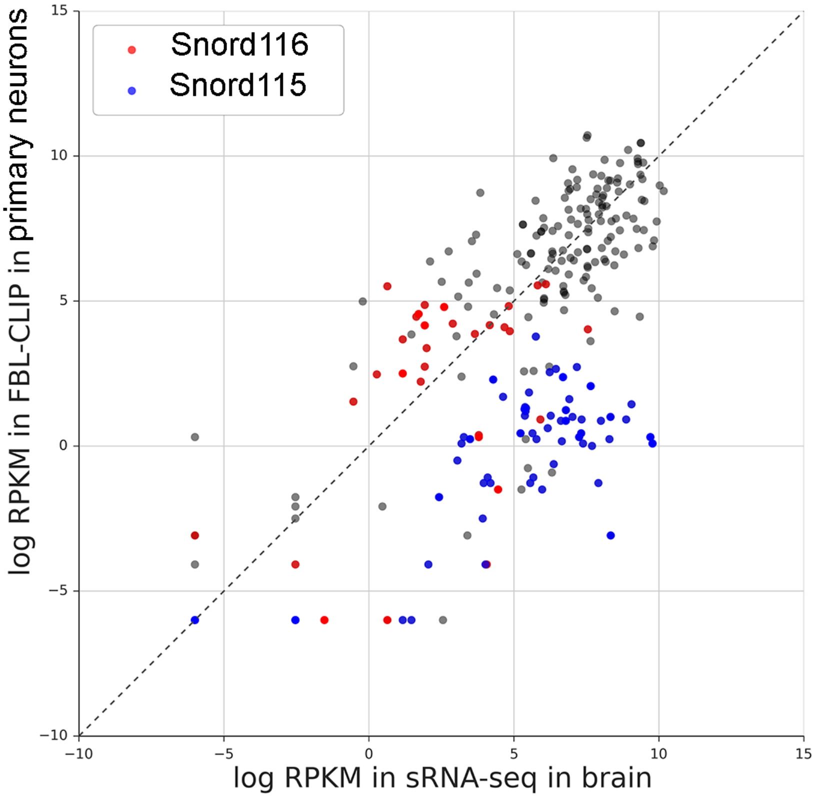


Figure 5.11 **Comparison of snoRNA expression levels calculated from Fibrillarin-CLIP** experiments (in primary neurons) and from small RNA-seq (mouse brain). Snord116 (red dots) are clearly enriched over Snord115 (blue dots) in the CLIP experiments, although they are expressed at lower levels in the sRNA-seq. Black dots represent all other snoRNAs.

FIGURE 5.11 depicts the capture of Snord116 and Snord115 by Fibrillarin-HITS-CLIP in mouse primary neurons. Snord116 are more efficiently captured than Snord115. However, canonical snoRNAs, although expressed at similar levels, represent the majority of the CLIP reads.

5.2.7 Combined analysis of RiboMeth-seq and CLIP-seq data

It is important to confirm putative 2’-O-methylation sites obtained from the chimeras in CLIP with RiboMeth-seq and *vice versa.* To this end we have intersected our mouse CLIP experiments with RiboMeth-seq. i.e. for each of the extracted sites obtained from the chimeric reads (irrespectively of the probability) we have checked if the site is also supported by RiboMeth-seq. The thresholds that we used were 0.1 for the probability in CLIP and 280 for the angle score in RiboMeth-seq in wt mice. FIGURE 5.12 and FIGURE 5.13 show all the 2’-O-Me sites found in rRNAs by CLIP and RiboMeth-seq along 18S and 28S rRNAs. Several sites are unknown in the literature but seem to be methylated based on the RiboMeth-seq data and their guides identified from chimeric reads. FIGURE 5.14 shows examples of individual 2’-O-Me sites.

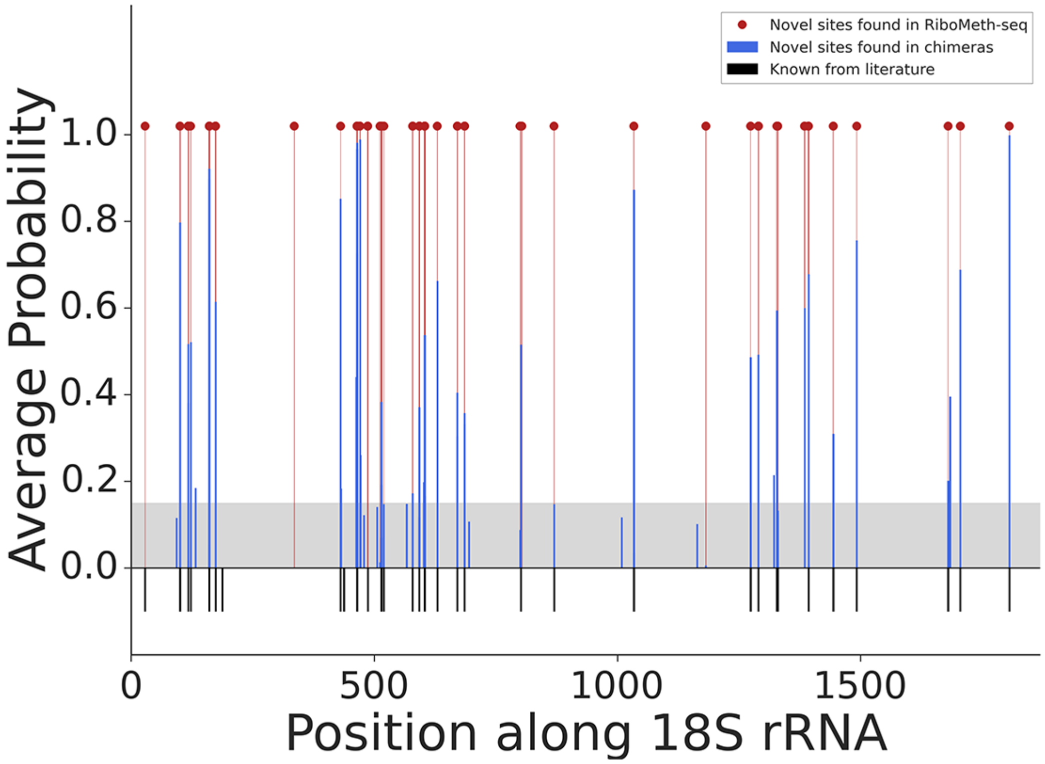


Figure 5.12 **2’-O-methylation sites and CLIP interactions identified in 18S rRNA.**

