

IDENTIFICATION OF SMALL NON-CODING RNA TARGETS  
USING COMPUTATIONAL PREDICTIONS AND HIGH  
THROUGHPUT SEQUENCING DATA

**Inauguraldissertation**

zur

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

von

RAFAŁ WOJCIECH GUMIENNY  
aus Polen

Basel, 2018

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel  
[edoc.unibas.ch](http://edoc.unibas.ch)

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät  
auf Antrag von

Prof. Dr. Mihaela Zavolan  
*Fakultätsverantwortliche und Dissertationsleiterin*

Prof. Dr. Witold Filipowicz  
*Korreferent*

Basel, den 22.03.2016

Prof. Dr. Jörg Schibler  
*Dekan*

Only death can finish the fight, everything else only interrupts the  
fighting.

— Andrzej Sapkowski

Dedicated to my newborn son Mieszko and my wife Ola.



## ABSTRACT

---

Although non-coding RNAs have been known for a relatively long time, they have largely been viewed as having a passive role in cellular processes. Ribosomal RNAs were thought to serve as a scaffold of the protein production machinery, tRNAs as transporters for amino acids and even the protein-coding mRNAs were seen as a passive template for protein synthesis. A sign of a revolution was perhaps visible with the discovery of small nuclear and nucleolar RNAs (snRNAs and snoRNAs), but it was not until the year 2000 with the discovery of the let-7 microRNAs that the revolution began. In microRNAs, a totally new layer of gene regulation was uncovered, leading to the revision of our understanding of the types of RNA molecules and their roles in the cell: RNAs are not viewed anymore as passive, but capable of regulating a vast number of cellular processes. Most often they serve as guides for ribonucleoprotein complexes that regulate the processing or expression of target RNAs. Recently developed high-throughput technologies enabled identification of many long and small non-coding RNAs. However, the identification of their targets has remained challenging, in spite of the recently proposed high-throughput sequencing-based or computational approaches. In this work, we aimed to identify the targets of two large groups of RNAs: miRNAs (as well as their exogenous counterparts, the small interfering RNAs (siRNAs)) and snoRNAs.

MicroRNAs (miRNAs) are ~21 nucleotides long non-coding RNAs that induce gene expression silencing by guiding Argonaute proteins to target mRNAs. This pathway is exploited to silence gene expression by means of siRNAs, that are designed to silence the expression of specific genes. Functioning similar to miRNAs, siRNAs act not only on the intended target, but also other transcripts called off-targets. In a first sub-project we combined the MIRZA biophysical model of miRNA-target interaction that was previously developed in the group with structural and sequence features of putative target sites to efficiently predict both miRNA and siRNA targets on a genome-wide scale. Starting from the observation that guide RNAs can be captured bound to their targets in high-throughput data sets, we then revisited the identification of snoRNA targets. Although snoRNAs are known for more than 30 years, some of them do not have known targets. To reveal these, we have developed novel methods to analyse the data obtained by crosslinking and immunoprecipitation of core snoRNP proteins as well as by the RiboMeth-seq method that detects 2'-O-methylation sites. This work provides high-quality sets of miRNA

and snoRNA targets and sets the ground for further analysis of their complex network of interactions.

## PUBLICATIONS

---

This PhD thesis is based on the following publications:

1. **Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G.**

Authors: RAFAL GUMIENNY and Mihaela Zavolan

Gumienny and Zavolan *Nucleic Acids Res.* (2015)

2. **Quantifying the strength of miRNA-target interactions.**

Authors: Jeremie Breda, Andrzej J. Rzepiela, RAFAL GUMIENNY, Erik van Nimwegen, Mihaela Zavolan

Breda et al. *Methods* (2015)

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

Authors: RAFAL GUMIENNY, Dominik J Jedlinski, Georges Martin, Arnau Vina-Vilaseca, Mihaela Zavolan

Gumienny et al. *Nucleic Acids Res.* (2016)



## ACKNOWLEDGMENTS

---

As Newton said we can see further because we stand on the shoulders of giants. It would not be possible to reach the place I am now without the support of my family, friends, colleges and mentors.

First of all, I would like to express my deepest gratitude to Mihaela Zavolan for giving me the opportunity to work in her group. Her constant support, never-ending stream of ideas and passion for the science gave me the courage and motivation to work better every day.

I would like to express my special thanks to Prof. Witold Filipowicz for his time and effort put into reviewing this work.

I am grateful to all current and past members of the Zavolan Lab: Joao, Hadi, Jean, Andrea, Andrzej, Souvik, Alex, Andreas J and R, Afzal, Jeremie, Foivos, Dominik, Ralf, Alexandra, Joana, Nitish, Beatrice and Georges for sharing with me their skills and knowledge that were essential to this work.

Many thanks to all our collaborators from the RNPnet Marie Curie Initial Training Networks. I point my gratitude especially towards Jernej Ule and Ina Hupertz for letting me learn new techniques and to see my work from a different perspective. I would like to take, as well, the opportunity to thank many unnamed people working with the Marie Curie Actions in European Commission.

My work would be impossible without workable PC and computer cluster. Thus, I would like to thank our IT department, especially Jan Welker and Pablo Escobar for endless patience in answering my (sometimes) silly questions about another bash script.

I am really thankful to Yvonne Steger, Rita Manohar and Sarah Güthe for the great and hard job they make on administrative site thereby making our life much easier.

I am much obliged to my friends here in Basel and there in Poland. Great and long non-scientific discussions had unprecedented influence on my ideas and they allowed me to stay sane through these years.

I would like to express my gratitude to my parents, who always encouraged me to take independent paths and gave me all possible support I could think of. The same gratitude I have for my brothers and my sister. I would like to thank my parents- and brother-in-law for all the encouragement they gave me.

Last but not least I would like to thank my wife Ola. She is my best friend and my salt of the earth.

To everyone I have mentioned and those I have forgotten: Thank you!



# CONTENTS

---

1	INTRODUCTION	1
1.1	Gene expression regulation by RNA interference	1
1.2	miRNAs	2
1.2.1	Biogenesis of miRNAs	2
1.2.2	Target recognition by miRNAs	4
1.2.3	Experimental identification of miRNA targets	4
1.2.4	Computational identification of miRNA targets	5
1.2.5	Prediction of siRNA off-targets	7
1.3	snoRNAs	7
1.3.1	The enigmatic orphan snoRNAs	9
1.3.2	Role of snoRNAs and 2'-O-methylations in ribosome biogenesis and function	9
2	MIRZA-G	11
2.1	Introduction	11
2.2	Materials and Methods	13
2.2.1	miRNA and siRNA transfection data	13
2.2.2	miRNA and siRNA sequences	14
2.2.3	3' UTR sequences	15
2.2.4	Comparisons with other miRNA target prediction methods	15
2.2.5	Prediction of siRNA off-targets with DIANA-microT and TargetScan Context+	16
2.2.6	Putative binding sites	16
2.2.7	Feature definition and computation	17
2.2.8	Training of the generalized linear model	19
2.2.9	Evaluation of model performance	20
2.2.10	Analysis of the siRNA screen	21
2.3	Results	21
2.3.1	Features of miRNA binding sites that are active in mRNA degradation	21
2.3.2	Performance of the model in predicting the response of mRNAs to miRNA transfection	24
2.3.3	Prediction of siRNA off-target effects	26
2.3.4	Analysis of siRNA screening results with MIRZA-G	28
2.4	Discussion	30
2.5	Supplementary Data	32
2.6	Funding	32
2.7	Acknowledgments	32
2.8	Conflict of interest statement	32
3	QUANTIFYING THE STRENGTH OF MIRNA-TARGET INTERACTIONS	33

3.1	Introduction	33
3.2	Inferring the strength of miRNA–target interactions	37
3.2.1	Input data: Argonaute-bound RNA fragments. Output: general model of miRNA–target interaction MIRZA–CLIP	38
3.2.2	Input data: chimeric miRNA–mRNA sequence reads. Output: general model of miRNA–target interaction MIRZA–CHIMERA	40
3.2.3	Input data: chimera of a specific miRNA with target sites. Output: miRNA-specific model of interaction with the target	40
3.3	Results	41
3.3.1	Evaluating the models on biochemical data	41
3.3.2	Genome-wide prediction of miRNA targets	43
3.3.3	Wide range of MIRZA quality scores across the targets of a given miRNA	45
3.3.4	Evaluation of the MIRZA models on miRNA transfection data	46
3.3.5	Inferring a MIRZA model from biochemical data	48
3.4	Discussion and perspective	50
3.5	Acknowledgments	52
4	SNORNA CHIMERAS	53
4.1	Introduction	53
4.2	Materials and Methods	55
4.2.1	CLIP of snoRNP core proteins	55
4.2.2	Identification of snoRNA–target chimera	55
4.2.3	Feature definition and computation	57
4.2.4	RiboMeth-seq	59
4.2.5	Validation of 2′-O-methylation sites with RTL-P	62
4.2.6	Validation of 2′-O-methylation at G2435 in 28S with mass spectrometry	62
4.3	Results	62
4.3.1	Crosslinking and immunoprecipitation of core snoRNPs captures snoRNA–target site chimeras	62
4.3.2	A model to identify high-confidence snoRNA–target chimeras	63
4.3.3	Chimeric reads reveal novel C/D box snoRNA target sites within structural RNAs	64
4.3.4	Redundant targeting of known sites of 2′-O-ribose methylation by multiple snoRNAs	66
4.3.5	Identification of snoRNA-guided 2′-O-Me sites with RiboMeth-seq	67

4.3.6	Position G2435 in the 28S rRNA, captured in interaction with SNORD2, is partially methylated	70
4.3.7	mRNAs captured in chimeras with snoRNAs do not show evidence of 2'-O-methylation	71
4.4	Discussion	71
4.5	Supplementary Data	74
4.6	Funding	74
4.7	Acknowledgements	74
5	SUMMARY	75
A	MIRZA-G SUPPLEMENTARY MATERIALS	79
B	SNORNA CHIMERAS SUPPLEMENTARY MATERIALS	87
	BIBLIOGRAPHY	93

## LIST OF FIGURES

---

Figure 1	Simplified biogenesis of miRNAs.	3	
Figure 2	Value of t-statistic in comparing the mean values of features used in the model.	23	
Figure 3	Comparative evaluation of various models.	25	
Figure 4	Relationship between the prediction scores obtained with different target prediction methods and the extent of down-regulation of target mRNAs upon siRNA transfections.	27	
Figure 5	SiRNA off-targets in the TGF- $\beta$ pathway.	29	
Figure 6	Accumulation of miRNA targets as a result of increasing transcription.	35	
Figure 7	Distribution of the number of targets of individual miRNAs that were captured from individual ESCs.	36	
Figure 8	Crystal structure of the human AGO-2 protein.	38	
Figure 9	The 27 parameters of various MIRZA model variants.	39	
Figure 10	Relationship between the nucleotide composition of the miRNA and the type of hybrids in which the miRNA was captured.	42	
Figure 11	Ratio of binding free energies of mismatched and perfectly matched hybrids.	43	
Figure 12	Diagram of the approach for predicting miRNA targets with MIRZA-G.	45	
Figure 13	Distribution of the MIRZA quality scores of target sites of individual miRNAs.	46	
Figure 14	Relationship between prediction score and the extent of mRNA downregulation.	49	
Figure 15	Root mean square difference (RMSD).	51	
Figure 16	Features that are relevant for the identification snoRNA-target interactions based on chimeric reads.	64	
Figure 17	Characterization of the model for inferring snoRNA-target interactions from chimeric reads.	65	
Figure 18	Schematic representation of snoRNA-target interactions that are predicted based on chimeric reads from CLIP experiments.	66	
Figure 19	Schematic representation of the data supporting the interaction of SNORD80 and SNORD118	67	
Figure 20	Analysis of RiboMeth-seq data.	68	

Figure 21	Location of snoRNA interaction sites and 2'-O-ribose methylation. 69
Figure 22	SNORD2-guided 2'-O-methylation of G2435 in the 28S rRNA 70
Figure S1	Empirical cumulative distribution function of MIRZA target quality scores 79
Figure S2	Comparison of the down-regulation of targets predicted with RNAup or CONTRAfold 80
Figure S3	Optimization of scaling factor (K) and threshold ( $\tau$ ). 81
Figure S4	Comparative evaluation of the performance of various models on proteomics data. 82
Figure S5	Comparison of performance of the different models in predicting off-targets of siRNAs 82
Figure S6	Performance comparison of various models on individual siRNA transfections from Birmingham et al. [22] 83
Figure S7	Performance comparison of various models on individual siRNA transfections from Jackson et al. [23] 84
Figure S8	Correlation between the z-score of an siRNA in the TGF- $\beta$ -screen and the average score 85
Figure S9	Distribution of the Smith-Waterman scores 87
Figure S10	Absolute value of t-statistic. 88
Figure S11	Targeted LC-MS/MS analysis of the G2435 site in 28S rRNA 89
Figure S12	Intersection of interactions. 91
Figure S13	Positive and negative controls for RTL-P 92

## LIST OF TABLES

---

Table 1	Summary of the experimental data sets that were used to train the model and evaluate its performance. 13
Table 2	Four alternative MIRZA-G models. 26
Table 3	Chimeras of the indicated miRNAs, obtained from the data set of Grosswendt et al. [69] were used to infer MIRZA-Class I and MIRZA-Class IV models. 41
Table 4	Summary of the experimental data sets that were used to train the model and evaluate its performance. 47



## INTRODUCTION

---

The flow of information in the cell, described by so-called “Central Dogma of Molecular Biology” which was proposed by Francis Crick in 1958 [4], links the DNA code to the protein effectors through transcription and translation. In this view, RNAs molecules were considered mainly as various types of intermediates: mRNA serving as a template, tRNA as transporting molecule and rRNA performing the translation. This simple perspective dominated molecular biology for more than 30 years and was supported by much research in enzymology and gene regulation [5, 6]. The more recently discovered snRNAs and snoRNAs [7–9], which were subsequently associated with splicing [10] and with modification of RNA nucleotides [11], respectively, fell into the same paradigm. However, everything changed with the discovery of small regulatory RNAs and RNA interference [12, 13]. In contrast to the previously known classes of RNAs, the miRNAs had a regulatory role, a function previously associated with proteins. However, similar to other classes of relatively small RNAs, miRNAs guide effector complexes to target RNAs through sequence complementarity. This is reminiscent of small nucleolar RNAs (snoRNAs) guiding 2'-O-methylation and pseudouridylation of rRNAs, or small nuclear RNAs (snRNAs) guiding splicing reactions.

### 1.1 GENE EXPRESSION REGULATION BY RNA INTERFERENCE

RNA interference (RNAi) as an inactivation of gene expression by homologous RNA sequences was first observed in plants and in neurospora [14, 15]. However, this multi-faceted mechanism only started to be understood with the work that was carried out by Fire and Mello in the worm *Caenorhabditis elegans*, which served as an important experimental system for the characterization of the RNAi pathway. In worm, double-stranded RNAs are more effective in down-regulating gene expression than antisense RNA alone [13]. The effector molecules are small interfering RNAs (siRNAs). SiRNAs mostly derive from exogenous sources like viral dsRNA or chemically synthesized hairpins (shRNA). In plants it was shown that siRNAs can act as antiviral agents [16]. However, they can be generated endogenously from long or short hairpins, transcripts of inverted repeats, double-stranded RNA generated by convergent transcription or other small-non coding RNAs [17]. Although RNA interference exists in many species in all kingdoms of life, there are vast differences between the number of variants of this pathway, the biogenesis of the

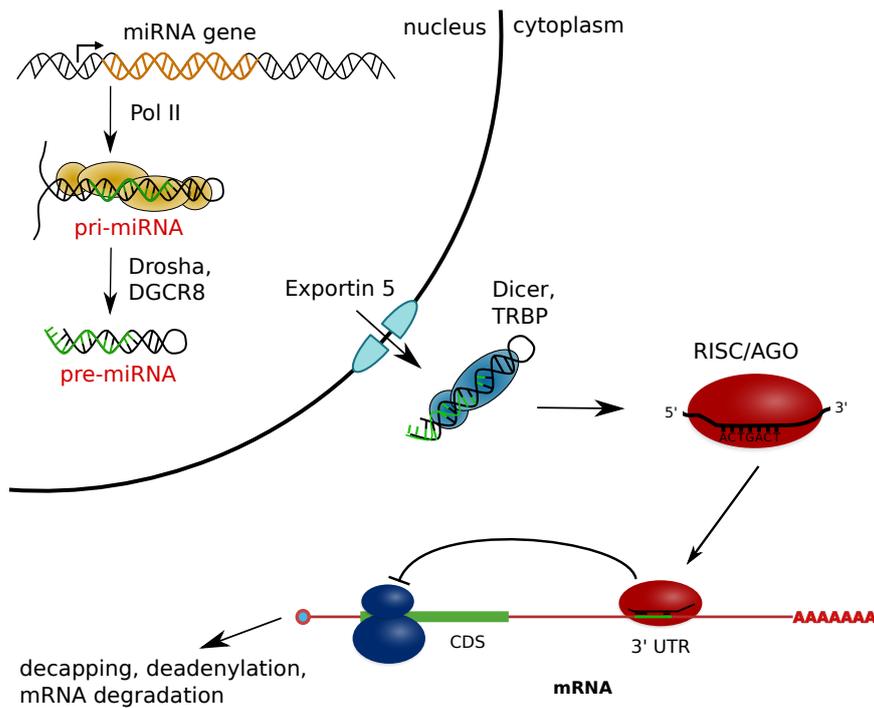
small RNAs, their mechanism of action and downstream consequences. Here, I will focus on mammals and on the specific class of small RNAs which are called microRNAs. However, I will make a short detour into siRNAs that were important in characterizing the pathway. After introduction into the cell, siRNA precursors are cleaved by the enzyme Dicer into 21 double stranded RNAs [18]. One strand of the duplex (called guide strand) is typically loaded into the RNA-induced silencing complex (RISC) and the other one (passenger strand) is degraded [19]. Which strand will be selected as a guide is determined by the thermodynamic asymmetry of the RNA duplex; the strand with less stable pairing at its 5' end becomes the guide [20]. The main component of the RISC complex is an Argonaute (Ago) protein, that may have endonuclease activity [21]. In human Due to their ability to selectively down-regulate gene expression siRNAs can be used as potential drugs. However, because siRNAs share the much of their biogenesis and effector machinery with endogenously-encoded small RNA, namely the miRNA, they are not as specific as desired, but rather act on many transcripts with which they have only partial complementarity. These off-target effects are an important stumbling block in the use of siRNAs as therapeutic agents [22, 23].

## 1.2 MIRNAS

MiRNAs are small non-coding RNAs of 21-22 nucleotides in length, which in the past 10 years became one of the most intensely researched molecular entities. They were first described in *C. elegans* in 1991 and 1993 [12, 24]. They are transcribed from genes that are present in the genomes of almost every species, including viruses [25] and guide the RISC to target RNAs, directing their degradation or interfering with protein translation ultimately leading to down-regulation of target expression [26]. MiRNAs have been found important for many biological processes including cell proliferation, differentiation and apoptosis [27–29], organism development, cancer or schizophrenia [30–32]. It is estimated that more than half of the human genes can be regulated by miRNAs [33]. A miRNA has, on average, over a hundred target mRNAs and each mRNA can be regulated by many miRNAs [33]. Through target mRNA down-regulation, miRNAs may be involved in many dynamic behaviors: fine-tuning gene expression, repressing them strongly and counter-acting 'leakage' in transcription. Many of these effects would lead to miRNAs increasing the robustness and precision of gene expression [34, 35].

### 1.2.1 Biogenesis of miRNAs

Mammalian miRNAs are encoded in the introns of other genes or as independent genes [36, 37], although other pathways can generate



**Figure 1:** Simplified biogenesis of miRNAs.

miRNA-like structures as well [38]. MiRNAs are typically transcribed by RNA polymerase II into long transcripts (pri-miRNA), in which the miRNA fold into stem-loop structures [39]. Pri-miRNAs, which may contain more than one miRNA precursor, are processed, leading to shorter fragments corresponding to individual miRNAs which are called precursor miRNAs (pre-miRNAs). This process is performed by the so-called microprocessor complex, which is composed, in most basic form, of the Drosha RNase III enzyme together with DiGeorge Syndrome Critical Region 8 (DGCR8) protein that recognizes the pri-miRNA structure and facilitates the reaction [40–42]. The released pre-miRNAs are 70 nucleotides long hairpins and possess characteristic 2 nucleotide-long overhangs on their 3' end. This feature is recognized by Exportin-5 that subsequently transports pre-miRNA from nucleus to cytoplasm [43–45]. Some pre-miRNAs, known as mirtrons, are generated directly through splicing of host genes, bypassing the multiprocessor complex, and loading onto Exportin-5 [37]. In the cytoplasm pre-miRNAs are recognized and cleaved by RNase III enzyme Dicer which liberates ~21 nucleotide-long duplex RNAs with characteristic 3' overhangs [18, 46]. As already discussed in the case of siRNAs, one strand (called guide strand) of the duplex is typically loaded into RISC, while the other (passenger strand) is degraded [19]. Simplified pathway of miRNA biogenesis is depicted on Figure 1.

### 1.2.2 Target recognition by miRNAs

MiRNA guide RISC to complementary sequences located usually in the 3' UTRs of protein-coding mRNAs, leads to down-regulation of gene expression through the inhibition of translation initiation and/or mRNA degradation via poly(A) tail cleavage by deadenylase CCR4-NOT [47, 48]. However, miRNAs, besides mRNAs, can bind other RNA species like long non-coding RNAs or circular RNAs [49–52]. The minimal mammalian RISC complex can be composed of only one protein from Argonaute family (AGO). In human those can be one of the four subtypes: AGO1, AGO2, AGO3 and AGO4. From these four, only AGO2 is able to cleave the target directly [53]. Experiments showed that different AGOs can take part in different cellular processes however more investigation is needed to elucidate their specific roles [53]. Early investigations revealed structural features of the interaction between miRNA and its target. In animals, miRNAs interact with the target primarily via so-called 'seed' region which is composed of 6-7 consecutive nucleotides in the 5' end of miRNA that perfectly match their complementary sequences in the target mRNA<sup>1</sup>. In metazoans miRNA seed-complementary motifs are much more conserved across species compared to other 3'UTR motifs of the same length [54, 55]. AGO effectively scans the RNA searching for short 2-4 nucleotides long complementary regions and the interaction becomes stable only when the complementarity extends from 2 to 8 nucleotides [56]. The role of the seed was explained additionally by the structural biology studies of the Argonaute protein. In complex with AGO2 the seed region of miRNA is already in pre-helical form, exposed to the interaction with the target [57]. It was shown, however, that there exist target sites that have no apparent seed match or where seed match is imperfect (referred to as non-canonical target sites). For example, 20% of miR-124 targets have bulged nucleotide in the seed [58, 59]. A recent study suggested that as many as 60% of miRNA targets are non-canonical [60]. Most of the predicted miRNA target sites lack the complementarity to the 3' end of the miRNA. However, it is believed that this fragment might confer specificity to the interaction with the target and can be responsible for the differences in target sets between miRNAs within the same seed families. On the other hand, another school of thought is that members of the same seed family share the majority of their targets [61].

### 1.2.3 Experimental identification of miRNA targets

Currently, the repertoire of experimental methods to identify miRNA targets experimentally is quite extensive, ranging from basic genetic

<sup>1</sup> miRNAs in plants interact with their whole sequence which makes target prediction easier [34]

screening to direct capture of miRNAs ligated to their target. With genetic screening, which one searches for genes that rescue the miRNA loss-of-function phenotype, which also help characterize the function of the miRNA [62, 63]. Another approach is to overexpress or repress a miRNA in a population of cells and globally quantify gene expression changes between treated and untreated cells using microarrays, mRNA-seq [64] or shot-gun proteomics approaches [65]. These approaches are particularly useful in dissecting the pathways that are targeted by miRNAs and in providing the data on which bioinformatics algorithms for miRNA target prediction can be trained and tested [1]. The most up-to-date high-throughput experimental approach to miRNA target identification is based on AGO crosslinking and immunoprecipitation (AGO-CLIP) [66, 67], in which AGO is first crosslinked with UV light (254 nm for 'standard' CLIP and 365 nm for a variant method called photoactivatable ribonucleoside-enhanced CLIP or PAR-CLIP [68]) to targets and/or miRNAs and then pulled down with a specific antibody. Bound RNAs are then sequenced and analysed in order to obtain the AGO-binding sites [35]. The most recent advance in CLIP consists in the capture of AGO loaded with a miRNA which in turn is covalently linked to the target site. This method is called crosslinking, immunoprecipitation and sequencing of hybrids (CLASH) [60]. Chimeric, miRNA-target, reads were also observed also in PAR-CLIP [69]. This strategy of capturing guide RNA-target RNA chimera can, in principle, be applied to other guide RNAs, as I have also done in my work. As one might expect, all of these methods have their pros and cons, in terms of efficacy, coverage, ease of use, etc. Which one should be used in a particular setting depends on all of these factors as well as the biological question pursued [35].

#### 1.2.4 *Computational identification of miRNA targets*

The availability of high-throughput data sets have prompted an explosion in the number of studies that exploit these data to construct empirical models that can be used to uncover novel molecules that share a specific set of properties. In the field of miRNAs, computational approaches have been employed from the very beginning and have strongly facilitate the efforts of identifying miRNA targets [54, 70]. Computational methods have the advantage that they help narrow the search for targets. However, they may be limited by the knowledge of what constitutes a functional target. Most of the current miRNA target prediction algorithms that are in use today enforce the principle that targets should have miRNA seed-complementary regions [71]. However, being so short, the miRNA seed-complementary motifs occur in the transcriptome in large numbers simply by chance and experimental studies found that the mere presence of a miRNA

seed match in an mRNA does not guarantee its repression [61, 72]. Thus, developing algorithms for miRNA target prediction is of great interest. However, because relatively little is known about what contributes to a functionally relevant miRNA-target interaction, miRNA target prediction methods have limited accuracy and the problem remains challenging. The most important improvement in the accuracy of computational miRNA target predictions came with the introduction of structural and sequence features beyond the miRNA seed match into the prediction algorithms. In fact, the determinants of miRNA-target interaction remain largely unknown, because the degree of evolutionary conservation of the seed match (which is taken into account in algorithms like EIMMo or TargetScan [73, 74]) provides the greatest improvement in accuracy, much higher than specific features such as the accessibility of the target site or the nucleotide composition of its neighborhood [33, 54, 73, 75]. Nevertheless, structure and sequence determinants that were found to be important for miRNA target recognition are as follows. Functional target sites tend to be located in the edges of 3'UTR of the mRNA, they are usually surrounded by AU-rich region and they are located on an accessible part of an mRNA [74, 76]. Even with this information at hand slight differences in utilization of aforementioned features lead to different algorithms predicting largely different target gene set. It is therefore necessary to combine computational target prediction with experimental validation by artificially increasing or decreasing miRNA activity in the cell combined with target expression measurements.

Ultimately, the interaction of small RNAs with their targets should obey physico-chemical laws and one would like to be able to explain and predict interactions in terms of biophysical models, rather than effective models, which contain arbitrary features. In our group, we made a first attempt at this with MIRZA [59], a model that is similar to biophysical models of RNA-RNA interactions but aims to capture the fact that in the case of small RNAs, the RNA-RNA interactions take place within ribonucleoprotein complexes, where the proteins modify the energy of interaction between the RNAs. This has been now shown experimentally in a series of studies from the Zamore group [77]. MIRZA provides a rigorous approach to infer principles of small RNA-target interaction and to identify targets without relying on an arbitrary definition of 'seed types'. However, as we mentioned already, we do not fully understand the parameters that are relevant to miRNA-target interactions. Thus, in our models we can attempt to represent explicitly the features that we do know have an effect and rely on evolutionary conservation to capture those parameters that are, at the moment, poorly understood. My first PhD project, described in the second chapter of the thesis, had as goal to supplement the MIRZA biophysical model of miRNA-target interaction

with sequence context information, structural accessibility and evolutionary conservation and to build a novel miRNA target prediction method. The resulting model is MIRZA-G, which was published in 2015 in the journal *Nucleic Acids Research* [1]. As mentioned above, recent studies managed to capture miRNAs complexed with their target sites in the form of chimeric sequence reads in Ago-CLIP data sets. This provides us with the unique opportunity to infer a MIRZA model from experimentally validated miRNA-target interactions, and to evaluate whether such a model improves miRNA target prediction. Thus, in chapter 3 we infer a MIRZA-CHIMERA model from chimeric miRNA-mRNA sequences obtained from CLIP or CLASH experiments and discuss its usefulness for miRNA target site predictions [2].

#### 1.2.5 Prediction of siRNA off-targets

As discussed in the introduction, siRNAs are commonly introduced into cells as double-stranded RNAs, the guide strands being incorporated into RISCs to guide Ago to complementary target mRNA, which are then cleaved and degraded [78]. Because siRNAs and miRNAs share the same downstream machinery that allows them to function, siRNAs act on many so-called off-targets, and not only on their intended on-targets. This complicates substantially interpretation of large-scale siRNA screens and usage of siRNAs in therapies. Thus, there is a great demand to predict the siRNA off-targets. One of the main goals in developing MIRZA-G was to be able to accurately predict siRNA off-targets. We have indeed shown that our model is the most accurate to date, but, as the siRNAs do not act on off-target sites that were subjected to evolutionary selection, the siRNA off-target predictions are not as accurate as miRNA target predictions and there remain opportunities for improving siRNA off-target prediction considerably. Nevertheless, MIRZA-G can be used as a starting point in the interpretation of siRNA screening data, which is the topic of one of the on-going projects in our group.

### 1.3 SNORNAS

The small nucleolar RNAs (snoRNAs) are an abundant class of small non-coding RNAs of length ~70-150 nucleotides that were discovered together with the of small nuclear RNAs (snRNAs) in nuclear extracts [7]. They participate in ribosome biogenesis, guiding the cleavage and site-specific modifications of the ribosomal RNAs (rRNAs) [79, 80]. Based on their sequence and structural characteristics snoRNAs can be divided into two major groups: C/D and H/ACA box snoRNAs. C/D box snoRNAs that guide rRNA cleavage and 2'-O-methylation of riboses are characterized by the presence of C (RUGAUGA, R =

A or G) and D (CUGA) box motifs [11, 81, 82]. H/ACA box snoRNAs (separated as a subgroup in the mid-90s) that guide conversion of uridine to pseudouridine ( $\Psi$ ) are characterized by conserved H (ANANNA, N = any nucleotide) and ACA (ACA trinucleotide) motifs [83–85]. C/D box snoRNAs will be discussed in more detail below. There is another closely related group of snoRNA-like molecules called small Cajal body-specific RNAs (scaRNAs). They have both C/D and H/ACA box motifs and guide 2'-O-methylation and/or pseudouridylation of snRNAs [86, 87].

In addition to their canonical C and D boxes, C/D box snoRNAs contain sometimes additional and less conserved copies of these boxes called the C' and D' boxes [88]. C and D boxes are usually located at the ends of a snoRNA and fold into a characteristic kink-turn motif. These structural and sequence features are essential for the maturation and function of snoRNAs, including binding to core snoRNP proteins, stabilizing the structure and guiding the localization of mature snoRNAs [89]. The 9-21 nucleotide-long antisense elements (ASEs) that guide the methylation of ribose are located 1 or 2 nucleotides upstream of the D and/or D' box sequences. The methyl group will be attached to the target nucleotide that is complementary to the fifth nucleotide upstream of the given box. However, there might be some exceptions and additions to this canonical view of snoRNA-target interaction. SNORD14A possesses ASEs that confer pre-rRNA cleavage and also guide 2'-O-methylation [90, 91]. Additionally some snoRNAs have supporting complementary sequences in other parts of snoRNA that can increase the efficiency of methylation up to five fold [92]. It was shown, however, that only canonical elements are sufficient to induce 2'-O-methylations guided by artificially expressed snoRNA [93]. C/D box snoRNAs are loaded in so-called snoRNP complex composed of four core proteins: fibrillarin (FBL), Nop56, Nop58 and the 15.5kDa protein that binds to C/D motif. These components are universal among eukaryotes [89]. Fibrillarin function as the methyltransferase, catalyzing the transfer of a methyl group to 2'-O-position on the ribose of the targeted nucleotide. Other proteins might be also involved. For example, a recent study showed, using CLIP, that the RBFOX2 binds to many snoRNAs [94].

Over 500 C/D box snoRNAs are currently known (Jorjani et al submitted), which guide the modification of over 100 sites in rRNAs. However, for a substantial number of snoRNAs computational and experimental approaches failed to detect any reliable targets [95, 96]. These, so-called orphan snoRNAs, may have functions beyond guiding rRNA modifications. Recent studies have started to construct a picture of snoRNAs as a diverse group of small non-coding RNAs, possessing diverse expression and processing patterns across tissues, binding a wider range of proteins than those that have been described as core snoRNP components, and taking part in many biological pro-

cesses including cancers [97–102]. All these features make them a current conundrum in biology - although known for such a long time they still seem to have secrets. Thus, it is important to develop new techniques for identification of their targets.

