Université de Montréal

Nucleotide Complementarity Features in the Design of Effective Artificial miRNAs

par Yifei Yan

Département de biochimie et médecine moléculaire

Faculté médecine

Thèse présentée
en vue de l'obtention du grade de doctorat
en biochimie
option biologie structurale

avril, 2018

爸爸、妈妈，谢谢你们

To mom and dad, with my deepest love

Résumé

L'importance du miARN dans la régulation des gènes a bien été établie. Cependant, le mécanisme précis du processus de reconnaissance des cibles n'est toujours pas complètement compris. Parmi les facteurs connus, la complémentarité en nucléotides, l'accessibilité des sites cibles, la concentration en espèces d'ARN et la coopérativité des sites ont été jugées importantes. En utilisant ces règles connues, nous avons précédemment conçu des miARN artificiels qui inhibent la croissance des cellules cancéreuses en réprimant l'expression de plusieurs gènes. De telles séquences guides ont été délivrées dans les cellules sous forme de shARN.

Le VIH étant un virus à ARN, nous avons conçu et testé des ARN guides qui inhibent sa réplication en ciblant directement le génome viral et les facteurs cellulaires nécessaires au virus dans le cadre de mon premier projet. En utilisant une version mise à jour du programme de conception, *miRBooking*, nous devenons capables de prédire l'effet de concentration des espèces à ARN avec plus de précision. Les séquences guides conçues fournissaient aux cellules une résistance efficace à l'infection virale, égale ou meilleure que celles ciblant directement le génome viral par une complémentarité quasi-parfaite. Cependant, les niveaux de répression des facteurs viraux et cellulaires ne pouvaient pas être prédits avec précision. Afin de mieux comprendre les règles de reconnaissance des cibles miARN, les règles de couplage des bases au-delà du « seed » ont été approfondies dans mon deuxième projet. En concevant des séquences guides correspondant partiellement à la cible et en analysant le schéma de répression, nous avons établi un modèle unificateur de reconnaissance de cible par miARN via la protéine Ago2. Il montre qu'une fois que le « seed » est appariée avec l'ARN cible, la formation d'un duplex d'ARN est interrompue au niveau de la partie centrale du brin guide mais reprend plus loin en aval de la partie centrale en suivant un ordre distinct. L'implémentation des règles découvertes dans un programme informatique, *MicroAlign*, a permis d'améliorer la conception de miARN artificiels efficaces.

Dans cette étude, nous avons non seulement confirmé la contribution des nucléotides non-germes à l'efficacité des miARN, mais également défini de manière quantitative la manière dont ils fonctionnent. Le point de vue actuellement répandu selon lequel les miARN peuvent cibler efficacement tous les gènes de manière égale, avec uniquement des correspondances de semences, peut nécessiter un réexamen attentif.

Mots-clés: le miARN, la semance, Ago2, le VIH, répression, le miARN artificiel, le multi-ciblage, la reconnaissance des cibles, la complémentarité des nucléotides

Abstract

The importance of miRNA in gene regulation has been well established; however, the precise mechanism of its target recognition process is still not completely understood. Among the known factors, nucleotide complementarity, accessibility of the target sites, and the concentration of the RNA species, and site cooperativity were deemed important. Using these known rules, we previously designed artificial miRNAs that inhibit cancer cell growth by repressing the expression of multiple genes. Such guide sequences were delivered into the cells in the form of shRNAs.

HIV is an RNA virus. We designed and tested guide RNAs that inhibit its replication by directly targeting the viral genome and cellular factors that the virus requires in my first project. Using an updated version of the design program, *miRBooking*, we become capable to predict the concentration effect of RNA species more accurately. Designed guide sequences provided cells with effective resistance against viral infection. The protection was equal or better than those that target the viral genome directly via near-perfect complementarity. However, the repression levels of the viral and cellular factors could not be precisely predicted. In order to gain further insights on the rules of miRNA target recognition, the rules of base pairing beyond the seed was further investigated in my second project. By designing guide sequences that partially match the target and analysing the repression pattern, we established a unifying model of miRNA target recognition via Ago2 protein. It shows that once the seed is base-paired with the target RNA, the formation of an RNA duplex is interrupted at the central portion of the guide strand but resumes further downstream of the central portion following a distinct order. The implementation of the discovered rules in a computer program, *MicroAlign*, enhanced the design of efficient artificial miRNAs.

In this study, we not only confirmed the contribution of non-seed nucleotides to the efficiency of miRNAs, but also quantitatively defined the way through which

they work. The currently popular view that miRNAs can effectively target all genes equally with only seed matches may require careful re-examination.

Keywords: miRNA, seed, Ago2, HIV, repression, artificial miRNA, multi-targeting, target recognition, nucleotide complementarity

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADAR: adenine deaminase

AGO (Ago): Argonaute

ARE: AU-rich elements

BBR: bicoid binding region

BMP: bone morphogenetic protein

bp: base pair

CAF1: CCR4-associated factor 1

CAT-1: catecholamine transferase 1

CCR4-NOT1: carbon catabolite repression 4-negative on TATA-less

ceRNA: competing endogenous RNA

CHX: cyclohexamide

CMV: cytomegalovirus

CrPV: Cricket paralysis virus

cS7: conserved segment 7

DGCR8: DiGeorge Syndrom critical region 8

dsRBDs: double strand RNA binding domain

DUF: domain of unknown function

eEF: enkaryote Elongation Factor

eIF: eukaryote Initialtion Factor

eIF4E-BP: eukaryotic Initiation Factor 4E binding protein

EMCV: encephalomyocarditis virus

ES cells: embryonic stem cells

FACS: Fluorescence activated cell sorting

FFluc: firefly luciferase

GDP: guanosine diphosphate

GFP: green fluorescence protein

GTP: guanosine triphosphate

HCV: hepatitis C virus

HEK 293T: human embryonic kidney 293 transformed

HITS-CLIP: high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation

miRNA: microRNA

MRE: miRNA response elements

MSCV: murine stem cell virus

mTOR: mammalian target of rapamycin

ncRNA: non-coding RNA

nt: nucleotide

ORF: open reading frame

PABP: poly(A) binding protein

PARN: poly(A)-specific ribonuclease

PCR: polymerase chain reaction

PI3K: Phosphoinositide-3 kinase

piRNA: PIWI domain interacting RNA

PIWI: P-element induced wimpy testis

Pri-miRNA: primary transcript of miRNA

PTEN: phosphatase and Tensin homolog

puro: puromycin

PVDF: Polyvinylidene fluoride

Rap: rapamycin

RIPA: Radio Immunoprecipitation Assay

RISC: RNA-induced silencing complex

RLC: RISC loading complex

Rluc: renilla luciferase

RNAi: RNA interference

RRM: RNA recognition motif

shRNA: small hairpin RNA

siRNA: small interfering RNA

TBF-β: transforming growth factor-β

TF: transcription factor

TNRC6: trinucleotide repeat containing 6

TRBP: TAR RNA-binding protein

tRNA: transfer RNA

TU: transcription unit

UTR: untranslated region

wt: wild type

Xrn1: Exoribonuclease 1

Contribution of Authors

Chapter 2:

I performed all experiments and data analyses except for:

The design of all SM shRNAs using *miRBooking* was performed by Nicolas Scott (Table I-III). One repeat of the Western blot of Rela, Akt, and tubulin proteins was performed by Roqaya Imane. The predicted alignment between the designed SM shRNA guide strands and the target sites were computed by Albert Feghaly (Table V). Etienne Gagnon provided stable cell generation strategies, technical training in tissue culture, as well as some of the stable cell lines. Gerardo Ferbeyre designed and directed the project, and proofread the manuscript. Francois Major designed and directed the project, defined the categories of the small RNA guides to be tested, supervised the adaptation of *miBooking* by *mirDesign*, which was used for the design of guide RNA sequences; in addition, Francois proofread and edited the manuscript.

Chapter 3:

I performed all experiments, wrote the *MicroAlgin* program, and conducted structural modeling and analyses except for:

Maria Acevedo generated stable cell lines that express the SM1-5 shRNA, performed Western blots on the targeted proteins, and conducted the cell growth assay to validate the efficiency and function of the knockdown of the E2Fs (**Fig. 17 and 19**). Lian Mignacca performed the RT-qPCR to validate the primers and quantitate the levels of mature miB-A to D guide RNAs (**Fig. 12G**). Philippe Desjardins made single nucleotide mutations in the seed nucleotides of miB guide strand and tested their activities using luciferase reporter assay (**Fig. 12A**). Nicolas Scott utilized external (Robertson and Wee) datasets and verified *MicroAlign* algorithm using them (**Fig. 17E**). Julie Robitaille cloned and extracted pPRIME

empty and miB-mod1 to 4 DNA plasmids (**Fig. 17C and D**). Jordan Quenneville performed cloning of the pPRIME-miB-D plasmid that I used when testing modified target sites (part of **Fig. 13H** and **Fig. 14**). Roqaya Imane performed cloning of pPRIME miB, miB A-C, as well as technical repeats of luciferase assay (**Fig. 14**); she also performed one repeat of the RT-qPCR to quantitate the level of mature miB guide strand (**Fig. 12E**). Albert Feghaly performed two repeats of the bioinformatics validation of *MicroAlign* outputs. Etienne Gagnon designed some of the validation strategies, provided technical training in cell culture, and proofread and edited the manuscript as well as the figures. Gerardo Ferbeyre directed the experimental validation of the model, and proofread the manuscript. Francois Major directed the bioinformatics validation of the model, designed computational approaches, and proofread and edited the manuscript.

Original Contribution to Knowledge

Chapter 2:

- High density of coding sequence in the HIV genome is one main reason for the ineffectiveness of siRNA, in addition to hindrance due to RNA structure.
- Puromycin can enhance miRNA-mediated repression of HIV when coding region is targeted.
- Artificial miRNAs targeting both the viral and the cellular genes provide cells with equal or better protection against invading viral particles than the perfectly complementary ones.

Chapter 3:

- The positional mismatches beyond the seed display a silencing pattern
- This pattern is only observable for 3'UTR target sites.
- The cooperativity of multiple sites enhances silencing but not as much as base pairing at key positions.
- The enhancement by 3' mismatches is only observable when no other mismatch occurs.
- Concentration of the mature guide RNA was not the cause of the silencing profile
- Base pairing propagates from the seed, skips the central part, and resumes downstream, then comes back to the central part to form the completely paired guide-target RNA duplex.
- The base pairing rule is sufficient to help design effective shRNAs.

Acknowledgement

I would like to thank my supervisor, Dr. François Major, for his support and guidance during my training. I am especially thankful to François for allowing me to pursue my own scientific interest and ideas, allowing the full development of a rather complex and fundamental science project. His emphasis on the quality of research sets a standard of undoubtedly strong work ethics in science. I feel well prepared for a career in science because of that. I would like to thank my co-supervisor, Dr. Gerardo Ferbeyre, for his guidance and motivation that helped me overcome every formidable challenge. I am especially grateful for the crucial support that he provided, together with the members of his lab, to appropriately and timely address the key issues that the referees raised during the review process of my articles.

I would like to thank Dr. Etienne Gagnon, for his unparalleled scientific intuition, shrewd foresights, and dedicated commitment that provided invaluable guidance for my training. His contribution to both manuscripts is integral and essential.

I would also like to thank the members of the Major and Ferbeyre lab, both past and present, for their contributions to my development in research projects. In addition to the acknowledgements inside the chapters, I would like to thank Julie Pelloux, for great discussions and suggestions that were later proved valuable. I would like to thank Julie Robitaille, for inspiring discussions and technical help.

I need to thank the members of my thesis committee, Dr. Pascal Chartrand and Dr. Eric Lecuyer, in addition to my supervisor and co-supervisor, for their great involvement and advices. I thank the Jury Members, Dr. Phillip Zamore, Dr. Franz Lang, and Dr. Eric Lecuyer for generously offering their time to evaluate this dissertation and my work.

In addition, I would like to express my special thank my Master supervisor, Dr. Gertraud Burger, who initially accepted me for graduate studies and inspired me with her patience and extraordinary visions about the wonders of the RNA world.

I would like to express my deep thanks to Dr. Jerry Pelletier and his lab, where I obtained invaluable training that transformed my way of thinking and prepared me for scientific research. He and his lab provided crucial theoretical and technical support for some of the experiments in this thesis.

And I would also like to thank Dr. Raquel Aloyz, Dr. Lawrence Panasci, and Dr. Jerry Price, who helped me to set my first footsteps into the realm of research.

I thank my friends Dr. John Mills and Dr. Abba Malina, whose insights and advices helped me enormously with my project. To John and Isabelle, you guys are great; and to Lillian and Amelia, thank you for supporting “uncle Yifei”.

Lastly, I thank my parents, Dr. Ju Yan and Bingzhen Chen, for their unconditional love and support throughout the years.

GENERAL INTRODUCTION

MicroRNAs are genome-encoded small RNA molecules that regulate gene expression post-transcriptionally. Mature microRNAs (miRNA) are single stranded RNA molecules of 21 nucleotides in length. Argonaute (Ago) protein associates with miRNA to form an essential component of the miRISC (miRNA-induced silencing complex), which downregulates target gene expression by cleaving the mRNA at the binding site, removing its poly-A tail, removing the 5' cap structure, or repressing its translation (Fabian et al., 2010).

First discovered as a gene that does not encode any protein but control the larval development in *Caenorhabditis elegans* by the Ambros lab (Lee et al., 1993), lin-4 was the first functional microRNA molecule identified. Its sequence is complementary to that of the 3' untranslated region (UTR) of the *lin-14* RNA and its regulatory roles were confirmed by the Ruvkun lab (Wightman et al., 1993). In the past two decades, miRNAs were found to play important roles in cell growth, division and differentiation, as well as metabolism and development.

As each metazoan miRNA is predicted to target hundreds of mRNAs due to the promiscuous base pairing between their seeds and multiple gene sequences, a large proportion of the human transcriptome is suggested to be under the control of miRNAs (Bushati and Cohen, 2007; Carthew and Sontheimer, 2009). Over half of the human genes are predicted to be directly regulated by miRNAs and hence the unique combination of miRNAs in each cell type is likely to be a determinant for the fate of thousands of mRNAs (Friedman et al., 2009; Kim et al., 2009). Supporting this view, genomic approaches such as Ago-CLIP and its derived methods identified 17,000 miRNA-target interactions in human, as well as new types of miRNA target sites (Chi et al., 2009; Grosswendt et al., 2014; Pasquinelli, 2012).

Based on the understanding of the miRNA machinery, RNAi technology has been widely applied as a gene knockdown method. However, off-targeting represent one of the main challenges due to the lack of accuracy in the prediction algorithms. Undesired gene knockdown can cause cell death and prevented its application on a larger scale. Correlations were found with base complementarity, target site location,

AU-content, secondary structure, and sequence conservation.(Agarwal et al., 2015; Grimson et al., 2007); yet precise quantifications are still required to improve the prediction algorithms. We started a project with an in-house algorithm, which implements the known targeting rules, to design RNA guide sequences that target the HIV genomic RNA as a proof of concept for the application of our artificial miRNA design strategy. Some of the guide strand showed significant protective effects against HIV infection; however, others showed activities that do not correspond well with the computer predictions. Upon analysis of the results, we further investigated features that are essential for the design of artificial miRNAs. We demonstrated that base pairs beyond the seed in the guide-target RNA duplex are formed following a particular order. In effect, such ordered base pairing has hierarchical impacts on the efficiency of the guide RNA. We validated this rule both experimentally and computationally and proposed a unifying model that describes the roles of non-seed base pairing in Ago2-mediated silencing.

CHAPTER 1: LITERATURE REVIEWED

1.1 MicroRNA Biogenesis

1.1.1 Encoding gene structure

The *miRBase* database (http://www.mirbase.org/), release of June 2013, contains 24,521 microRNA loci from 206 species, processed to produce 30,424 mature microRNA products (Kozomara and Griffiths-Jones, 2014). MiRNAs were also identified in simple multi-cellular organisms, such as poriferans, cnidarians (Grimson et al., 2008), as well as protists (such as *Dictyostelium*) (Avesson et al., 2012). Except for the placozoan *Trichoplax*, miRNAs have been identified in every animal species with a sequenced genome (Maxwell et al., 2012). According to current (at the time of writing this thesis) *miRBase*, there are 1,917 miRNA genes in humans, 1,234 in mouse, 258 in fly, 253 in worm, and 326 in *Arabidopsis thaliana*. Conserved through evolution, around 55% of *C. elegans* miRNAs have homologs in human (Kim et al., 2009). Considering the advantages of short hairpins for generating guide RNAs in gene silencing, it is suggested that miRNAs have arisen more than once in eukaryotic evolution (Bartel, 2018).

The location of miRNA-encoding sequences relative to the transcription units yielded information about miRNA biogenesis. Bradley and colleagues found that about 70% of mammalian miRNA genes (161 out of 232) are located in defined transcription units, and 117 of them are located in introns. Among the 117 intronic miRNA genes, 90 of them are located in protein-coding genes, and 27 are located in non-coding RNA genes (ncRNAs) (Rodriguez et al., 2004). Later studies revealed that miRNA genes can occur in four types of transcripts: 40% occurs in the introns of non-coding RNA ncRNA transcripts units (TU) (**Fig. 1a**); 10% occurs in the exons of ncRNAs (**Fig. 1b**); 40% occurs in the introns of coding RNA (**Fig. 1c**); some occur in either the introns or the exons (**Fig. 1d**) depending on the result of alternative splicing (Kim et al., 2009). About 50% of microRNA genes are endogenous genes that occur in clusters as polycistronic genes. They are transcribed as a single TU. Further processing is needed to generate the mature miRNA sequences. In rare cases, individual miRNA genes occur with their own promoters.

Figure 1. Four types of genomic locations of miRNA genes.

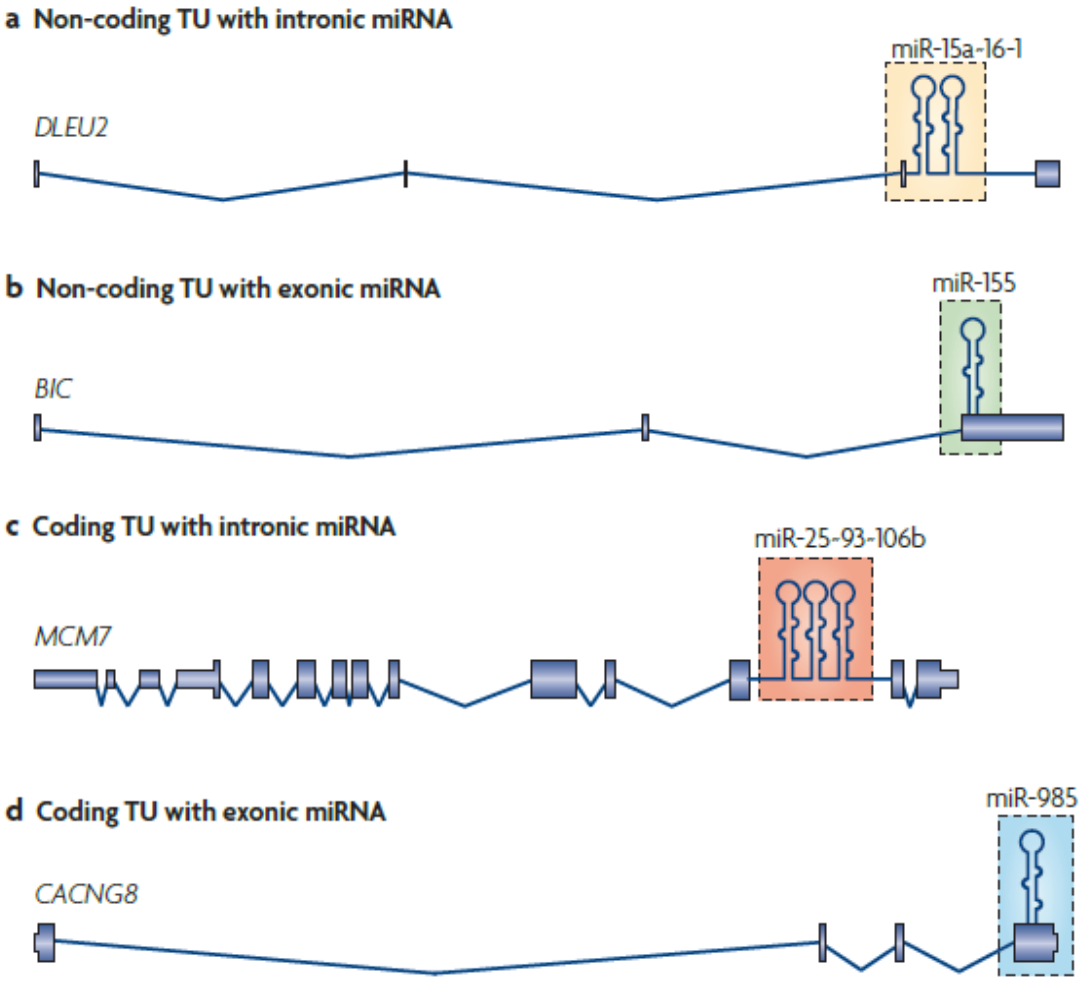


Figure 1. There are four types of genomic locations of miRNA genes. a. Intronic miRNAs in non-coding transcripts, exemplified by the miR-15a~16-1 cluster. **b**. Exonic miRNAs in non-coding transcripts. This is shown by miR-155, which was found in a non-coding RNA gene, BIC198. **c**. Intronic miRNAs in protein-coding transcripts. An example is the miR-25~93~106b cluster, which is embedded in the intron of the DNA replication licensing factor MCM7 transcript. **d**. miRNAs located in exons of protein-coding transcripts. The last exon of CACNG8 mRNA contains the miR-985 hairpin. (Kim 2009, Nature Molecular Cell Biology)

1.1.2 Biosynthesis/Transcription

Since miRNA mature sequence is only ~21 nt in length, it was originally thought that the RNA polymerase that transcribes it should belong to the RNA pol III family, which transcribes short RNA genes. However, primary transcripts of miRNA (pri-miRNAs) were later identified to be rather long, often containing thousands of nucleotides (Lee et al., 2002b). It was later confirmed that RNA pol II was mainly responsible for its transcription (Lee et al., 2004). This conclusion is supported by the fact that pri-miRNA transcripts are capped and poly-adenylated, which is the signature characteristics of the pol II transcribed genes; in addition, α-amanitin, which specifically inhibits pol II, greatly reduces the pri-miRNA levels (Lee et al., 2004). Mature miRNA can also be generated by polymerase III using transgenic constructs (Zhou et al., 2008a; Zhou et al., 2005).

1.1.3 Nuclear Processing

The pri-miRNAs are usually very long (up to several kilo bases), capped and poly-adenylated. Stem-loop structures that contain miRNA sequences need to be recognized and processed to eventually yield a mature ~21nt miRNA. Processing occurs in two steps: the first step occurs in the nucleus where the stem-loop structure of ~75nt will be cut out (pre-miRNA); the second step occurs in the cytoplasm, where the loop is removed and the double-stranded RNA molecule will dissociate and load into the RNA-induced silencing complex (RISC) to function as a guide for target sequences. We will look at these two steps as well as the transport step from the nucleus to the cytoplasm in detail.

Figure 2. The overall biogenesis pathway of miRNA.

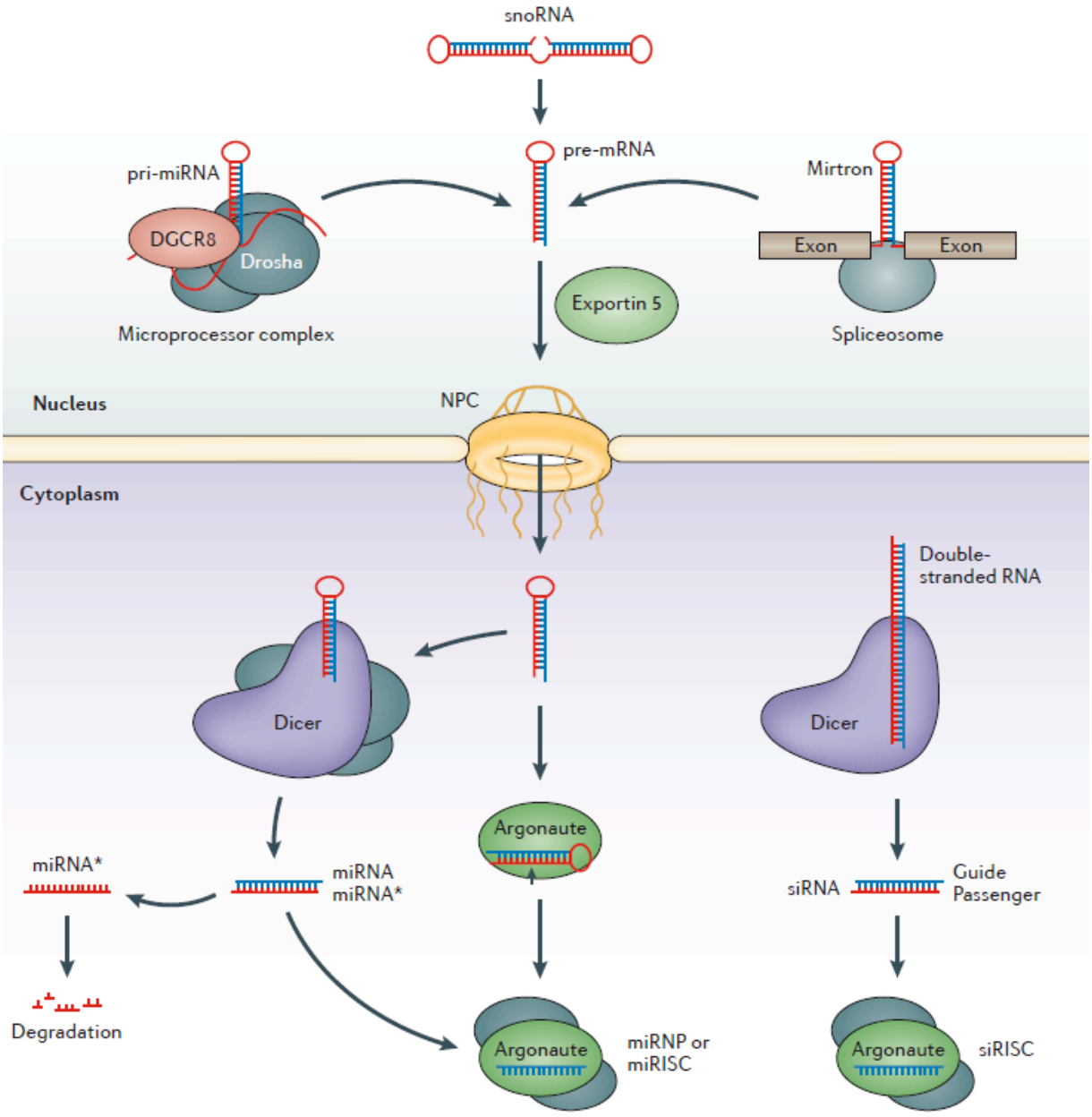


Figure 2. The overall biogenesis pathway of miRNA. The primary transcripts of microRNAs are called pri-miRNA that are processed into pre-miRNA hairpins within the nucleus. Processing by the nuclear microprocessor complex, which contains the RNase III enzyme Drosha, releases the hairpin: this part is referred to as the precursor miRNA (pre-miRNA). The primary transcript can also be generated from Mirtrons and some of the snoRNA. Mirtrons are processed by the spliceosome, while the processing machinery of snoRNA is unclear. The generated pre-miRNAs are exported into the cytoplasm by Exportin-5. They are processed there into mature miRNA by Dicer and loaded into the RNA induced silencing complex, where it directly binds to the AGO protein. The other strand is referred to as miRNA* and is normally degraded. Mature miRNAs then guide the RISC to target select mRNA transcripts for translational silencing or degradation. In the case of siRNA processing, the loaded functional strand is referred as the guide; the other, the passenger strand. Figure adapted from a 2013 review article by Meister (Meister, 2013).

As aforementioned, pri-miRNA is usually a long (up to several kilo bases) capped and poly-adenylated molecule that contains stem-loop structures. RNAse III-type protein, Drosha, which recognizes the stem of the hairpin structure, cuts out the ~75nt stem-loop from its primary transcript (Lee et al., 2003). The released stem-loop structure is termed the pre-miRNA (Lee et al., 2002b).

Drosha requires a cofactor called DiGeorge syndrome critical region 8 (DGCR8) in humans, and Pasha, in *D. melanogaster* and *C. elegans* (Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Landthaler et al., 2004). In humans, Drosha and DGCR8 form a large protein complex of ~650 kD, called the Microprocessor complex (Gregory et al., 2004; Han et al., 2004). DGCR8 recognizes two features in pri-miRNA: the single stranded base segments and stem of about 33 bp. With its assistance, Drosha is able to cleave the substrate at ~11 bp away from the ssRNA-dsRNA junction (Han et al., 2006; Zeng and Cullen, 2005). Mouse embryonic stem (ES) cells that fail to produce miRNAs suffer from defects in proliferation and differentiation when they are deficient in the *Dgcr8* gene. This phenomenon establishes the necessity of DGCR8 in the miRNA pathway as well as miRNA function in ES cells (Wang et al., 2007).

As mentioned in Section 1.1.1, many miRNA genes are located in the introns of the coding and non-coding RNAs. This suggests the possible coordination between transcription, miRNA processing, and splicing. Indeed, studies revealed that pri-miRNA processing is a co-transcriptional process (Kim and Kim, 2007b). Moreover, by mutating the Drosha recognition sequence in miR-26 or depleting Drosha in cells, Drosha processing is shown to precede intron splicing, and cleavage of the stem-loop structures within the intron does not impair splicing (Kim and Kim, 2007a). The "exon-tethering" model is favoured where the exons of Pol II transcripts are co-transcriptionally assembled into the spliceosome; then, Drosha complex processing takes place before the intron is excised (Dye et al., 2006). Evidence from chromatin precipitation and nuclear run-on assays support this model (Morlando et al., 2008; Pawlicki and Steitz, 2008).

1.1.4 Transport

Exportin-5 is the main transporter protein that is responsible for delivering pre-miRNA molecules out of the nucleus (**Fig. 2**). It mediates the transport of pre-miRNAs across the nuclear membrane by cooperatively binding to a cofactor called Ran. Upon the completion of transport, a molecule of GTP is hydrolyzed to GDP to release the cargo molecule into the cytoplasm (Bohnsack et al., 2004; Kim, 2005).

Exportin-5 was originally discovered as a minor transporter for tRNA molecules when Exportin-t, the main transporter for tRNA, is knocked down or overloaded. Later studies showed that Exportin-5 is indeed the main transporter protein of miRNA (Lund et al., 2004; Yi et al., 2003) because its depletion causes a significant decrease of the pre-miRNA level in cytoplasm.

1.1.5 Cytoplasmic processing by Dicer

Further processing of the ~70 nt stem-loop structure of pre-miRNA is carried out by the RNase III family protein, Dicer (Bernstein et al., 2001; Grishok et al., 2001; Hutvagner et al., 2001; Ketting et al., 2001; Knight and Bass, 2001). It was originally discovered as the processing protein required for siRNA generation (Bernstein et al., 2001; Hammond et al., 2000). Dicer is a highly conserved protein of about 200 kD across all Eukaryotes. It cuts away the loop structure and generates the ~22 nt double-stranded RNA molecule with two nucleotides of 3' overhang. One of the two strands becomes the mature miRNA. Knocking down Dicer causes the accumulation of pre-miRNA and the diminishment of the 22 nt mature miRNA (Grishok et al., 2001; Hutvagner et al., 2001; Ketting et al., 2001; Knight and Bass, 2001)

Dicer interacts with protein partners to carry out its dicing function. In *C. elegans*, it interacts with RDE-4 to convert the precursor RNA into a short dsRNA (Tabara et al., 2002). In *D. melanogaster*, there are two isoforms of Dicer, Dcr-1 and Dcr-2. Dcr-1 interacts with Loquacious (Loqs) and converts pre-miRNA to a miRNA/miRNA* duplex (Forstemann et al., 2005). Dcr-2 interacts with R2D2 to process long dsRNA into short siRNA duplex

(Forstemann et al., 2007; Liu et al., 2003). Human Dicer interacts with TRBP (TAR RNA-binding protein; also known as TRBP2) (Chendrimada et al., 2005; Haase et al., 2005) and PACT (also known as PRKRA) proteins (Lee et al., 2006).

1.1.6 Argonaute loading

The selection of strand to be loaded into Argonaute is not random. The selected strand, called the guide strand or anti-sense strand, will be functional and the other strand, the passenger strand or the miRNA* strand, will be degraded or discarded through an ATP-dependent process (Hammond et al., 2000; Nykanen et al., 2001; Suzuki and Miyazono, 2010, 2011). The mechanism of strand selection is based on thermodynamic stability of the RNA duplex. It was first discovered in *D. melanogaster*, where R2D2 protein senses the asymmetry in stability of the RNA duplex and binds to the more stable end of the duplex, while it forms a stable heterodimer with Dicer 2 protein and orients Ago2 on the RNA duplex (Liu et al., 2003; Siomi and Siomi, 2010; Tomari et al., 2004). It was shown that the flyAgo1-RLC operates in a similar manner via the interaction between Dcr-1 and Loqs (Chendrimada et al., 2005; Gregory et al., 2005; MacRae et al., 2008; Maniataki and Mourelatos, 2005; Tomari et al., 2004). Together they sense central mismatches in fly miRNA and preferentially load the guide strand into Ago1 (Tomari et al., 2007). Later studies demonstrated that both fly Ago1 and 2 are also able to sense the thermodynamic asymmetry as well as mismatches at key positions in the RNA duplex; strand separation subsequently takes place within the Argonaute protein (Iwasaki et al., 2009).

A similar postulate was originally made about the mammalian Ago-loading process, thinking that AGO isoforms may distinguish miRNA and siRNA. Yet studies revealed that such mechanism is partially lost in mammals. Ago1-4 do not have distinguishable preference for miRNAs; on the other hand, Ago1 and 2 have preference for siRNA duplexes compare to Ago3 and 4 (Su et al., 2009). Evidence suggests that for mammalian siRNAs, after Dicer cleavage, the siRNA duplex is released from Dicer and the more stable end binds to TRBP in the RLC and its less stable end binds to the Argonaute (Aza-Blanc et al., 2003; Khvorova et

al., 2003; Preall and Sontheimer, 2005; Schwarz et al., 2003; Tomari et al., 2004). It has also been shown that Dicer, TRBP (and/or PACT) help orient the dsRNA by the asymmetry rule when loading dsRNA into Argonaute (Noland et al., 2011). However, other studies indicate that Dicer and TRBP may not be the only ones responsible for the implementation of the asymmetry rule (Betancur and Tomari, 2012; Murchison et al., 2005). For siRNA, endonucleolytic activity of the AGO2 cleaves the passenger strand and, as the consequence, the guide strand remains in the AGO2 protein while the cleaved passenger strand undergoes degradation (Leuschner et al., 2006; Matranga et al., 2005; Miyoshi et al., 2005). Moreover, overexpression of mammalian Ago1-4 causes increase in mature miRNA levels, indicating that they all contribute to miRNA processing (Diederichs and Haber, 2007). It was later demonstrated that each human Ago isoform alone is sufficient to implement the asymmetry rule in strand selection (Suzuki et al., 2015). However, these pieces of evidence do not dismiss the importance of Dicer-interacting partners in RLC since without them loading will be inefficient (Cenik and Zamore, 2011).

1.1.7 Regulation of miRNA biogenesis

There are three main ways to regulate miRNA biogenesis: transcriptional control, post-transcriptional control, and feedback circuits (Kim et al., 2009).

1.1.7.1 Transcriptional control

Among the known miRNAs, miR-1 and miR-133 are specifically expressed in adult cardiac and skeletal muscle tissues (Horak et al., 2016). They are transcriptionally regulated. Myogenic TFs, such as myogenin and myoblast determination 1 (MYOD1), bind upstream of miR-1 and miR-133 loci and promote the transcription of these two miRNA genes (Chen, 2006; Rao et al., 2006). Some miRNAs can be potentially used to assess cancer progression due to their correlations with transcription levels of factors involved in tumour. For instance, some tumour suppressor TF can also regulate miRNA gene expression. Tumour suppressor p53 activates the transcription of the miR-34 family (He et al., 2007). On the other hand, MYC, an oncogenic protein, activates or represses a number of miRNAs that are involved in

the cell cycle and apoptosis (Chang, 2008; He, 2005). Interestingly, epigenetic control such as DNA methylation contributes in the regulation of miR-203 locus in the T-cell lymphoma but not in normal T-cells (Bueno, 2008).

1.1.7.2 Post-transcriptional control

Drosha processing represents the first post-transcriptional control point of miRNA biogenesis. As an example, in the case of the induction of miR-21, bone morphogenetic protein (BMP)/transforming growth factor-β (TBF-β) activates SMAD protein, which interacts with Drosha and DDX5 (also known as p68) and enhances Drosha processing (Davis et al., 2008). An additional post-transcriptional regulation point is the nuclear transport step. In some human cell types, miR-31, miR-128, and miR-105 precursors are retained in the nucleus while mature miRNAs are not produced, suggesting that they are regulated at the transport step (Lee, 2008).

1.1.7.3 Feedback loop control

Two types of feedback circuits are usually observed: single-negative feedback and double-negative feedback loops. Drosha and Dicer levels are regulated by the former type of feedback (Forman et al., 2008; Tokumaru et al., 2008). Drosha and DGCR8 form a single-negative feedback loop: Drosha downregulates DGCR8 by cleaving DGCR8 mRNA, while DGCR8 upregulates Drosha by stabilizing its Drosha protein (Han, 2009; Yeom et al., 2006). Human Dicer, on the other hand, constitutes a single-negative feedback with its product, let-7 miRNA, which binds to the 3'UTR of Dicer mRNA and represses its expression (Forman et al., 2008; Tokumaru et al., 2008).

Double negative feedback loops are often regarded as an efficient genetic switch of specific miRNAs during differentiation (Kim et al., 2009). They are also referred to as bistable switches in biochemical networks (Tyson and Novak, 2010). The interaction between let-7 and LIN28 falls into this category: let-7 represses LIN28 mRNA expression while LIN28 represses let-7 maturation (Newman et al., 2008; Viswanathan et al., 2008). Another example of this type of feedback is the miR-200 family and the transcriptional repressors ZEB1 and ZEB2.

Their repressive action upon each other constitutes an important switch in the epithelial-mesenchymal transition (Bracken, 2008).

1.1.8 Biogenesis from engineered constructs

Custom synthesis of RNA oligos are relatively expensive and not every lab can afford to test them in large quantities. Cloning the guide sequence of interest into an expression construct becomes a cost-effective choice for most researchers. The guide sequences are usually designed to appear on one arm of the stem of a small hairpin which can be recognized and processed by endogenous miRNA processing machinery; such engineered RNA species is termed "small hairpin RNA" (shRNA). Upon transcription, the hairpin structure of RNA is recognized by the microRNA processing machinery and subsequently processed and loaded into Argonaute. Among all constructs tested, the engineered construct based on miR-30 backbone from the Hannon lab has been most widely accepted as an efficient method of producing mature guide sequences (Dickins et al., 2005; Paddison et al., 2004; Stegmeier et al., 2005).

1.1.9 miRNA definition and annotation conventions

Given the description of miRNA gene structure, biogenesis, and function, we now know that miRNAs are almost exclusively endogenous in origin, and possibly goes through more concerted processing steps, in which both Drosha and Dicer are essential. But one may still ask the question: "How are the miRNA genes annotated?" The definition of miRNA determines its annotation in the genome. Researchers who greatly contributed to the study of miRNA across different species reached an agreement on the traits that miRNAs should possess (Ambros et al., 2003); consequently, these agreed criteria shaped the definition of miRNA. The annotation standards set in 2003 is still in use in most of the miRNA databases as well as prediction software such as *MirBase* and *TargetScan*.

In brief, five criteria were agreed upon (listed below). The first two are expression criteria that verify the existence of the miRNA, and the other three are called biogenesis criteria. They verify the authenticity of miRNA via their biogenesis characteristics.

Expression criteria:

A. Detection of a distinct ~22-nt RNA transcript by hybridization to a size-fractionated RNA sample, often by Northern blotting.
B. Identification of the ~22-nt sequence in a library of cDNA made from size-fractionated RNA. Such sequences must precisely match the genomic sequence of the organism from which they were cloned.

Biogenesis criteria:

C. Prediction of a potential fold-back precursor structure that contains the ~22-nt miRNA sequence within one arm of the hairpin.
D. Phylogenic conservation of the ~22-nt miRNA sequence and its predicted fold-back precursor secondary structure.
E. Detection of increased accumulation of organisms with reduced Dicer function.

Since it is not always possible to verify all five criteria for a particular candidate miRNA, some relaxation in the criteria is allowed. For example, A+D+E, A+D, A+C, B+D, D+E are all accepted as sufficient to annotate an miRNA.

According to the above definition, the conservation criterion plays an important role in the identification of miRNA, while the function of the miRNA is not mentioned. This leads to the fact that many of the miRNA collected in the existing databases do not have any known function, or any well validated target. Experimental validation of the targets is currently an on-going process for many annotated miRNA genes. Hence the targeting rules that facilitate the prediction of target for a given miRNA became crucial to the study of miRNA functions.

1.1.10 siRNA discovery, definition, and functions

RNA interference (RNAi) is a phenomenon by which double-stranded (ds) RNA induces sequence-specific post-transcriptional gene silencing. The term RNAi came into existence after Fire and Mello confirmed that dsRNA in both sense and anti-sense transcripts were responsible for the silencing in *C. elegans* (Fire et al., 1998). RNAi was later observed in various organisms including plants, Drosophila, nematodes and protozoa (Hannon, 2002;

Mello and Conte, 2004). It was first demonstrated in plants that the long dsRNAs were converted to small ones of ~25 nt RNA molecules in order to function as silencing triggers (Hamilton and Baulcombe, 1999). Supplying long dsRNAs to mammalian cells induces non-specific suppression of gene expression; this is because the host defense system against viral infections is activated when dsRNAs are introduced into the cell (Manche et al., 1992; Minks et al., 1979). Elbashir et al. resolved this problem by utilizing small (21-23 nucleotide) dsRNAs instead of long dsRNAs to avoid the non-specific gene suppression (Elbashir et al., 2001a). They named such small dsRNA as small interfering RNA (siRNA). With this method, they confirmed in animal cell extracts that the small RNAs were derived from the long dsRNAs and functioned as silencing triggers (Elbashir et al., 2001a). Similar siRNAs were later identified in *Drosophila* S2 (Hammond et al., 2000; Tuschl et al., 1999) as well as human HeLa cells (Martinez et al., 2002; Schwarz et al., 2003), where both sense and anti-sense strands were processed into 21-23 nt segments (Zamore et al., 2000). It was later found that Dicer processes the long dsRNA and generate siRNAs to direct silencing of specific targets (Meister and Tuschl, 2004; Tomari and Zamore, 2005).

The original inducers of RNAi were long, perfectly base-paired, linear dsRNA species and they were exogenously supplied to the cell or taken up from the environment. They are hence referred to as exo-siRNAs. Exo-siRNAs were identified in flies as a defense mechanism against invading viruses that produces long strands of dsRNA during infection (van Rij et al., 2006; Wang et al., 2006). Though the initial discovery is based on exogenous RNA, the origin of siRNA could also be endogenous. Heterochromatin sequence, including centromeres, transposons, and other repetitive sequences were found to give rise to a variety of siRNA (Lippman and Martienssen, 2004). Functional studies in plants also identified trans-acting siRNAs (ta-siRNAs) that are generated from well-defined transcription units and regulate the expression of specific genes (Allen et al., 2005; Vazquez et al., 2004). Deep sequencing revealed that in somatic tissue, cultured cells, and ovaries of *D. melanogaster*, siRNAs are derived from transposon transcripts, sense-antisense transcript pairs and long stem-loop structures (Babiarz et al., 2008; Chung et al., 2008; Czech et al., 2008; Ghildiyal et al., 2008;

Kawamura et al., 2008; Okamura et al., 2008a; Okamura et al., 2008b). In mouse, numerous types of endo-siRNAs were identified in oocytes (Tam et al., 2008; Watanabe et al., 2008), and, to a lesser extent, in ES cells (Babiarz et al., 2008). In other species such as plants and *C. elegans*, endo-siRNAs were also discovered (Chapman and Carrington, 2007). In plants, RNA-dependent RNA polymerases (RdRPs) are required to generate functional endo-siRNAs, adding more complexity to its pathway (Tang et al., 2003); on the other hand, fly and mammalian generate endo-siRNAs in an RdRP-independent manner.