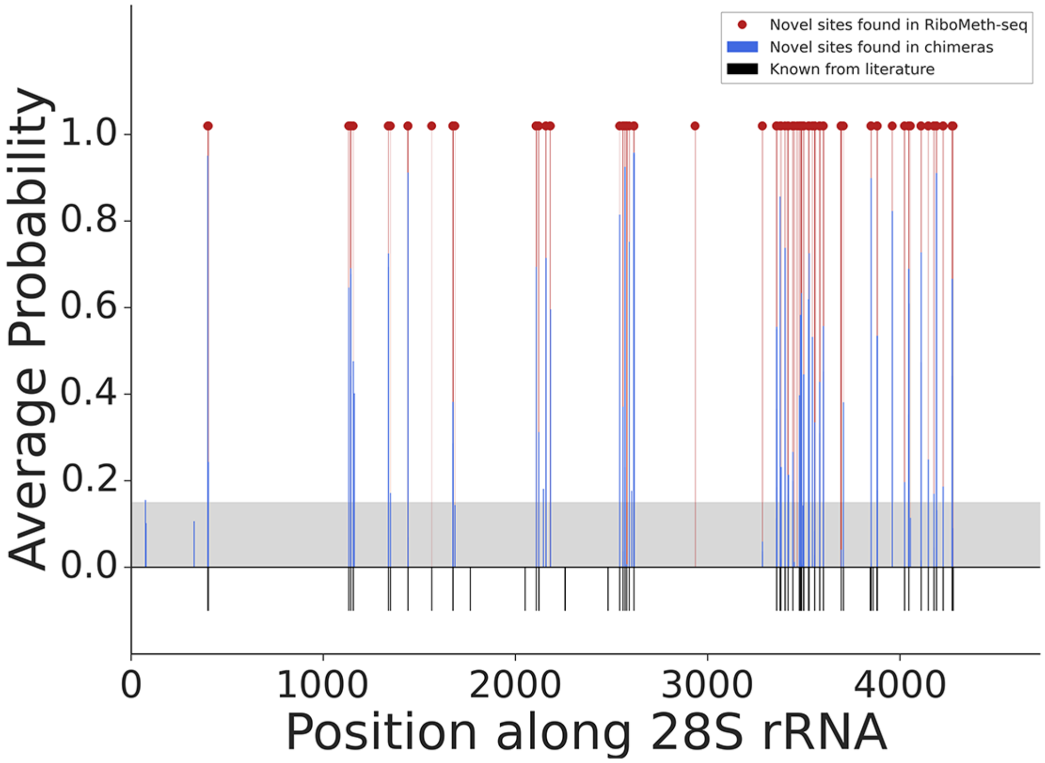


Figure 5.13 **2’-O-methylation sites and CLIP interactions found in 28S rRNA.**

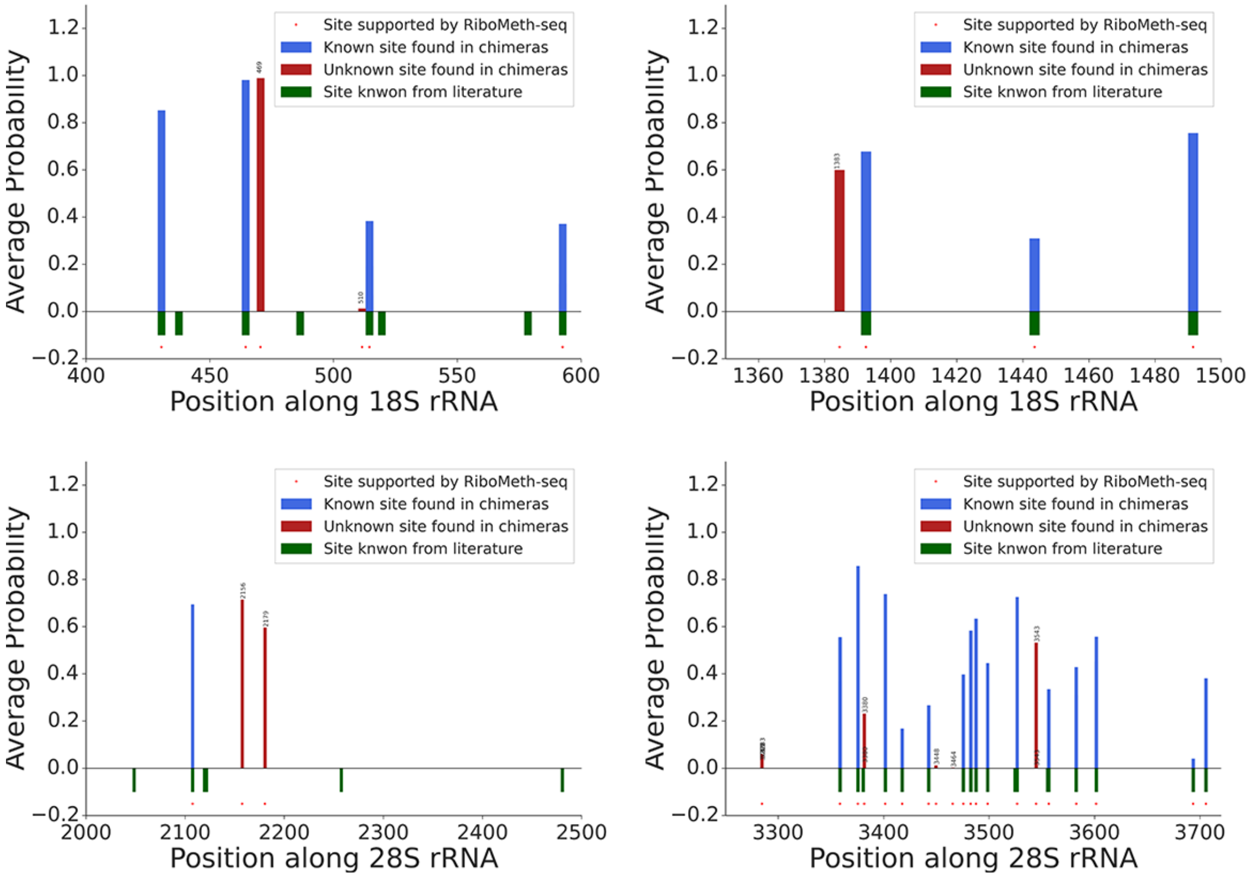


Figure 5.14 **Regions of 18S and 28S rRNAs showing known as well as novel positions of 2’-O-methylation identified with RiboMeth-seq and CLIP.**

Taken together, using RiboMeth-seq and CLIP-seq we have recovered 73% of the known methylation sites in mouse. In addition to these sites we have found 15 novel 2’-O-methylation sites whose guides were apparent in chimeric reads, seven of which with a high confidence (probability over 0.2). Most of these sites seem to interact with snoRNAs that already have assigned targets (i.e. they are not orphan snoRNAs). We have not found any non-canonical methylation site in the intersection of the RiboMeth-seq and CLIP-seq data sets. This does not exclude that non-canonical target sites exist. It is important to bear in mind that the formation of chimeras in CLIP experiments is very inefficient [47]. This means that chimeras involving targets with low expression are very difficult to capture and deeper sampling will be necessary.

5.2.8 Nop58-CLIP in N2A cells

Analysis is still in progress.

5.3 Discussion

PWS is a congenital disease that is caused by the loss of paternal expression of genes from the maternally imprinted chromosomal region 15q11-13. This region comprises multi-copy snoRNAs as well as several protein-coding genes. The latter do not appear to contribute to the development of the syndrome and all the evidence points to SNORD116 snoRNAs playing a major role in PWS. It is striking that SNORD116 snoRNAs display most features of canonical C/D box snoRNAs, yet targets with which these snoRNAs interact, as canonical snoRNAs do, have not been identified. However, their unusual but conserved multi-copy genomic arrangement and the high stability of SNORD116 found in the brain are indicative of an important role that is still to be uncovered. We here tried to examine comprehensively whether SNORD116 snoRNAs do play the role of guides that direct 2'-O-methylation. Advances in high-throughput analysis have greatly facilitated this effort in the recent years.

First, we set out to find out whether sno-lncRNAs emerging from the PWS locus can regulate alternative splicing of Rbfox2-targets in the murine brain. Our analysis does not suggest that this mechanism is prevalent in adult mouse brain. These results are consistent with our observation, that sno-lncRNA are only faintly expressed in brain tissue and neuron-derived cell lines that we tested and that mature SNORD116/Snord116 represent the predominant form (FIGURE 5.6, FIGURE 5.7, and FIGURE 5.8) . We then thoroughly examined SNORD116 expression in various cellular systems and observed that the human SH-SHY5Y cell line exhibited a significantly higher level of expression of SNORD116 compared to the murine cell lines, yet still much smaller than primary murine neurons. This indicates that the function of SNORD116/Snord116 should be studied in primary cells.

We then decided to examine a canonical function of SNORD116 by analyzing 2'-O-methylation transcriptome-wide in PWScr$^{m+/p+}$ wt mice and PWScr$^{m+/p-}$ mice, in which the SNORD116 cluster was deleted. To this end, we generated RiboMeth-seq libraries from brain tissue obtained from wt mice and mice carrying the deletion. The analysis of the 2'-O-methylation sites enabled a thorough characterization of 2'-O-Me's in mouse 18S and 28S rRNA as well as in snRNA. Interestingly, the analysis of the 2'-O-methylation landscape in brains of wt mice and PWScr$^{m+/p-}$ mice with RiboMeth-seq revealed reproducible differences in some transcripts. One of these corresponds e.g. to the cholecystokinin triacontatriapeptide (Cck). CCK is an endocrine gut hormone that is released by epithelial cells in the mucosal lining of the small intestine following food intake and is responsible for mediating satiety by acting on CCK receptors distributed throughout the central nervous system, thus regulating the metabolic homeostasis [232]. Since lack of satiety linked with excessive eating behavior are some of the hallmarks of PWS, CCK presents a plausible target that may be disregulated in this disorder. We are currently validating these candidate target sites with mass spectrometry and RTL-P [188] in both wt and knock-out conditions. To further explore a possible guide function of SNORD116/Snord116 we performed snoRNP-CLIP experiments in neuronal cell lines as well as in primary mouse neurons and examined PWS-related snoRNA incorporation into snoRNPs. As presumed, PWS-snoRNA-loading was particularly high in CLIPs conducted in primary neurons and the human SH-SY5Y cell line, while Snord116 snoRNAs were not captured in the CLIPs that were carried out in murine cell lines. In fact, the capture of SNORD116 by CLIP was best in the human SH-SY5Y cells, indicating that there may be a species-specific component regulating the loading of these snoRNAs into the snoRNP.

Surprisingly, we observed that SNORD115, a C/D box snoRNA that was initially thought to be one of the key players in PWS, was captured only minimally in our snoRNP CLIP experiments. Although both SNORD115 and SNORD116 are similarly expressed in the tested cell lines and tissues, SNORD116 seem to be preferentially loaded into the snoRNP complex compared to SNORD115. This observation goes along with the proposed alternative role of SNORD115 in the modulation of alternative splicing of the 5-HT2C serotonin receptor, which would not require a primary binding of SNORD115 to the proteins of the snoRNP

Next, we analyzed the guide RNA-target hybrids that usually form in CLIP experiments. We identified an extensive number of snoRNA-rRNA chimeras which allowed us to largely reconstruct the 2'-O-methylation landscape in 18S and 28S rRNA as well as snRNA in mouse and human, as we have previously done in human HEK cells (Gumienny and Jedlinski et al., CHAPTER 3). These results imply that the snoRNP-CLIPs are suitable for the further analysis of potential SNORD116-target hybrids. However, the in-depth analysis of the hybrids did not reveal any reliable targets neither in rRNA and snRNA nor in other RNAs. This could be due the fact that chimeras represent only a small fraction of all the reads that are usually obtained in CLIP experiments, which is a serious limitation of this method in regard to guide-target assignment for low-abundance targets [47]. Additionally, Snord116 are captured at lower efficiency than canonical snoRNAs. As a result of the relatively low capture and the aforementioned limitation of the CLIP method one may miss out on potential Snord116-target chimeras.

Taken together, we have re-evaluated here a canonical role of the PWS-associated SNORD116 snoRNAs. After failing to reproduce the reported impact of sno-lncRNAs on splicing modulation of Rbfox2-regulated targets in a mouse model that is lacking expression of the PWS critical region [219], we have extensively examined the role of SNORD116 in 2'-O-methylation with various high-throughput approaches. We have thus identified three putative 2'-O-Me sites in mRNAs with RiboMeth-seq that may be attributed to Snord116. These candidate sites are currently being individually validated. Intersections of the RiboMeth-seq data with CLIP-seq data could not confirm these sites, however, we have identified 15 new 2'-O-Me's in both 18S and 28S rRNA that were supported by both approaches. Hence we expand here the mouse snoRNA atlas by providing new 2'-O-Me sites.

Uncovering unconventional targets will enable one to estimate to what extent SNORD116 participate in cellular networks that regulate gene expression and will therefore provide novel insights into snoRNA-mediated gene regulation. Most importantly, identification of targets of the PWS-associated SNORD116 would mean a major advancement towards the development of a suitable treatment.

5.4 Methods

5.4.1 Splicing Analysis

Reads were mapped to the genome using the STAR aligner (version 2.5.0a) [187] with the following command: STAR --runMode alignReads --runThreadN 10 --genomeDir <genome dir> --sjdbGTFfile <annotation file> --readFilesIn <reads> --outFileNamePrefix <our prefix> --outSAMtype BAM SortedByCoordinate --twopassMode Basic --alignIntronMax 200000 --readFilesCommand zcat. Then rMATs (version 3.2.0 beta) [228] software was used to calculate the PSI scores (Percent Spliced In) for all the different splicing events using the following command: python RNASeq-MATS.py -b1 <ko_rep1.bam>,<ko_rep2.bam>,<ko_rep3.bam> -b2 <wt_rep1.bam>,<wt_rep2.bam>,<wt_rep3.bam> -gtf <annotation file> -o <output dir> -t single -len 63.