### 1.3.1 *The enigmatic orphan snoRNAs*

Many studies have attempted to identify snoRNA targets [92, 103]. They were especially focused on the targets of orphan snoRNAs and on precisely identifying the guides for positions that are already known to undergo modifications. Up to date, there is almost no known 2'-O-methylation position for which the guide has not been assigned. However, these studies predicted complementarities between orphan snoRNAs and rRNA positions not known to be methylated. It is unknown whether these predicting binding events really take place and whether they result in methylation under specific conditions (cell type, developmental stage etc.), contributing, for example, to so-called specialization of ribosomes [104]. SnoRNA expression profiling in mammals showed that many orphan snoRNAs are differentially expressed in tissues with impressive upregulation of SNORD115 and SNORD116 in the brain [105, 106]. SNORD116 has been also directly linked to the Prader-Willi Syndrome which is a genetic disease caused by the deletion of parental chromosome region containing among the others aforementioned SNORD116 [107–110]. It is characterised by behavioral problems, muscle development abnormalities, mental retardation and in most cases morbid obesity caused by chronic hunger. This strongly suggests that at least some snoRNAs might have tissue-specific functions. Alternatively, binding of snoRNA to predicted targets serves a function different than 2'-O-ribose methylation.

### 1.3.2 *Role of snoRNAs and 2'-O-methylations in ribosome biogenesis and function*

Biogenesis of ribosomes is one of the most complex and most fundamental processes in the cell. It requires lots of energy and more than 200 assembly factors. The main component of the ribosome is the ribosomal RNA (rRNA) that catalyzes the most important steps in mRNA translation [111]. 18S, 5.8S and 28S ribosomal RNAs are organised in the genome in clusters and are transcribed by polymerase I into one large transcript called 47S pre-rRNA that is subsequently cleaved into separate rRNAs [112]. The 5S rRNA is transcribed from separate genes by Pol III [112]. The basic steps of transcription and assembly of rRNAs are facilitated by small nucleolar RNAs that contribute to the processing of rRNA through cleavage and 2'-O-methylation. The first snoRNA that was discovered was U3, which plays a crucial role in the first endonucleolytic cleavages of pre-rRNA, acting as an RNA chap-

erone [113]. This exceptional CD-box snoRNA is longer than “normal” snoRNAs and it is conserved in many species [114]. However the primary role of C/D-box snoRNAs, as mentioned earlier, is to guide 2'-O-methylations of rRNA. Ribosomal RNAs are known to possess more than 100 methylations guided by snoRNAs. 2'-O-methylation is known to play a role in the stability and folding of RNA by decreasing conformational flexibility [115]. It can also alter hydrogen bonding potential and is known to protect the internucleotide bond against hydrolysis, a property which is used in the identification of 2'-O-methylation sites [116]. 2'-O-methylation changes the conformational energy of the nucleotide in solution and also affects the stability of the structure that is assumed by the oligonucleotide chain [117]. These effects depend on the nucleotide and structural context of the modified nucleotide. The fact that the sequences as well as the modifications are evolutionarily conserved indicates that they are functionally relevant [118]. Consistently, lack or aberrant presence of 2'-O-methylations has been described in the context of some diseases [119].

In the fourth chapter of this thesis we have combined high-throughput experimental approaches with novel computational analysis methods to develop a new method to globally assign guide snoRNAs to 2'-O-Me sites in human cells.

## ACCURATE TRANSCRIPTOME-WIDE PREDICTION OF MICRORNA TARGETS AND SMALL INTERFERING RNA OFF-TARGETS WITH MIRZA-G

---

### ABSTRACT

Small interfering RNA (siRNA)-mediated knock-down is a widely used experimental approach to characterizing gene function. Although siRNAs are designed to guide the cleavage of perfectly complementary mRNA targets, acting similarly to microRNAs (miRNAs), siRNAs down-regulate the expression of hundreds of genes to which they have only partial complementarity. Prediction of these siRNA ‘off-targets’ remains difficult, due to the incomplete understanding of siRNA/miRNA–target interactions. Combining a biophysical model of miRNA–target interaction with structure and sequence features of putative target sites we developed a suite of algorithms, MIRZA-G, for the prediction of miRNA targets and siRNA off-targets on a genome-wide scale. The MIRZA-G variant that uses evolutionary conservation performs better than currently available methods in predicting canonical miRNA target sites and in addition, it predicts non-canonical miRNA target sites with similarly high accuracy. Furthermore, MIRZA-G variants predict siRNA off-target sites with an accuracy unmatched by currently available programs. Thus, MIRZA-G may prove instrumental in the analysis of data resulting from large-scale siRNA screens.

*The work presented in this chapter was originally published in Nucleic Acids Research [1]*

### 2.1 INTRODUCTION

MicroRNAs (miRNAs) are ~22 nucleotides long non-coding RNAs that guide Argonaute proteins to RNA targets. By silencing target expression [120], miRNAs take part in the regulation of many processes including cell differentiation and development [121]. Aberrant miRNA expression has been implicated in many diseases, notably in carcinogenesis [122]. The miRNA’s 5’ end, particularly nucleotides 2–7 which are known as the ‘seed’ region [123, 124], is thought to nucleate the miRNA–target interaction. Much experimental and computational work has established that perfect complementarity between the miRNA seed and the target site is important for the interaction (see Pasquinelli et al. for a recent review). Target sites that satisfy this constraint are known as ‘canonical’ while those that do not as ‘non-canonical’. High-throughput experimental studies point to a relatively high preponderance of non-canonical sites [58–60, 69].

Exploiting the miRNA-dependent gene silencing pathway, exogenous small interfering RNAs (siRNAs) have been used as a tool to rapidly silence gene expression [126]. Although an siRNA is designed to be perfectly complementary to its mRNA target, it rapidly became apparent that the transfection of the siRNA affects the expression of many other RNAs that are complementary to the siRNA seed region [22, 23]. These siRNA seed-dependent, ‘off-target’ interactions are frequently responsible for the observed phenotypes, and hamper the use of siRNAs for gene targeting. Nonetheless, large siRNA screens continue to be used to elucidate gene function, and therefore accurate prediction of siRNA off-targets has great practical importance.

One step in this direction has been made by approaches that uncover siRNA ‘off-target’ signatures from mRNA expression data [127, 128]. Prediction of siRNA off-targets has also been attempted [129] although stand-alone programs are not generally available. However, because siRNA off-target effects occur through the miRNA pathway, tools for miRNA target site prediction [75, 130] can also be used to predict siRNA off-targets. An important limitation for this approach is that the strongest indicator of functionality of a putative miRNA target site, namely its evolutionary conservation [124], is unlikely to be relevant for the off-target sites of exogenous siRNAs. Yet it is precisely this feature that is exploited by the most accurate miRNA target prediction methods [33, 73, 131]. Thus, the accuracy of siRNA off-target prediction is probably lower than the accuracy of miRNA target prediction, although such comparisons have not been carried out systematically. Interestingly, a tendency of active siRNA off-target sites to reside in transcript regions that are evolutionarily conserved has been noted [132].

The goal of our work was to develop a method that can predict canonical and non-canonical miRNA targets and siRNA off-targets with comparable accuracy. An important ingredient of our model is the miRNA–target interaction energy predicted by the MIRZA biophysical model that we previously inferred from Argonaute 2 crosslinking and immunoprecipitation (Ago2-CLIP) data [59]. In addition to the MIRZA-predicted energy of interaction, the model includes features that we and others have shown to be predictive for functional miRNA target interactions, such as the nucleotide (nt) composition around putative target sites, their structural accessibility and location within 3′ untranslated regions (3′ UTRs) [73, 74, 76, 133]. We called the resulting miRNA target prediction method MIRZA-G (from MIRZA-Genome-wide). We illustrate the performance of the model on several large-scale data sets and demonstrate that MIRZA-G can help in the interpretation of large-scale siRNA screens.

## 2.2 MATERIALS AND METHODS

2.2.1 *miRNA and siRNA transfection data*

To train the model and evaluate its performance we made use of an extensive set of 26 experiments, carried out by seven different groups, in which the gene expression changes that were induced by the transfection of individual miRNAs were measured [65, 134–139]. A summary of the experimental data sets is given in Table 1. Data were processed as described previously [59] to obtain the  $\log_2$  fold changes in gene expression levels upon transfection of individual miRNAs. The  $\log_2$  fold changes for all used experiments can be found in Supplementary Table S1.

**Table 1:** Summary of the experimental data sets that were used to train the model and evaluate its performance.

REFERENCE	DATA SOURCE (GEO ACCESSION / URL)	MIRNAS IN THE DATA SET
Dahiya et al. [135]	GSE10150	miR-200c, miR-98
Frankel et al. [136]	GSE31397	miR-101
Gennarino et al. [137]	GSE12100	miR-26b, miR-98
Hudson et al. [134]	GSE34893	miR-106b
Leivonen et al.[138]	GSE14847	miR-206, miR-18a mir-193b, miR-302c
Linsley et al. [139]	GSE6838	miR-103, miR-215, miR-17, miR-192, let-7c, miR-106b, lmiR-16, miR-20, miR-15a, miR-141, miR-200a
Selbach et al.[65]	psilac.mdc-berlin.de/download/	miR-155, let-7b, miR-30a, miR-1, miR-16

The gene expression changes induced by 12 different siRNA transfected individually were measured by Birmingham et al. et al. [22] and processed by Dongen, Abreu-Goodger, and Enright et al. [127] to infer siRNA off-target signatures. We obtained the processed data from the supplementary material of this latter study.

Microarray-based measurements of gene expression changes that were induced by the transfection of individual siRNA were also carried out in the study of Jackson et al. (13). From the Gene Expression

Omnibus database <sup>1</sup>, we obtained the gene expression data as SOFT-formatted files (accession GSE5814). The data correspond to transfections of 10 distinct siRNAs:

- PIK<sub>3</sub>CB-6338
- PIK<sub>3</sub>CB-6340
- MAPK<sub>14</sub>-193
- MAPK<sub>14</sub>-pos2-mismatch
- MAPK<sub>14</sub>-pos3-mismatch
- MAPK<sub>14</sub>-pos4-mismatch
- MAPK<sub>14</sub>-pos5-mismatch
- MAPK<sub>14</sub>-pos6-mismatch
- MAPK<sub>14</sub>-pos7-mismatch
- MAPK<sub>14</sub>-pos8-mismatch

The samples were prepared 24 h after transfection. From this study, we also obtained the RefSeq annotations of the probes that were present on the microarray. Each probe was mapped to a RefSeq identifier and subsequently to Entrez Gene <sup>2</sup> identifier. If there were multiple probes per gene, the expression was averaged. For each gene, fold-changes were averaged over replicate experiments.

A more recent siRNA screen aiming to identify regulators of the TGF- $\beta$  pathway [132] used a library of ~21000 siRNAs that were designed to target approximately 6000 human genes that have been previously connected to cancers, including all known phosphatases, kinases and more generally, components of signal transduction pathways. The sequences of these siRNAs were obtained from the supplementary material of the paper. We scanned the set of 3' UTRs (obtained as described in the Section 2.2.3) for matches to the seed regions of all siRNAs included in this screen, obtaining ~50 million distinct matches. For each of these putative target sites, we calculated the associated features, as described below. Finally, we determined per-gene scores for all siRNAs as described in the section 'Computing Transcript/Gene Scores'.

### 2.2.2 *miRNA and siRNA sequences*

miRNA sequences were downloaded from miRBase [140] version 20. The sequences of siRNAs that were used in the experiments described

<sup>1</sup> <http://www.ncbi.nlm.nih.gov/geo/>

<sup>2</sup> <http://www.ncbi.nlm.nih.gov/gene>

above were obtained directly from the supplementary material of the studies that described the data [23, 127, 132]. Some siRNA sequences were shorter than 21 nucleotides (nts). Because the MIRZA model assumes a small RNA sequence of 21 nts, we extended the sequences of these siRNAs to 21 nts with adenines which have been shown to be favorable for the functionality of the siRNA [141]. For the miRNAs whose sequence in miRBase was shorter than 21 nts (a relatively uncommon situation), we extended to 21 nts based on the genomic locus of the miRNA. The correspondence between the names of the miRNAs that were used in the transfection experiments that we analyzed and those in the current version of miRBase is provided in Supplementary Table S2, together with the miRNA sequences.

### 2.2.3 3' UTR sequences

A common stumbling block in comparing the accuracy of miRNA target prediction methods is that stand-alone versions of the software are not always available. Directly comparing the sets of predictions made by different methods is problematic because the set of transcripts/3' UTRs that served as input for target prediction differed from study to study. Because TargetScan was the baseline algorithm with which we compared our results, we used human 3' UTR sequences downloaded from TargetScan v6.2<sup>3</sup> [124] for our predictions.

### 2.2.4 Comparisons with other miRNA target prediction methods

MiRNA target predictions were obtained from the websites corresponding to each of the tools as follows: TargetScan<sup>4</sup>, DIANA-microT<sup>5</sup>, MiRanda mirSVR<sup>6</sup>. Version v3.0 of DIANA-microT <sup>7</sup> allows prediction of targets of individual small RNAs (miRNAs and siRNAs). Therefore, we used version v3.0 of the software to predict siRNA off-targets. We downloaded the predictions generated with DIANA-microT v5.0 (CDS) for the comparative analysis of mRNA and protein-level prediction of miRNA targets. To obtain a gene-level target score for methods that only score individual target sites (TargetScan and mirSVR), we summed up the scores of the target sites predicted in each individual gene.

<sup>3</sup> [http://www.targetscan.org/cgi-bin/targetscan/data\\_download.cgi?db=verte\\_61](http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=verte_61)

<sup>4</sup> [http://www.targetscan.org/cgi-bin/targetscan/data\\_download.cgi?db=verte\\_61](http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=verte_61)

<sup>5</sup> <http://www.microrna.gr/webServer>

<sup>6</sup> <http://www.microrna.org/microrna/getDownloads.do>

<sup>7</sup> <http://www.microrna.gr/microT>

### 2.2.5 Prediction of siRNA off-targets with DIANA-microT and TargetScan Context+

Of the miRNA target prediction tools that have been reported to have high accuracy, DIANA-microT and TargetScan Context+ are accessible and allow prediction of targets not only for miRNAs but also for siRNAs. Therefore, for TargetScan Context+ we downloaded scripts provided on the website<sup>8</sup> and predicted target sites for all siRNAs from the Birmingham et al. et al. [22] and Jackson et al. et al. [23] studies. As for miRNAs, we obtained gene-level target scores by summing up the scores of individual sites within each gene. For DIANA-microT we used the available web server<sup>9</sup> to obtain directly gene-level predictions of siRNA targets. Because some siRNAs yielded no predictions with DIANA-microT (one of the siRNA from Birmingham et al. et al. [22] and five siRNAs from Jackson et al. et al. [23]), in our comparisons of the performance of the methods we used only siRNAs for which all methods tested yielded predictions.

### 2.2.6 Putative binding sites

We focused our analysis and prediction on the following types of binding sites. First, we considered canonical sites in the sense used by TargetScan [124]. Thus, we scanned the 3' UTRs for miRNA seed matches (defined as exact match to the nucleotides 2–8 of mature miRNA or match to nucleotides 2–7 and followed by an adenine). Second, we sought to identify non-canonical sites that would interact strongly with miRNAs. We scanned the entire 3' UTRs with MIRZA<sup>10</sup> using a window of 50 nts, sliding by 30 nts at a time. Validated miRNA target sites in the literature do not surpass a length of 50 nucleotides and at the same time, it is relatively unlikely that such regions contain multiple sites because sites that are too close to each other presumably 'interfere' with each other [142]. We then identified windows with a MIRZA target quality score of at least 50, a score threshold that we chose based on the distribution of MIRZA scores among Ago2-CLIP sites (see Section 2.2.7.1 below). Then, we calculated the best miRNA-mRNA hybrid structure and inferred the region in the mRNA that would hybridize with the miRNA seed. We used this anchor region in the mRNA to define the full miRNA target site, comprising the miRNA seed match and the upstream 21 nts. For each of these sites, we computed the set of features described below. We applied the same procedure to the prediction of siRNA off-target sites.

<sup>8</sup> [http://www.targetscan.org/cgi-bin/targetscan/data\\_download.cgi?db=vert\\_61](http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61)

<sup>9</sup> <http://diana.cslab.ece.ntua.gr/microT/>

<sup>10</sup> Current version at <http://www.clipz.unibas.ch/index.php?r=tools/mirza/Submission/index>

### 2.2.7 Feature definition and computation

#### 2.2.7.1 MIRZA target quality score

Computing the MIRZA target quality score, defined as in [59], was the first step in our transcriptome-wide prediction of miRNA/siRNA target sites. Because the target quality score depends on the length of the putative target site, we used windows of fixed length, 50 nts, in 3' UTRs. To define a minimum target quality score, we reanalyzed the 2998 sites that were previously used by Khorshid et al. et al. [59] to train the MIRZA model. For each site, we identified the miRNA that had the highest target quality score and then computed the highest-scoring hybrid structure between this miRNA and the CLIPed site. After classifying the sites into canonical/non-canonical, we determined the distributions of target quality score for these two categories of sites. We found as before, that the target quality scores were, on average, higher for canonical compared to non-canonical sites (316 compared to 15). The cumulative density function of the scores for the two types of sites showed that a score of 50 allows us to retain most (92%) of the canonical sites and a substantial proportion (18%) of the non-canonical sites, and we therefore chose 50 as a minimum target site quality score (Supplementary [Figure S1](#)).

#### 2.2.7.2 Position of the target site in 3' UTRs

We determined the distance to the closest 3' UTR boundary as the minimum between the distance from the beginning of the seed complementary region to the stop codon and to the poly-A tail.

#### 2.2.7.3 Nucleotide content

The 'Flanks G content' and 'Flanks U content' features were defined as the proportion of G and U nts, respectively, within 50 nt upstream and 50 nt downstream of the miRNA seed-matching region.

#### 2.2.7.4 Accessibility

The structural accessibility of the target site was defined as the probability that the 21 nucleotide long region (anchored on the right-hand side by the nucleotide matching the 5'-most nucleotide of the miRNA seed) is in single-stranded conformation, across all possible secondary structures. This probability was computed with CONTRAfold, a method for RNA secondary structure prediction that is based on conditional log-linear models (CLLMs) [143]. CONTRAfold was applied to the region covering the miRNA seed match, and the 50 nucleotides upstream and 50 nucleotides downstream of the seed match. Computing the partition function over structures in which the

target region was either constrained to be in single-stranded conformation or not (running CONTRAfold with `-partition` and `-constraints` flags, all other parameters left to default values) we could obtain the log-probability that the target site is in single-stranded conformation. We also carried out the entire model training and target prediction procedure using the energy necessary to open the secondary structure of the target region (computed with the RNAup program from the Vienna package [144], as described before in [76]) as a measure of target site accessibility. The results are comparable (Supplementary Figure S2, see Section 2.2.9 for more details), although the top CONTRAfold-based predictions are slightly more down-regulated than the RNAup-based predictions. Thus, we used the CONTRAfold-based measure in the final model.

#### 2.2.7.5 Branch length score

We quantified the selection pressure on putative target sites in terms of a ‘branch length score’ [145], defined as described below. The 3’ UTR sequences were aligned to the human genome (hg19) with GMAP [146]. The pairwise alignments of the human genome (hg19) to the genomes of 41 other species were obtained from UCSC<sup>11</sup>, and then anchored alignments (with the genomic region of the human 3’ UTRs serving as anchor) were constructed as described before [73]. These alignments were used to assess the degree of evolutionary conservation of putative target sites.

The phylogenetic tree of 46 species (including *Homo sapiens*) was downloaded from the UCSC database<sup>12</sup> and the species for which pairwise alignments to human were not available were pruned. For each putative target site in human we carried out the following computation. Based on the alignment of the human 3’ UTRs with all the other species, we extracted the region that corresponded to the putative target site in the human 3’ UTR in all other species. Because the MIRZA target quality score depends on the length of the site, we either padded or trimmed the putative target sites in all of these species to precisely 50 nts. We then computed the target quality score of the putative target sites with the human miRNA, and we considered the target site to be conserved in a species when the target quality score was at least 50. Then, based on the evolutionary distances along the tree provided by UCSC, we computed the fraction of the total evolutionary distance in the phylogenetic tree along which the site was conserved. We called this measure branch length score. All manipulations of the phylogenetic tree were performed with DendroPy package [147]. To assess the accuracy of this measure, we compared the estimates of selection pressure obtained in the manner described above with the posterior probabilities that individual putative target

<sup>11</sup> <http://hgdownload.cse.ucsc.edu/downloads.html#human>

<sup>12</sup> <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/>

site are under evolutionary selective pressure, calculated with the EIMMo method [73]. Because EIMMo only handles canonical sites, we did this comparison for seed-matching miRNA-complementary sites only. The two methods had comparable ability to distinguish between functional and non-functional sites (not shown).

### 2.2.8 Training of the generalized linear model

To train the model, we used only putative canonical sites of miRNAs in the test set (see below). Furthermore, to ensure that the impact of the miRNA can be attributed to specific sites, we analyzed only transcripts that contained a single putative canonical site for the transfected miRNA. For each experiment we extracted the 100 most downregulated and the 100 least-changing (whose log fold-change was closest to 0) transcripts with a single putative miRNA binding site in the 3' UTR. These transcripts provided the 100 positive and the 100 negative target sites in the respective experiment.

For each site we then calculated the features described above: MIRZA target quality score, distance to the 3' UTR boundary, G/U composition of flanking regions, structural accessibility, and branch length score. To assess the prediction power of these features, we carried out two-sample t-tests for the difference of the mean values of a given feature between the positive and negative target sites in each experiment.

Although the experiments show consistent differences between the positive and negative sites for all features that we used in our model, the significance of the difference differs to some extent between experiments (Figure 2). We used the subset of experiments in which the differences between the positive and the negative subsets of sites were most significant (labeled with blue in Figure 2) to train the model. The other subset of experiments (labeled with red) was used for testing the performance of the model.

We trained two generalized linear models (GLMs) with the logit link function (logistic regression) to classify the training data using the Statsmodels python library [148]. The first model included the branch length score as a feature, the second model did not. The fitted parameters for both models can be found in Supplementary Table S3.

When predicting miRNA-binding sites transcriptome-wide we expect that the non-functional miRNA-complementary sites vastly outnumber those that are functional in gene repression, in contrast to our model training set-up, where we used an equal number of positive and negative sites. Thus, the value of the feature-dependent score at which a site has a 0.5 probability to be functional will likely be higher than the value inferred based on the training set. Formally, this would be equivalent to shifting the scores that we obtain from the linear predictor by a constant value  $\Delta S$ , such that the probabil-

ity of a site being bound changes from  $p = \frac{e^S}{e^S+1}$  to  $p' = \frac{e^{S+\Delta S}}{e^{S+\Delta S}+1}$ . This leads to the transformation  $p' = \frac{Kp}{Kp+1-p}$  where  $K = e^{\Delta S}$ . To determine an appropriate value for the constant  $K$ , we computed an overall measure of down-regulation of the predicted miRNA targets upon transfection. That is, for a given  $K$ , we computed the score  $T$  of an individual target  $M$  for a given miRNA as the expected number of bound sites in this target  $T(M) = \sum_{s \in \text{sites}(M)} p'(s)$ , sorted the predicted targets of the miRNA from highest-to-lowest scoring, calculated the sum of fold-changes of top  $n$  targets for all values of  $n$ , and finally averaged these values over all miRNAs in the training set. To allow for the possibility that a minimum binding probability  $\tau$  needs to be reached for a site to have a functional impact, we carried out the above calculations also allowing for different probability thresholds (Supplementary [Figure S3](#)). The optimized parameter values are  $K = 0.24$  and  $\tau = 0.12$ . The resulting model was used to predict miRNA target sites and siRNA off-target sites across the entire set of 3' UTRs.

### 2.2.9 Evaluation of model performance

#### 2.2.9.1 Median fold changes

We compared the performance of various miRNA/siRNA target prediction methods as follows. For each miRNA, and for each method, we sorted all predicted target genes by their score, from highest to lowest. We determined the fold-change for each gene in each experiment and, when more than one experiment was available for a particular miRNA/siRNA, we computed the average fold-change in these experiments. Genes for which no expression estimates were available were filtered out. We then evaluated the median log fold-change of the targets predicted by a method  $lm(n)$  as a function of the number  $n$  of top predicted targets. Lower median log fold-changes indicate a stronger down-regulation of the targets predicted by a given method upon miRNA/siRNA transfection. Finally, we calculated average median log fold-changes  $\langle lm(n) \rangle$  for all the miRNAs/siRNAs under consideration by averaging the functions  $lm(n)$  over the considered miRNA/siRNA.

#### 2.2.9.2 Estimating the number of functional targets

The number of functional targets predicted by each method for each miRNA was estimated as follows. For each miRNA transfection data set, we calculated the fraction  $ftot$  of downregulated transcripts among all transcripts. This value is usually around 0.5. Then, considering the top  $n$  targets predicted by a given method for the transfected miRNA, we determined the fraction  $f(n)$  of these predicted targets that are downregulated upon transfection. An  $f(n)$  significantly larger than

$f_{\text{tot}}$ , indicates the presence of ‘true’ targets among the  $n$  predicted targets, as all of the true targets are expected to be downregulated. The total fraction  $f(n)$  can be written as  $f(n) = \rho(n) + f_{\text{tot}}(1 - \rho(n))$ , where  $\rho(n)$  is the fraction of  $n$  predicted targets that are true targets. From this we can estimate the number of true, functional targets among to top  $n$  predicted by the method as  $n_{\text{func}}(n) = n \times \rho(n) = n \frac{f(n) - f_{\text{tot}}}{1 - f_{\text{tot}}}$ . To summarize the data from all transfection experiments, we then determined the average number of functional targets over all considered experiments  $\langle n_{\text{func}}(n) \rangle$ . A similar approach was used previously in Khorshid et al. et al. [59].

### 2.2.10 Analysis of the siRNA screen

#### 2.2.10.1 siRNA-specific targeting score per gene

The score of a given siRNA for a given target gene was calculated as the sum of the scores of all unique target sites identified in the 3’ UTRs associated with the gene.

#### 2.2.10.2 KEGG pathway analysis

For the 100 siRNAs with the strongest effect in the screen [132], we obtained seed-MIRZA-G (see Table 2) off-target predictions. Then, for each gene that was predicted to be targeted by at least one of the 100 siRNAs, we calculated the average prediction score over all of these 100 siRNAs. Additionally, we determined the number of siRNAs (from the 100 with the highest score in the screen) that were predicted to target each individual gene. We sorted genes based on the number of targeting siRNAs and extracted the top 1000 for further analysis. We performed the same analysis considering all siRNAs in the libraries, not only the 100 that were found active in the screen. KEGG pathways analysis was performed using DAVID [149, 150]. As background we used the human genes whose 3’ UTRs we used for target site prediction.

## 2.3 RESULTS

### 2.3.1 Features of miRNA binding sites that are active in mRNA degradation

In line with previous studies [74, 76, 131], we sought to combine in our model a small number of sequence and structure features that are known to affect the efficacy of miRNA binding sites in mRNA degradation. These features were as follows:

- MIRZA quality score of the target site – reflects the free energy of binding between a miRNA and a target site and has been

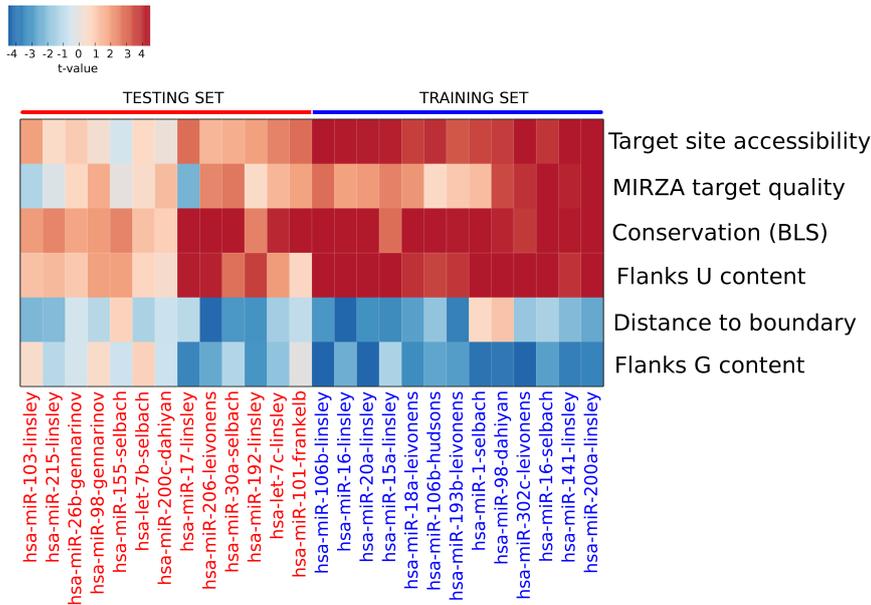
shown to enable identification of non-canonical binding sites that are effective in mRNA degradation [59].

- Accessibility of the target site – defined as the probability that the target site (defined as 7 nucleotide seed match plus 14 nucleotides upstream) is in single-stranded conformation within the mRNA [76, 151].
- Nucleotide composition of regions flanking the miRNA binding site – effective miRNA binding sites have been shown to reside in G-poor and U-rich sequence environments [76].
- Evolutionary conservation – this feature has been repeatedly shown to be highly informative for functional miRNA binding sites [73, 124], capturing probably a variety of distinct factors that have not been characterized yet.
- Distance to the boundary – functional miRNA binding sites tend to be located at the beginning and at the end of 3' UTRs [73, 74, 133] and this seems to be the case for siRNA target sites as well (data not shown).

The computation of these features is described in the Methods. To demonstrate that these features are informative for the prediction of functional miRNA target sites we used a set of 26 experimental data sets consisting of mRNA expression measurements before and after the transfection of individual miRNAs, that were obtained by seven different laboratories. From each experiment, we determined the 100 most downregulated (positive, effective sites) and the 100 least-changing (negative, ineffective sites) transcripts that had in the 3' UTR a single canonical match to the transfected miRNA. We then computed the features of the corresponding sites as described in the Methods section, and we evaluated the significance of the difference between the means of each feature's values in the positive and negative sets with the t-test. The results, shown in Figure 2, indicate that the features that we selected indeed distinguish the positive from the negative sites consistently, across the entire set of experiments.

In particular, the feature with the most consistent predictive power is the branch length score, that reflects the evolutionary conservation of miRNA–target interaction. We used this measure of selection pressure rather than the EIMMo score that we developed previously developed [73] because although the two measures have comparable predictive power (not shown), the branch length score can be more readily be computed for non-canonical sites compared to the EIMMo score, that was designed specifically for miRNA seed matches.

Also consistent with previous results [76], the sequence composition of the flanking regions is highly predictive for their responding in miRNA transfection experiments, to an extent comparable with the branch length score. Among the features that describe structural



**Figure 2:** Value of t-statistic in comparing the mean values of features used in the model (rows) among functional and non-functional miRNA seed-complementary sites across 26 experiments (columns). The data from the experiments labeled in blue were used to train the model and those from experiments labeled in red were used in testing the model.

accessibility (accessibility of the seed-complementary region, target site, extended target site), the accessibility of the target site (probability that a 21 nucleotides long target site anchored on the right-hand side by the match to the miRNA seed region is in single stranded conformation) has the most consistent performance across data sets (not shown). The accessibility of an RNA fragment for interaction with cognate factors can be defined in various ways. For example, the RNAup program from the Vienna package [152] calculates the energy that is necessary to generate a single-stranded conformation for the RNA sequence of interest, whereas the CONTRAfold program [143] computes the probability that the RNA sequence is in single-stranded conformation in the ensemble of all possible structures that it can assume. Because the CONTRAfold-based model appears to have slightly better performance than the RNAup-based model in predicting transcript down-regulation (Supplementary Figure S2), we used the CONTRAfold-based accessibility in our generalized linear model.

As shown in Figure 2, the experimental data sets appear to separate into two clusters that differ in the t-values of the differences between the feature values of positive and negative sites. To train our model we decided to use the set of experiments that gave the most significant t-values in the t-tests comparing feature values among the positive

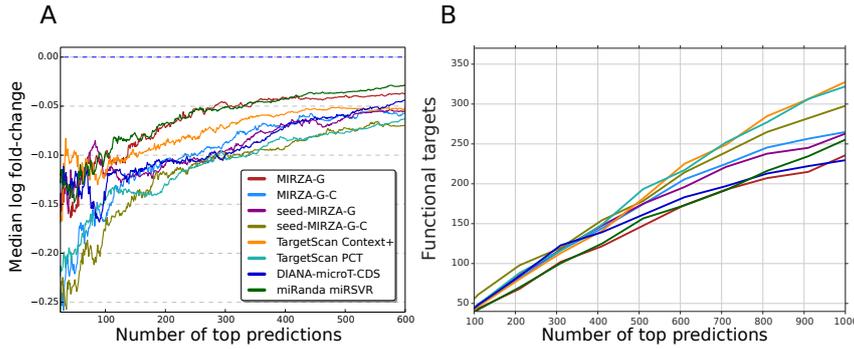
and negative sites (labeled in blue in [Figure 2](#)). The remaining set of experiments (labeled in red in [Figure 2](#)) were used for testing.