So far, only flies are known to differentiate between miRNA- and siRNA-like precursor molecules by the two isoforms of Dicer. For siRNAs, Dicer 2 was the choice of processing nuclease. With the help of R2D2, the resulting guide strand RNA is preferentially loaded in Ago2 (Forstemann et al., 2007; Tomari et al., 2007). On the other hand, endogenous miRNA precursors, especially those that contain mismatched bulges, were preferentially cleaved by Dicer 1, with the help of Loquacious (LOQS; also known as R3D1) (Czech et al., 2008; Kawamura et al., 2008; Okamura et al., 2008b; Tomari et al., 2007). Such origin-dependent preference was not detected in mammals and it was suggested that siRNAs and miRNAs can functionally mimic each other depending on their complementarity with targets (Doench et al., 2003; Hutvágner and Zamore, 2002). Plant miRNA are usually highly complementary to their targets and preferentially mediates silencing through cleavage of the target mRNA. Due to the processing and targeting processes closely resemble those of siRNAs, plant miRNAs are suggested to be able to function as siRNAs (Tang et al., 2003).

In *S. pombe,* siRNAs were found to induce heterochromatin formation and lead to transcriptional gene silencing (Lippman and Martienssen, 2004). Similar observations were made in plants, animals, and ciliates. Transcriptional gene silencing is mediated via a complex, called the RNA-induced transcriptional silencing (RITS) complex, which contains Ago1 loaded with the siRNA. The interaction between RITS complex and RNA polymerase II facilitates siRNA's recognition of the nascent transcript (Buhler et al., 2006; Djupedal et al., 2005; Kato et al., 2005). RITS association promotes histone H3 methylation on lysine 9 (H3K9) by histone methyltransferases (MHTs); as a consequence, the chromodomain-

containing protein Swi6 is recruited and chromatin compaction takes place (Lippman and Martienssen, 2004).

1.2 Messenger RNA (mRNA) architecture

As mRNA is targeted by miRISC, understanding the functional organization of mRNA is rudimentary to the study of functions of miRNA. Messenger RNA is transcribed in the nucleus by RNA polymerase II (Pol II). It is used as blue print for protein synthesis in the cytoplasm. In some cases, mRNA is targeted to specific subcellular locations for translation or temporary storage (Rodriguez et al., 2008). In Eukaryotes, it usually consists of the following regions: the 5' cap structure, the 5'untranslated region (5'UTR), the start codon that marks the start of the coding region (CD), the stop codon which signals the end of the coding region and the start of the 3' untranslated region (3'UTR), and the poly-A tail. The 5' cap and the 5' UTR are important for the recruitment of initiation factors that signals the ribosomes to start translation. Coding region is located between the start codon and the termination codon and it contains a series of codons that encode the amino acid sequence. When a ribosome scans over the start codon in a suitable context, methionine initiator $tRNA_i$ is brought to the P site of the ribosome and it initiates protein synthesis. In mammalians, the start codon (AUG) itself encodes the methionine residue (Sonenberg and Hinnebusch, 2009).

Further downstream of the stop codon, the mRNA is not translated. This region is called the 3'UTR. The primary role of the 3'UTR is to regulate mRNA stability. The AU-rich elements (AREs) control mRNA degradation and translation via interactions with specific binding proteins (ARE-BP) (Helfer et al., 2012). The AREs usually contain AUUUA pentamers and U-rich sequences; however, there is little sequence homology among AREs besides those motifs (Chen and Shyu, 1995). For instance, AUF1 protein binds to the ARE and, depending on the mRNA it binds, it could either stabilize or destabilize the mRNA (Loflin et al., 1999; Xu et al., 2001). On the other hand, HuR binds AREs and universally stabilizes mRNA by inhibiting 3'-5' degradation (Brennan and Steitz, 2001). In *Drosophila*, the Caudal mRNA contains a bicoid binding region (BBR) in its 3'UTR, which recruits the Bicoid protein and causes the tethering of the 5' end to the 3'end of the mRNA. As the result, the Caudal mRNA is maintained in a translationally inactive state (Cho et al., 2005).

One important way that 3'UTR contribute to the stability of mRNA is via the presence of miRNA response elements (MREs). As elaborated later in this thesis (section 1.5.2.5), the miRNAs prefer certain regions of the 3'UTR as their targets for effective repression of the encoded gene. The reason why miRNAs preferentially target the 3'UTR is thought to avoid the impeding ribosomes (Guo et al., 2010). The MREs present in the coding regions are likely protected by the traveling ribosomes, which can even displace bound miRISC. Recent report has shown that the coding region can also be targeted by miRNA and consequently repressed via slicer-independent mechanisms (Zhang et al., 2018).

When RNA polymerase II completes the transcription of precursor mRNAs, a process called polyadenylation takes place (Albert L. Lehninger, 1993; Colgan and Manley, 1997). It refers to the addition of adenosine to the 3'end of the mRNA (normally 200-300 adenosines) and the added long tract of adenosines is called the poly(A) tail. Nuclear poly(A) binding protein (PABP) binds to the nascent poly(A) tail and enhances the polymerase activity of PAP, facilitating further extension of the poly(A) tail. Cytoplasmic PABP binds to 3'UTR and it inhibits mRNA degradation as well as promotes translation (Bernstein and Ross, 1989; Borman et al., 2000) by facilitating the circularization of mRNA through its binding to the eIF4G scaffolding protein (Sonenberg and Hinnebusch, 2009).

1.3 The Argonaute genes

1.3.1 The Argonaute genes

Many Argonaute orthologs are found in metazoan and plant genomes. The orthologs are usually identified in piRNA (PIWI domain interacting RNA) pathways in germ cells or other classes of small RNAs. Only the original AGO proteins identified in miRNA and siRNA pathways mediate gene silencing (Peters and Meister, 2007). Argonaute proteins are classified into three orthologous groups: Argonaute-like proteins are similar to *Arabidopsis thaliana* AGO1; Piwi-like proteins are closely related to *D. melanogaster* PIWI (P-element induced wimpy testis); and the last one contains the *C. elegans*-specific group 3 Argonautes (Yigit et al., 2006). Argonaute-like and Piwi-like proteins are found in bacteria, archaea, and eukaryotes, implying their ancient origin (Cerutti and Casas-Mollano, 2006). The number of Argonaute genes varies depending on the species. In human, there are 8 Argonaute paralogous genes, including four Argonaute-like and four Piwi-like genes; five were found in *D. melanogaster*, including two Argonaute-like and three Piwi-like proteins; in *A. thaliana*, 10 Argonaute-like were identified; only one Argonaute-like protein was identified in *Schizosaccharomyces pombe*; at least 26 Argonaute genes in *C. elegans* (5 Argonaute-like, 3 Piwi-like and 18 group 3 Argonautes) (Hutvagner and Simard, 2008).

The human Argonaute has four paralogs, hAgo1-4, which are ubiquitously expressed. Human AGO1, 3, and 4 genes are located next to each other on chromosome 1, while hAGO2 gene is located on chromosome 8 separately (Nakanishi et al., 2013). AGO2 is well known for being the only cleavage-active form of the four. It was later shown that each one of Ago1-4 is essential in mouse embryonic stem (ES) cells as they all prevent cells from undergoing apoptosis (Su et al., 2009).

Ago1-4 were thought to have redundant functions for repression of translation; it was later demonstrated that Ago3 is a more potent translational repressor than other non-cleaving AGOs when tethered to the 3'UTR of a reporter transcript (Wu et al., 2008). Loss of Ago2 in

hematopoietic cells results in deficiency of B-cell and red blood cell development (O'Carroll et al., 2007). Germ line Ago2 deficiency is embryonically lethal (Liu, 2004; Meister, 2004). Interestingly, male mouse germ line cells express high levels of Ago4 as well as Ago3 (Gonzalez-Gonzalez et al., 2008). Ago4 is demonstrated to be specifically responsible for the silencing of many sex-linked transcripts in male germ line. Loss of Ago4 results in fertility defects including reduced testis sized and lower sperm counts in male mice (Modzelewski et al., 2012). On the other hand, Ago1 and Ago3 are required for RNAi pathways of cellular defence against influenza A viral infection. Ago1 and Ago3 double-knockout cells are significantly more vulnerable to this RNA virus (Van Stry et al., 2012). Hence recent evidence may suggest that each Ago paralog may take specialized role when acting as the repressor of translation.

As the overexpression of each of the four paralogs is demonstrated to enhance the production of mature miRNAs, it was postulated that they may also have redundant functions in miRNA maturation (Diederichs and Haber, 2007). However, it was shown in mouse ES cells that Ago1 and 2 have preferences for siRNA duplexes during Ago-loading comparing to Ago3 and 4. This indicates that the four paralogs have non-redundant roles in miRNA maturation because they may distinguish RNA duplexes based on their complementarity (Su et al., 2009).

In addition, a recent study has revealed that Ago proteins can be differentially regulated by phosphorylation. Phosphorylation at Ago2 Y529 inhibits it from being loaded with small RNAs (Rudel et al., 2011); on the other hand, EGFR-dependent phosphorylation of Ago2 Y393 hinders the processing of looped precursor RNAs (Shen et al., 2013). Moreover, Akt3 is shown to phosphorylate Ago2 S387 and alters its activity from cleavage toward translational repression (Horman et al., 2013; Zeng et al., 2008). It was recently shown that phosphorylation of Ago2 S387 by Akt3 induces LIMD1 binding, which enables the recruitment of TNRC6 by Ago2. The assembly of Ago2-TNRC6 complex switches the Ago2 activity to favour translational repression. In the absence of LIMD1, Ago2 miRNA-silencing function is lost and translational repression is mainly mediated by Ago3 (Bridge et al., 2017).

1.3.2 Structural organization of the Argonaute protein

Eukaryotic Argonaute proteins usually consist of four domains: the N-terminal, PAZ (PIWI-Argonaute-Zwille), MID and PIWI domains (**Fig. 3**). Between the N-terminal and PAZ domains, there is a loop called L1; similarly, between PAZ and the MID domain, there is another loop called L2. These two flexible loops render the PAZ domain more flexible relative to the rest of the Argonaute protein.

The N-terminal domain is commonly found in animal Argonautes, unlike the other three domains, which are ubiquitously found in all species. The N-terminal domain was thought to prevent base pairing of the target to the guide beyond nt 16 in the guide sequence (denoted as g16) (Faehnle et al., 2013; Kwak et al., 2012).

PAZ domain is found in both Argonautes and Dicer. It contains an OB-like fold (oligonucleotide/oligosaccharide binding fold). Data from structural as well as biochemical studies shows that it binds to single-stranded (ss) nucleic acids (Lingel et al., 2003, 2004; Ma et al., 2004; Song, 2003; Yan, 2003). The binding appears to be sequence-independent; however, an intriguing feature is that PAZ recognizes the 3' end overhang of ssRNAs, where two such overhang nucleotides are typically produced after Dicer processing of the pre-miRNA. For human Ago2, the entire protein structure looks like a bird with two wings (MID and N domains) spread out and the head (PAZ) tilted upward, holding the 3' end of the guide RNA in its bill. The RNA guide strand is threaded through the N-PAZ channel.

MID domain of Argonaute protein binds to the 5' end of the small RNA guides , with some preference to U or A at position 1 (Boland et al., 2010; Frank et al., 2010; Parker et al., 2005). The nucleotide preference is achieved via specific contacts with amino acid residues in the MID lobe. Nucleotide 2-10 of the RNA guide strand is threaded through the RNA binding groove between PIWI and MID lobes (Elkayam et al., 2012).

PIWI domain contains the catalytic residues that cleave the target RNA. The catalytic center is a RNase-H-like fold (containing the DEDH catalytic tetrad), which was originally found to cleave RNA that are base-paired with DNA(Ma et al., 2005; Parker et al., 2004; Song et al., 2004; Yuan et al., 2005). In human, Ago2 is the only isoform that is capable of cleaving

target RNA. Ago1, 3, and 4 mediate repression by slicer-independent pathways. The requirement for divalent cation as well as 5'-phosphate and 3' OH detected in the products confirmed its RNase H characteristic (Tolia and Joshua-Tor, 2007). Such catalytic tetrad is conserved in hAgo3 but altered to be DEDR in hAgo1and hAgo4, which are compromised in catalytic activity. Another characteristic of the PIWI domain is that it accommodates the GW182 protein by binding to its WG/GW repeats (Till et al., 2007).

Figure 3. The human Ago2 protein with a modeled guide-target RNA duplex bound to it.

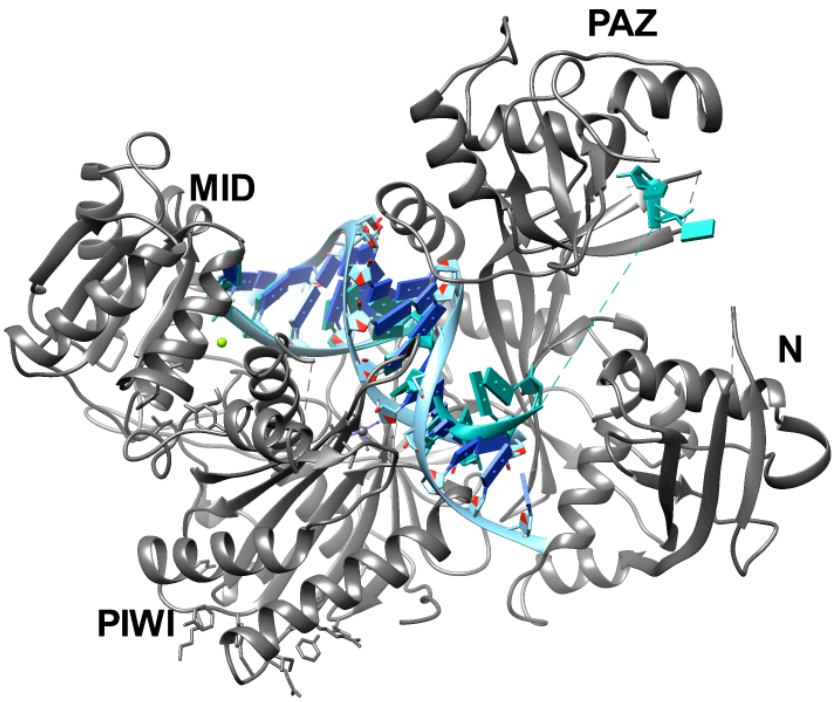


Figure 3. The human Ago2 protein with a modeled guide-target RNA duplex bound to it. The domains are labeled in bold letters. The Ago2 structure is adapted from PDB 4F3T from Elkayam et al. (Elkayam et al., 2012). The 3' end of the guide RNA is bound to the PAZ domain while the nucleotides between the nt15 and the 3' end could not be resolved.

1.3.3 Argonaute slicer activity

The Argonaute protein is a multi-functional protein. It plays versatile roles in small RNA biogenesis and gene silencing. The most prominent and potent effect is its RNase activity. All AGOs are not slicer-active. In humans, Ago2 is the only isoform that is capable of cleaving target mRNA. It cleaves the mRNA via RNase H-like mechanism, utilizing Mg^{2+} as a co-factor. Mutational experiments were conducted to understand the protein features that are important for nuclease reaction to occur as well as the reasons why some Argonautes are RNase-inactive. The first structural and functional study on hAgo2 was conducted by Liu et al., who investigated the mammalian slicer Ago2 (Liu, 2004). When a human RISC is assembled with a guide RNA, the RISC becomes an enzyme for the highly complementary MREs (Haley and Zamore, 2004).

Guide-loaded RISC is a multiple-turnover enzyme; after cleaving a perfectly paired target, it leaves with the guide strand intact and becomes ready to bind next target (Haley and Zamore, 2004; Hutvágner and Zamore, 2002; Martinez and Tuschl, 2004). In Zamore's study, they measured values of K_m and k_{cat}, which are measurements of the affinity and the turnover rate of an enzyme, respectively. In addition, the ratio between k_{cat} and K_m, which is the "specificity constant", is a classical measure of catalytic efficiency and corresponds to the second order rate constant of the reaction when the substrate concentration is much lower than the K_m. In the study of wild-type *D. melanogaster* RISC loaded with *let-7*, k_{cat}/K_m is ~$8.4X10^{-4}$ $nM^{-1}S^{-1}$ (Haley and Zamore, 2004). The measured values are much slower than the expected rate of collision of RISC with mRNA (>= 10^{-2} $nM^{-1}S^{-1}$) (Haley and Zamore, 2004). This indicates that there are factors that rate-limit the reaction, possibly due to the conformational changes required during the target recognition and cleavage (Haley and Zamore, 2004).

This suspicion eventually led to the proposition of the "two-state model" of Ago2's cleavage mechanism. Cross-linking experiments suggests that the 3' end of the guide strand binds PAZ domain (Tomari et al., 2004), confirmed by structural studies (Lingel et al., 2003, 2004; Ma et al., 2004; Song, 2003; Yan, 2003). Upon seed binding to the target, the guide

RNA becomes extended along the positively charged binding cleft in the N-PIWI channel to complete the full complementarity with the target, as predicted by the structure of *P. furiosus* Argonaute (Song et al., 2004). Hence the guide RNA can bind to the target only by seed pairing, with its 3' end remains bound to PAZ but not paired with the target, or with 3' end released from PAZ and fully paired with target. It was suggested that the 3' end-released state is the one in which Ago2 becomes capable of cleaving the target; the implication is that the time required to achieve this pre-cleavage conformation could be rate limiting.

Kinetic studies suggested that the product release can be the rate-limiting step (Haley and Zamore, 2004; Rivas et al., 2005; Tang et al., 2003; Wee et al., 2012), similar to the general pattern of RNase P cleavage (Tallsjo and Kirsebom, 1993). Single-molecule techniques confirm that the target release process can be rate-limiting for siRNA-directed cleavage of RISC (Salomon et al., 2015). Interestingly, using population fluorescence based methods, the formation of additional base pairs beyond the seed is suggested to require longer "dwell-time" before cleavage happens (Deerberg et al., 2013).

1.3.4 Structural studies of Argonaute proteins

Early structural studies focused on the specificity of loaded guide strand and understanding of the origin of the RNase activity. They sought to confirm the existence of an RNase H-like reaction center. Before eukaryotic structure became available, hints were taken from bacterial (Wang et al., 2008b; Wang et al., 2009b; Wang et al., 2008c) and archaeal (Rashid et al., 2007; Song et al., 2004) AGOs. The first full-length Argonaute structure became available was PfAgo (Song et al., 2004).

Interestingly, these structures showed that there is a deep pocket between the MID and PIWI domains and the 5'end of the guide RNA binds to it. In the pocket, four residues that are highly conserved across archaeal and eukaryotic interact with the 5' phosphate of the first nucleotide of the guide (Willkomm et al., 2015). In *A. fulgidus*, *A. thaliana*, and human MID, such pocket contains Y-K-Q-K tetrad (Elkayam et al., 2012; Frank et al., 2012; Frank et al., 2010; Ma et al., 2005; Schirle and MacRae, 2012). Another finding from these structures is

that there is generally no specific contact between the bases and the amino acid side-chains of the Argonaute proteins; however, in eukaryotic Argonautes, a nucleotide specificity loop can be found to recognize the first nucleotide of the guide (PDB:3LUD), giving it a preference for A or U nucleotide (Frank et al., 2010). In both TtAgo and hAgo2 structures, when the guide strand is loaded, a kink is introduced in the seed region (nt 6 for hAgo2 and nt10 for TtAgo).

Beyond the seed, the structure of the guide RNA is highly disordered, except for the last 2-3 nucleotides that are bound to the PAZ domain (Willkomm et al., 2015). The last two nucleotides of the guide are in contact with aromatic and basic residues in the PAZ domain (Elkayam et al., 2012; Lingel et al., 2003; Ma et al., 2004; Ma et al., 2005; Nakanishi et al., 2013; Schirle and MacRae, 2012; Wang et al., 2008c). Though there is no structural data for archaeal Ago loaded with guide RNA, single-molecule approaches provided evidence for the 3' anchorage of guide strand in the archaeal Argonaute's PAZ domain (Zander et al., 2014). Combined with kinetic studies, the observation of the bound 3' end of the guide RNA led to a "two-state model" for mammalian Argonautes (Rashid et al., 2007; Wang et al., 2009b).

In this model, the PAZ domain can be either in a state where the 3' end of the guide is bound or in a state where it is released. Subtlety arose from the structural analysis about the target recognition and cleavage process using bacteria *Thermus thermophilus* Argonaute (TtAgo) protein structures (Wang et al., 2008b). In contrast to the two-state model, a "nucleation, propagation, and cleavage model" for bacteria Argonaute was proposed (Wang et al., 2009b). Bacterial AGO requires sequential binding to the target in the 5' to 3' direction along the guide. In crystal structure determination study, a more stable complex can be visualized by using an RNA-DNA target duplex to bind to TtAgo, preventing target cleavage. The structure of a well accommodated duplex was resolved and clearly indicated that the flexibility of PAZ is important for the step-wise accommodation of the target to occur; moreover, recent structural studies confirmed that the correct positioning of the scissile nucleotides relative to the RISC reaction center rate-limits the cleavage process (Deerberg et al., 2013; Salomon et al., 2015; Zander et al., 2014).

However, one recent structural study raised concern about the propagation model by showing that the existence of a structural element called α-7 helix, which is unique to archaeal and eukaryotic Argonautes, can prevent base pairing further downstream of the seed (Schirle et al., 2014). In their study, structures of human Ago2 loaded with duplexes of a target strand and complementary strands of different lengths were determined. This study confirmed the importance of the seed pairing, and the α-7 helix causes a narrowing of the PAZ-MID cleft, through which the guide strand is threaded. The narrowing would not allow the accommodation of a double-stranded RNA and hence base-pairing is not favoured immediately downstream of the seed. This observation is consistent with kinetic data collected using single-molecule approaches, where a "second seed" is detected in the 3'-supplementary region (Salomon et al., 2015). Moreover, Schirle's study showed that the seed pairing opens the N-PAZ channel, which facilitates supplemental pairing at nt11-16. This is suggested by the observation that g11-g16 nucleotides shifted to adopt a near A-form conformation, which is the favourable conformation in RNA duplex. It then became puzzling how the pre-cleavage complex can form inside the eukaryotic AGO2 accommodation site when base-pairing in the scissile nucleotides are prohibited.

1.4 miRNA functional overview

Argonaute, is the core component of the miRISC (miRNA-induced silencing complex) and it is responsible for mediating both the slicer-dependent and slicer-independent pathways of silencing. Once loaded with miRNA, the Argonaute protein acquires specificity towards the RNA target (Bartel, 2004). In mammals, miRNAs were predicted to have many conserved genes targets (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Xie et al., 2005), as more than half of the protein-coding targets are under its regulation (Friedman et al., 2009) and almost every cellular process up to date is subject to the regulation of miRNAs (Bartel, 2009; Bushati and Cohen, 2007; Friedman et al., 2009).

Argonaute uses a small RNA, which could be either a mature miRNA or siRNA, as a guide to find the target RNA molecule via base complementarity. In plants, miRNAs usually base-pair with their targets via near-perfect base-pairing and it elicits endonucleolytic cleavage of the target mRNA (Bartel, 2009). This "slicer" activity is usually referred as the mechanism for RNAi, though the term "RNAi" does not explicitly name the nuclease activity as the only mechanism of regulation. In this thesis, this pathway is called the "slicer-dependent" pathway of target gene silencing. In animals, perfect base-pairing between miRNA and target is rare; mismatched central region of different sizes are often found. The target mRNAs are mostly recognized by base-pairing between nt 2-8, call the "seed", of the miRNA and the 3'UTR of the target mRNA (Lewis et al., 2003). RNA duplex formation in the seed alone contributes the most to the specificity of target recognition (Bartel, 2009; Bushati and Cohen, 2007; Filipowicz et al., 2008; Friedman et al., 2009; Ghildiyal and Zamore, 2009). Due to the mismatched nucleotides, slicer activity was absent and the target mRNA is regulated either by repression of translation or by triggering deadenylation and decay of the target mRNA. This pathway is often referred as the "slicer-independent" mechanism (Kawamata and Tomari, 2010). I will cover both classes of silencing below in more detail. Regardless the mode of silencing, miRNAs have profound impact on protein levels (Selbach et al., 2008); moreover,

their effects on target mRNA levels were more predominant than those on the protein levels (Baek et al., 2008; Guo et al., 2010).

1.4.1 Slicer-dependent silencing

Slicer activity is mediated by the Argonaute protein. The only slicer-active human AGO in RNAi pathway is hAgo2 (Song et al., 2004). The cleavage requires perfect or near-perfect alignment between the RNA guide and the target, so that the guide strand can bring the target RNA close to the DEDH-motif reaction center. An endonucleolytic incision is made on the phosphate back bone in the target strand between two bases that face nt10 and 11 of the guide strand, counting from the 5' end of the guide (Wang et al., 2009b). Once the target strand is cleaved, the products are released from AGO2 and the release step is the rate-limiting step under normal conditions. Slicer-dependent activity is rare for miRNA-mediated silencing in metazoans.

1.4.2 Slicer-independent silencing

Though some Ago isoforms can cleave the target mRNA, the degradation of target mRNA is not always due to the catalytic activity of the Ago protein (Behm-Ansmant et al., 2006; Giraldez et al., 2006; Wu et al., 2006). In the slicer-independent action of miRNA, protein output of targeted genes is reduced as the stability of mRNAs or its translation being compromised. Slicer-independent mRNA degradation process requires GW182, as well as the decapping, deadenylation, and exonucleolytic machineries in addition to the Ago protein. Later studies suggest that such degradation is miRNA-dependent, but not always dependent upon active translation (Wakiyama et al., 2007; Wu et al., 2006), and it is not a secondary effect of translation shut-down (Wu et al., 2010). It was shown that the complementarity between miRNA and mRNA also likely affects the decision process of translation arrest or degradation (Aleman et al., 2007).

1.4.2.1 Translational repression

Translation refers to the process in which ribosome synthesize proteins according to the mRNA blue print. Inhibition of this step has a reversible but immediate effect on gene expression. MicroRNA-mediated repression of translation was first reported when both mono- and bi-cistronic reporter mRNA's 3'UTRs were targeted by endogenous let-7 (Pillai et al., 2005). Artificial CXCR4 miRNAs can also repress the translation of reporter mRNA (Humphreys et al., 2005). In these studies, the targeted mRNA shifted to the lighter fraction of the ribosomal density gradient, which indicates the repression of translation. Similar shifts were reported in Huh7 cells for CAT-1 mRNA targeted by miR-122 (Bhattacharyya et al., 2006), in HEK 293T cells for a reporter mRNA targeted by miR-16 (Huang et al., 2007a), and in *C. elegans* for mRNAs, such as those of daf-12 and lin-41, targeted by multiple miRNAs (Ding and Grosshans, 2009).

Several groups have shown that repression of translation by miRNA is cap-dependent. They demonstrated that in cap-independent translation or when the 5'-cap structure is non-functional, miRNA-mediated repression becomes refractory (Humphreys et al., 2005; Pillai et al., 2005). In contrast, other studies demonstrated that miRNA-mediated repression takes place at post-initiation steps (Gu et al., 2009; Maroney et al., 2006; Nottrott et al., 2006; Olsen and Ambros, 1999; Petersen et al., 2006). The most convincing evidence originates from investigations of in *C. elegans*, where lin-14 and lin-28 mRNAs remain associated with polysomes during larval development under the repression of lin-4 miRNA. However, given the miRNA-mediated repression is dependent on the target site location in the 3'UTR as well as its repeat numbers, the association should not be taken as a definitive proof that repression occurs during the elongation in mammalians (Fabian et al., 2010). It remains debatable whether miRNA mediates repression through eIF4E (Eulalio et al., 2008; Kinch and Grishin, 2009) and the 80S ribosome (Fabian et al., 2009; Wang et al., 2008a).

1.4.2.2 Non-translational repression mediated by miRNA

The decay of mRNA is known to go through several mechanisms that involve the removal of the poly(A) tail in the 3' to5' direction by exoribonucleases. The first mechanism involves the CCR4-NOT1 (carbon catabolite repression 4-negative on TATA-less) complex, which contains the deadenylases CCR4/CNOT6 and CAF1 (CCR4-associated factor 1)/CNOT7. CAF1 is an RNase D family deadenylase. The second involves poly(A)-specific ribonuclease (PARN) (Korner and Wahle, 1997; Virtanen et al., 2013). The third one is by poly(A) nuclease (Meyer et al., 2004; Yamashita et al., 2005). Without the poly(A) tail, stability of the mRNA is greatly reduced. The 5'-3' direction decay of mRNA starts by the removal of the 5' cap structure by decapping enzymes such as DCP1-DCP2 complex. The cap-less mRNA is then degraded in the 5'-3' direction by exoribonuclease Xrn1 (Coller and Parker, 2004).

Interestingly, in miRNA-mediated mRNA decay, both deadenylation and decapping activities were detected. Following deadenylation, decapping leads to rapid degradation of mRNA in 5'-3' direction. Supporting evidence was found across different species. In zebra fish, miR-430 elicits the deadenylation of hundreds of maternal transcript at the early stage of embryo development (Giraldez et al., 2006). In P19 embryonic carcinoma cells, *lin-28* mRNA is deadenylated under the effect of miR-125, of which the level increases during retinoic acid-induced neuronal differentiation (Wu and Belasco, 2005). In mammalian and *Drosophila* cell-free extracts, miRNA-mediated deadenylation was also observed (Fabian et al., 2009; Iwasaki et al., 2009; Wakiyama et al., 2007). In *C. elegans* embryos, the deadenylation of the 3'UTR was found to be pervasive and cooperative (Wu et al., 2010).

1.4.2.2.1 Role of GW182 in deadenylation and decay

The deadenylation process requires the miRISC that contains both AGO and GW182 (Behm-Ansmant et al., 2006). GW182 proteins are crucial for miRNA-mediated repression (Eulalio et al., 2009) as they interact with all mammalian AGO proteins as well as Drosophila AGO1. They are required for miRNA-mediated deadenylation and decapping (Behm-Ansmant et al., 2006; Eulalio et al., 2008; Eulalio et al., 2007; Fabian et al., 2009; Iwasaki et al., 2009;

Wakiyama et al., 2007). Knocking down or immunodepleting human AGO2 or Drosophila AGO1 (Behm-Ansmant et al., 2006) abolishes miRNA-mediated deadenylation and stabilizes miRNA-targeted mRNA. Knocking down GW182 in *Drosophila* S2 (Behm-Ansmant et al., 2006; Chekulaeva et al., 2009; Iwasaki et al., 2009), mammalian (Zipprich et al., 2009), as well as *C. elegans* cells (Ding et al., 2005; Ding and Grosshans, 2009) diminished both translational repression and mRNA decay.

The mammalian counterparts of the GW182 protein are TNRC6A, B, and C. They contain glycine (G)-tryptophan (W) repeats in the N-terminal portion, followed by a glutamine (Q)-rich domain, a domain of unknown function (DUF), and an RNA recognition motif (RRM) domain. The *Drosophila* homolog is known as Gawky. In *C. elegans*, its homolog is called AIN-1 and AIN-2, which contain the GW domain while lack the DUF and RRM domains (Ding et al., 2005; Zhang et al., 2007). The GW rich domain is responsible for interaction with the AGO proteins, and the region extending from the N-terminus to the Q-rich domain targets dGW182 to P-bodies (Behm-Ansmant et al., 2006).

GW182 interacts with the AGO proteins via GW repeats in its N-terminus through binding to the MID/PIWI domain of the AGO protein (Behm-Ansmant et al., 2006; El-Shami et al., 2007; Lian et al., 2009; Takimoto et al., 2009; Till et al., 2007). To confirm this interaction, the GW-rich fragment of GW182 protein, called the "GW hook", was expressed in *Drosophila* cells. The GW hook competes with the GW182 protein and hampered miRNA-mediated repression (Eulalio et al., 2008). In addition, purified GW hook peptide is shown to block miRNA-mediated translational repression or deadenylation *in vitro* (Fabian et al., 2009; Takimoto et al., 2009; Till et al., 2007). Tethering of GW182 to the mRNA repressed translation and causes mRNA to decay even in the absence of AGO protein, further confirming that AGO acts as a scaffold to recruit GW182 (Behm-Ansmant et al., 2006; Chekulaeva et al., 2009; Eulalio et al., 2008; Li et al., 2008). Taken together, GW182 is involved in both mRNA decay and translational repression pathways (Fabian et al., 2010).

1.4.2.2.2 The deadenylation complexes

CCR4-NOT complex is involved in miRNA-mediated deadenylation. Deadenylation is considered to be the first step in mRNA decay. In yeast, deadenylation precedes mRNA decay and it involves collaboration between Ccr4 and Pan2/Pan3 (Brown et al., 1996; Chen and Shyu, 2011; Chen et al., 2002; Dupressoir et al., 2001; Tucker et al., 2002; Weyand et al., 2001; Yamashita et al., 2005). GW182 protein is shown to recruit this complex to the mRNA to help it carries out its function on a specific mRNA, as shown in the tethering experiment where GW182 is linked to the 3'UTR (Behm-Ansmant et al., 2006).

1.4.2.2.3 The decapping enzymes

The removal of poly(A) tail disrupts the circularization of the mRNA and hence exposes the 5' cap structure, which is then removed by the decapping enzymes Dcp1/Dcp2. Degradation can then proceed in the 5'-3' direction by an exonuclease, Xrn1 (Houseley et al., 2006; Steiger et al., 2003; Wang et al., 2002). Using RNAi technology, factors that are required for RNA-mediated deadenylation, decapping, and decay were screened in Drosophila S2 cells (Behm-Ansmant et al., 2006; Eulalio et al., 2008; Eulalio et al., 2007; Rehwinkel et al., 2005). Decapping complex proteins DCP1/DCP2, along with their enhancer proteins Ge-1, EDC3, HPat, and Me31B, were identified. Knockdown experiments targeting these factors showed that mRNA was stabilized in spite of deadenylation. This means that deadenylation alone is not sufficient to elicit the decay of target mRNA and subsequent decapping needs to take place to warrant efficient repression of target expression (Eulalio et al., 2007; Fabian et al., 2010).

1.5 Factors that influence the efficiency of silencing

1.5.1 The intrinsic factors for guide RNA-mediated silencing

Once a small RNA is loaded into RISC, its origin cannot be discerned as whether it came from miRNA or siRNA. Directed by base complementarity, the guide RNA-bound Argonaute 2 cleaves the target mRNA at a single phosphodiester bond between the ribonucleotides opposing position 10 and 11 of the guide strand (Elbashir et al., 2001b; Hammond et al., 2001; Meister and Tuschl, 2004; Rand et al., 2004; Rivas et al., 2005; Song et al., 2004; Tomari and Zamore, 2005).

1.5.1.1 Amino acid sequence, substrate, and cofactors of human Argonaute2 slicer

The slicing activity of the hAgo2 was shown to be Mg^{2+}-dependent. Moreover, an siRNA with a 5' phosphate directs target cleavage more efficiently than that with a 5'-hydroxyl. The slicer activity is RNA-specific; in other words, hAgo2 is unable to cleave DNA target or use DNA as guide (Rivas et al., 2005).

The amino acid sequence of the Ago protein is the primary factor for slicer activity. The PIWI domain contains the DDE motif of RNase H. Altering some key residues of Ago leads to the conversion of inactive hAgo1 into an enzymatically active form that is comparable to hAgo2 (Faehnle et al., 2013). The PIWI domains of hAgo1, 3, and 4 contain inserted amino acids near the catalytic center called the "conserved Segment 7" (cS7). This segment hinders the correct positioning of scissile nucleotides in the active site, rendering these homologs inactive in RNase function (Nakanishi et al., 2013). Removal of the cS7 and converting the tetrad to the DEDH motif activates the RNase function of hAgo1 (Faehnle et al., 2013).

1.5.1.2 Nucleotide base pairing between guide and target at the seed

Among the ~22 nucleotides of the guide strand, the "seed" region is identified as the most important in conservational studies (Bartel, 2009; Lewis et al., 2003). Based on the number of matching nucleotides in the seed and overall complementarity with the target, the

target sites were categorized into three main types. The first type of sites is called the canonical sites (**Fig. 4A-C**). They perfectly base pair with the 7 nt of the seed. The canonical sites consist of three subtypes: 7mer-A1 sites match nt2-7 by Watson-Crick pairing and contain "A" residue at position 1; 7mer-m8 sites that match nt2-7 and position 8; and lastly, 8mer sites that match nt2-8 and contain A at position 1. The second main type of sites is called the marginal sites (**Fig. 4DE**). They consist of 6mer sites (nt2-7) and offset 6mer sites, which match at nt3-8. The third type of sites is called the atypical sites that constitute base pairs beyond the seed (**Fig. 4FG**). They are subdivided into 3'-supplementary sites and 3'-compensatory sites. The 3'-supplementary sites contain seed matching nucleotides 2-8 and supplementary 3-4 region base pairs (among nt12-17). The 3'-compensatory sites usually contain mismatched nucleotides with the seed and extensively paired nucleotides (more than 4 pairs) in among nt12-18. The third type is sometimes referred to as "non-canonical sites". Later, the Bartel group also identified centrally matched sites (Shin et al., 2010).

In addition to the sites that the Bartel group discovered, the Hannon group identified an alternative type of site that can be recognized and silenced with high efficiency (Chi et al., 2012). These sites contain G-bulges that are not complementary to the seed at nt5-6. They are predominantly present in miR-124 sites in the mouse brain. Due to the existence of the bulge, a pivot occurs in the seed region of the duplex, hence this rule is dubbed "pivot pairing rule" (Broughton et al., 2016). Recently, CDS-specific 3'-supplementary-pairing sites were also identified (Zhang et al., 2018).

Figure 4. Types of target sites of miRNA.

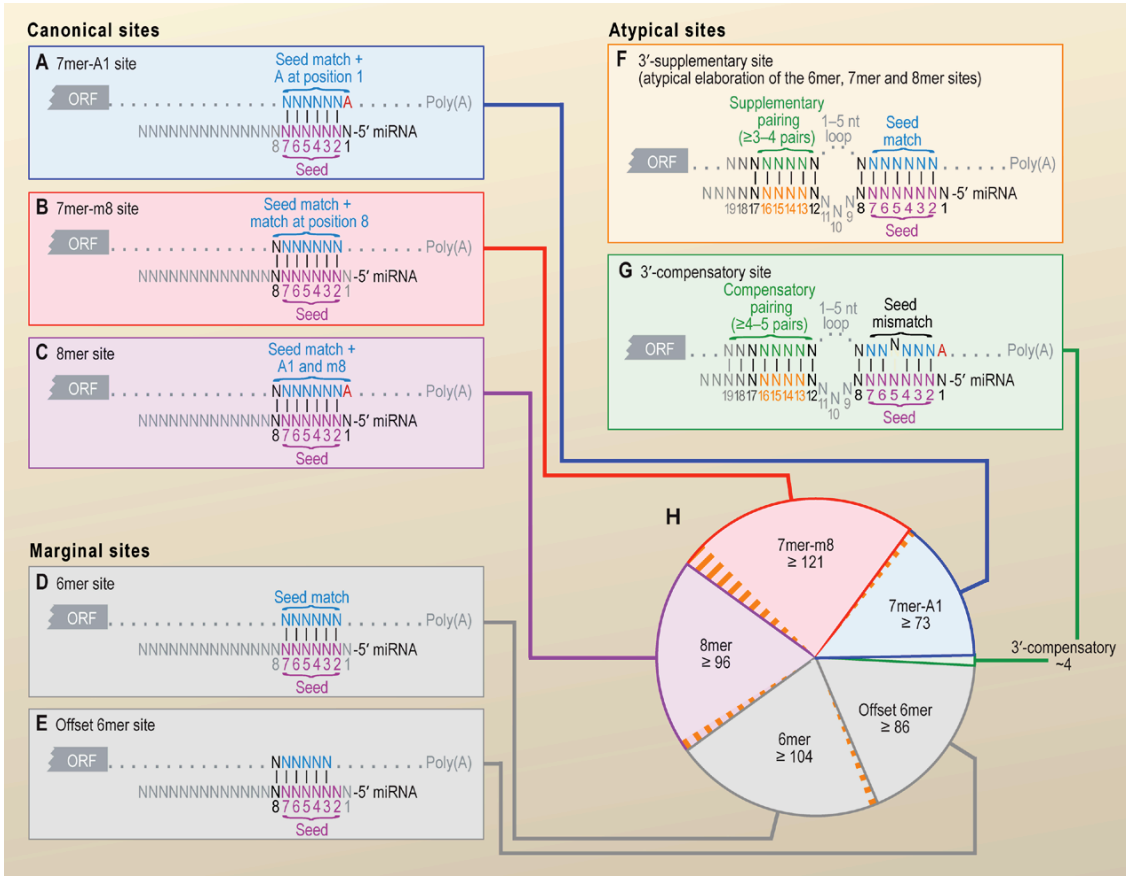


Figure 4. Types of target sites of miRNA. (A-C) Canonical sites with 7-8 nt seed match. Vertical lines indicate Watson-Crick pairing. (D-E) Marginal sites with 6 nt matching the seed. (F-G) Atypical sites: sites with productive 3' pairing. (F) 3'-supplementary sites. (G) 3'-compensatory sites. (H) Number of preferentially conserved mammalian sites matching a typical highly conserved miRNA (Friedman et al., 2008). For each site matching the seed region, orange-hatched subsectors indicate the fraction of conserved sites with preferentially conserved 3'-supplementary pairing. Figure adapted from Bartel's 2009 Cell Review (Bartel, 2009).

1.5.1.3 Nucleotide pairing beyond the seed

Plant miRNA usually base pair with the targets with high complementarity and mediate their cleavage (Bartel, 2004; Tang et al., 2003). Mammalian miRNAs, however, are rarely fully complementary to their targets beyond the seed (Doench and Sharp, 2004). As computational studies discovered, non-seed base pairing between nt 13-16 of the guide strand are important for silencing, especially when the complementarity of the seed was somewhat weakened (Friedman et al., 2009; Grimson et al., 2007). Substantial efforts were made to delineate the roles of base pairing beyond the seed in the hope to make precise predictions about miRNA targets based on sequence complementarity.

The first mutational studies were carried out at cell level using siRNAs. Tolerance for single nucleotide mutation 5'or 3' ends is tolerated in siRNAs against Tissue Factor (hTF) when they are delivered into human keratinocyte cell line (HeCaT) (Amarzguioui et al., 2003; Holen et al., 2002). Systematic alterations of guide or target nucleotides at specific positions were carried out *in vitro* and *in vivo*. They reached similar conclusions about differentiated contributions of non-seed nucleotides (Broderick et al., 2011; Deerberg et al., 2013; Du et al., 2005; Schwarz et al., 2003; Wee et al., 2012).

In one study, Du et al. performed a reporter assay after mutating each position of an siRNA to all three possible bases and tested their efficiencies (Du et al., 2005). They demonstrated that single nucleotide mismatches are generally tolerated beyond the seed and mRNA degradation can still be observed in most cases. Coherent with the analysis results from the Bartel lab (Grimson et al., 2007), single nucleotide mutations abolished repression the most in the seed region, followed by those made between nucleotides 13 and 17, though less in magnitude. Though the target construct consists of a target site that is located in the 5'UTR instead of the now commonly accepted 3'UTR, the pattern of effects on repression is largely consistent.

A few groups have suggested that base pairs in the 3' supplementary region contributes to the catalytic activity of Ago2's RNase; on the other hand, seed base pairs contribute mostly

to the affinity between the guide and the target (Ding et al., 2003; Haley and Zamore, 2004; Martinez and Tuschl, 2004). The Zamore group elucidated the specific contribution of each base pair by kinetic studies. Tiling the entire guide sequence with dinucleotide mismatches systematically, the cleavage efficiency of Ago2 was tested *in vitro*. They showed that mismatches in the 3' supplementary region perturbed predominantly k_{cat}, and to some extent the K_m, especially those between nt13 and 16 (Wee et al., 2012). The last three or four nucleotides, on the other hand, have no significant effects according to kinetic studies (Wee et al., 2012). In a single-molecule study, the Zamore group has shown that the seed binding contributes the most to the affinity between miRNA and its target. The measured K_m for seed-only binding sites is very close to that of the seed plus 3'-supplementary sites (Salomon et al., 2015). Several groups that performed kinetic studies overlooked the contribution of base pairs in the 3' supplementary region of the duplex (Brennecke et al., 2005; Elbashir et al., 2001c; Gu et al., 2014; Lin et al., 2005; Miranda et al., 2006; Saraiya et al., 2013); rather, most of them consider the RNase activity of RISC as a black and white phenomenon, switch-controlled by the central region base-pairs.