5.4.2 HITS-CLIP experiments

HITS-CLIP was performed in the following cells: N2A(ATCC), SH-SY5Y (ATCC), and N1E-115 (ATCC), and neurons extracted from mouse cerebellum (C57BL/6JRj). HITS-CLIPs in N2A and SH-SY5Y cells were performed as previously described using 'mild RNase T1' digestions conditions [130]. HITS-CLIPs in N1E-115 and neurons were performed alike with the following modifications: after the overnight 3'-adapter ligation the RNA was not PAGE-purified and immediately carried over to the next step: before 5'-adapter ligation, the reverse transcription primer was added (at equimolar amounts to the the 3'- and 5'-adapter (1µl of 10µM)) to the reaction and annealed by denaturing at 90°C for 30 seconds, 65°C for 5 minutes, and then chilled on ice for at least 1 minute. Subsequently, the 5' adapter was added and ligated at 20°C for 1 hour, followed by 37°C for 30 minutes. The RNA was then reverse transcribed with Superscript IV (Invitrogen) according to manufacturer's instructions. After pilot PCR and final PCR as described in [130], the PCR product was then purified using a non-denaturing 8% PAGE purification instead of using an agarose gel electrophoresis-based purification.

For immunoprecipitation, antibodies were coupled to protein-G dynabeads (Invitrogen #10003D). Antibodies used against endogenous proteins were α-Nop58 (sc-23705) and α-Fibrillarin (Bethyl labs A303-891A). After SDS-PAGE, gels were blotted onto nitrocellulose membranes to reduce the background from free RNAs [151]. N2A and SH-SY5Y libraries were sequenced on a HiSeq 2000 system (Illumina), N1E-115 and the neuron libraries were sequenced on a HiSeq 2500 system (Illumina).

5.4.3 Northern Blots

Northern blots were performed as described in [177]. The probes used to detect Snord115 and Snord116 in N1E-115 cells and neurons/mouse brain were as follows:

Snord115: Sequence (5'-3'): CCT CAG CGT AAT CCT ATT GAG CAT GAA
Snord116: Sequence (5'-3'): TTC CGA TGA GAG TGG CGG TAC AGA

The probes used to detect Snord116/SNORD116 in N2A, mouse brain, and SH-SY5Y were as follows

Snord116/SNORD116: Sequence (5'-3'): TCA CTC ATT TTG TTC AGC TTT

The Snord115/116 family members were manually aligned (not shown). Both Snord115 and Snord116 display highest conservation among family members near their 3' ends. The probes were designed in a way to target as many of the family members as possible.

5.4.4 RiboMeth-seq

RiboMeth-seq was previously described in **CHAPTER 3**.

The following adaptations were made here for the analysis of the RiboMeth-seq data:

To avoid false positives we have applied an angle score threshold of 290 for each of the experiment, as this seem to lead to the most accurate identification of rRNA sites (see Chapter 3). Next, for all sites that had an angle score above the threshold in at least one experiment, we calculated the average angle score across all experiments. Furthermore, to avoid noisy data coming from low expressed targets, we have also calculated the local coverage of the putative methylation positions and we used a threshold of at least 5 reads on average per neighboring nucleotide for a site to be considered in a given experiment. We have assessed the threshold for the angle score using the precision-recall (PR) curve and Matthews coefficient on ribosomal RNAs (**FIGURE 5.15**). As can be seen the optimal threshold for the averaged angle score is 290. Overall there are 90 known 2'-O-methylation sites listed in the snOPY database and this threshold recovers 70% of the known sites. We have decreased this value to 280 which has lower precision (~50%) but substantially higher recall (77%). Score C was computed as described in [65] and we used a score threshold of 0.7 to select putative methylation sites.

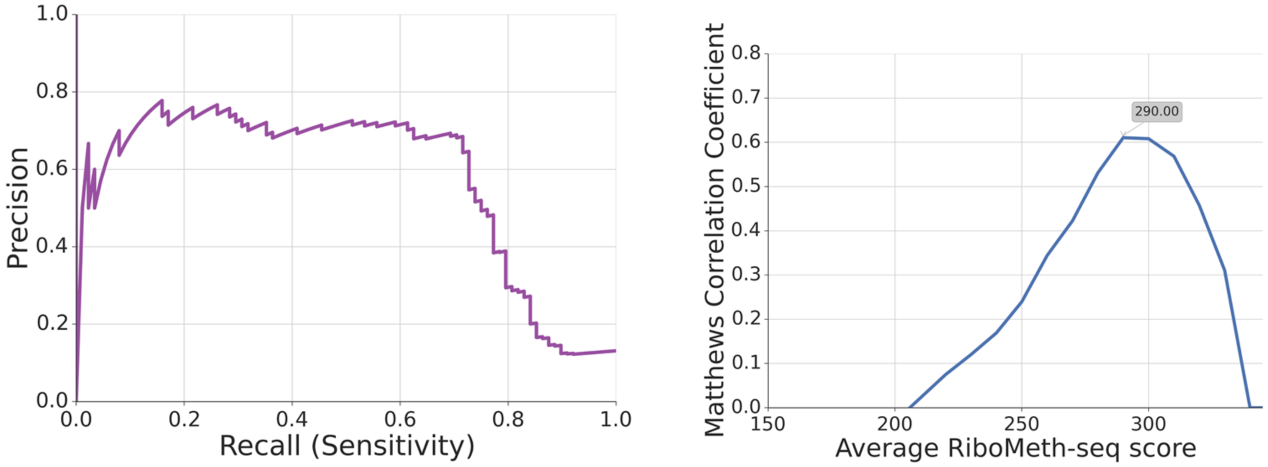


Figure 5.15 **PR curve (A) and Matthews coefficient plot (B) of the angle score.**

5.4.5 Cell culture and differentiation

N2A, SH-SY5Y, and N1E-115 samples, cells were grown according to the supplier's instructions. Thirty 15-cm tissue culture plates were used for N2A and SH-SY5Y cells, five 15-cm tissue culture plates were used for N1E-115 cells.

N2A cells were plated at 30-40% confluency and were differentiated in DMEM (Sigma #D5796), 2mM L-glutamine (#25030024), 0.1% BSA (Sigma #A9418), and no FBS for the first day, the second day with the same medium including 20µM all-*trans*-retinoic acid (Sigma #R2625). After 48 hours of differentiation they were crosslinked with UV 254 nm, harvested and snap-frozen for CLIP.

SH-SY5Y cells were plated at 30-40% differentiated in DMEM with 2mM L-glutamine and, 2% FBS, and 20µM all-*trans*-retinoic acid for 48 hours, UV-crosslinked, harvested and snap-frozen.

N1E-115 were grown according to the supplier's instructions. Differentiation of N1E-115 cells for Snord116 probing was performed as follows: at approx. 50% confluency cells were washed once with prewarmed sterile PBS (Gibco #10010023) and the medium was changed to Neurobasal Medium (Gibco #21103-049) supplemented with 2mM GlutaMax (Gibco #35050-061). Cells were differentiated for 72 hours.
N1E-115 were not differentiated for CLIP, they were grown under standard conditions and at 80% confluency they were UV-crosslinked and snap-frozen for CLIP.

5.4.6 Preparation of cerebellar granule cells for CLIP-experiments

5.4.6.1 Tissue dissection

Mice (C57BL/6JRj) at P5 or P6 were used for the extraction of the neurons (cerebellar granule cells). The animal was decapitated, the brain removed and transferred to a 60mm dish with cold HBSS (Gibco #14175). HBSS was supplemented with 5ml of 1M Hepes (Gibco #15640-080) to 500 ml of HBSS. The meninges were removed from the cerebellum, the cerebellum and brainstem were separated from the mid- and forebrain, then the cerebellum was removed from the brainstem. The meninges in the cerebellar folds were removed as well as possible. The dissected cerebellum was transferred to a new dish and cut into 10-15 pieces.

5.4.6.2 Tissue dissociation

A sterile, BSA (Sigma #A9418 4% BSA in HHBS)-coated pipette was used to to transfer the tissue pieces to a coated 15ml Falcon tube containing 10ml of HBSS on ice. After the tissue pieces settled to the bottom, most of the HBSS solution was removed, only enough HBSS was left to cover the cells. 100µl of DNase I (Roche #10104159001, grade II from bovine pancreas) at a concentration of 1mg/ml (in DMEM Sigma #D5796) and 1ml of 1x Trypsin-EDTA (Gibco #25300-054) (containing 7mM Hepes pH7.25) were added. The mixture was incubated for 10-15 minutes at 37°C. Subsequently, the following was added to the dissociation mixture:

- 100µl soy bean trypsin inhibitor (Sigma #T6522, type I-S from soybean)
- 250µl DNase I (1mg/ml in DMEM)
- 320µl HBSS containing 4 ml/ml BSA

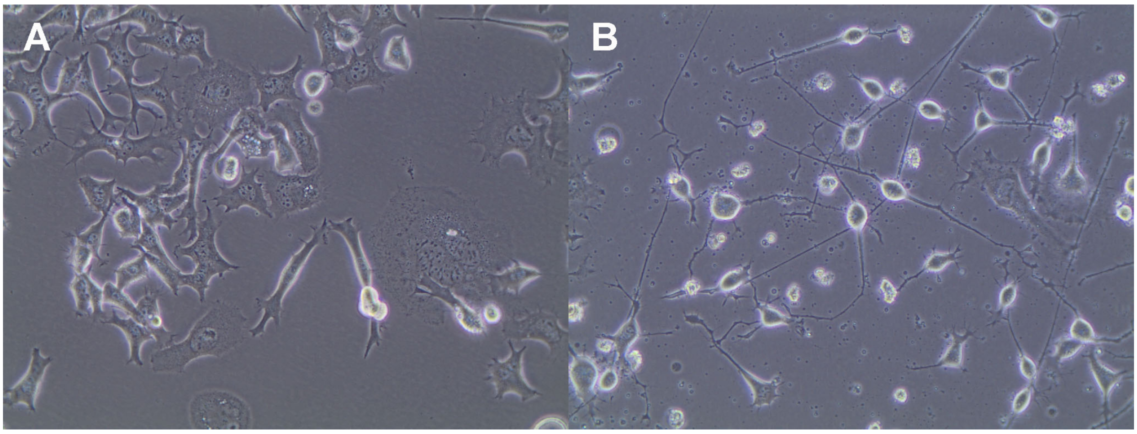
The mixture was incubated for 1 minute at 37°C, then cooled on ice. The mixture was carefully triturated four times with a coated 1ml pipette tip (coated by pipetting 0.1% BSA und and down several times). A 200µl pipette tip was put on top of the 1ml pipette tip, and undissociated material was further triturated approximately 10 times. HBSS was added to a final volume of 10ml. The cell suspension was filtered through a 40µm nylon mesh cell filter (BD Flacon #352349) into a 50ml falcon tube. If material from several pups was used, the cell suspensions were combined. Cells were centrifuged in 15ml flacons for 10 minutes at 250g and 4°C. The supernatant was discarded. The cell pellet was resuspended in 1ml of ice-cold HBSS using a coated 1ml pipette. The volume was filled up to 5ml with ice-cold HBSS. The cell suspensions were filtered a second time using a 40µm nylon mesh cell filter and the cells were counted. 10cm-Petri dishes were filled with 8ml of ice-cold HBSS. 5ml of filtered cell suspension were added per Petri dish. The cell suspensions were UV-crosslinked (254nm) 3x0.1J/cm^2. Crosslinked cell suspensions were filled into coated 15 ml polypropylene tubes. The plates were washed with 2ml ice-cold HBSS and remaining cells (after UV-crosslinking cells tend to attach to the Petri-dish surface) were collected using a rubber cell scraper (BD Falcon #353086) and transferred to the 15 ml tube as well. Cells were spun for 10 minutes at 250g and 4°C. The supernatant was discarded, the cell pellets were snap-frozen at -80°C until use.