### 2.3.2 *Performance of the model in predicting the response of mRNAs to miRNA transfection*

We used the features defined above and the ‘training set’ of miRNA transfection experiments to construct a generalized linear model to predict positive sites – that confer downregulation to the host mRNA upon transfection of the cognate miRNA – and negative sites – that do not confer increased decay rate to the host mRNA – as described in the section ‘Training of the Generalized Linear Model’.

We used the ‘test set’ of miRNA transfection experiments ([Figure 2](#)) and a procedure that we described before [[59](#)] to evaluate the performance of our model. Briefly, we sorted the putative targets of a miRNA in the order of the scores assigned to them by a given prediction method and then we traversed the list of targets from top to bottom, computing, at each target rank  $x$ , the median fold change of all top  $x$  targets in response to miRNA transfection. Although many miRNA target prediction methods have been proposed, the benchmarking studies that are available [[59](#), [153](#)] consistently identify a few methods that yield consistently good results. We included these methods here and further refer the reader to the above-mentioned benchmarking studies for additional comparisons. One of the most widely used miRNA target prediction methods is TargetScan which consistently shows close-to-best performance [[59](#), [153](#)]. We therefore used TargetScan as the base-line for our assessment of algorithms’ performance. TargetScan has two variants, one that relies on the evolutionary conservation of the putative target sites (TargetScan PCT) [[33](#)] and one that uses information about the context in which the target site resides (TargetScan Context+) [[74](#), [75](#)]. We used both of these variants in the initial testing of our model’s performance. We further included DIANA-microT [[130](#)], which has also been reported to have high accuracy [[153](#)] and miRanda-mirSVR [[131](#)], which has been proposed for the prediction of both canonical and non-canonical sites.

We constructed and compared the accuracy of two types of MIRZA-based models: one that uses the branch length scores of sites in training and prediction and one that does not. Furthermore, we considered predicting only canonical targets or targets that possibly contained non-canonical sites. In the first case, we scanned the 3’ UTRs for canonical miRNA seed matches, while in the latter case we scanned the 3’ UTRs for 50 nts-long putative binding regions whose target quality score for a given miRNA was at least 50 (as described in the Materials and Methods). These models are summarized in [Table 2](#), and the performance evaluations in [Figure 3A](#).



**Figure 3:** Comparative evaluation of various models. (A) Models' performance in predicting mRNA down-regulation following miRNA transfection. The expectation is that a model performs well when its top predicted targets undergo the strongest downregulation after miRNA transfection. (B) Estimated number of functional targets for different methods as the function of the number of top predictions. Variants of the MIRZA-G model are described in Table 2. The other tested models are TargetScan Context+, TargetScan PCT, DIANA-microT-CDS and miRanda-mirSVR (the most conservative predictions). See text for additional details on these methods.

We found that models that take into account evolutionary conservation perform distinctly better than those that do not (Figure 3A). When considering evolutionary conservation, the targets predicted by the model that only considers canonical sites (seed-MIRZA-G-C) undergo the strongest down-regulation in response to miRNA transfection, followed by targets predicted by DIANA-microT, TargetScan PCT, our model that also considers non-canonical sites (MIRZA-G-C) and finally those predicted by miRanda-mirSVR. Among models that do not consider evolutionary conservation, our model that only takes into account canonical sites (seed-MIRZA-G) has by far the best performance followed by our model that includes non-canonical sites (MIRZA-G), and TargetScan Context+. The top targets of MIRZA-G respond stronger to miRNA transfection compared to those of TargetScan Context+, but for targets with mid-range scores, the relative magnitude of the response is reversed. The results are comparable when we assess the performance of the models in predicting protein-level changes (measured in [65]) in response to miRNA perturbations (Supplementary Figure S4).

For each method, we also estimated the number of functional targets, comparing the proportion of predicted targets that are down-regulated with the proportion of all genes that are downregulated in the transfection experiment. The relative performance of the methods, shown in Figure 3B, shows a pattern similar to that shown in Figure 3A.

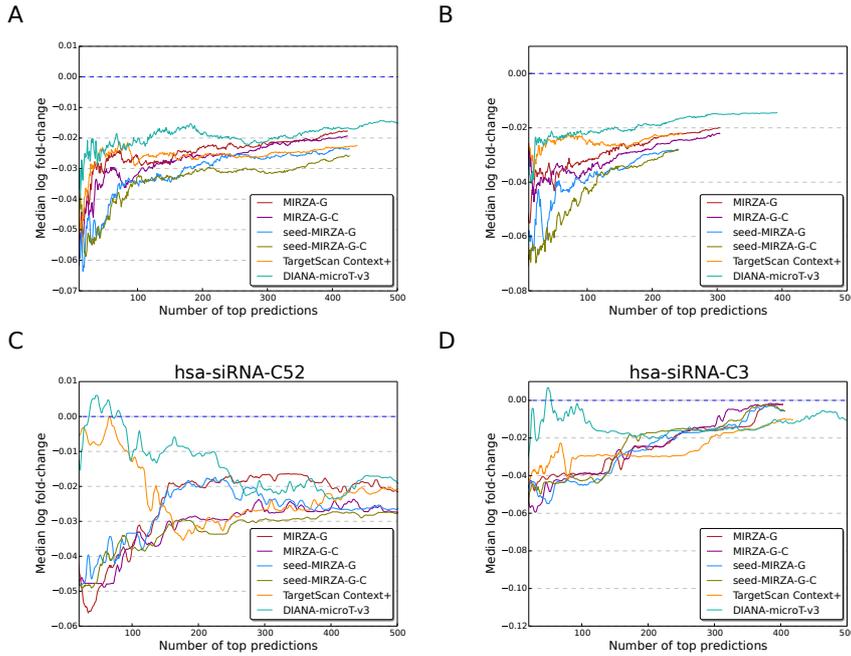
**Table 2:** Four alternative MIRZA-G models.

MODEL NAME	FEATURES	TARGET SITE TYPE
seed-MIRZA-G	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary	canonical
seed-MIRZA-G-C	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary evolutionary conservation	canonical
MIRZA-G	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary	canonical and non-canonical
MIRZA-G-C	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary evolutionary conservation	canonical and non-canonical

### 2.3.3 Prediction of siRNA off-target effects

Small interfering RNAs (siRNAs) have become a very important tool for studying gene function. Many studies have employed siRNAs or short hairpin RNAs (shRNAs) to screen for genes that are relevant to specific phenotypes [154–156]. It is not trivial to interpret the outcomes of these screens, due to a large extent to the so-called ‘off-target’ effects that the siRNAs have because they act through the miRNA effector pathway. SiRNAs being exogenous molecules, the feature that is most informative in the prediction of functional miRNA target sites, namely their strong evolutionary conservation, is unlikely to be informative. Thus, accurate prediction of siRNA off-target effects has remained challenging. As a the main aim of our study was to improve the prediction of siRNA off-target effects, we next tested our models on siRNA transfection data sets.

The first siRNA transfection data set that we used covered 12 distinct siRNAs [22] and previously used in the development of the Sylamer tool for the detection (though not prediction) of siRNA off-target effects [127]. Figure 4A shows that our models clearly outperform TargetScan Context+ and DIANA-microT in the prediction of off-target effects of these siRNAs, whether we consider only canonical or both canonical and non-canonical sites. Interestingly, when



**Figure 4:** Relationship between the prediction scores obtained with different target prediction methods and the extent of down-regulation of target mRNAs upon siRNA transfections. (A) Average over the siRNAs in the data set of Birmingham et al. [22]. (B) Average over the siRNAs from Jackson et al. [23]. (C) Data from an individual siRNA identified by Dongen, Abreu-Goodger, and Enright [127] to have prominent off-target effects. (D) Data from an individual siRNA identified by Dongen, Abreu-Goodger, and Enright et al. [127] to have modest off-target effects. See also Table 2 and the text for details on the methods.

we take into account the evolutionary conservation of the siRNA-complementary sites, we observe a somewhat stronger downregulation of the predicted mRNA targets, consistent with prior observations [132]. This does not appear to be the result of siRNAs acting on the target sites of miRNAs with the same seed sequence, because we obtain similar results when we use only siRNAs that do not share six or more contiguous seed nucleotides with any of the known miRNAs (Supplementary Figure S5A and B). The results obtained for each individual siRNA in this set are given in Supplementary Figure S6A–I. In Figure 4C and D we show two examples, one corresponding to an siRNA that was inferred [127] to have strong off-target effects (siRNA-C52), and the other to an siRNA with small off-target signature (siRNA-C3). In contrast to the targets predicted by TargetScan Context+ and DIANA-microT, the top targets that are predicted by our models consistently show stronger down-regulation compared to targets with lower prediction scores.

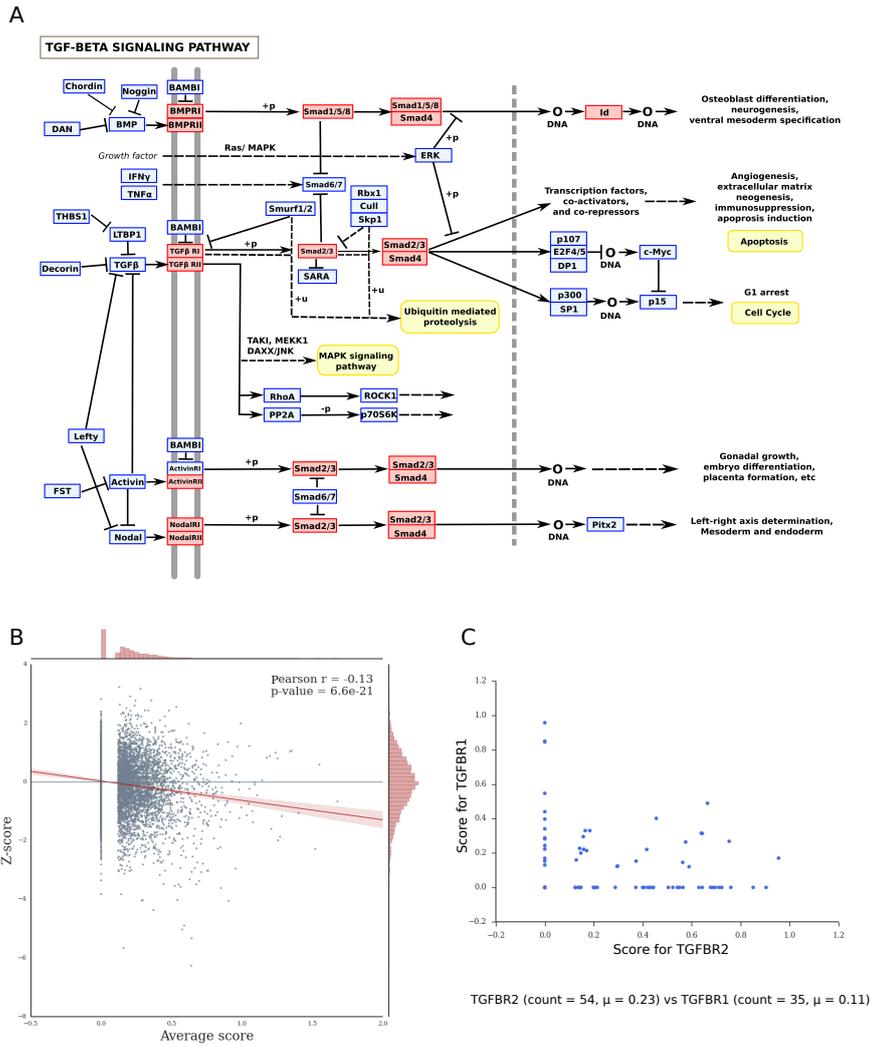
We further analyzed the data set obtained in one of the first studies that showed that the siRNA off-target effects are mediated by

the siRNA seed, similarly to miRNAs [23]. This study measured the transcriptome-wide response induced by mutants of an siRNA that was designed to target the MAP kinase. As shown in Supplementary Figure S7A–G and summarized in Figure 4B, in 6 of the 7 siRNA transfections the highest-scoring predictions of our models show a stronger down-regulation compared to TargetScan Context+ or DIANA-microT-predicted targets.

#### 2.3.4 Analysis of siRNA screening results with MIRZA-G

SiRNAs have been used in many high-throughput screens to identify key regulators or components of various biological processes. Most of these studies do not specifically investigate the off-target effects. However, a recent study found that of the ~20 000 siRNAs that were designed, in an ‘unbiased’ manner, to target the coding sequence (CDS) of 6000 distinct genes (phosphatases, kinases, signal transducers and cell-surface receptors) previously implicated in cancer, a large proportion had off-target effects on the TGF- $\beta$  pathway [132]. We sought to determine whether the results of the screen could be interpreted in light of MIRZA-G’s prediction of off-target effects.

From the screening results we identified the 100 siRNAs with the strongest phenotypic readout of TGF- $\beta$  pathway inhibition, which was the translocation of a GFP-SMAD2 reporter to the nucleus. For each gene in our 3’ UTR set we calculated the average MIRZA-G targeting score over all of these siRNAs as described in Methods. We repeated this procedure using predictions from all MIRZA-G variants, as well as from TargetScan and DIANA-microT. We found that TGFBR2 is the gene with the highest seed-MIRZA-G-C and MIRZA-G-C average score for the siRNAs that were most active in the screen (Supplementary Tables S4 and S5), consistent with previous results [132]. It is also a top target (3rd and 2nd, respectively) in the MIRZA-G and seed-MIRZA-G predictions. In contrast, the rank of TGFBR2 based on the TargetScan and DIANA-microT predictions is 13 and 43, respectively (Supplementary Table S5). We further used the 1000 genes with the highest average score to determine whether specific KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) [157] are targeted by the active siRNAs. In this test again, the TGF- $\beta$  pathway is most enriched among the prediction of the MIRZA-G variants compared to the other methods (Supplementary Table S6). These are in fact the pathways that should be targeted through on-target effects, guided by the perfect complementarity between the siRNAs and the coding regions of the mRNAs. Interestingly, these pathways are also predicted to be targeted through off-target effects, the reason being that all of these pathways contain TGF- $\beta$ . These results are consistent with the phenotypic readout of the screen as well as with our predictions (Supplementary Table S7). Figure 5A shows a sketch of the



**Figure 5:** SiRNA off-targets in the TGF- $\beta$  pathway. (A) Schema of the TGF- $\beta$  pathway drawn based on the figure provided by the DAVID server (42,43). Genes predicted to be off-targets of the top 100 siRNAs with the strongest effect in the screen are marked with red boxes. (B) Correlation between the z-score of an siRNA in the screen (y-axis) and the score that our model assigns to the interaction of the siRNA with TGFBR2 (x-axis). (C) Scatter plot of the predicted activities of the top 100 most active siRNAs on TGFBR1 and TGFBR2.

TGF- $\beta$  pathway with the genes predicted to be targeted by the active siRNAs labeled with a red.

We further found a significant anti-correlation between the z-score, that quantifies the magnitude of the cellular response to an siRNA in the screen, and the score that our model gives to the interaction of the siRNA with TGFBR2 (Figure 5B). This anti-correlation is weaker to absent when we include more genes of the TGF- $\beta$  pathway (TGFBR1, SMAD2 and SMAD4, Supplementary Figure S8) to compute an aver-

age score of interaction of the siRNA with TGF- $\beta$  pathway components.

These results suggest that the gene that most responsible for the observed phenotype is TGFBR2. Although TGF- $\beta$  has two main receptors, TGFBR1 and TGFBR2, it has been remarked that these two receptors do not appear to be similarly targeted [132]. Indeed, we found that more of the top 100 most active siRNAs are predicted to target TGFBR2 (Figure 5C) and with higher MIRZA-G off-target scores compared to TGFBR1.

In the above analysis, we started from siRNAs that were identified in the screen to be effective in modulating the response to TGF- $\beta$ . However, a question of high relevance in an experimental setting is whether relevant off-targets could be predicted a priori. To address this question, we computed, for each human gene, an average seed-MIRZA-G targeting score across all the siRNAs of this library (Supplementary Table S8). We then determined the enrichment of KEGG pathways among the top 1000 genes with the highest average score (Supplementary Table S9). Taking all human genes as the background set, the TGF- $\beta$  pathway shows the 12th most significant enrichment. Other pathways that are even more enriched than TGF- $\beta$  and would thus be expected to confound screening studies are the MAPK, neurotrophin, insulin, mTOR and ErbB pathways. Relevant for siRNA screening could be that the siRNAs in this library are also predicted to affect endocytosis.

## 2.4 DISCUSSION

Computational prediction of miRNA targets has progressed at a fast pace after the discovery of miRNAs, aiming to facilitate functional characterization of the thousands of miRNA genes that emerged from next-generation sequencing-based studies. Many methods are now available [153]. However, a tendency to converge on a small number of determinants has been apparent, even for tools that have been in use for almost a decade. Although increasingly large numbers of non-canonical miRNA binding sites have been reported in the recent years, it is clear that many miRNA target sites are perfectly complementary to miRNA seed regions and that the degree of evolutionary conservation of the miRNA-seed complementary region is a strong predictor of target site functionality. In our study, we took advantage of a biophysical model<sup>13</sup> of miRNA–target interaction that is able to identify not only canonical but also non-canonical interactions that are effective in mRNA destabilization from CLIP data [59] to predict such sites genome-wide. On its own, the biophysical model can be used to identify the miRNAs that guided the interaction of the Argonaute protein with CLIP-identified sites. However, for an accurate

<sup>13</sup> <http://www.clipz.unibas.ch/index.php?r=tools/sub/mirza>

prediction of miRNA as well as siRNA binding sites at a genome-wide scale, features beyond the energy of the small RNA–target site interaction need to be taken into consideration. This was the motivation for developing MIRZA-G. We have shown that MIRZA-G improves to some extent the genome-wide prediction of miRNA targets and substantially the prediction of siRNA off-targets<sup>14</sup>. The pipeline was implemented with the ruffus framework<sup>15</sup> [158].

Our analysis indicates that the features that were previously found to characterize effective miRNA target sites, whether they are located in the 3' UTRs or coding regions [159], are also informative for predicting siRNA off-target sites, as has been argued before [132]. Overall, this is not unexpected because that siRNAs and miRNAs use the same effector pathway. What may be surprising is that taking evolutionary conservation into account improves the prediction of siRNA target sites (compare the results of seed-MIRZA-G-C with those of seed-MIRZA-G in Figure 3A and B). This is consistent with the results of a previous study which found that conserved siRNA seed matches are more likely to be effective than non-conserved seed matches [132]. Although a trivial explanation could be that some siRNAs share the seed sequence with endogenous miRNAs, excluding these siRNAs from the analysis does not completely eliminate the signal (Supplementary Figure S6A and B). A possible explanation is that the conservation of a 3' UTR region, indicative of its relevance for some biological process, is correlated with other properties, such as its structural accessibility and nucleotide composition, that support targeting by siRNAs or miRNAs. The same reasoning may explain why functional miRNA-complementary sites preferentially emerge at the beginning and end of long 3' UTRs [73].

Although much work has been invested in computational miRNA target prediction, there remains substantial room for improvement. This may come from improved estimates of the rates of interaction between miRNAs and targets, from the inclusion of context-dependent effects such as 3' UTR isoforms [160], modulation of miRNA–target interactions by RNA-binding proteins [161] and others. Computational modeling of the miRNA-induced effects in systems in which measurements of relevant rate constants and abundances of relevant molecular species are available, will provide further insights into this mode of regulation [35]. Predictions generated by models such as MIRZA-G can provide essential entry points into such studies. Specifically in the analysis of siRNA screens, an avenue that has not been explored yet, is to use siRNA off-target predictions in conjunction with the measured phenotypic effects to infer the contribution of individual genes to the measured phenotype. This approach has been suc-

<sup>14</sup> The software is accessible at [http://www.clipz.unibas.ch/index.php?r=tools/sub/mirza\\_g](http://www.clipz.unibas.ch/index.php?r=tools/sub/mirza_g)

<sup>15</sup> Recently the pipeline was rewritten using in-house workflow management system Jobber.

cessfully used in the identification of transcription factors and miRNAs that have an important contribution to the pattern of mRNA expression in individual cell types [162]. It would be interesting to apply this methodology to a large number of siRNA screens to further unravel the contributions of individual molecular pathways to phenotypes.

## 2.5 SUPPLEMENTARY DATA

Supplementary Tables are available at NAR Online<sup>16</sup> or can be requested from the author.

## 2.6 FUNDING

Marie Curie Initial Training Network, RNPnet (project no. 289007), from the European Commission [to R.G.]. Funding for open access charge: ITN-RNPnet.

## 2.7 ACKNOWLEDGMENTS

We are grateful to Jean Hausser for sharing the pipeline for calculating miRNA target site features and to Erik van Nimwegen for discussions on the miRNA target prediction model. We also thank members of the Zavolan group for their comments on the manuscript.

## 2.8 CONFLICT OF INTEREST STATEMENT

Conflict of interest statement. None declared.

---

<sup>16</sup> <http://nar.oxfordjournals.org/content/suppl/2015/01/27/gkv050.DC1/nar-02756-n-2014-File007.xlsx>

## QUANTIFYING THE STRENGTH OF MIRNA–TARGET INTERACTIONS

---

### 3.1 INTRODUCTION

MicroRNAs (miRNAs) have emerged as important regulators of gene expression across a wide range of species. They are endogenously encoded small RNAs that are incorporated in ribonucleoprotein complexes also containing an Argonaute (Ago) protein, which they guide to other RNA targets to modulate their expression [120]. Although comparative genomic analyses indicate that a miRNA has on average hundreds of targets [124], how these predicted targets respond to changes in miRNA concentration is not entirely clear. The best-documented outcome of miRNA–target interaction is target destabilization [163], which is typically modest, but can give rise to interesting behaviors of miRNA-containing regulatory networks. These include the ‘threshold–linear’ response of miRNA targets to their transcriptional induction [164, 165] and the ultrasensitivity of target expression to the miRNA concentration [166]. The steady-state level of a given mRNA reflects the balance between transcription and decay. If the mRNA decay rate were constant, not modulated by miRNAs, the mRNA level would be expected to increase linearly with the transcription rate. However, if transcriptional induction occurs in the presence of a cognate miRNA, the target is expected to respond in a ‘threshold–linear’ manner: when the transcription rate is low, the few mRNA molecules that are produced are bound by the cognate miRNA and degraded. Once the transcription rate is sufficiently high for the mRNAs to saturate the miRNA–Ago complexes, the mRNAs escape the miRNA-induced repression and accumulate at a rate proportional to their transcription rate. The location of the threshold depends on the abundance of miRNA–Ago complexes, while the steepness of the transition between the two regimes depends additionally on the affinity of miRNA–target interaction.

We can illustrate these concepts with a simple model that focuses on the interaction of a single miRNA target with the miRNA and on the effect of this interaction on the rate of target decay, ignoring the possible effect of miRNAs on translation, the possible competition between targets for miRNAs and vice versa, other secondary effects such as feedbacks on target transcription rates, etc. Although these aspects most likely are relevant in in vivo situations, they go beyond the scope of our present study. Let us consider a miRNA target that is transcribed at rate  $\alpha$  [ $\frac{\text{mol}}{\text{s}}$ ] and decays with rate  $\delta$  [ $\frac{1}{\text{s}}$ ]. The free miRNA

*The work presented  
in this chapter was  
originally published  
in Methods [2]*

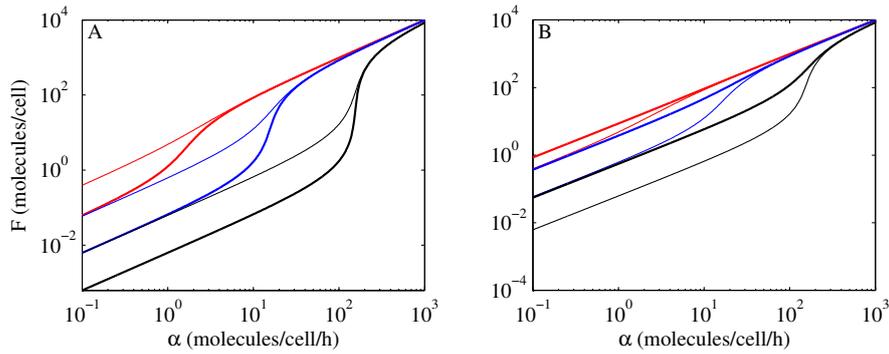
target  $F$  [mol] associates at rate  $\beta$  [ $\frac{1}{\text{mol}\cdot\text{s}}$ ] with miRNA-Ago complexes whose total concentration in a cell we assume to be constant,  $\Sigma$  [mol]. This leads to the formation of ternary target-miRNA-Ago complexes whose concentration we denote by  $A$  [mol], which can either dissociate into their components with rate  $\rho$  [ $\frac{1}{\text{s}}$ ], or fall apart due to the degradation of the miRNA target, which occurs at rate  $d\delta$  [ $\frac{1}{\text{s}}$ ]. The dynamics of these molecular species can then be described by the following equations:

$$\frac{dF}{dt} = \alpha - \delta F - \beta(\Sigma - A)F + \rho A \quad (1)$$

$$\frac{dA}{dt} = \beta(\Sigma - A)F - \rho A - d\delta A \quad (2)$$

Solving this system of differential equations we obtain the dependency between the concentration of the free (and total) target and its transcription rate, which has the threshold-linear form. [Figure 6](#) shows how the concentration of the free mRNA target responds to changes in target transcription rate, assuming values for the parameters  $\delta = 0.1 \frac{1}{\text{hour}}$  and  $d = 1.55$ , which we have recently estimated [\[35\]](#). To illustrate the expected behavior of high and low affinity targets we use two distinct values of the rate of ternary complex formation  $\beta$ , namely 0.24 and 2.4 cell/molecule/hour, and two distinct values of the rate of ternary complex dissociation  $\rho$ , namely 2.16 and 21.6  $\frac{1}{\text{hour}}$ . To further explore the behavior of targets of low, intermediate and high abundance miRNAs, we consider three total concentrations  $\Sigma$  of miRNA-Ago complexes, namely 10, 100 and 1000 molecules/cell. Our model thus assumes that the total concentration of miRNA-Ago complexes (free or bound to targets) is constant and does not respond to changes in miRNA target concentration. Although it remains unclear whether this assumption holds in vivo, data showing that the targets of endogenous miRNAs are up-regulated in response to transfection of exogenous siRNAs [\[167\]](#) suggest that at least the number of Argonaute molecules in a cell does not scale with the number of small RNAs that are present in cells. It can be observed that the transcription rate at which the target escapes miRNA regulation and accumulates rapidly depends on the total concentration of miRNA-Ago complexes, and that the transition is sharper for targets that have a higher rate of association with miRNA-Ago complexes. These behaviors have been observed in experiments with reporter constructs [\[165, 168\]](#).

So far we discussed the expected behavior of an individual miRNA target. However, because a miRNA probably has hundreds of targets, one of the strongly debated questions in the field is whether changes in expression of one of these targets affects the expression of the others by modulating their interaction with the common targeting miRNA. Computational studies have shown that the targets of a

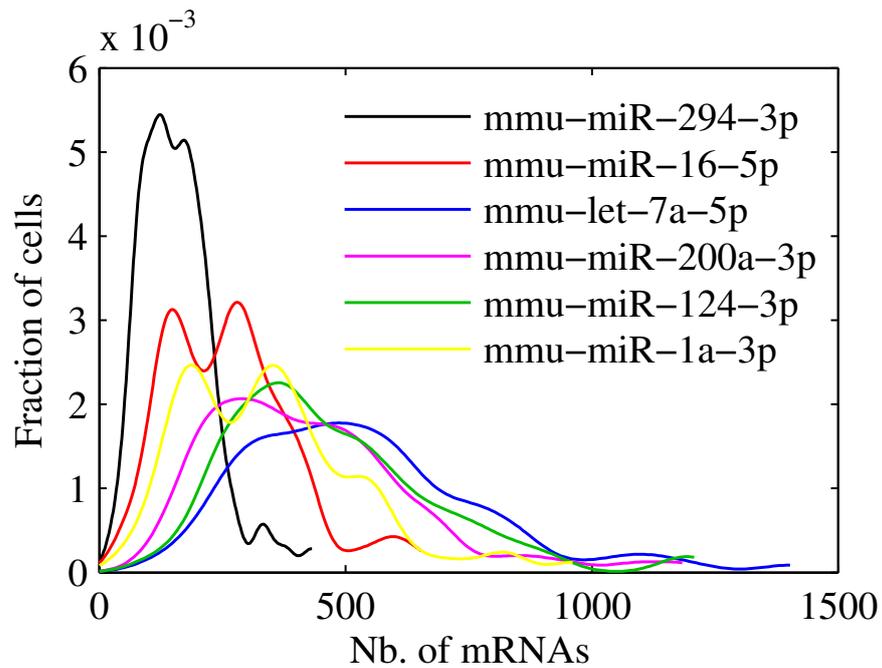


**Figure 6:** Accumulation of miRNA targets as a result of increasing transcription, in the presence of miRNAs, based on the steady state solution of Equation 1 and Equation 2. The three colors correspond to three total concentrations of miRNA–Ago complexes of 10 (red), 100 (blue) and 1000 (black) molecules/cell. (A) Thin lines correspond to low rates of target–miRNA–Ago association  $\beta = 0.24$  cell/molecules/hour, and thick lines to 10-fold higher association rates,  $\beta = 2.4$  cell/molecules/hour, with  $\rho = 2.16 \frac{1}{\text{hour}}$ . (B) Thin lines correspond to low rates of target–miRNA–Ago dissociation of  $\rho = 2.16 \frac{1}{\text{hour}}$ , and thick lines to 10-fold higher dissociation rates,  $21.6 \frac{1}{\text{hour}}$ , with  $\beta = 0.24$  cell/molecules/hour.

miRNA are expected to respond in an asymmetrical manner, changes in expression of high-affinity targets affecting the binding of the lower affinity targets but not the other way around [169, 170]. Whether these behaviors indeed occur in vivo is largely unknown. Rather, it has become clear that progress in understanding the impact of miRNAs on gene expression requires accurate measurements of miRNA abundance in single cells, estimates of the number of binding sites that a miRNA typically accesses within a cell, and estimates of the affinity of interaction between a miRNA and its multiple targets.

The abundance of individual miRNAs in mammalian cells varies over orders of magnitude (see for e.g. [171]). MiR-122, a highly expressed, hepatocyte-specific miRNA can reach 66,000 copies per cell in mouse liver cells and 135,000 in primary human hepatocytes [172]. The more typical range for well-expressed miRNAs is 1000–10,000 molecules per cell [171], which can probably be accommodated by the population of Ago proteins, whose abundance per cell has been estimated to be  $\sim 140,000$ – $170,000$  molecules (in a mouse epidermis and a human melanoma cell) [173].

The number of target sites that a miRNA can access within an individual cell remains hotly debated [168]. Recently developed methods have enabled quantification of mRNA species within single cells, although the mRNA capture rate appears to be low, around 10% [174]. A cursory analysis of the published mouse embryonic stem cell (ESC) single cell data shows that among the mRNAs that were captured, miRNA targets occur in a handful of copies such that the top 100 predicted targets of individual miRNAs yield a few hun-



**Figure 7:** Distribution of the number of targets of individual miRNAs that were captured from individual ESCs [174]. For each miRNA, the number of molecules of top 100 targets that were predicted with the seed-MIRZA-G-C miRNA target prediction program [1] were counted. The actual number of molecules was probably 10-fold higher (assuming that the capture rate of mRNA molecules in mRNA-seq is  $\sim 10\%$ ).