The importance of non-seed base pairing is also demonstrated in miRNA-inhibiting oligonucleotides, which comprise "miRNA-sponges" and "antagomirs" that work by complementarily binds to miRNA molecules and prevent their functions (Robertson et al., 2010). Some studies suggested subtle differences in cleavage efficiency with dinucleotide mismatches in the 3'supplementary regions; however, due to the experimental setup, the measurements were not performed at an optimal scale and most of the data points (>75%) were suppressed to a level that cannot be effectively discerned from the negative controls (Jo et al., 2015). In a recent study of *C. elegans* miRNA target sites, individual nucleotide resolution cross-linking immunoprecipitation (iCLIP) experiment on miRNA-mRNA chimeras identified 7 classes of non-seed pairings target sties, based on RNA hybrid predictions and *k*-means clustering (Broughton et al., 2016). Though this study confirms the roles of non-seed pairing in the determination of specificity of miRNA targeting, the classification was derived from genome-wide study and alignment-based, which is statistical in nature. Such

classification provides limited functional and mechanistic insights of miRNA-mediated silencing pathway. In all of the studies mentioned above, mismatches generally hampered the efficiency of guide-RNA mediated repression. To simplify, they are referred as “bad mismatches”.

Curiously, in one study demonstrated by the Sharp group using siRNAs, loops formed by mismatches at the center of the guide-target duplex are well tolerated (Doench et al., 2003). Centrally mismatched sites can elicit repression of target gene expression even when target RNA cleavage is compromised due to imperfect pairings (Doench et al., 2003). Based on this work, the classical model of miRNA targeting is depicted to consist of a “central mismatch”, which typically spans between nt 9 and 12 of the miRNA. This characteristic was further investigated by engineering mismatched loops of different sizes within the miRNA-target duplex in the non-seed regions. Using reporter assay loops of certain sizes and patterns were found optimal for miRNA-mediated repression. Incorporating “loop rules” in their own computational approach, the researcher group improved prediction of miRNA targets (Kiriakidou et al., 2004). Later, such centre-mismatched sites were taken as the common mode of repression for endogenous miRNAs (Martin et al., 2014). As the currently prevalent model for miRNA-mediated repression, centrally mismatched nucleotides were purposely designed in most studies that tend to investigate the silencing effects of miRNA.

A perplexing role of mismatched nucleotides was reported at the 3’ end of the guide, where introduction of mismatches enhanced silencing by facilitating the release of the target (De et al., 2013; Salomon et al., 2015; Tang et al., 2003; Wee et al., 2012). In this study, the observed enhancement only occurs when the rest of the guide perfectly matches the target, and cooperativity of repeated sites can be observed as the number of repeats increases. Moreover, the effect is more pronounced when the target reporter concentration was high and the effect diminishes as the target concentration decreases. Such effects were only observed for the last four nucleotides at the 3’ end of the guide RNA of 21nt in length. Since mismatched nucleotides in these above cases were found to be favourable for guide RNA-mediated silencing, they are collectively referred as “good mismatches”.

A few studies that systematically mutating nucleotides in siRNA sequences revealed more surprising facts about the "good" and "bad" mismatches. By profiling the efficiency of siRNAs of different sequences, base pairing at every third nucleotide positions was found to contribute more significantly than the others (Katoh and Suzuki, 2007). The Crooke lab reached a similar conclusion from experiments that use triplet mutations generated in siRNA sequences. Generating mismatches of three nucleotides, it was shown that the cleavage activity of Ago2 remains when positions 9-11 and 12-14 were mutated. However, mutating nt13 individually greatly compromised the cleavage activity (Lima et al., 2009). Though not explicitly indicated, the results from the Zamore lab well agree with these findings (Wee et al., 2012).

1.5.1.4 The nature of mismatched nucleotides

Initial studies on the effects of mismatched nucleotides were conducted using siRNAs. From an siRNA that perfectly matches its target sequence of a reporter mRNA, mutations can be made at specific positions and the new siRNAs were tested for their repression efficiency.

To systematically investigate how the nature of the nucleotides in a mismatched region affects repression efficiency, the Zamore group synthesized siRNAs that contain all four versions of the nucleotide at each position and assessed their repression efficiency on the target reporter gene expression (Schwarz et al., 2006). They observed that mismatches of different positions have different abolishing effects on repression. A mutation in the seed generally abolishes repression by 40-60%, while mutations that disrupt pairing with the 3' end of the miRNA do not have strong abolishing effects. In addition, the purine:purine and pyrimidine:pyrimidine mismatches are not as well tolerated as purine:pyrimidine mismatches at these positions (Fig. 5). Similar studies were carried out on different siRNAs. Notably, Du et al. mutated nucleotides at each position of the target site instead of in the siRNA. This way, they excluded the possibility that the observed effects are due to the difference in the level of mature miRNA copies. Though the positional effects slightly differ, the overall repression profile is similar in the way that purine mismatches are more deleterious to repression

efficiency than other types (Du et al., 2005). This is thought to be due to the fact that purine nucleotides are larger than pyrimidines in size.

Figure 5. The identity of mismatched nucleotides affects repression efficiency.

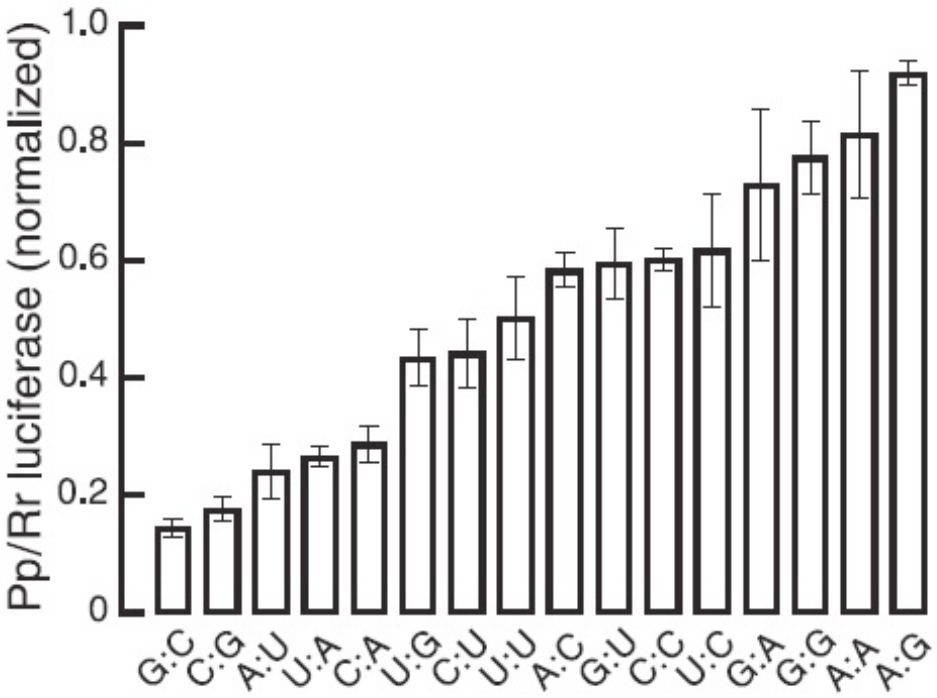


Figure 5. The identity of mismatched nucleotides affects repression efficiency. All possible single nucleotide pairs were examined for position 10 in siRNA. G:C complementary pair elicits more efficient silencing than an A:U. Purine:pyrimidine and pyrimidine:pyrimidine mismatches displayed intermediate levels of silencing. Silencing efficiency was compromised the most with purine:purine mismatches. Figure is adapted from Schwarz et al. (Schwarz et al., 2006).

1.5.2 Extrinsic factors

1.5.2.1 The stoichiometry of miRNA and targets

Since the "seed rule" has been taken as the predominant principle of miRNA target recognition, it is natural to speculate from a thermodynamic point of view that no substantial difference exists among the affinities of a miRNA for targets that share the same seed-binding sequence. Taking this idea further, it was postulated that under these premises, the stoichiometry of the miRNA and target RNA species plays an important role of determining which targets are preferentially regulated. This model is called the "competing endogenous RNA" (ceRNA) model and it was demonstrated in the regulation of PTEN gene regulation (Tay et al., 2011). In this study, the co-expression of the ceRNAs and PTEN was observed in agreement with predictions using stoichiometry computations. However, the ceRNA theory does not take the biochemical nature of the pathway into account and its correctness has been intensively questioned.

Recent studies suggest that the initial claim could have ignored a few significant factors. First, the available quantity of Argonaute proteins needs to take into consideration. Only Argonaute2 protein is slicer-active and other isoforms can only repress target via slicer-independent pathways The stoichiometry calculations, hence, also need to consider the Ago protein isoform levels regarding whether the repression is mainly through slicer-dependent or independent pathway (Flores et al., 2014). As the Zamore group demonstrated the roles of base pairing beyond the seed in slicer activity, the original assumption that only the "seed rule" is sufficient cannot be fully justified (Wee et al., 2012). In addition, the slicer-independent activity of miRNA, which was thought to rely more heavily on the stoichiometry of the miRNA and the target RNA, has been shown to be restricted by miRNA identity and RISC availability. *In vivo* study shows that only large changes in miRNA target concentration can detectably influence miRNA-mediated repression; therefore, ceRNA theory is unlikely to widely apply to most mRNAs under biological conditions (Denzler et al., 2014).

Consequently, the levels of miRNA and target RNA alone could not be used as sufficient criteria to predict miRNA targets based on ceRNA theory (Mayya and Duchaine, 2015).

1.5.2.2 Factors in the biogenesis pathway

As mentioned in the biogenesis section, correct processing of the pri-miRNA and pre-miRNA requires conserved nucleotide sequences flanking the stem part of the RNA duplex in the precursor molecules. Both the loop nucleotide sequence and structure were found to control the production of primary and mature miRNA (Yue et al., 2011). A thorough study was conducted by the Bartel group to identify effects of nucleotide changes on the processing of precursor miRNA. From their study, sequences of several pri-miRNA molecules were mutated in triplets that tile the stem region. Mutations were generated using all possible combinations of bases on both strands. Their analyses identified a mismatch motif in the basal stem region, a preference for maintaining base pairing in the rest of the stem, and a stringent stem-length requirement of 35 ± 1 bps (Fang and Bartel, 2015).

1.5.2.3 Thermodynamic environment of the target site

Seed base pairing is a energetically favourable process that brings guide and target RNA together; therefore, competing nucleotide species that bind to the target RNA can potentially hinder the access of miRNA guides. The AU-rich sequences were identified to be enriched in miRNA target sites in the conservational study using microarray (Grimson et al., 2007). This observation was confirmed by testing artificially engineered structures in *lin-41* target 3'UTR in *C. elegans*. Using a program called *sFold* (Ding et al., 2004), the mutant constructs were calculated to have lower ΔG, which makes them more stable, correspond to less efficient repression by *let-7* (Long et al., 2007). This principle was also shown by high-throughput shRNA library screens of anti-HIV sequences. Due to the unusual nature of the lentiviral genome, secondary structures are abundant in the HIV-1 RNA (Watts et al., 2009; Wilkinson et al., 2008). Designing shRNA molecules against HIV genome by tiling the entire genome with shRNA sequences of 22nt, one can obtain a map of RNAi-accessible target regions in the HIV-1 genome (Tan et al., 2012). Such accessible sites showed high correlation with the

structure-free regions identified in a technique called selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE), further confirming the inhibitory role of RNA secondary structures in guide RNA-mediated repression (Wilkinson et al., 2008). RNA structure software has been developed and made available to the public by our lab; moreover, our software is capable of constructing 3D structures of RNA from the 2D predictions, making further calculation and modeling feasible (Parisien and Major, 2008).

1.5.2.4 Site-cooperativity

When multiple target sites occur in the vicinity of each other, the bound Argonaute proteins exhibit cooperativity in their silencing efficiency. However, such cooperativity is not observed when the target site perfectly base-pair with the guide; rather, in the imperfectly paired sites, the cooperative effect is prominent. This is possibly because for a single perfectly matched site, the IC_{50} of the target RNA is already very low (IC_{50}=0.63±0.25 nM), indicating maximal silencing efficiency has already been reached without much room of improvement. On the other hand, single copies of bulged sites, seed plus nt13-16 sites, and seed-only sites all have IC_{50}>20 nM and when 6 sites are placed in close proximity, IC50's are lowered by >2-20 fold (Broderick et al., 2011).

1.5.2.5 Target site location

Even if the sites are identical in sequence, their location in the mRNA can affect the extent to which they are silenced by miRNA. In order to identify target sites that are effectively silenced by miRNA, high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation (HITS-CLIP) was carried out to map functional Ago2-RNA interactions in the mouse brain. By looking at the binding sites of miR-124 as well as other 20 major miRNAs, The Ago HITS-CLIP showed that 63% of the target sites are located in the 3'UTR while 37% in the coding region of the mRNA. Among the 3'UTR sites, a peak of sites was identified within 50nt downstream of the STOP codon. The number of sites decreases further downstream (toward the 3' end) and peaked again within 70nt of the start of the poly-A tail (Chi et al., 2009)..

In vitro and *in vivo* studies confirmed that target sites in the coding region cannot be effectively targeted by the miRNA machinery (Chi et al., 2009; Gu et al., 2009). Repression of the target is restored once a STOP codon is placed upstream of the target site and, in effect, moves the target site into the 3'UTR. The authors hence concluded that the incoming ribosome displaced the Ago2-miRNA complex from the mRNA during translation, and sites in the 3'UTR would be more readily accessible to RISC (Gu et al., 2009).

A novel type of CDS-specialized target site has been recently reported, challenging the classical view that translational state has decisive role on the accessibility of miRISC (Zhang et al., 2018). This finding supports one particular class of target sites identified by the Pasquinelli lab using Ago-iCLIP method in *C. elegans*. However, taking a closer look into this isolated report by Zhang et al., which generated and mined Ago-CLIP data, one finds that necessary positive controls are missing to confirm that such sites are sufficient and necessary to allow the RISC-mediated specific repression to occur. Moreover, evidence that directly links RISC with these sites is still absent. The evidence that the authors presented to support the involvement of RISC is another round of Ago-CLIP experiment, which is exactly the same technique with which they originally identified candidate sites. In a way, using Ago-CLIP again as the verification method may defeat the purpose of validating those candidate sites. Not surprisingly, the authors could only vaguely link the observed repression to a generally accepted slicer-independent pathway; yet its mechanistic details remain unclear, similar to many published work on "slicer-independent" repression. Though it represents a novel class of target sites, the number of identified target sites of this type is still fractional comparing 3'UTR sites. Ago-CLIP studies suggest that such coding region target sites accounts for 20% of the total target sites in *C. elegans* (Broughton et al., 2016) and about one-third in human (Chi et al., 2009). Hence, the number of sites that contain perfect 3'supplementary pairing without canonical seed pairing among the CDS sites is likely to be even less.

1.6 Computational approaches to study miRNA targets

The most intriguing claim about miRNA is their capability to regulate more than half of the coding genes (Bartel, 2004). This makes the prediction of their targets an attractive subject of research for the past 25 years. As mentioned above, miRISC-targeting specificity is mainly determined by sequence complementarity between the mRNA target site and nucleotides 2-8, termed the "seed". This simple "seed rule" formed the computational basis for search algorithms that identify miRNA target sites. However, the large number of candidate target genes represents a challenge to computationally discern the effective from the ineffective ones. Computer programs were developed to make such predictions at the genomic level.

To develop an algorithm of miRNA target prediction, patterns observed from experimental data are normally expressed as a set of rules and a computational model need to be built to apply the rules consistently. Computational model forms the core component of a prediction algorithm, which needs to be validated using data that are not included in the training set. A prediction program will implement the input and output interfaces for the core prediction algorithm, and store results in an appropriate format.

1.6.1 Classification of prediction programs

There are two ways to classify the available computer programs that predict miRNA targets. The first way is by the factors considered in the computation, as summarized in Table I (Saito and Saetrom, 2010). As listed in the table, each computational program considers a subset of the intrinsic and extrinsic factors listed in the table. The subset of factors can differ from program to program; however, they always use seed base pairing as predictor, except for Stanhope's program. The other two most commonly used predictors are site accessibility and conservation.

The second way to classify these programs is by computational principles that they use (Reyes-Herrera and Ficarra, 2012). Three categories of principles emerged: *Ab initio*, machine learning, and hybrid methods. *Ab initio* means "based on the first principle". Algorithm in this

category simulates the underline rules that govern the natural phenomenon and predicts the outcomes. To achieve this goal, the intrinsic and extrinsic factors were combined in its computational model to produce a predicted score. Many early programs, such as *TargetScan*, *PicTar*, and *Miranda*, belong to this category. On the other hand, machine learning algorithms use experimental measurements to train the computational model so that the model can be corrected to better mimic the outcomes, regardless its underline principles. Adding more training data to the learning improves the computational model by correcting errors in the previous predictions. Hybrid models combine both approaches so that the underlying principles can be implemented and further corrected by machine learning method using additional data collected. In Table I, asterisks next to the names of the programs indicate that machine learning methods or hybrid methods were used in its implementation.

Table I. List of tools used for miRNA target prediction by features that they include in computation.

Tool	Pair[a]	Site[b]	Consv[c]	Access[d]	Multi[e]	Expr[f]	Refs
TargetScan	y	x	x	y	y		(Friedman et al., 2009; Grimson et al., 2007; Lewis et al., 2005)
PicTar	x		y	x	x		(Chen and Rajewsky, 2006; Grun et al., 2005; Krek et al., 2005; Lall et al., 2006)
miRanda	x		y	x	y		(Betel et al., 2008; John et al., 2004)
MicroCosm Targets	x		y	x	y		(Enright et al., 2003; Griffiths-Jones et al., 2006; Griffiths-Jones et al., 2008)
RNAhybrid	x			x			(Kruger and Rehmsmeier, 2006; Rehmsmeier et al., 2004)
PITA	x		x	x	y		(Kertesz et al., 2007)
STarMir	x			x			(Long et al., 2007)
Rajewsky & Socci	x			x			(Rajewsky and Socci, 2004)
Robins	x			x	y		(Robins et al., 2005)
mirWIP	x		y	x	y	x	(Hammell et al., 2008)
MicroInspector	x			x			(Rusinov et al., 2005)
MicroTar	x			x			(Thadani and Tammi, 2006)
MirTarget2*	y	x	x	x			(Wang and El Naqa, 2008)
miTarget*	x			x			(Kim et al., 2006a)
TargetMiner*	x		y	x		x	(Bandyopadhyay and Mitra, 2009)
EIMMo	x		y		y		(Gaidatzis et al., 2007)
NbmiRTar*	x		y	x			(Yousef et al., 2007)
TargetBoost*	x						(Saetrom et al., 2005)
RNA22	x		y	x	x		(Miranda et al., 2006)
TargetRank	y		x	y			(Nielsen et al., 2007)
EMBL	x		y	x	y		(Brennecke et al., 2005; Stark et al., 2005; Stark et al., 2003)
MovingTarget	x		y	x	y		(Burgler and Macdonald, 2005)
DIANA-microT	x		y	x			(Kiriakidou et al., 2004)
HOCTAR	x		y	x		x	(Gennarino et al., 2009)
Stanhope						x	(Stanhope et al., 2009)
GenMiR++	y		y			x	(Huang et al., 2007b)
HuMiTar	x						(Ruan et al., 2008)
MirTif*	x						(Yang et al., 2008)
Yan et al.*	x		y	x			(Yan et al., 2007)
Xie et al.	y		y				(Xie et al., 2005)

Table I. List of tools used for miRNA target prediction by features that they include in computation (adapted form Saito and Saetrom) (Saito and Saetrom, 2010).

***** Programs implemented with machine learning methods. **a**. miRNA:mRNA pairing. x: stringent seeds, y: moderately stringent seeds, Blank: seed sites not considered. **b**. Site location. x: target positions considered, Blank: target positions not considered. **c.** Conservation. x: with/without conservation filter, y: with conservation filter, Blank: conservation not considered. **d.** Site accessibility. x: site accessibility with minimum free energy considered, y: A:U rich flanking considered, Blank: site accessibility not considered. **e.** Multiple sites. x: multiple sites considered, y: the number of putative sites considered, Blank: multiple co-operability not considered. **f.** Expression profile. x: expression profiles used, Blank: expression profiles not used.

1.6.1.1 *Ab initio* algorithms

The dominant and most widely accepted principle of miRNA targeting is the seed rule. It states that nt2-8 of the miRNA must be well complemented by the target sequence in order to be functional. Besides this rule, the nucleotide composition, possible presence of secondary structures, and sequence conservation of the target sites were also considered as determinants of targeting efficiency of microRNAs (Agarwal et al., 2015). Each principle was weighted differently when combined in the prediction program. The programs that are not annotated with asterisks in Table I belong to this category. Some examples of prediction programs of this category are listed as the following:

TargetScan (Friedman et al., 2009; Grimson et al., 2007; Lewis et al., 2003) is a program that requires the seed complementary at least for 6 nucleotides (Bartel, 2009). Moreover, it uses a context score that describes seed complementarity, conservation and AU content of the sites' surroundings to rank the target sites. In the recent release of the latest version of *TargetScan* (Garcia et al., 2011), additional determinants are incorporated. For instance, a multiple linear regression trained on 74 filtered datasets was used to integrate determinants such as seed-pairing stability (SPS) and target-site abundance (TA). *TargetScan* can be accessed online (http://www.targetscan.org/).

The program *miRanda* (Betel et al., 2008; John et al., 2004) aligns candidate target sequence with miRNA using a weighted dynamic programming algorithm. The predicted scores are calculated using a weighted sum based on matches, mismatches and G:U wobbles. In a more recent update, this program takes into account a conservation measure based on the *PhastCons* conservation score, in addition to its original seed complementarity and duplex free energy (http://www.microrna.org).

PITA (Kertesz et al., 2007) considers not only the specific duplex interaction information, but also takes the accessibility of the target site. Accessibility is defined as the difference between the minimum free energy of the guide-target RNA duplex and the energy of the target region mRNA in the absence of the guide strand, $\Delta\Delta G$. The user can impose

different restrictions to reduce the resultant set of candidates (minimum seed size, G:U bobbles and unpaired bases). The accessibility calculation feature can also be conveniently switched off using its online user interface (http://genie.weizmann.ac.il/pubs/mir07/).

PicTar (Lall et al., 2006) algorithm has strict requirements for the seed and it also considers the overall duplex stability based on free energy. Once the sites are aligned, the targets are ranked based on a score derived from a hidden Markov model that considers the site conservation (http://pictar.mdc-berlin.de/).

Other *Ab initio* algorithms include: DIANA (Maragkakis et al., 2009), RNA22 (Miranda et al., 2006), RNAhybrid (Kruger and Rehmsmeier, 2006), EiMMo (Gaidatzis et al., 2007), etc. Among the listed programs in Table I, only a few programs utilize expression profile for the prediction. In other words, existing algorithms largely ignored the context of competing RNA species. To address this shortcoming, the *miRBooking* program was developed and it includes cell-type-dependent RNA quantities in its calculations (Weill et al., 2015). Seed complementarity and concentration of the MREs were combined in computation, and predicted repression levels are calculated for each gene that contains the MRE. This enables the program to make predictions according endogenous RNA levels. Though *ab initio* programs are capable of including more factors by adding weighted terms in regression, unexpected factors often interfere with the accuracy of prediction and lead to errors. Moreover, most of the programs combined thermodynamic contribution at each base pair position in a linear fashion, assuming that they are additive. To address errors that may arise from these assumptions, machine learning methods were implemented to improve the rate of accuracy *ad hoc*.

1.6.1.2 Machine learning and hybrid algorithms

Machine learning methods emerged after the limitations were reached by the *ab initio* approaches. The importance of these methods has grown since the data with experimental support started to grow significantly. Representatives from this category are indicated by an asterisk in Table I. One example is *TargetBoost* (Saetrom et al., 2005), which uses a boosting

algorithm that assigns weights to sequence patterns of 30 nucleotides to allow the emergence of a strong learner. The training set consists of 300 randomly-generated sequences for negative results and 36 interactions with experimental support for positive results. Another example is *miTarget* (Kim et al., 2006a), which is an algorithm that uses support vector machine (SVM) to make target predictions. Structural, thermodynamic and positional features were combined to form the support vectors. Training was achieved on a negative set of 83 interactions with experimental support, 163 negative interactions inferred from experimental data, and 152 positive interactions with experimental support. As SVMs were demonstrated as an efficient way to build predictive models, Ensemble Algorithm (Yan et al., 2007) uses 10 SVMs (polynomial kernels) as a post-processing step for miRanda, The prediction is based on features from the miRNA-MRE interactions, combining features from the mRNA targets. The negative and positive datasets used for training consist of 16 and 48 experimentally-verified interactions, respectively. Other programs that uses machine learning methods include *MirTarget2* (Wang and El Naqa, 2008), *MiRTif* (Yang et al., 2008), *TargetMiner* (Bandyopadhyay and Mitra, 2009), *MTar* (Chandra et al., 2010), *TargetSpy* (Sturm et al., 2010), *mirSVR* (Betel et al., 2010), *miRror* (Friedman et al., 2010), *miREE* (Reyes-Herrera et al., 2011), etc.

In essence, the core algorithms of the programs in this category do not differ from the *ab initio* methods; the only difference is that machine learning methods added a means to perform self-correction based on additional training data. Several of them even make the use of existing *ab initio* algorithms as their core algorithms. The majority of the machine learning approaches is based on SVM, which combines the intrinsic and extrinsic factors into support vectors to build classifiers. The fact that kernels were often used means that features and experimental data need to be transformed into higher dimensions. Though this is a powerful feature of SVM in machine learning, it also indicates the difficulty of correlating experimental data and the potential danger of over-fitting.

1.6.2 Limitations of existing prediction algorithms

All popular prediction programs use guide-target RNA duplex stability as the major predictor of targeting efficiency (Gaidatzis et al., 2007; Lewis et al., 2003). Quickly, researchers realized that pure thermodynamics calculations of free energy of RNA duplex are not sufficient to accurately correlate with repression levels. To better correlate, seed nucleotides were given more weights than those beyond the seed. A linear combination, or vector-based methods of the same nature, was used to sum up the total contributions of base pairing (Agarwal et al., 2015; Friedman et al., 2009; Grimson et al., 2007; Grosswendt et al., 2014; Majoros et al., 2013).

Realizing that the base-complementarity approach alone was not sufficient to accurately predict miRNA targets, additional consideration of extrinsic factors were combined in computation by many recently developed programs. The results were subsequently corrected with respect to experimental data. This combinatorial approach demonstrated the importance of requirements such as target site accessibility and A/U context (Grimson et al., 2008; Kertesz et al., 2007). Other extrinsic factors included by various computational models are: preference of asymmetry in base composition during the processing and loading (Fellmann et al., 2011; Knott et al., 2014), location of target site (Chi et al., 2009; Gu et al., 2009), number of repeats and proximity of multiple target sites (Bartel, 2009; Doench and Sharp, 2004; Saetrom et al., 2007; Wu et al., 2010), accessibility of the target site (Kertesz et al., 2007), and concentration of competing endogenous RNA species (Ala et al., 2013; Ragan et al., 2011; Salmena et al., 2011; Tay et al., 2011; Weill et al., 2015). Incorporating these extrinsic factors into the prediction programs greatly increased the complexity of them while led to moderate improvements observable under specific criteria in large-scale studies (Grimson et al., 2007; Lewis et al., 2003). Unexplained discrepancies persist between predictions and experimental results, such as those obtained by HITS-CLIP (Grosswendt et al., 2014) and SILAC (Baek et al., 2008) despite drastic increase in complexity. Most of the currently available programs do not offer predicted efficiency in the output; rather, a ranked

list of predicted targets is displayed, indicating the challenge of establishing a precise quantitative correlation.

1.7 Development of RNAi strategy against HIV

The original biological function of RNAi was thought to defend against viral infections as the consequence of evolution. Engineering anti-viral RNA guides that take the advantage of the miRNA machinery to defend the host became a seemingly plausible strategy. Among the viruses, HIV genome as the target of RNAi has been well studied.

1.7.1 HIV is an RNA virus

The human immunodeficiency virus (HIV) is a RNA virus that requires integration into the human host genome in order to replicate. Upon membrane fusion with the target cell, the viral core, which contains the genome, and viral proteins such as reverse transcriptase, integrase, and protease, are released into the cytoplasm. Together forming a reverse transcription complex, the viral proteins synthesize a double-stranded DNA, which will to be transported into the nucleus for integration by microtubules through the nuclear pore (Brass et al., 2008). The viral genome encodes only 15 proteins (Frankel and Young, 1998) and it heavily relies on the endogenous proteins for its gene expression and genome replication (Goff, 2007). Current treatments for HIV-1 has greatly improved the prognosis of patients infected by HIV; however, the life-long antiretroviral regimen comes with disadvantages of toxicity and resistance, in addition to patient non-compliance and high cost (Rossi et al., 2007).

1.7.2 RNAi technology against HIV

The specificity, reversibility, and cost-effectiveness of the RNAi approach made it a potential strategy against HIV (ter Brake et al., 2009). Some endogenous miRNAs were shown to directly repress HIV genome (Ahluwalia et al., 2008); in addition, efficient shRNAs targeting the HIV genome were identified, such as the shRNA *miB* against the *tat* (Boden et al., 2003; Boden et al., 2004). To identify effective RNAi guide molecules, high throughput screens were carried out using shRNA or siRNA libraries targeting every segment of 21 nt tiling the viral genome (Tan et al., 2012) or against all cellular proteins (Brass et al., 2008). However,

the disadvantages of the shRNA approach were also revealed through previous studies. First, the repression level of an shRNA was shown to correlate with its complementary level to the target (Houzet et al., 2012). As the viral reverse transcriptase generates mutations in the viral genome at a rapid rate, the virus can quickly evolve to evade the suppression by mutations (Boden et al., 2003; Das et al., 2004; Lee et al., 2005; Sabariegos et al., 2006). In addition, substantial secondary structures are present in the HIV RNA genome and the efficiency of targeting by endogenous miRNA is believed to be low (Watts et al., 2009; Wilkinson et al., 2008), as the incoming HIV genome has been shown to be resistant RNAi (Westerhout et al., 2006). For these reasons, endogenous miRNAs from the mammalian genome are unlikely to mediate innate immunity against retroviral infection (Cullen et al., 2013).

A multi-targeting artificial miRNA strategy may address these pitfalls to repress HIV replication more effectively. Using an in-house program called *MultiTar*, we are able to design shRNA sequences that target multiple genes by complementarity with the target at the seed and nt13-16. Targeting both viral genes and cellular factors that are known to be essential to the virus simultaneously, it would be much more difficult for the virus to escape. Mutations are very unlikely to occur simultaneously at all target sites. Secondly, viral infection may be hindered by targeting the cellular factors. Once a viral particle releases its contents, the viral genome will be exposed to less cellular factor to assist its reverse transcription and transport to the nucleus. Thirdly, with only one shRNA, the off-targets will be much less than using multiple shRNAs and the cytotoxicity is expected to be much lower. The design software *MiRBooking* replaced the original *MultiTar*. *MiRBooking* was originally developed to be an miRNA target prediction software using seed complementarity and endogenous RNA concentrations as the main input. It has been adapted to design artificial miRNAs against HIV.

1.8 Rationale of the thesis: refocusing on non-seed base-pairing to gain mechanistic insights

We intend to design multi-targeting shRNAs that mimic the action of natural miRNAs to repress HIV. Due to the limitations in the current knowledge about targeting principles and the lack of accurate computational tools, choosing an efficient design approach became a challenge. Manually covering more possibilities during the design phase became the way to address unknown factors; however, without further knowledge in the targeting rules, the cost of design and testing would become difficult to manage. Further filtering was required to limit our search space. From the results obtained in the anti-HIV shRNA design project, better understanding about the targeting rules was achieved and the necessity to improve our knowledge of the targeting process became apparent.

Close examination of the way that base pairings were considered in prediction tools, it is clear now that a paradox exists. Mismatches in the duplex at certain positions (especially in the seed) were found unfavourable for silencing, while neutral or favourable at other positions (central bulge and the last 3 nucleotides). All conclusions about the "good" and "bad" mismatches, as described in Section 1.5.1.3, were drawn from experiments that tested individually mutated nucleotides or regions. What if we combine the favourable and non-favourable base pairs? It will be hard to deduce the outcome because two contradicting situations may occur. The first one is that an average of the favourable and unfavourable effects may occur, resulting in repression efficiency that is between those when individual mismatches occur in the duplex. The second possibility is that the resulting efficiency is worse than the duplex with either mismatch alone, due to an overall reduced affinity between the guide and the target. This paradox reveals that the mechanistic details were not fully addressed in the previous research of miRNA targeting process. As long as it is not explicitly and mechanistically addressed, sophisticated data fitting techniques would not be able to make substantial advances in prediction accuracy. Based on these thoughts, we refocused our attention to address the core mechanism of the targeting process.

CHPATER 2: APPLYING *MIRBOOKING* AS A DESIGN TOOL

Yifei Yan, Nicolas Scott, Roqaya Imane, Albert Feghaly, Etienne Gagnon, Gerardo Ferbeyre, and François Major

Efficient small artificial RNAs against HIV

Manuscript in preparation

2.1 Abstract

The human immunodeficiency virus (HIV) is a retrovirus with an RNA genome. RNA interference (RNAi) based strategies to inhibit its replication have been developed over the years. However, they were shown to be inefficient in maintaining target gene's repression. One of the difficulties resides in the stringent base pairing complementarity required by small hairpin RNAs (shRNA) behind the RNAi approach. In addition, the use of shRNAs is incompatible with the fast mutation rate and hard-to-reach targets of the HIV genome. Here, we addressed these issues by designing small artificial (smart) RNAs that can repress the expression of multiple genes simultaneously, mimicking the function of naturally occurring microRNAs (miRNAs). Using the *miRBooking* algorithm that we developed to predict microRNA interactions transcriptome-wide, we designed smart RNAs to target multiple predetermined HIV genes. Here, we demonstrate their efficiency to strongly inhibit the expression of these viral genes (>60%) and provide a robust protection against incoming viral particles, comparatively to a previously proven shRNA against the viral Tat protein. Moreover, in stably transduced cells, we show that more than half of tested smart RNAs protect against infecting viral particles and half of them also provide equivalent or stronger protection than the previously proven shRNA.

2.2 Introduction

The human immunodeficiency virus (HIV) is an RNA virus that requires integration into the human host genome to replicate. The predominant, earliest, and most commonly referred to virus is HIV-1, which accounts for around 95% of all HIV infections worldwide (Kiwanuka et al., 2008). Current treatments for HIV-1 have greatly improved the prognosis of patients infected by HIV. However, the life-long antiretroviral regimen comes with disadvantages of toxicity and resistance, in addition to patient non-compliance and high cost (Desai et al., 2012; Higaki et al., 2018; Rossi et al., 2007). With the emergence of RNA-interference (RNAi) technology, microRNAs (miRNAs) were shown to have the potential to regulate HIV viral replication by targeting either the viral RNA directly, or cellular factors that are required by the virus (Capodici et al., 2002; Coburn and Cullen, 2002; Klase et al., 2012; Lee et al., 2002a; Novina et al., 2002; Qin et al., 2003; Wang et al., 2000).

RNAi-mediated anti-viral therapeutics have identified a number of effective candidate small-interfering RNAs (siRNAs) targeting either viral or cellular genes (Brass et al., 2008; Dziuba et al., 2012; Espeseth et al., 2011; Konig et al., 2008; Liu et al., 2011; Yeung et al., 2009; Zhou et al., 2008b). Though the intersection of the precise identity of the cellular factors identified by these high-throughput screenings is small (only about 20%), their functions are very similar (Bushman et al., 2009). Among them, transcription factors (TF) dominate, and thus represent valuable targets for viral inhibition since viral transcription efficiency largely depends on the availability of these TFs from the host (Cusanovich et al., 2014).

Among the identified TFs, RelA (p65 subunit) is a member of the NF-κB protein family that activates over 100 genes involved in inflammation, immune and acute phase responses, as well as cell growth and differentiation (Baeuerle, 1991; Libermann and Baltimore, 1990). To promote its own gene replication and prevent apoptosis of the host cell, the HIV virus purposely activates NF-κB, either by the viral product Tat (Pahl, 1999), or by the process of viral and cellular membrane fusion (Hiscott, 2001). The endogenous NF-κB

(p65) then binds to the viral promoter, LTR, which contains two adjacent NF-κB binding sites, and activates viral gene transcription (Kretzschmar et al., 1992).

Mediator subunits belong to another class of TFs that activate viral transcription (Boyer et al., 1999; Fang et al., 2004; Gwack et al., 2003; Mittler et al., 2003; Yang et al., 2004). The Mediator complex was shown to be required for HIV infection and replication (Bushman et al., 2009; Fahey et al., 2011), as well as (along with TFIIH) to re-activate latent HIV-1 transcription with the stimulation by NF-κB (Kim et al., 2006b). In particular, MED7 enhances early HIV reverse transcription (Konig et al., 2008), and MED4, MED6, MED7, MED14, and MED28 are required for HIV infection (Brass et al., 2008). Other Mediator subunits were also linked to HIV replication and Tat-activated transcription (Zhou et al., 2008b). Cell signalling proteins such as c-MYC, cyclin T1, and Akt-1 represent other possible targets (Brass et al., 2008). Hence, targeting all or a subgroup of these TFs, Mediator subunits, and cell signalling proteins represents a potentially effective strategy against the HIV.

While the RNAi approach looks promising, it still has several limitations. The virus has evolved to evade from being targeted by endogenous miRNAs as the viral sequences is depleted in conserved miRNA targets (Boden et al., 2003; Das et al., 2004; Lee et al., 2005; Sabariegos et al., 2006). The likelihood that endogenous miRNAs could target the HIV is thus believed to be very low (Wilkinson et al., 2008), and the incoming HIV genome has been shown to be resistant to RNAi (Westerhout et al., 2006). Attempts to overcome these limitations, for instance using combinatorial RNAi approaches targeting multiple sites in the viral RNA and cellular factors, have shown improvements in the targeting efficiency (Anderson et al., 2009; ter Brake et al., 2006; ter Brake et al., 2009). However, multiple RNAi at high dosage, at which the combinatorial approach elicits inhibitory effect, are often cytotoxic due to the knockdown, through off-target effects, of functionally important genes (Fedorov et al., 2006). Further, miRNAs in the mammalian genome are not triggering the necessary innate immunity to refrain retroviral infection (Cullen et al., 2013). The multiple-targeting approach in RNAi against HIV made little progress in the past decade.

To address these limitations, we devised a multi-targeting approach using an engineered shRNA-expression construct (Boden et al., 2004; Stegmeier et al., 2005) test computer-aided designs of guide RNA molecules with optimized on-target and controlled off-target effects. Mammalian miRNAs rarely base pairs with targets extensively, but are capable of downregulating multiple targets simultaneously. *MultiTar* was a computer program that we previously developed to construct effective small artificial (smart) RNAs targeting multiple messenger RNAs (mRNAs) simultaneously (De Guire et al., 2010). The idea here was to carefully design and coordinate the repression of both viral and cellular factors by using a single smart RNA with limited and controlled off-targets to avoid viral escape and cytotoxicity problems. Since RNA stoichiometry and genome-wide silencing were not considered in *MultiTar*, we implemented an improved algorithm, *miRBooking* (Weill et al., 2015), which now allows us to reduce unwanted off-target effects. It requires the input of the concentrations of all RNA species of a particular cell type as input (Houzet et al., 2012; Wee et al., 2012).

The mRNA targets that are higher in concentration are generally more likely to be targeted by miRNAs, and, similarly, the miRNAs that are present at higher concentrations are more likely to be active and repress their targets than those in lower concentrations (Mukherji et al., 2011). The interplay between miRNA and mRNA target concentrations is simulated by *miRBooking*, which implements the Gale-Shapley algorithm for stable matching (Gale and Shapley, 1962). Using a series of *miRBooking* invocations, we developed a strategy, *mirDesign*, to identify efficient and specific smart RNA sequences for a given set of targets and cell type.

Using *mirDesign*, we engineered anti-HIV smart RNAs, which we tested upon their ability to inhibit HIV propagation. The effect of the smart RNAs on the endogenous targets was quantified to confirm their knockdown efficiencies. We found that some smart RNAs can inhibit viral gene expression under transient conditions almost as efficiently as a previously proposed shRNA, miB, which targets the viral *tat* gene (Boden et al., 2004). Moreover, the smart RNAs were able to confer stronger resistance against viral infection than miB. Our

design pipeline hence offers a means to enhance the use of the RNAi approach against the HIV.

2.3 Results

2.3.1 Using *mirDesign* for Smart RNA design and selection

Previous shRNA library screens have identified endogenous genes that help HIV propagation (Brass et al., 2008). Among these required genes, ten were chosen as our targets, in addition to the HIV-1 mRNA, design artificial miRNAs (smart RNAs). Target genes are listed in the first row of Table 1. Seed complementarity and RNA concentrations were considered by *mirDesign* to predict the number of binding sites ("Disturbance") and repression efficiency in "Fold Change" of target expression (See Supplementary Information for definitions). Since seed pairing is the prevalent feature that determines miRNA's ability to recognize the target, *mirDesign* only predicts target sites and repression efficiency basing on seed base pairing. As the result, three categories of seeds are designed: the "best fold change" class consists of seeds with highest overall fold change values predicted (see Materials and Methods); the "low disturbance" category consists of seeds with the lowest total number of binding sites in all targets, and the "high disturbance" category comprises seeds with the highest total number of binding sites in all targets. The top five seeds in each category were selected for engineering and experimental validation (Table 1).

With each seed sequence, we designed three guide sequences by appending to it three different non-seed sequences ("tails") designed to yield the following classes of guides: "HIV full-complement", "artificial miRNA", and "scramble" guides. In the first design approach, the tails were designed to perfectly complement an optimal HIV target site. For this class of guides, complementarity with cellular targets were not optimized (Table S1, Seed 1-5). In the second approach, tails were optimized to complement all desired targets of each seed using algorithms implemented in *mirDesign*. Appending these tails to the seeds, we obtained the class of artificial miRNAs against HIV (Table S1, Seed 6-10). To control for the effects of the

tails, random sequences were appended to the seeds to generate the "scramble" class of guides (Table S1, Seed11-15).

Two representative seeds from each category were selected (best fold change, high, or low disturbance) and tested by experiment (Table S1, asterisks). For the "best fold change" category, seed 1 and 5 correspond to the highest and lowest predicted efficiency against HIV among this category (Table 1). We chose these two to verify the ability of *mirDesign* to differentiate repression levels. For the "low disturbance" category, we chose seed 7 and 9 since they have the lowest disturbance values among this category (Table 1). Seed 7 ties with seed 6 for disturbance; but seed 7 has better predicted efficiency on HIV (Table 1) and was hence chosen for testing. Substantial non-specific targets are predicted for the seeds in the "high disturbance" category. In order to confirm their efficiency against cellular targets, seed 14 and 15 were selected for testing since they were predicted with the highest efficiency against one specific endogenous protein, RelA, among all protein targets (Table 1). As the result, 18 guide sequences with seed 1, 5, 7, 9, 14, and 15 were selected for further testing. In addition, a negative control guide that has the least seed-binding probability for all endogenous genes, according to *mirDesign* calculations, was designed. We included one additional non-targeting negative control, shRNA (NCS), which is the shFF3 construct that came with the pPRIME vector, and it was originally used in the Elledge lab (Stegmeier et al., 2005). A total of 20 guide sequences were cloned for experiments (Table 2). We pre-fixed the IDs of these designs as *SM* (Small Multi-targeting RNAs) and they were numbered in the following way: SM1-6 are the "HIV full-complements", SM7-10 are "artificial miRNAs", SM16-21 are "scramble" guides, and SM23 is the "non-targeting control".