For the CLIP experiment that was conducted in primary neurons we extracted cells from 20 animals. The cell number was not determined.

5.5 Author's Contributions

DJJ contributed the most. DJJ, RG, and MZ conceived the project. DJJ co-wrote the introduction and discussion with MZ, DJJ, RG, and MZ cowrote the results part. DJJ performed all experiments with help of HW (primary neurons extraction), RG performed all bioinformatic analyses with help of FG (alternatvie splicing analysis). BS provided murine brains from PWScr$^{m+/p+}$ and PWScr$^{m+/p-}$ mice.

5.6 Supplementary Figures



Supplementary Figure 5.1 **Murine neuroblastoma N1E-115 cells** (A) undifferentiated at 72 hours and (B) differentiated at 72 hours of differentiation.

Chapter 6 Conclusion

In this thesis we have thoroughly characterized snoRNAs and the landscape of snoRNA-mediated RNA modifications in several cellular systems as well as in murine brain tissues with various high-throughput sequencing approaches.

In CHAPTER 2 we have extensively studied snoRNA processing and investigated the potential role of snoRNA-derived processing products in post-transcriptional gene silencing. In particular we have shown that virtually all C/D box snoRNAs follow a defined processing pattern: in the mature form the C/D box snoRNAs retain very sharply defined 5' and 3' termini, ending four to five nucleotides upstream of the C box and two to five nucleotides downstream of the D box, respectively. This finding suggests that, similar to other small RNAs, snoRNAs are trimmed presumably by exonucleases, to boundaries that are determined by the proteins with which these small RNAs are complexed, which is in line with the current understanding of snoRNA biogenesis. Furthermore, we have identified several novel snoRNA loci in the human genome including C/D box-like snoRNAs that do not exhibit the typical C-D'-C'-D box architecture, but rather lack mostly the D' and C' boxes. These loci give rise to small RNAs that are bound by the core snoRNPs. Interestingly, some C/D box-like snoRNAs are very short, 27-46 nucleotides in length, and we termed them mini-snoRNAs. To gain insight into the function of the novel snoRNAs - both highly conserved C/D box as well as H/ACA box snoRNAs - that we identified, we sought to determine whether they have canonical targets. Indeed, for some of them, we could identify canonical target sites in rRNA and snRNA. We were especially interested to find out whether the non-conserved C/D box-like snoRNAs particularly the mini-snoRNAs, could guide 2'-O-methylation. Although we did identify several putative interaction sites with a computational screen, none of these correspond to known modification sites and thus, the targets of these C/D box-like snoRNAs, if there are any, remain to be validated.

Although snoRNAs are best known for guiding modifications of rRNAs and snRNAs, in our snoRNP-PAR-CLIP datasets we repeatedly identified several non-coding RNAs including vault RNA1-2, 7SK RNA, and 7SL, that are not known to be subject of snoRNA-guided modification. We tested these targets with additional experimental methods, confirming that they carry 2'-O-methylations and pseudouridines. We tried to identify C/D box snoRNAs that could guide the observed 2'-O-methylation, but we did not find sequences complementary to the modification sites. For pseudouridine sites found in 7SK RNA and 7SL RNA we predicted potential guiding H/ACA box snoRNAs. Further we identified cross-linked H/ACA box snoRNAs in PAR-CLIP experiments conducted with C/D box snoRNP core proteins, and similarly we observed cross-linking of several C/D box snoRNAs in the Dyskerin PAR-CLIP. This could be caused by the close spatial arrangement of snoRNPs on the target molecule, or could indicate that H/ACA box snoRNAs and C/D box snoRNAs guide modifications on each other. Supporting this notion, by primer extension assays we identified several 2'-O-methylations and pseudouridines in C/D-and H/ACA box snoRNAs. However, a drawback of the primer extension assay is that stoppages could be caused not only by 2'-O-methylation but can also have other causes. Therefore, the 2'-O-methylations that we identified with this assay need to investigated further. In contrast, we can be fairly certain that we identified bona fide pseudouridylation sites. Particularly, in the case of the C/D box SNORD35A we were able to identify five putative pseudouridylated residues but no convincing guiding sequence in a known H/ACA box snoRNA. This suggests either that even more snoRNAs remain to be identified or that these pseudouridylations are caused by a protein-only mechanism not requiring guidance by H/ACA box snoRNAs. Taken together, these results indicate that snoRNAs may guide RNA modifications in a wide spectrum of RNAs that goes beyond the canonical rRNA and snRNA sites. However, more work needs to be done to validate this hypothesis.

Another question that we wanted to answer was whether snoRNA give rise to processing products that function as miRNAs in human cells, as this phenomenon was observed in *Giardia lamblia* [233]. To this end we first quantified the abundance of snoRNA-derived small RNAs in HEK293 cells. Consistent with other studies [125] we found that these processing products are indeed abundant. However, when we performed Ago2-IP-seq experiment we did not find evidence that these sdRNAs efficiently associate with Ago2 to act as miRNAs. We thus conclude that a microRNA-like function of snoRNA-derived small RNAs is an exception rather than a rule. Although here we did not confirm a general miRNA-like function of sdRNAs in human cells (albeit validating the one case that was described in the literature before), it is important to keep in mind that snoRNAs can assume alternative roles, as was previously and also most recently demonstrated for some C/D box snoRNAs that are involved in the modulation of alternative pre-mRNA splicing [92, 234]. This is achieved through direct RNA-RNA interactions without the methylation of the target RNA [234]. In the context of the snoRNA-derived small RNAs future work will clarify whether they are mere by-products of snoRNA biogenesis or whether they carry out some function in the cell.

In **Chapter 3** we have combined two recent experimental high-throughput methods (CLIP-seq and RiboMeth-seq) with computational modelling to extensively study and describe C/D box snoRNA-target interactions and 2’-O-methylation transcriptome-wide in human cells. We determined that many snoRNAs that were previously considered as orphan indeed guide 2’-O-methylation. Interestingly, they seem to act on canonical rRNA and snRNA sites that are targeted by other C/D box snoRNAs. We also observed numerous interactions between C/D box snoRNAs and mRNAs in CLIP-seq experiments. Subsequently, we conducted RiboMeth-seq experiments to test whether these interactions also result in 2’-O-methylation. Although the majority of the canonical 2’-O-Me sites were detected with RiboMeth-seq, the sites that we identified from the snoRNA-mRNA chimeras do not seem to undergo 2’O-methylation that is detectable with RiboMeth-seq. One explanation for this observation is that a low expression of the examined 2’-O-methylated RNA could hinder their identification. This could be a problem for the identification of modification sites in targets that are expressed in a tissue-specific manner, in mRNAs or lncRNAs, whose abundance per cell is much lower compared to the rRNAs. Further data that we obtained with the RTL-P approach also suggests that at least some of the novel sites of snoRNA-target interaction that emerged from our study are not methylated with 100% efficiency. Thus, an increased sensitivity of detection of 2’-O-Me sites may confirm the role of newly discovered interactions in 2’-O-ribose methylation. It is also possible that the captured snoRNA-mRNA chimera are an artefact of the CLIP protocol, where the guide snoRNA is ligated to an RNA without this being functionally relevant. It could however also be that these interactions do occur *in vivo* but do not result in 2’-O-methylation. This possibility does not seem unreasonable in the light of a very recent study that identified C/D box snoRNAs with dual functions: 2’-O-methylation and alternative pre-mRNA splicing [234]. Taken together, the significance of snoRNA expression changes can be studied with the techniques we established here. Our approach would be particularly interesting to apply to study the role of C/D box snoRNAs in development, across cell types or tissues, or in health/disease contexts where snoRNAs are deregulated.

In a broader sense, it is fascinating to see how the task of identifying 2’-O-methylations and the assignment of their snoRNA guides, once a very tedious endeavor that involved lengthy gene-by-gene analyses, has now been dramatically facilitated by high-throughput technologies. Although identified sites still need to be validated in low-throughput, e.g. by reverse transcription-based methods or by mass spectrometry, *de novo* identification of 2'-O-Me's and their guide RNAs has been significantly simplified.

Notably, other RNA modifications such as pseudouridines or N6-methyladenosines can now also be studied in high-throughput. Advancements of high-throughput sequencing protocols will further stimulate research in the relatively young field of epitranscriptomics. **CHAPTER 4** illustrates how we developed a high-throughput method (RiM-seq) to validate 2'-O-methylations. It can be considered as an alternative independent validation method, since all three approaches (CLIP-seq, RiboMeth-seq, and RiM-seq) are based on different biochemical principles.