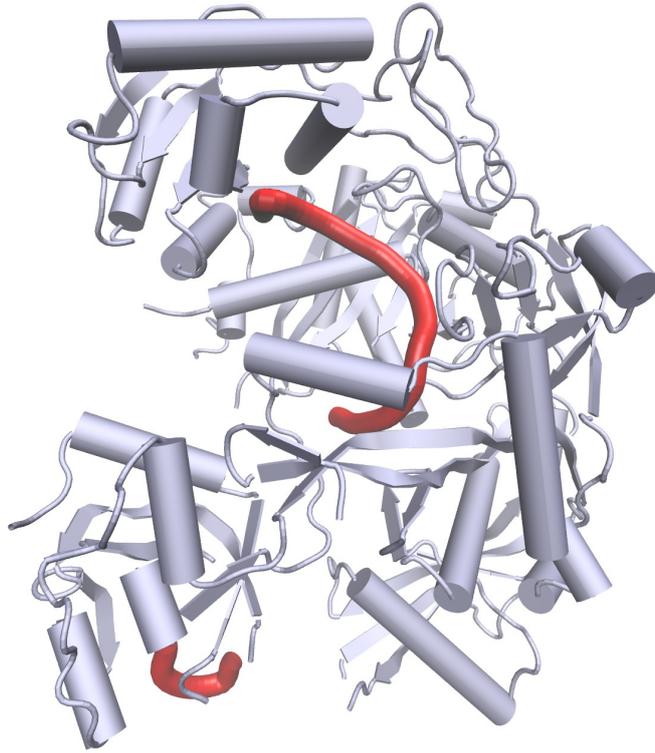
dred captured target molecules per cell [Figure 7](#). The targets of the mouse ESC-specific miR-294 are less abundant,  $\sim 1$  captured mRNA per cell, compared to targets of the ubiquitously expressed miR-16 and of some miRNAs that are expressed in differentiated tissues (e.g. the general differentiation marker let-7, the neuron-specific miR-124, the muscle-specific miR-1 and the epithelia-specific miR-200a), which were captured in 2–5 copies, on average. Assuming a capture rate of 10%, a mouse ESC thus expresses on average 10–50 molecules per miRNA target. The argument can be made that our estimation ignores the fact that ESCs already contain miRNAs which have reduced the levels of their targets and that we have thus underestimated the number of miRNA targets. Indeed, to improve these estimates we would need to quantify mRNA abundance in ESCs devoid of miRNAs (Drosha/Dicer knock-out ESCs). However, many studies in which miRNAs have been transfected in cells in which they were not previously expressed found only modest changes (less than 2-fold) in target levels and thereby decay rates (see for e.g. [159]). If a miRNA does target over a hundred distinct mRNA species, binding to perhaps multiple sites within a mRNA, the number of putative binding sites of a miRNA in a single cell can reach  $10^3$ – $10^4$ . Precise estimates of the number of binding sites and the ratio of binding sites

to miRNA-Ago molecules are essential for understanding the behavior of the targets *in vivo*, in individual cells.

### 3.2 INFERRING THE STRENGTH OF MIRNA-TARGET INTERACTIONS FROM EXPERIMENTALLY-DETERMINED TARGET SITES; THEORY

An important breakthrough in the experimental identification of miRNA targets came with the development of methods based on the crosslinking and immunoprecipitation of Argonaute proteins (Ago-CLIP) [175, 176], which enabled the capture of *in vivo* miRNA targets in high-throughput. The basic principle is to crosslink proteins to RNAs *in vivo* with ultraviolet light, immunoprecipitate the protein of interest and associated RNAs with a specific antibody, and prepare the protein-bound RNA fragments for deep sequencing. The resulting reads can be used not only to identify the mRNAs that were bound by miRNA-guided Argonaute proteins, but also to learn more about how miRNAs interact with their targets. For example, to describe this interaction, in previous work we introduced a model (MIRZA) that includes besides parameters for A-U, G-C, and G-U base pairs, for symmetrical and asymmetrical loops, a set of parameters corresponding to miRNA position-dependent contributions to the interaction energy [59]. The latter could result from the interaction taking place within the context of the Argonaute protein (Figure 8). Parameter values were inferred within a probabilistic framework, by maximizing the likelihood of the CLIP data. They confirmed the known importance of the miRNA 5' end (also known as 'seed' [124]) in the interaction with the target. However, application of the model to the CLIP sites suggested that many are bound in a 'non-canonical' manner (i.e. without perfect complementarity to the miRNA seed) and that the proportion of non-canonical sites that were captured for a given miRNA with CLIP increased with the abundance of the miRNA [59]. Because MIRZA provides a quantitative measure of the strength of interaction of miRNAs with target sites, it can be used not only for genome-wide prediction of binding sites but also to study miRNA-dependent regulation in deeper quantitative detail. In a parallel development, a next step in the experimental identification of miRNA target sites has been taken with the simultaneous capture of interacting miRNAs and target sites as chimeric sequence reads [60, 69]. Initial analysis of these data suggested that miRNAs may differ in their mode of interaction with the targets.

Thus, important open questions for the quantitative modeling of miRNA-target interactions are: what approach yields the most predictive model; what structure does this model have; are miRNA-specific models necessary to explain the experimental data? In the following we describe the miRNA-target interaction models that we inferred

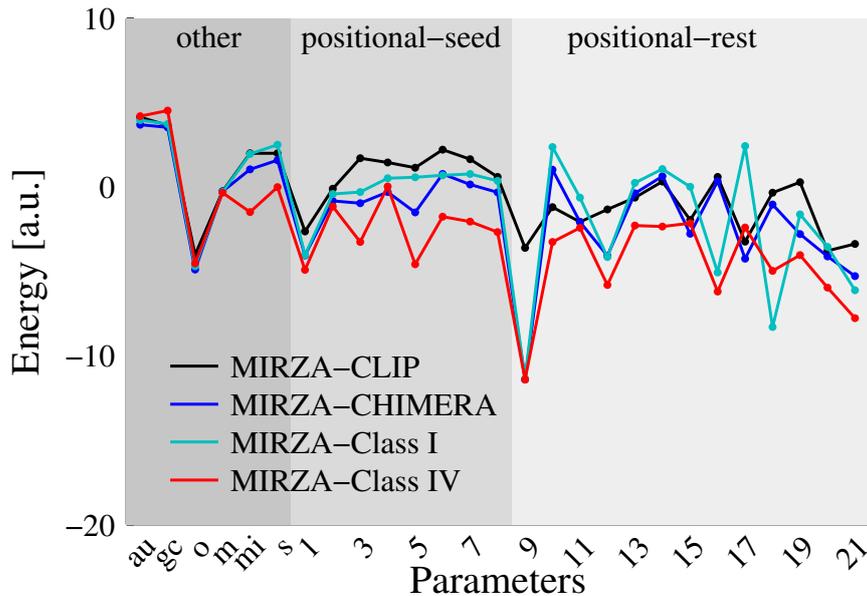


**Figure 8:** Crystal structure of the human AGO-2 protein (silver) in complex with miR-20a (red) [22]. The ‘seed’ nucleotides are visible in the structure because the conformational entropy of the miRNA 5’ end in the binding pocket of AGO-2 is limited. The residues 11–16 of the miRNA are not resolved due to their conformational freedom. The terminal 3’ end nucleotides, that contribute to the anchoring of the miRNA within AGO-2, are again visible.

with the MIRZA approach from various types of high-throughput data, and we evaluate their ability to identify functional miRNA targets, that are destabilized upon transfection of the cognate miRNA.

### 3.2.1 *Input data: Argonaute-bound RNA fragments. Output: general model of miRNA–target interaction MIRZA–CLIP*

A target site  $m$  of a miRNA  $\mu$  can be in one of two states, namely bound or unbound to the miRNA. Denoting the energies of the bound and unbound states by  $E_B$  and  $\bar{E}_B$ , the probability to find the site in bound state will be given by  $P_B = \frac{e^{E_B}}{e^{E_B} + e^{\bar{E}_B}}$ . The ‘bound’ state consists in fact of all ways in which the miRNA is hybridized with the target in the context of the Ago protein. Denoting by  $E(m, \mu, \sigma)$  the energy of the state in which site  $m$  is bound to miRNA  $\mu$  in configuration  $\sigma$ ,  $e^{E_B}$  is proportional to  $\sum_{\sigma} e^{E(m, \mu, \sigma)}$ . Similar to the standard



**Figure 9:** The 27 parameters of various MIRZA model variants. From left to right, base-pair parameters (A-U, G-C, G-U = 0), loop parameters (o: opening a loop, m: looped out mRNA nucleotide, mi: looped out miRNA nucleotide, s: symmetrical loop) and the 21 positional parameters are shown. The parameters of the MIRZA-CLIP model are shown in black, those of the MIRZA-CHIMERA model in blue, those of the MIRZA-Class I model in cyan and those of the MIRZA-Class IV model in red.

model of RNA-RNA interaction [177],  $E(m, \mu, \sigma)$  can be written in terms of a small number of parameters such as the energy of A-U, G-C and G-U base pairs, the energy for opening a loop in the miRNA-target hybrid, energies for extending a loop by a nucleotide in the miRNA, or in the mRNA, or by two unpaired nucleotides in the miRNA and target. In addition, specific to the MIRZA model of miRNA-target interaction [59] is a set of miRNA-position-specific energies (Figure 9). The logarithm of the ‘quality score’ of a site for a miRNA that MIRZA computes can be viewed as the energy of interaction between the miRNA and the target. An efficient dynamic programming algorithm for computing target quality scores has been proposed [59]. This enables one to infer the parameters of the MIRZA model by maximizing the likelihood of the Ago-CLIP data. Here we have repeated the analysis of the ~3000 Ago2-CLIP sites that were reproducibly isolated in multiple CLIP experiments [59, 178] to derived the baseline MIRZA-CLIP model shown in Figure 9.

3.2.2 *Input data: chimeric miRNA-mRNA sequence reads. Output: general model of miRNA-target interaction MIRZA-CHIMERA*

As mentioned in the Introduction, Helwak et al. [60] designed the Crosslinking and Sequencing of Hybrids approach (CLASH), in which the interacting RNAs are ligated prior to sequencing, thereby enabling the simultaneous capture of interacting miRNAs and target sites. These appear as “chimeric reads” each composed partly of a miRNA and partly of the miRNA target. Grosswendt et al. [69] subsequently reported that a substantial number of ligated miRNA-target site chimeras can be found even in samples prepared with a standard CLIP protocol. In contrast to Ago-CLIP, in these data sets there is no uncertainty about the miRNA that guided the interaction with each target site captured in the chimeras. Thus, in maximizing the likelihood of the data to infer a MIRZA-type model, one only needs to sum over all the ways in which the miRNA and target site in each chimera hybridizes with each other (and not over the miRNAs that could have interacted with the target site, as in the case of Ago-CLIP sites). We used the miRNA-target site pairs that were inferred by Grosswendt et al. from various PAR-CLIP and HITS-CLIP experiments (Table 3 and Supplementary Table 3 in Grosswendt et al. ) to construct a general model that could explain all these interactions. We called this model MIRZA-CHIMERA. Compared to the MIRZA-CLIP model that we inferred from Ago-CLIP data, MIRZA-CHIMERA seems to put less emphasis on the miRNA seed (Figure 9). The functional relevance of these differences will be discussed in the following sections.

3.2.3 *Input data: chimera of a specific miRNA with target sites. Output: miRNA-specific model of interaction with the target*

The CLASH study reported that some miRNAs, such as miR-92a and miR-181b, interact with their targets predominantly through their 3' rather than the 5' end, yielding ‘Class IV’ chimeras [60]. Other miRNAs such as those of the let-7 family were captured rather in ‘Class I’ chimeras, in which the miRNA presumably interacted through the ‘seed’. These observations suggest that the accuracy of miRNA target prediction could be improved through the use of miRNA-specific models of interaction. We decided to test this hypothesis here. However, because the available data sets [60, 69] contain a limited number of distinct target sites ligated to individual miRNAs, we inferred ‘Class’-specific rather than miRNA-specific models. Concretely, from the data of Grosswendt et al. [69] we selected a total 2589 chimeras of 24 miRNAs (those that yielded predominantly Class I chimeras in the data of Helwak et al. [60]) to train the “MIRZA-Class I” model and 949 chimeras of 8 miRNAs (those that yielded predominantly Class IV chimeras) to train the “MIRZA-Class IV” model. The cor-

**Table 3:** Chimeras of the indicated miRNAs, obtained from the data set of Grosswendt et al. [69] were used to infer MIRZA-Class I and MIRZA-Class IV models.

MIRZA-CLASS I	et-7a-5p, let-7e-5p, let-7f-5p, miR-10a-5p, miR-10b-5p, miR-125a-5p, miR-125b-5p, miR-1260b, miR-1301-3p, miR-15b-5p, miR-17-5p, miR-183-5p, miR-185-5p, miR-23a-3p, miR-27b-3p, miR-31-5p, miR-324-3p, miR-339-5p, miR-34a-5p, miR-423-5p, miR-455-3p, miR-484, miR-744-5p, miR-130b-3p
MIRZA-CLASS IV	miR-181b-5p, miR-221-3p, miR-30c-5p, miR-30d-5p, miR-320a, miR-361-5p, miR-92a-3p, miR-92b-3p

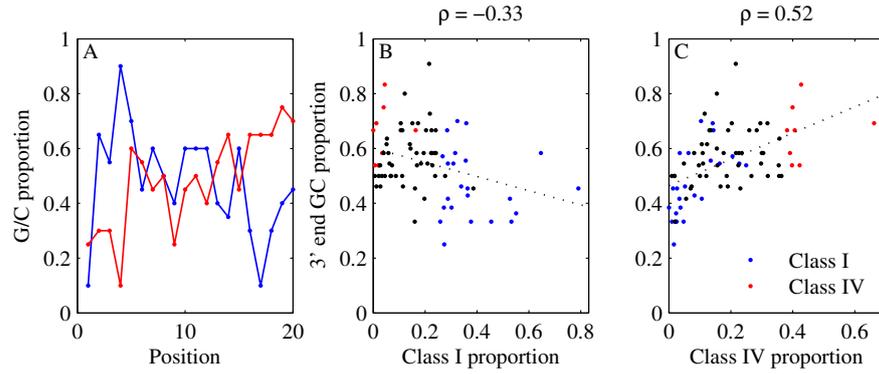
responding miRNAs are listed in Table 3. The parameters of these models, shown in Figure 9, indicate a positive contribution of the seed positional parameters in the MIRZA-Class I model, but not in the MIRZA-Class IV model. However, Figure 9 also shows a trend of positional parameters to progressively decrease from the seed to the 3' end in the MIRZA-Class IV model, but not in the MIRZA-Class I model. We test the functional relevance of these differences in a subsequent section.

It has been recently observed that the miRNAs that were reported to form Class IV hybrids have G/C-rich 3' ends [179]. We reproduced these observations here (Fig. 5). Furthermore, we found that the proportion of Class I hybrids that were captured for a miRNA decreases with the G/C content of the miRNA 3' end, while the proportion of Class IV hybrids shows the opposite trend (Fig. 5). A possible explanation behind the different propensities of different miRNAs to yield Class I or Class IV chimeras is that the G/C-content of the miRNA 3' end stabilizes the interaction with the target site, facilitates ligation and leads to an over-representation of this type of sites among the chimeric sequences. This possibility would need to be investigated in more detail before miRNA-specific modes of interaction are inferred from chimera data.

### 3.3 RESULTS

#### 3.3.1 Evaluating the models on biochemical data

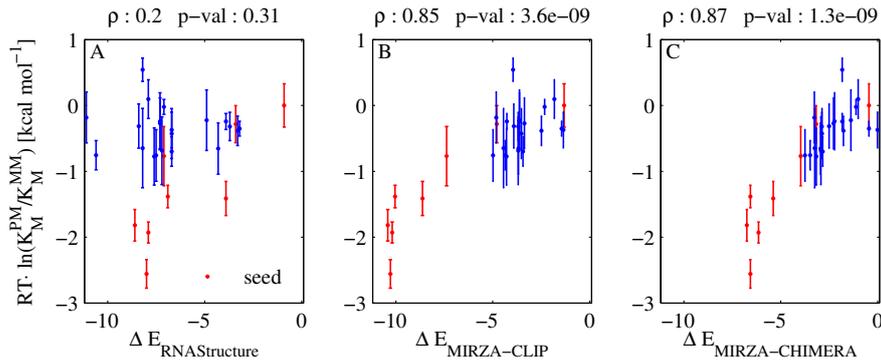
The 'quality score' assigned to a site by the MIRZA model takes into account all possible configurations in which the miRNA can hybridize to the target site within the ternary miRNA–target site–Ago complex, and provides an estimate of the binding energy between the miRNA and the target site. Thus, if the model is accurate, it should be able to predict the free energy of interaction determined with biochemical approaches. The dissociation constant  $K_D$ , which is the ratio of the rates of dissociation ( $k_{off}$ ) and association ( $k_{on}$ ) of molecules



**Figure 10:** Relationship between the nucleotide composition of the miRNA and the type of hybrids in which the miRNA was captured. The miRNAs used to infer the MIRZA-Class I model are shown in blue, the miRNAs used to infer the MIRZA-Class IV model are shown in red and other miRNAs are shown in black. Data for analysis taken from Helwak et al. [60]. (A) Proportion of G/C nucleotides at different positions along miRNAs that yield predominantly Class I and IV hybrids/chimeric reads in the data set of Helwak et al. [60]. (B) Correlation between the proportion of G/C nucleotides at the 3' end of a miRNA and the proportion of captured Class I chimeras. (C) Correlation between the proportion of G/C nucleotides at the 3' end of a miRNA and the proportion of detected Class IV chimeras.

in a complex,  $K_D = \frac{k_{off}}{k_{on}}$ , should be related to the Gibbs free energy of interaction through the relationship  $\Delta G = -k_B T \log\left(\frac{1}{K_D}\right)$ , where  $k_B$  is the Boltzmann constant and  $T$  is the temperature. Although only few measurements of miRNA–target dissociation constants are available, particularly for mammalian systems, Wee et al. [77] measured a related constant, namely the Michaelis–Menten constant. This is defined as  $K_M = \frac{k_{cat} + k_{off}}{k_{on}}$ , thus including besides the dissociation and association rates the rate with which the miRNA catalyzes the target cleavage. Wee et al. measured for  $K_M$ 's for perfectly complementary sequences (PM) and for sequences that have mismatches at different positions along the miRNA (MM) in the context of Argonaute 1 protein of *Drosophila melanogaster* [77] and then correlated  $\log \frac{K_M^{PM}}{K_M^{MM}}$  source with the difference in the free energy of interaction of the perfectly matched and mismatched hybrids given by the RNAstructure software [180]. Computing this correlation separately for duplexes in which mismatches were located at the 5' and 3' ends of the miRNA, respectively, Wee et al. concluded that the standard base pairing rules apply to miRNA–Ago2–target complexes [77]. We thus sought to use the measurements of Wee et al. [77] to further validate the MIRZA models that we inferred from CLIP data sets.

First, we compared the energy differences inferred from measurements of  $K_M$ 's with those predicted with the current version (5.7) of the RNAstructure software and with those predicted with MIRZA-type models. As described by Wee et al. [77], we found relatively



**Figure 11:** Ratio of binding free energies of mismatched and perfectly matched hybrids. The Spearman correlation was computed between the values estimated based on biochemical measurements (energy of interaction  $\sim \ln(\frac{1}{K_M})$ ) and values predicted with three distinct models: RNAStructure 5.7 (left), MIRZA-CLIP (center) and MIRZA-CHIMERA (right). Data points in red correspond to hybrids with mismatches in the miRNA seed region, those in blue to hybrids with mismatches in the 3' region.

good correlations between RNAStructure-based predictions and experimental measurements, if we consider separately hybrids with mismatches in the miRNA seed region (Spearman correlation coefficient  $\rho = 0.81$ , p-value = 0.015) and in the miRNA 3' end (Spearman correlation coefficient  $\rho = 0.4$ , p-value = 0.20). However, considering all the hybrids together, the correlation is rather poor (Spearman correlation coefficient  $\rho = 0.20$ ), presumably because the nearest neighbor model implemented in RNAStructure does not appropriately describe interactions that take place within RNA-protein complexes, where different nucleotides in the RNA can have disproportionate contributions to the energy of interaction.

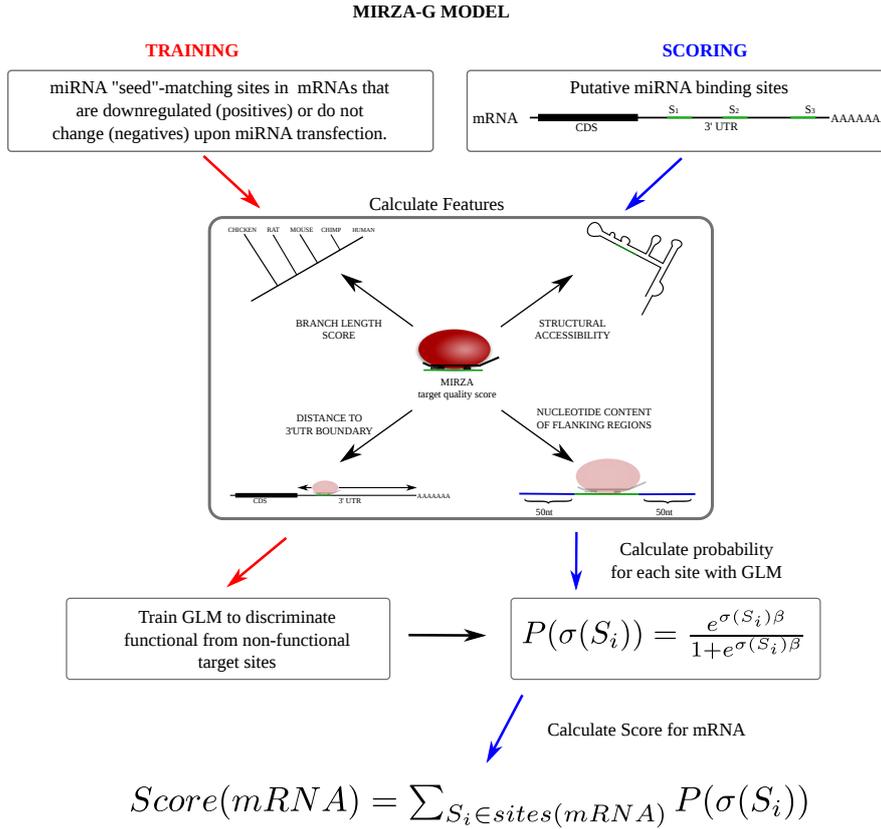
In contrast, evaluating all of the hybrids within the MIRZA-CLIP model yields predictions that are strongly correlated with the experimental results (Spearman correlation coefficient  $\rho = 0.85$ , p-value =  $3.6e-09$ , 95% confidence interval = [0.71, 0.93]). Interestingly, the MIRZA-CHIMERA model gives a slightly higher correlation with the experimental data (Spearman correlation coefficient  $\rho = 0.87$ , p-value =  $3e-09$ , 95% confidence interval = [0.73, 0.94]), although the difference is not significant. Thus, these two models, that were inferred from different types of sequenced miRNA target sites, predict remarkably well the energies of interaction between miRNAs and target sites that are inferred from biochemical measurements (Figure 11).

### 3.3.2 Genome-wide prediction of miRNA targets

One of the main applications of these models is in the genome-wide prediction of miRNA binding sites. However, the predicted energy of interaction between a miRNA and a target site is only one of the

factors that contributes to a functional interaction. Other features of the target site have also been shown to be important [76]. Thus, in recent work we sought to build on MIRZA and develop a model that is suitable for accurate prediction of miRNA binding sites genome-wide. The resulting MIRZA-G model combines the MIRZA target quality score with the accessibility of the target site, the G/U content of the region in which the site is embedded, the relative location of the site in the transcript and, optionally, with the degree of evolutionary conservation of the putative target site (Figure 12). MIRZA-G is trained by fitting a generalized linear model with a logit function to discriminate between miRNA-complementary sites located in mRNAs that do and mRNAs that do not respond to the transfection of the cognate miRNAs [1]. Furthermore, because high-throughput studies evaluate the effects of miRNAs at the level of transcripts and genes rather than individual sites, MIRZA-G computes transcript/-gene scores, summing up the probabilities that individual target sites have a functional impact. Using different MIRZA variants to compute target quality scores for the MIRZA-G model we can test the ability of these variants to predict which transcripts are most affected by the transfection of individual miRNAs. Thus, we employed the MIRZA-CLIP/CHIMERA/Class I/Class IV models individually within the MIRZA-G framework to predict and rank targets of individual miRNAs. Because different MIRZA variants yield different distributions of target quality scores and in the genome-wide prediction of target sites we only consider putative sites with a minimal 'target quality' score, we have used different thresholds for different models. The weight of different features of target sites within the MIRZA-G model were kept unchanged.

To determine a target quality score threshold for different MIRZA variants we noted that 'canonical' interactions that involve perfect pairing of the miRNA seed have the highest scores with all MIRZA variants. Thus, we employed the procedure that we used before for MIRZA-CLIP [1]. That is, with each MIRZA variant, we assigned to each of the 2998 CLIPed sites from Khorshid et al. [59] the most likely guiding miRNA. This was the miRNA with the highest target quality score for the site given under the considered MIRZA model. We then predicted the structure of the most likely hybrid between the target site and the guiding miRNA, and divided the sites into canonical – those with perfect base-pairing over nucleotides 2–8 of the miRNA or perfect pairing over nucleotides 2–7 followed by an adenine (opposite position 1 in the miRNA) – and non-canonical – all other sites. Based on the cumulative distribution of target quality scores for canonical and non-canonical sites, we set a threshold that allowed us to capture the majority of canonical sites without including too many non-canonical sites, that may be artifactually captured. For MIRZA-CLIP a threshold of 50 captures 91% of canonical

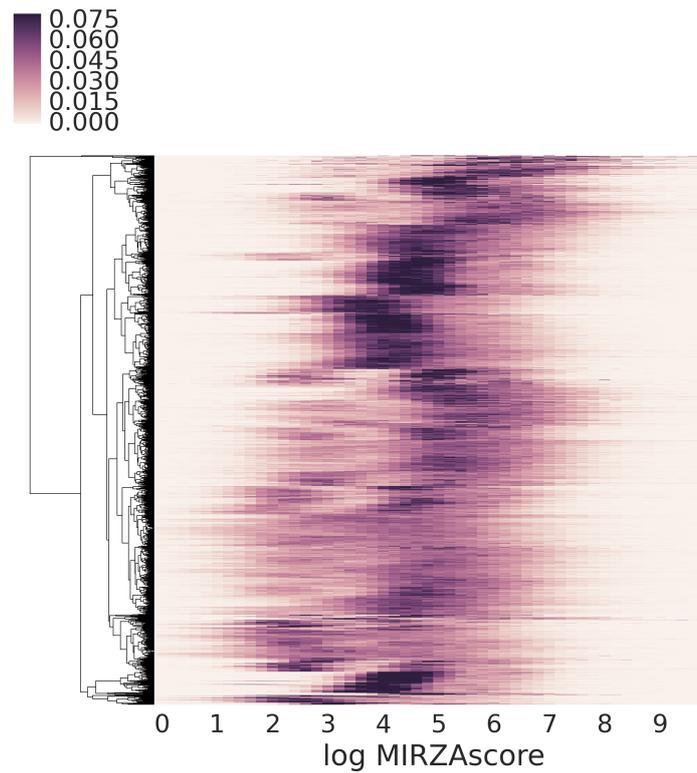


**Figure 12:** Diagram of the approach for predicting miRNA targets with MIRZA-G.

sites and 18% non-canonical sites, for MIRZA-CHIMERA a threshold of 20 captures 97% canonical and 20% of non-canonical sites, for MIRZA-Class I a threshold of 30 leads to the capture of 94% of the canonical and 18% of non-canonical sites, while for MIRZA-Class IV a threshold of 20 captures 94% of canonical target sites and 20% of the non-canonical target sites.

### 3.3.3 Wide range of MIRZA quality scores across the targets of a given miRNA

Although we do not focus on this aspect here, it has been proposed that differences in affinity between targets may underlie asymmetries in the crosstalk of mRNAs that bind the same miRNAs [169]. Thus, having shown that the target quality scores computed with MIRZA models correlate very well with the affinities of miRNA-target interactions measured with biochemical methods, we wondered how much variation there is in the affinity of different target sites for a miRNA. Therefore, we determined the MIRZA target quality score for all the sites of all miRNAs that were considered in the genome-wide predictions with MIRZA-G. These had a probability of being



**Figure 13:** Distribution of the MIRZA quality scores of target sites of individual miRNAs. Each line corresponds to one miRNA and the intensity of the color indicates the density of target sites within a particular range of target quality scores, computed with MIRZA-CLIP.

functional of at least 0.12 (see Gumienny and Zavolan [1] for details). For each miRNA we have divided the 0–10 range of MIRZA target quality scores into bins of 0.2 and have shown the distribution of the target sites of each miRNA as a heat-map, which each line corresponding to a miRNA and the intensity of the color indicating the density of target sites within a bin (Figure 13). It can be seen that the target sites of an individual miRNA span a range of  $\sim 4$  log units or they can differ by  $\sim 50$ -fold in the predicted affinity.

#### 3.3.4 Evaluation of the MIRZA models on miRNA transfection data

miRNAs have been reported to destabilize their mRNA targets, inhibit their translation [120], and even to increase transcript stability under specific circumstances [181]. Of these, perhaps the least controversial is mRNA destabilization, which has been argued to be the dominant mechanism behind the repressive effect of miRNA, with translational repression playing a small, perhaps more transient role

**Table 4:** Summary of the experimental data sets that were used to train the model and evaluate its performance.

REFERENCE	DATA SOURCE (GEO ACCESSION / URL)	MIRNAS IN THE DATA SET
Dahiya et al. [135]	GSE10150	miR-200c, miR-98
Frankel et al. [136]	GSE31397	miR-101
Gennarino et al. [137]	GSE12100	miR-26b, miR-98
Hudson et al. [134]	GSE34893	miR-106b
Leivonen et al. [138]	GSE14847	miR-206, miR-18a mir-193b, miR-302c
Linsley et al. [139]	GSE6838	miR-103, miR-215, miR-17, miR-192, let-7c, miR-106b, lmiR-16, miR-20, miR-15a, miR-141, miR-200a
Selbach et al. [65]	psilac.mdc-berlin.de/download/	miR-155, let-7b, miR-30a, miR-1, miR-16
Olive et al. [182]	GSE53225	miR-92a

[163]. The importance of this mechanism is further underscored by observations that miRNA-complementary sites that are conserved in evolution and sites that induce strongest downregulation of their host transcripts upon miRNA transfection have similar properties [76]. Furthermore, acting through the miRNA pathway, small interfering RNAs (siRNA) also destabilize many transcripts (the so-called “off-target” mRNAs) [23]. Thus, it is reasonable to expect that the extent of mRNA destabilization upon miRNA transfection is a robust measure of the strength of interaction between a miRNA and the mRNA. Consequently, the ranking assigned by a computational miRNA target prediction method to mRNAs should correlate well with their change in expression upon miRNA transfection. This is the assumption that we make in discussing the relative performance of various models for miRNA target prediction.

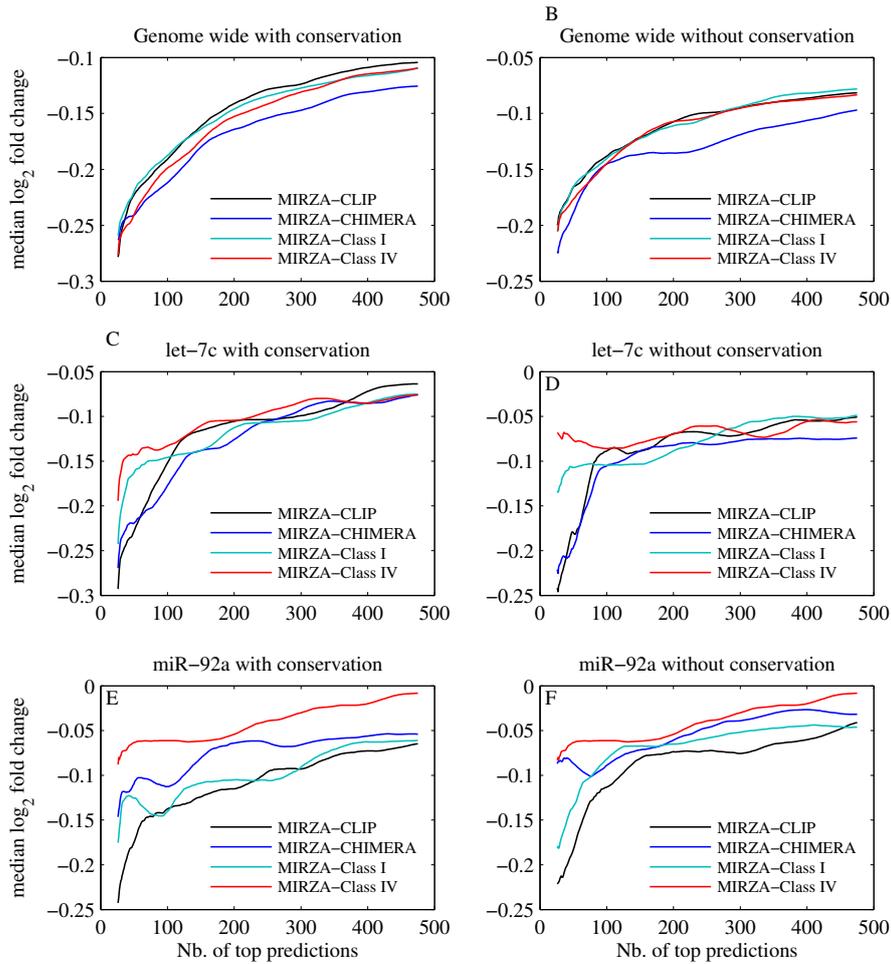
First, we tested whether the models can predict the mRNA expression changes that were induced by individual transfections of miRNAs. To this end, we used data corresponding to 26 miRNA transfections into human cells and one transfection into mouse cells (Table 4). The processing of the transfection data was described extensively in [1]. For each type of MIRZA model of miRNA–target interaction we used two variants of the genome-wide MIRZA-G prediction model

to predict sites. One of these considered the evolutionary conservation of the sites and the other did not [1] (see Figure 12). We sorted targets predicted by each of these models in the order of their prediction score. We then computed the median  $\log_2$  fold-change of the top  $N$  predicted transcripts as a function of the number  $N$  of top targets considered. The average profiles, computed over the 26 data sets, are shown in Figure 14A–B. We found that all four models perform as expected in predicting miRNA targets genome-wide. Consistent with its slightly better performance in predicting the in vitro-measured free energy of interaction between miRNAs and target sites, the targets predicted by the MIRZA-CHIMERA model are somewhat more downregulated compared to the targets predicted with MIRZA-CLIP, particularly when the evolutionary conservation of the sites is not taken into account.