Table II. *MiRBooking* predicted repression effects on each target gene

	Best fold change													
Seed ID	Seed	Disturbance	HIV1_pNL	RELA	MED6	Cyclin_T1	MED4_var2	MED7	AKT1	JAK1	MED28	MED14	MED4_var1	Best of 5	Max FC
1	GGUCCCC	381	0.92	0.665	1	0.74	0.486	1	0.911	0.966	1	0.977	0.588	0.833	0.92
2	CCCAUCU	131	0.923	0.908	1	0.678	0.593	0.88	0.804	0.983	0.634	1	1	0.848	0.923
3	UUCCCCU	139	0.923	0.795	1	0.975	0.581	0.44	0.911	0.902	1	1	1	0.86	0.923
4	UCUUUCC	50	0.935	0.795	0.813	0.928	0.987	0.861	0.781	0.872	0.783	1	1	0.882	0.935*
5	CCUCUGU	136	0.936	0.486	0.931	0.821	1	0.821	0.715	0.787	0.913	1	1	0.847	0.936*

	Low disturbance													
Seed ID	Seed	Disturbance	HIV1_pNL	RELA	MED6	Cyclin_T1	MED4_var2	MED7	AKT1	JAK1	MED28	MED14	MED4_var1	Best of 5	Max FC
6	UUCCUUU	20	0.963	0.874	0.882	0.824	0.75	1	0.974	0.73	0.937	1	1	0.897	0.963
7	UUUUCCU	20	0.954	0.906	0.965	0.975	0.739	1	0.954	0.899	1	1	1	0.944	0.965*
8	CUUUUCU	30	0.966	0.86	0.763	0.728	0.618	0.856	0.966	0.968	0.802	1	1	0.856	0.966
9	UACUUCU	11	0.967	0.971	0.887	0.951	0.825	0.935	0.929	0.904	0.978	1	1	0.938	0.967*
10	CUGCACU	49	0.968	0.786	0.993	0.851	0.959	1	0.993	0.9	0.943	1	1	0.943	0.968

	High disturbance													
Seed ID	Seed	Disturbance	HIV1_pNL	RELA	MED6	Cyclin_T1	MED4_var2	MED7	AKT1	JAK1	MED28	MED14	MED4_var1	Best of 5	Max FC
11	CCCUGCG	340	0.975	0.912	1	0.981	0.893	1	0.891	0.924	0.986	0.951	0.905	0.944	0.975
12	UGGUCCC	307	0.976	0.432	0.982	0.962	0.965	1	0.712	0.987	1	0.966	0.954	0.896	0.976
13	GCUCGCC	354	0.978	0.732	0.967	0.749	1	1	0.616	0.729	1	0.771	1	0.856	0.978
14	GUCCCGC	339	0.98	0.239	1	1	0.506	1	0.962	0.97	1	1	0.607	0.829	0.98*
15	UCCCGCU	328	0.976	0.379	1	0.982	0.859	1	0.644	1	0.98	0.99	0.901	0.873	0.98*

The seeds are divided into three categories: Best fold change, and Low and High disturbance. The predicted repression from miRBooking simulations are given for the HIV genome (pNL reporter) and ten host mRNA targets (including two variants of MED4). Two seeds per category were selected for engineering (indicated in bold).

Table III. *MirBooking* designed guide sequences

Three categories of seeds, Best Fold Change, and Low and High Disturbance, were considered. Each category follows three different design for the sequence beyond the seed: HIV full-complement, small artificial RNA, and scramble. Asterisks (*) marks the six seed groups that were tested.

Seed	Seed ID	Category	HIV full complement	Artificial miRNA	Non-seed nt Scramble
GGUCCCC	1	Best Fold Change	UGGUCCCCCCACUCCCUGACAU	UGGUCCCCUGGAACCCAGGUCA	UGGUCCCCCGUCACAUAAACCC
CCCAUCU	2	Best Fold Change	UCCCAUCUCUCUCCUUCUAGCC	UCCCAUCUCCUGAUUCUCAGGC	UCCCAUCUACUAUAUUCGCCAC
UUCCCCU	3	Best Fold Change	UUUCCCCUUGGUUCUCUCAUCU	UUUCCCCUGUAUUGUCCCCUUC	UUUCCCCUGAAGCCUCCUACAU
UCUUUCC*	4	Best Fold Change	UUCUUUCCCCCUGGCCUUAACC	UUCUUUCCGCUGUAAUUCAUUC	UUCUUUCCUCCUCACCCUCCGU
CCUCUGU*	5	Best Fold Change	UCCUCUGUUAAUUGUUUCACAU	UCCUCUGUCCUUUGCUCACGUU	UCCUCUGUCGCCCUCCUACCCU
UUCCUUU	6	Low Disturbance	UUUCCUUUGGUCCUUGUCUUAU	UUUCCUUUGGAUCAAACUACAG	UUUCCUUUCCCCGCCGCGCAAG
UUUUCCU*	7	Low Disturbance	UUUUUCCUAGGGGCCCUGCAAU	UUUUUCCUUUUCAACAGAAACA	UUUUUCCUAACACCCCAUGUGA
CUUUUCU	8	Low Disturbance	UCUUUUCUGGCAGCACUAUAGG	UCUUUUCUACACACCACCGCGG	UCUUUUCUGGACAUUUCCCCCA
UACUUCU*	9	Low Disturbance	UUACUUCUGGGCUGAAAGCCUU	UUACUUCUGGAUACACUGAUCA	UUACUUCUGAACUCGGCUGUUA
CUGCACU	10	Low Disturbance	UCUGCACUAUAGGGUAAUUUUG	UCUGCACUUUGCUCACGUUGGC	UCUGCACUUCCCUUCCUCCAUG
CCCUGCG	11	High Disturbance	UCCCUGCGUCCCAGAAGUUCCA	UCCCUGCGACUGUUAAAGAAAC	UCCCUGCGAUUCUCUUUCCUCU
UGGUCCC	12	High Disturbance	UUGGUCCCAGUGCUUUUAAAAU	UUGGUCCCCUGGAACCCAGGUC	UUGGUCCCAUUCUGCCCUACAC
GCUCGCC	13	High Disturbance	UGCUCGCCACUCCCCAGUCCCG	UGCUCGCCCAUCUCCUGUUUCC	UGCUCGCCCUAGCAUUCCCUCA
GUCCCGC*	14	High Disturbance	UGUCCCGCCCAGGCCACGCCUC	UGUCCCGCCACCGGCGCACGCG	UGUCCCGCCCCCUACACGUACG
UCCCGCU*	15	High Disturbance	UUCCCGCUACUACUAUUGGUAU	UUCCCGCUACCGGCGCACGCGC	UUCCCGCUCGUUCGUCGCGACA

2.3.2 Optimization of the renilla luciferase construct for dual luciferase assay

We used pNL4.3R-E-luc plasmid (a gracious gift of the Cohen lab) containing the HIV-1 genome as the target reporter in our assay. This plasmid is based on the proviral clone pNL4-3 and it contains a firefly luciferase reporter in the Nef gene near the 3' end (**Fig. 6A**). The Env gene is mutated due to a frameshift near its 5'-end. Another frameshift was made in the Vpr gene to prevent its replication (Connor et al., 1995). To control for the quantity of DNA transfected, a renilla luciferase construct is co-transfected. The ratio of firefly (FF) to renilla light intensity ratio was used as the measure of the HIV reporter expression level. We originally used a CMV-RlucII plasmid (a gracious gift of the Mader lab), which contains a renilla luciferase (**Fig. 6B**), as a transfection control. However, the renilla luciferase was found to be co-activated by the HIV-containing plasmid more than 20 fold **(Fig. 6C**). This makes the CMV-Rluc plasmid unsuitable as a control plasmid.

We replaced the CMV promoter in the renilla construct with either simian virus 40 (SV40) or thymidine kinase (TK) promoter, making an SVR or TKR renilla reporter construct, respectively (**Fig. 6D and E**). Using either new renilla construct, the observed potent activation of the renilla gene in the presence of the pNL plasmid was abolished (**Fig. 6F**). We suspected that the co-activation is due to the interaction between the *tat* protein produced from the pNL4-3 plasmid and the CMV promoter of the renilla gene. To test that hypothesis, we modified the pNL4-3-luc R-E- vector by removing the tat gene using restriction endonucleases followed by ligation. The resulting vector, pNL tat-, was co-transfected with the CMV-Rluc plasmid and the co-activation of renilla gene was abolished (**Fig. 6F**).

To be sure that the two new renilla constructs, SVR and TKR, can be used as transfection controls, we co-transfected either one of them with the pNL vector. We observed a linear relationship between the in FF/ren raito and the amount of pNL vector transfected (**Fig. 6G and H**). This indicates that our control renilla plasmid can be used to quantitatively monitor the reporter expression.

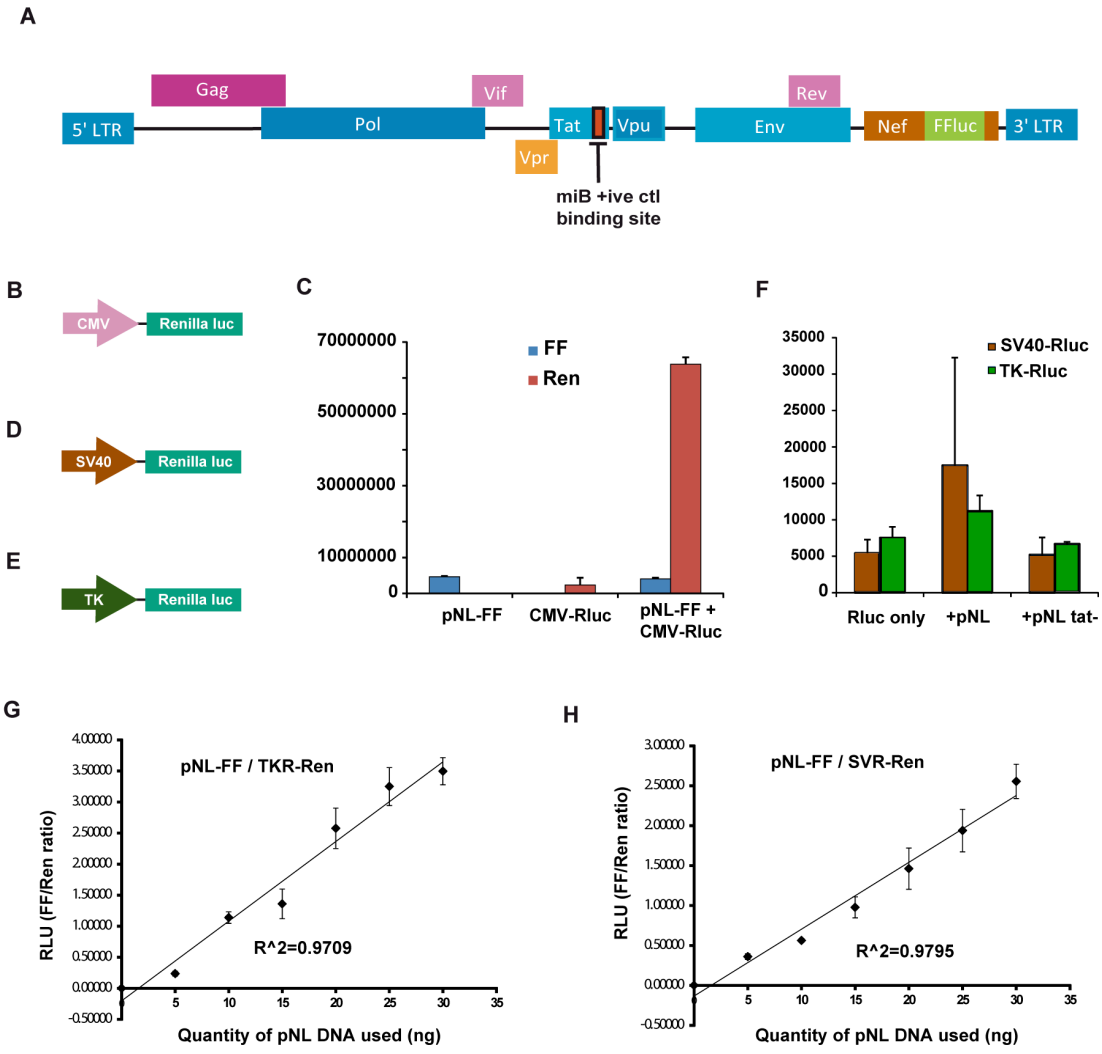


Figure 6. Testing and selection of the transfection controls for dual luciferase assay.

Figure 6. Testing and selection of the transfection control for dual luciferase assay.

(A) The pNL4.3 R-E-luc construct as a HIV-1 target reporter. The firefly luciferase gene is inserted in the Nef gene.

(B) The pcDNA3-CMV-RLuc II construct as originally used as a transfection control for the dual luciferase assay.

(C) The CMV-Rluc is activated by the pNL4.3 R-E- luc plasmid and resulted in increased renilla luciferase assay.

(D) The SVR construct that contains a SV40 promoter instead of the CMV promoter.

(E) The TKR construct has a TK-promoter that replaces the CMV promoter.

(F) Renilla luciferase light intensity of SVR and TKR when the pNL construct in present. Both renilla constructs showed much less activation than the CMV-RlucII. No activation was observed when the tat gene is removed from the pNL plasmid (pNL tat-).

(G) Co-transfection of the TKR construct with different amount of pNL4.3 R-E-luc construct shows a linear relationship between the FF/Ren ratio and the amount of pNL reporter used.

(H) Co-transfection of SVR with different amount of pNL4.3R-E-luc shows a good linear relation.

2.3.3 *mirDesign* smart RNAs inhibit HIV gene expression

We co-transfected pNL4.3 R-E-luc plasmid and mirDesign-designed shRNA constructs to establish a quantitative assay (**Fig. 7A**). The pPRIME vector was used to deliver the small RNAs into the cells in the form of shRNAs in the miR-30 backbone (**Fig. 7B**). We used the pPRIME empty vector and two previously mentioned non-targeting shRNAs as negative controls. We used miB, which is an shRNA that potently inhibits HIV tat gene (Boden et al., 2004), as the positive control. A second positive control that we used was a mutated version of the miB shRNA. Four nucleotides at the 3' end of miB were mutated to mismatch the tat target gene of HIV. Mismatches at the 3' end of a shRNA were known to be tolerated with moderately reduced efficiency of repression. Using this positive control enables us to test whether our assay is precise enough to identify subtle differences in the efficiency of the guide RNAs. Three smart RNA sequences did not propagate well or resulted in frequent mutations during cloning and were dropped from testing (asterisks in Table IV).

We co-transfected three quantity combinations of pNL, renilla, and pPRIME-guide constructs into the cells, and then we performed the reporter assay. The ratio between shRNA and the pNL reporter was fixed at 2:1. We found that the combination of 5 ng of pNL, 4 ng of renilla, and 10 ng of the shRNA construct gave the largest difference between positive and negative controls with the lowest percentage error (**Fig. 7C**). This combination was used for subsequent assays.

We performed at least five individual repeats of reporter assays (each in technical triplicates or quadruplicates) using shRNAs that are fully complementary to the HIV RNA sequence. Though SM2, 5, and 6 showed repressive activity comparing to the negative controls, only SM5's repression is statistically significant (**Fig. 7D**, p-value obtained by two-tailed Student's t-test assuming unequal variance). With guides that are partially complementary to HIV (**Fig. 7E**), the mean reporter expression levels were reduced for SM10, 12, and 13; however, the most statistically significant one, SM12, has a p-value above 0.1. Further, we noticed that

the SM12 shares the same seed sequence as SM5, indicating that the tail complementarity is also important in for repression. This conclusion was further confirmed by the observation that none of the guide sequences that containing random tail sequence showed better effect than the negative controls (**Fig. 7F**). All partial complementary guides were tested at least three times in triplicates or quadruplicates except for SM9, which was tested twice in quadruplicates.

To confirm that the knockdown that we observed was against the expression from the viral genome and not the luciferase or the basal functions of the cell, a parallel experiment was performed using a firefly luciferase reporter vector (pGL3 reporter from Promega) that does not contain the viral genome sequence. We selected the HIV-complementary shRNAs and the best partial complementary shRNA, SM12, in the test. With a repeat, we confirmed that the inhibitory effect by SM5 and SM12 were the only significant ones among all the guide sequences tested (**Fig. 7G**).

Table IV. Selected guide RNA sequences against HIV to be tested.

Seed ID	*Seed sequence*	*Seed Category*	*Guide RNA sequence*	*Guide RNA ID*
1	GGUCCCC	Best Fold Change	UGGUCCCCCCACUCCCUGACAU	SM1*
			UGGUCCCCUGGAACCCAGGUCA	SM7*
			CGGUCCCCCGUCACAUAAACCC	SM16
5	CCUCUGU	Best Fold Change	UCCUCUGUUAAUUGUUUCACAU	SM2
			UCCUCUGUCCUUUGCUCACGUU	SM8
			UCCUCUGUCGCCCUCCUACCCU	SM17*
7	UUUUCCU	Low Disturbance	UUUUUCCUAGGGGCCCUGCAAU	SM3
			UUUUUCCUUUUCAACAGAAACA	SM9
			UUUUUCCUAACACCCCAUGUGA	SM18
9	UACUUCU	Low Disturbance	UUACUUCUGGGCUGAAAGCCUU	SM4
			UUACUUCUGGAUACACUGAUCA	SMM10
			UUACUUCUGAACUCGGCUGUUA	SM19
14	GUCCCGC	High Disturbance	UGUCCCGCCCAGGCCACGCCUC	SM5
			UGUCCCGCCACCGGCGCACGCG	SM12
			UGUCCCGCCCCCUACACGUACG	SM20
15	UCCCGCU	High Disturbance	UUCCCGCUACUACUAUUGGUAU	SM6
			UUCCCGCUACCGGCGCACGCGC	SM13
			UUCCCGCUCGUUCGUCGCGACA	SM21
52	CGUAUGC	Control	UACGUAUGCCCUACCUAACUCU	SM23

*Mutations occurred during the cloning of these shRNAs and they were subsequently dropped from testing. F.C.: full-complementary to HIV sequence; A.M.: artificial miRNAs. Scr.: scramble tail sequences.

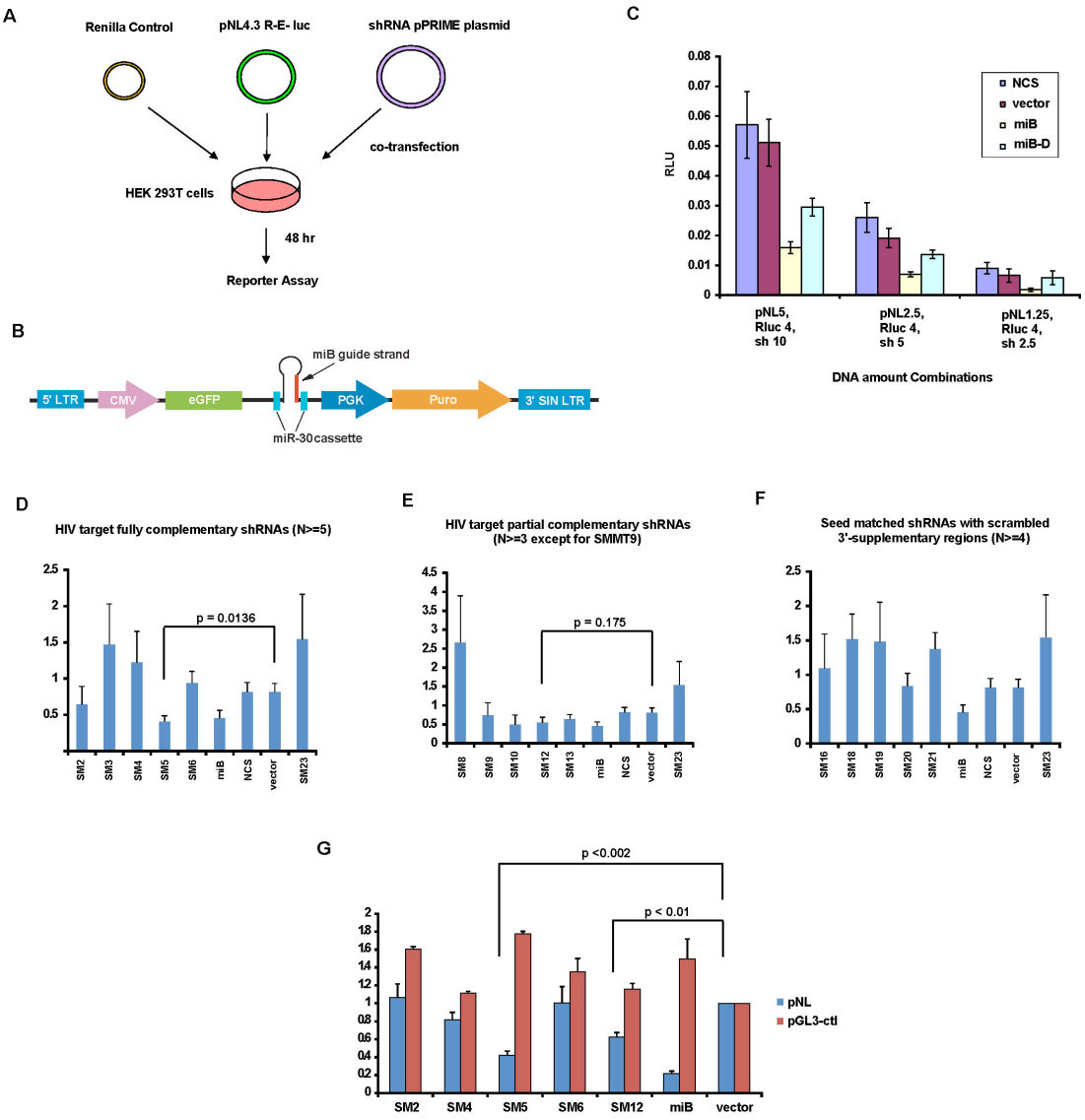


Figure 7. Reporter assay identifies RNA guides that inhibit HIV gene expression.

Figure 7. **Reporter assay identifies RNA guides that inhibit HIV gene expression**.

(A) Luciferase reporter assay following the co-transfection of HIV and shRNA.

(B) The pPRIME construct is used to deliver designed guide RNAs.

(C) Optimization of the amount of plasmid DNA to be co-transfected.

(D) Reporter levels using shRNA of the first design approach: RNA guides are fully complementary to HIV but partially complementary to cellular targets in the 3'-supplementary regions. The repeat numbers were indicted on top of the panel (D-F). All experiments were conducted in technical triplicates or quadruplicates.

(E) HIV reporter levels using shRNAs of the second design approach: guides are partially complementary to the 3'-supplementary regions of HIV and the cellular RNA target sequences. Complementarity with the cellular targets was optimized in the design.

(F) HIV reporter levels using the third design approach: only the seed matches the HIV and cellular target sites. A randomly generated sequence is used for the rest of the RNA guide.

(G) A counter-screen of *MiRBooking* designed shRNAs shows that they are not targeting the luciferase gene. Error bars represent standard deviation (N=2 in triplicates).

2.3.4 Protective effect against viral infection in transiently transduced cells

To test whether the designed guides provides cells with protection against viral infection, we packaged the pNL4.3-R-E-luc genome with coat proteins by co-transfecting it with a VSV-G plasmid that contains the Env gene and the RRE plasmid into the HEK293T cells. The target cells were transduced with pPRIME vector to express the guide 24 hours before infection. With the collected viral sup, we infected the transduced target cells and measured luciferase activity 48 hours post-infection (**Fig. 8A**).

In order to establish a quantitative assay, we titrated the viral supernatant to identify the range of quantities in which the reporter expression level increases linearly with respect to it (**Fig. 8B**). We found that when the viral supernatant used is less than 100 uL per 24-well, there is a linear relationship between the amount of the supernatant used and the light intensity. We tested the two consistently identified guide RNAs, SM5 and 12, using 10, 20, and 100 μL of the viral supernatant.

At all volumes of viral supernatant used, the level of inhibition by *mirDesign*-designed guide RNAs was stronger than the anti-tat shRNA positive control, miB. At the highest volume of 100 μL, miB-transduced cells were not resistant to incoming viral particles while SM5 and 12 still provided significant protection (**Fig. 8C**).

To see whether the predicted cellular target was successfully knocked down along with the viral genes, we performed Western blot to detect the Rela (P65) and Akt1, as well as the P24 viral capsid protein levels in the stably transduced cell lines (**Fig. 8D**). As comparisons, we extracted protein from SM6 and 13 expressing cells lines for Western blot because they were also predicted to target Rela protein more potently than other shRNAs. As the result, the viral P24 protein was reduced in agreement with the luciferase readings in the co-transfected cells. Moreover, SM6 significantly knocked down the viral protein, which was not significantly detected by previous luciferase assay. However, the differences in expression of Rela and Akt1 proteins could not be clearly discerned by analyzing the immunoblots (**Fig.**

8E). More distinguishable effects were detected at the RNA level when we performed RT-qPCR experiments. The pNL RNA levels were in agreement with the Western blot SM5, 6, 12, and 13. However, SM5 and 12 had moderate or no effect on Rela and Akt1 RNA; in contrast, SM6 and 13 knocked them down more efficiently. The same is true for Jak1 and Med14 target genes: the seed sequences that provided the most protection against viral infection are not the ones that most efficiently knocked down predicted targets. SM5 and 6, which are fully complementary to the HIV sequences and knocked down the HIV RNA more efficiently than their artificial miRNA designs (SM12 and 13), also knocked down the Rela and Akt1 RNA more efficiently. We hence conclude that the inhibitory effects in transiently transduced cells are mostly determined by the extent of complementarity with the HIV RNA. We could not clearly observe the advantage of simultaneously targeting endogenous genes.

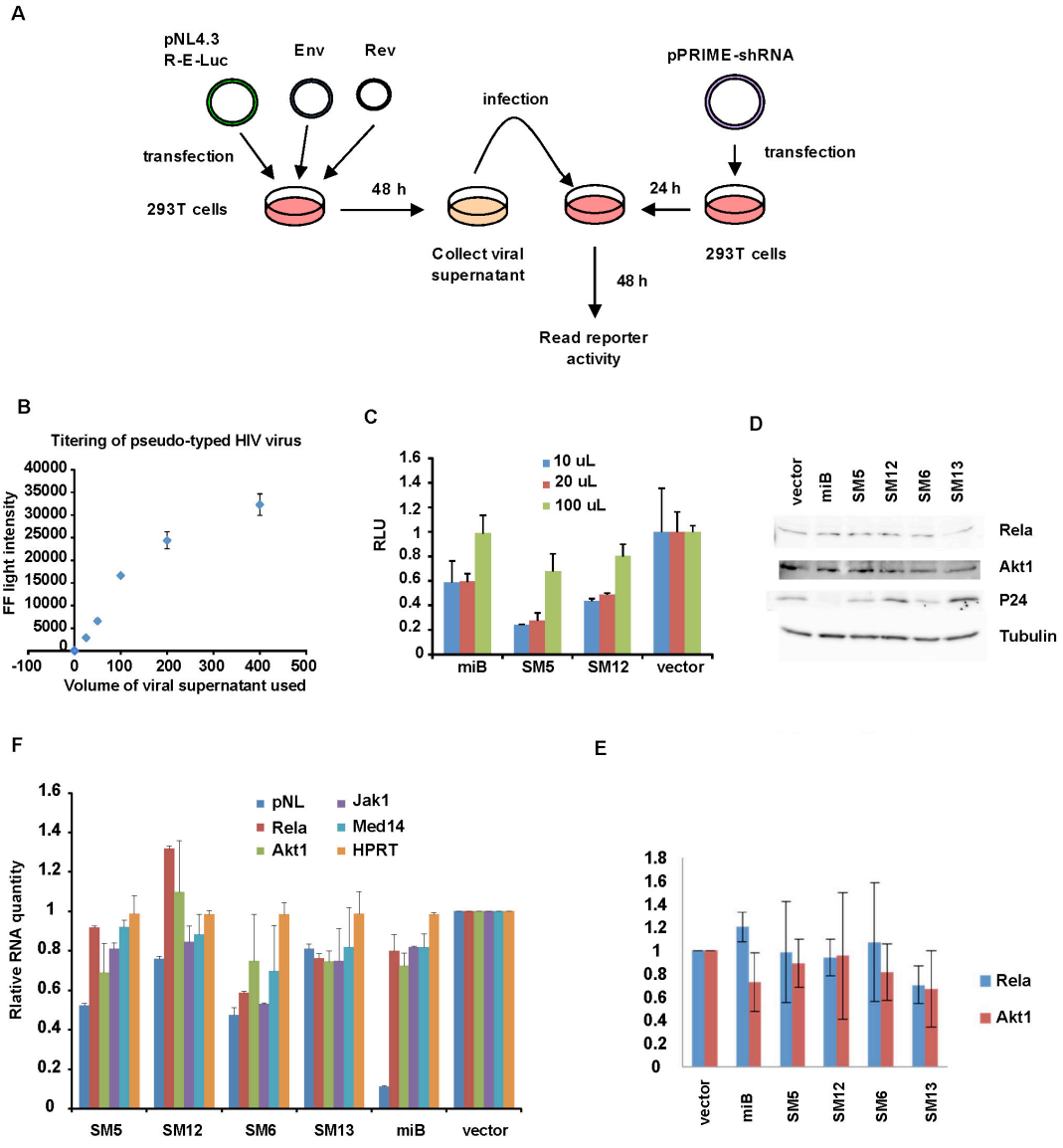


Figure 8. Protection against infection in transiently transduced cells.

Figure 8. Protection against infection in transiently transduced cells.

(A) Experimental layout to test *MiRBooking* designed shRNAs for their ability to protect cells from viral infection. The target cells were transiently transduced to express the guide RNA from the pPRIME vector.

(**B**) Titration of the viral supernatant for infection assay. When the amount of viral supernatant used is below 100 uL, the firefly light intensity linearly increases with the amount of supernatant used.

(**C**) SM5 and 12 reduced viral gene expression upon infection by the VSV-G pseudo-typed virus. At high volume of the viral supernatant, the miB shRNA failed to inhibit the virus while SM5 and 12 still provide protection.

(**D**) Protein quantification of two endogenous and one viral target: Rela, Akt1, and p24. Samples from cells that were transiently transduced with SM5, 12, 6, and 13 were tested.

(**E**) Measured protein levels from Western blot of the cell samples (N=2). The intensities of the protein band on the blot were normalized to those of tubulin. Error bar represents one standard deviation.

(**F**) RT-pPCR quantification of RNA levels of the target genes. The viral RNA, pNL, and the endogenous Real, Akt1, Jak1, and Med14 were quantified and normalized to HPRT RNA levels. Error bar represents one standard deviation.

2.3.5 Protective effects in stably transduced cells

In order to assess the protective effects in the stably transduced cells that express the designed guides (**Fig. 9A**), we used the third generation lentiviral packaging system to produce the pPRIME-shRNA viruses in 293TC17 cells (a gracious gift of the Gagnon lab). Stable expression of some of the guides caused cell loss due to possibly cytotoxicity. We had to drop them from subsequent assays, in addition to the ones that were dropped due to cloning problems (blank columns in **Fig. 9B**). Consistently, SM5 and SM12 were the most efficient guide sequences and provided better protection than the miB positive control; meanwhile, three additional guides (SM2, 3, and 4) that are perfectly complementary to HIV RNA also reduced viral gene expression (30-40%). Nevertheless, with repeats in triplicates, statistically significant protection was mediated by SM3, 5, and 6 for HIV full complementary guides; SM 12 and 13, which were purposely designed to mismatch the HIV as artificial miRNAs, also mediated protection. Interestingly, SM 18, 19, 20, and 21, which were designed with scrambled tail sequences, also showed protective effects.

To assess the knockdown efficiency of the intended endogenous targets, we performed Western blot to detect the Rela and Akt1 protein levels in the stable cells (**Fig. 9C**). SM2, 12, 13, 4, and 19 all showed significant repression on Rela; while SM2 and 19 showed some reduction in Akt1 levels. This indicates that tail complementarity is playing an important role in the knockdown efficiency of endogenous genes. Comparing to the guides that perfectly base pair with the HIV RNA for their entire length of 22 nt, the same or even better protection can be achieved by partially complementary guide sequences that share the same seed sequences (comparing SM3 with 18, SM 4 with 19, SM5 with 12 and 20, SM 6 with 13 and 21).

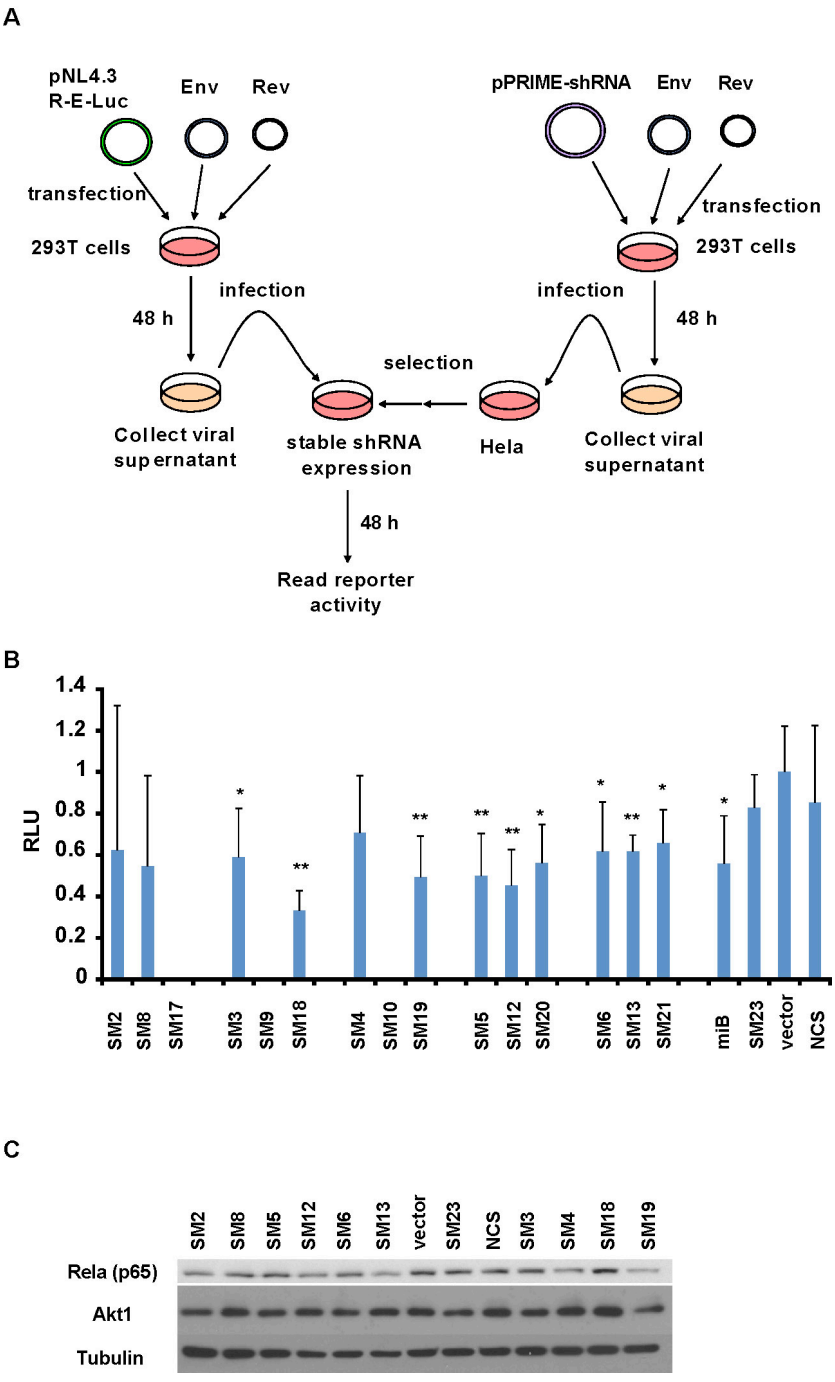


Figure 9. Cells stably transduced with *MiRBooking* designed shRNAs showed protection against viral infection

Figure 9. Cells stably transduced with *MiRBooking* designed shRNAs showed protection against viral infection.

(**A**) Experimental layout for testing the protective effects of the shRNAs using stably transduced cells. The pPRIME lentiviral vector was co-transfected with the third generation packaging system. HEK293T cells were used to package the shRNA-expressing virus. Hela cells were stably transduced by the harvested viral supernatant. The pNL4.3-R-E- was packaged using the same protocol as in **Fig. 6A**.

(B) Reporter assay result (N=3, each with technical triplicates) of stably transduced HeLa cells that expressed the designed shRNAs upon the challenge of pNL4.3 R-E-Luc pseudo-typed virus. Columns indicating the reporter expression levels were grouped together by the same seed number in the guide RNA used. SM5 and 12 are consistently the best guide sequences; meanwhile, three additional shRNAs (SM2, 3, and 4) that are perfectly complementary to HIV RNA also reduced viral gene expression. Note: ** indicates $p < 0.005$ and * indicates $p < 0.05$ when comparing to SM23, the non-targeting negative control using two-tailed Student t-test. Cell loss occurred during the transduction of SM9 and 10; these two shRNAs were dropped from testing in addition to SM7, which we had difficulty to clone.

(**C**) Protein levels of Rela and Akt1, the two key targets, were quantitated and visualized by Western blot. Tubulin was quantitated as the control.

2.3.6 Assessment of the effects of mismatched nucleotides in the non-seed region

As we found out in the viral infection assay, the guides that are partially base pairing with the HIV target RNA showed effective inhibition of the viral expression comparing to the perfectly matching guides. To see whether the shRNAs that partially base pair with the HIV target RNA in the tail region could repress viral reporter by direct targeting alone, we mutated the nucleotides in the non-seed region of the positive control guide RNA, miB. Three nucleotides were mutated to its Watson-Crick complementary version at a time, starting from nt 9 toward the 3' end of the guide strand. Four nucleotides at the 3' end were mutated altogether. We name these modified miB sequences miB-A, -B, -C, and -D, in which each hyphenated letter represents a "module" that is mutated. Next, we mutated two modules at a time, in all possible combinations. We named them miB-AC, -BD, -AD, -AB, -BC, and -CD, respectively (**Fig. 10A**). We co-transfected these shRNAs into the cells and measured their inhibitory effects on the pNL reporter. Except for miB-D, all modifications abolished repression comparing miB (**Fig. 10B**). This indicates that an artificial miRNA that targets the HIV RNA with mismatches in the tail regions is unlikely to act as an efficient guide RNA that mediates repression. Hence the inhibitory effects we observed with partially complementary guide RNAs (such as SM 12, 13, 18, 19, 20, and 21) were unlikely due to the repression of the HIV expression.

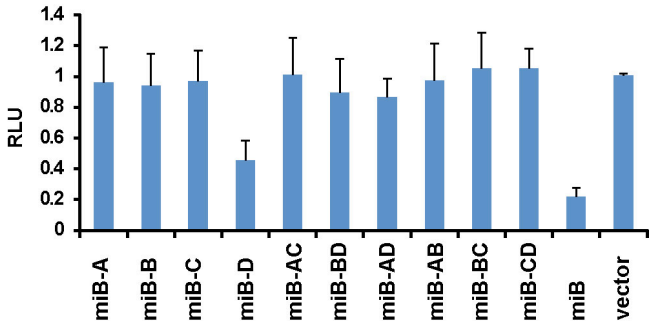
Previous report indicated that miRNA target sites were preferentially found within the 3'UTR of the genes (Chi et al., 2009). To test whether the inefficient targeting by the artificial miRNAs we made from miB was due to the target location, which is in the coding region of the tat gene, we repeated the reporter assay in the presence of puromycin at different concentrations. Using puromycin, we enhanced repression mediated by mismatched guides modified from miB by up to 20% (**Fig. 10C**). The enhancing effect of puromycin on repression confirms that the modified versions of miB mimicked natural miRNA for their repression levels on HIV. In

contrast, such enhancing effect was not observed in the presence of rapamycin or cycleohexamide (data not shown).

A

Oligo Name	5'->3' guide sequences
miB	UGUUCUGAUGAGCUCUUCGUC
miB-A	UGUUCUGAACUGCUCUUCGUC
miB-B	UGUUCUGAUGACGACUUCGUC
miB-C	UGUUCUGAUGAGCUGAACGUC
miB-D	UGUUCUGAUGAGCUCUUGCAG
miB-AC	UGUUCUGAACUGCUGAACGUC
miB-BD	UGUUCUGAUGACGACUUGCAG
miB-AD	UGUUCUGAACUGCUCUUGCAG
miB-AB	UGUUCUGAACUCGACUUCGUC
miB-BC	UGUUCUGAUGACGAGAACGUC
miB-CD	UGUUCUGAUGAGCUGAAGCAG

B



C

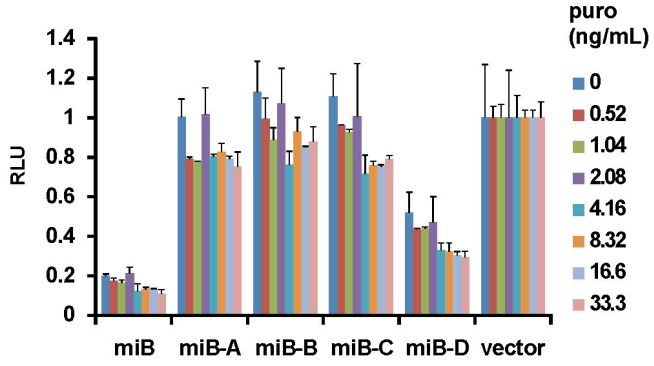


Figure 10. Non-seed nucleotide complementarity is important for HIV-targeting shRNAs

Figure 10. Non-seed nucleotide complementarity is important for HIV-targeting shRNAs.

(A) Mutated versions of the miB guide RNA. Nucleotides are changed to its Watson-Crick complementary nucleotide. Each module contains three or four consecutive nucleotides. Single- and double-module mutations were generated for miB. Mutated nucleotides are in red and underlined.

(B) Dual luciferase assay results show that only the D-module mutations are tolerated. All other mutations abolished repression. Error bars represents standard deviation from three individual repeats (with technical quadruplicates).

(C) The abolishing effect of mismatched modules can be moderately alleviated by puromycin.

2.4 Discussion

In our study, we demonstrated that designing smart RNA guides that mimic miRNAs by partially base pairing with the viral and cellular RNA targets. These smart RNAs repressed the viral gene expression from invading viral particles as efficiently as the fully complementary guide sequences. This is a surprising finding because, corroborating with previous reports (Liu et al., 2009), our experiments showed that mutations in the perfectly complementary guide sequence should not be well tolerated when directly targeting the viral genomic RNA. For these guide RNAs to achieve such efficient repression, simultaneous repression of endogenous factors must have occurred. To be sure, we verified the knockdown of the Rela protein, an endogenous transcription factor that is selectively activated by the HIV for its gene expression, in the presence of the most effective smart RNA (SM12). Similarly, Akt1 knockdown was observed when the smart RNA guide, SM13, was stably expressed. Other intended targets were knocked down at the RNA level, but significant reduction in the protein levels could not be observed, possibly due to protein stability and precision limit of the detection techniques.

Previous studies showed that directly targeting incoming viral genome using RNAi was inefficient because the viral genome was protected by a protein coat (Cullen et al., 2013). However, the expression of viral genes still depends on endogenous factor. By simultaneous inhibition of multiple cellular factors, the viral genome will be exposed to limited amount of required factors once the protein coating dissociates in the intracellular environment. Protection provided by *mirDesign* guide RNAs thus represents an alternative RNAi design strategy to address such known shortcoming of anti-HIV shRNAs.

In transiently transduced cells, the efficiency of guide RNAs with the three classes of "tails" designed for each seed follow a consistent pattern. The "HIV fully-complementary" sequence is always more efficient than "artificial miRNA", which is more efficient than the "scramble". Interestingly, when cells were stably transduced with these guides, the efficiency of all classes of guides increased;

consequently, ordered efficiency pattern of the three classes diminished. This observation is consistent with the mechanism of miRNA-mediated slicer-independent silencing, by which downstream protein factors are recruited by Argonaute protein to elicit mRNA deadenylation, decapping, and degradation processes (Fabian et al., 2009). We argue that because this is a slower process than slicer-dependent mechanism, a stably transduced system allows cellular targets for longer exposure to smart RNAs such that stronger repression can be achieved.

As *mirDesign*'s first application to design smart RNA guides, there are still a few limitations to the program. The first limitation is that it only considers the seed complementarity and relative concentration of each mRNA species in the cell type in question. Base pairing beyond the seed region was not considered. In the design phase of the guide sequences, we addressed this issue by designing three different tail sequences for each seed. As we observed, in the transiently transduced system, repression of HIV was always more efficient with fully complementary sequences. Though clearly tail sequence base pairing plays a role in repression (Table V), its contribution could not yet be predicted by *mirDesign*, Further improvement can be made by incorporating their contributions into *mirDesign* according to published studies (Broughton et al., 2016).

Another challenge is that sequence requirement for efficient guides cloning, propagation, and guide strand biogenesis were not taken into consideration. We addressed this issue manually in the design phase according to known rule for guide design (Moore et al., 2010). However, the seed sequences were still mostly enriched in G/C nucleotides because *mirDesgin*'s calculation favors seed pairing that are thermodynamically stable. As we used miR-30 backbone to express the smart RNAs, long stretches of "G/C" are generally not well tolerated for shRNA expression; moreover, correct RISC loading is likely interfered due to altered asymmetry of thermodynamic profile of the duplex RNA trigger (Suzuki et al., 2015). Consequently, three of the six guides of the "best fold change" category could not propagate during cloning. In addition, "best fold change" is associated with cytotoxicity, which caused one additional guide RNA to fail to be stably expressed.

These shows that using *mirDesign* as a design tool, "best fold change" may not be the best choice for the best anti-HIV guide RNA. In contrast, smart RNAs (SM5, 6, 12, and 13) of the "high disturbance" category showed most consistently potent repression against HIV; meanwhile, "best fold change" and "low disturbance" categories only showed significant repression in stably transduced cells. Nevertheless, comparison between three categories of guide sequences demonstrates that the number of target sites is a key factor to the success of efficient anti-HIV guide RNAs. It is a proof-of-concept of the principles behind *MirDesgin*: targeting as many required factors as possible simultaneously with moderate potency, one can achieve efficient repression of viral gene expression while minimizing cytotoxicity.