In **CHAPTER 5** we have extensively studied the orphan SNORD116 which are implicated in a rare neurodevelopmental disorder called Prader-Willi syndrome. Previously it was reported that in pluripotent cells, sno-lncRNA emerging from the PWS locus associate with the RBFOX2 splicing regulator, thereby regulating splicing of RBFOX2-targets. We did not observe a change in splicing in Rbfox2-regulated targets in brains obtained from wild-type (wt) and PWScr$^{m+/p-}$ mice that do not express Snord116 and decided to evaluate the canonical functions of the PWS-associated SNORD116. Because SNORD116 snoRNAs have most of the features of a genuine C/D box snoRNA we re-considered the question whether they have a snoRNP guide function in 2'-O-methylation. As previously reported [93], we have shown in snoRNP-CLIP experiments that SNORD116 snoRNAs associate with the snoRNP core proteins. We also observed, that SNORD116 association with Fibrillarin is more pronounced in human neuronal cell lines when compared with neuronal cell lines of murine origin. Interestingly SNORD115, also a snoRNA with C/D box architecture that was initially thought to be one of the key players in PWS, was not captured extensively in our snoRNP-CLIP experiments. In spite of their association with core snoRNPs, neither RiboMeth-seq nor analysis of guide RNA-target hybrids from CLIP-seq experiments revealed convincing sites of 2'O-methylation guided by SNORD116. Nevertheless analysis of the 2'-O-methylation landscape in brains of wt mice and PWScr$^{m+/p-}$ mice with RiboMeth-seq revealed reproducible differences in some transcripts. One of these corresponds to the cholecystokinin triacontatriapeptide (Cck), a hormone that mediates satiety after food intake. We are currently validating these candidate target sites with mass spectrometry and RTL-P in both wt and knock-out conditions.

Additionally we have identified 15 new 2'-O-Me's in both 18S and 28S rRNA, guided by canonical snoRNA guides, that were supported by both CLIP-seq and RiboMeth-seq. Hence we expand here the mouse snoRNA atlas by providing new sites of 2'-O-methylation.

It is plausible that SNORD116 also interacts with other proteins than the core snoRNP proteins. To identify further potential protein binding partners of SNORD116, snythetic oligonucleotides that are coupled to biotin at their 5' or 3' ends to facilitate their purification on streptavidin-coupled beads can be used. The immobilized synthetic oligonucleotides then can be incubated with cellular extracts and protein(s) bound on the oligos could be then identified by mass spectrometry. To confirm that these interactions occur *in vivo*, the identified proteins can be co-immunoprecipitated from tissue/cell line lysate and the bound RNA is isolated and analyzed by northern blot or qPCR.

An alternative approach to CLIP-seq that can reveal direct guide RNA-target RNA interactions is to use synthetic RNA oligonucleotides corresponding to SNORD116 that are conjugated with psoralen and biotin at various positions within the oligonucleotide. Psoralen has been previously efficiently used to induce covalent bonds between RNA hybrids [235]. Such chemically conjugated oligos can be transfected in neuron-derived cell lines, e.g. SH-SY5Y cells. Upon UV exposure specific adducts between the SNORD116-mimic and their target RNAs are formed. These cross-linked hybrids are purified on streptavidin beads and RNA isolated using standard a Trizol protocol. Target RNA thus obtained is initially fragmented to facilitate sequencing. Sequence reads obtained are mapped to the genome and computational tools can be used to determine the exact binding site of the snoRNA to its target. This approach may be able to reveal RNA targets of SNORD116 without a prior assumption about the ribonucleoprotein complex that is involved. However, experience with this approach applied to miRNAs indicates that a double purification, which requires knowledge of the proteins in the complex, is much more efficient for target enrichment. This may be especially important if the small RNA has a small number of targets, as would be expected for SNORD116.

Upon discovery of an RNA target(s) their biological relevance will have to be confirmed. Therefore, several functional assays will clarify the consequence of the snoRNA interaction with its target, namely, whether upon binding of the snoRNA the target undergoes modification or cleavage or whether snoRNAs simply act as chaperones in assisting proper folding of their target RNAs. Similarly, if the target is an mRNA it is curcial to determine whether SNORD116 affects the translation rate of its target. Depending on the nature of the target RNA it is likely that diverse experimental strategies will have to be pursued. It is important to realize that not all C/D box snoRNA necessarily are involved in RNA nucleotide modification. For instance, U3 snoRNA, which is a C/D box snoRNA that localizes to the nucleolus [93, 107], is involved in rRNA processing. It has been suggested that snoRNAs may act like chaperones to assist folding of target RNA and hence facilitate RNA processing [236].

Recent novel findings have highlighted the role of IPW116 in the PWS disorder [218]. IPW116 is a lncRNA hosting in its intron the SNORD116 snoRNA genes. IPW116 was shown to regulate the expression of an imprinted locus distinct from the PWS region. Interestingly, aberrant expression from that locus results in a disorder with a PWS-like phenotype [237, 238]. Since in most PWS patients the loss of SNORD116 expression is also accompanied by a loss of IPW116 expression, it will be important to decouple IPW116 and SNORD116 contributions to PWS development to fully understand this disorder.

Overall, deciphering of the molecular mechanisms behind the SNORD116 function has potentially many implications. The gained experimental data would be critical in training predictive models to identify non-canonical snoRNA-target interactions, because only a handful of non-canonical targets, insufficient to derive general computational models, are available today. Uncovering unconventional targets would enable one to estimate to what extent snoRNAs participate in cellular networks that regulate gene expression and would therefore provide novel insights into snoRNA-mediated gene regulation. Most importantly, identification of targets of the snoRNAs implicated in PWS would mean a major advancement towards the development of a suitable treatment.

References

1. Collins, F.S., et al., *A vision for the future of genomics research.* Nature, 2003. **422**(6934): p. 835-47.
2. Tripp, S. and M. Grueber, *Economic Impact of the Human Genome Project - Battelle Memorial Institute*. 2011.
3. Hardison, R.C., *Comparative genomics.* PLoS Biol, 2003. **1**(2): p. E58.
4. Consortium, I.H., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
5. Maher, B., *Personal genomes: The case of the missing heritability.* Nature, 2008. **456**(7218): p. 18-21.
6. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease.* N Engl J Med, 2010. **363**(2): p. 166-76.
7. Mitchell, K.J., *What is complex about complex disorders?* Genome Biol, 2012. **13**(1): p. 237.
8. Miko, I. and L. LeJenua, *Essentials of Genetics*. 2009, NPG Education: Cambridge, MA.
9. *The Human Genome Project Timeline*. last reviewed April 2013, National Human Genome Research Institute.
10. Bártová, E., et al., *Histone modifications and nuclear architecture: a review.* J Histochem Cytochem, 2008. **56**(8): p. 711-21.
11. Chambeyron, S. and W.A. Bickmore, *Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription.* Genes Dev, 2004. **18**(10): p. 1119-30.
12. Cao, R., et al., *Role of histone H3 lysine 27 methylation in Polycomb-group silencing.* Science, 2002. **298**(5595): p. 1039-43.
13. Bannister, A.J., et al., *Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain.* Nature, 2001. **410**(6824): p. 120-4.
14. Lachner, M., et al., *Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins.* Nature, 2001. **410**(6824): p. 116-20.
15. Schotta, G., et al., *A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin.* Genes Dev, 2004. **18**(11): p. 1251-62.
16. Smith, Z.D. and A. Meissner, *DNA methylation: roles in mammalian development.* Nat Rev Genet, 2013. **14**(3): p. 204-20.
17. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control.* Nat Rev Genet, 2012. **13**(9): p. 613-26.
18. Ogbourne, S. and T.M. Antalis, *Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes.* Biochem J, 1998. **331 (Pt 1)**: p. 1-14.
19. Pennacchio, L.A., et al., *Enhancers: five essential questions.* Nat Rev Genet, 2013. **14**(4): p. 288-95.
20. Wittkopp, P.J. and G. Kalay, *Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence.* Nat Rev Genet, 2012. **13**(1): p. 59-69.
21. Merkhofer, E.C., P. Hu, and T.L. Johnson, *Introduction to cotranscriptional RNA splicing.* Methods Mol Biol, 2014. **1126**: p. 83-96.
22. Chen, M. and J.L. Manley, *Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches.* Nat Rev Mol Cell Biol, 2009. **10**(11): p. 741-54.
23. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.
24. *List of human proteins in the Uniprot Human reference proteome, accessed February 2016*.
25. Lodish, H., et al., *Molecular Cell Biology 7th*. August 2012, W. H. Freeman, Palgrave Macmillan.
26. Gruber, A.R., et al., *Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors.* Wiley Interdiscip Rev RNA, 2014. **5**(2): p. 183-96.
27. Chalfie, M., H.R. Horvitz, and J.E. Sulston, *Mutations that lead to reiterations in the cell lineages of C. elegans.* Cell, 1981. **24**(1): p. 59-69.
28. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
29. Grosshans, H., et al., *The temporal patterning microRNA let-7 regulates several transcription factors at the larval to adult transition in C. elegans.* Dev Cell, 2005. **8**(3): p. 321-30.
30. Giraldez, A.J., et al., *MicroRNAs regulate brain morphogenesis in zebrafish.* Science, 2005. **308**(5723): p. 833-8.
31. Jiang, Q., et al., *miR2Disease: a manually curated database for microRNA deregulation in human disease.* Nucleic Acids Res, 2009. **37**(Database issue): p. D98-104.
32. Ha, M. and V.N. Kim, *Regulation of microRNA biogenesis.* Nat Rev Mol Cell Biol, 2014. **15**(8): p. 509-24.
33. Wilczynska, A. and M. Bushell, *The complexity of miRNA-mediated repression.* Cell Death Differ, 2015. **22**(1): p. 22-33.

34. Londin, E., et al., *Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs.* Proc Natl Acad Sci U S A, 2015. **112**(10): p. E1106-15.

35. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.

36. Khorshid, M., et al., *A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets.* Nat Methods, 2013. **10**(3): p. 253-5.

37. Gumienny, R. and M. Zavolan, *Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G.* Nucleic Acids Res, 2015. **43**(18): p. 9095.

38. Agarwal, V., et al., *Predicting effective microRNA target sites in mammalian mRNAs.* Elife, 2015. **4**.

39. Paraskevopoulou, M.D., et al., *DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W169-73.

40. Enright, A.J., et al., *MicroRNA targets in Drosophila.* Genome Biol, 2003. **5**(1): p. R1.

41. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.* Cell, 2010. **141**(1): p. 129-41.

42. Chi, S.W., et al., *Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.* Nature, 2009. **460**(7254): p. 479-86.

43. Farazi, T.A., et al., *Identification of distinct miRNA target regulation between breast cancer molecular subtypes using AGO2-PAR-CLIP and patient datasets.* Genome Biol, 2014. **15**(1): p. R9.

44. Leung, A.K., et al., *Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs.* Nat Struct Mol Biol, 2011. **18**(2): p. 237-44.

45. Zhang, X., et al., *MicroRNA directly enhances mitochondrial translation during muscle differentiation.* Cell, 2014. **158**(3): p. 607-19.