Next we asked whether Class I and Class IV-specific models are more accurate in predicting targets of miRNAs that have been found to yield predominantly Class I and Class IV chimeras, respectively. As representatives of the first we chose the let-7 family of miRNAs and as a representative of the latter the miR-92a. Because we did not find transfection data for Class IV-chimera forming human miRNAs, we used a data set obtained from mouse cells transfected with the mouse miR-92a. The results, shown in Figure 14, panels C–D for let-7 and E–F for miR-92a, clearly indicate that the general MIRZA-CLIP and MIRZA-CHIMERA models are more accurate in predicting transcript downregulation upon miRNA transfection than Class I/IV-specific models. Together with the fact that the sites that are predicted with these models tend to be canonical sites, these results indicate that the origin and relevance of Class IV hybrids needs to be further investigated. As mentioned above, a possibility that needs to be ruled out is that the experimental procedure for isolating miRNA-target hybrids via chimeric sequences enriches for non-canonical hybrids that have increased stability prior to ligation.

### 3.3.5 *Inferring a MIRZA model from biochemical data*

The results presented above indicate that the MIRZA-CLIP/CHIMERA models explain well both the biochemical data as well as the response of mRNAs to miRNA transfection. However, given the complexity of CLIP experiments and the indirect nature of the resulting data, one wonders whether an even more accurate model of miRNA-target interaction could not be derived from in vitro measurements of interaction affinity as obtained in the study of Wee et al. [77]. To gain further insight into the design of an efficient experiment, we generated synthetic data sets of hybrids, computed their pseudo-energies of interaction with MIRZA-CLIP, and then asked how our ability to recover the model parameters from the synthetic data sets depends



**Figure 14:** Relationship between prediction score and the extent of mRNA downregulation. Genome-wide target predictions were carried out with the MIRZA-G generalized linear model [1], within which the target quality scores were calculated with different MIRZA variants: MIRZA-CLIP, MIRZA-CHIMERA, MIRZA-Class I and MIRZA-Class IV. Measurements of mRNA expression in control and miRNA-transfected cells were used to determine the  $\log_2$  fold-changes of predicted miRNA targets. (A) Median  $\log_2$  fold-change of the top  $N$  targets of the transfected miRNA, in function of  $N$ , were averaged over a data set of 26 miRNA transfection experiments. (C) Same procedure, but showing the median  $\log_2$  fold-change of predicted let-7 targets upon let-7 transfection (Table 4, data from [65, 139]). (E) Same procedure, but showing the median  $\log_2$  fold-change of predicted targets of the mouse miR-92a upon miR-92a transfection in mouse cells (Table 4, data from [182]). For (A), (C) and (E), genome-wide predictions were carried out including evolutionary conservation whereas for (B), (D) and (F), without [1].

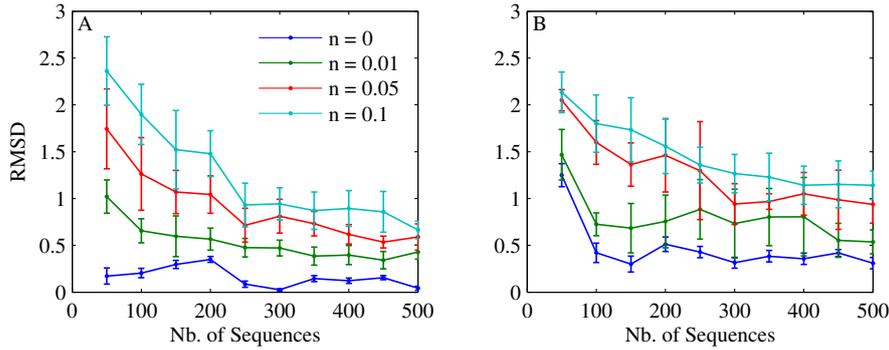
on the number and type of hybrids and the accuracy of the provided pseudo-energies.

First, we simulated the experimental design of Wee et al. [77], in which energies of interaction between close variants of a single miRNA

(let-7) and their perfectly complementary sequences were measured. There are 1890 possible two point-mutants of let-7, from which we sampled datasets of different sizes. An alternative design is to measure the energies of interaction between ‘random’ small RNAs and their partially complementary sequences. In this approach the small RNA is an entirely ‘random’ sequence whereas the interacting site is a sequence whose complementarity to the small RNA varies. To construct it, we first chose the average number of complementary nucleotides. With probabilities of complementarity chosen uniformly between 0.25 and 1, we can simulate from interactions of random RNA fragments to interactions of perfectly complementary sequences. This second approach is meant to provide datasets containing more information in terms of pairs of interacting nucleotides than the first approach. For both methods, while constructing subsets of various sizes, we aimed to cover uniformly the space of interaction energies and of nucleotide positions involved in the binding. Finally, we considered the possibilities that the measurements are not entirely accurate. To simulate this, we added gaussian noise to the computed interaction energy for each hybrid with a standard deviation of 0 (no noise), 1%, 5% and 10% of the predicted energy of interaction. For each data set size and each noise level we generated 100 synthetic data sets. To each synthetic data set we applied the simulated annealing procedure that was described in Khorshid et al. [59] to recover the parameters of the MIRZA model used to generate the pseudo-energies. The results, averaged over the 100 replicates of each setting, are shown in [Figure 15](#). They indicate that if the measurement noise is less than 10%, ~250 hybrids, chosen from across the entire range of expected affinities would be sufficient to recover the model parameters with reasonable accuracy (root mean square difference, RMSD, between recovered and input parameters  $< 1$ ). If the measurements were very precise (relative error of a few percent), the number of hybrids necessary to recover a model with  $\text{RMSD} < 1$  is considerable smaller, ~100, which is within reach with the technology available today. The experimental design of measuring closely related variants of a single miRNA does not yield equally accurate parameter values from a comparable number of hybrids, presumably due to the limited sampling of nucleotide/position combinations.

### 3.4 DISCUSSION AND PERSPECTIVE

That miRNAs are important for the proper development and function in a large number of species is undisputed. Similar to transcription regulation by transcription factors, miRNA-dependent regulation is ‘combinatorial’. That is, a regulator typically has many targets and a target is affected by many regulators. In contrast to transcription factors, miRNAs induce milder changes in target expression, which



**Figure 15:** Root mean square difference (RMSD) between the MIRZA parameters used to generate the training set and the MIRZA parameters inferred from the training data, as a function of the size of the training set. The colors correspond to the noise added to the training set data (0%, 1%, 5% and 10% of the predicted energy value). For (A), the data sets were generated with the ‘randomized’ procedure, whereas for (B), the data sets were generated through mutations of the let-7 miRNA.

makes it more difficult to distinguish bona fide regulatory effects from biological or experimental variability. Consequently, a number of distinct directions are pursued in the field. Many groups have started to explore functional consequences of miRNA–target interaction that go beyond the repression of a single miRNA target into dynamical aspects of the response of a larger network, containing multiple miRNAs and multiple targets [168–170, 183, 184]. Such a network is quite complex and can exhibit very rich behaviors. For example, a recent study emphasized that even an increased expression of some miRNA targets can be expected in response to the increased expression of a miRNA. This could happen if miRNAs with different efficiencies in target down-regulation compete for the same sites on the target, because over-expression of the miRNA that is less effective in repressing the target could lead to the displacement of the miRNA that is more effective and thus to a net increase in target expression [185]. Additional experiments are necessary to determine whether this behavior occurs *in vivo*.

More generally, given the wide range of behaviors that computational models can predict, it is important to sufficiently constrain them with accurate parameters. Indeed, as described in previous sections, recent studies have started to provide measurements of the concentrations and the rate of interactions between the relevant molecular players. Our work shares this aim. Up to this point we used high-throughput data sets of miRNA binding sites that were derived with various approaches to parameterize a model of miRNA–target interaction. This model allows us to compute the energy of interaction between miRNAs and arbitrary target sites and to carry out genome-wide predictions of miRNA targets. We have shown that the model inferred from sequenced Argonaute/miRNA binding sites predicts

quite accurately hybrid energies that are measured with biochemical methods *in vitro*. Furthermore, we have proposed a strategy for deriving a MIRZA-like model from biochemical measurements that can be obtained with the technology available today.

Although on its own, the energy of miRNA–target interaction is not sufficiently predictive of functional interactions, it is one of several informative features that together enable fairly accurate transcriptome-wide predictions. These additional features reflect the secondary structure of the target mRNA, its interactions with RNA-binding proteins, as well as other factors that are yet not understood but can be captured in the degree of evolutionary conservation of the putative miRNA binding site. Dynamical changes in the miRNA targetome between cell types or cell states will remain difficult to model computationally, but they may be important for the interpretability of experimental data. For example, it has been shown that taking into account tissue/condition-specific isoform expression can improve the prediction of miRNA targets [160], because alternative polyadenylation can change the susceptibility of transcripts to miRNA regulation. Conversely, miRNA stability is also subject to regulation, e.g. by addition of nucleotides (especially of uridine and adenine) at the 3' end [186]. Argonaute protein modifications, mainly phosphorylation, provide another layer of regulation, relieving target repression or changing the subcellular localization [17]. Nevertheless, the approach that we presented here provides the basis on which more complex, context-specific and even dynamical models describing the impact of miRNA regulation on cellular function can be developed.

### 3.5 ACKNOWLEDGMENTS

Jeremie Breda is a Werner-Siemens fellow at Biozentrum and Rafal Gumienny is supported by the Marie Curie Initial Training Network RNPnet project (no. 289007) from the European Commission. This work was also supported by SystemsX.ch, the systems biology initiative in Switzerland through the RTD project StoNets.

## HIGH-THROUGHPUT IDENTIFICATION OF C/D BOX SNORNA TARGETS WITH CLIP AND RIBOMETH-SEQ

---

### 4.1 INTRODUCTION

RNAs are extensively modified in all living organisms [187]. Recently, high-throughput approaches have been developed to map 2'-O-methylated riboses (2'-O-Me, [116]) and nucleobases carrying the most frequent modifications, including N6-methyladenosine (m6A, [188]), pseudouridine ( $\Psi$ , [189]) and 5-methylcytosine (m5C, [190]), transcriptome-wide. These studies have catalyzed the birth of “epitranscriptomics” [191] and have rekindled the interest in the functions of RNA modifications and their relevance for human diseases [192, 193]. Whereas 2'-O-ribose methylation has long been implicated in the stability and structure of ribosomal RNAs (reviewed in [194]) and m6A appears to modulate the rate of mRNA translation [195–198], 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) [199–201], 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 [202], SNU13/Snu13 [203], NOP56/Nop56 and NOP58/Nop58 [204]. As summarized in [205] 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 H/ACA boxes and signals that direct their localization to Cajal bodies (therefore called small Cajal body-associated RNAs or scaRNAs [206]) 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 [207]. “Anti-sense” elements located upstream of the D and/or

*The work presented in this chapter was originally published in Nucleic Acids Research [3]*

D' box, 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 [208]. Many studies have investigated snoRNA-guided modifications, particularly in yeast [209–212]. As a result, features that define snoRNA target sites have been identified and incorporated into computational methods for snoRNA target prediction [213, 214]. They include a high complementarity to the 3' end of the anti-sense box, with no more than one mismatch over at least 7 nucleotides, and no bulges [214]. A few snoRNAs including U<sub>3</sub>, U<sub>8</sub>, U<sub>13</sub> have been found to be essential for the processing of rRNA precursors in multiple species, whereas U<sub>14</sub> functions in both guiding 2'-O-methylation as well as rRNA precursor processing [90, 215–217]. Until the introduction of the crosslinking, ligation and sequencing of hybrids (CLASH) [218], experimental characterization of snoRNA target sites was laborious and addressed only a few sites at a time [219]. Progress on method development was further driven by the need to generalize target identification approaches to other guide RNAs, such as the miRNAs [60]. 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 [69]. 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 [2]. 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 [116]. 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 *S. cerevisiae* and is now applied to the profiling of rRNA modifications in human cells as well [220]. Studies from various groups have recently expanded the set of human snoRNAs, beyond those that are catalogued in snoRNAbase (<https://www-snoRNA.biotoul.fr/> [91]) [221–224]. Taking advantage of the processing pattern that most C/D-box snoRNAs seem to follow [222] and of the small RNA sequencing data sets generated by the ENCODE consortium, we have recently constructed an updated catalog of human snoRNAs [224]. 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 [116]. Our data supports the concept that some rRNA sites are only partially methylated [220] 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.

## 4.2 MATERIALS AND METHODS

### 4.2.1 CLIP of snoRNP core proteins

To identify chimeric snoRNA-target reads, we analyzed 5 CLIP data sets that were published before [222]: 2 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 [225] (GEO accession GSE77027).

### 4.2.2 Identification of snoRNA-target chimera

#### 4.2.2.1 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 [224]. We downloaded the sequences of known snoRNA targets (rRNA and snRNA) from the snoRNA database [91] and we further obtained tRNA sequences from GtRNAdb [226]. 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.

#### 4.2.2.2 Computational analysis of chimeric reads

Analogous to a previous study that developed a strategy to uncover chimeric miRNA-target reads from Argonaute-CLIP data [69], 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.

#### 4.2.2.3 *Read selection*

We carried out an initial annotation of CLIP data sets with the CLIPZ web server [227], 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 [176]. Thus, to reduce the search space, we used clusters of at least 2 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.

#### 4.2.2.4 *Detection of snoRNAs subsequences in unmapped reads*

To speed up the identification of snoRNA subsequences within unmapped reads we first generated all possible sub-sequences 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<sup>1</sup> (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 (Figure S9A). 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.

---

<sup>1</sup> <https://pypi.python.org/pypi/swalign>

#### 4.2.2.5 *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 [69]. For the mapping we used Bowtie2 [228] 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 1.

#### 4.2.2.6 *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 [213] 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.

### 4.2.3 *Feature definition and computation*

#### 4.2.3.1 *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 [213, 229], 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.

#### 4.2.3.2 *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 [143].

#### 4.2.3.3 *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 (Figure S10) we only used this feature in the model.

#### 4.2.3.4 *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 16) and we therefore trained a generalized linear model (GLM) with the logit link function (logistic regression) using these features, with the Statsmodels python library [148]. 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<sup>2</sup> and additional information is available on the accompanying web site<sup>3</sup>.

#### 4.2.3.5 *Annotation of modification sites*

We annotated the biotypes of the targets in which predicted modification sites resided based on the ENSEMBL version 75 [230] and the RMSK table from University of California Santa Cruz genome

<sup>2</sup> <https://github.com/guma44/snoRNAHybridSearchPipeline>

<sup>3</sup> <http://www.clipz.unibas.ch/snoRNAchimeras>

browser [231], 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 Figure S9B-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.

#### 4.2.4 *RiboMeth-seq*

##### 4.2.4.1 *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 [116]. 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 minutes. 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  $\mu$ g of total RNA as starting material. To explore the existence of 2'-O-methyl sites in mRNAs, we used poly(A)-selected RNA (200ng). In both protocols, the RNA was partially degraded under alkaline conditions in a sodium carbonate/bicarbonate buffer at pH 9.2 for 14 minutes and

then put on ice. Samples were separated parallel to a low molecular weight marker ladder (10-100nt) on a 15% denaturing polyacrylamide gel for 1 hour 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  $\mu$ l of coprecipitant (GlycoBlue) in 75% ethanol at -20°C for 2 hours 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 one hour 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  $\mu$ l mix (4  $\mu$ l H<sub>2</sub>O, 1  $\mu$ l 10x truncated T<sub>4</sub> RNA Ligase 2 buffer, 1  $\mu$ l 100  $\mu$ M 3' rApp-adapter (5' adenylated 3' adapter, 5'-App-TGGAATTCTCGGGTGCCAAGG-amino-3'), 2  $\mu$ l 50% DMSO), denatured at 90°C for 30 seconds and chilled on ice. Next, RNasin Plus RNase inhibitor (Promega) and T<sub>4</sub> RNA Ligase 2 truncated were added to a final concentration of 2 U/ $\mu$ l and 30 U/ $\mu$ l, respectively, and the reaction was incubated at 4°C for 20 hours over night. The next day, 1  $\mu$ l of RT primer (100  $\mu$ M; 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 seconds, at 65°C for 5 minutes, then placed on ice. A 5'-adapter ligation mix was then directly added to the sample (1.5  $\mu$ l 10 mM ATP, 1  $\mu$ l 100  $\mu$ M 5' RNA Adapter RA5 (Illumina TruSeq RNA sample prep kit), 1  $\mu$ l T<sub>4</sub> RNA Ligase 1 (20 U/ $\mu$ l), 0.5  $\mu$ l RNasin Plus RNase inhibitor (40 U/ $\mu$ l) and reactions were incubated at 20°C for 1 h and 37°C for 30 minutes. The RNA was then directly reverse transcribed in a 30  $\mu$ l reaction by adding dNTPs to 0.5 mM, DTT to 5 mM, 1x SSIV buffer, RNasin to 2 U/ $\mu$ l and 1  $\mu$ l 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  $\mu$ l 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  $\mu$ l cDNA in a 100  $\mu$ l 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 hour at 250 V, 20 W. The gel was dismantled and stained for three minutes with GR Green. PCR products between 125 bp and 175 bp were cut out, the gel piece was mashed and DNA was eluted overnight into 400

$\mu\text{l}$  of  $\text{H}_2\text{O}$ . 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  $\mu\text{l}$   $\text{H}_2\text{O}$ . Libraries were sequenced on an Illumina HiSeq-2500 deep sequencer (GEO accession GSE77024). Their summary can be found in Supplementary Table 2.

#### 4.2.4.2 Mapping of RiboMeth-seq reads

We obtained  $\sim 50$  million reads for each of the RiboMeth-seq samples. We removed adaptors with Cutadapt (`--minimum-length 15`, other parameters left with default values) [232] and mapped the reads with STAR (parameters: `-outFilterMultimapNmax 20 -outFilterMismatchNoverLmax 0.05 -scoreGenomicLengthLog2scale 0 -outSAMattributes All`) [233] to a human GRCh37 assembly version-based transcriptome composed of rRNAs, snRNAs, tRNAs and snoRNAs (see section SnoRNA and target sets) as well as to lincRNAs, miscRNAs, and all unspliced protein coding genes (obtained from GRCh37 version of ENSEMBL, <http://grch37.ensembl.org/index.html> [230]).

#### 4.2.4.3 Computation of the RiboMeth-seq score

For each target of interest such as the 18S rRNA, we calculated the  $\log_2$  normalized (to a total library size of 106 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  $180^\circ$  indicates that the frequency of cleavage at the three adjacent positions is identical,  $0^\circ$  indicates that the central position has very high coverage compared to the neighboring positions (and is therefore not protected from cleavage) and  $360^\circ$  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  $290^\circ$  for predicting sites in individual RiboMeth-seq experiments, favoring slightly recall over precision. Detailed statistics for individual experiments can be found in Supplementary Table 2. 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 7 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.

#### 4.2.5 Validation of 2'-O-methylation sites with RTL-P

Similar to the classic primer extension assays [234], the “Reverse Transcription at Low deoxy-ribonucleoside triphosphate (dNTP) followed by polymerase chain reaction” method (RTL-P, [235]) 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 3).

#### 4.2.6 Validation of 2'-O-methylation at G2435 in 28S with mass spectrometry

The rRNA fragment isolation for mass spectrometry analysis (MS) was adapted from [236]. 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 pmoles 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 [237]. 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 [Figure S11](#).

### 4.3 RESULTS

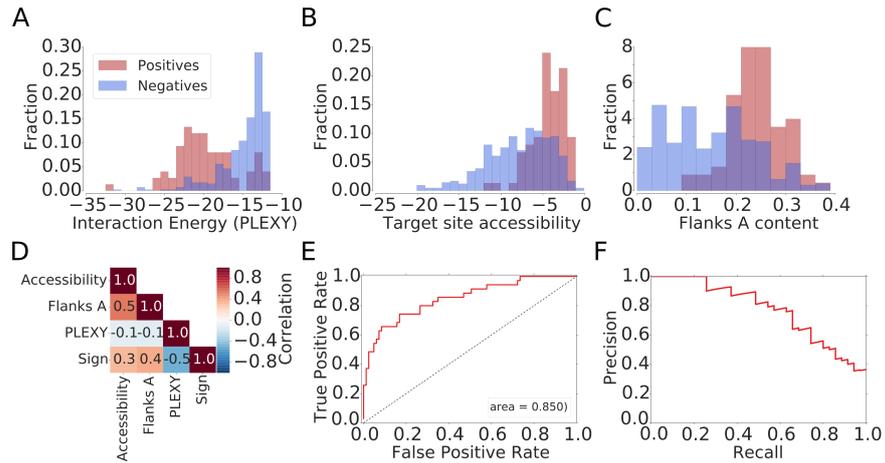
#### 4.3.1 Crosslinking and immunoprecipitation of core snoRNPs captures snoRNA-target site chimeras

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 [69], we hypothesized that chimeric molecules, composed of snoRNAs and their targets, are cap-

tured 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 6 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 ~5% 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 1. 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 [213]. 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 17A). This indicated that a more sophisticated approach is needed to reliably identify snoRNA-target interactions from these data, which likely contains a large number of false positives.

#### 4.3.2 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 [1] between the two sets. The PLEXY interaction score [213] discriminated best these two data sets (as shown in Figure 16A and D). However, known snoRNA target sites also reside in structurally accessible regions (Figure 16B), rich in adenines (Figure 16C). 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 16E,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,

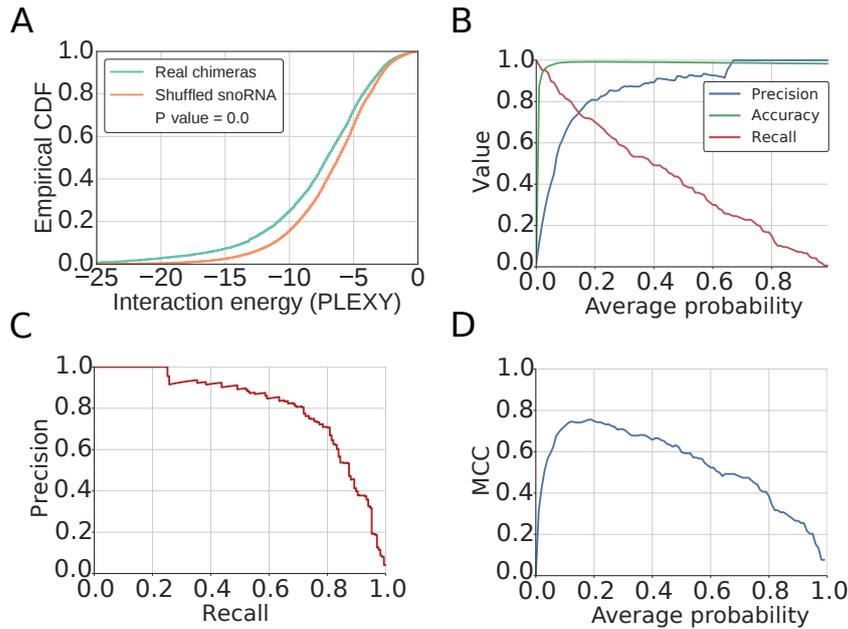


**Figure 16:** Features that are relevant for the identification snoRNA-target interactions based on chimeric reads. Distributions of (A) the interaction energy calculated with PLEXY [213], (B) the target site accessibility calculated with CONTRAfold [143] and (C) the A nucleotide composition of the 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.

with a Matthews correlation coefficient (MCC) of  $\sim 0.75$ , precision of 0.75 and recall value of 0.74 (Figure 17B-D). Our predictions finally consisted of putative interactions that were supported by chimeric reads from at least 2 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 Figure S12.

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

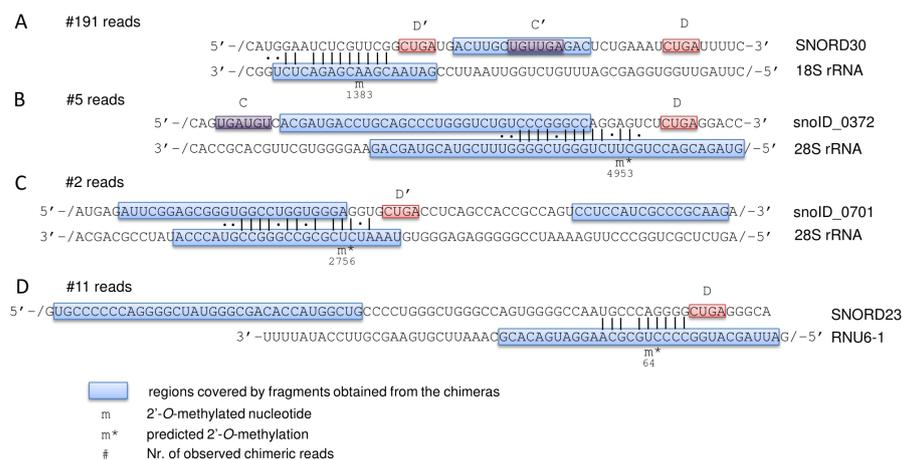
We applied the derived model to the full chimeric read data and identified 980 putative interactions, involving 852 unique target sites. We focused on the snoRNA interactions with structural RNAs, including not only the rRNAs, but also snRNAs, tRNAs and the snoRNAs themselves. Only one of the 2'-O-Me sites in rRNAs that have been mapped so far is "orphan", meaning that its guide snoRNA is unknown. Our data indicates that this modification, located at position A1383 in the 18S rRNA [238], is guided by SNORD30 (Figure 18A), a snoRNA which was reported to guide the 2'-O-methylation at position A3804 in 28S rRNA [239]. The chimeric reads also revealed 35 potentially novel 2'-O-Me sites in rRNAs (13 in 18S rRNA, 21 in 28S



**Figure 17:** 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 across CLIP data sets. (C) Precision-Recall curve for the method. (D) Matthews correlation coefficient (MCC) as a function of the minimum average probability of the considered sites and the derived optimal threshold.

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 4). 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 18B). Similarly, in two experiments we found the recently uncovered snoID\_0701 (family unknown) orphan snoRNA, which has low but broad expression across tissues [224], 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 18C).

SnRNAs are also known targets of scaRNA-guided 2'-O-methylation. Of the 9 such sites that are known, we were able to recover 4 over our prediction threshold. Additionally, we identified chimeric reads of the SNORD23, a snoRNA that is currently considered orphan, with the U6 snRNA (Figure 18D) In previous work [222] we have studied the



**Figure 18:** 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'.

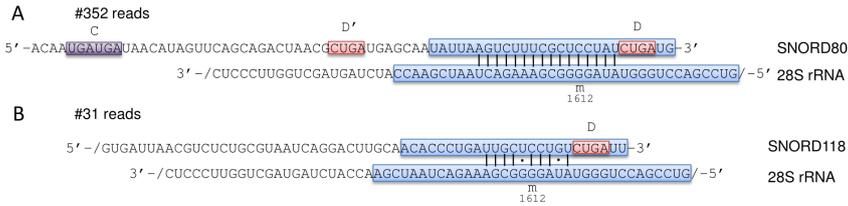
methylation pattern of this snRNA by primer extension. We found evidence of 2'-O-methylation at positions 60, 62 and 63 of U6, but not at position 64, which is predicted to be modified as a result of the interaction with SNORD23. Thus, the significance of this interaction, supported by 11 reads in our data, remains to be determined.

Additionally, we identified 3 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 4 and all alignments of putative chimeric reads to putative target sites and snoRNAs can be viewed at CLIPz website<sup>4</sup>.

#### 4.3.4 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 [224]. 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

<sup>4</sup> <http://www.clipz.unibas.ch/snoRNAchimeras>



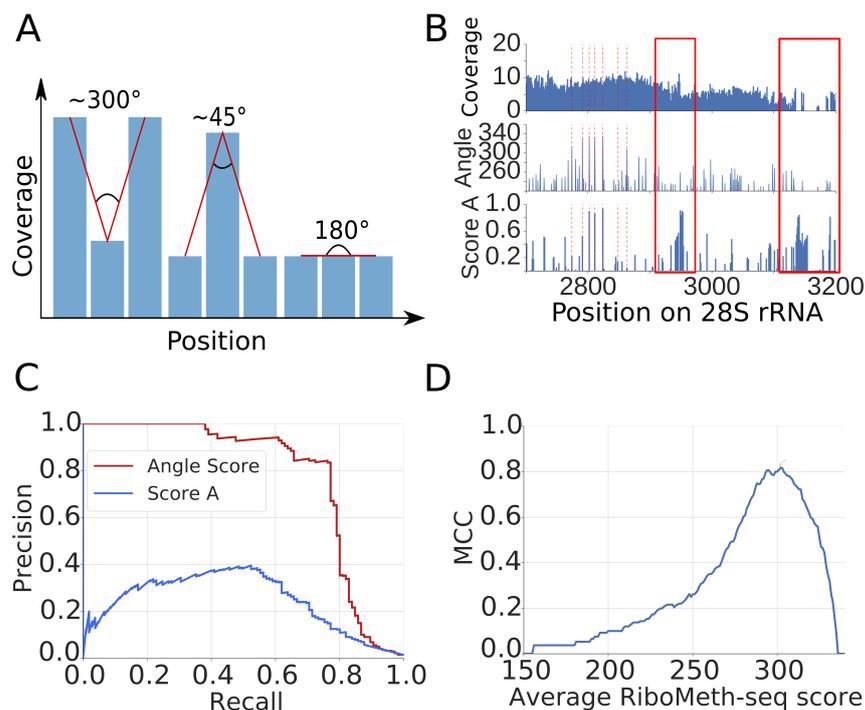
**Figure 19:** Similar to [Figure 18](#), representation of the data supporting the interaction of both SNORD80 and SNORD118 with the 28S rRNA, around the known position of 2'-O-methylation at G1612.

U8, a snoRNA which is necessary for the proper maturation of 5.8S and 28S rRNAs [216], interacts with the region of the 28S rRNA where the SNORD80 was known to guide the modification of G1612. The evidence for this interaction is very strong, chimeras having been captured in 6 distinct experiments ([Figure 19](#)). 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.

#### 4.3.5 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 [116]. 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.

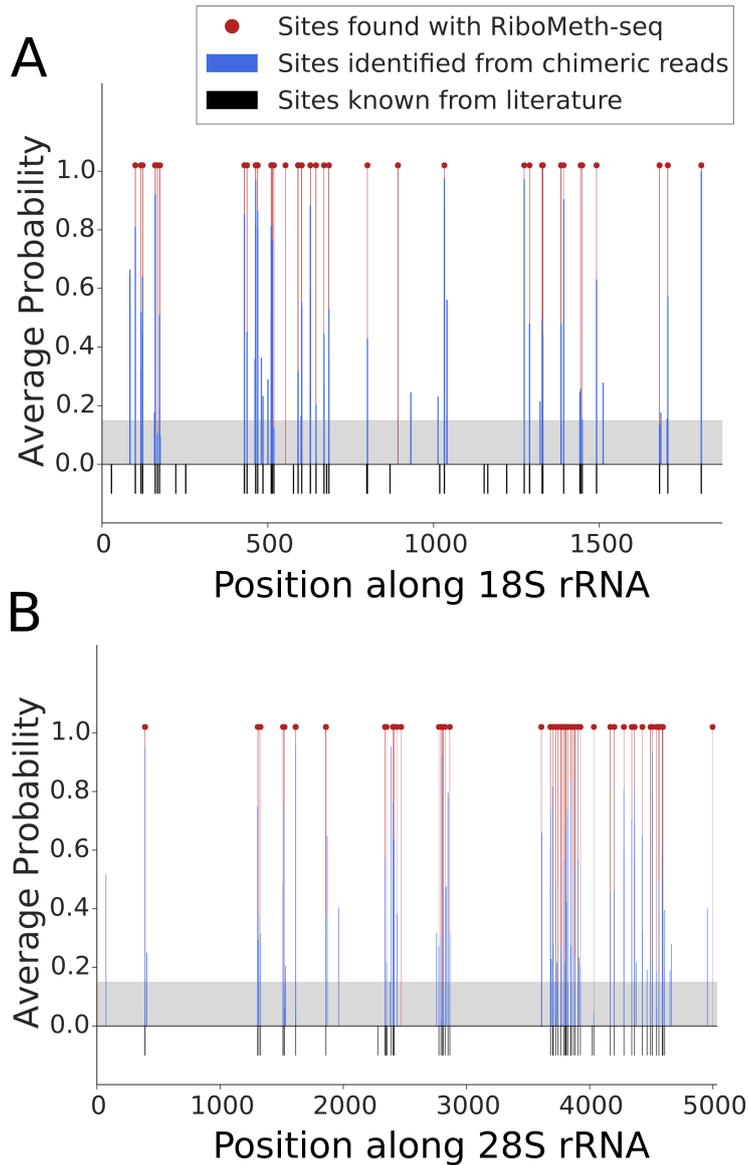
2'-O-Me sites were previously identified from RiboMeth-seq by comparing the number of reads ending at a particular position in the target with the average number of reads ending at the flanking regions ("score A" in [116]). Reasoning that 2'-O-methylation of adjacent nucleotides is very rare and that 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<sub>2</sub> normalized read coverage at the specific position and the immediately adjacent positions



**Figure 20:** Analysis of RiboMeth-seq data. (A) Strategy for evaluating the RiboMeth-seq data. The score was calculated based on the normalized  $\log_2$  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  $\log_2$  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 RiboMeth-Seq\_HEK\_totalRNA\_8min experiment. (D) Matthews correlation coefficient (MCC) plot of average RiboMeth-seq score indicating the optimal angle score.