Lastly, *mirDesign* does not check the target site accessibility and binding site complementarity. Local structural context and accessibility were deemed to be important for the efficiency of repression by multiple reports (Grimson et al., 2007; Kertesz et al., 2007). This issue is also addressed manually at the moment by referencing previous anti-HIV shRNA library screening data (Tan et al., 2012). Identifying accessible regions of the HIV RNA may serve as an additional improvement step of the *mirDesign* pipeline. Targeting the coding region is generally disfavored for partial complementary guide sequences due to the translating ribosomes. Mutations in anti-HIV shRNAs are not well tolerated is likely due to the fact that most of the HIV genome is coding. We confirmed this by treating cells with puromycin, an inhibitor of translation, when smart RNAs were delivered. The enhancing effect of puromycin on repression by miRNA mimetics indicates a possible solution to the accessibility issue when targeting coding sequences. Puromycin is a structural mimetic to the 3'-end of aminoacylated tRNA and enter the A-site of the translating ribosome. It causes ribosome to terminate translation and subsequently dissociate from mRNA (Blobel and Sabatini, 1971). This facilitates RISC to access the target and resulted in more efficient repression. Such effect was not observed when we treated the cells with rapamycin or cyclohexamide, which are inhibitors of translation that do not cause ribosomes to dissociate from mRNA, indicating that the enhancement is specifically associated with ribosome clearing.

Table V. Alignment of some *MiRBooking* designs with their target sites.

	SM5	*SM12*	*SM6*	*SM13*
HIV	>YH_000001,8844 5' CAGGGAGGCGUGGCCUGGGCGGGACU 3' 3' CUCCGCACCGGACCCGCCCUGU 5' >YH_000001, 4004 5' AAGGCCGCCU-GUUGGUGGGCGGGAAU 3' 3' CUCCG-C--ACCGG—ACCCGCCCUGU 5' >YH_000001, 7482 5' UAACA-UGACCUGGAUGGAGUGGGACA 3' 3' CUCCGCACCGGACCCGCCCUGU 5' >YH_000001,7017 5' ACCUGGAGGA—GGCGAUAUGAGGGACA 3' 3' CUCCGCACCG-GACCCGCCCUGU 5'	>YH_000001,8844 5' CAGGGAGGCGUG-GCCUG-GGCGGGACU 3' 3' GCGCACGCGGCCACCGCCCUGU 5' >YH_000001,4004 5' AAGGCCGCCUGUUGGUGGGCGGGAAU 3' 3' GCG-CACGCGGCCAC-CGCCCUGU 5' >YH_000001,7482 5' UAACAUGACCUGGAUGGAGUGGGACA 3' 3' GCGCACGCGGCC-ACC--GCCCUGU 5' >YH_000001,7017 5' ACCUGGAGGAGGCGAUAUGAGGGACA 3' 3' G-CGCACGCGGCCAC-CGCCCUGU 5'	>YH_000001,6028 5' ACUAAUACCAAUAGUAGUAGCGGGAG-3' 3' UAUGGUUAUCAUCAUCGCCCUU-5' >YH_000001,4003 5'-UAAGGCCGCCUGUUGGUGGGCGGGAA-3' 3'-UAUGGUUAUCA-UCA-UCGCCCUU-5' >YH_000001,H8843 5'-CCA-GGGAGGCGUGGCCUGGGCGGGAC-3' 3'-UAUGGUUAUCAUC---AUCGCCCUU-5' >YH_000001,192 5'-GCGAGAGCGUCGGUAUUAAGCGGGGG-3' 3'-UAUGGUUAUCAUCA-UCGCCCUU-5' >YH_000001,2707 5'-CAAUGACAUACAGAAAUUAGUGGGAA-3' 3'-UA-UGGUUAUCAUC-AUCGCCCUU-5' >YH_000001,3165 5'-AAGCCACCUGGAUUCCUGAGUGGGAG-3' 3'-U---AUGGUUAUCAUCA-UCGCCCUU-5' >YH_000001,7128 5'-GCAGAGAGAAAAAAGAGCAGUGGGAA-3' 3'-UAUGGUUAUCAUCAUCGCCCUU-5' >YH_000001,7481 5'-AUAACAUGACCUGGAUGGAGUGGGAC-3' 3'-UAUGGUUAU---CAU-CAUCGCCCUU-5'	>YH_000001,6028 5' ACUAAUACCAAUAGUAGUAGCGGGAG 3' 3' CGCGCACGC--GGC--CAUCGCCCUU 5' >YH_000001,4003 5' UAAGGCCGCCU--GUUGGUGGGCGGGAA 3' 3' C-GCGCACGCGGCCAU-CGCCCUU 5' >YH_000001, 8843 5' CCAGGGAGGCGUG-GCCUG-GGCGGGAC 3' 3' C-GCGCACGCGGCCAUCGCCCUU 5' >YH_000001 ,192 5' GCGAGAGCGUCGGUAUUAAGCGGGGG 3' 3' CGCGCACGCGGCC---AU-CGCCCUU 5' >YH_000001,2707 5'-CAAUGACAUACAGAAAUUAGUGGGAA-3' 3'-CGC-GCACG-CGGCCAUCGCCCUU-5' >YH_000001,3165 5' AAGCCACCUGGAUUCCUG-AGUGGGAG 3' 3' CG-CGCAC--GCGGCCAUCGCCCUU 5' >YH_000001,7128 5' GCAGAGAGAAAAAAGAGCAGUGGGAA 3' 3' CG-CGCACGC--GGC-CAUCGCCCUU 5' >YH_000001,7481 5' AUAACAUGACCUGGAUGGAGUGGGAC 3' 3' CGCGCACGCGGCC---AUCGCCCUU 5'
Rela	>NM_021975,509 5’ AAUCCA-GUGUGUGAAGAAGCGGGACC 3’ 3’ CUCCGCAC-CGGACCCGCCCUGU 5’ >NM_021975,635 5’ CUGCUUCCAGGUGACAGUGCGGGACC 3’ 3’ CUCCG—CACCGGACCCGCCCUGU 5’ >NM_021975,227 5’ CAUUGAGCAGCCCAAGCA-GCGGGGCA 3’ 3’ CU-CCGCA-CCGGA-C—CCGCCCUGU 5’	>NM_021975, 509 5’ AAUCCAGUGUGUGAAGA-AGCGGGACC 3’ 3’ G--CGCACGCGGCCACCGCCCUGU 5’ >NM_021975, 635 5’ CUGCUUC-CAGGUGACAGUG-CGGGACC 3’ 3’ GCGCACGCGGCCACCGCCCUGU 5’ >NM_021975 227 5’ CAUUGAGCAGCCCAAGCAGCGGGGCA 3’ 3’ G-CGCACG-CGGC--CACCGCCCUGU 5’	>NM_021975,508 5’ GAAUCCAGUGUGUGAAGAAGCGGGAC 3’ 3’ UAUGGUUAU-CA--UCAUCGCCCUU 5’ >NM_021975, 226 5’ UCAUUGAGCAGCCCAAGCAGCGGGGC 3’ 3’ U--AUGGUUAUCAUCAUCGCCCUU 5’ >NM_021975,634 5’ UCUGCUUCCAGGUGACAGU-GCGGGAC 3’ 3’ UAUGGUUAUC-AUCAUCGCCCUU 5’	>NM_021975, 508 5’ GAAUCCAGUGUGUGA-AGAAGCGGGAC 3’ 3’ C---GCGCACGCG—GCCAUCGCCCUU 5’ >NM_021975, 226 5’ UCAUU--GAGCAGCCCAAGCAGCGGGGC 3’ 3’ CG---CGCA---CGCGGCCAUCGCCCUU 5’ >NM_021975, 634 5’ UCUGCUUCCAGGUGACAGU-GCGGGAC 3’ 3’ CG--CGCACGCGGCCAUCGCCCUU 5’

Akt1	>NM_005163, 2594 5′ UGGGCCAGGGUUUACCCAGUGGGACA 3′ ||| | ||||| |||||||| 3′ CUCCGCACCGGAC---CCGCCCUGU 5′	>NM_005163, 2594 5′ UGGGCCAGGGUUUACCC--AGUGGGACA 3′ || | || || |||||||| 3′ GCG--CACG--CGGCCACCGCCCUGU 5′	>NM_005163, 2169 5′ GGCAGCACCCUCCCCCGCAGCGGGGU 3′ | ||| | |||||| 3′ U-AUGGUUAUCAUCAUCGCCCUU 5′ >NM_005163, 2593 5′ AUGGGCCAGG-GUUUACCCAGUGGGAC 3′ | |||| || |||||| 3′ UAUGGUUAUCAUC---AUCGCCCUU 5′	>NM_005163, 838 5′ AUGUGGAGACUCCUGAGGAGCGGGAG 3′ |||| | || | | |||||||| 3′ CGCGCA-C-GCGG-C-CAUCGCCCUU 5′ >NM_005163, 2169 5′ GGCAGCAC--CCUCCCCCGCAGCGGGGU 3′ || | || |||||| 3′ CG--CGCACGCGGCCAUCGCCCUU 5′ >NM_005163, 2593 5′ AUGGGCCAGGGUUUACC--CAGUGGGAC 3′ | || | || || |||||| 3′ CGCG--CACGC--GGCCAUCGCCCUU 5′
MED 4	>NM_001270629,47 5′ UCUGCGCGUGCGCCGGUGGCGGGACU 3′ | |||| ||| | |||||||| 3′ CUC-CGCAC-CGGAC-CCGCCCUGU 5′ >NM_014166,10 5′ G---CG---CCGGUGGCGGGACU 3′ | || || | |||||||| 3′ CUCCGCACCGGAC-CCGCCCUGU 5′	>NM_001270629, 47 5′ UCUGCGCGUGCGCCGGUGGCGGGACU 3′ |||||||||||||||||||||| 3′ GCGCACGCGGCCACCGCCCUGU 5′ >NM_014166 10 5′ G-C--GC--CGGUGGCGGGACU 3′ | || |||||||||||| 3′ GCGCACGCGGCCACCGCCCUGU 5′	>NM_001270629, 46 5′ CUCUGCGCGUGCGCCGGUGGCGGGAC 3′ ||| ||| | |||||||||| 3′ UAUGGUUAU-CA-UCAUCGCCCUU 5′ >NM_001270629, 79 5′ AAAAUGGCUGCG-U-CUUCGAGUGGGAA 3′ || |||| | | |||||||| 3′ UA-UGGU—UAUCAUCAUCGCCCUU 5′ >NM_014166, 9 5′ GCGCC---GGU---GGCGGGAC 3′ | ||| ||| ||||||| 3′ UAUGGUUAUCAUCAUCGCCCUU 5′	>NM_001270629,46 5′ CUCUGCGCGUGCGCCGGUGGCGGGAC 3′ |||||||||||||||||||| 3′ CGCGCACGCGGCCAUCGCCCUU 5′ >NM_001270629, 79 5′ AAAAUGGC-UGCGUCUUCGAGUGGGAA 3′ | || ||||| |||||||| 3′ CG-CGCACGCGGC-CAUCGCCCUU 5′ >NM_014166, 9 5′ GCGC---C---GGUGGCGGGAC 3′ |||| | ||||||||| 3′ CGCGCACGCGGCCAUCGCCCUU 5′
Jak1	>NM_002227,86 5′ CGCGCACGCUGGGGCCC-CGCGGGGCU 3′ | | || | |||| ||||||| 3′ C-UCCG-CACCGGACCCGCCCUGU 5′	>NM_002227, 86 5′ CGCGCACGCUGGGGCCCCGCGGGGCU 3′ |||| ||||| | ||||||| 3′ GCGCACGCGGCCAC----CGCCCUGU 5′		

Note: The alignment, the target gene's accession number, and the target site position of the smart RNA guides. Predicted Watson-Crick and GU Wobble base pairs are shown using bars, '|'; target strand on top. The target strand is on top, in 5'-3' direction and the guide strand is at the bottom.

2.5 Materials and methods

2.5.1 Design three classes of "tail sequences" for each guide RNA seed

Complete cDNA sequences of the ten target genes, in addition to HIV pNL4-3 sequence, were given to the bioinformatician in the lab as the input for the *mirDesign* program. As the output, 15 seed sequences of three different categories were taken to the next design phase. For each seed sequence, three different "tail" sequences, according to the aforementioned principles, were appended to complete the design of 21 nt guide RNA. For anti-HIV SmartRNAs, we chose to target regions in the HIV genome that were mapped to be accessible to shRNAs by Tan et al (Tan et al., 2012). Among 15 seed sequences identified, we chose six seeds that perfectly complement the HIV genome at more than 3 sites and of which the log score greater than or equal to 0 in Tan's data.

2.5.2 Categorization of seeds *mirDesign*-predicted seeds

Designed seeds are divided into three categories based on *mirDesign* predictions: best fold change, low disturbance, and high disturbance. Fold change of repression is defined as the ratio of the repression level under normal conditions and that when the designed guide sequence is present. For each guide sequence, a fold change is predicted for each target gene. For 10 target genes, the lowest fold change of the top 5 (i.e. the fifth best fold change) is used as the representative of the overall fold change of the seed on all cellular targets. Combined with efficient repression of HIV predicted, seeds with best representative fold change are classified under the category of "best Fold Change". The total number of target genes of which the expression level is altered more than two fold is termed the "disturbance" in *mirDesign*. Combined with the prediction of efficient repression of HIV, seeds with the lowest and highest disturbance form the categories of Low Disturbance and High Disturbance, respectively. Top five seeds are designed for each category.

2.5.3 Cloning of designed smart RNAs

Using the same design pipeline, we have cloned the multi-targeting siRNAs into the miR-30 backbone-containing shRNA constructs, pPRIME (a gracious gift from Dr. Abba Malina of the Pelletier lab) (Dow et al., 2012). Using an in-house developed program, "m2sh" (http://www.major.iric.ca/~dallaire/m2sh/), we converted the 21 nt guide RNA sequence into a 97-nt PCR template that was used for cloning into the miR-30 based lentiviral vectors. A "U" residue was added at the 5'-end of each guide sequence when generating the PCR template sequences to enhance its binding to the Ago2 protein.

2.5.4 Plasmid Construction

The vector pPRIME (a gift from Jerry Pelletier's lab) has been previously optimized for shRNA cloning. Designed guide-RNAs were cloned into the vector following miR-30-based shRNA cloning protocols. Briefly, complementary oligonucleotides that contain the shRNA sequences (Biocorp, oligos are listed in Table IV) were diluted to 100 μM in deionized water. Annealing reaction was carried out at 95˚C in annealing buffer for 5 minutes followed by slow cooling to room temperature. The annealed double-stranded oligonucleotides were then phosphorylated by T4 PNK (NEB). Ligation reaction was performed by combining doubly digested pPRIME by *XhoI* and *EcoRI* with the phosphorylation product of annealed oligonucleotides in T4 DNA ligase (NEB) reaction mix at 16 ˚C overnight. For all the shRNA trigger sequences designed for this study, see Table III.

The renilla luciferase control vector, SVR, was obtained by replacing the CMV promoter in the pcDNA3-RlucII plasmid (a gift from Sylvie Mader's lab) with an SV40 promoter. Briefly, the CMV promoter was removed by restriction enzymes *SpeI* and *HindIII* (New England Biolabs). The resulting linearized vector was gel-purified with QIAEX II ® Gel Extraction Kit. The SV40 promoter from the pGL3-control luciferase vector fragment was obtained by digesting the vector with *NheI*

and *HindIII*. Gel purified SV40 promoter fragment was inserted upstream of the RlucII gene in pcDNA-RLucII vector by ligation using T4 DNA ligase (NEB).

Similarly, the renilla luciferase control vector TKR was constructed by replacing the CMV promoter of pcDNA3-RlucII plasmid with the thymidine kinase (TK) promoter from the pRL-TK plasmid (Promega) by restriction endonucleases *BglII* and *HindIII* (NEB). Compatible ends were generated, as such, the promoter fragment can be ligated with the promoter-less pcDNA3-RlucII fragment as previously described for SVR construction.

The pNL tat- plasmid was generated by removing the tat gene from pNL4-3luc R-E- plasmid. Briefly, NheI and EcoRI (NEB) were used to digest the purified plasmid following supplier's instructions. The large fragment obtained was treated with Klenow (NEB) and subsequently ligated using T4 DNA ligase (NEB). The ligation product was then transformed into *E. Coli* for amplification and purification.

2.5.5 Cell culture and transduction of gene expression

HEK 293T (c17) cells (from ATCC) were maintained according to established conditions. Briefly, cells were grown in DMEM (+L-glutamine) (Life Technologies) supplemented with 10% FBS, 100 U/mL penicillin/streptomycin at 37 ˚C and 5% CO_2. Cells were grown to confluency before plating. For testing the efficiencies of mismatched guides, cells were plated in 96-well plates at ~20,000 cells per well 24 hours prior to the transfection. For assays that required growth in 24-well plates, cells were plated at ~100,000 cells per well. The reporter plasmids and the shRNA plasmids were co-transfected into the cells using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. Along with 10 ng of shRNA plasmid, 5 ng of pNL-luc and 2 ng of SVR control vector were co-transfected into each 96 well; alternatively, 50 ng of the shRNA construct, 20 ng of the pNL-luc, and 10 ng of the SVR control vector were co-transfected into each 24-well.

2.5.6 Establishing stable cell lines that express designed smart RNAs

In a 6-well plate, we plated 8X105 cells/well 24 hours before transfecting pPRIME-shRNA, VSV-G, RRE, and REV DNA at 2:2:1:1 ratio. The total amount of DNA is 2 μg per well. Viral supernatant was collected and filtered at 48 hours post transfection. Supernatant was then flash-frozen in aliquots and the viral titer was tested. Target cells were infected at low MOI (<0.15). Cells were analyzed using the BD FACSCanto™ II flow cytometer. Infected cells were selected in puromycin media (1 μg/mL) for one week to obtain a stable cell line. Stable cells were counted and seeded 24 hours before infection and luciferase assay was performed 48 hours post infection of the pNL4.3R-E-luc virus.

2.5.7 Pseudoviral particle packaging using pNL4.3-luc

To package viral genome using pNL4-3lucR-E- plasmid, the VSV-G and RRE plasmids of the third generation lentiviral packaging system were co-transfected into 293T cells. The ratio between the pNL, VSV-G, and RRE plasmid was 2:2:1. The viral supernatant was collected 48 hours post-transfection and was aliquoted and flash frozen for long term storage at -80 ˚C.

2.5.8 Dual luciferase assay

Luciferase assays were performed accordingly to established protocols adapted from the Duo-Glo Luciferase System (Promega). 48 hours post-transfection, cells were lysed with 1× Passive lysis buffer (Promega) and luciferase activity was assayed using the Dual-Glo Luciferase System (Promega). Luminescent light was measured on Veritas Microplate Luminometer (Turner Biosystems) (a gift from the Michel Bouvier's Lab). The ratio between the reporter and the control luciferase bioluminescence light was taken and then normalized to that of the negative control shRNA or empty vector, resulting in the percentage residual expression of the reporter gene.

2.5.9 Immunoblot Analysis

Cells were washed with cold PBS and then scraped on ice into 500 µl of PBS buffer containing 1X Complete-EDTA free Protease Inhibitor Cocktail (Roche Applied Science) and 1X PhosSTOP Phosphatase Inhibitor Cocktail (Roche Applied Science). Cells were spun at maximum speed for 5 min. Protein extracts were prepared in RIPA lysis buffer (20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 0.1% SDS, 1% Nonidet P-40, 0.5% sodium deoxycholate, 1 mM β-glycerophosphate, 1 mM PMSF, 1 µg/ml leupeptin, 10 µg/ml aprotinin, and 2.5 µM pepstatin A) at 10 days post-transduction. PVDF membranes were probed with the indicated primary antibodies and HRP-conjugated secondary antibodies (rabbit or mouse (Cell Signaling)) and visualized using enhanced chemiluminescence (ECL) (PerkinElmer Life Sciences). Proteins were quantified with the Bradford reagent and 30 µg were loaded on a 10% SDS–PAGE and transferred to Immobilon-P PVDF membranes (Millipore). Membranes were blocked 1 hour at room temperature in PBS containing 0.1% Tween 20 (PBS-T) and 5% dry milk and then washed for 5 min 3 times with PBS-T. The membranes were incubated with the primary antibodies diluted in PBS-T + 3% BSA + 0.05% Na-azide overnight at 4ºC. The following primary antibodies were used in this study: P24 (ab9071), NF-κB (Santa Cruz, sc-71675), Akt1 (Santa Cruz, sc-5298), β-tubulin (sc-5274). Quantification of Western blot band intensities was carried out using the ImageJ software (National Institutes of Health). Membranes were washed three times 5 min with PBS-T and then incubated with the secondary antibodies diluted in PBS-T + 5% dry milk 1 hour at room temperature. Finally, the membranes were washed three times 5 min with PBS-T. Immunoblots were visualized using enhanced chemiluminescence (ECL) detection systems and Super RX X-Ray films (Fujifilm) or a ChemiDocTM MP system (Bio-Rad). Band quantification was done using ImageJ or Image Lab 4.0 (Bio-Rad).

2.5.10 Measuring reporter transcript and mature RNA guide abundance using RT-qPCR

RNA extraction was performed using TRIzol® reagent following manufacturer's protocol. RNA was extracted from the same cells used in the luciferase assay. Either oligo-dT primer or random primer were used for the synthesis of cDNA from total RNA extracted according to previously established protocols (Kiethega et al., 2013). 800 ng of total RNA was used for each synthesis reaction in 20 µL of total volume using Invitrogen reagents (M-MLV Reverse Transcriptase, Cat. No. 28025-021, Invtitrogen^{TM}). RNA was extracted from the same cells that were used in the luciferase assay and M-MLV was used to perform cDNA synthesis.

The newly synthesized cDNA was diluted by a factor of 100 prior to real-time PCR. Each real-time PCR reaction mixture contained the diluted cDNA (1 µl), forward and reverse primers (250 nM), $MgCl_2$ (2.5 mM), dNTPs (0.2 mM), SYBR green (0.33X), buffer for Jumpstart *Taq* DNA polymerase and Jumpstart *Taq* DNA polymerase (0.25 U; Sigma) in a final volume of 10 µl. After denaturation at 95 °C for 6 min, samples went through 50 cycles of amplification (20 s at 95 °C, 20 s at 58 °C and 30 s at 72 °C). Melt curves were determined for each reaction and qPCR was performed using a LightCycler 480 (Roche Applied Science, Canada). Data was normalized using Renilla and HPRT as controls.

2.6 Acknowledgements

We thank Jerry Pelletier, Abba Malina, John Mills, Regina Cencic, Francis Robert, David Cotenoir-White, Khalid Hilmi, and Justina Kulpa for discussions, cloning and cell culture materials, as well as assistance; Julie Pelloux, Jean Paquette, and Angelique Bellmare-Pelletier for discussions; Sylevie Mader and Eric Cohen for constructs; and, André Laperrière for assistance. Grants to François Major from the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery grant program), the Canadian Institutes of Health Research (CIHR) [MOP-93679], and the National Institutes of Health [R01GM088813] supported this work.

CHAPTER 3: A NEW MODEL FOR BASE PAIRING BEYOND THE SEED

Yifei Yan, Mariana Acevedo, Lian Mignacca, Philippe Desjardins, Nicolas Scott, Roqaya Imane, Jordan Queneville, Julie Robitaille, Albert Feghaly, Etienne Gagnon, Gerardo Ferbeyre, and François Major**.**

The sequence features that define efficient and specific hAGO2-dependent miRNA silencing guides

Nucleic Acids Res. 2018 Sep 19; 46(16): 8181–8196. Published online 2018 Jun 22. doi: 10.1093/nar/gky546

3.1 Abstract

MicroRNAs (miRNAs) are ribonucleic acids (RNAs) of ~21 nucleotides that interfere with the translation of messenger RNAs (mRNAs) and play significant roles in development and diseases. In bilaterian animals, the specificity of miRNA targeting is determined by sequence complementarity involving the seed. However, the role of the remaining nucleotides (non-seed) is only vaguely defined, impacting negatively on our ability to efficiently use miRNAs exogenously to control gene expression. Here, using reporter assays, we deciphered the role of the base pairs formed between the non-seed region and target mRNA. We used molecular modeling to reveal that this mechanism corresponds to the formation of base pairs mediated by ordered motions of the miRNA-induced silencing complex. Subsequently, we developed an algorithm based on this distinctive recognition to predict from sequence the levels of mRNA downregulation with high accuracy ($r^2 > 0.5$, p-value $< 10^{-12}$). Overall, our discovery improves the design of miRNA-guide sequences used to simultaneously downregulate the expression of multiple predetermined target genes.

Key words: MicroRNA; beyond-the-seed; Argonaute; RISC; interference; silencing.

3.2 Introduction

A microRNA (miRNA) and an Argonaute (AGO) protein associate to form an essential component of the miRNA-induced silencing complex (miRISC). The miRISC-targeting specificity is mainly determined by sequence complementarity between the miRNA seed (nucleotides 2-8) and target RNA. The miRNA complementary region in the target RNA is called the miRNA regulatory element (MRE) (John et al., 2004; Lewis et al., 2003). Base complementarity between seeds and MREs is the predominant feature of most miRNA target prediction algorithms (Reyes-Herrera and Ficarra, 2012; Saito and Saetrom, 2010). When the base complementarity between a miRNA and its target mRNA is complete, they form a

perfect duplex and the miRISC cleaves the mRNA (Fabian et al., 2010; Lewis et al., 2003). However, beyond the seed, bilaterian animal miRNAs are rarely fully complementary to their targets (Doench and Sharp, 2004). Therefore, in most cases, the miRISC downregulates the expression of a gene by either removing its poly-A tail or 5' cap structure (Chen et al., 2009; Eulalio et al., 2007; Fabian et al., 2010; Giraldez et al., 2006), or repressing its translation (Fabian et al., 2010; Lewis et al., 2003).

A loop in the central region of the miRNA-mRNA duplex is tolerated (Doench et al., 2003). Studies revealed that its size and sequence influence the silencing efficiency, and hence loop scores have been assigned to improve miRNA target prediction (Kiriakidou et al., 2004; Ye et al., 2008). Mismatches at the 3' end of the miRNA-mRNA duplex (miRNA 3' end positions) were found to facilitate the release of the miRNA from the miRISC and to enhance gene silencing (De et al., 2013). Base pairs (bps) involving miRNA positions 13-16 were shown to rescue silencing when the base complementarity in the seed region is low (Friedman et al., 2009; Grimson et al., 2007). To delineate the precise contribution of all bps in the duplex, researchers performed mutagenesis studies introducing mismatches along the entire duplex (Boden et al., 2004; Du et al., 2005; Hibio et al., 2012; Holen et al., 2002; Houzet et al., 2012; Kamola et al., 2015; Robertson et al., 2010; Saxena et al., 2003). Beyond confirming that non-seed bps contribute to silencing efficiency, *in vitro* experiments further revealed that many of them along the duplex play different roles in bilaterian animals AGO2-mediated cleavage and silencing (Lima et al., 2009; Wee et al., 2012). While the seed bps were shown to affect K_m, which is a measure of affinity of the targeting process, those in the central region were found to contribute mostly to K_{cat}, which is a measure of its endonuclease activity.

These findings were sought to resolve the controversy surrounding mismatches and their effects on the efficiency of silencing. Several attempts were made to incorporate them into prediction programs. This involved calibrating each position and trying to fit a free-energy model accordingly (Agarwal et al., 2015; Friedman et al., 2009; Grimson et al., 2007; Kiriakidou et al., 2004; Lewis et al.,

2003; Majoros et al., 2013). Despite providing some additional predictive power, considerable discrepancies still exist between experimental silencing measurements and predictions. As such, the intrinsic requirements of the miRNA-induced silencing mechanism remained elusive.

Structural studies of AGO became essential to provide mechanistic details that may help reveal miRNA-induced inhibitory action. AGO is found in all three domains of life (Swarts et al., 2014). Even though AGO was initially discovered in eukaryotes (Bohmert et al., 1998), the first structural insights originated from their prokaryotic orthologs (Willkomm et al., 2015). The AGO structure is well conserved and display a bi-lobed conformation that consists of the N, PAZ, MID, and PIWI domains. The PAZ domain is connected to the N- and MID domain by two loops, L1 and L2, respectively. This renders the PAZ domain flexible enough to move as a rigid body relative to all other domains (Elkayam et al., 2012; Wang et al., 2009b; Willkomm and Restle, 2015; Willkomm et al., 2015; Yuan et al., 2005). The 5' end of the guide RNA is bound to the MID domain in a pre-shaped A-form (Wang et al., 2009b). A kink occurs at nucleotide (nt) 6, and the guide assumes an extended form all the way to the PAZ domain. The PAZ domain through hydrogen bonds and ring-stacking interactions holds the 3' extremity of the guide (Elkayam et al., 2012; Schirle and MacRae, 2012; Wang et al., 2009b).

Target recognition by the miRNA seed has been well characterized in crystal structures (Elkayam et al., 2012; Schirle and MacRae, 2012; Wang et al., 2009b). The PIWI-PAZ channel accommodates the seed region of the guide-target duplex, strongly favoring perfect complementarity at nts 2-6. In all crystal structures with the bound RNA guide (Elkayam et al., 2012; Schirle and MacRae, 2012; Schirle et al., 2014; Wang et al., 2008b; Wang et al., 2009b), the nts 2-5 in the seed are exposed to the solvent and hence available for recognition.

As the formation of the RNA duplex proceeds in the 5' to 3' direction along the guide, the 3' end of the guide strand is released from the PAZ domain, relieving the steric hindrance of the two intertwining strands. This model is supported by a crystal structure in which a bacterial AGO (*T. thermophilus*) accommodates a guide-

target duplex of 15 bps in its MID-PIWI cradle (Wang et al., 2009b). The duplex formation allows for the correct positioning of the scissile phosphate (between nts 10-11) in the nuclease active site so that the cleavage of the target strand takes place precisely. This model, called the nucleation and propagation model, was proposed to be also true for human AGO2 (hAGO2), which is the only RNase-capable isoform among the four human AGO proteins. It entails that hAGO2 exists in two states in action: the 3' end bound and the 3' end released states (Elkayam et al., 2012; Schirle and MacRae, 2012; Wang et al., 2009b). Transition between these two states is achieved through the base pairing propagation.

However, the bacterial model is incompatible with experimental data collected from bilaterian animals as their AGO protein tolerates a central loop in the guide::target duplex (Doench and Sharp, 2004). Moreover, it would also imply that base pairing downstream of the seed should increase the guide-target affinity, which was not observed at significant levels (Wee et al., 2012). The advent of human AGO2 protein structures offered an explanation (Schirle and MacRae, 2012). When hAGO2 is only loaded with a guide strand, the seed region of the guide is bound to highly conserved residues at the 5' end, while nts 9-11 are occluded by the α-7 helix and a ten-residue loop (residues 600-609, PDB4W5N), and the 3' end is bound to the PAZ domain by ring stacking and hydrogen bonds (Schirle and MacRae, 2012).

Using miB, a perfectly complementary small hairpin RNA (shRNA) against the *tat* gene of HIV (Boden et al., 2003; Boden et al., 2004), we demonstrate that base pairing beyond the seed exerts a spectrum of effects on reporter gene downregulation. A distinctive pattern linked to AGO-binding events allows us to predict induced silencing efficiency from mutated miB sequences. We used this model to develop a rule-based algorithm to compute the silencing efficiency of guide RNA sequences, validated against mRNAs in a pooled dataset from published data. We depicted this pattern at the molecular level and deduced the motions in the AGO2 upon RNA binding using molecular modeling.

3.3 Results

3.3.1 Mismatched modules cause disturbance in silencing efficiency

To study the role of base pairing as a determinant of the efficiency of AGO2-mediated silencing, we chose an shRNA, miB, which was reported to target a structurally open region of the HIV genome and inhibit viral gene expression (Boden et al., 2004; Tan et al., 2012; Wilkinson et al., 2008). Its MRE is located in exon 1 of the HIV-1 *tat* gene (nts 5993-6013). We mutated miB's nts in the non-seed region in short stretches of 3 or 4 nts at a time (modules) such that they mismatch the corresponding nts in the target sequence: from 5' to 3' module **A** (nts 9-11), **B** (nts 12-14), **C** (nts 15-17), and **D** (nts 18-21) (**Fig. 11A**). The mismatched positions were engineered by copying the nt from the target strand (A:A, G:G, C:C, and U:U). We named the guide strands containing these mismatches miB-A, -B, -C, and -D, respectively, and cloned them into pPRIME (Dickins et al., 2005), an shRNA expression vector based on the miR-30 backbone (**Fig. 11B**).

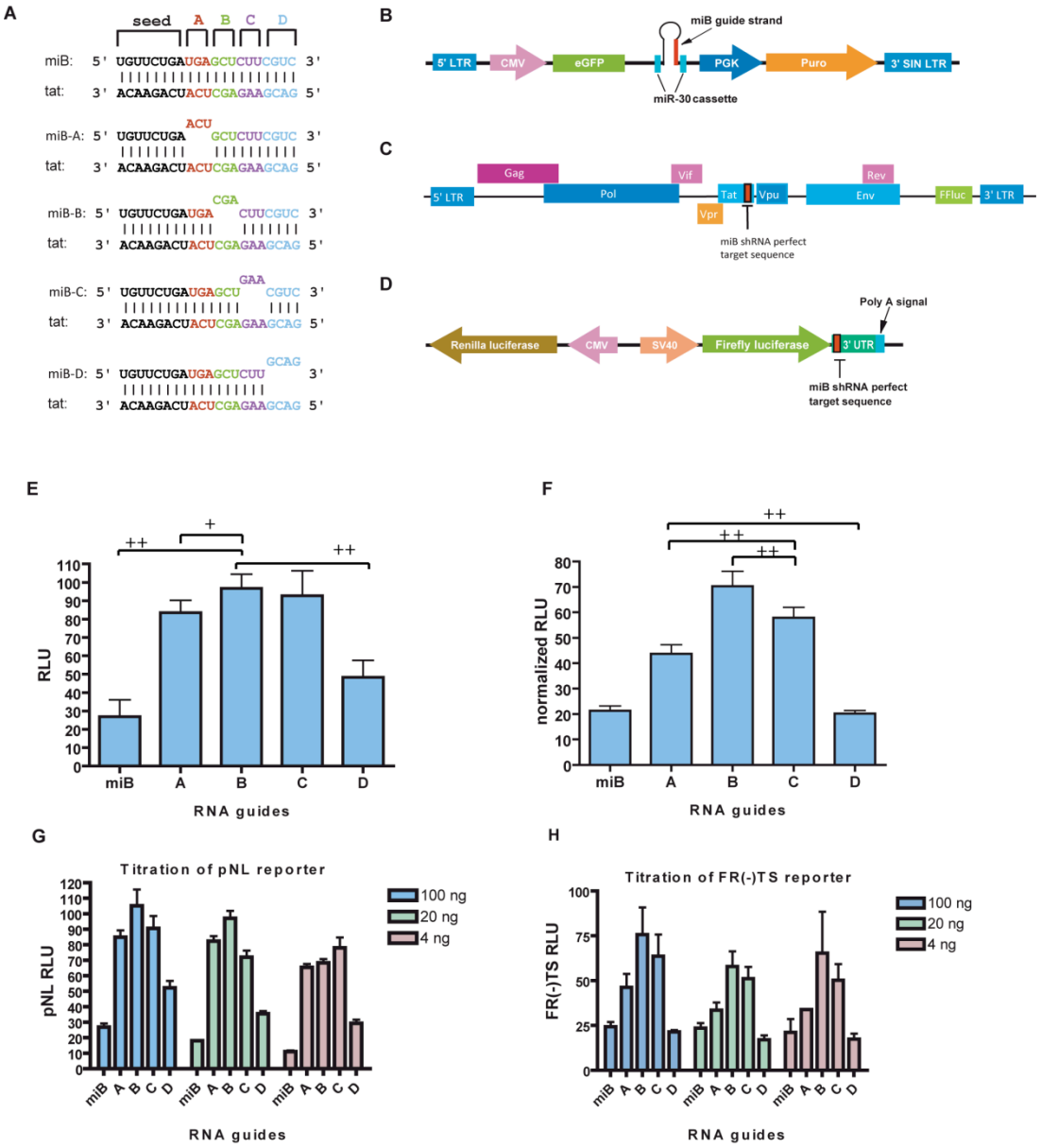


Figure 11. Silencing profile of the coding region and 3'UTR sites in the reporter plasmid

Figure 11. Silencing profile of the coding region and 3'UTR sites in the reporter plasmid

(**A**) MiB and designed single-module guides. The perfect complementary shRNA miB; then, from top to bottom, the mismatched guide RNAs miB-A (nts 9-11); miB-B (nts 12-14); miB-C (nts 15-17); and, miB-D (nts 18-21).

(**B**) The pPRIME vector used to clone all shRNAs. The guide strand is located at the 3' half of the stem loop structure (red).

(**C**) The pNL-luc plasmid is a luciferase reporter that contains the HIV-1 genome. The luciferase reporter gene (green) is located near the 3' LTR of the viral genome.

(**D**) The dual luciferase reporter plasmid FR(-)TS in which the cloning site of the target sequence is located in the 3' UTR of the firefly reporter gene. The firefly and renilla luciferases are transcribed in opposite directions.

(**E**) The repression profile of mismatched shRNA on the pNL-luc reporter. The plus sign (+) indicates the student t-test for the comparing columns yields $p < 0.05$; double plus signs (++) $p < 0.01$. The same convention is followed for panels **F-H**.

(**G**) Titration of the 3'UTR FR(-)TS reporter has limited effects on silencing.

(**H**) Titration of the HIV pNL-luc target reporter has limited effects on silencing.

To test the mismatched shRNAs, we used the pNL4.3-luc reporter construct, which contains the complete HIV genome with a disabled *env* gene (**Fig. 11C**). Since effective endogenous MREs are often located in the 3'UTR of their mRNA (Gu et al., 2009), we constructed a dual luciferase reporter, FR(-)TS, which embeds the miB MRE in the 3'UTR of the firefly luciferase (**Fig. 11D**). The MRE is located 29 nts downstream of the firefly luciferase stop codon, which is within a region (15-300 nts from the stop codon) associated with a high density of mRNA-bound AGO2 protein in the HITS-CLIP assays conducted by the Darnell group (Chi et al., 2009). To test whether this reporter construct functions properly, we mutated individual nts to their complementary nts in the seed of miB between position 1 and 6. As a result, we observed a significant abolishing effect of the repression compared to miB (**Fig. 12A**), confirming the reporter system is capable of measuring one-nt mismatch effects.

Consistent with the previous report, miB effectively repressed pNL-4.3 reporter gene expression, with a 75-80% knockdown efficiency relative to vector-only transfected cells (Boden et al., 2004). However, reporter gene silencing by mismatched small RNA guide was greatly abolished except for miB-D, which retained more than 50% of the silencing capability (**Fig. 11E**). ShRNAs (or miRNAs) that partially base pair with the HIV target sequences in the non-seed regions were strikingly ineffective in repressing the viral target (Boden et al., 2003; Houzet et al., 2012). As previously reported, we observed at least 80% loss of repression due to a mismatch of three nts in module A, B, or C. When FR(-)TS was used as the target construct, all guide strands showed improved silencing efficiency compared to the pNL-luc reporter construct (**Fig. 11F**). Also, a profile of repression efficiency emerges: miB and miB-D were the most efficient, followed by miB-A, then miB-C and miB-B, with more than 60% remaining expression. These results corroborate the findings that some non-seed nts are important for silencing (Friedman et al., 2009; Grimson et al., 2007), as well as the results of mono- and di-nucleotide mismatching guide RNAs (De et al., 2013; Doench and Sharp, 2004; Wee et al., 2012).

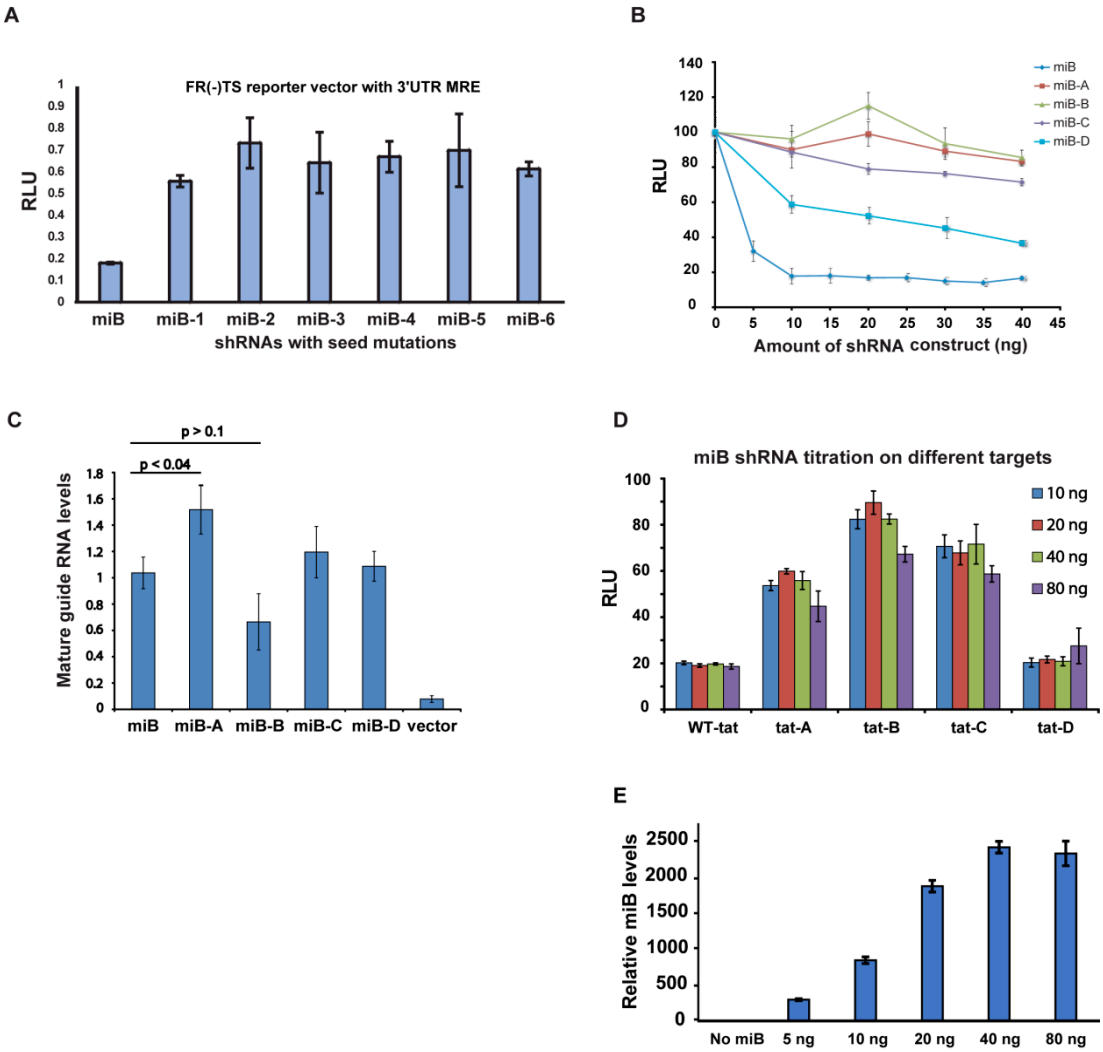


Figure 12. Verification of our assay system being reliable to assess effects of base pairing

Figure 12. Verification of our assay system being reliable to assess effets of base pairing

(**A**) Single-nt mutations in the seed of miB shRNA abolish its ability to repress FR(-)TS reporter expression.

(**B**) Titration of four different shRNA constructs (miB-A, -B, -C, and -D), across eight-fold concentration difference.

(**C**) Quantification of mature guides of miB and its module-altered variants by TaqMan RT-qPCR. Significant differences are shown on as p-values on top of horizontal bars for significant or near-significant differences.

(**D**) Titration of miB shRNA construct concentrations, covering 8-fold difference, used in combination with four different MRE reporters.

(**E**) TaqMan RT-qPCR determination of the level of mature miB at the transfection concentrations used in reporter assay in (**D**). The relative quantities are normalized to the background levels of miB in the negative control.