46. Krell, J., et al., *TP53 regulates miRNA association with AGO2 to remodel the miRNA-mRNA interaction network.* Genome Res, 2016. **26**(3): p. 331-41.

47. Grosswendt, S., et al., *Unambiguous identification of miRNA:target site interactions by different types of ligation reactions.* Mol Cell, 2014. **54**(6): p. 1042-54.

48. Kudla, G., et al., *Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast.* Proc Natl Acad Sci U S A, 2011. **108**(24): p. 10010-5.

49. Helwak, A., et al., *Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding.* Cell, 2013. **153**(3): p. 654-65.

50. Schwartz, S., *Cracking the epitranscriptome.* RNA, 2016. **22**(2): p. 169-74.

51. TB, J. and C. RD, *Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculnic acid, the nucleic acid of the tubercle bacillis.* 1925, J. Am. Chem. Soc. p. 47 (11), pp 2838–2844.

52. WYATT, G.R. and S.S. COHEN, *A new pyrimidine base from bacteriophage nucleic acids.* Nature, 1952. **170**(4338): p. 1072-3.

53. Korlach, J. and S.W. Turner, *Going beyond five bases in DNA sequencing.* Curr Opin Struct Biol, 2012. **22**(3): p. 251-61.

54. Chadwick, L.H., *The NIH Roadmap Epigenomics Program data resource.* Epigenomics, 2012. **4**(3): p. 317-24.

55. *Blueprint Epigenome*.

56. Dai, B. and T.P. Rasmussen, *Global epiproteomic signatures distinguish embryonic stem cells from differentiated cells.* Stem Cells, 2007. **25**(10): p. 2567-74.

57. Khoury, G.A., R.C. Baliban, and C.A. Floudas, *Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database.* Sci Rep, 2011. **1**.

58. J, A., *The Proteome: Discovering the Structure and Function of Proteins.* 2008, Nature Education.

59. Meyer, K.D., et al., *Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons.* Cell, 2012. **149**(7): p. 1635-46.

60. Dominissini, D., et al., *Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq.* Nature, 2012. **485**(7397): p. 201-6.

61. Meyer, K.D. and S.R. Jaffrey, *The dynamic epitranscriptome: N6-methyladenosine and gene expression control.* Nat Rev Mol Cell Biol, 2014. **15**(5): p. 313-26.

62. Linder, B., et al., *Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome.* Nat Methods, 2015. **12**(8): p. 767-72.

63. Schaefer, M., *RNA 5-Methylcytosine Analysis by Bisulfite Sequencing.* Methods Enzymol, 2015. **560**: p. 297-329.

64. Schwartz, S., et al., *Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA.* Cell, 2014. **159**(1): p. 148-62.

65. Birkedal, U., et al., *Profiling of ribose methylations in RNA by high-throughput sequencing.* Angew Chem Int Ed Engl, 2015. **54**(2): p. 451-5.
66. Dieci, G., M. Preti, and B. Montanini, *Eukaryotic snoRNAs: a paradigm for gene expression flexibility.* Genomics, 2009. **94**(2): p. 83-8.
67. Karijolich, J. and Y.T. Yu, *Spliceosomal snRNA modifications and their function.* RNA Biol, 2010. **7**(2): p. 192-204.
68. Filipowicz, W. and V. Pogacić, *Biogenesis of small nucleolar ribonucleoproteins.* Curr Opin Cell Biol, 2002. **14**(3): p. 319-27.
69. Kiss-László, Z., Y. Henry, and T. Kiss, *Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA.* EMBO J, 1998. **17**(3): p. 797-807.
70. Lapinaite, A., et al., *The structure of the box C/D enzyme reveals regulation of RNA methylation.* Nature, 2013. **502**(7472): p. 519-23.
71. Lin, J., et al., *Structural basis for site-specific ribose methylation by box C/D RNA protein complexes.* Nature, 2011. **469**(7331): p. 559-63.
72. Reichow, S.L., et al., *The structure and function of small nucleolar ribonucleoproteins.* Nucleic Acids Res, 2007. **35**(5): p. 1452-64.
73. Tollervey, D., et al., *The small nucleolar RNP protein NOP1 (fibrillarin) is required for pre-rRNA processing in yeast.* EMBO J, 1991. **10**(3): p. 573-83.
74. Kiss, T., *Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs.* EMBO J, 2001. **20**(14): p. 3617-22.
75. Tycowski, K.T., M.D. Shu, and J.A. Steitz, *A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3.* Genes Dev, 1993. **7**(7A): p. 1176-90.
76. Dennis, P.P., A. Omer, and T. Lowe, *A guided tour: small RNA function in Archaea.* Mol Microbiol, 2001. **40**(3): p. 509-19.
77. McPheeters, D.S., P. Fabrizio, and J. Abelson, *In vitro reconstitution of functional yeast U2 snRNPs.* Genes Dev, 1989. **3**(12B): p. 2124-36.
78. McPheeters, D.S. and J. Abelson, *Mutational analysis of the yeast U2 snRNA suggests a structural similarity to the catalytic core of group I introns.* Cell, 1992. **71**(5): p. 819-31.
79. Fabrizio, P., D.S. McPheeters, and J. Abelson, *In vitro assembly of yeast U6 snRNP: a functional assay.* Genes Dev, 1989. **3**(12B): p. 2137-50.
80. Yu, Y.T., M.D. Shu, and J.A. Steitz, *Modifications of U2 snRNA are required for snRNP assembly and pre-mRNA splicing.* EMBO J, 1998. **17**(19): p. 5783-95.
81. Zhao, X. and Y.T. Yu, *Pseudouridines in and near the branch site recognition region of U2 snRNA are required for snRNP biogenesis and pre-mRNA splicing in Xenopus oocytes.* RNA, 2004. **10**(4): p. 681-90.
82. Ségault, V., et al., *In vitro reconstitution of mammalian U2 and U5 snRNPs active in splicing: Sm proteins are functionally interchangeable and are essential for the formation of functional U2 and U5 snRNPs.* EMBO J, 1995. **14**(16): p. 4010-21.
83. Decatur, W.A. and M.J. Fournier, *rRNA modifications and ribosome function.* Trends Biochem Sci, 2002. **27**(7): p. 344-51.
84. King, T.H., et al., *Ribosome structure and activity are altered in cells lacking snoRNPs that form pseudouridines in the peptidyl transferase center.* Mol Cell, 2003. **11**(2): p. 425-35.
85. Tollervey, D., et al., *Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly.* Cell, 1993. **72**(3): p. 443-57.
86. Zebarjadian, Y., et al., *Point mutations in yeast CBF5 can abolish in vivo pseudouridylation of rRNA.* Mol Cell Biol, 1999. **19**(11): p. 7461-72.
87. Auffinger, P. and E. Westhof, *Rules governing the orientation of the 2'-hydroxyl group in RNA.* J Mol Biol, 1997. **274**(1): p. 54-63.
88. Helm, M., *Post-transcriptional nucleotide modification and alternative folding of RNA.* Nucleic Acids Res, 2006. **34**(2): p. 721-33.
89. Davis, D.R., *Stabilization of RNA stacking by pseudouridine.* Nucleic Acids Res, 1995. **23**(24): p. 5020-6.
90. Ballarino, M., et al., *The cotranscriptional assembly of snoRNPs controls the biosynthesis of H/ACA snoRNAs in Saccharomyces cerevisiae.* Mol Cell Biol, 2005. **25**(13): p. 5396-403.
91. Tycowski, K.T., M.D. Shu, and J.A. Steitz, *A mammalian gene with introns instead of exons generating stable RNA products.* Nature, 1996. **379**(6564): p. 464-6.
92. Kishore, S. and S. Stamm, *The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C.* Science, 2006. **311**(5758): p. 230-2.

93. Vitali, P., et al., *ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs.* J Cell Biol, 2005. **169**(5): p. 745-53.
94. Scott, M.S. and M. Ono, *From snoRNA to miRNA: Dual function regulatory non-coding RNAs.* Biochimie, 2011. **93**(11): p. 1987-92.
95. Lestrade, L. and M.J. Weber, *snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs.* Nucleic Acids Res, 2006. **34**(Database issue): p. D158-62.
96. Montanaro, L., D. Treré, and M. Derenzini, *Nucleolus, ribosomes, and cancer.* Am J Pathol, 2008. **173**(2): p. 301-10.
97. Stepanov, G.A., et al., *Regulatory role of small nucleolar RNAs in human diseases.* Biomed Res Int, 2015. **2015**: p. 206849.
98. Gallagher, R.C., et al., *Evidence for the role of PWCR1/HBII-85 C/D box small nucleolar RNAs in Prader-Willi syndrome.* Am J Hum Genet, 2002. **71**(3): p. 669-78.
99. Sahoo, T., et al., *Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster.* Nat Genet, 2008. **40**(6): p. 719-21.
100. Darzacq, X., et al., *Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs.* EMBO J, 2002. **21**(11): p. 2746-56.
101. Clouet d'Orval, B., et al., *Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp.* Nucleic Acids Res, 2001. **29**(22): p. 4518-29.
102. Tollervey, D. and T. Kiss, *Function and synthesis of small nucleolar RNAs.* Curr Opin Cell Biol, 1997. **9**(3): p. 337-42.
103. Darzacq, X. and T. Kiss, *Processing of intron-encoded box C/D small nucleolar RNAs lacking a 5',3'-terminal stem structure.* Mol Cell Biol, 2000. **20**(13): p. 4522-31.
104. Brown, J.W., M. Echeverria, and L.H. Qu, *Plant snoRNAs: functional evolution and new modes of gene expression.* Trends Plant Sci, 2003. **8**(1): p. 42-9.
105. McKeegan, K.S., et al., *A dynamic scaffold of pre-snoRNP factors facilitates human box C/D snoRNP assembly.* Mol Cell Biol, 2007. **27**(19): p. 6782-93.
106. Kiss, T., E. Fayet-Lebaron, and B.E. Jády, *Box H/ACA small ribonucleoproteins.* Mol Cell, 2010. **37**(5): p. 597-606.
107. Lafontaine, D.L. and D. Tollervey, *Birth of the snoRNPs: the evolution of the modification-guide snoRNAs.* Trends Biochem Sci, 1998. **23**(10): p. 383-8.
108. Richard, P., et al., *A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs.* EMBO J, 2003. **22**(16): p. 4283-93.
109. Nicoloso, M., et al., *Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2'-O-ribose methylation of rRNAs.* J Mol Biol, 1996. **260**(2): p. 178-95.
110. Kiss-László, Z., et al., *Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs.* Cell, 1996. **85**(7): p. 1077-88.
111. Cavaillé, J., M. Nicoloso, and J.P. Bachellerie, *Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides.* Nature, 1996. **383**(6602): p. 732-5.
112. Ganot, P., M.L. Bortolin, and T. Kiss, *Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs.* Cell, 1997. **89**(5): p. 799-809.
113. Bortolin, M.L., P. Ganot, and T. Kiss, *Elements essential for accumulation and function of small nucleolar RNAs directing site-specific pseudouridylation of ribosomal RNAs.* EMBO J, 1999. **18**(2): p. 457-69.
114. Lee, Y.S., et al., *A novel class of small RNAs: tRNA-derived RNA fragments (tRFs).* Genes Dev, 2009. **23**(22): p. 2639-49.
115. Haussecker, D., et al., *Human tRNA-derived small RNAs in the global regulation of RNA silencing.* RNA, 2010. **16**(4): p. 673-95.
116. Nicolas, F.E., et al., *Biogenesis of Y RNA-derived small RNAs is independent of the microRNA pathway.* FEBS Lett, 2012. **586**(8): p. 1226-30.
117. Persson, H., et al., *The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs.* Nat Cell Biol, 2009. **11**(10): p. 1268-71.
118. Zywicki, M., K. Bakowska-Zywicka, and N. Polacek, *Revealing stable processing products from ribosome-associated small RNAs by deep-sequencing data analysis.* Nucleic Acids Res, 2012. **40**(9): p. 4013-24.
119. Kawaji, H., et al., *Hidden layers of human small RNAs.* BMC Genomics, 2008. **9**: p. 157.
120. Ender, C., et al., *A human snoRNA with microRNA-like functions.* Mol Cell, 2008. **32**(4): p. 519-28.
121. Taft, R.J., et al., *Small RNAs derived from snoRNAs.* RNA, 2009. **15**(7): p. 1233-40.