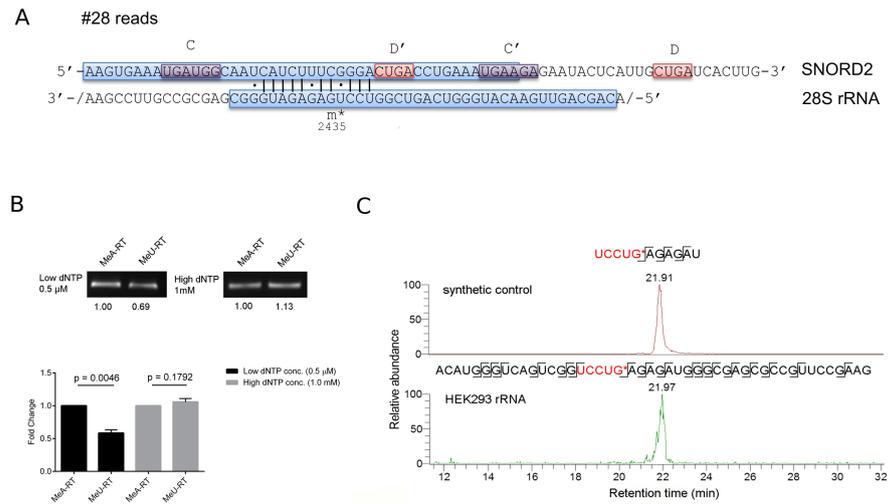
(Figure 20A). We found that this score yields a higher precision compared to the 'score A' proposed before [116] (Figure 20B and C) and a very high Matthews correlation coefficient in classifying the sites (Figure 20D).

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 21 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 tar-



**Figure 21:** 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 grey 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.

gets - 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.



**Figure 22:** 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 18](#)). (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 3 replicate experiments, each with 3 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.

#### 4.3.6 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 22A](#)).

We applied the recently published method “Reverse Transcription at Low deoxy-ribonucleoside triphosphate concentrations followed by polymerase chain reaction” (RTL-P) [235], 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) ([Figure S13](#)), 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 22B](#)), 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 [237]. Figure 22C 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 (Figure S11A) and found that they matched those obtained from the HEK rRNA (Figure S11B), confirming the presence of UCCUG\* in the HEK sample. The LC-MS analysis also identified the unmodified fragment UCCUG from HEK rRNA (Figure S11C), albeit at a lower level than UCCUG\* (Figure S11D). These results show that the G2435 28S rRNA site identified among the chimeric reads is predominantly 2'-O-methylated.

#### 4.3.7 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).

## 4.4 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 led to the realization that the non-coding transcriptome is as large as the protein-coding fraction [240]. New members of all classes of RNAs, including miRNAs and snoRNA have also been discovered [241, 242]. 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.

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 [69]. 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 [69]), the large "background" of CLIP [176], 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 [216], 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 [116]. 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 [220, 224]. Of note, the interaction of SNORD48 with C1868 in the 28S rRNA, presumed to lead to the observed partial methylation of this site [220] 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 [211, 215–217, 243, 244]. 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. Although PLEXY enforces the snoRNA interaction with the target to take place close to already

annotated D boxes and we do not expect such cases in our final list of candidates, a careful inspection of the hybrids and chimeric read alignments that we provide on the accompanying web site should help identify them.

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 [245].

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 [246]. Interestingly, a recent study reported that G3771 in the 28S rRNA is 2'-O-methylated, guided by SNORD15A. 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 [11] 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.

#### 4.5 SUPPLEMENTARY DATA

Supplementary Tables are available at NucleicAcidsResearch website<sup>5</sup> or can be requested from the author.

#### 4.6 FUNDING

Rafal Gumienny was supported by the Marie Curie Initial Training Network RNPnet project (No. 289007) from the European Commission. Foivos Gypas was supported by the Marie Curie Initial Training Network RNAtrain project (No. 607720) from the European Commission.

#### 4.7 ACKNOWLEDGEMENTS

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

---

<sup>5</sup> <https://doi.org/10.1093/nar/gkw1321>

SUMMARY

---

A vast body of work in the recent years has come to the conclusion that the non-coding fraction of the transcriptome is as big as the protein-coding fraction. This shifted the research focus from proteins to non-coding RNAs to such an extent that sometimes, this period it is referred to as the non-coding revolution. New members of the existing families and totally new families of molecules are being constantly discovered. This rapid pace of discovery brings with it the challenge of developing novel, high-throughput methods to functionally characterize these new players. In this thesis I have discussed novel methods that I and other members of the group have developed to identify targets of two important classes of small non-coding RNAs: the miRNAs that regulate gene expression and the snoRNAs that guide 2'-O-methylation of rRNAs. We approached this problem using a wide variety of techniques that ranged from novel high-throughput experimental techniques for which appropriate analysis methods had to be developed, to computational modelling.

Our new method to predict miRNA targets, MIRZA-G, incorporates several sequence and structure features of miRNA binding sites together with a biophysical model of miRNA-target interaction into an improved algorithm that can be used not only for miRNA target prediction but also siRNA off-target searches. We have also demonstrated that improved data sets of validated miRNA targets, as those that can be obtained from CLIP experiments, further enable improvements of the MIRZA model. And yet, the need for improvements in small RNA target prediction approaches remains. In our model we have used fairly understandable and interpretable miRNA-target interaction features like energy of interaction, features of the structure and sequence of the target and, as an effective parameter that captures a lot of unknowns, the degree of evolutionary conservation of the putative target site. In the future, what we would most like to do is the be able to replace this parameter with others, that are more directly reflecting the mechanism of miRNA-target interaction and that derive from a better understanding of miRNA silencing mechanism. As several groups are working in measuring the affinity of miRNA-target interaction in high throughput, this aspect may be accessible in the very near future. A second source of uncertainty in the accuracy of miRNA-target interaction, that we currently do not take into account in our models, is the cellular context in which the interaction takes place. It is known that the transcriptome vastly differs between tissues, as does the set and concentration of RNA-binding proteins that

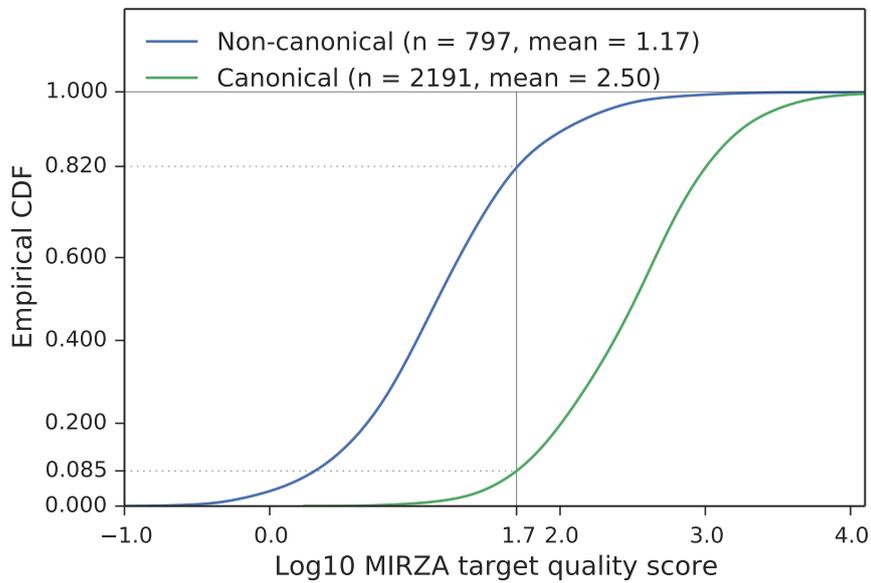
can promote or hinder the interaction of miRNAs with targets. Incorporating such information into models eg. tissue-specific transcript isoform expression, relative concentration of RNA binding proteins and knowledge about their sequence specificity is expected to greatly enhance the prediction of miRNA targets [160]. Applying our miRNA target prediction method to the interpretation of siRNA screening results is promising, and is one of the directions of work in the lab. A server for predicting off-targets of siRNAs would probably be of use to the community as well. The server could be used to predict targets of newly discovered miRNAs, although at the moment, computing all the features that go into the construction of the model, in particular the degree of evolutionary conservation and the accessibility of the target site are quite time consuming, and not easy to do 'on-line'.

Although a lot of efforts were made to decipher the functions and mechanisms of snoRNAs by many scientists, there are still various open questions awaiting to be answered. One of the most obvious ones is to find the targets of the highly expressed and brain-specific snoRNAs, like SNORD116, whose lack of expression leads to the Prader-Willi Syndrome [107–109]. A related question is whether the currently orphan snoRNAs have other, non-canonical, functions; do they really influence alternative splicing or are they processed into smaller RNAs and act as miRNAs as have been proposed [98, 247, 248]? Attempting to find answers to these questions we have combined two high-throughput approaches, PAR-CLIP, which yields chimeric reads that reveal direct interactions between snoRNAs and their targets, and RiboMeth-seq, which enables the mapping of 2'-O-methylations genome-wide. Our aim was to build the tools that in the future will help us investigate more deeply unexpected roles of snoRNAs and take a step into global analysis of snoRNAs interactions, including interactions that do not obviously result in 2'-O-methylation. We have identified many novel canonical 2'-O-methylation sites that were supported by chimeric reads. Additionally, we have shown that many of the snoRNAs previously considered orphan can be found within chimeras with known 2'-O-methylation sites, showing a substantial redundancy in targeting. We have demonstrated that most of the high confidence non-canonical target sites found in chimeric reads are not detected to be methylated suggesting another role for snoRNAs interacting with mRNAs. It is important to mention that our experiments were performed in HEK cells, therefore the detected sites might be specific to this cell type. It is highly probable that the pattern of 2'-O-methylation changes during development and/or cell type. This can significantly contribute to the cell identity and have influence on carcinogenesis as it is known that snoRNAs could be dysregulated in the different cancers (Jojani et al. submitted). Recently, it has been realized that the ribosomes itself are not so static and they can occur in the cell in many flavours differing in eg. protein content

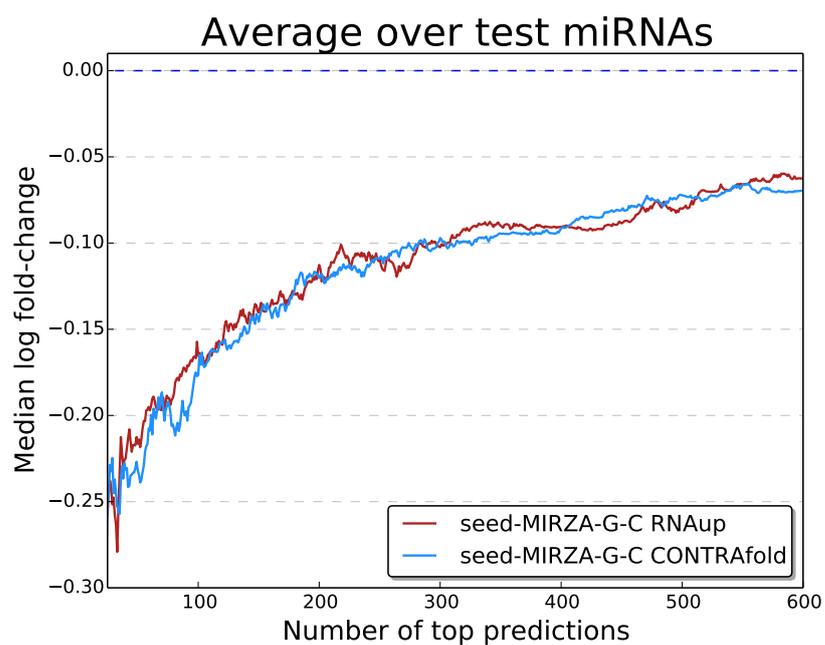
[104]. It is likely that differences in 2'-O-methylation and other rRNA modifications can contribute to this phenomenon known as ribosome heterogeneity. These hypotheses could be answered in the future with the techniques we established in this thesis.

Although they are small, non-coding RNAs play important roles in biology, which have been underscored recently by the discovery of many new small RNAs. They participate in almost, if not all, cellular processes. We hope that with the techniques and insights established here we brought the community a little step further in the understanding and appreciation of these fascinating molecules and that we provided helpful tools to make further progress possible.

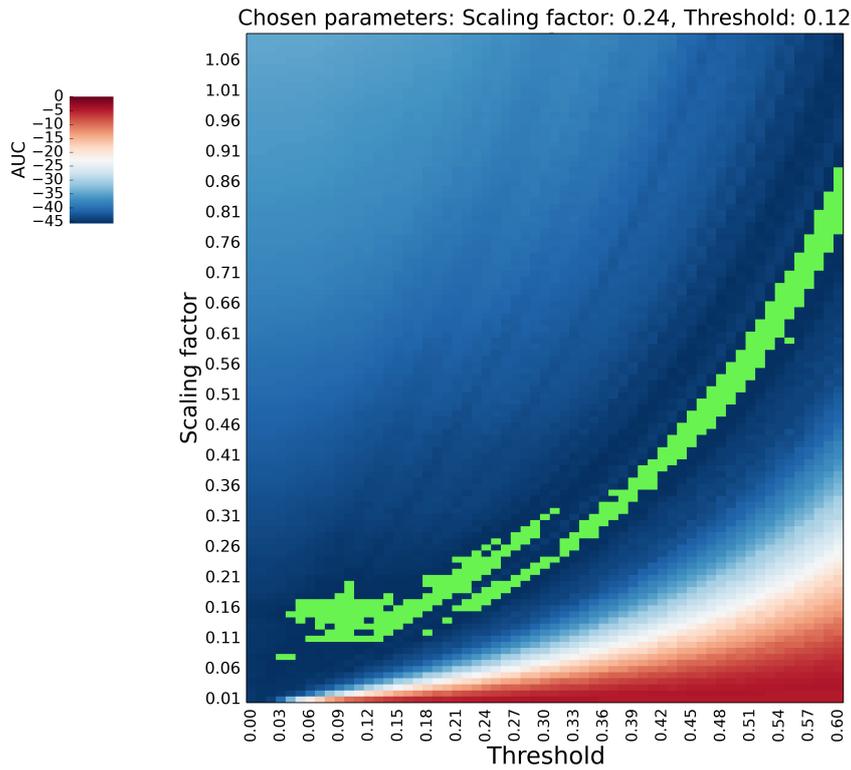




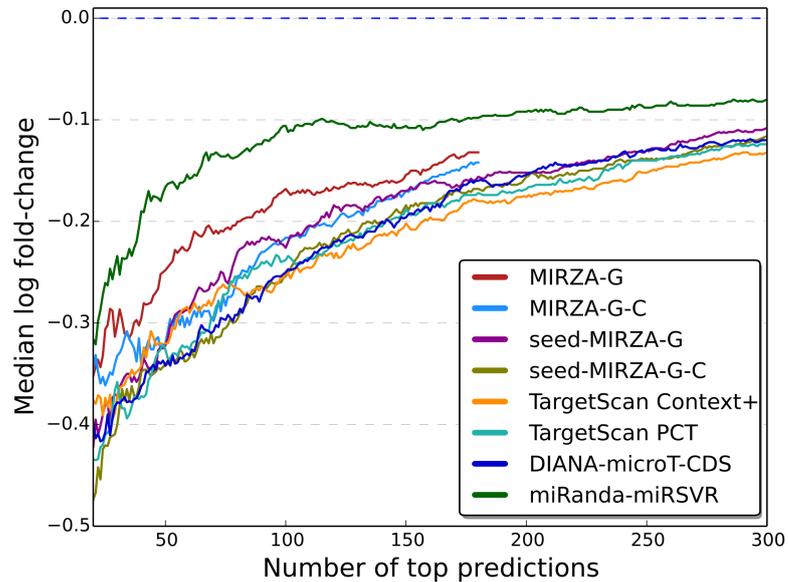
**Supplementary Figure S1:** Empirical cumulative distribution function of MIRZA target quality scores for canonical (green) and non-canonical (blue) miRNA binding sites. The binding sites were obtained with Argonaute 2 crosslinking and immunoprecipitation interactions and binding sites of individual miRNAs were predicted with the MIRZA method [59].



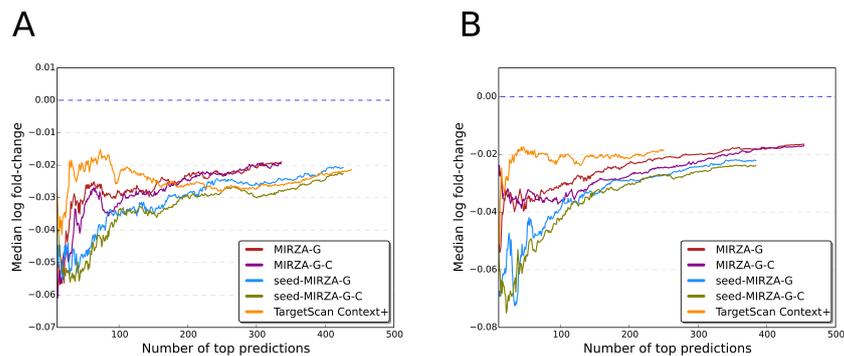
**Supplementary Figure S2:** Comparison of the down-regulation of targets predicted with models that either used RNAup or CONTRAfold to estimate target site accessibility upon miRNA transfection. Both models were trained as described in the Methods section on the ‘training set’ of miRNA transfection data and were then tested on the ‘test’ set of experiments. Y-axis shows  $\log_2$  fold-changes.



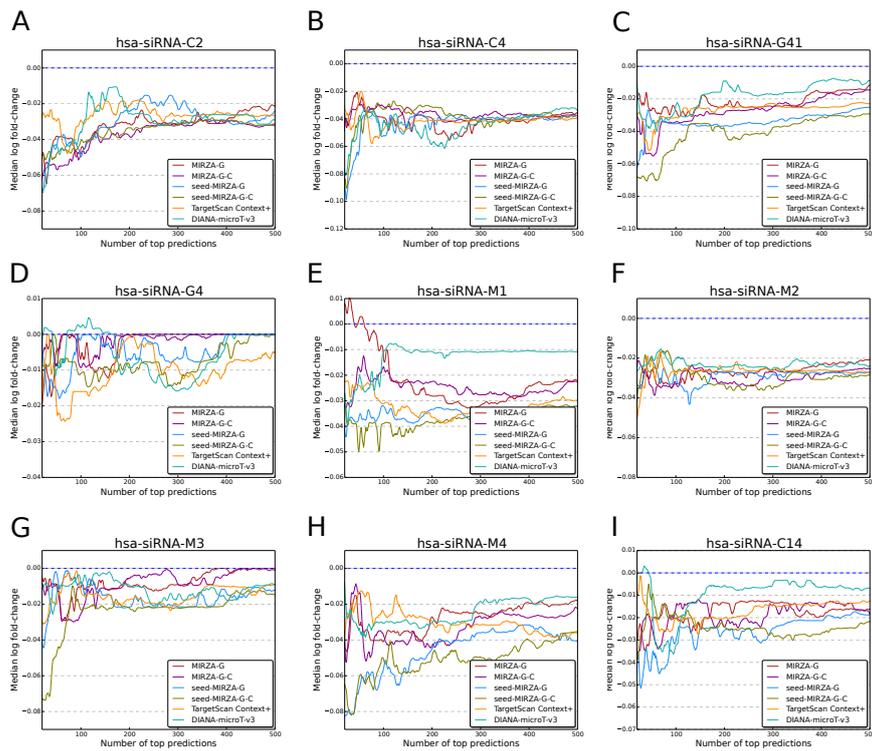
**Supplementary Figure S3:** Optimization of scaling factor ( $K$ ) used to score individual sites and threshold ( $\tau$ ) used to compute gene-level scores. Predictions were made using individual  $K$ - $\tau$  pairs, the median down-regulation as a function of the number of top predictions considered was computed, and then the total down-regulation over the entire range of targets ('AUC') was calculated. The optimal values of the parameters were considered those that lead to predictions with the strongest overall down-regulation and were highlighted in green.



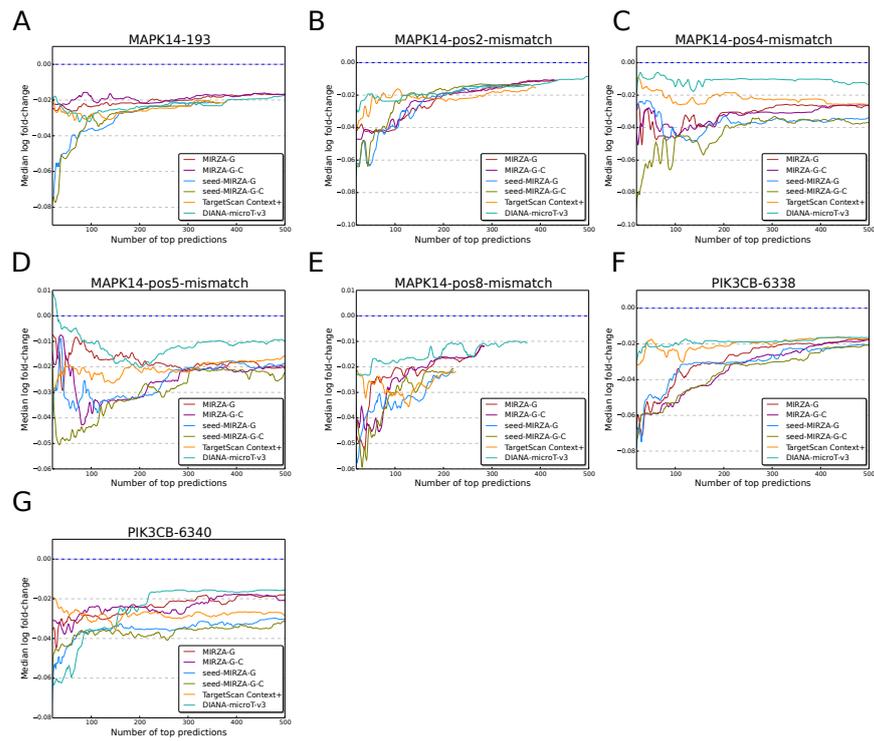
**Supplementary Figure S4:** Comparative evaluation of the performance of various models in predicting protein down-regulation following miRNA transfection. Variants of the MIRZA-G model are described in [Table 2](#). The other tested models are TargetScan Context+, TargetScan PCT, DIANA-microT (the newest version), and miRanda-miRSVR (the most conservative predictions). Y-axis shows  $\log_2$  fold-changes.



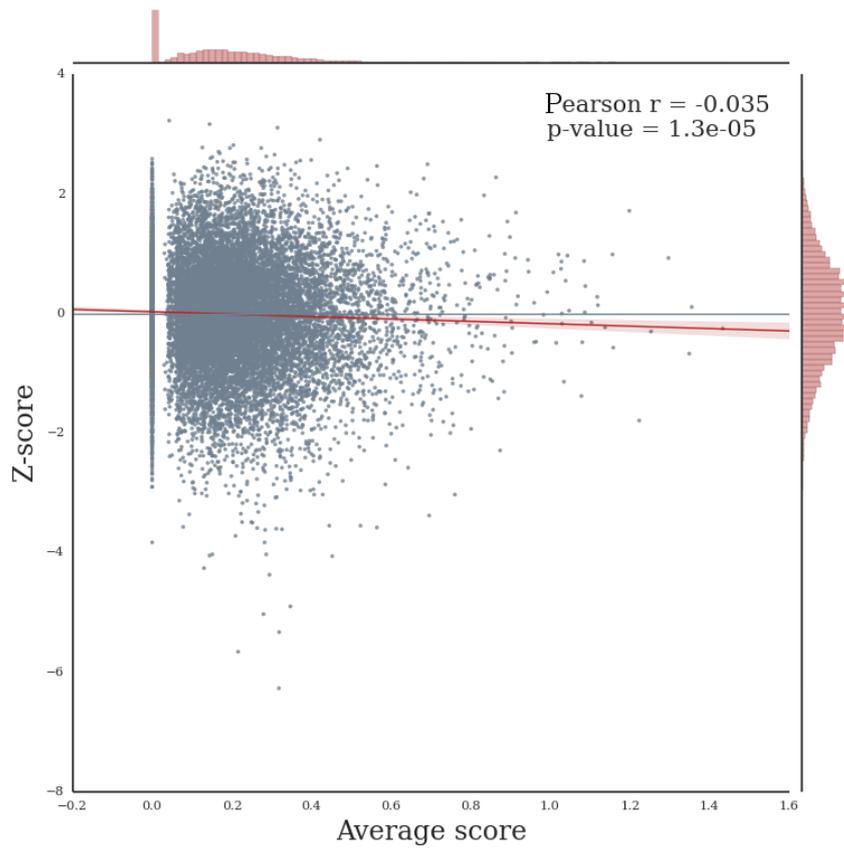
**Supplementary Figure S5:** Comparison of performance of the different models in predicting off-targets of siRNAs from the Birmingham et al. [22] (A) and Jackson et al. [23] (B) studies. Only siRNAs that did not share 6 or more nucleotides in the seed region with a known miRNA were used. Y-axis shows  $\log_{10}$  fold-changes.



**Supplementary Figure S6:** Performance comparison of various models on individual siRNA transfections (siRNAs labeled on the top of each panel) from the Birmingham et al. [22] dataset. Y-axis shows  $\log_{10}$  fold-changes.

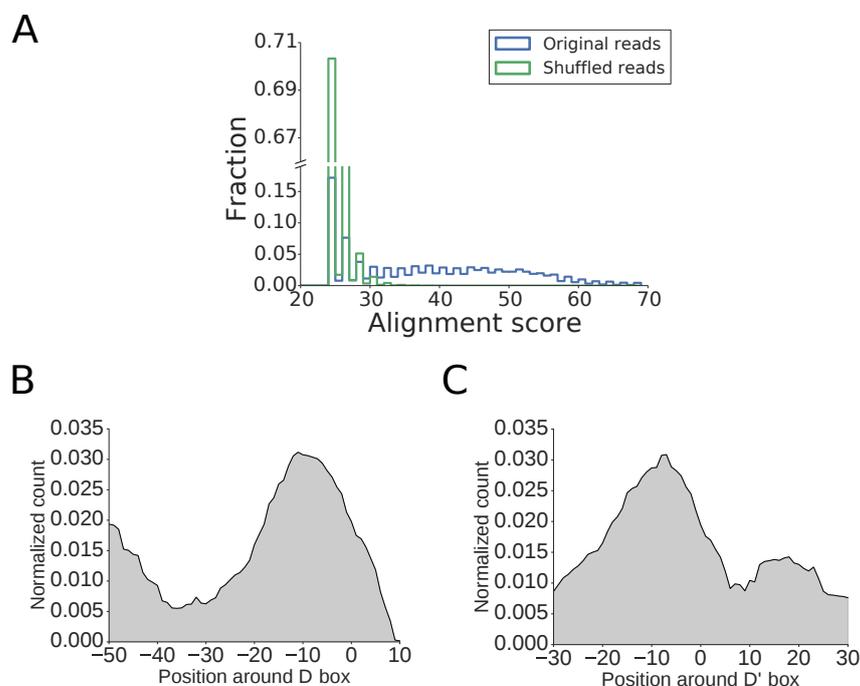


**Supplementary Figure S7:** Performance comparison of various models on individual siRNA transfections (siRNAs labeled on the top of each panel) from the Jackson et al. [23] dataset. Y-axis shows  $\log_{10}$  fold-changes.

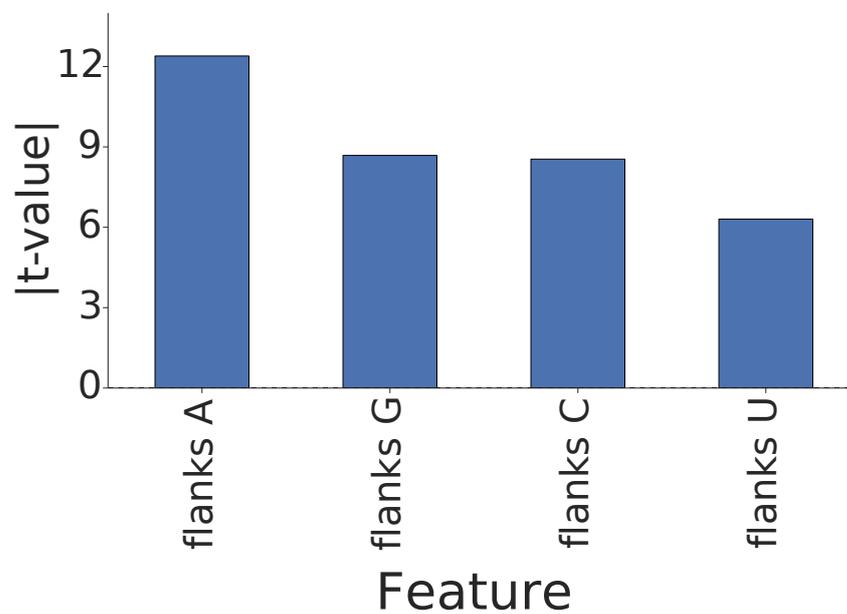


**Supplementary Figure S8:** Correlation between the z-score of an siRNA in the TGF-screen and the average score that our model assigns to the interaction of core components of the TGF- $\beta$  pathway (TGFB $\beta$ 2, TGFB $\beta$ 1, SMAD2 and SMAD4) with the siRNA.