3.3.2 Variation in target concentration is not a dominant factor that perturbs the silencing efficiencies

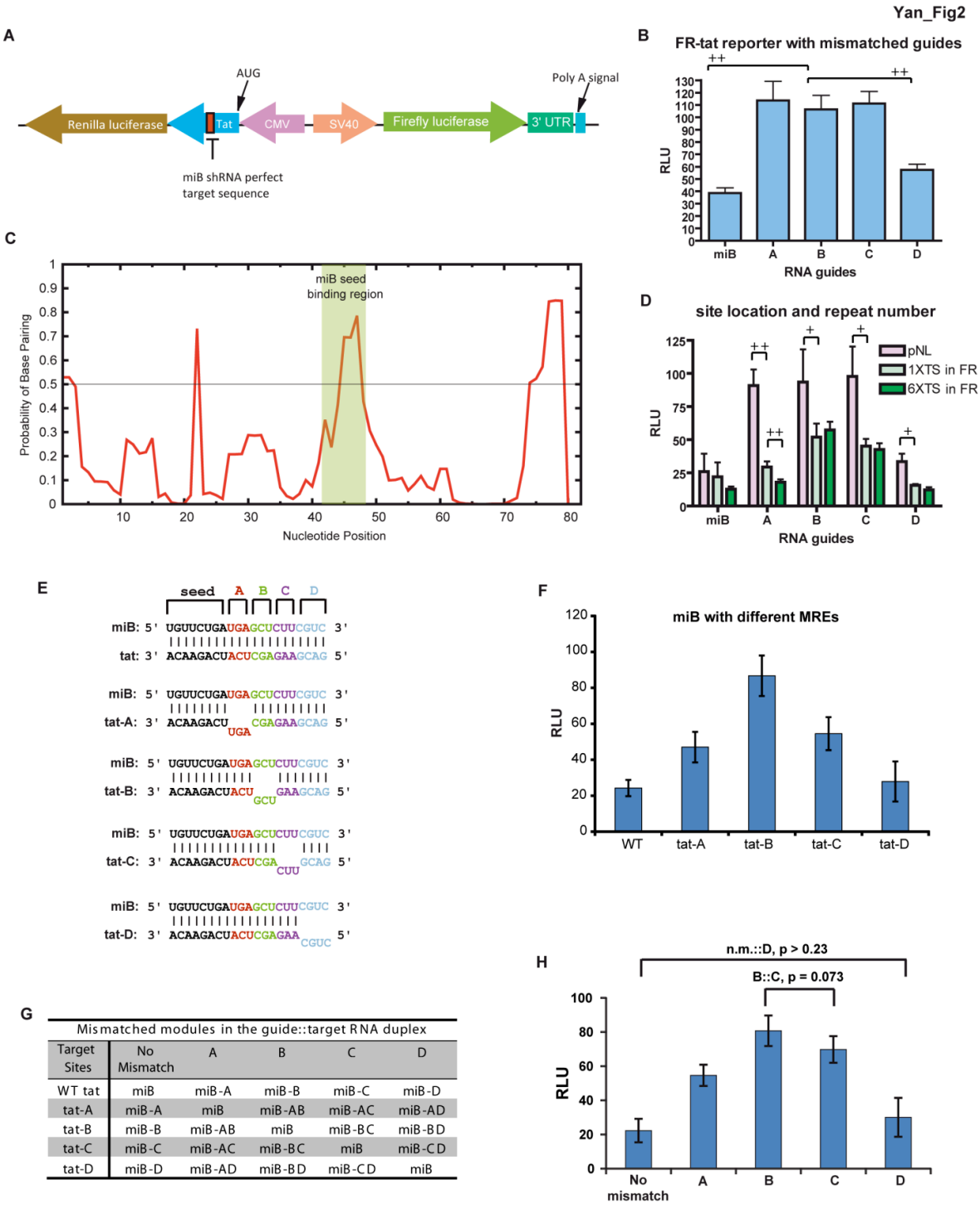
Previous studies have shown that concentration of the target or the miRNA affects the repression efficiency due to threshold effects and competition from ceRNAs (Bosson et al., 2014; Houzet et al., 2012). We optimized our assay conditions so that target concentration will not affect repression significantly in this study. Both pNL-luc and FR(-)TS reporters were titrated at a concentration range of 25-fold difference (4 ng, 20 ng, and 100 ng) with no significant alteration of the repression pattern. At higher concentrations of the target, the downregulation is less efficient in general (**Fig. 11G**). However, the efficiency is maintained with the FR(-)TS reporter in cells transfected with miB, miB-A or -D (**Fig. 11H**), even at the highest level. For these three guide sequences, the enhancement of repression did not exeed 20% for any guide even when the target concentration decreased 25-fold.

3.3.3 Confirmation of the effects of MRE location, accessibility, and repeats

Local structures in the target RNA may hinder the action of miRNA (Grimson et al., 2007), and the RNA genome of the HIV is known to contain rich secondary structure (Watts et al., 2009). To make sure that the improvement of silencing efficiency when moving the MRE from the pNL-4.3 to FR(-)TS construct is not due to the removal of global structure of the viral mRNA, we cloned exon1 of the *tat* gene into the dual luciferase construct after removing the *miB* MRE. Exon 1 of *tat* is inserted in-frame with and upstream of the renilla luciferase (**Fig. 13A**). We name this vector FR-*tat*. As the result, a fusion protein of *tat* and renilla luciferase is synthesized upon translation. Despite reduced light intensity, the renilla luciferase remains active and its expression is still sensitive to the downregulation of miB shRNA (**Fig. 13B**).

Mismatches in module A, B, or C greatly abolish silencing. This resembles the repression pattern displayed when pNL4.3-luc viral genome construct was used as a target. Energy calculation following the approach developed by the Ding group using *sFold* (Long et al., 2007) indicated the absence of stable local RNA structure (**Fig. 13C**, $p > 0.5$), rendering the MRE accessible (seed position 41-47). This corroborates with the high throughput screen results from the Elledge Lab, where an shRNAs library tiling the entire genome of the HIV was screened to probe for the accessible regions of the viral RNA genome (Tan et al., 2012). Therefore, the enhancement of repression reflects the fact that the MRE has been moved from the coding to a non-coding region of the mRNA, rather than the removal of either global or local secondary structure.

Multiple MREs in close proximity were shown to have enhancing effect on miRNA-mediated repression (Broderick et al., 2011; Doench et al., 2003; Doench and Sharp, 2004). To see whether the number of MREs on each target RNA could alter the repression profile significantly, we inserted the miB target site into the FR(-)TS vector six times in tandem, and tested the 1- and 6-MRE target constructs side by side with the pNL-luc reporter (**Fig. 13D**). The 6-MRE in the 3'UTR has enhancing effects on silencing for miB, miB-A, and miB-D. However, no significant changes were observed for miB-B and miB-C. We concluded that the number of MREs in the 3'UTR influences the silencing efficiency, but to a much less extent than their location.



Mismatched modules in the guide::target RNA duplex					
Target Sites	No Mismatch	A	B	C	D
WT tat	miB	miB-A	miB-B	miB-C	miB-D
tat-A	miB-A	miB	miB-AB	miB-AC	miB-AD
tat-B	miB-B	miB-AB	miB	miB-BC	miB-BD
tat-C	miB-C	miB-AC	miB-BC	miB	miB-CD
tat-D	miB-D	miB-AD	miB-BD	miB-CD	miB

Figure 13. Silencing profile in FR(-)*tat* and pNL-luc reporters resemble

Figure 13. Silencing profile in FR(-)*tat* and pNL-luc reporters resemble

(A) Dual luciferase construct FR(-)*tat* contains the first exon of the *tat* gene of HIV upstream of the renilla luciferase, creating a fusion protein of *tat* and renilla luciferase; the miB MRE is in the *tat* gene.

(B) Repression profile on the FR(-)*tat* reporter. The plus sign (+) indicates the student t-test for the comparing columns yields $p < 0.05$; double plus signs (++) $p < 0.01$. The same convention is followed for panel D.

(C) Secondary structure calculation of the miB target sequence within the *tat* gene by *SFold*. The vertical axis indicates the probability of being single-stranded. A horizontal line indicates the threshold of $p = 0.5$.

(D) The silencing profile is more sensitive to MRE location than MRE repeat numbers. Reporter expression levels of pNL-luc vector (blue bars), or FR(-)TS vector that contains the target sequence either one time (red bars) or six times in tandem (green bars) in the presence of mismatched miB variants.

(E) Base pairing between engineered sites that mismatch the miB shRNA at modules A-D. Each site is cloned into the same FR-reporter with the flanking regions from the *tat* gene.

(F) Dual luciferase assay when miB shRNA construct is used in combination with all four site reporters. Firefly luciferase levels were first normalized to renilla light, then normalized to the non-repressed level of each particular reporter construct (N=4).

(G) A table of shRNAs used in combination with each target site reporter to reconstitute the same mismatched positions in modules A, B, C, and D. The first row of the table indicates which module is mismatched in the guide::target duplex. The first column on the left is a list of mismatched module site reporters. Each entry in the table is a miB-modified shRNA with mutated modules used in combination with the target site of that row. Sequences of guide::target duplexes are listed in **Fig. 14**.

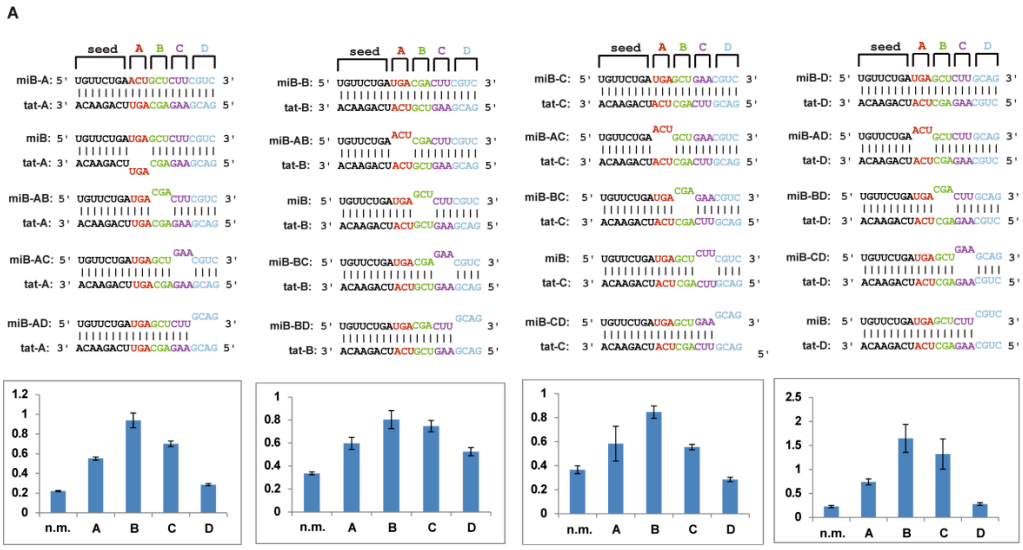
(H) Synthesized repression profile from reporter assays results by testing the 25 guide-target combinations in the table. Indistinguishable columns heights are indicated by bars on top of the figure.

3.3.4 The pattern of repression levels is not associated with the levels of mature guide RNAs

We transfected shRNA constructs with eight-fold differences in quantity. Downregulation levels appeared to be resistant to such perturbations, indicating that the guide-AGO2 biogenesis pathway was already saturated at half of the amount of guide RNA constructs used (i.e. 20 ng) (**Fig. 12D**). To further confirm that the pattern is not due to differences in mature guide RNA levels, we measured them using TaqMan RT-qPCR (Luo et al., 2012) (**Fig. 12 E**). We observed no significant differences for miB, miB-C, or miB-D. However, the levels of miB-A and miB-B are significantly different, respectively 1.5 and 0.4 times that of miB.

To address the concern of whether the profile of repression efficiency truthfully reflects the positional effects of the mismatched nts during the targeting process rather than the efficiency of processing and AGO-loading, we altered the sequence in the target, rather than the guide, to create the same mismatches in the four modules when using miB as a guide. Using the same design rationale for mismatches in the guide, four mutated target sequences, tat-A, -B, -C, and -D, were cloned into the 3'UTR of the same dual luciferase reporter (**Fig. 13E**), and we observed a similar profile (**Fig. 13F**). To confirm that this profile is stable with different amounts of mature miB guide RNA, we titrated the miB construct at eight different fold concentrations. Again, the same profile emerged (**Fig. 12F**). Then, using TaqMan RT-qPCR, we quantified the mature miB at each transfected concentration, and found that variations in mature miB abundance is not related to the observed pattern (**Fig. 12G**). We used 20 ng of each shRNA construct for transfection, where the mature levels can vary linearly with that of tranfected DNA. However, within the variation range, no significant difference in repression levels could be detected. This confirmed that although the mature levels of the guide RNAs may differ by up to 1.5 times, such as in the cases of miB-A and miB-B in the previous experiments, the repression profile is not affected and is solely due to the positional effects of the mismatches in the targeting process.

To make sure that these observations were not biased by a particular guide RNA or a particular MRE, we reconstituted the mismatches of the four modules in different target sites and shRNA-target combinations. Along with the wild-type, four additional sites were tested in combination with five sets of guides. Along with the fully complementary guide for each site, 25 different combinations were tested in total (**Fig. 13G**). The nts at the mismatches as well as the surrounding sequences of the modules differ in each combination (**Fig. 14A**). For each module, we averaged the repression values obtained from all sites to produce a synthesized repression profile (**Fig. 13H**). Again, the four modules are pair-wise distinguishable (**Fig. 14B**).



p-values of pair-wise comparisons in the synthesized profile

Modules	A	B	C	D	n.m.
A	NA	0.00107	0.00949	0.00545	5.27E-05
B		NA	0.0729	5.1E-05	2.82E-06
C			NA	0.000354	7.26E-06
D				NA	0.233

Figure 14. Different combinations of guides and target sites to generate mismatches at modules A-D to produce the repression profile

Figure 14. Combinations of guides and target sites to generate mismatches at modules A-D

(A) Combining different sets of shRNAs with the no mismatch sequence (n.m.) and four mismatched modules in four different sites in the reporter constructs. Each site was tested to derive its own repression profile. These four profiles were combined to the wild-type profile to generate the synthesized profile shown in **Fig. 13G**.

(B) Statistical significance of pairwise comparison of the mismatched expression levels from mismatched modules in **Fig. 13H**. Each entry is a p-value obtained by two-tailed unpaired student t-test assuming unequal variance.

3.3.5 Sequence alterations in the non-seed region display a decidable pattern in repression levels

To grade the relative importance of each module in their ability to influence gene silencing, we combined the wild-type site with all six possible double-module variants of miB: miB-AC, -BD, -AD, -AB, -BC, and -CD (**Fig. 15A**). This produced a spectrum of silencing effects when luciferase expression was monitored (**Fig. 15B**). We grouped the reporter level of each single-module with those of the double-module variants that contain it (**Fig. 16A**). A pattern of indistinguishable reporter levels emerged from these expression levels (**Fig. 15B**; compare the columns of the same color), as well as from their associated *p*-values (**Fig. 15C**; Student's *t*-test; p-values in **Fig. 16B**). We deciphered the following information: when the seed is perfectly matched, the B-module has the most decisive effect on silencing because it determines how base pairing in the rest of the non-seed nts contribute to silencing. Such decisive power of the modules decreases following the order of module C, A, and D.

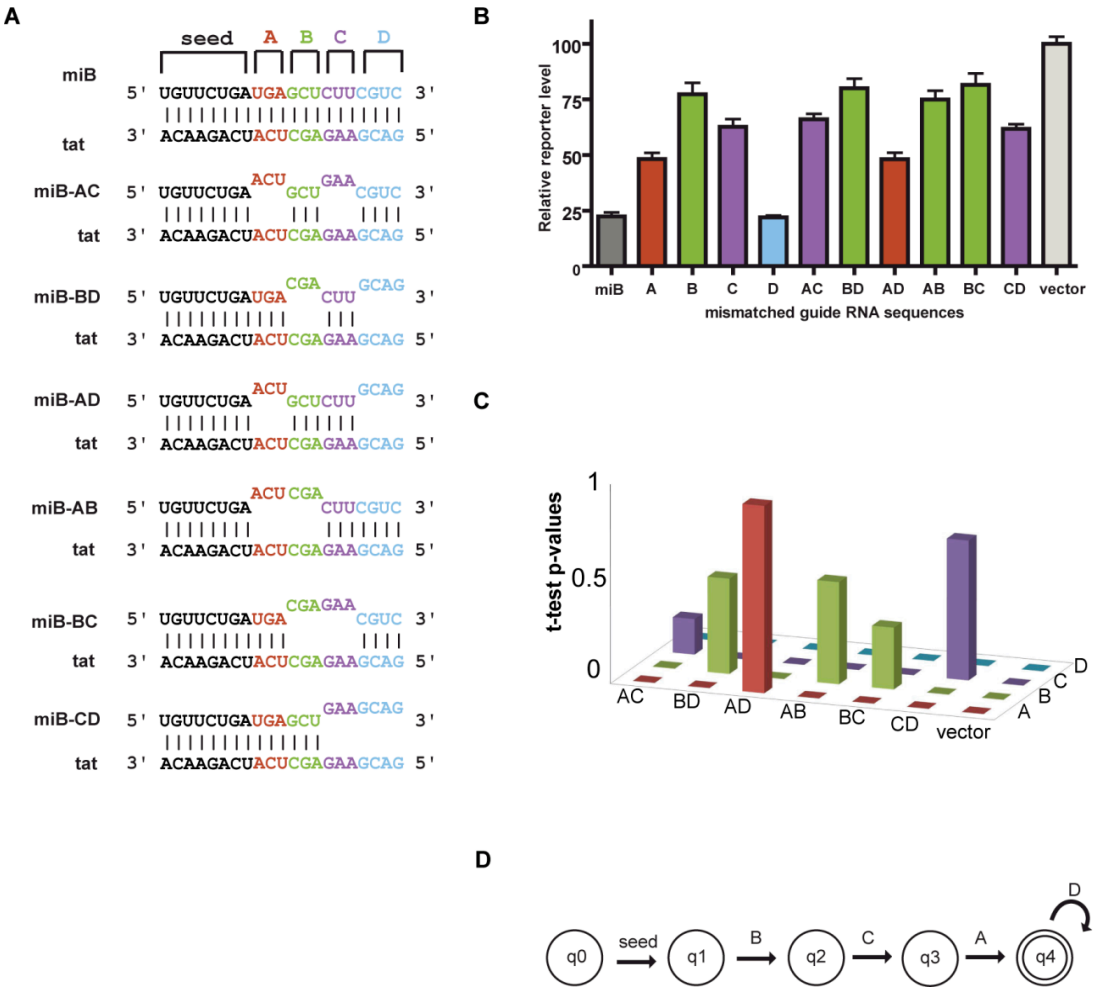


Figure 15. Combined effects of mismatches reveal the interdependency between the modules.

Figure 15. Combined effects of mismatches reveal the interdependency between the modules.

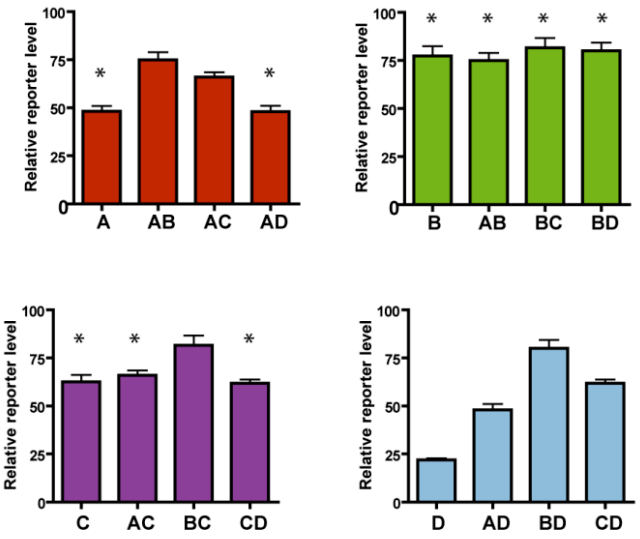
(**A**) MiB and double-module guides. All possible combinations of two mismatched modules are listed.

(**B**) The repression profile of all miB variants (N=4). The columns where the expression levels cannot be distinguished are of the same color.

(**C**) A 3-D representation of the p-values of the student t-test results of comparing the efficiencies of all miB variants. When the pairwise comparison is not able to distinguish the two guide RNAs by their residual reporter expression levels, a large p-value shows up as a tall column on the graph.

(**D**) State diagram of the proposed sequential recognition model for AGO2 slicing. Double circles represent the accepted state, which is defined as the most efficient slicing state.

A



B

Table of p-values of two-tailed t-test, unequal variance

	miB-AC	miB-BD	miB-AD	miB-AB	miB-BC	miB-CD	Vector
miB	6.161E-20	1.281E-15	1.702E-12	2.100E-15	1.649E-13	9.882E-22	2.448E-22
miB-A	1.578E-09	1.913E-11	9.404E-01	1.640E-10	5.500E-10	1.116E-07	1.194E-19
miB-B	1.791E-03	4.777E-01	5.350E-09	5.124E-01	3.101E-01	6.153E-05	4.796E-07
miB-C	1.777E-01	4.800E-06	4.483E-06	2.565E-04	1.203E-05	7.019E-01	1.016E-14
miB-D	9.833E-17	6.257E-14	3.358E-11	8.063E-14	7.403E-13	1.195E-18	1.119E-18

Figure 16. Repression profile of miB target sorted by mismatched modules

Figure 16. Repression profile of miB target sorted by mismatched modules

(**A**) Student's t-test p-values are obtained by comparing the repression levels using single- and double-module mismatched guide RNAs**;** double-module shown with each corresponding single-module. Asterisks on top indicate the expression levels that are indistinguishable.

(**B**) Tabulated p-values used in the above graph from the pairwise comparisons.

3.3.6 Establishing a computational model using the pattern

To consistently apply this rule to evaluate the targeting efficiency of miRNA-mediated repression, we built a computational tool that emulates the decision-making process of AGO2. AGO2 can be modeled as a multi-state machine, depicted in a Deterministic Finite Automaton (DFA) (**Fig. 15D**). The guide-loaded AGO2 first recognizes bps in the seed. Seed pairing is followed by base pairing of the nts in module-B. When the bps in the module-B are recognized, AGO2 transitions to the next state, allowing base pairing of the nts in the C-module to be recognized, followed by the A-module. Since a mismatched module-D is indistinguishable from miB, the slicer activity is likely to be fully functional once modules A, B, and C are all base paired. For this reason, we defined the accepted state of the DFA, q_4, i.e. where slicing can occur. This is consistent with the fact that the miRISC tolerates a loop in module-A, and that mismatches in module-D enhances the release of the miRNA from the miRISC (De et al., 2013), which is an independent step of the mechanism of the slicer activity of AGO2. The DFA describes a recursive algorithm that asserts the rule of evaluating the efficiency of a guide RNA. We implemented this model in a program called *MicroAlign* as a stand-alone Windows application. The first step of the program is to align the guide and the target strands to make sure that a reasonable conformation of the duplex is scored. Then, the miScore, which quantitatively reflects the silencing efficiency, is calculated. We observed a very strong correlation between miScores and the expression levels of our reporters (**Fig. 17A**; $r^2 > 0.98$, $p < 2.6\times10^{-10}$).

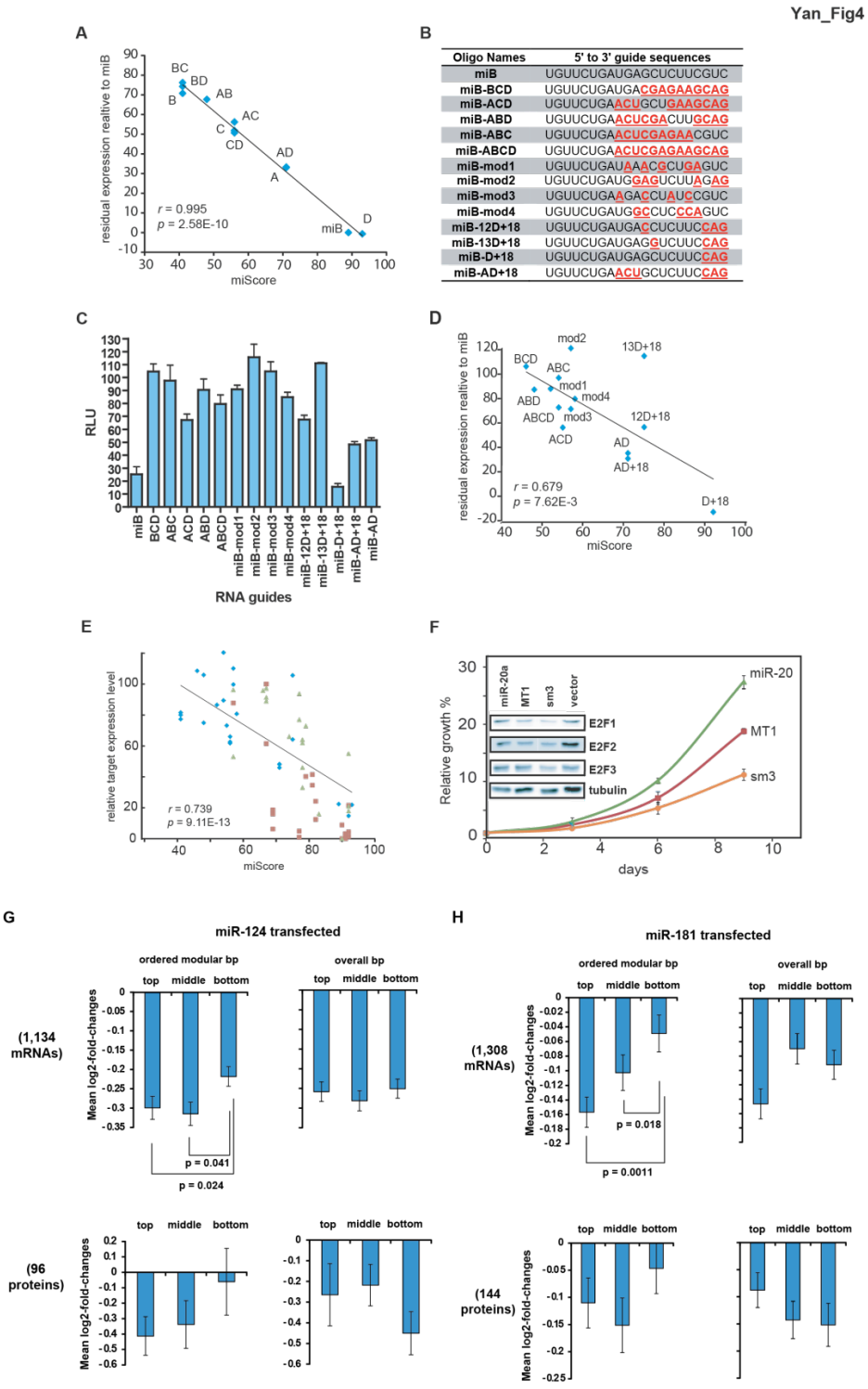


Figure 17. Validation of the non-seed model

Figure 17. Validation of the non-seed model

(A) Pearson correlation between reporter assay results and miScores (single- and double-module guides).

(**B**) miB and additional 13 guide sequences. Mismatched nts are in red and underlined.

(**C**) RLU values of the guides in (C) as measured in the FR(-)TS reporter assay.

(**D**) Pearson correlation between the reporter assay results and the miScores without alignment (13 additional guides).

(E) Pearson correlation between published silencing efficiencies and miScores. miB-based guides (blue diamonds); Wee et al. dataset (green triangles); and, Robertson et al. dataset (red squares).

(**F**) PC3 cell growth curves of miR-20a, MT1, and sm3. Juxtaposed are the Western blots of E2F factors when miR-20a, MT1, and sm3 were present.

(**G**) Mean $\log_2$ fold changes of the targeted mRNAs (top two panels) and proteins (bottom two panels) in miR-124 transfected cells binned by miScores (left) and scores not considering the modular order of the base pairs (right).

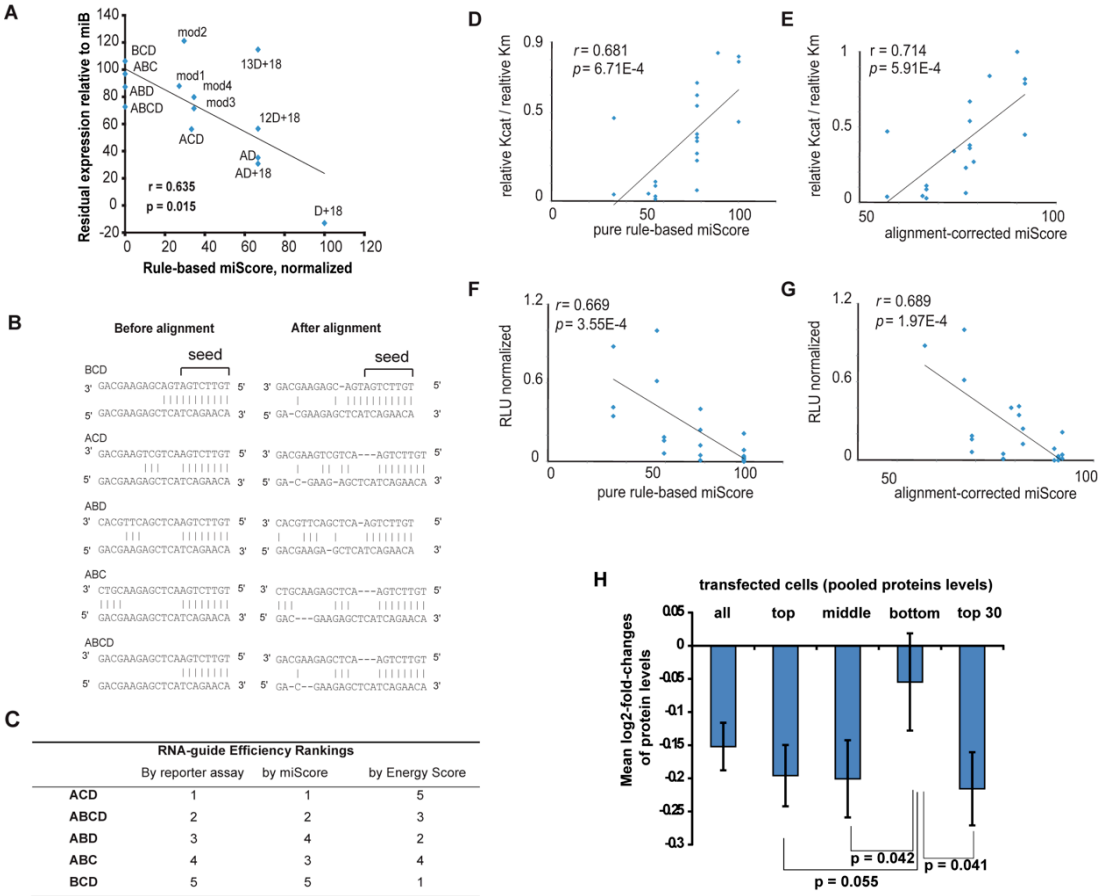
(**H**) Same as (G), but for miR-181 transfected cells.

3.3.7 Correlation with larger mismatched regions

We engineered RNA guides that contain at least three of the four mismatched modules (**Fig. 17B**, rows 2-6), as well as combinations of random mismatches (**Fig. 17B**, rows 7-14). From the reporter assay results (**Fig. 17C**), we observed again a high accuracy of the miScores (**Fig. 17D**, $r^2 \sim 0.50$, $p < 0.01$). This holds even when no alignment is performed (**Fig. 18A**). Inaccuracies of the free energy model mostly occur when the mismatches are in more than two modules, whereas our alignment algorithm identifies alternative bps (**Fig. 18B**) that improves the ranking of predicted activities of such guides (**Fig. 18C**).

3.3.8 Correlation with other siRNA studies

The analysis of third-party published data further confirmed the strong correlation between miScores and silencing. We used: i) the catalytic efficiency (K_{cat}/K_m) measured for AGO2 *in vitro*, where mismatches were systematically generated in the guide RNA (Wee et al., 2012); and, ii) miRNA sponges engineered with dinucleotide mismatches tiling the entire non-seed region (Robertson et al., 2010) (see **Fig. 18E** and **F**; and with alignment **Fig. 18G** and **H**). We pooled these data and computed the Pearson correlation between miScores and experimental expression levels (**Fig. 17E**; $r^2 > 0.5$; $p < 10^{-13}$).



RNA-guide Efficiency Rankings			
	By reporter assay	by miScore	by Energy Score
ACD	1	1	5
ABCD	2	2	3
ABD	3	4	2
ABC	4	3	4
BCD	5	5	1

Figure 18. Alignment step improves efficiency prediction

Figure 18. Alignment step improves efficiency prediction

(**A**) The *MicroAlign* algorithm robustly predicts the efficiency of guide strands with good accuracy without the alignment step.
(**B**) When more than two trinucleotide modules are mismatched at the same time, alternative bps can occur between the strands in each alignment. The guide sequence is listed on top and the target sequence at the bottom. The guide sequence is written in the 3' to 5' direction and the target sequence in the 5' to 3' direction.
(**C**) Comparing the measured and the predicted silencing efficiency of the guide sequences that mismatch in at least three modules, our program produced a better ranking than the conventional free-energy model. (**D**) Pearson correlation between miScores and k_{cat} /K_m values (Wee et al.) calculated without the alignment step.
(**E**) Pearson correlation calculated with the alignment step for the same guide sequences in (**D**).
(**F**) Pearson correlation between miScores and luciferase assay results (Robertson et al.) calculated without the alignment step.
(**G**) Pearson correlation calculated for the same guide sequences in (**F**) calculated alignment step.
(**H**) Mean $\log_2$ fold changes in protein levels of the miScore-evaluated targets were pooled from cells that were transfected with miR-124, miR-181, and miR-1, respectively. The mean target protein level was taken for 293 proteins. The pooled protein levels were placed in three equal size bins: top, mid, and bottom. The mean value of the top 30 targets evaluated by miScore was also computed (top 30).

3.3.9 Enrichment in designing effective artificial miRNAs

We further validated the model by showing how it enriches the design of efficient smartRNAs. As previously established, when the synthesis of multiple isoforms of the E2F protein is inhibited using smartRNAs, PC3 cell growth and proliferation are compromised (De Guire et al., 2010). In the previous study, we used the program *MultiTar* developed in our laboratory to obtain a list of possible guide sequences against three E2F isoforms (E2F1-3). Here, with the same design principles of *MultiTar*, we used *MicroAlign* to score the efficiency of the designed anti-E2F smartRNAs. We then tested the top five scored designed smartRNAs, sm1-5 (**Fig. 19A**), alongside with the previous best smartRNA we tested, MT1. We compared the protein levels of the E2Fs (**Fig. 19B**) and found that three smartRNAs, sm3-5, significantly knockdown (> 30%) all three isoforms (**Fig. 19C**). Plotting relative protein levels against the predicted miScores, we found that a cut-off score of 55 selects efficient guide strands (**Fig. 19E**). Comparing to the positive control, MT1, three of the five new smartRNAs knocked down E2F1 to a similar degree or more, while four of the five new ones knocked down E2F2 or E2F3 more effectively. Following a nine-day growth assay of PC3 cells, sm3 inhibited cell growth more efficiently than MT1 (**Fig. 19F**), while sm4 and sm5 inhibit cell growth comparably to MT1 (**Fig. 19D**).

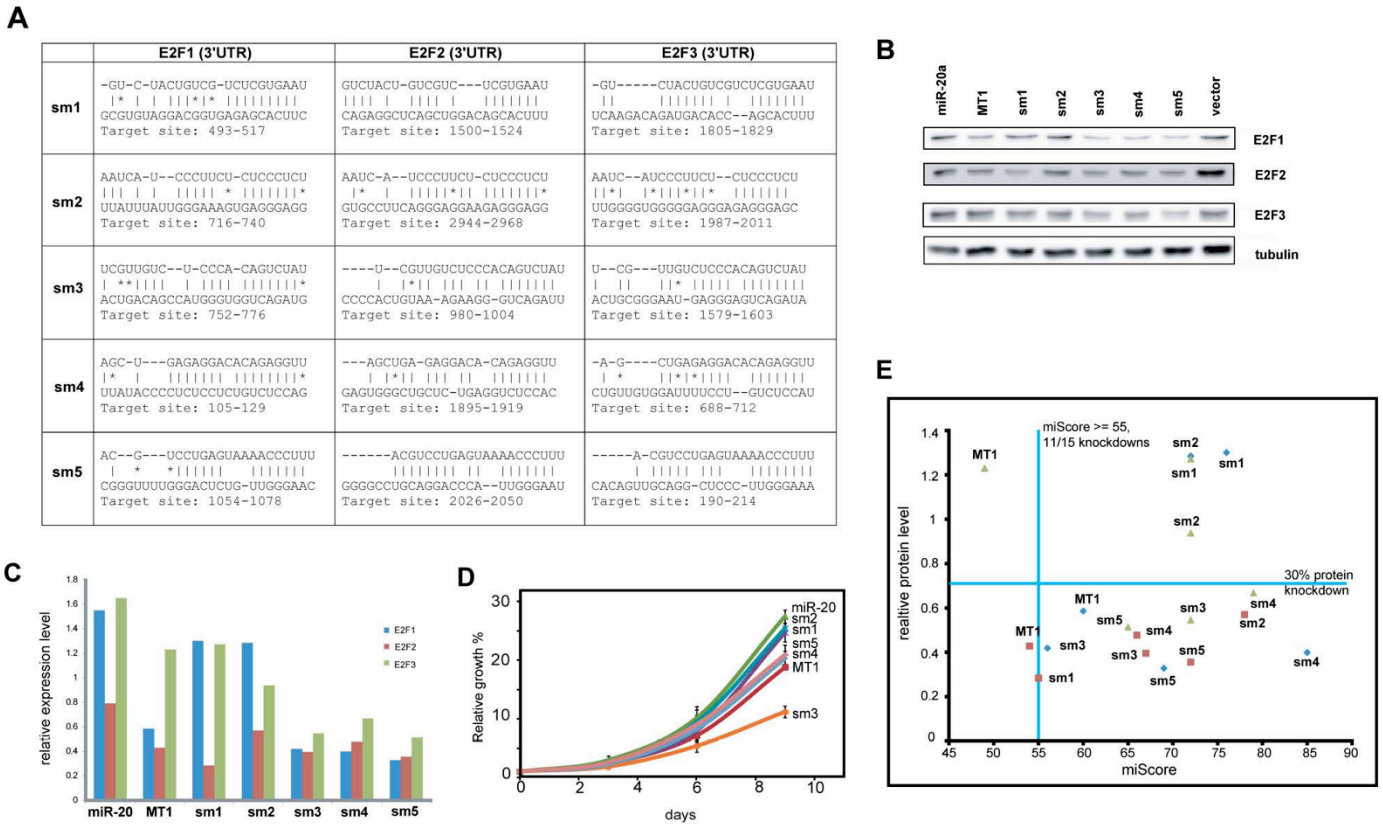


Figure 19. Validation of the model by designing efficient artificial miRN

Figure 19. Validation of the model by designing efficient artificial miRNAs

(**A**) The enrichment of smartRNA designs further validates the model. Top smartRNAs designed by MultiTar then selected using miScores. The guide RNA sequence is on top, in 3' to 5' direction, and the target sequence is at the bottom in 5' to 3' direction. Predicted Watson-Crick (|) and Wobble (*) base pairs were indicated. (**B**) Western blot of the target E2F proteins in the presence of each designed multi-targeting guide RNAs.
(**C**) The quantification of target protein levels from the Western blot in the top right corner, which shows all the target protein levels of the five tested top designs.

(**D**) Growth curves of all five tested new multi-targeting guides comparing to best previously tested ones.

(**E**) Relative protein levels are plotted against the predicted *miScores*. Data point shape corresponds to the target gene: blue diamond represents E2F-1, red square represents E2F-2, and green triangle represents E2F-3. The horizontal line represents 70% expression level threshold of an "effective" knockdown. The vertical line is the suggested *miScore* cut-off for selecting efficient guide designs.

3.3.10 Enrichment effect in public data from genome-wide studies

To further validate the suggested modular base pairing mechanism beyond the seed is playing a significant role in the targeting process of cellular miRNAs, we used standard public data used to benchmark miRNA target prediction programs (Baek et al., 2008). These data were generated by transfecting cells with three miRNAs, miR-124, miR-181, and miR-1, followed by mRNA and protein quantification, using, respectively, expression profiling and Stable Isotope Labeling with Amino acids in Cell culture (SILAC) and LC-MS/MS (liquid chromatography-mass spectrometry/mass-spectrometry). If the modular order of base pairing beyond the seed is a significant factor in target repression efficiency, then the targets that are top-ranked by the *MicroAlign* program should be enriched by effectively repressed mRNAs and proteins.

We pooled mRNA and protein levels of the three miRNA target genes, and calculated the mean differential repression levels as $\log_2$ fold changes. We first established the mean of the 293 protein targets, which is -0.15 (**Fig. 18H**). Then, we sorted the target protein levels by their *miScores* and split them into three equal sized bins, which we labeled “top”, “mid”, and “bottom”. We calculated the mean of each bin (**Fig. 18H**), and observed enriched repression efficiencies in the top and mid bins (near -0.2). The mid bin significantly differs from the bottom bin ($p < 0.05$). This shows that the *miScores* significantly enrich for more effectively repressed targets in the top two bins. Then, we took the mean of the top-30 proteins from each transfected sample, and we consistently observed the enrichment (mean < -0.22, $p < 0.05$, Mann-Whitney U test). Previously, similar mean repression at the protein level was achieved by the top-scored target predictions from PicTar and PITA; and an even better mean was observed from those of *TargetScan* (near -0.28). As for the programs that do not consider evolutionary conservation, they mostly yielded less signficant means (> -0.1) (Baek et al., 2008).

To confirm that the enrichment is due to the base pairing order beyond the seed, we modified the *MicroAlign* program so it calculates scores according to the total number of base pairs, without considering their order. We considered the enrichment for the miR-124 and miR-181 mRNA targets in the three bins. Using the non-modified *MicroAlign* to analyse 1,334 miR-124 targets, we consistently observed top and middle bins enriched in more efficiently repressed targets (**Fig. 17G**, top left panel), as well as a significant difference between the bottom and the top two bins ($p < 0.05$ in both cases, Mann-Whitney test). When we removed the base pairing order, the enrichment of efficiently repressed targets weakened in the top and middle bins, while the bottom bin got more efficiently repressed targets (**Fig. 15G**, top right panel). The same pattern was observed for the 98 protein levels measured by SILAC; however, statistical significance could not be established due to the low number of data points (**Fig. 17G**, two bottom panels).

For the 1,308 miR-181 mRNA targets, the same gradual enrichment from the bottom to the top bin was observed (**Fig. 17H**, top left), with a significant difference between the bottom and the two top bins ($p < 0.02$ and $p < 0.002$, respectively). Once again, more efficiently repressed targets are found in the bottom bin when the base pairing order was not considered (**Fig. 17H**, top right). The same pattern was observed at the protein expression levels (**Fig 17H**, bottom panels).

Taken together, when miRNAs were ectopically expressed, *MicroAlign* resolves the difference in repression efficiency of the targets solely based on the hierarchical order of base pairing beyond the seed. Hence, this modular base pairing mechanism beyond the seed is playing a significant role in the targeting process of cellular miRNAs and can be used to determine the repression efficiency of AGO2-dependent miRNA silencing guides.

3.3.11 Structural analysis supports the modular functioning of AGO2

Published data provided us with underlying AGO2 structural information to further substantiate our hierarchical model. We found that, in addition to the seed (**Fig. 21A**, left; PDB 4W5N), nts in positions 13-15 (B-module) are also exposed to the solvent when the seed of the RNA guide is annealed to an mRNA target (**Fig. 21A**, right). To see how base pairing occurs with the nts in modules B and C, we compared the structures of the AGO2 with and without the seed of the RNA guide annealed to an mRNA target (PDB 4F3T and 4W5R). In the bound structure, we observed that the PAZ domain pivots as a rigid body around the base of the α-7 helix (Ser371) by approximately 13˚ (**Fig. 20A**, angle θ). The PAZ-MID channel opens as the α-7 helix is displaced by 4 to 6 Å. The displacement is amplified at the 3' end binding site of the PAZ domain to 9.3 Å (**Fig. 20A**, top). Meanwhile, the number of hydrogen bonds between the PAZ domain and the 3' end of the guide RNA is reduced from five to two, or even zero in some structures (**Fig. 21B-D**). This indicates that a promoted release of the 3' end of the guide RNA from AGO2, which is required for miRISC-mediated cleavage. With its 3' end liberated, the guide strand is free to skip the central cleft of the AGO2, and progressively base pairs with the target strand in module-B, toward C-module-C, allowing for the formation of the RNA duplex beyond one turn (Wang et al., 2009b).