122. Shen, M., et al., *Direct cloning of double-stranded RNAs from RNase protection analysis reveals processing patterns of C/D box snoRNAs and provides evidence for widespread antisense transcript expression.* Nucleic Acids Res, 2011. **39**(22): p. 9720-30.

123. Jung, C.H., et al., *Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data.* BMC Genomics, 2010. **11**: p. 77.

124. Langenberger, D., et al., *deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns.* Bioinformatics, 2012. **28**(1): p. 17-24.

125. Li, Z., et al., *Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs.* Nucleic Acids Res, 2012. **40**(14): p. 6787-99.

126. Scott, M.S., et al., *Human box C/D snoRNA processing conservation across multiple cell types.* Nucleic Acids Res, 2012. **40**(8): p. 3676-88.

127. Li, W., A.A. Saraiya, and C.C. Wang, *The profile of snoRNA-derived microRNAs that regulate expression of variant surface proteins in Giardia lamblia.* Cell Microbiol, 2012. **14**(9): p. 1455-73.

128. Kishore, S., et al., *The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing.* Hum Mol Genet, 2010. **19**(7): p. 1153-64.

129. Brameier, M., et al., *Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs.* Nucleic Acids Res, 2011. **39**(2): p. 675-86.

130. Kishore, S., et al., *A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins.* Nat Methods, 2011. **8**(7): p. 559-64.

131. Khorshid, M., C. Rodak, and M. Zavolan, *CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins.* Nucleic Acids Res, 2011. **39**(Database issue): p. D245-52.

132. Hertel, J., I.L. Hofacker, and P.F. Stadler, *SnoReport: computational identification of snoRNAs with unknown targets.* Bioinformatics, 2008. **24**(2): p. 158-64.

133. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

134. org, h.w.e., *ENSEMBL release 65*.

135. Burge, S.W., et al., *Rfam 11.0: 10 years of RNA families.* Nucleic Acids Res, 2013. **41**(Database issue): p. D226-32.

136. Yan, D., et al., *Identification and analysis of intermediate size noncoding RNAs in the human fetal brain.* PLoS One, 2011. **6**(7): p. e21652.

137. Zhang, Y., et al., *Systematic identification and characterization of chicken (Gallus gallus) ncRNAs.* Nucleic Acids Res, 2009. **37**(19): p. 6562-74.

138. Marz, M., et al., *Animal snoRNAs and scaRNAs with exceptional structures.* RNA Biol, 2011. **8**(6): p. 938-46.

139. Yang, J.H., et al., *snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome.* Nucleic Acids Res, 2006. **34**(18): p. 5112-23.

140. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.

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

142. Tafer, H., et al., *RNAsnoop: efficient target prediction for H/ACA snoRNAs.* Bioinformatics, 2010. **26**(5): p. 610-6.

143. Kehr, S., et al., *PLEXY: efficient target prediction for box C/D snoRNAs.* Bioinformatics, 2011. **27**(2): p. 279-80.

144. Berninger, P., et al., *Conserved generation of short products at piRNA loci.* BMC Genomics, 2011. **12**: p. 46.

145. Valen, E., et al., *Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes.* Nat Struct Mol Biol, 2011. **18**(9): p. 1075-82.

146. Cole, C., et al., *Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs.* RNA, 2009. **15**(12): p. 2147-60.

147. Yamasaki, S., et al., *Angiogenin cleaves tRNA and promotes stress-induced translational repression.* J Cell Biol, 2009. **185**(1): p. 35-42.

148. Liao, J.Y., et al., *Deep sequencing of human nuclear and cytoplasmic small RNAs reveals an unexpectedly complex subcellular distribution of miRNAs and tRNA 3' trailers.* PLoS One, 2010. **5**(5): p. e10563.

149. Bernhart, S.H. and I.L. Hofacker, *From consensus structure prediction to RNA gene finding.* Brief Funct Genomic Proteomic, 2009. **8**(6): p. 461-71.

150. Schubert, T., et al., *Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin.* Mol Cell, 2012. **48**(3): p. 434-44.

151. Ule, J., et al., *Nova regulates brain-specific splicing to shape the synapse.* Nat Genet, 2005. **37**(8): p. 844-52.

152. Hafner, M., et al., *Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing.* Methods, 2008. **44**(1): p. 3-12.

153. Hoffmann, S., et al., *Fast mapping of short sequences with mismatches, insertions and deletions using index structures.* PLoS Comput Biol, 2009. **5**(9): p. e1000502.

154. html, h.g.u.e.E.d., ***ENCODE data coordination center at UCSC***.

155. Nawrocki, E.P., D.L. Kolbe, and S.R. Eddy, *Infernal 1.0: inference of RNA alignments.* Bioinformatics, 2009. **25**(10): p. 1335-7.

156. Maden, B.E., et al., *Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA.* Biochimie, 1995. **77**(1-2): p. 22-9.

157. Maden, B.E., *Mapping 2'-O-methyl groups in ribosomal RNA.* Methods, 2001. **25**(3): p. 374-82.

158. Ofengand, J., M. Del Campo, and Y. Kaya, *Mapping pseudouridines in RNA molecules.* Methods, 2001. **25**(3): p. 365-73.

159. Morla, A.O., et al., *Reversible tyrosine phosphorylation of cdc2: dephosphorylation accompanies activation during entry into mitosis.* Cell, 1989. **58**(1): p. 193-203.

160. Pines, J. and T. Hunter, *Isolation of a human cyclin cDNA: evidence for cyclin mRNA and protein regulation in the cell cycle and for interaction with p34cdc2.* Cell, 1989. **58**(5): p. 833-46.

161. Elvin, P. and C.W. Evans, *Cell adhesiveness and the cell cycle: correlation in synchronized Balb/c 3T3 cells.* Biol Cell, 1983. **48**(1): p. 1-9.

162. Li, S. and C.E. Mason, *The pivotal regulatory landscape of RNA modifications.* Annu Rev Genomics Hum Genet, 2014. **15**: p. 127-50.

163. Saletore, Y., et al., *The birth of the Epitranscriptome: deciphering the function of RNA modifications.* Genome Biol, 2012. **13**(10): p. 175.

164. Lee, M., B. Kim, and V.N. Kim, *Emerging roles of RNA modification: m(6)A and U-tail.* Cell, 2014. **158**(5): p. 980-7.

165. Sun, W.J., et al., *RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data.* Nucleic Acids Res, 2016. **44**(D1): p. D259-65.

166. Karijolich, J., A. Kantartzis, and Y.T. Yu, *RNA modifications: a mechanism that modulates gene expression.* Methods Mol Biol, 2010. **629**: p. 1-19.

167. Heilman, K.L., R.A. Leach, and M.T. Tuck, *Internal 6-methyladenine residues increase the in vitro translation efficiency of dihydrofolate reductase messenger RNA.* Int J Biochem Cell Biol, 1996. **28**(7): p. 823-9.

168. Tuck, M.T., P.E. Wiehl, and T. Pan, *Inhibition of 6-methyladenine formation decreases the translation efficiency of dihydrofolate reductase transcripts.* Int J Biochem Cell Biol, 1999. **31**(8): p. 837-51.

169. Wang, X., et al., *N6-methyladenosine-dependent regulation of messenger RNA stability.* Nature, 2014. **505**(7481): p. 117-20.

170. Wang, X., et al., *N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency.* Cell, 2015. **161**(6): p. 1388-99.

171. Tycowski, K.T., et al., *Modification of U6 spliceosomal RNA is guided by other small RNAs.* Mol Cell, 1998. **2**(5): p. 629-38.

172. Caffarelli, E., et al., *Processing of the intron-encoded U16 and U18 snoRNAs: the conserved C and D boxes control both the processing reaction and the stability of the mature snoRNA.* EMBO J, 1996. **15**(5): p. 1121-31.

173. Fatica, A. and D. Tollervey, *Insights into the structure and function of a guide RNP.* Nat Struct Biol, 2003. **10**(4): p. 237-9.

174. Chen, C.L., et al., *Exploration of pairing constraints identifies a 9 base-pair core within box C/D snoRNA-rRNA duplexes.* J Mol Biol, 2007. **369**(3): p. 771-83.

175. Breda, J., et al., *Quantifying the strength of miRNA-target interactions.* Methods, 2015. **85**: p. 90-9.

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

177. Kishore, S., et al., *Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing.* Genome Biol, 2013. **14**(5): p. R45.

178. Machyna, M., et al., *The coilin interactome identifies hundreds of small noncoding RNAs that traffic through Cajal bodies.* Mol Cell, 2014. **56**(3): p. 389-99.