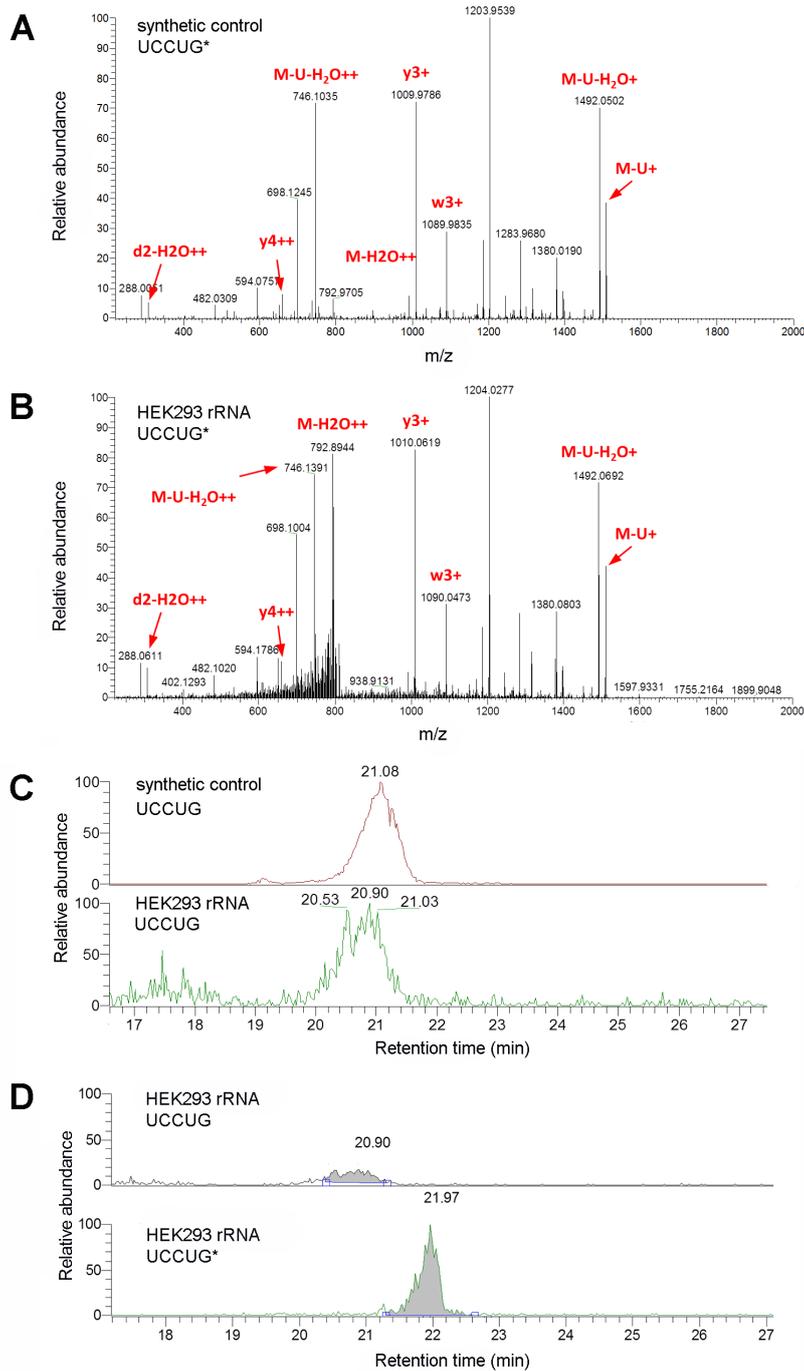


SUPPLEMENTARY MATERIALS FOR SNORNA  
CHIMERAS

**Supplementary Figure S9:** (A) Distribution of the Smith-Waterman score of the alignment between known snoRNAs and unmapped reads obtained in the experiment, as well as shuffled versions of the unmapped reads. Real sequences are enriched in matches to snoRNAs compared to randomized sequences. (B-C) Coverage profiles of snoRNAs that guide a known 2'-O-methylation through only one of their anti-sense boxes located either at the D box (B) or at the D' box (C), by fragments found in chimeric reads. For each snoRNA the profile was normalized such that the integral of the read density was 1, and then the normalized read density profiles were cumulated over the snoRNAs. For snoRNAs whose guide sequence is located at the D' box, only snoRNAs that were long enough to cover all the positions indicated in the figures were used.



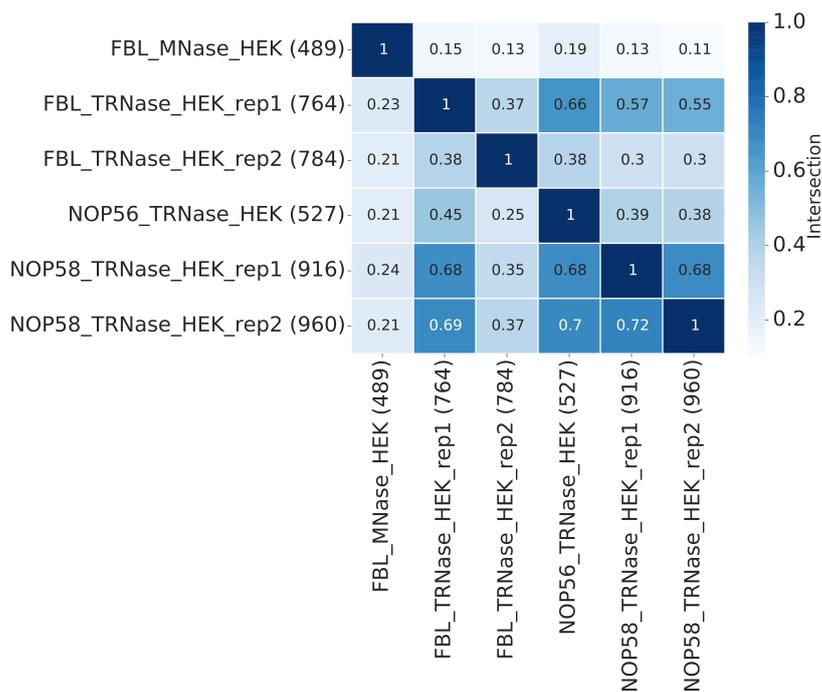
**Supplementary Figure S10:** Absolute value of t-statistic obtained in comparing the frequency of individual nucleotides in flanking regions (30 nts upstream/downstream) of chimeric read-supported-sites (7 nt sites anchored at the 5' end by the nucleotide hybridizing with the snoRNA nucleotide immediately upstream of the D box) where 2'-O-methylation is known and not known, respectively, to take place.



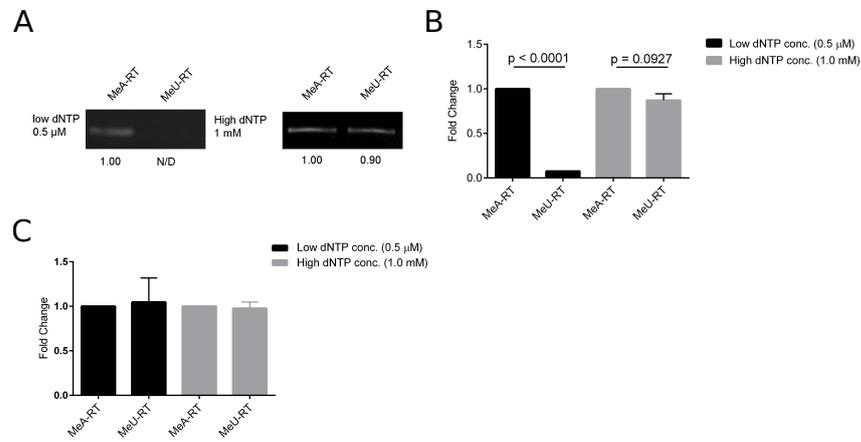
**Supplementary Figure S11:** Targeted LC-MS/MS analysis of the G2435 site in 28S rRNA (A) MS/MS spectra of the UCCUG\* fragment acquired from a synthetic RNA oligonucleotide control and (B) from the 28S rRNA A2416-G2461 fragment. MS/MS fragments generated after collision induced dissociation were manually assigned and are indicated.

**Supplementary Figure S11:** These assigned and specific fragments were selected as transitions for targeted LC-MS/MS analysis. (C) Targeted LC-MS/MS analysis of the unmodified UCCUG confirms the presence of un-methylated G2435 in 28S rRNA. (D) Targeted LC-MS/MS analysis shows that G2435 is predominantly 2'-O-methylated in the HEK293 sample. The LC-MS/MS experiment was performed as follows. 1.5 nmoles of a synthetic oligodeoxynucleotide (CTT CGG AAC GGC GCT CGC CCA TCT CTC AGG ACC GAC TGA CCC ATG T) complementary to A2416-G2461 of 28S rRNA were incubated with 150 µg of total RNA isolated from HEK 293 cells in 0.3 volumes of hybridization buffer (250 mM HEPES, 500 mM KCl at pH 7) in 75 µl total volume. The mixture was incubated for 5 min at 90°C and then allowed to hybridize during slow cooling to 45°C over 2.5 h. Mung bean nuclease buffer was added to a final concentration of 50 mM NaOAc (pH 5) at 25°C, 30 mM NaCl, and 1 mM ZnCl<sub>2</sub> together with 60 units of mung bean nuclease (New England Biolabs) and 1.5 µg of RNase A (Sigma-Aldrich), followed by a 60-min incubation at 35°C. The reaction was phenol/chloroform-extracted and the RNA:DNA hybrid was ethanol-precipitated. The precipitate was redissolved in a 1X formamide-gel loading buffer. The specific rRNA sequence was purified using a 12% polyacrylamide gel containing 7M urea. The bands were visualized by SYBR Gold Nucleic Acid Gel Stain (Thermo Fisher Scientific). The excised rRNA fragment was eluted overnight at 4°C in 2 M NH<sub>4</sub>Ac (pH 5.3). The RNA was precipitated from the eluate by adding 1 volume of ethanol and 1 volume of isopropanol. The pellet was dissolved in 12 µl H<sub>2</sub>O and 40 units of RNase T1 and digested at 37°C for 3.5 h. The sample was dephosphorylated by adding of 1.5 µL of alkaline phosphatase buffer (0.5 M Tris-HCl, 1 mM EDTA at pH 8.5 and 20°C) and 0.2 units of FastAP Thermosensitive Alkaline Phosphatase (Thermo Fisher Scientific) and incubating at 37°C for 45 min. The samples were purified on C18 Microspin columns (The Nest Group, Inc.) according to manufacturers, dried under vacuum and subjected to LC-MS analysis. 1 pmol of RNA oligos of each sample were subjected to LC-MS analysis using a dual pressure LTQ-Orbitrap Elite mass spectrometer connected to an electrospray ion source (both Thermo Fisher Scientific) as described recently [249] with a few modifications. In brief, RNA oligo separation was carried out using an EASY nLC-1000 system (Thermo Fisher Scientific) equipped with a RP-HPLC column (75µm × 30cm) packed in-house with C18 resin (ReproSil-Pur C18-AQ, 1.9µm resin; Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) using a linear gradient from 95% solvent A (0.15% formic acid, 2% acetonitrile) and 3% solvent B (98% acetonitrile, 0.15% formic acid) to 20% solvent B over 40min at a flow rate of 0.2µl/min.

**Supplementary Figure S11:** The data acquisition mode was set to obtain one high resolution MS scan in the FT part of the mass spectrometer at a resolution of 240000 full width at half-maximum (at  $m/z$  400) followed by targeted MS/MS scans in the linear ion trap of UCCUG ( $m/z = 753.619, 780.07$  (adduct)) and UCCUG\* ( $m/z = 760.627, 811.05$  (adduct)). The ion accumulation time was set to 300ms (MS) and 200ms (MS/MS). The collision energy was set to 35%, mass selection window was set to 2Th and one microscan was acquired for each spectrum. AGC settings were 3E4 for MS2 scans and 1E6 for MS1 scans. Data analysis was carried out using the Qual Browser tool of the Xcalibur software (version: 3.0.63). RNA oligo precursor and fragment monoisotopic masses were calculated using the Mongo Oligo Mass Calculator v2.06 (<http://mods.rna.albany.edu/masspec/Mongo-Oligo>).



**Supplementary Figure S12:** Intersection of interactions (over the threshold probability 0.15) captured in chimeras in pairs of CLIP experiments. The matrix shows the fraction of common interactions and the numbers in parentheses correspond to the number of interactions above the threshold in each experiment.



**Supplementary Figure S13:** Positive and negative controls for RTL-P (A) Detection of known Am1031 in 18S rRNA by RTL-P followed by agarose gel analysis (B) and qPCR analysis. (C) Negative control: Absence of 2'-O-methylation in U1991 28S rRNA is demonstrated by qPCR analysis.

## BIBLIOGRAPHY

---

- [1] Rafal Gumienny and Mihaela Zavolan. "Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G." In: *Nucleic Acids Res.* (Jan. 2015).
- [2] Jeremie Breda, Andrzej J Rzepiela, Rafal Gumienny, Erik van Nimwegen, and Mihaela Zavolan. "Quantifying the strength of miRNA-target interactions." In: *Methods* 85 (Sept. 2015), pp. 90–99.
- [3] Rafal Gumienny, Dominik J Jedlinski, Alexander Schmidt, Foivos Gypas, Georges Martin, Arnau Vina-Vilaseca, and Mihaela Zavolan. "High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq." en. In: *Nucleic Acids Res.* (Dec. 2016).
- [4] F H Crick. "On protein synthesis." In: *Symp. Soc. Exp. Biol.* 12 (1958), pp. 138–163.
- [5] F Jacob and J Monod. "Genetic regulatory mechanisms in the synthesis of proteins." In: *J. Mol. Biol.* 3 (June 1961), pp. 318–356.
- [6] W Gilbert and B Müller-Hill. "Isolation of the lac repressor." In: *Proc. Natl. Acad. Sci. U. S. A.* 56.6 (Dec. 1966), pp. 1891–1898.
- [7] R A Weinberg and S Penman. "Small molecular weight monodisperse nuclear RNA." In: *J. Mol. Biol.* 38.3 (Dec. 1968), pp. 289–304.
- [8] A W Prestayko, M Tonato, and H Busch. "Low molecular weight RNA associated with 28 s nucleolar RNA." In: *J. Mol. Biol.* 47.3 (1970), pp. 505–515.
- [9] H Busch, R Reddy, L Rothblum, and Y C Choi. "SnRNAs, SnRNPs, and RNA processing." In: *Annu. Rev. Biochem.* 51 (1982), pp. 617–654.
- [10] M R Lerner, J A Boyle, S M Mount, S L Wolin, and J A Steitz. "Are snRNPs involved in splicing?" In: *Nature* 283.5743 (1980), pp. 220–224.
- [11] Z Kiss-László, Y Henry, J P Bachellerie, M Caizergues-Ferrer, and T Kiss. "Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs." en. In: *Cell* 85.7 (June 1996), pp. 1077–1088.
- [12] R C Lee, R L Feinbaum, and Ambros, V. "The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*." In: *Cell* 75.5 (Dec. 1993), pp. 843–854.

- [13] A Fire, S Xu, M K Montgomery, S A Kostas, S E Driver, and C C Mello. "Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*." In: *Nature* 391.6669 (1998), pp. 806–811.
- [14] C Napoli, C Lemieux, and R Jorgensen. "Introduction of a Chimeric Chalcone Synthase Gene into *Petunia* Results in Reversible Co-Suppression of Homologous Genes in trans." In: *Plant Cell* 2.4 (Apr. 1990), pp. 279–289.
- [15] N Romano and G Macino. "Quelling: transient inactivation of gene expression in *Neurospora crassa* by transformation with homologous sequences." In: *Mol. Microbiol.* 6.22 (Nov. 1992), pp. 3343–3353.
- [16] Ming-Bo Wang and Michael Metzloff. "RNA silencing and antiviral defense in plants." In: *Curr. Opin. Plant Biol.* 8.2 (Apr. 2005), pp. 216–222.
- [17] Minju Ha and V Narry Kim. "Regulation of microRNA biogenesis." In: *Nat. Rev. Mol. Cell Biol.* 15.8 (Aug. 2014), pp. 509–524.
- [18] E Bernstein, A A Caudy, S M Hammond, and G J Hannon. "Role for a bidentate ribonuclease in the initiation step of RNA interference." In: *Nature* 409.6818 (2001), pp. 363–366.
- [19] Richard I Gregory, Thimmaiah P Chendrimada, Neil Cooch, and Ramin Shiekhattar. "Human RISC couples microRNA biogenesis and posttranscriptional gene silencing." In: *Cell* 123.4 (2005), pp. 631–640.
- [20] Dianne S Schwarz, György Hutvagner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D Zamore. "Asymmetry in the assembly of the RNAi enzyme complex." In: *Cell* 115.2 (2003), pp. 199–208.
- [21] Traci M Tanaka Hall. "Structure and function of argonaute proteins." In: *Structure* 13.10 (Oct. 2005), pp. 1403–1408.
- [22] Amanda Birmingham et al. "3' UTR seed matches, but not overall identity, are associated with RNAi off-targets." In: *Nat. Methods* 3.3 (2006), pp. 199–204.
- [23] Aimee L Jackson, Julja Burchard, Janell Schelter, B Nelson Chau, Michele Cleary, Lee Lim, and Peter S Linsley. "Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity." In: *RNA* 12.7 (2006), pp. 1179–1187.
- [24] B Wightman, T R Bürglin, J Gatto, P Arasu, and G Ruvkun. "Negative regulatory sequences in the *lin-14* 3'-untranslated region are necessary to generate a temporal switch during *Caenorhabditis elegans* development." In: *Genes Dev.* 5.10 (Oct. 1991), pp. 1813–1824.

- [25] Sébastien Pfeffer et al. "Identification of virus-encoded microRNAs." In: *Science* 304.5671 (2004), pp. 734–736.
- [26] Marc Robert Fabian, Nahum Sonenberg, and Witold Filipowicz. "Regulation of mRNA translation and stability by microRNAs." In: *Annu. Rev. Biochem.* 79 (Jan. 2010), pp. 351–379.
- [27] H-W Hwang and J T Mendell. "MicroRNAs in cell proliferation, cell death, and tumorigenesis." In: *Br. J. Cancer* 96 Suppl (2007), R40–4.
- [28] Kathryn N Ivey and Deepak Srivastava. "MicroRNAs as regulators of differentiation and cell fate decisions." In: *Cell Stem Cell* 7.1 (2010), pp. 36–41.
- [29] M Jovanovic and M O Hengartner. "miRNAs and apoptosis: RNAs to die for." In: *Oncogene* 25.46 (2006), pp. 6176–6187.
- [30] Ramiro Garzon, George a Calin, and Carlo M Croce. "MicroRNAs in Cancer." In: *Annu. Rev. Med.* 60 (Jan. 2009), pp. 167–179.
- [31] Victor Ambros. "MicroRNAs and developmental timing." In: *Curr. Opin. Genet. Dev.* 21.4 (Aug. 2011), pp. 511–517.
- [32] N J Beveridge, E Gardiner, A P Carroll, P A Tooney, and M J Cairns. "Schizophrenia is associated with an increase in cortical microRNA biogenesis." In: *Mol. Psychiatry* 15.12 (Dec. 2010), pp. 1176–1189.
- [33] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. "Most mammalian mRNAs are conserved targets of microRNAs." In: *Genome Res.* 19.1 (Jan. 2009), pp. 92–105.
- [34] David P Bartel. "MicroRNAs: target recognition and regulatory functions." In: *Cell* 136.2 (2009), pp. 215–233.
- [35] Jean Hausser and Mihaela Zavolan. "Identification and consequences of miRNA-target interactions [mdash] beyond repression of gene expression." In: *Nat. Rev. Genet.* 15.9 (2014), pp. 599–612.
- [36] Antony Rodriguez, Sam Griffiths-Jones, Jennifer L Ashurst, and Allan Bradley. "Identification of mammalian microRNA host genes and transcription units." In: *Genome Res.* 14.10A (Oct. 2004), pp. 1902–1910.
- [37] J Graham Ruby, Calvin H Jan, and David P Bartel. "Intronic microRNA precursors that bypass Drosha processing." In: *Nature* 448.7149 (2007), pp. 83–86.
- [38] Keita Miyoshi, Tomohiro Miyoshi, and Haruhiko Siomi. "Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production." In: *Mol. Genet. Genomics* 284.2 (Aug. 2010), pp. 95–103.

- [39] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. "MicroRNA genes are transcribed by RNA polymerase II." In: *EMBO J.* 23.20 (2004), pp. 4051–4060.
- [40] Yoontae Lee et al. "The nuclear RNase III Droscha initiates microRNA processing." In: *Nature* 425.6956 (2003), pp. 415–419.
- [41] Kyu-Hyeon Yeom, Yoontae Lee, Jinju Han, Mi Ra Suh, and V Narry Kim. "Characterization of DGCR8/Pasha, the essential cofactor for Droscha in primary miRNA processing." In: *Nucleic Acids Res.* 34.16 (2006), pp. 4622–4629.
- [42] Ahmet M Denli, Bastiaan B J Tops, Ronald H A Plasterk, René F Ketting, and Gregory J Hannon. "Processing of primary microRNAs by the Microprocessor complex." In: *Nature* 432.7014 (2004), pp. 231–235.
- [43] Rui Yi, Yi Qin, Ian G Macara, and Bryan R Cullen. "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs." In: *Genes Dev.* 17.24 (2003), pp. 3011–3016.
- [44] Markus T Bohnsack, Kevin Czaplinski, and Dirk Gorlich. "Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs." In: *RNA* 10.2 (Feb. 2004), pp. 185–191.
- [45] Elsebet Lund, Stephan Güttinger, Angelo Calado, James E Dahlberg, and Ulrike Kutay. "Nuclear export of microRNA precursors." In: *Science* 303.5654 (2004), pp. 95–98.
- [46] Xinhua Ji. "The mechanism of RNase III action: how dicer dices." In: *Curr. Top. Microbiol. Immunol.* 320 (2008), pp. 99–116.
- [47] Joerg E Braun, Eric Huntzinger, Maria Fauser, and Elisa Izauralde. "GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets." In: *Mol. Cell* 44.1 (2011), pp. 120–133.
- [48] Ramesh S Pillai, Suvendra N Bhattacharyya, Caroline G Artus, Tabea Zoller, Nicolas Cougot, Eugenia Basyuk, Edouard Bertrand, and Witold Filipowicz. "Inhibition of translational initiation by Let-7 MicroRNA in human cells." In: *Science* 309.5740 (2005), pp. 1573–1576.
- [49] Ashwini Jeggari, Debora S Marks, and Erik Larsson. "miRcode: a map of putative microRNA target sites in the long non-coding transcriptome." In: *Bioinformatics* 28.15 (2012), pp. 2062–2063.
- [50] Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. "A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?" In: *Cell* 146.3 (2011), pp. 353–358.

- [51] Marcella Cesana, Davide Cacchiarelli, Ivano Legnini, Tiziana Santini, Olga Sthandier, Mauro Chinappi, Anna Tramontano, and Irene Bozzoni. "A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA." In: *Cell* 147.2 (2011), pp. 358–369.
- [52] José Manuel Franco-Zorrilla, Adrián Valli, Marco Todesco, Isabel Mateos, María Isabel Puga, Ignacio Rubio-Somoza, Antonio Leyva, Detlef Weigel, Juan Antonio García, and Javier Paz-Ares. "Target mimicry provides a new mechanism for regulation of microRNA activity." In: *Nat. Genet.* 39.8 (Aug. 2007), pp. 1033–1037.
- [53] Gunter Meister. "Argonaute proteins: functional insights and emerging roles." In: *Nat. Rev. Genet.* 14.7 (July 2013), pp. 447–459.
- [54] Benjamin P Lewis, I-Hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. "Prediction of mammalian microRNA targets." In: *Cell* 115.7 (Dec. 2003), pp. 787–798.
- [55] Eric C Lai. "Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation." In: *Nat. Genet.* 30.4 (Apr. 2002), pp. 363–364.
- [56] Stanley D Chandradoss, Nicole T Schirle, Malwina Szczepaniak, Ian J MacRae, and Chirlmin Joo. "A Dynamic Search Process Underlies MicroRNA Targeting." In: *Cell* 162.1 (2015), pp. 96–107.
- [57] Nicole T Schirle and Ian J MacRae. "The crystal structure of human Argonaute2." In: *Science* 336.6084 (2012), pp. 1037–1040.
- [58] Sung Wook Chi, Gregory J Hannon, and Robert B Darnell. "An alternative mode of microRNA target recognition." In: *Nat. Struct. Mol. Biol.* 19.3 (2012), pp. 321–327.
- [59] Mohsen Khorshid, Jean Hausser, Mihaela Zavolan, and Erik van Nimwegen. "A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets." In: *Nat. Methods* 10.3 (2013), pp. 253–255.
- [60] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. "Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding." In: *Cell* 153.3 (Apr. 2013), pp. 654–665.
- [61] Julius Brennecke, Alexander Stark, Robert B Russell, and Stephen M Cohen. "Principles of microRNA-target recognition." In: *PLoS Biol.* 3.3 (Mar. 2005), e85.

- [62] B J Reinhart, F J Slack, M Basson, A E Pasquinelli, J C Bettinger, A E Rougvie, H R Horvitz, and G Ruvkun. "The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*." In: *Nature* 403.6772 (2000), pp. 901–906.
- [63] Matyas Ecsedi, Magdalene Rausch, and Helge Großhans. "The let-7 microRNA directs vulval development through a single target." In: *Dev. Cell* 32.3 (2015), pp. 335–344.
- [64] Daniel W Thomson, Cameron P Bracken, and Gregory J Goodall. "Experimental strategies for microRNA target identification." In: *Nucleic Acids Res.* 39.16 (2011), pp. 6845–6853.
- [65] Matthias Selbach, Björn Schwanhäusser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. "Widespread changes in protein synthesis induced by microRNAs." In: *Nature* 455.7209 (2008), pp. 58–63.
- [66] Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B Darnell. "CLIP identifies Nova-regulated RNA networks in the brain." In: *Science* 302.5648 (2003), pp. 1212–1215.
- [67] Chaolin Zhang and Robert B Darnell. "Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data." In: *Nat. Biotechnol.* 29.7 (July 2011), pp. 607–614.
- [68] Markus Hafner et al. "PAR-CLIP—a method to identify transcriptome-wide the binding sites of RNA binding proteins." In: *J. Vis. Exp.* 41 (Jan. 2010), pp. 2–6.
- [69] Stefanie Grosswendt, Andrei Filipchuk, Mark Manzano, Filippos Klironomos, Marcel Schilling, Margareta Herzog, Eva Gottwein, and Nikolaus Rajewsky. "Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions." In: *Mol. Cell* (May 2014), pp. 1–13.
- [70] Alexander Stark, Julius Brennecke, Robert B Russell, and Stephen M Cohen. "Identification of *Drosophila* MicroRNA targets." In: *PLoS Biol.* 1.3 (Dec. 2003), E60.
- [71] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. "Predicting effective microRNA target sites in mammalian mRNAs." In: *Elife* 4 (2015).
- [72] Dominic Didiano and Oliver Hobert. "Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions." In: *Nat. Struct. Mol. Biol.* 13.9 (Sept. 2006), pp. 849–851.
- [73] Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. "Inference of miRNA targets using evolutionary conservation and pathway analysis." In: *BMC Bioinformatics* 8 (2007), p. 69.

- [74] Andrew Grimson, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett-Engele, Lee P Lim, and David P Bartel. "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." In: *Mol. Cell* 27.1 (2007), pp. 91–105.
- [75] David M Garcia, Daehyun Baek, Chanseok Shin, George W Bell, Andrew Grimson, and David P Bartel. "Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lsy-6* and other microRNAs." In: *Nat. Struct. Mol. Biol.* 18.10 (Oct. 2011), pp. 1139–1146.
- [76] Jean Hausser, Markus Landthaler, Lukasz Jaskiewicz, Dimos Gaidatzis, and Mihaela Zavolan. "Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets." In: *Genome Res.* 19.11 (2009), pp. 2009–2020.
- [77] Liang Meng Wee, C Fabián Flores-Jasso, William E Salomon, and Phillip D Zamore. "Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties." In: *Cell* 151.5 (Nov. 2012), pp. 1055–1067.
- [78] S M Hammond, E Bernstein, D Beach, and G J Hannon. "An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells." In: *Nature* 404.6775 (2000), pp. 293–296.
- [79] Nicholas J Watkins and Markus T Bohnsack. "The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA." In: *Wiley Interdiscip. Rev. RNA* 3.3 (May 2012), pp. 397–414.
- [80] P Ganot, B E Jády, M L Bortolin, X Darzacq, and T Kiss. "Nucleolar factors direct the 2'-O-ribose methylation and pseudouridylation of U6 spliceosomal RNA." In: *Mol. Cell. Biol.* 19.10 (Oct. 1999), pp. 6906–6917.
- [81] S Kass, K Tyc, J A Steitz, and B Sollner-Webb. "The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing." In: *Cell* 60.6 (1990), pp. 897–908.
- [82] K Tyc and J A Steitz. "U3, U8 and U13 comprise a new class of mammalian snRNPs localized in the cell nucleolus." In: *EMBO J.* 8.10 (Oct. 1989), pp. 3113–3119.
- [83] P Ganot, M L Bortolin, and T Kiss. "Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs." In: *Cell* 89.5 (1997), pp. 799–809.
- [84] J Ni, A L Tien, and M J Fournier. "Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA." In: *Cell* 89.4 (1997), pp. 565–573.

- [85] A G Balakin, L Smith, and M J Fournier. "The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions." In: *Cell* 86.5 (1996), pp. 823–834.
- [86] Xavier Darzacq, Beáta E Jády, Céline Verheggen, Arnold M Kiss, Edouard Bertrand, and Tamás Kiss. "Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs." In: *EMBO J.* 21.11 (2002), pp. 2746–2756.
- [87] Patricia Richard, Xavier Darzacq, Edouard Bertrand, Beáta E Jády, Céline Verheggen, and Tamás Kiss. "A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs." In: *EMBO J.* 22.16 (2003), pp. 4283–4293.
- [88] Z Kiss-László, Y Henry, and T Kiss. "Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA." In: *EMBO J.* 17.3 (1998), pp. 797–807.
- [89] Steve L Reichow, Tomoko Hamma, Adrian R Ferré-D'Amaré, and Gabriele Varani. "The structure and function of small nucleolar ribonucleoproteins." In: *Nucleic Acids Res.* 35.5 (2007), pp. 1452–1464.
- [90] H D Li, J Zagorski, and M J Fournier. "Depletion of U<sub>14</sub> small nuclear RNA (snR<sub>128</sub>) disrupts production of 18S rRNA in *Saccharomyces cerevisiae*." In: *Mol. Cell. Biol.* 10.3 (Mar. 1990), pp. 1145–1152.
- [91] Laurent Lestrade and Michel J Weber. "snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs." In: *Nucleic Acids Res.* 34.Database issue (Jan. 2006), pp. D158–62.
- [92] Robert Willem van Nues, Sander Granneman, Grzegorz Kudla, Katherine Elizabeth Sloan, Matthew Chicken, David Tollervey, and Nicholas James Watkins. "Box C/D snoRNP catalysed methylation is aided by additional pre-rRNA base-pairing." In: *EMBO J.* 30.12 (2011), pp. 2420–2430.
- [93] J Cavailé, M Nicoloso, and J P Bachellerie. "Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides." In: *Nature* 383.6602 (1996), pp. 732–735.
- [94] Gene W Yeo, Nicole G Coufal, Tiffany Y Liang, Grace E Peng, Xiang-Dong Fu, and Fred H Gage. "An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells." In: *Nat. Struct. Mol. Biol.* 16.2 (Feb. 2009), pp. 130–137.

- [95] Jana Hertel, Ivo L Hofacker, and Peter F Stadler. "SnoReport: computational identification of snoRNAs with unknown targets." In: *Bioinformatics* 24.2 (2008), pp. 158–164.
- [96] Patrice Vitali, H el ene Royo, Herv e Seitz, Jean-Pierre Bachelier, Alexander H uttenhofer, and J er ome Cavall e. "Identification of 13 novel human modification guide RNAs." In: *Nucleic Acids Res.* 31.22 (2003), pp. 6543–6551.
- [97] Zurab Siprashvili et al. "The noncoding RNAs SNORD50A and SNORD50B bind K-Ras and are recurrently deleted in human cancer." In: *Nat. Genet.* (2015).
- [98] Shivendra Kishore, Amit Khanna, Zhaiyi Zhang, Jingyi Hui, Piotr J Balwiercz, Mihaela Stefan, Carol Beach, Robert D Nicholls, Mihaela Zavolan, and Stefan Stamm. "The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing." In: *Hum. Mol. Genet.* 19.7 (Apr. 2010), pp. 1153–1164.
- [99] Qing-Fei Yin, Li Yang, Yang Zhang, Jian-Feng Xiang, Yue-Wei Wu, Gordon G Carmichael, and Ling-Ling Chen. "Long non-coding RNAs with snoRNA ends." In: *Mol. Cell* 48.2 (2012), pp. 219–230.
- [100] Tanmoy Mondal, Markus Rasmussen, Gaurav Kumar Pandey, Anders Isaksson, and Chandrasekhar Kanduri. "Characterization of the RNA content of chromatin." In: *Genome Res.* 20.7 (July 2010), pp. 899–907.
- [101] Carlos I Michel, Christopher L Holley, Benjamin S Scruggs, Rohini Sidhu, Rita T Brookheart, Laura L Listenberger, Mark A Behlke, Daniel S Ory, and Jean E Schaffer. "Small nucleolar RNAs U32a, U33, and U35a are critical mediators of metabolic stress." In: *Cell Metab.* 14.1 (2011), pp. 33–44.
- [102] Kaiissar Mannoor, Jipei Liao, and Feng Jiang. "Small nucleolar RNAs in cancer." In: *Biochim. Biophys. Acta* 1826.1 (Aug. 2012), pp. 121–128.
- [103] Peter S Bazeley, Valery Shepelev, Zohreh Talebizadeh, Merlin G Butler, Larisa Fedorova, Vadim Filatov, and Alexei Fedorov. "snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions." In: *Gene* 408.1–2 (2008), pp. 172–179.
- [104] Shifeng Xue and Maria Barna. "Specialized ribosomes: a new frontier in gene regulation and organismal biology." In: *Nat. Rev. Mol. Cell Biol.* 13.6 (June 2012), pp. 355–369.
- [105] Boris Rogelj. "Brain-specific small nucleolar RNAs." In: *J. Mol. Neurosci.* 28.2 (2006), pp. 103–109.