We docked a guide of fifteen bps by performing a structural alignment between the AGO2 (PDB 4W5O) and a non-cleaving mutant of the *Thermus thermophilus* AGO (PDB 3HJF) (Wang et al., 2009b) (**Fig. 20B**; RMSD ~ 1.15 Å). In our model, AGO2 is capable of accommodating the annealed duplex, consistent with the published model (Willkomm and Restle, 2015). Moderate clashes between the side chains of the α-7 helix and the RNA guide strand occurred. However, a rotation of the PAZ domain of a few degrees can remove the clashes. Indeed, the maximum rotation of the PAZ domain was evaluated to be approximately 25˚

(Robertson et al., 2010), which largely suffices. The coulomb potential surface of the AGO2 shows that the RNA-binding pocket has a natural tendency to open due to the presence of repulsive electrostatic charges lining its interior (**Fig. 21A**; blue regions). This view is also in agreement with a previous report stating that the AGO2 can recognize preformed duplexes and induce cleavage (Janas et al., 2012).

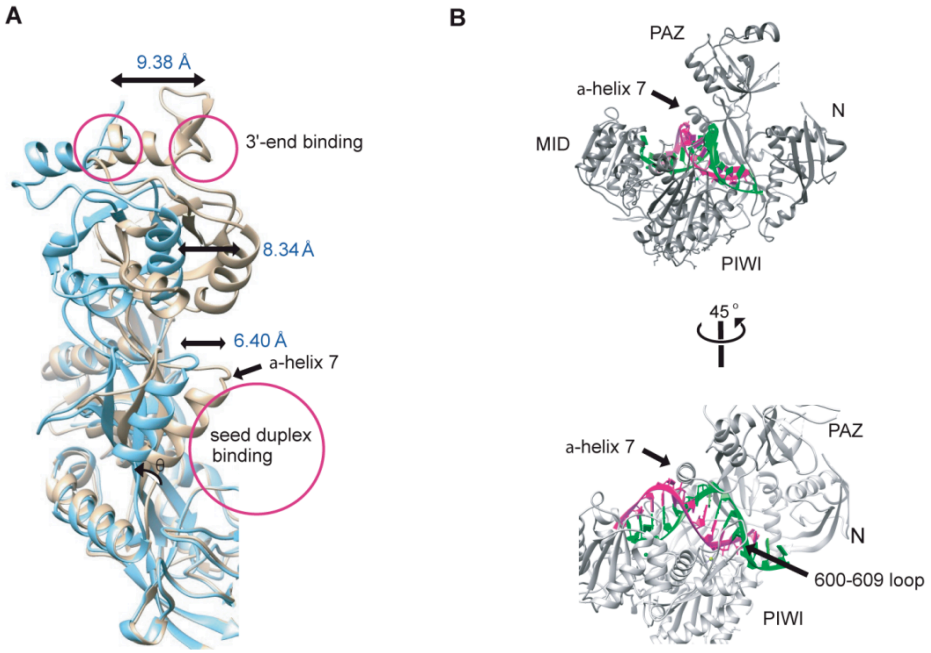


Figure 20. Structural analysis supports the proposed mechanism

Figure 20. Structural analysis supports the proposed mechanism

(**A**) Amplified view of α-7 helix. The rotation originates at the base of α-7 helix, with a visible angle θ of about 13° between the structure before and after seed pairing.

(**B**) The modeled accommodation of guide-target duplex of 15 bp in AGO2. The α-7 helix and the loop 600-609, which cause the narrowing of the central cleft of the AGO2, are interacting closely with the minor groove of the duplex. The guide strand is colored green; the target pink. The docking simulation was performed between a 15-bp guide-target duplex from *T. thermophilus* Argonaute protein (PDB 3HJF) and the human AGO2 structure (PDB 4W5T).

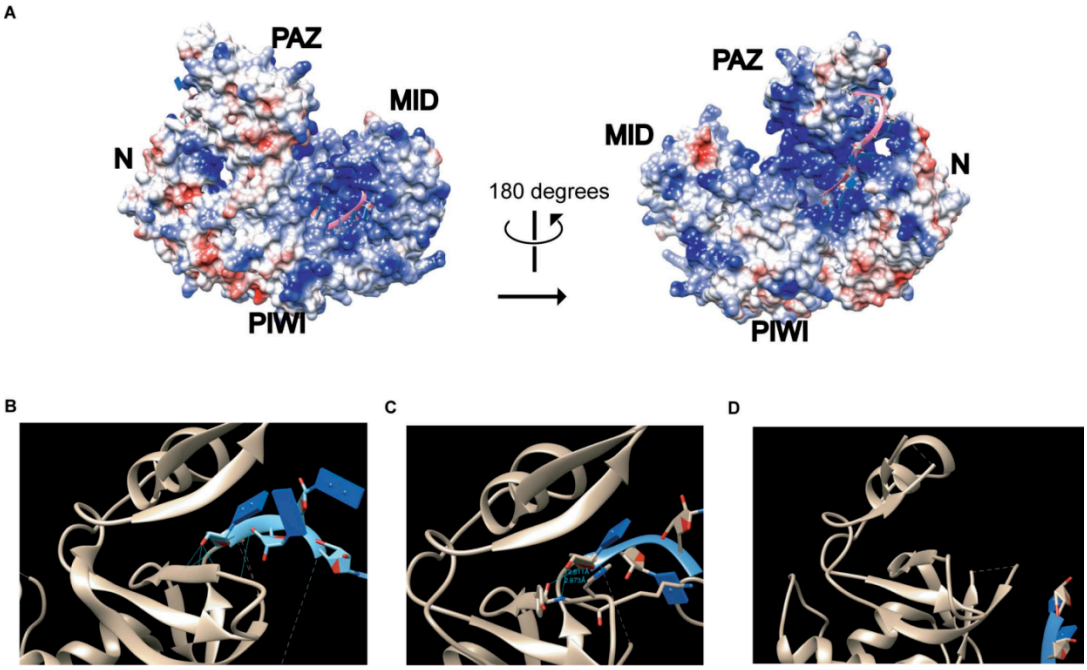


Figure 21. Additional features of the interaction between Ago2 and the guide strand

Figure 21. Additional features of the interaction between Ago2 and the guide strand

(**A**) The AGO2 structure with guide RNA (pink) were modeled and analyzed. We calculated the molecular surface and surface potential of AGO2 (PDB 4W5N): Positive surface potentials in blue; negative in red; neutral in white.

(**B**) Visualization of the 3'-end binding site of the PAZ domain (PDB 4F3T). Five hydrogen bonds can be observed between the RNA molecule and the peptide sequence.

(**C**) Visualization of the 3' end binding to the PAZ domain in 4W5N, in which a duplex is formed at the seed in the AGO2 structure. Only two hydrogen bonds could be detected.

(**D**) Binding of the 3' end to the PAZ domain in 4W5Q, in which a seed duplex was bound. In this case, the resolution of the 3' end of the guide strand is completely lost, indicating its high degree of freedom.

3.3.12 A possible model for non-seed nucleotide binding to AGO2

A recent structural study reported that base pairing at positions 9-11 is hindered regardless whether the 3' end is released or not due to the location of the α-7 helix and the 600-609 loop in protein hAGO2 (Schirle et al., 2014). Consequently, continued base pairing in the 5' to 3' direction along the guide RNA is interrupted at the central cleft of the hAGO2. Interestingly, the enhancement of cleavage activity by base pairing beyond the central loop indicates that some degree of base pairing is beneficial in the 3' supplementary region (Wee et al., 2012). Yet no such intermediate structure of 3' supplementary base pairing has been resolved, due to poor visibility. Slicing activity immediately following base pairing with the target may have made it difficult to observe a conformation bound to the target (Schirle et al., 2014). Structural data eventually became available for AGO2 bound to a guide strand (Elkayam et al., 2012; Schirle et al., 2014), a duplex of the guide RNA with a partial target (Schirle and MacRae, 2012; Schirle et al., 2014), and a catalytic mutant AGO2 bound with a guide-target duplex of 15 nts (Wang et al., 2009a).

Combining our experimental with the structural data, we suggest that the following sequence of events takes place for miB guide RNA to achieve AGO2-mediated silencing. As the seed of miB base pair with the target, the seed duplex is accommodated in the PAZ-MID channel. Suggested by previous studies, the narrowing of the channel forms a cleft and prohibits further base pairing immediately downstream of the seed. However, while the duplex pushes the α-7 helix outward and causes the PAZ domain to pivot, the 3' end of the RNA guide becomes less tightly bound to the PAZ domain, and thus more prone to be released. The free 3' end facilitates the base pairing process to "skip" the cleft and resumes in module-B, and then propagates to module-C. Once the RNA guide is bound to the mRNA target between nts 12 and 18, the duplex is formed on both sides across the narrow cleft of the PAZ-MID channel. The nts around the scissile phosphate (nts 10-11) eventually anneal with the RNA guide and "fit" into the cleavage site, either by further pivoting of the PAZ domain, which opens the channel, or with the help of twisting motions of the duplex formation on both flanks of the channel. When the

bases are complementary in the central module-A, efficient cleavage occurs. On the other hand, in the presence of mismatches, the cleavage efficiency depends on the protein tolerance for them. Moving along the steps of the duplex formation, the mRNA target becomes less and less likely to dissociate from the AGO2 complex. As the "dwell time" of the AGO2 complex on target gets greater, so as its chances to recruit protein factors for slicer-independent repression (**Fig. 22**).

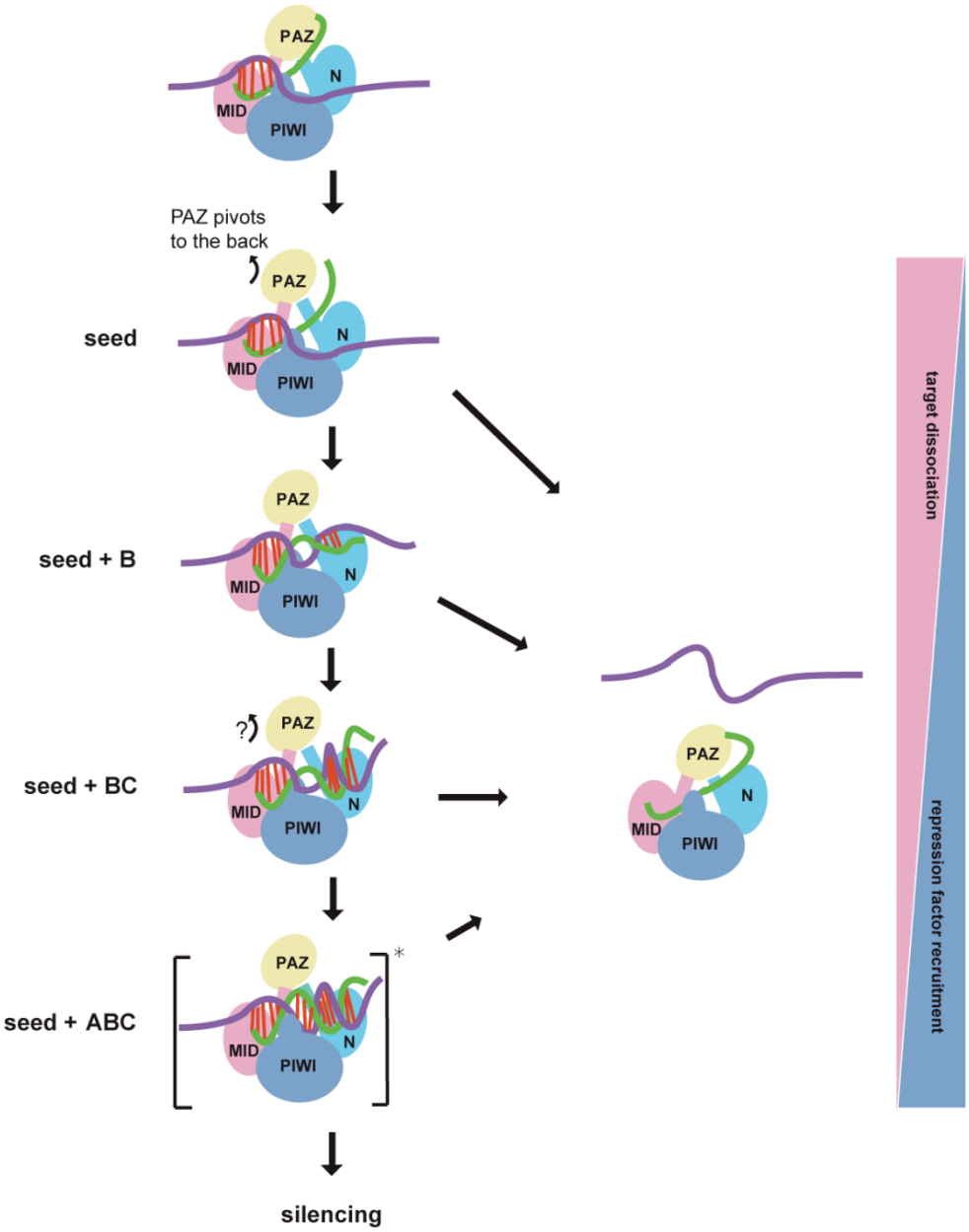


Figure 22. Summary of the skipped-propagation and coordinated annealing model

Figure 22. Summary of the skipped-propagation and coordinated annealing model

The step-wise nature of the AGO2 slicer activation process entails that it is a multi-state machine. In every step along its structural change, the efficiency of silencing becomes gradually enhanced and to the next complex structure (down arrows), or, in lack of base pairing, to the dissociation of the complex (right arrows). The greater the base pairing the most stable is the complex and longer the dwell time, which increases the probabilities to recruit repression factors (blue triangle), and decreases those of target dissociation (pink triangle).

3.4 Discussion

Mismatches introduced by a miRNA near the scissile phosphate of a target are generally tolerated in spite of a compromised endonucleolytic activity. In bilaterian animals, miRISC mediates repression predominantly via slicer-independent pathways. We found that a number of mismatches further downstream of the central region can impair repression to a greater extent following a hierarchical pattern, which became apparent when triplet mismatches were made in the non-seed region. This unique experimental design explains why this pattern remained elusive in the past despite several systematic investigations.

From this pattern, we built a computational model that evaluates the importance of base pairing beyond the seed in AGO2-mediated repression. This model suggests that when the seed region is perfectly base paired with the target, then the bps in module-B (nts 12-14) play a decisive role. This rule applies recursively to modules C (nts 15-17), A (nts 9-11), and D (nts 18-21), in this order. The idea that miRNA/guide-RNA pairing to targets is modular has been proposed and tested through structural, computational and reporter assays by other groups (described in the following section). In addition to this growing body of work, our model suggests that base pairing contributions follow a hierarchical decision-making process, which resolves the subtlety of the sequential base pairing events beyond the seed.

3.4.1 Simplicity and consistency of the sequential recognition model

Comparing with the nucleation-propagation model where the AGO2 is regarded as a two-state machine (Knott et al., 2014; Wee et al., 2012), our model also depicts it as a state machine with more states. State transitions take place in a stepwise fashion following a specific precedence of nt positions. Our results agree well with triplet mismatches made in a previous study of the cleavage activity of hAGO2 (Lima et al., 2009). The reason for the conservation signals detected around nt 13 (Grimson et al., 2007) and the outstanding contributions of the “supplementary” base pairing

measured by Wee and coworkers (Wee et al., 2012) hence become clear: the base pairing in this module mechanically determines whether other non-seed bps should be accounted for silencing, rather than providing a large contribution in free energy of binding. More recently, the Pasquinelli group suggested that certain classes of miRNAs are capable of 3' end pairing interactions (13-16) to outcompete miRNAs that support only seed pairing for a given site (Broughton et al., 2016). This observation provides additional evidence for the proposed model and the role of the central loop.

Previous reports showed that elongating the central loop enhances repression of engineered miRNAs. Loop scores were then assigned to computational models to evaluate the effects of such central bulges (Kiriakidou et al., 2004). Our model suggests that the enlargement of the central loop, which corresponds to module-A, is likely to relax the central portion of the target so to bypass the protein structure blockage more easily and promote downstream base pairing. It has been shown that the target release was the rate-limiting step for AGO2 slicer activity (De et al., 2013; Deerberg et al., 2013; Willkomm and Restle, 2015) and that mismatches at the 3' end of the guide RNA enhances slicer function. Our data corroborate these enhancing effects. However, we only observed enhancement under the premise that modules A, B, and C are paired. This may indicate that mismatches in these modules significantly slow down the formation of the pre-cleavage complex so that base pairing becomes the new rate-limiting step. With over 60 data points, including about half from third parties, the experimental measurements agree with the model with high confidence (**Fig. 17G**; p = 9.11E-13), indicating that the order of base-pairing beyond the seed can resolve difference in silencing efficiency of the RNA guides even when the free energy of base pairing is the same.

Using microarray and SILAC data, this ability to determine silencing efficiency among over 2,500 targets became evident at the mRNA level. Using three bins, the repression levels of the top ranked targets by *MicroAlign* were consistently higher than those in the bottom bin. *TargetScan*, *miRanda*, *PITA*, and *PicTar* were also able to bring similar enrichments. However, the scoring functions of these

programs were derived in part from statistics including many additional factors, and in particular evolutionary conservation of the target sites. *MicroAlign* is solely based on our AGO2 mechanistic model. Its ability to enrich effectively repressed targets at the protein level was statistically less significant (**Fig. 15GH**). This might be due to the fact that less data points were generated from SILAC. In support to this argument, the enrichment became statistically significant when we combined the protein levels from all three overexpressed miRNAs (**Fig. 16H**).

In total agreement with our model, it has recently been suggested that pairing in the miRNA:mRNA duplex does not move forward from the seed to the 3' end, and the idea of a second nucleation site is defensible since miRNAs prefer 3' supplementary pairing (Bartel, 2018). Our experimental results and our structural modeling provide the first evidence of this "skipping" model during target recognition by the miRISC. Moreover, our model further defines the order in which target recognition occurs beyond the seed, and module-B as the second nucleation site. Interestingly, a study using a massively parallel experiment reporter assay identified that miRNA position 14 (the last base in module-B) as the saturation point of mismatched bases that reduce repression (Vainberg Slutskin et al., 2018).

3.4.2 Limitations of the current model

First, the goal of the *MicroAlign* program is not to predict miRNA targets, but rather to calculate the silencing efficiency of possible guide::target duplexes. *MicroAlign* does not consider extrinsic factors such as target site location, AU content, target site accessibility, abundance, and evolutionary conservation. Statistical training and combinations of these factors were shown by genome-wide analysis to have similar predictive power (Baek et al., 2008). Since *MicroAlign* was designed to evaluate 7-8mer sites, its average performance of enrichment is upper-bounded by that of using these sites, which in the genome-wide analysis include false positives that introduce noise. Nevertheless, we observed enrichment of effectively regulated targets without combining or optimizing any additional factor. This enrichment was greater than those obtained by prediction programs that use

free energy without site conservation (Baek et al., 2008). This suggests that the order of base pairing beyond the seed plays a significant role in the regulation of the expression of the targets. Though this feature alone is not sufficient to predict mRNA targets, it can be used in the design of effective RNA guide sequences to inhibit simultaneously multiple targets.

Second, contributions to silencing by other mechanisms, such as deadenylation, decapping, and translational repression (Fabian et al., 2010; Wee et al., 2012) could not be clearly discerned individually in this study. We observed that mismatches at the scissile phosphate (miRNA positions 10 and 11; module-A) impair the slicer activity, which can be made even worse if combined with mismatches in modules B or C. This suggests that the pattern we detected also reflects silencing efficiency related to slicer-independent mechanisms. For instance, by comparing miB-A, -AB, -AC, and -AD, we observed the same hierarchical pattern of B > C > D, indicating that module-B directly contributes to slicer-independent repression as well. The precise mechanism by which these mismatches affect the slicer-independent pathways remains unclear. Perhaps the non-seed mismatches are capable of altering the AGO2's ability to recruit protein cofactors by changing its "dwell time" spent on targets (Chandradoss et al., 2015). Non-seed mismatches could thus produce different effects on the slicing and repression mechanisms. Simultaneously, they could reduce the slicing rate and define the time spent on a target, which determines the recruitment of slicer-independent repression factors. How this interplay produces the particular hierarchical pattern we observed is yet to be found.

3.5 Materials and Methods

3.5.1 Plasmid Construction

The Renilla luciferase control vector, SVR, was obtained by replacing the CMV promoter in the pcDNA-RlucII plasmid (a gift from the Mader lab) with an SV40 promoter. Briefly, the CMV promoter was removed by restriction enzymes *SpeI* and *HindIII* (New England Biolabs). The resulting linearized vector was gel-purified with QIAEX II ® Gel Extraction Kit. The SV40 promoter from the pGL3-control luciferase vector fragment was obtained by digesting the vector with *NheI* and *HindIII*. Gel purified SV40 promoter fragment was inserted upstream of the RlucII gene in pcDNA-RLucII vector by ligation using T4 DNA ligase (NEB).

The firefly-renilla opposite-sense target site reporter is referred to as the FR(-)TS construct, which contains both firefly and renilla luciferase reporter genes oriented in the opposite directions. In addition, a 76 bp region of the HIV genome containing the miB shRNA target site in the center (pNL4-3 vector, Accession number: AF324493, nts 5968-6044) was inserted into the 3'UTR of the firefly gene. Cloning of the target site was carried out by inserting the annealed oligonucleotides into the *XbaI* site upstream of the poly-A signal in the pGL3-Ctl reporter. For FR(-)TS vector, the annealed oligos are the following: the forward oligo sequence is CTAGAATGGCAGGAAGAAGCGGAGACAGC**GACGAAGAGCTCATCAGA ACA**GTCAGACTCATCAAGCTTCTCTATCAAAGCAT; and, the reverse oligo sequence is CTAGATGCTTTGATAGAGAAGCTTGATGAGTCTGAC**TGTTCTGATGAGC TCTTCGTC**GCTGTCTCCGCTTCTTCCTGCCATT. Bold letters represent the miB binding site. The reporter that contains six times of the target site does not include the flanking regions; rather, the 3'UTR insert is a tandem repeat of the target site only. Renilla luciferase gene was removed from pcDNA-RlucII plasmid by digesting the vector with *SpeI* and *XbaI* restriction enzymes. The gel purified (QIAEX II ® Gel Extraction Kit) renilla luciferase fragment was then inserted in the *NheI* site in Promega pGL3-control luciferase vector.

The FR(-)*tat* dual luciferase vector was constructed as follows. The FR(-)TS vector without insertion of the miB shRNA binding site from the previous step was used as a starting material. The vector was digested with restriction enzymes *XbaI* and *HindIII* from NEB, which creates a linearized vector for upstream insertion of the renilla gene. Subsequent gel purification was performed using QIAEX II ® Gel Extraction Kit. The first exon of the *tat* gene was amplified from pNL4.3-luc vector (a gift from the Cohen lab) with forward primer (5' to 3'): ATCCAAGCTTCCCGCCACCATGGCAGGAAGAAGCGGA, and reverse primer (5' to 3'): CGACTCTAGATGCTTTGATAGAGAAGCT. The PCR was carried out using 55 ˚C as annealing temperature. The amplified fragment was ethanol precipitated and digested with restriction enzymes *XbaI* and *HindIII*. Upon gel purification, the fragment was ligated with the digested vector at 16˚C overnight. The ligation mix was transformed into DH10B.

The vector pPRIME (a gift from the Pelletier lab) has been previously optimized for shRNA cloning (Lee et al., 2014; Malina et al., 2013; Mills et al., 2013). Designed guide-RNAs were cloned into the vector following miR-30-based shRNA cloning protocols (Dickins et al., 2005). Briefly, complementary oligonucleotides that contain the shRNA sequences were diluted to 100 μM in deionized water. Annealing reaction was carried out at 95 ˚C in annealing buffer for 5 minutes followed by slow cooling to room temperature. The annealed double-stranded oligonucleotides were then phosphorylated by T4 PNK (NEB). Ligation reaction was performed by combining doubly digested pPRIME by *XhoI* and *EcoRI* with the phosphorylation product of annealed oligos in T4 DNA ligase (NEB) reaction mix at 16˚C overnight.

3.5.2 Cell culture and monitoring shRNA efficiencies

HEK 293T (c17) cells (from ATCC) were maintained according to established conditions (De Guire et al., 2010). Cells were grown in DMEM (+L-glutamine) (Life Technologies) supplemented with 10% FBS, 100 U/mL penicillin/streptomycin at 37 ˚C and 5% CO_2. Cells were grown to confluence before

plating. For testing the efficiencies of mismatched guides, cells were plated in 96-well plates at ~20,000 cells per well 24 hours prior to the transfection. For assays that required growth in 24-well plates, cells were plated at ~100,000 cells per well.

The reporter plasmids and the shRNA plasmids were co-transfected into the cells using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. Along with 10 ng of shRNA plasmid, 5 ng of pNL-luc and 2 ng of SVR control vector were co-transfected into each 96-well; alternatively, 50 ng of the shRNA construct, 20 ng of the pNL-luc, and 10 ng of the SVR control vector were co-transfected into each 24-well. When an AGO2 expression construct is used, 25 ng of the AGO2D597A vector (Diederichs et al., 2008) (A gift from the Diederichs lab) was combined with the DNA mix described above and subsequently co-transfected into the cells

Luciferase assays were performed accordingly to established protocols adapted from the Duo-Glo Luciferase System (Promega). 48 hours post-transfection, cells were lysed with 1× Passive lysis buffer (Promega) and luciferase activity was assayed using the Dual-Glo Luciferase System (Promega). Luminescent light was measured on Veritas Microplate Luminometer (Turner Biosystems) (a gift from the Bouvier Lab). The ratio between the reporter and the control luciferase bioluminescence light was taken and then normalized to that of the negative control shRNA or empty vector, resulting in the percentage residual expression of the reporter gene.

3.5.3 Measuring reporter transcript and mature RNA guide abundance using qRT-PCR

RNA extraction was performed using TRIzol® reagent following manufacturer’s protocol. RNA was extracted from the same cells used in the luciferase assay. Either oligo-dT primer or random primer were used for the synthesis of cDNA from total RNA extracted according to previously established protocols (Kiethega et al., 2013). 800 ng of total RNA was used for each synthesis reaction in 20 μL of total volume using Invitrogen reagents (M-MLV Reverse

Transcriptase, Cat. No. 28025-021, Invtitrogen™). RNA was extracted from the same cells that were used in the luciferase assay and M-MLV was used to perform the cDNA synthesis.

The newly synthesized cDNA was diluted by a factor of 100 prior to real-time PCR. Each real-time PCR reaction mixture contained the diluted cDNA (1 µl), forward and reverse primers (250 nM), $MgCl_2$ (2.5 mM), dNTPs (0.2 mM), SYBR green (0.33X), buffer for Jumpstart *Taq* DNA polymerase and Jumpstart *Taq* DNA polymerase (0.25 U; Sigma) in a final volume of 10 µl. After denaturation at 95 °C for 6 min, samples went through 50 cycles of amplification (20 s at 95 °C, 20 s at 58 °C and 30 s at 72 °C). Melt curves were determined for each reaction and qPCR was performed using a LightCycler 480 (Roche Applied Science, Canada). Data was normalized using Renilla and HPRT as controls.

The detection of mature RNA guide molecules was performed following the polyA-based RT-qPCR protocol established previously (Luo et al., 2012; Zhang et al., 2008). Briefly, 20 µL of reaction contained 1 µL of reverse transcription products diluted 10-fold, 10 µM of forward primer, and 10 µM of universal reverse primer, 2 µL of Taq polymerase buffer (10X), 4 µL of 2.5 mM each dNTP, 0.6 U Taq and 10 µM of universal TaqMan probe. The mix is heated to 95 ˚C for 2 minutes prior to entering 45 cycles of 95 ˚C for 15 seconds followed by 60 ˚C for 1 minute. The reactions were carried out and measurements were taken on a StepOnePlus™ Real-Time System from Applied Biosciences. The forward primer sequences are as follows: miB: GTGCTGTTCTGATGAGCTCTTCGTC; miB-A: GTGCTGTTCTGAACTGCTCTTCGTC; miB-B: GTGCTGTTCTGATGACGACTTCGTC; miB-C: GTGCTGTTCTGATGAGCTGAACGTC; miB-D: GTGCTGTTCTGATGAGCTCTTGCAG; U6: ACGCAAATTCGTGAAGCGTTCCAT; Puromycin: TGACCGAGTACAAGCCCAC.

3.5.4 Cells and Retroviral-Mediated Gene Transfer

PC3 were obtained from American Type Culture Collection (ATCC) and cultured in RPM1 (Wisent) supplemented with 10% FBS (Wisent), 1% penicillin/streptomycin sulfate (Wisent), and 2 mmol/L L-glutamine (Wisent) at 37°C and 5% CO_2. Gene transfer was performed using retroviral particles produced in Phoenix packaging cells. Phoenix cells were transfected by calcium-phosphate precipitation with 20 μg of a retroviral plasmid (15 hrs at 37°C). The plasmids used were: shNTC (non-targeting control), MiR20, MT E2F(1), E2F Afa, E2F Afb, E2F Afc, E2F Afd and E2F Afe. After 48 hrs, the virus-containing medium was filtered (0.45 μm filter, Millipore) and supplemented with 4 μg/ml polybrene (Sigma) (first supernatant). Viruses were collected for an additional 8 hrs as before (second supernatant). For infections, the culture medium was replaced by the appropriate first and second supernatant on PC3 cells. Sixteen hours later, infected cell populations were purified by selection with 2 μg/ml puromycin for 48 hours.

3.5.5 Growth Curve

Twenty thousand cells per well were plated into 6 well plates. At the indicated times, cells were washed with PBS, fixed in 4% formaldehyde, and rinsed with distilled water. Cells were stained with 0.1% crystal violet (Sigma) for 30 min, rinsed extensively, and dried. Cell-associated dye was extracted with 2.0 ml 10% acetic acid. Aliquots were diluted 1:4 with H_2O, transferred to 96-well microtiter plates, and the optical density at 590 nm was determined. Values were normalized to the optical density at day 0 for the appropriate condition. Within an experiment, each point was determined in triplicate.

3.5.6 Western blot

PC3 cells were washed with cold PBS and then scraped on ice into 500 μl of PBS buffer containing 1X Complete-EDTA free Protease Inhibitor Cocktail (Roche Applied Science) and 1X PhosSTOP Phosphatase Inhibitor Cocktail (Roche Applied Science). Cells were spun at maximum speed for 5 min. The pellet was re-suspended

in 100 µl of Laemmli-β-Mercaptoethanol buffer, sonicated 5 seconds at a low intensity, heated 5 min at 95ºC and then cleared by centrifugation at 13 000 RPM for 10 min. The proteins were quantified with the Bradford reagent and 30 µg were loaded on a 10% SDS–PAGE and transferred to Immobilon-P PVDF membranes (Millipore). Membranes were blocked 1 hour at room temperature in PBS containing 0.1% Tween 20 (PBS-T) and 5% dry milk and then washed for 5 min 3 times with PBS-T. The membranes were incubated with the primary antibodies diluted in PBS-T + 3% BSA + 0.05% Na-azide overnight at 4ºC. The following primary antibodies were used: anti E2F1 (1:1000, clone H-137; rabbit polyclonal; #SC22820, Santa Cruz); anti E2F2 (1:1000, clone L-20; rabbit polyclonal; #SC632, Santa Cruz); anti E2F3 (1:1000; clone PG-37, mouse monoclonal, #5551, Millipore); anti-α-tubulin (1:20000, mouse monoclonal clone B-5-1-2, T6074, Sigma-Aldrich). Membranes were washed three times 5 min with PBS-T and then incubated with the secondary antibodies diluted in PBS-T + 5% dry milk 1 hour at room temperature. The following secondary antibodies were used: goat anti-rabbit IgG conjugated to HRP (1:3000, #170-6515, Bio-Rad) or goat anti-mouse IgG conjugated to HRP (1:3000, #170-6516, Bio-Rad). Finally, the membranes were washed three times 5 min with PBS-T. Immunoblots were visualized using enhanced chemiluminescence (ECL) detection systems and Super RX X-Ray films (Fujifilm) or a ChemiDocTM MP system (Bio-Rad). Band quantification was done using ImageJ or Image Lab 4.0 (Bio-Rad).

3.5.7 Molecular modeling of AGO protein structures

Protein structure files were downloaded from the Protein Data Bank (PDB) website. Modeling was performed in UCSF Chimera version 1.10.2 following the software documentation. Molecular surface calculation, sequence alignments, and distance measurements were performed according to established procedures (Burger et al., 2009). Briefly, the default settings were used for all the calculations: molecular surface computation and distance measurement was performed using the built-in function of Chimera. The coulomb potential surface was calculated using the built-in function under the “surface/binding analysis” option among the “Tools”.

Default parameters were used: dielectric constant = 4.0, distance from surface = 1.4, and histidine protonation was assumed for structures without explicit hydrogens. Hydrogen bond predictions were performed with the "FindHBond" function in "Structural Analysis" with default parameter settings to relax H-bond constraints by 0.4 angstroms and 20 degrees. The MatchMaker function of Chimera performed the structural alignment between the two protein structures. The function's default settings include using Needelman-Wunsch algorithm and BLOSUM-62 matrix for sequence alignment, where gap-opening penalties for intra-helix and intra-strand are both 18, and 6 for any others, and the program iterates by pruning long atom pairs until non pair exceeds 2.0 angstroms.

3.5.8 Implementation and validation of *MicroAlign* and the *miScore*

The evaluation program *MicroAlgin* was implemented in MicroSoft Visual Studio Express 2012 C++ as a stand-alone windows application. Experimental measurements were plotted against the predicted *miScore*s (see code below) to calculate Pearson correlations.

Fold inhibition of miR-21, miR-122 and miR-22 were taken from (Robertson et al., 2010). For each miRNA, the dataset chosen represented what the authors defined as the inhibitor concentration whose efficacy most accurately captured the effects of the dinucleotide mismatches. The concentrations were: 20 nM for miR-21, 2 nM for miR-122 and 0.3 nM for miR-22. As a pre-filtering step, mismatched inhibitors in the first position and seed region (nts 2-8) were excluded. The data were transformed into residual target proportions (1 / Fold inhibition), and because all 3 miRNAs do not share the same concentration, the residual target proportions had to be linearly scaled to give relative target expression levels. The linear scaling was performed by fixing the lowest residual target proportion to 0 and the positive control value, represented by the fully matched inhibitor, to 100.

The catalytic efficiency (k_{cat}/Km) measured for AGO2 was used as a proxy to infer relative target expression levels (Wee et al., 2012). The less efficient the catalysis, the higher is the expression of the siRNA target. As for the Fold inhibition

dataset, mismatched guide siRNAs in the first position and seed region were excluded from this dataset. For the sake of uniformity, the catalytic efficiency values were normalized to the most efficient siRNA guide to get k_{cat}/Km percentage values comparable to the other datasets. Relative target expression was defined as 100 minus the percentage catalytic efficiency of the siRNA guide.

The datasets generate by Baek et al. were downloaded from the Nature journal website. The 3'UTR sequences were obtained using methods established previously in the lab (Weill et al., 2015). A script was generated to run *MicroAlign* program on each 3'UTR sequence downloaded from the UCSC. The list of hits were stored in a text file and compared with the Baek data set for entries with the same gene symbols and RefSeq ID. The protein and mRNA expression levels were extracted for analysis.

The pseudocode of *MicroAlign* evaluation algorithm implements the DFA described in **Fig. 15D**. The set of transition functions (δ) of the DFA is described in the figure, where state set Q = [q_0, q_1, q_2, q_3, q_4], alphabet set Σ = [seed, A, B, C, D], and transition function set δ: Q X Σ -> Q. The start state is q_0 and the accepted state is q_4. The configuration of bps between the miRNA and the target now can be regarded as a regular expression that is recognized by this DFA, simulating the AGO2 mechanism.

The bps are predicted by Needleman-Wunch algorithm and evaluated by regions following the discovered order. We described this DFA using a recursive algorithm. The "bottom" of the recursion is the evaluation of region-D, where the contribution of bps is little for accessible 3'UTR sites. The algorithm is implemented as a Windows application and a copy of it is available online: http://major.iric.ca/MajorLabEn/MiR-Tools.html.

```
pairing_score = 3;
list_of_regions = ( B, C, A, D )
Evaluate_score( list_of_regions ) {
    current_region = car( list_of_regions )
    %pair = paired bases in current_region / number of bases in current_region
```

```
    if( current_region == D )
        score = 0
    else if( %pair > 0 )
        score = %paire * ( pairing_score + Evaluate_score( cdr( list_of_regions )))
    return score
}
```

3.6 Acknowledgements

We thank Jerry Pelletier, Abba Malina, John Mills, Regina Cencic, Francis Robert, David Cotenoir-White, Khalid Hilmi, and Justina Kulpa for discussions, cloning and cell culture materials, as well as assistance; Julie Pelloux, Jean Paquette, and Angelique Bellmare-Pelletier for discussions; Sven Diederichs and Eric Cohen for constructs; and, André Laperrière for assistance. Grants to François Major from the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery grant program), the Canadian Institutes of Health Research (CIHR) [MOP-93679], and the National Institutes of Health [R01GM088813] supported this work.

CHAPTER 4: GENERAL DISCUSSION

4.1 The multiple-target approach in designing anti-HIV shRNAs

In the search of efficient RNAi molecules that inhibit HIV replication and expression, prioritization of features to be taken into consideration became the focus of my study. The early version of *miRBooking*, which was developed to predict endogenous miRNA targets based on seed complementarity and intracellular concentration of RNA species, was adapted to design multi-targeting guide RNA molecules as a replacement of the *MultiTar* program (De Guire et al., 2010). Some efficient repressor RNA molecules were designed and provided cells with significant protection against invading viral particles; however, the lack of thorough understanding of the targeting principles had hindered a precise rationalization of the observed differences in efficiency among the designs.

As shown in Chapter 2, SM5 shRNA elicited stronger repression than SM12 on HIV while they share the same seed sequence in transiently transduced cells. This is corroborates with a previous report that the efficiency of targeting directly the viral genome correlates with the overall complementarity between the guide-target duplex (Houzet et al., 2012). We confirmed the reason why complementarity is important by treating the cells with puromycin, a chemical that causes the dissociation of ribosomes, and enhanced repression. We concluded that the translating ribosome on the coding region of the viral genome is the main reason why partially complementary RNA guides are much less effective. We confirmed this is not a phenomenon due to general shut-down of translation by treating the same cells with cyclohexamide and rapamycin, which failed to elicit enhancing effect on RSIC-mediated repression (data not shown). Cyclohexamide and rapamycin, though reduces protein synthesis in general, do not specifically reduce the number of ribosomes on mRNA. In the case of rapamycin, the repression levels were even reduced, indicating that some basal levels of protein synthesis is required for miRNA-mediated repression to occur.

As SM12 targets endogenous Rela gene more potently than SM5, it becomes hard to discern whether synergistic inhibition of HIV gene expression took place when SM12 was stably transduced in cells and provided similar level of protection as SM5. The discrepancy between repression efficiencies of the guide RNAs SM5, 12, and 20, which share a common seed sequence, was surprising; but so was the unanimous efficiency of them when they were stably transduced in a different cell type. For some seeds designed with "best fold-change" and "low disturbance" principles, guide RNAs with randomly generated non-seed sequence elicited even stronger repression than the perfect matching ones. By testing rationally designed mismatches in the non-seed region, we confirmed that such enhanced repressive effect could not come from direct inhibition of HIV RNA alone. It is possible that SM12 off-targets unintendedly other genes and enhances the repression. To confirm that, a more sophisticated computational tool is required to narrow down the off-targets of shRNAs. Moreover, the features that determine the efficiency of knockdown by partially matched guide RNA against the HIV genome are yet to be thoroughly examined.

4.2 Essential features of a guide RNA for effective silencing

In Chapter 2, we showed that both non-seed complementarity and target site location play deterministic roles for repression efficiency when we target the HIV genome. Existing algorithms are not able to satisfactorily address these factors with high accuracy. Aiming to identify the key factors that are most pertinent to the design of artificial miRNAs, I conducted experiments described in Chapter 3. Target site location, site repeats, seed complementarity, non-seed complementarity, and target and guide RNA concentrations were tested side by side. The results showed that the determinant intrinsic factor was base complementarity, and the extrinsic one, target site accessibility. This conclusion corroborates with those of Zamore and Segal (Vainberg Slutskin et al., 2018; Wee et al., 2012).

Nevertheless, this conclusion should not be interpreted as other factors can be safely ignored when predicting genomic targets of miRNAs. Designed guide RNA

molecules are meant to be delivered into cells and are expressed at higher levels comparing to those of endogenous miRNAs. As Sharp et al. pointed out that endogenous miRNA generates "threshold" (Mukherji et al., 2011), RNA species within specific concentration range will manifest competitive effects as ceRNAs (Ala et al., 2013). The significance of ceRNA concentrations has also been confirmed by *miRBooking* in miRNA target prediction (Weill et al., 2015).

Artificial miRNAs studied in this thesis could be less sensitive to concentration variations because their operating concentrations are generally far beyond the "thresholds". This was confirmed in the titration of miB levels using RT-qPCR in Chapter 3. The repressive power of the miB guide RNA was not significantly shifted when its concentration varied up to 8 fold. The same may not be true for endogenous miRNAs which operate at a wider range of concentrations. Competition between endogenous miRNAs as well as target species that are close in concentrations are likely to cause significant fold changes in the target gene expression when the concentration of one species changed.

To illustrate this point, a simulation of the mass-action law was conducted and the fractional increase was plotted against the initial concentration of a hypothetical molecular species (Fig. 23). Assuming that a molecular species accounts for more than 80% of the original population while the rest 20% are all its competitors, after 10X increase in its concentration, its fractional increase in the total population of competing RNA species does not exceed 20%. This is a close resemblance to what was observed in the titration of miB in Chapter 3. On the other hand, the fractional increase always peaked around initial concentrations of 30-50%. This means that for miRNA of which initial concentration is similar to the competing RNA species, somewhat significant alteration in total fraction can be achieved by altering its concentration. Our conclusion from the simulation as well as experimental results agrees well with that presented by the Stoffel and Bartel labs (Denzler et al., 2014). In their study, the derepression of targets of miR-122, which is the most abundant miRNA in hepatocytes (Landgraf et al., 2007), requires more than 1.5 X 10^5 added MREs per cell; moreover, the target abundance of miR-122,

miR-33, and miR-16 which are the most abundant miRNA species in liver, were shown to be not altered more than 25% in any liver disease model. As the consequence, they did not observe ceRNA effect for miR-122 *in vivo*. Their conclusion is that the changes in ceRNAs must begin to approach the target abundance of miRNA before their effects can be observed. They also pointed out though miRNA and their target abundance alone may not be sufficient predictors, the concentration effect is still significant considering the involvement of unknown non-coding RNA species that might contribute to the pool of binding sites within the transcriptome. Contrasting their results with miR-122, the Steitz group showed that miR-1a/miR-106 and miR-133 can be effectively titrated up to 50% using addition MREs in C2C12 cells (Pinzon et al., 2017). However, in most cases, the reduction of miRNA efficiency is between 10-25%, consistent with the "best cases" in our simulation results.

Furthermore, all concentration simulations are performed under the assumption that there is no significant difference in binding affinity between miRNAs with their targets in the presence of Argonaute. The Zamore lab has shown that miRISC's binding affinity is greatly affected by base pairing positions (Wee et al., 2012). The affinity between miRNA and their targets may further complicate the prediction yet must not be omitted due to its deterministic power.

The determinant features that need to be prioritized in genome-wide miRNA target prediction are very likely to differ from those in the design of artificial miRNAs. In order to predict endogenous miRNA targets more accurately, the intrinsic factors in the targeting process need to be understood thoroughly to allow its integration into a miRNA target prediction algorithm.