179. Martin, G., et al., *Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length.* Cell Rep, 2012. **1**(6): p. 753-63.

180. Chan, P.P. and T.M. Lowe, *GtRNAdb: a database of transfer RNA genes detected in genomic sequence.* Nucleic Acids Res, 2009. **37**(Database issue): p. D93-7.

181. Tafer, H. and I.L. Hofacker, *RNAplex: a fast tool for RNA-RNA interaction search.* Bioinformatics, 2008. **24**(22): p. 2657-63.

182. Do, C.B., D.A. Woods, and S. Batzoglou, *CONTRAfold: RNA secondary structure prediction without physics-based models.* Bioinformatics, 2006. **22**(14): p. e90-8.

183. Seabold, S.a.P., J, *Statsmodels: Econometric and statistical modeling with python.* 2010, *of the 9th Python in Science Conference*.
184. Cunningham, F., et al., *Ensembl 2015.* Nucleic Acids Res, 2015. **43**(Database issue): p. D662-9.
185. Rosenbloom, K.R., et al., *The UCSC Genome Browser database: 2015 update.* Nucleic Acids Res, 2015. **43**(Database issue): p. D670-81.
186. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* 2011, EMBnet. Journal. p. 17, -10.
187. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
188. Dong, Z.W., et al., *RTL-P: a sensitive approach for detecting sites of 2'-O-methylation in RNA molecules.* Nucleic Acids Res, 2012. **40**(20): p. e157.
189. Pintard, L., et al., *Trm7p catalyses the formation of two 2'-O-methylriboses in yeast tRNA anticodon loop.* EMBO J, 2002. **21**(7): p. 1811-20.
190. Wang, J., et al., *WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W77-83.
191. Enright, C.A., et al., *5'ETS rRNA processing facilitated by four small RNAs: U14, E3, U17, and U3.* RNA, 1996. **2**(11): p. 1094-9.
192. Fujita, T., et al., *Identification of non-coding RNAs associated with telomeres using a combination of enChIP and RNA sequencing.* PLoS One, 2015. **10**(4): p. e0123387.
193. Bailey, T.L., et al., *MEME: discovering and analyzing DNA and protein sequence motifs.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W369-73.
194. Hofacker, I.L., *Vienna RNA secondary structure server.* Nucleic Acids Res, 2003. **31**(13): p. 3429-31.
195. Mückstein, U., et al., *Thermodynamics of RNA-RNA binding.* Bioinformatics, 2006. **22**(10): p. 1177-82.
196. Mattick, J.S., *Non-coding RNAs: the architects of eukaryotic complexity.* EMBO Rep, 2001. **2**(11): p. 986-91.
197. Morin, R.D., et al., *Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.* Genome Res, 2008. **18**(4): p. 610-21.
198. Chen, H.M. and S.H. Wu, *Mining small RNA sequencing data: a new approach to identify small nucleolar RNAs in Arabidopsis.* Nucleic Acids Res, 2009. **37**(9): p. e69.
199. Granneman, S., et al., *Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs.* Proc Natl Acad Sci U S A, 2009. **106**(24): p. 9613-8.
200. Guy, M.P., et al., *Defects in tRNA Anticodon Loop 2'-O-Methylation Are Implicated in Nonsyndromic X-Linked Intellectual Disability due to Mutations in FTSJ1.* Hum Mutat, 2015. **36**(12): p. 1176-87.
201. Motorin, Y., et al., *Identification of modified residues in RNAs by reverse transcription-based methods.* Methods Enzymol, 2007. **425**: p. 21-53.
202. Kiss, T. and W. Filipowicz, *Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns.* Genes Dev, 1995. **9**(11): p. 1411-24.
203. Bratkovič, T. and B. Rogelj, *The many faces of small nucleolar RNAs.* Biochim Biophys Acta, 2014. **1839**(6): p. 438-43.
204. Dupuis-Sandoval, F., M. Poirier, and M.S. Scott, *The emerging landscape of small nucleolar RNAs in cell biology.* Wiley Interdiscip Rev RNA, 2015. **6**(4): p. 381-97.
205. Yin, Q.F., et al., *Long noncoding RNAs with snoRNA ends.* Mol Cell, 2012. **48**(2): p. 219-30.
206. Nicholls, R.D. and J.L. Knepper, *Genome organization, function, and imprinting in Prader-Willi and Angelman syndromes.* Annu Rev Genomics Hum Genet, 2001. **2**: p. 153-75.
207. Buiting, K., *Prader-Willi syndrome and Angelman syndrome.* Am J Med Genet C Semin Med Genet, 2010. **154C**(3): p. 365-76.
208. Runte, M., et al., *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, 2001. **10**(23): p. 2687-700.
209. de los Santos, T., et al., *Small evolutionarily conserved RNA, resembling C/D box small nucleolar RNA, is transcribed from PWCR1, a novel imprinted gene in the Prader-Willi deletion region, which Is highly expressed in brain.* Am J Hum Genet, 2000. **67**(5): p. 1067-82.
210. Meguro, M., et al., *Large-scale evaluation of imprinting status in the Prader-Willi syndrome region: an imprinted direct repeat cluster resembling small nucleolar RNA genes.* Hum Mol Genet, 2001. **10**(4): p. 383-94.
211. Cavaillé, J., et al., *Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization.* Proc Natl Acad Sci U S A, 2000. **97**(26): p. 14311-6.
212. Heisler, L.K., H.M. Chu, and L.H. Tecott, *Epilepsy and obesity in serotonin 5-HT2C receptor mutant mice.* Ann N Y Acad Sci, 1998. **861**: p. 74-8.

213. Runte, M., et al., *Exclusion of the C/D box snoRNA gene cluster HBII-52 from a major role in Prader-Willi syndrome.* Hum Genet, 2005. **116**(3): p. 228-30.

214. Ding, F., et al., *Lack of Pwcr1/MBII-85 snoRNA is critical for neonatal lethality in Prader-Willi syndrome mouse models.* Mamm Genome, 2005. **16**(6): p. 424-31.

215. de Smith, A.J., et al., *A deletion of the HBII-85 class of small nucleolar RNAs (snoRNAs) is associated with hyperphagia, obesity and hypogonadism.* Hum Mol Genet, 2009. **18**(17): p. 3257-65.

216. Bieth, E., et al., *Highly restricted deletion of the SNORD116 region is implicated in Prader-Willi Syndrome.* Eur J Hum Genet, 2015. **23**(2): p. 252-5.

217. Galiveti, C.R., et al., *Differential regulation of non-protein coding RNAs from Prader-Willi Syndrome locus.* Sci Rep, 2014. **4**: p. 6445.

218. Stelzer, Y., et al., *The noncoding RNA IPW regulates the imprinted DLK1-DIO3 locus in an induced pluripotent stem cell model of Prader-Willi syndrome.* Nat Genet, 2014. **46**(6): p. 551-7.

219. Skryabin, B.V., et al., *Deletion of the MBII-85 snoRNA gene cluster in mice results in postnatal growth retardation.* PLoS Genet, 2007. **3**(12): p. e235.

220. Ding, F., et al., *SnoRNA Snord116 (Pwcr1/MBII-85) deletion causes growth deficiency and hyperphagia in mice.* PLoS One, 2008. **3**(3): p. e1709.

221. Gabriel, J.M., et al., *A transgene insertion creating a heritable chromosome deletion mouse model of Prader-Willi and angelman syndromes.* Proc Natl Acad Sci U S A, 1999. **96**(16): p. 9258-63.

222. Yang, T., et al., *A mouse model for Prader-Willi syndrome imprinting-centre mutations.* Nat Genet, 1998. **19**(1): p. 25-31.

223. Tsai, T.F., et al., *Paternal deletion from Snrpn to Ube3a in the mouse causes hypotonia, growth retardation and partial lethality and provides evidence for a gene contributing to Prader-Willi syndrome.* Hum Mol Genet, 1999. **8**(8): p. 1357-64.

224. Nicholls, R.D., *Incriminating gene suspects, Prader-Willi style.* Nat Genet, 1999. **23**(2): p. 132-4.

225. Bortolin-Cavaillé, M.L. and J. Cavaillé, *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, 2012. **40**(14): p. 6800-7.

226. Weyn-Vanhentenryck, S.M., et al., *HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism.* Cell Rep, 2014. **6**(6): p. 1139-52.

227. Jaskiewicz, L., et al., *Argonaute CLIP--a method to identify in vivo targets of miRNAs.* Methods, 2012. **58**(2): p. 106-12.

228. Shen, S., et al., *rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.* Proc Natl Acad Sci U S A, 2014. **111**(51): p. E5593-601.

229. Christov, C.P., et al., *Functional requirement of noncoding Y RNAs for human chromosomal DNA replication.* Mol Cell Biol, 2006. **26**(18): p. 6993-7004.

230. Gardiner, T.J., et al., *A conserved motif of vertebrate Y RNAs essential for chromosomal DNA replication.* RNA, 2009. **15**(7): p. 1375-85.

231. Yoshihama, M., A. Nakao, and N. Kenmochi, *snOPY: a small nucleolar RNA orthological gene database.* BMC Res Notes, 2013. **6**: p. 426.

232. Côté, C.D., et al., *Hormonal signaling in the gut.* J Biol Chem, 2014. **289**(17): p. 11642-9.

233. Saraiya, A.A. and C.C. Wang, *snoRNA, a novel precursor of microRNA in Giardia lamblia.* PLoS Pathog, 2008. **4**(11): p. e1000224.

234. Falaleeva, M., et al., *Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing.* Proc Natl Acad Sci U S A, 2016.

235. Imig, J., et al., *miR-CLIP capture of a miRNA targetome uncovers a lincRNA H19-miR-106a interaction.* Nat Chem Biol, 2015. **11**(2): p. 107-14.

236. Borovjagin, A.V. and S.A. Gerbi, *U3 small nucleolar RNA is essential for cleavage at sites 1, 2 and 3 in pre-rRNA and determines which rRNA processing pathway is taken in Xenopus oocytes.* J Mol Biol, 1999. **286**(5): p. 1347-63.

237. Hosoki, K., et al., *Maternal uniparental disomy 14 syndrome demonstrates prader-willi syndrome-like phenotype.* J Pediatr, 2009. **155**(6): p. 900-903.e1.

238. Hordijk, R., et al., *Maternal uniparental disomy for chromosome 14 in a boy with a normal karyotype.* J Med Genet, 1999. **36**(10): p. 782-5.