- [106] J Cavallé, K Buiting, M Kiefmann, M Lalande, C I Brannan, B Horsthemke, J P Bachelierie, J Brosius, and A Hüttenhofer. "Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization." In: *Proc. Natl. Acad. Sci. U. S. A.* 97.26 (2000), pp. 14311–14316.
- [107] Boris V Skryabin, Leonid V Gubar, Birte Seeger, Jana Pfeiffer, Sergej Handel, Thomas Robeck, Elena Karpova, Timofey S Rozhdestvensky, and Jürgen Brosius. "Deletion of the MBII-85 snoRNA gene cluster in mice results in postnatal growth retardation." In: *PLoS Genet.* 3.12 (2007), e235.
- [108] Trilochan Sahoo, Daniela del Gaudio, Jennifer R German, Marwan Shinawi, Sarika U Peters, Richard E Person, Adolfo Garnica, Sau Wai Cheung, and Arthur L Beaudet. "Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster." In: *Nat. Genet.* 40.6 (June 2008), pp. 719–721.
- [109] Feng Ding, Yelena Prints, Madhu S Dhar, Dabney K Johnson, Carmen Garnacho-Montero, Robert D Nicholls, and Uta Francke. "Lack of Pwcr1/MBII-85 snoRNA is critical for neonatal lethality in Prader-Willi syndrome mouse models." In: *Mamm. Genome* 16.6 (June 2005), pp. 424–431.
- [110] Adam J de Smith et al. "A deletion of the HBII-85 class of small nucleolar RNAs (snoRNAs) is associated with hyperphagia, obesity and hypogonadism." In: *Hum. Mol. Genet.* 18.17 (2009), pp. 3257–3265.
- [111] Rebecca M Voorhees and V Ramakrishnan. "Structural basis of the translational elongation cycle." In: *Annu. Rev. Biochem.* 82 (2013), pp. 203–236.
- [112] Denis L J Lafontaine. "Noncoding RNAs in eukaryotic ribosome biogenesis and function." In: *Nat. Struct. Mol. Biol.* 22.1 (2015), pp. 11–19.
- [113] Vera Atzorn, Paola Fragapane, and Tamás Kiss. "U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for 18S rRNA production." In: *Mol. Cell. Biol.* 24.4 (Feb. 2004), pp. 1769–1778.
- [114] Manja Marz and Peter F Stadler. "Comparative analysis of eukaryotic U3 snoRNA." In: *RNA Biol.* 6.5 (Nov. 2009), pp. 503–507.
- [115] Darrell R Davis. "Biophysical and Conformational Properties of Modified Nucleosides in RNA (Nuclear Magnetic Resonance Studies)." In: *Modification and Editing of RNA*. Ed. by Henri Grosjean and Rob Benne. American Society of Microbiology, 1998, pp. 85–102.

- [116] Ulf Birkedal, Mikkel Christensen-Dalsgaard, Nicolai Krogh, Radhakrishnan Sabarinathan, Jan Gorodkin, and Henrik Nielsen. "Profiling of Ribose Methylations in RNA by High-Throughput Sequencing." In: *Angew. Chem. Int. Ed Engl.* (2014).
- [117] D J Williams, J L Boots, and K B Hall. "Thermodynamics of 2'-ribose substitutions in UUCG tetraloops." In: *RNA* 7.1 (Jan. 2001), pp. 44–53.
- [118] Sunny Sharma and Denis L J Lafontaine. "'View From A Bridge': A New Perspective on Eukaryotic rRNA Base Modification." In: *Trends Biochem. Sci.* 40.10 (Oct. 2015), pp. 560–575.
- [119] Mary McMahon, Adrian Contreras, and Davide Ruggiero. "Small RNAs with big implications: new insights into H/ACA snoRNA function and their role in human disease." In: *Wiley Interdiscip. Rev. RNA* (2014).
- [120] Eric Huntzinger and Elisa Izaurralde. "Gene silencing by microRNAs: contributions of translational repression and mRNA decay." In: *Nat. Rev. Genet.* 12.2 (Feb. 2011), pp. 99–110.
- [121] Ramesh A Shivdasani. "MicroRNAs: regulators of gene expression and cell differentiation." In: *Blood* 108.12 (2006), pp. 3646–3653.
- [122] George A Calin and Carlo M Croce. "MicroRNA signatures in human cancers." In: *Nat. Rev. Cancer* 6.11 (Nov. 2006), pp. 857–866.
- [123] Nikolaus Rajewsky and Nicholas D Socci. "Computational identification of microRNA targets." In: *Dev. Biol.* 267.2 (2004), pp. 529–535.
- [124] Benjamin P Lewis, Christopher B Burge, and David P Bartel. "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." In: *Cell* 120.1 (Jan. 2005), pp. 15–20.
- [125] Amy E Pasquinelli. "MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship." In: *Nat. Rev. Genet.* 13.4 (Apr. 2012), pp. 271–282.
- [126] Sayda M Elbashir, Jens Harborth, Klaus Weber, and Thomas Tuschl. "Analysis of gene function in somatic mammalian cells using small interfering RNAs." In: *Methods* 26.2 (Feb. 2002), pp. 199–213.
- [127] Stijn van Dongen, Cei Abreu-Goodger, and Anton J Enright. "Detecting microRNA binding and siRNA off-target effects from expression data." In: *Nat. Methods* 5.12 (2008), pp. 1023–1025.

- [128] Bahar Yilmazel, Yanhui Hu, Frederic Sigoillot, Jennifer A Smith, Caroline E Shamu, Norbert Perrimon, and Stephanie E Mohr. "Online GESS: prediction of miRNA-like off-target effects in large-scale RNAi screen data by seed region analysis." In: *BMC Bioinformatics* 15 (2014), p. 192.
- [129] Shaoli Das, Suman Ghosal, Jayprokas Chakrabarti, and Karol Kozak. "SeedSeq: off-target transcriptome database." In: *Biomed Res. Int.* 2013 (2013), p. 905429.
- [130] Maria D Paraskevopoulou, Georgios Georgakilas, Nikos Kostoulas, Ioannis S Vlachos, Thanasis Vergoulis, Martin Reczko, Christos Filippidis, Theodore Dalamagas, and A G Hatzigeorgiou. "DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows." In: *Nucleic Acids Res.* 41.Web Server issue (July 2013), W169–73.
- [131] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. "Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites." In: *Genome Biol.* 11.8 (2010), R90.
- [132] Nikolaus Schultz, Dina R Marenstein, Dino A De Angelis, Wei-Qing Wang, Sven Nelander, Anders Jacobsen, Debora S Marks, Joan Massagué, and Chris Sander. "Off-target effects dominate a large-scale RNAi screen for modulators of the TGF- $\beta$  pathway and reveal microRNA regulation of TGFBR2." In: *Silence* 2 (2011), p. 3.
- [133] William H Majoros and Uwe Ohler. "Spatial preferences of microRNA targets in 3' untranslated regions." In: *BMC Genomics* 8 (2007), p. 152.
- [134] R S Hudson et al. "MicroRNA-106b-25 cluster expression is associated with early disease recurrence and targets caspase-7 and focal adhesion in human prostate cancer." In: *Oncogene* 32.35 (2012), pp. 4139–4147.
- [135] Neetu Dahiya, Cheryl A Sherman-Baust, Tian-Li Wang, Ben Davidson, Ie-Ming Shih, Yongqing Zhang, William Wood 3rd, Kevin G Becker, and Patrice J Morin. "MicroRNA expression and identification of putative miRNA targets in ovarian cancer." In: *PLoS One* 3.6 (2008), e2436.
- [136] Lisa B Frankel, Jiayu Wen, Michael Lees, Maria Høyer-Hansen, Thomas Farkas, Anders Krogh, Marja Jäätelä, and Anders H Lund. "microRNA-101 is a potent inhibitor of autophagy." In: *EMBO J.* 30.22 (2011), pp. 4628–4641.
- [137] Vincenzo Alessandro Gennarino, Marco Sardiello, Raffaella Avelino, Nicola Meola, Vincenza Maselli, Santosh Anand, Luisa Cutillo, Andrea Ballabio, and Sandro Banfi. "MicroRNA target

- prediction by expression analysis of host genes." In: *Genome Res.* 19.3 (2008), pp. 481–490.
- [138] S-K Leivonen et al. "Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines." In: *Oncogene* 28.44 (2009), pp. 3926–3936.
- [139] Peter S Linsley et al. "Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression." In: *Mol. Cell. Biol.* 27.6 (2007), pp. 2240–2252.
- [140] Ana Kozomara and Sam Griffiths-Jones. "miRBase: integrating microRNA annotation and deep-sequencing data." In: *Nucleic Acids Res.* 39.Database issue (Jan. 2011), pp. D152–7.
- [141] S M Elbashir, J Martinez, A Patkaniowska, W Lendeckel, and T Tuschl. "Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate." In: *EMBO J.* 20.23 (2001), pp. 6877–6888.
- [142] Pål Sætrom, Bret S E Heale, Ola Snøve, Lars Aagaard, Jessica Alluin, and John J Rossi. "Distance constraints between microRNA target sites dictate efficacy and cooperativity." In: *Nucleic Acids Res.* 35.7 (2007), pp. 2333–2342.
- [143] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. "CONTRAFold: RNA secondary structure prediction without physics-based models." In: *Bioinformatics* 22.14 (July 2006), e90–8.
- [144] Ivo L Hofacker. "Vienna RNA secondary structure server." In: *Nucleic Acids Res.* 31.13 (2003), pp. 3429–3431.
- [145] Pouya Kheradpour, Alexander Stark, Sushmita Roy, and Manolis Kellis. "Reliable prediction of regulator targets using 12 *Drosophila* genomes." In: *Genome Res.* 17.12 (Dec. 2007), pp. 1919–1931.
- [146] Thomas D Wu and Colin K Watanabe. "GMAP: a genomic mapping and alignment program for mRNA and EST sequences." In: *Bioinformatics* 21.9 (2005), pp. 1859–1875.
- [147] Jeet Sukumaran and Mark T Holder. "DendroPy: a Python library for phylogenetic computing." In: *Bioinformatics* 26.12 (2010), pp. 1569–1571.
- [148] S Seabold and J Perktold. "Statsmodels: Econometric and statistical modeling with python." In: *of the 9th Python in Science Conference* (2010).
- [149] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." In: *Nat. Protoc.* 4.1 (2009), pp. 44–57.

- [150] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." In: *Nucleic Acids Res.* 37.1 (Jan. 2009), pp. 1–13.
- [151] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. "The role of site accessibility in microRNA target recognition." In: *Nat. Genet.* 39.10 (2007), pp. 1278–1284.
- [152] Ulrike Mückstein, Hakim Tafer, Jörg Hackermüller, Stephan H Bernhart, Peter F Stadler, and Ivo L Hofacker. "Thermodynamics of RNA-RNA binding." In: *Bioinformatics* 22.10 (2006), pp. 1177–1182.
- [153] Panagiotis Alexiou, Manolis Maragkakis, Giorgos L Papadopoulos, Martin Reczko, and Artemis G Hatzigeorgiou. "Lost in translation: an assessment and perspective for computational microRNA target identification." In: *Bioinformatics* 25.23 (Dec. 2009), pp. 3049–3055.
- [154] Sonia Sharma, Ariel Quintana, Gregory M Findlay, Marcel Mettlen, Beate Baust, Mohit Jain, Roland Nilsson, Anjana Rao, and Patrick G Hogan. "An siRNA screen for NFAT activation identifies septins as coordinators of store-operated Ca<sup>2+</sup> entry." In: *Nature* 499.7457 (2013), pp. 238–242.
- [155] Honglin Zhou et al. "Genome-scale RNAi screen for host factors required for HIV replication." In: *Cell Host Microbe* 4.5 (2008), pp. 495–504.
- [156] Dimitri Moreau, Pankaj Kumar, Shyi Chyi Wang, Alexandre Chaumet, Shin Yi Chew, Hélène Chevalley, and Frédéric Bard. "Genome-wide RNAi screens identify genes required for Ricin and PE intoxications." In: *Dev. Cell* 21.2 (2011), pp. 231–244.
- [157] M Kanehisa and S Goto. "KEGG: kyoto encyclopedia of genes and genomes." In: *Nucleic Acids Res.* 28.1 (2000), pp. 27–30.
- [158] Leo Goodstadt. "Ruffus: a lightweight Python library for computational pipelines." In: *Bioinformatics* 26.21 (2010), pp. 2778–2779.
- [159] Jean Hausser, Afzal Pasha Syed, Biter Bilen, and Mihaela Zavolan. "Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation." In: *Genome Res.* 23.4 (2013), pp. 604–615.
- [160] Jin-Wu Nam, Olivia S Rissland, David Koppstein, Cei Abreu-Goodger, Calvin H Jan, Vikram Agarwal, Muhammed a Yildirim, Antony Rodriguez, and David P Bartel. "Global analyses of the effect of different cellular contexts on microRNA targeting." In: *Mol. Cell* 53.6 (Mar. 2014), pp. 1031–1043.

- [161] Suvendra N Bhattacharyya, Regula Habermacher, Ursula Martine, Ellen I Closs, and Witold Filipowicz. "Relief of microRNA-mediated translational repression in human cells subjected to stress." In: *Cell* 125.6 (2006), pp. 1111–1124.
- [162] Piotr J Balwiercz, Mikhail Pachkov, Phil Arnold, Andreas J Gruber, Mihaela Zavolan, and Erik van Nimwegen. "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs." In: *Genome Res.* 24.5 (May 2014), pp. 869–884.
- [163] Stephen W Eichhorn, Huili Guo, Sean E McGeary, Ricard A Rodriguez-Mias, Chanseok Shin, Daehyun Baek, Shu-Hao Hsu, Kalpana Ghoshal, Judit Villén, and David P Bartel. "mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues." In: *Mol. Cell* 56.1 (2014), pp. 104–115.
- [164] Erel Levine, Zhongge Zhang, Thomas Kuhlman, and Terence Hwa. "Quantitative characteristics of gene regulation by small RNA." In: *PLoS Biol.* 5.9 (Sept. 2007), e229.
- [165] Shankar Mukherji, Margaret S Ebert, Grace X Y Zheng, John S Tsang, Phillip A Sharp, and Alexander van Oudenaarden. "MicroRNAs can generate thresholds in target gene expression." In: *Nat. Genet.* 43.9 (Sept. 2011), pp. 854–859.
- [166] Nicolas E Buchler and Matthieu Louis. "Molecular titration and ultrasensitivity in regulatory networks." In: *J. Mol. Biol.* 384.5 (2008), pp. 1106–1119.
- [167] Aly A Khan, Doron Betel, Martin L Miller, Chris Sander, Christina S Leslie, and Debora S Marks. "Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs." In: *Nat. Biotechnol.* 27.6 (June 2009), pp. 549–555.
- [168] Andrew D Bosson, Jesse R Zamudio, and Phillip A Sharp. "Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition." In: *Mol. Cell* 56.3 (2014), pp. 347–359.
- [169] Matteo Figliuzzi, Enzo Marinari, and Andrea De Martino. "MicroRNAs as a selective channel of communication between competing RNAs: a steady-state theory." In: *Biophys. J.* 104.5 (2013), pp. 1203–1213.
- [170] Carla Bosia, Andrea Pagnani, and Riccardo Zecchina. "Modelling Competing Endogenous RNA Networks." In: *PLoS One* 8.6 (2013), e66609.
- [171] Ute Bissels, Stefan Wild, Stefan Tomiuk, Angela Holste, Markus Hafner, Thomas Tuschl, and Andreas Bosio. "Absolute quantification of microRNAs by using a universal reference." In: *RNA* 15.12 (Dec. 2009), pp. 2375–2384.

- [172] Jinhong Chang et al. "miR-122, a Mammalian Liver-Specific microRNA, is Processed from hcr mRNA and May Downregulate the High Affinity Cationic Amino Acid Transporter CAT-1." In: *RNA Biol.* 1.2 (2004), pp. 106–113.
- [173] Dongmei Wang, Zhaojie Zhang, Evan O'Loughlin, Thomas Lee, Stephane Houel, Dónal O'Carroll, Alexander Tarakhovsky, Natalie G Ahn, and Rui Yi. "Quantitative functions of Argonaute proteins in mammalian development." In: *Genes Dev.* 26.7 (2012), pp. 693–704.
- [174] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. "Validation of noise models for single-cell transcriptomics." In: *Nat. Methods* 11.6 (June 2014), pp. 637–640.
- [175] Sung Wook Chi, Julie B. Zang, Aldo Mele, and Robert B. Darnell. "Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps." In: *Nature* 460.7254 (2009), pp. 479–486.
- [176] Markus Hafner et al. "Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP." In: *Cell* 141.1 (Apr. 2010), pp. 129–141.
- [177] T Xia, J SantaLucia Jr, M E Burkard, R Kierzek, S J Schroeder, X Jiao, C Cox, and D H Turner. "Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs." In: *Biochemistry* 37.42 (1998), pp. 14719–14735.
- [178] Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid, and Mihaela Zavolan. "A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins." In: *Nat. Methods* 8.7 (July 2011), pp. 559–564.
- [179] Xiaowei Wang. "Composition of seed sequence is a major determinant of microRNA targeting patterns." In: *Bioinformatics* 30.10 (2014), pp. 1377–1383.
- [180] Jessica S Reuter and David H Mathews. "RNAstructure: software for RNA secondary structure prediction and analysis." In: *BMC Bioinformatics* 11 (2010), p. 129.
- [181] Shobha Vasudevan, Yingchun Tong, and Joan A Steitz. "Switching from repression to activation: microRNAs can up-regulate translation." In: *Science* 318.5858 (2007), pp. 1931–1934.
- [182] Virginie Olive et al. "A component of the mir-17-92 polycistronic oncomir promotes oncogene-dependent apoptosis." In: *Elife* 2 (2013), e00822.
- [183] Laura Poliseno, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J Haveman, and Pier Paolo Pandolfi. "A coding-independent function of gene and pseudogene mRNAs regulates tumour biology." In: *Nature* 465.7301 (2010), pp. 1033–1038.

- [184] Rémy Denzler, Vikram Agarwal, Joanna Stefano, David P Bartel, and Markus Stoffel. "Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance." In: *Mol. Cell* 54.5 (2014), pp. 766–776.
- [185] Dimpal Nyayanit and Chetan J Gadgil. "Mathematical modeling of combinatorial regulation suggests that apparent positive regulation of targets by miRNA could be an artifact resulting from competition for mRNA." In: *RNA* 21.3 (Mar. 2015), pp. 307–319.
- [186] Young-Kook Kim, Inha Heo, and V Narry Kim. "Modifications of small RNAs and their associated proteins." In: *Cell* 143.5 (2010), pp. 703–709.
- [187] Sheng Li and Christopher E Mason. "The pivotal regulatory landscape of RNA modifications." In: *Annu. Rev. Genomics Hum. Genet.* 15 (June 2014), pp. 127–150.
- [188] Dan Dominissini et al. "Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq." In: *Nature* 485.7397 (May 2012), pp. 201–206.
- [189] Schraga Schwartz et al. "Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA." In: *Cell* 159.1 (Sept. 2014), pp. 148–162.
- [190] Matthias Schaefer. "RNA 5-Methylcytosine Analysis by Bisulfite Sequencing." In: *Methods Enzymol.* 560 (May 2015), pp. 297–329.
- [191] Yogesh Saletore, Kate Meyer, Jonas Korlach, Igor D Vilfan, Samie Jaffrey, and Christopher E Mason. "The birth of the Epitranscriptome: deciphering the function of RNA modifications." In: *Genome Biol.* 13.10 (Oct. 2012), p. 175.
- [192] Mihye Lee, Boseon Kim, and V Narry Kim. "Emerging roles of RNA modification: m(6)A and U-tail." In: *Cell* 158.5 (Aug. 2014), pp. 980–987.
- [193] Wen-Ju Sun, Jun-Hao Li, Shun Liu, Jie Wu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. "RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data." In: *Nucleic Acids Res.* (Oct. 2015).
- [194] John Karijolich, Athena Kantartzis, and Yi-Tao Yu. "RNA modifications: a mechanism that modulates gene expression." In: *Methods Mol. Biol.* 629 (2010), pp. 1–19.
- [195] K L Heilman, R A Leach, and M T Tuck. "Internal 6-methyladenine residues increase the in vitro translation efficiency of dihydrofolate reductase messenger RNA." In: *Int. J. Biochem. Cell Biol.* 28.7 (July 1996), pp. 823–829.

- [196] M T Tuck, P E Wiehl, and T Pan. "Inhibition of 6-methyladenine formation decreases the translation efficiency of dihydrofolate reductase transcripts." In: *Int. J. Biochem. Cell Biol.* 31.8 (Aug. 1999), pp. 837–851.
- [197] Xiao Wang et al. "N6-methyladenosine-dependent regulation of messenger RNA stability." In: *Nature* 505.7481 (Jan. 2014), pp. 117–120.
- [198] Xiao Wang, Boxuan Simen Zhao, Ian A Roundtree, Zhike Lu, Dali Han, Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He. "N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency." In: *Cell* 161.6 (June 2015), pp. 1388–1399.
- [199] K T Tycowski, Z H You, P J Graham, and J A Steitz. "Modification of U6 spliceosomal RNA is guided by other small RNAs." In: *Mol. Cell* 2.5 (Nov. 1998), pp. 629–638.
- [200] P P Dennis, A Omer, and T Lowe. "A guided tour: small RNA function in Archaea." In: *Mol. Microbiol.* 40.3 (May 2001), pp. 509–519.
- [201] Anja Zemmann, Anja op de Bekke, Martin Kiefmann, Jürgen Brosius, and Jürgen Schmitz. "Evolution of small nucleolar RNAs in nematodes." In: *Nucleic Acids Res.* 34.9 (May 2006), pp. 2676–2685.
- [202] D Tollervey, H Lehtonen, R Jansen, H Kern, and E C Hurt. "Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly." In: *Cell* 72.3 (Feb. 1993), pp. 443–457.
- [203] Nicholas J Watkins, Ségault Véronique, Charpentier Bruno, Nottrott Stephanie, Fabrizio Patrizia, Bachi Angela, Wilm Matthias, Rosbash Michael, Branlant Christiane, and Lührmann Reinhard. "A Common Core RNP Structure Shared between the Small Nucleolar Box C/D RNPs and the Spliceosomal U4 snRNP." In: *Cell* 103.3 (2000), pp. 457–466.
- [204] T Gautier, T Bergès, D Tollervey, and E Hurt. "Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis." In: *Mol. Cell. Biol.* 17.12 (Dec. 1997), pp. 7088–7098.
- [205] Marc Quinternet, Quinternet Marc, Chagot Marie-Eve, Rothé Benjamin, Tiotiu Decebal, Charpentier Bruno, and Manival Xavier. "Structural Features of the Box C/D snoRNP Pre-assembly Process Are Conserved through Species." In: *Structure* 24.10 (2016), pp. 1693–1706.

- [206] B E Jady and T Kiss. "A small nucleolar guide RNA functions both in 2'-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA." en. In: *EMBO J.* 20.3 (Feb. 2001), pp. 541-551.
- [207] E Caffarelli, A Fatica, S Prislei, E De Gregorio, P Fragapane, and I Bozzoni. "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." In: *EMBO J.* 15.5 (Mar. 1996), pp. 1121-1131.
- [208] Robert Willem van Nues, Granneman Sander, Kudla Grzegorz, Katherine Elizabeth Sloan, Chicken Matthew, Tollervey David, and Nicholas James Watkins. "Box C/D snoRNP catalysed methylation is aided by additional pre-rRNA base-pairing." In: *EMBO J.* 30.12 (2011), pp. 2420-2430.
- [209] D Tollervey, H Lehtonen, M Carmo-Fonseca, and E C Hurt. "The small nucleolar RNP protein NOP1 (fibrillarin) is required for pre-rRNA processing in yeast." In: *EMBO J.* 10.3 (Mar. 1991), pp. 573-583.
- [210] D Tollervey and T Kiss. "Function and synthesis of small nucleolar RNAs." In: *Curr. Opin. Cell Biol.* 9.3 (June 1997), pp. 337-342.
- [211] D L Lafontaine and D Tollervey. "Birth of the snoRNPs: the evolution of the modification-guide snoRNAs." In: *Trends Biochem. Sci.* 23.10 (Oct. 1998), pp. 383-388.
- [212] Alessandro Fatica and David Tollervey. "Insights into the structure and function of a guide RNP." In: *Nat. Struct. Biol.* 10.4 (Apr. 2003), pp. 237-239.
- [213] Stephanie Kehr, Sebastian Bartschat, Peter F Stadler, and Hakim Tafer. "PLEXY: efficient target prediction for box C/D snoRNAs." In: *Bioinformatics* 27.2 (Jan. 2011), pp. 279-280.
- [214] Chun-Long Chen, Roland Perasso, Liang-Hu Qu, and Laurence Amar. "Exploration of Pairing Constraints Identifies a 9 Base-pair Core within Box C/D snoRNA-rRNA Duplexes." In: *J. Mol. Biol.* 369.3 (June 2007), pp. 771-783.
- [215] S Kass, K Tyc, J A Steitz, and B Sollner-Webb. "The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing." en. In: *Cell* 60.6 (Mar. 1990), pp. 897-908.
- [216] B A Peculis and J A Steitz. "Disruption of U8 nucleolar snRNA inhibits 5.8S and 28S rRNA processing in the *Xenopus* oocyte." en. In: *Cell* 73.6 (June 1993), pp. 1233-1245.

- [217] J Cavail , A A Hadjiolov, and J P Bachellerie. "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." en. In: *Eur. J. Biochem.* 242.2 (Dec. 1996), pp. 206–213.
- [218] Grzegorz Kudla, Sander Granneman, Daniela Hahn, Jean D Beggs, and David Tollervey. "Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast." In: *Proc. Natl. Acad. Sci. U. S. A.* 108.24 (June 2011), pp. 10010–10015.
- [219] B E Maden. "Mapping 2'-O-methyl groups in ribosomal RNA." In: *Methods* 25.3 (Nov. 2001), pp. 374–382.
- [220] Nicolai Krogh, Martin D Jansson, Sophia J H fner, Disa Tehler, Ulf Birkedal, Mikkel Christensen-Dalsgaard, Anders H Lund, and Henrik Nielsen. "Profiling of 2'-O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity." en. In: *Nucleic Acids Res.* (June 2016).
- [221] Be ta E J dy, Amandine Ketele, and Tam s Kiss. "Human intron-encoded Alu RNAs are processed and packaged into Wdr79-associated nucleoplasmic box H/ACA RNPs." In: *Genes Dev.* 26.17 (Sept. 2012), pp. 1897–1910.
- [222] Shivendra Kishore, Andreas R Gruber, Dominik J Jedlinski, Afzal P Syed, Hadi Jorjani, and Mihaela Zavolan. "Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing." In: *Genome Biol.* 14.5 (Sept. 2013), R45.
- [223] Martin Machyna, Stephanie Kehr, Korinna Straube, Dennis Kappei, Frank Buchholz, Falk Butter, Jernej Ule, Jana Hertel, Peter F Stadler, and Karla M Neugebauer. "The coilin interactome identifies hundreds of small noncoding RNAs that traffic through Cajal bodies." In: *Mol. Cell* 56.3 (Nov. 2014), pp. 389–399.
- [224] Hadi Jorjani, Stephanie Kehr, Dominik J Jedlinski, Rafal Gumieny, Jana Hertel, Peter F Stadler, Mihaela Zavolan, and Andreas R Gruber. "An updated human snoRNAome." In: *Nucleic Acids Res.* (May 2016).
- [225] Georges Martin, Andreas R Gruber, Walter Keller, and Mihaela Zavolan. "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." In: *Cell Rep.* 1.6 (June 2012), pp. 753–763.

- [226] Patricia P Chan and Todd M Lowe. "GtRNAdb: a database of transfer RNA genes detected in genomic sequence." In: *Nucleic Acids Res.* 37.Database issue (Jan. 2009), pp. D93–7.
- [227] Mohsen Khorshid, Christoph Rodak, and Mihaela Zavolan. "CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins." In: *Nucleic Acids Res.* 39.Database issue (Jan. 2011), pp. D245–52.
- [228] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2." en. In: *Nat. Methods* 9.4 (Apr. 2012), pp. 357–359.
- [229] H Tafer and I L Hofacker. "RNAplex: a fast tool for RNA–RNA interaction search." In: *Bioinformatics* (2008).
- [230] Fiona Cunningham et al. "Ensembl 2015." In: *Nucleic Acids Res.* 43.Database issue (Jan. 2015), pp. D662–9.
- [231] Kate R Rosenbloom et al. "The UCSC Genome Browser database: 2015 update." In: *Nucleic Acids Res.* 43.Database issue (Jan. 2015), pp. D670–81.
- [232] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads." In: *EMBnet. journal* 17.1 (2011), pp. –10.
- [233] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. "STAR: ultrafast universal RNA-seq aligner." In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21.
- [234] B E Maden, M E Corbett, P A Heeney, K Pugh, and P M Ajuh. "Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA." In: *Biochimie* 77.1-2 (1995), pp. 22–29.
- [235] Zhi-Wei Dong, Peng Shao, Li-Ting Diao, Hui Zhou, Chun-Hong Yu, and Liang-Hu Qu. "RTL-P: a sensitive approach for detecting sites of 2'-O-methylation in RNA molecules." In: *Nucleic Acids Res.* 40.20 (Nov. 2012), e157.
- [236] Thomas Emil Andersen, Bo Torben Porse, and Finn Kirpekar. "A novel partial modification at C2501 in Escherichia coli 23S ribosomal RNA." en. In: *RNA* 10.6 (June 2004), pp. 907–913.
- [237] Manuel Bauer, Erik Ahrné, Anna P Baron, Timo Glatter, Luca L Fava, Anna Santamaria, Erich A Nigg, and Alexander Schmidt. "Evaluation of data-dependent and -independent mass spectrometric workflows for sensitive quantification of proteins and phosphorylation sites." en. In: *J. Proteome Res.* 13.12 (Dec. 2014), pp. 5973–5988.

- [238] B E Maden. "Identification of the locations of the methyl groups in 18 S ribosomal RNA from *Xenopus laevis* and man." en. In: *J. Mol. Biol.* 189.4 (June 1986), pp. 681–699.
- [239] K T Tycowski, M D Shu, and J A Steitz. "A mammalian gene with introns instead of exons generating stable RNA products." en. In: *Nature* 379.6564 (Feb. 1996), pp. 464–466.
- [240] J S Mattick. "Non-coding RNAs: the architects of eukaryotic complexity." In: *EMBO Rep.* 2.11 (Nov. 2001), pp. 986–991.
- [241] Ryan D Morin et al. "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells." In: *Genome Res.* 18.4 (Apr. 2008), pp. 610–621.
- [242] Ho-Ming Chen and Shu-Hsing Wu. "Mining small RNA sequencing data: a new approach to identify small nucleolar RNAs in *Arabidopsis*." In: *Nucleic Acids Res.* 37.9 (May 2009), e69.
- [243] D Tollervey. "A yeast small nuclear RNA is required for normal processing of pre-ribosomal RNA." en. In: *EMBO J.* 6.13 (Dec. 1987), pp. 4169–4175.
- [244] C A Enright, E S Maxwell, G L Eliceiri, and B Sollner-Webb. "5'ETS rRNA processing facilitated by four small RNAs: U14, E3, U17, and U3." en. In: *RNA* 2.11 (Nov. 1996), pp. 1094–1099.
- [245] Eesha Sharma, Tim Sterne-Weiler, Dave O'Hanlon, and Benjamin J Blencowe. "Global Mapping of Human RNA-RNA Interactions." en. In: *Mol. Cell* 62.4 (May 2016), pp. 618–626.
- [246] Michael P Guy, Marie Shaw, Catherine L Weiner, Lynne Hobson, Zornitza Stark, Katherine Rose, Vera M Kalscheuer, Jozef Gecz, and Eric M Phizicky. "Defects in tRNA Anticodon Loop 2'-O-Methylation Are Implicated in Nonsyndromic X-Linked Intellectual Disability due to Mutations in FTSJ1." In: *Hum. Mutat.* 36.12 (Dec. 2015), pp. 1176–1187.
- [247] Marina Falaleeva, Justin Surface, Manli Shen, Pierre de la Grange, and Stefan Stamm. "SNORD116 and SNORD115 change expression of multiple genes and modify each other's activity." In: *Gene* (2015).
- [248] Shivendra Kishore and Stefan Stamm. "The snoRNA HBII-52 Regulates Alternative Splicing of the Serotonin Receptor 2C." In: *Science* 311.5758 (2006), pp. 230–232.
- [249] Erik Ahrné, Timo Glatter, Cristina Viganò, Conrad von Schubert, Erich A Nigg, and Alexander Schmidt. "Evaluation and Improvement of Quantification Accuracy in Isobaric Mass Tag-Based Protein Quantification Experiments." en. In: *J. Proteome Res.* 15.8 (Aug. 2016), pp. 2537–2547.

## DECLARATION

---

I hereby declare that this doctoral dissertation "Identification of small non-coding RNA targets using computational predictions and high throughput sequencing data" has been completed only with the assistance mentioned herein and that it has not been submitted for award to any other university nor any other faculty at the University of Basel.

*Basel, 2018*

---

Rafał Wojciech Gumienny