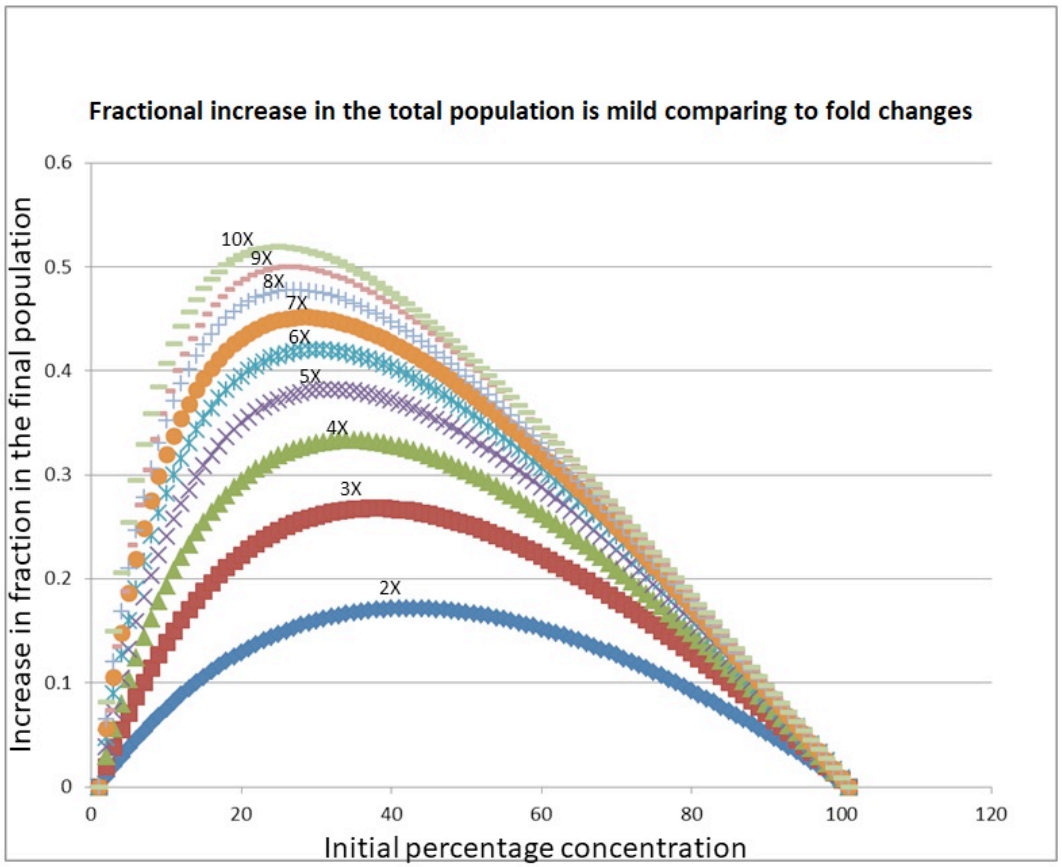


Figure 23. Simulation of Concentration Effect of miRNA.

We plotted the fractional increase of a particular molecular species among the total population against its initial percentage concentration at different numbers of fold changes. Each curve represents a specific number of fold change, the blue curve represents the doubling of concentration of species A, while other curves, from bottom to top, represent 2 to 10-fold of concentration changes.

Putting the three main factors listed above into broader perspective, I summarized the key features of the targeting process that determine the efficiency of miRNA/small RNA silencing:

A. Accessibility of the target site. This is mainly determined by three sub-categories of factors:
    a. RNA-RNA interactions, such as a local RNA secondary structure;
    b. RNA-protein interactions, such as RNA-binding proteins that are present in the 5'UTR and the polyA-tail;
    c. RNA-ribosome interactions, such as the translating ribosomes in the coding region.
B. Base pairing between the guide and the target:
    a. Positional-dependent effects of base pairs;
        i. Seed base pairing (the order is not clear);
        ii. Base pairing beyond the seed (in sequential and modular propagation);
        iii. Thermodynamic stability of key nucleotide positions.
    b. Position-independent effects of base pairs:
        i. Nature of the mismatched nucleotide (e.g. A/G mismatches are worse than C/U mismatches);
        ii. Central loop geometry: size and symmetry are important.
C. Concentrations of key components of the miRISC in competition. This feature is usually cell-type and cell-state dependent
    a. Mature miRNAs/guide RNAs;
    b. Target mRNAs;
    c. AGO protein.
D. Deadenylation, decapping, decay, and translational-repression factors (Reviewed in Section 1.4.2). The abundance of these proteins factors could explain the difference of slicer-independent repression efficiencies in different cell lines. As we proposed the model of target recognition, higher

concentration of these factors may lead to more rapid target decay or repression in the case of a partially complementary guide RNA.

For *ab initio* target prediction programs to achieve high accuracy, at least the base pairing features should be well simulated since it is the intrinsic factor that governs the repression mechanism.

Well-known miRNA target prediction programs, such as *TargetScan*, *miRanda*, *miRBase*, *PicTar*, and *PITA*, use site sequence conservation as an important factor in their prediction. Though it plays an important role in target prediction, the nature of this factor is more associative than mechanistic, hence not listed above as features for *ab initio* algorithms.

4.3 Analysis of the limitation of linear regression-based target prediction algorithms

The development of miRNA target prediction tools so far has revealed that a limitation exists in the implementation. Before the intrinsic factors were thoroughly understood and could be well considered in computation, the significance of extrinsic factors was identified. Because of their importance, extrinsic factors were often heavily weighted as predictors in the existing target prediction algorithms. When experimental data have grown over the years, it has become harder to manually optimize the way to combine these features. Substantial efforts were spent to automate the calibration process by combining evidence from experimental data to train prediction models (Agarwal et al., 2015; Davis et al., 2017). These automated approaches, which are collectively referred as "machine learning" methods usually do not address the underlying mechanistic nature of the targeting process.

It is a general issue in current computational approaches in biological research, which heavily relies on automation to handle large amount of data. To illustrate this challenge, it is necessary to clarify the two major goals of computation in biological research: inference and prediction. Inference builds a mathematical

model that generates data according to the formalized understanding of the underlying biological process. Predictions intend to forecast unobserved outcomes of future behaviour of a system without any requirement of knowledge of its underlying principles (Bzdok et al., 2017). Nearly all currently available miRNA target prediction programs are perfect examples of such algorithms. Though machine learning approaches can be used for both inference and prediction, due to its minimal assumption about the underlying mechanism as well as easiness of use (only general-purpose learning algorithms are needed to find patterns in large amount of data), it is predominantly used for target prediction in the miRNA research. Machine learning can be effective in cases where data are gathered without carefully controlled experimental design and when non-linear interactions exist in the system in question (Bzdok, 2018). However, despite the seemingly accurate perdition results, its characteristic "black-box" operating principle, which leads to the absence of an explicit model, makes its solution marginally relevant to existing biological knowledge and thus insufficient to help understand the mechanisms of the systems in medicine and biology (Ma et al., 2018).

At the surface, machine learning approaches seem to fit well with the concept of Turing Test; however, theory of the Turing Test only states the criteria for a machine to "pass". It did not restrict either the complexity or the computational approach that it uses, whether *ab initio* or machine learning in nature, of the machine in question. Hence the current problem is rooted in a bias of the interpretation of the Turing Test for machine learning approaches in general. With their growing complexity and building cost, pure machine-learning-based approaches for miRNA prediction tools are reaching their performance plateau in target prediction. In contrast, genome-scale, high-throughput, and high-content experimental approaches are becoming gradually more cost-effective and represent feasible alternatives to investigate the targets of miRNAs. Although the fast increase of biological data may bring extra support for machine-learning approaches, the training and correction process will always continue as all possible combinations of the determining factors have not been fully understood.

4.4 Recent updates of representative target prediction programs

Realizing the limitations of the previous versions of *TargetScan*, Agarwal et al. updated it as TargetScan7 by including more factors (called "features") into calculation (Agarwal et al., 2015). Fourteen different features were included in total. Among all the features, base pairing in the 3'-supplementary region (nt 13-16) were included in the score calculation. Regarding how non-seed base-pairing was addressed, two key differences between his approach and ours are listed as follows:

First, apart from the seed pairing, Agarwal et al. only considered base pairs within nt 13-16 and assigns a score to it. In our model, we considered the base pairs in B, C, and A regions collectively following a hierarchical rule (nt 9-17). Secondly, the score that Agarwal et al. assigns to nt13-16 is multiplied by a coefficient, and then combined with 13 other features, which are multiplied by their own coefficients. In that way, a total score was obtained by numeric combination without addressing the underlying molecular mechanism, as long as the final results fit the experimental data. Our model addresses only the molecular mechanism of target recognition and our computational process follows the step-wise biological process.

Meanwhile, the Segal lab's improvement upon the original predictive algorithm that was based on ΔΔG of the RNA duplex also showed some promising results in correlation with experimental data (Vainberg Slutskin et al., 2018). As introduced in *PITA* (Kertesz et al., 2007), this approach computes the difference between the predicted energy expenditure to access the target RNA and the energy gain during the guide-target annealing process. The predicted values were then compared to experimentally collected data from rationally designed guide RNA libraries using a massive parallel reporter system. Their study has identified base complementarity, target site thermodynamics, and miRNA concentration as the major determinants of miRNA targeting efficiency. Their conclusion agrees well with ours, indicating a convergence in understanding of miRNA targeting as well as the validity of the experimental approaches that we used. However, the way that they

addressed nucleotide complementarity still largely based on linear regression. Each nucleotide was assigned with a weight that has to be calibrated with each input data set. As the result, some experimental data showed high correlations with predicted values with $R^2>0.8$, while some correlations may be low as $R^2=0.12$, confirming what we described as a pitfall for all regression-based methods. Improvements based on regression, as well as the class of numerical methods represented by it, suffer from one common pitfall. Due to its *ad hoc* nature, unpredictable errors may arise given a new set of data even if there is no extra interfering factor in the data collection process.

It would be difficult to rationalize the discrepancy between the predicted value and experimental data. Programmers usually resolve this issue by a new round of calibrating the parameters in the numerical model; yet that merely initiates another cycle of *ad hoc* optimization by automated data fitting.

Such root problems with miRNA target prediction programs are known, as computational biologists already pointed out the major consequence of such prediction programs: large amount of false positives were predicted (Pinzon et al., 2017; Seitz, 2017). Scepticism was raised regarding Bartel's claim that the majority of the genome is under the regulation of miRNAs, Steitz group argued that most of the targets are repressed at insignificant levels to be biologically functional. Moreover, the conservation of miRNA target sequences is not always due to their complementarity with miRNA; rather, they are conserved in species without miRNA genes and hence might represent a deeper root in evolution. Interestingly, the Steitz lab also performed the titration experiment using miR-122 complementary MREs in C2C12 cells. They found that several mRNAs can effectively titrate (more than 10%, up to 50%) miR-1a/miR-206 and miR133. They argued that titration could be a potential mechanism of regulation and many MREs recognized by current prediction programs may be falsely identified as direct targets. Using data from inconsistent and biased experimental approaches, fitted with sophisticated mathematical models, the accuracy of current miRNA target prediction programs is significantly compromised.

4.5 Known limitations of the non-seed base pairing model we proposed.

In Chapter 3, experiments were designed to test the contributions of non-seed base pairing. From the experimental observations a mechanistic model was inferred to unify most of the published observations about miRNA targeting to date. However, several details of this proposition still need to be addressed.

It has been shown that target release was the rate-limiting step for AGO2 slicer activity (Deerberg et al., 2013; Willkomm and Restle, 2015) and the mismatching the 3 nucleotides at the 3' end of the guide RNA enhances slicer function. Our data also corroborate the enhancing effect. However, we only observed enhancement under the premise that the A-, B-, and C-modules are all base paired. This possibly indicates that mismatches in these modules significantly slow down the pre-cleavage complex formation so that base pairing becomes the new rate-limiting step. A general conclusion from this observation is that given a multi-step process such as RISC target recognition, the rate-determining step may be altered when mismatches occur in the RNA duplex comparing to the perfectly matching guide. Moreover, the extent to which these mismatches compromise silencing may change when slicer-independent protein factors participate in the process. These factors, exemplified the GW182 family proteins, may have a better chance to encounter a non-cleaving miRISC that dwells on the mRNA for a longer duration of time. As the consequence, slicer-independent pathway might be more readily activated when mismatches are present. Further combinatorial studies need to be performed to calibrate the effects of nucleotide mismatches on the two different pathways.

In addition, due to the experimental design, we could only addressed non-seed base mismatches that occur in a stretches of at least three consecutive nucleotides. Hence our "modular" conclusion should not be interpreted as the nature of the target recognition process; rather, it was due to the design of our experimental approach which aimed to reduce search space. Individual nucleotides were not

distinguished within each module and no overlapping modules were tested. This design had led to some gaps in the coverage of our investigation, namely, unequal contribution of base pairing at each position. As miB-mod2 and miB-13D+18 were shown to be outliers (**Fig. 17CD**), we noticed that they both contain a mismatch at position 13. Hence, the cursory assumptions of equal contribution of base pairs within a module, as implemented in *MicroAlign*, require further calibration. In addition to the positional effect, as mentioned in the *Introduction*, the nature of the mismatched nucleotides was also known to affect the silencing efficiency (Section 1.5.1.4). In particular, mismatches involving the purine bases abolish the silencing efficiency to a greater extent than those involving pyrimidines only. As we saw in the validation of *MicroAlign* algorithm (**Fig. 17G**) using third-party data, discrepancies occurred when the same key position was mismatched with different bases. In the next round of calibration of the algorithm, both nucleotide positions and the nature of the bases need to be calibrated.

4.6 Validation issues of *MicroAlign*

One of the major concerns with the *MicroAlign* algorithm is the low abundance of validation data. Comparing to existing miRNA target prediction software, which were trained and validated using data collected at genomic scale using high throughput methods, our validation using two external data sets and randomly generated mismatches based on miB guide strand seemed insufficient. In general, abundant validation data is preferred; however, it does not mean the validation for *MicroAlign* was insufficient.

The model derived from our experimental data is an inference one, not a predictive one. It describes the order by which Ago2 acknowledges the guide-target base pairs beyond the seed. The goal of this implementation is to make sure that we can apply the mechanistic rules consistently without human error when cross-checking with third party data, as such, we could demonstrate the truthfulness of the proposed mechanism. This determined the fact that the number of data sets that are suitable for a fair validation of *MicroAlign* is scarce despite the seemingly abundant

data collected from siRNA cleavage studies. To validate this mechanism using third party data, we need to be sure the data collection process was not tampered with factors uncontrolled for; otherwise, we risk rejecting a valid mechanism based on interference. In summary, the data that we can use must satisfy the following criteria:

A. The target site is located in an untranslated RNA that is relatively free of secondary structures (no interfering RNA structures, ribosomes, or binding proteins).

B. In each series of nucleotide mutations to be compared, the seed must always be perfectly paired with the target (our model does not emulate mismatches in the seed).

C. The mismatches must be in the form of symmetric or near-symmetric loops in the duplex. We did not test loops with large asymmetric bulges, though they are an interesting case for experiments and computations.

D. Di- or tri-nucleotide mismatches must be present in the duplex. Single nucleotide mismatches can produce kinks in the duplex and may alter the duplex geometry in the Ago2 binding site.

E. A positive (perfect match) and a negative control must always be present for the same target site in question.

When we looked into past literatures for data sets that satisfy all of the above criteria, we could only find those used in our paper, which are based on let-7 (Wee et al., 2012), miR-21, miR-22, and miR-122 siRNAs (Robertson et al., 2010). Combined with miB, we have shown that this mechanism is consistent with these five. Using four mutated target sites of miB, additional 16 combinations of guide-target duplex were produced and the same pattern was emerged. When we observed that a total of 10 different guide RNAs conform to the same proposed mechanism from multiple biochemical assays (15 repeats, with 93 total interactions including 15 at the protein levels, in our case), we could not reject the mechanistic model based on the lack of validation data of its predictive power. Comparing to some of machine learning prediction programs, for example, *TargetBoost* was trained upon 300

randomly generated negative sequences and 36 positive ones; *miTarget* was trained upon 246 negative and 152 positive interactions; the *Ensemble* algorithm, which uses 10 SVMs, was trained on 16 negative and 48 positive interactions from experiments (Reyes-Herrera and Ficarra, 2012).

Due to the lack of ideal third party data, we used the microarray and SILAC data from Baek et al. (Baek et al., 2008). The original publication benchmarked several available prediction programs that were available at the time of publication. Though *MicroAlign* is not a target prediction program, applying it to this data set lends an opportunity to verify how significant the role of the sequential base-pairing beyond the seed is for genomic targets. The Baek paper contains the evaluation of performance of algorithms that identify 7-8mer seed sites, which is a suitable control in our case since *MicroAlign* was only calibrated to evaluate sites of this type. Another reason for choosing this data set is that it has separated the evaluation of programs that consider conservation and those that do not. With genomic data, it became harder to observe clear correlations between the miScore and target gene expression levels, which is probably why the original paper did not present it either.

As described in the paper of Baek et al., we used the three-bin approach, which divides the ranked sites into top, middle, and bottom one-third. We observed the enrichment using *MicroAlign*. Its average performance is the same the 7-8mer sites, as expected; however, among the top ranked sites, verified target genes were enriched at better or the same level as programs that use both base pairing and conservation rules in computation, except *TargetScan*. *MicroAlign* outperforms all other programs that mainly based on thermodynamic scores and do not use conservation in computation. Since *MicroAlign* does not rely on such differential assignment and it enriches target sites purely based on the sequential rule of base pairing beyond the seed, it is quite possible that the modular and sequential base pairing beyond seed could be the main contributing factor to the site conservation pattern of miRNA targets.

4.7 Application of the *MicroAlign* algorithm

The latest version of *MicroAlign* is available through *ResearchGate* in addition to the lab website. This algorithm, though not suitable for the prediction of miRNA targets at the genomic level, is effective at filtering out low efficiency target sites, as shown by the absence of data points in the lower left corner of its validation plot using third party data (**Fig. 17G**). Using this property, two application programs can be developed.

The first development would be a program that designs multi-targeting guide RNA strands based on the *MicroAlign* algorithm. A list of target gene sequences will be used as input and a list of guide RNA sequences will output to the user. The list is enriched with guide RNAs that are capable of knocking down the target genes (**Fig. 19**). Since the goal of such program is to enrich effective RNA guides, its ability to filter out ineffective guides becomes suitable for this purpose.

Another utility for this algorithm is to identify effective targets of a given RNAi guide in the genome. The user would provide the guide RNA sequence of interest to the program and the program searches all the 3'UTR sequences for effective target sites using the *MicroAlign* algorithm. A list of target genes, their accession numbers, target site positions, predicted *miScores* (for knockdown efficiency), as well as the most-likely alignment between of the guide-target RNA duplex are produced the output. The user can further filter the output according to other bioinformatics or biological criteria to identify their targets of interest. During the validation process of *MicroAlign* in Chapter 3, we have implemented a prototype of this program in the form of an improvised pipeline to demonstrate the importance of the sequence of base pairing in the regions beyond the seed. An example of the output of this prototype program for miR-20 targets is included in Appendix A (**Fig. 24**). Though not a program to predict the entire set of targets for a given miRNA, it can be used to visualize the most effectively targeted genes as well as effectively filtering out the genes that are unlikely to be repressed in the genome.

4.8 Evolutionary perspectives

The most significant proposition that this thesis describes, perhaps, is that double-stranded RNA helix must form on both flanks of the cleavage site of Aog2 protein before slicing reaction takes place. This model proposes that recognition via base pairing occurs on both flanks of the scissile nucleotide in a stepwise fashion. It is not a coincidence. The *par* RNAI and RNAII in *Enterococcus faecalis* (Greenfield et al., 2001), clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated proteins (Cas) system in Bacteria and Archaea (Semenova et al., 2011), RNA-editing enzymes in the Euglenozoa phylum (Kiethega et al., 2013; Stuart et al., 1997), and hammerhead ribozymes in all domains of life (Perreault et al., 2011) follow similar step-wise binding pathways during their action, suggesting an evolutionary convergence in the mechanism of RNA-guided "natural genome editing" (Witzany, 2011). Our finding suggests that the human Ago2 protein also belongs to this class of genome editing machines.

4.9 RNAi in comparison with other genome editing methods

As a gene knockdown method, the RNAi technology has been the preferred method of choice due to its convenience and efficiency. Knockout technologies such as Cre/loxP recombination and TALEN (transcription activator-like effector nuclease) are highly effective and specific; however, clones are costly to generate and difficult to isolate. They are more frequently used for engineering cell lines and animal models for lab use. With the advent of CRISPR/Cas9 technology, which is based on guide RNA's recognition of genomic DNA followed by endonuclease cleavage of the target DNA, engineering genetic knockouts has become much more feasible and efficient. Comparing to RNAi-mediated knockdown, the main differences are that CRISPR/Cas9 approach is irreversible as it edits the genome, and that clone selection process is required due to the uncertainty in the repair process of the double stranded breaks produced by Cas9. The specificity of CRISPR/Cas9 was believed to be higher since a guide RNA length of 20nt with a PAM (protospacer adjacent motif) of NGG nucleotides is required. The "seed" of the

guide RNA is reported to be around 11nt for the Cas9 from *Streptococcus pyogenes*. Non-specific targeting in the genome were reported, indicating a similar off-targeting issue of shRNAs. However, with systematic study of the structure and function of the Cas9 proteins, later versions of the CRISPR/Cas9 technology showed significantly improved specificity for the target. Moreover, by carefully designing the assay, the intended clones can be specifically enriched with selective markers, hence repurposing this technology for high throughput studies (Malina et al., 2013).

The RNAi approach has been questioned for its lack of predictability in off-targets. This issue can be attributed to the lack of understanding of the structure-function relationship of the Ago2 protein, in particular, its base-pairing rules. Taking lessons from the development of the CRISPR/Cas9 technology, similar studies about the RISC proteins could be conducted to improve the targeting specificity and efficiency of the RNAi technology. Comparing to CRISPR/Cas9, RNAi approach still has its irreplaceable advantages in its ease of use (does not normally require clone selection) and reversibility. For potential therapeutic uses, RNAi technology may represent milder but safer method than the genomic DNA modifying approaches.

Science emphasizes evidence and logical deduction,

and is forever uncertain.

—Douglas J. Futuyama*

CONCLUSION

By designing shRNAs that target both the HIV genome and endogenous factors that help the virus, viral gene expression can be effectively inhibited and cells become more resistant to viral infection. Though some shRNA only partially base pair with the HIV genome sequence, they provided stronger protection against invading viral particles than those that perfectly complement the viral genome. This confirms the known importance of the seed complementarity rule, as well as the effectiveness of the strategy of targeting both the viral and endogenous RNAs simultaneously. Further verification and refinement in design will be carried out for this approach.

In order to improve the artificial miRNA approach for designing efficient inhibitors, we further investigated the base-pairing rules for miRNA target recognition. Combining the seed pairing and concentration rules was not sufficient to achieve the required precision. By rationally designing shRNAs that partially match the *tat* gene of HIV, we deduced the molecular mechanism by which the Ago2 protein forms the pre-cleavage complex with the guide and the target RNA. It shows that once the seed is base-paired with the target RNA, the formation of RNA duplex is interrupted at the central portion of the guide strand and skips to nt 12-13, where it resumes base-pairing and continues to nt 17. The central portion is base

*Futuyama, J. Douglas, *Science on Trial: The Case for Evolution*, *Chapter 1: Reason under Fire*, Sunderland, Mass.: Sinauer Associates, Inc., 1982.

paired last. From this distinct order of base pairing in the non-seed region, a parsimonious model that identifies with most of the published miRNA studies was proposed. Implementing the rules of the model as a computer program improved the prediction accuracy and enhanced the design of artificial miRNAs.

In agreement with current view (Bartel, 2018; Broughton et al., 2016; Vainberg Slutskin et al., 2018), base pairing and accessibility of the target site are the two key factors for efficient targeting process. To this growing body of work, we resolved a subtlety in the order of base pairing beyond the seed and demonstrated its effects on target gene silencing. Other factors were carefully controlled for, but not thoroughly investigated in our study. Though this rule alone is not sufficient to predict genomic targets of miRNAs with high confidence, it effectively enhances the identification of preferred targets of ectopically expressed shRNA/miRNA. Besides enabling *de novo* design of specific and efficient RNA silencing guides, incorporation of this rule into miRNA target prediction software will allow significant improvement in the decoding of miRNA targets at the genome level.

During this study, we have re-examined the complementarity requirement of miRNA target recognition. As the result, we discovered a sequential rule that has been unfortunately overlooked by previous studies. The reason why it was not revealed earlier as it should have been, as I concluded from published work in this field, is that the rise of big data techniques at the turn of the millennium has endowed researchers the ability to mine patterns without properly addressing the underline biases in data. For instance, Ago-CLIP data contain substantial noises due to the irreversible nature of the cross-linking reaction. Such reactions coerce RISC-target interaction far beyond equilibrium. Noises are not effectively filtered; rather, they were data-mined with sophisticated mathematical tools and treated as signals. Due to the overwhelming power of pattern identification and data-mining techniques, they gained tremendous popularity in recent publications. Consequently, more and more target sequence patterns have been discovered using similar Big Data generating techniques.

Such trend became apparent when the seed rule alone turned out to be insufficient to accurately predict miRNA targets. Many research groups resorted to the "Big Data" techniques in hope to find additional patterns to improve the predictive power. Novel patterns/motifs discovered from "big data" approaches were quickly incorporated into miRNA target prediction programs. Yet such patterns have limited predictive power because they lack causal relationship with the underline mechanism; instead, they are, if not astrological, purely statistical, or at best, correlational in nature.

Moreover, such massively generated data have diverted the research effort from intrinsic factor to extrinsic ones before the intrinsic properties were thoroughly examined. Despite multiple novel classes of target sites were identified, limited evidence is provided regarding the mechanism of targeting for each class. Taking these new patterns into consideration, computational predictions present significantly high proportion of false positives in the output. This phenomenon only confirms the popular expression in computational modeling that "all models are wrong, but some are useful". Recent identification of 3' supplementary seeds by Zhang et al. is not an exception to this process. As expected, the authors could not offer any mechanistic explanation on how such a controversial process could take place, violating the known site thermodynamics, enzyme kinetics, as well as profiling results. Unless such novel sites are mechanistically supported, rather than merely associated, intensive backtrack and verification would be required to validate their biological significance.

We have seen several similar back-and-forth incidents in miRNA research. The seed rule was claimed to be the only one that is important to determine target efficiency by the Bartel group, who showed that only less than 5% of the sites detected in CLIP data contain 3' supplementary base pairing. However, as shown by multiple studies, the 3' supplementary region makes an undeniable contribution to targeting efficiency when it was purposely and systematically investigated. Interestingly, the Bartel group did not reveal the precise scoring algorithms of the latest *TargetScan*, which clearly assigns weights to non-seed nucleotides under the

name of “context score” and “site conservation”. Another example is the claim that centrally mismatched RNA loop impairs target cleavage but does not impair repression. At first glance, this claim agreed very well with the “seed only” rule by dismissing the importance of base pairing immediately downstream of the seed. However, looking closely at the carefully engineered central mismatches, one could see that they often contain two nucleotides on one stand and three on the other (asymmetric loop). In contrast, the Crooke lab has demonstrated that when the loop contains three nucleotides on both strands (symmetric loop), the target mRNA is indeed cleaved by Ago2. This led to the next claim, which states that centrally mismatched target sites, regardless the symmetry, lead to deadenylation, decapping, and decay of mRNA. However, following those lines of evidence, the final executioners of RNA decay are rather general factors that are involved in the break down RNA (the CCR4-NOT complex and its downstream effectors), which is a long shot to tie them with specific and efficient RNAi effects. When logical conclusions point to translational repression as the main down-regulation mechanism, the factors identified greatly overlap with those involved in stress responses and it takes longer to manifest their effects. With non-specialized effector molecules that overlap other pathways, neither RNA decay nor translational repression could explain efficient miRNA-mediated down regulation claimed. Decay of mRNA is slower, and translational repression is weaker than one would expect from miRNA. To resolve the controversy between translational repression versus mRNA decay in the “slicer-independent” pathway, translational repression was shown to be a weak but early effect, while RNA decay eventually becomes the predominant mechanism of repression (Bethune et al., 2012; Djuranovic et al., 2012; Eichhorn et al., 2014). Bartel commented on such timely findings as “welcome news” (Bartel, 2018).

However, popularity has little with validity. By looking at the back-and-forth manner and uncertainties in the verification process of numerous claims about miRNA action, one could only conclude with confidence that, without inclining toward any particular school of theory, overall base complementarity alters the efficiency of miRNA-mediated repression and miRNA acts predominantly via

causing the degradation of target RNA. Such conclusion is not a popular one as it places fundamental doubt about the original claim that miRNAs can regulate more than half of the coding genes via promiscuous seed base pairing. The Pandolfi group tried to offer an explanation for miRNA's broad-spectrum effect. They extended the promiscuity idea further by demonstrating its inevitable consequence: competition between targets for miRNAs, and vice versa. However, that possibility has been diminished by the Bartel group using *in vitro* and *in vivo* data showing that concentrations of the competing species are not physiologically feasible to allow such observations, dismissing the ubiquitous potential of cross-talks between miRNAs and/or targets. Bartel's claim puts the original miRNA theory in a dilemma, if not a mystery: miRNAs must effectively regulate most of the genome through direct base pairing with their targets, rather than indirect cross-talks; however, none of the proposed mechanism can exclusively tie their specificity and efficiency to the 6-7 nt of base pairing in the seed.

As more and more labs blindly adopt high throughput and Big Data approaches to increase their chance of discovering new miRNA target site rules, many curious but unexplained claims might arise. The challenge of re-examining these claims is expected to increase in the near future. Einstein once said, "Science should be as simple as possible but not simpler." It is simple when collected information follows consistent logic; it is not when new pieces of information generated are fragmented, biased, and logically inconsistent. As "novel" claims are flooding current research in the Big Data era, remaining vigilant and sceptical and resorting to logic could be the way to help us stay clear of the tendency to trade validity for novelty. The work presented in this thesis is merely one logical step taken from existing studies of the non-seed nucleotides, aiming to simplify and unify the seemingly inconsistent opinions about their contribution. There are still many steps to follow to unveil the truth about miRNA actions.

REFERENCES

Agarwal, V., Bell, G.W., Nam, J.W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife *4*.

Ahluwalia, J.K., Khan, S.Z., Soni, K., Rawat, P., Gupta, A., Hariharan, M., Scaria, V., Lalwani, M., Pillai, B., Mitra, D.*, et al.* (2008). Human cellular microRNA hsa-miR-29a interferes with viral nef protein expression and HIV-1 replication. Retrovirology *5*, 117.

Ala, U., Karreth, F.A., Bosia, C., Pagnani, A., Taulli, R., Leopold, V., Tay, Y., Provero, P., Zecchina, R., and Pandolfi, P.P. (2013). Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. Proc Natl Acad Sci U S A *110*, 7154-7159.

Albert L. Lehninger, D.L.N., Michael M. Cox (1993). Principles of Biochemistry, Second Edition edn (New York: Worth Publishers).

Aleman, L.M., Doench, J., and Sharp, P.A. (2007). Comparison of siRNA-induced off-target RNA and protein effects. Rna *13*, 385-395.

Allen, E., Xie, Z., Gustafson, A.M., and Carrington, J.C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell *121*, 207-221.

Amarzguioui, M., Holen, T., Babaie, E., and Prydz, H. (2003). Tolerance for mutations and chemical modifications in a siRNA. Nucleic Acids Res *31*, 589-595.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M.*, et al.* (2003). A uniform system for microRNA annotation. RNA *9*, 277-279.

Anderson, J.S., Javien, J., Nolta, J.A., and Bauer, G. (2009). Preintegration HIV-1 inhibition by a combination lentiviral vector containing a chimeric TRIM5 alpha protein, a CCR5 shRNA, and a TAR decoy. Molecular therapy : the journal of the American Society of Gene Therapy *17*, 2103-2114.

Avesson, L., Reimegard, J., Wagner, E.G., and Soderbom, F. (2012). MicroRNAs in Amoebozoa: deep sequencing of the small RNA population in the social

amoeba Dictyostelium discoideum reveals developmentally regulated microRNAs. RNA *18*, 1771-1782.

Aza-Blanc, P., Cooper, C.L., Wagner, K., Batalov, S., Deveraux, Q.L., and Cooke, M.P. (2003). Identification of modulators of TRAIL-induced apoptosis via RNAi-based phenotypic screening. Molecular cell *12*, 627-637.

Babiarz, J.E., Ruby, J.G., Wang, Y., Bartel, D.P., and Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes & development *22*, 2773-2785.

Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71.

Baeuerle, P.A. (1991). The inducible transcription activator NF-kappa B: regulation by distinct protein subunits. Biochim Biophys Acta *1072*, 63-80.

Bandyopadhyay, S., and Mitra, R. (2009). TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. Bioinformatics *25*, 2625-2631.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20-51.

Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes & development *20*, 1885-1898.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363-366.

Bernstein, P., and Ross, J. (1989). Poly(A), poly(A) binding protein and the regulation of mRNA stability. Trends Biochem Sci *14*, 373-377.

Betancur, J.G., and Tomari, Y. (2012). Dicer is dispensable for asymmetric RISC loading in mammals. RNA *18*, 24-30.

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol *11*, R90.

Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. (2008). The microRNA.org resource: targets and expression. Nucleic Acids Res *36*, D149-153.

Bethune, J., Artus-Revel, C.G., and Filipowicz, W. (2012). Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells. EMBO Rep *13*, 716-723.

Bhattacharyya, S.N., Habermacher, R., Martine, U., Closs, E.I., and Filipowicz, W. (2006). Relief of microRNA-mediated translational repression in human cells subjected to stress. Cell *125*, 1111-1124.

Blobel, G., and Sabatini, D. (1971). Dissociation of mammalian polyribosomes into subunits by puromycin. Proc Natl Acad Sci U S A *68*, 390-394.

Boden, D., Pusch, O., Lee, F., Tucker, L., and Ramratnam, B. (2003). Human immunodeficiency virus type 1 escape from RNA interference. Journal of virology *77*, 11531-11535.

Boden, D., Pusch, O., Silbermann, R., Lee, F., Tucker, L., and Ramratnam, B. (2004). Enhanced gene silencing of HIV-1 specific siRNA using microRNA designed hairpins. Nucleic Acids Res *32*, 1154-1158.

Bohmert, K., Camus, I., Bellini, C., Bouchez, D., Caboche, M., and Benning, C. (1998). AGO1 defines a novel locus of Arabidopsis controlling leaf development. EMBO J *17*, 170-180.

Bohnsack, M.T., Czaplinski, K., and Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA *10*, 185-191.

Boland, A., Tritschler, F., Heimstadt, S., Izaurralde, E., and Weichenrieder, O. (2010). Crystal structure and ligand binding of the MID domain of a eukaryotic Argonaute protein. EMBO Rep *11*, 522-527.

Borman, A.M., Michel, Y.M., and Kean, K.M. (2000). Biochemical characterisation of cap-poly(A) synergy in rabbit reticulocyte lysates: the eIF4G-PABP interaction increases the functional affinity of eIF4E for the capped mRNA 5'-end. Nucleic Acids Res *28*, 4068-4075.

Bosson, A.D., Zamudio, J.R., and Sharp, P.A. (2014). Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. Molecular cell *56*, 347-359.

Boyer, T.G., Martin, M.E., Lees, E., Ricciardi, R.P., and Berk, A.J. (1999). Mammalian Srb/Mediator complex is targeted by adenovirus E1A protein. Nature *399*, 276-279.

Bracken, C.P. (2008). A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition. Cancer Res *68*, 7846-7854.

Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J., Lieberman, J., and Elledge, S.J. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. Science *319*, 921-926.

Brennan, C.M., and Steitz, J.A. (2001). HuR and mRNA stability. Cell Mol Life Sci *58*, 266-277.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA-target recognition. PLoS biology *3*, e85.

Bridge, K.S., Shah, K.M., Li, Y., Foxler, D.E., Wong, S.C.K., Miller, D.C., Davidson, K.M., Foster, J.G., Rose, R., Hodgkinson, M.R*., et al.* (2017). Argonaute Utilization for miRNA Silencing Is Determined by Phosphorylation-Dependent Recruitment of LIM-Domain-Containing Proteins. Cell Rep *20*, 173-187.

Broderick, J.A., Salomon, W.E., Ryder, S.P., Aronin, N., and Zamore, P.D. (2011). Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. RNA *17*, 1858-1869.

Broughton, J.P., Lovci, M.T., Huang, J.L., Yeo, G.W., and Pasquinelli, A.E. (2016). Pairing beyond the Seed Supports MicroRNA Targeting Specificity. Molecular cell *64*, 320-333.

Brown, C.E., Tarun, S.Z., Jr., Boeck, R., and Sachs, A.B. (1996). PAN3 encodes a subunit of the Pab1p-dependent poly(A) nuclease in Saccharomyces cerevisiae. Mol Cell Biol *16*, 5744-5753.

Bueno, M.J. (2008). Genetic and epigenetic silencing of microRNA-203 enhances ABL1 and BCR-ABL1 oncogene expression. Cancer Cell *13*, 496-506.

Buhler, M., Verdel, A., and Moazed, D. (2006). Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing. Cell *125*, 873-886.

Burger, G., Yan, Y., Javadi, P., and Lang, B.F. (2009). Group I-intron trans-splicing and mRNA editing in the mitochondria of placozoan animals. Trends Genet *25*, 381-386.

Burgler, C., and Macdonald, P.M. (2005). Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. BMC Genomics *6*, 88.

Bushati, N., and Cohen, S.M. (2007). microRNA functions. Annu Rev Cell Dev Biol *23*, 175-205.

Bushman, F.D., Malani, N., Fernandes, J., D'Orso, I., Cagney, G., Diamond, T.L., Zhou, H., Hazuda, D.J., Espeseth, A.S., Konig, R.*, et al.* (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. PLoS pathogens *5*, e1000437.

Bzdok, D. (2018). Points of significance: Statistics versus machine learning. Nat Methods *15*, 233-234.

Bzdok, D., Krzywinski, M., and Altman, N. (2017). Points of Significance: Machine learning: a primer. Nat Methods *14*, 1119-1120.

Capodici, J., Kariko, K., and Weissman, D. (2002). Inhibition of HIV-1 infection by small interfering RNA-mediated RNA interference. J Immunol *169*, 5196-5201.

Carthew, R.W., and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell *136*, 642-655.

Cenik, E.S., and Zamore, P.D. (2011). Argonaute proteins. Curr Biol *21*, R446-449.

Cerutti, H., and Casas-Mollano, J.A. (2006). On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet *50*, 81-99.

Chandra, V., Girijadevi, R., Nair, A.S., Pillai, S.S., and Pillai, R.M. (2010). MTar: a computational microRNA target prediction architecture for human transcriptome. BMC Bioinformatics *11 Suppl 1*, S2.

Chandradoss, S.D., Schirle, N.T., Szczepaniak, M., MacRae, I.J., and Joo, C. (2015). A Dynamic Search Process Underlies MicroRNA Targeting. Cell *162*, 96-107.

Chang, T.C. (2008). Widespread microRNA repression by Myc contributes to tumorigenesis. Nature Genet *40*, 43-50.

Chapman, E.J., and Carrington, J.C. (2007). Specialization and evolution of endogenous small RNA pathways. Nat Rev Genet *8*, 884-896.

Chekulaeva, M., Filipowicz, W., and Parker, R. (2009). Multiple independent domains of dGW182 function in miRNA-mediated repression in Drosophila. RNA *15*, 794-803.

Chen, C.Y., and Shyu, A.B. (1995). AU-rich elements: characterization and importance in mRNA degradation. Trends Biochem Sci *20*, 465-470.

Chen, C.Y., and Shyu, A.B. (2011). Mechanisms of deadenylation-dependent decay. Wiley Interdiscip Rev RNA *2*, 167-183.

Chen, J., Chiang, Y.C., and Denis, C.L. (2002). CCR4, a 3'-5' poly(A) RNA and ssDNA exonuclease, is the catalytic component of the cytoplasmic deadenylase. EMBO J *21*, 1414-1426.

Chen, J.F. (2006). The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. Nature Genet *38*, 228-233.

Chen, K., Maaskola, J., Siegal, M.L., and Rajewsky, N. (2009). Reexamining microRNA site accessibility in Drosophila: a population genomics study. PLoS ONE *4*, e5681.

Chen, K., and Rajewsky, N. (2006). Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet *38*, 1452-1456.

Chendrimada, T.P., Gregory, R.I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., and Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature *436*, 740-744.

Chi, S.W., Hannon, G.J., and Darnell, R.B. (2012). An alternative mode of microRNA target recognition. Nat Struct Mol Biol *19*, 321-327.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479-486.

Cho, P.F., Poulin, F., Cho-Park, Y.A., Cho-Park, I.B., Chicoine, J.D., Lasko, P., and Sonenberg, N. (2005). A new paradigm for translational control: inhibition via 5'-3' mRNA tethering by Bicoid and the eIF4E cognate 4EHP. Cell *121*, 411-423.

Chung, W.J., Okamura, K., Martin, R., and Lai, E.C. (2008). Endogenous RNA interference provides a somatic defense against Drosophila transposons. Curr Biol *18*, 795-802.

Coburn, G.A., and Cullen, B.R. (2002). Potent and specific inhibition of human immunodeficiency virus type 1 replication by RNA interference. Journal of virology *76*, 9225-9231.

Colgan, D.F., and Manley, J.L. (1997). Mechanism and regulation of mRNA polyadenylation. Genes & development *11*, 2755-2766.

Coller, J., and Parker, R. (2004). Eukaryotic mRNA decapping. Annu Rev Biochem *73*, 861-890.

Connor, R.I., Chen, B.K., Choe, S., and Landau, N.R. (1995). Vpr is required for efficient replication of human immunodeficiency virus type-1 in mononuclear phagocytes. Virology *206*, 935-944.

Cullen, B.R., Cherry, S., and tenOever, B.R. (2013). Is RNA interference a physiologically relevant innate antiviral immune response in mammals? Cell Host Microbe *14*, 374-378.

Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. PLoS Genet *10*, e1004226.

Czech, B., Malone, C.D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J.A., Sachidanandam, R.*, et al.* (2008). An endogenous small interfering RNA pathway in Drosophila. Nature *453*, 798-802.

Das, A.T., Brummelkamp, T.R., Westerhout, E.M., Vink, M., Madiredjo, M., Bernards, R., and Berkhout, B. (2004). Human immunodeficiency virus type 1 escapes from RNA interference-mediated inhibition. Journal of virology *78*, 2601-2605.

Davis, B.N., Hilyard, A.C., Lagna, G., and Hata, A. (2008). SMAD proteins control DROSHA-mediated microRNA maturation. Nature *454*, 56-61.

Davis, J.A., Saunders, S.J., Mann, M., and Backofen, R. (2017). Combinatorial ensemble miRNA target prediction of co-regulation networks with non-prediction data. Nucleic Acids Res *45*, 8745-8757.

De Guire, V., Caron, M., Scott, N., Menard, C., Gaumont-Leclerc, M.F., Chartrand, P., Major, F., and Ferbeyre, G. (2010). Designing small multiple-target artificial RNAs. Nucleic Acids Res *38*, e140.

De, N., Young, L., Lau, P.W., Meisner, N.C., Morrissey, D.V., and MacRae, I.J. (2013). Highly complementary target RNAs promote release of guide RNAs from human Argonaute2. Molecular cell *50*, 344-355.

Deerberg, A., Willkomm, S., and Restle, T. (2013). Minimal mechanistic model of siRNA-dependent target RNA slicing by recombinant human Argonaute 2 protein. Proc Natl Acad Sci U S A *110*, 17850-17855.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231-235.

Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P., and Stoffel, M. (2014). Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. Molecular cell *54*, 766-776.

Desai, M., Iyer, G., and Dikshit, R.K. (2012). Antiretroviral drugs: critical issues and recent advances. Indian J Pharmacol *44*, 288-298.

Dickins, R.A., Hemann, M.T., Zilfou, J.T., Simpson, D.R., Ibarra, I., Hannon, G.J., and Lowe, S.W. (2005). Probing tumor phenotypes using stable and regulated synthetic microRNA precursors. Nat Genet *37*, 1289-1295.

Diederichs, S., and Haber, D.A. (2007). Dual role for argonautes in microRNA processing and posttranscriptional regulation of microRNA expression. Cell *131*, 1097-1108.

Diederichs, S., Jung, S., Rothenberg, S.M., Smolen, G.A., Mlody, B.G., and Haber, D.A. (2008). Coexpression of Argonaute-2 enhances RNA interference toward perfect match binding sites. Proc Natl Acad Sci U S A *105*, 9284-9289.

Ding, H., Schwarz, D.S., Keene, A., Affar el, B., Fenton, L., Xia, X., Shi, Y., Zamore, P.D., and Xu, Z. (2003). Selective silencing by RNAi of a dominant allele that causes amyotrophic lateral sclerosis. Aging Cell *2*, 209-217.

Ding, L., Spencer, A., Morita, K., and Han, M. (2005). The developmental timing regulator AIN-1 interacts with miRISCs and may target the argonaute protein ALG-1 to cytoplasmic P bodies in C. elegans. Molecular cell *19*, 437-447.

Ding, X.C., and Grosshans, H. (2009). Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins. Embo j *28*, 213-222.

Ding, Y., Chan, C.Y., and Lawrence, C.E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res *32*, W135-141.

Djupedal, I., Portoso, M., Spahr, H., Bonilla, C., Gustafsson, C.M., Allshire, R.C., and Ekwall, K. (2005). RNA Pol II subunit Rpb7 promotes centromeric transcription and RNAi-directed chromatin silencing. Genes & development *19*, 2301-2306.

Djuranovic, S., Nahvi, A., and Green, R. (2012). miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. Science *336*, 237-240.

Doench, J.G., Petersen, C.P., and Sharp, P.A. (2003). siRNAs can function as miRNAs. Genes & development *17*, 438-442.

Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. Genes & development *18*, 504-511.

Dow, L.E., Premsrirut, P.K., Zuber, J., Fellmann, C., McJunkin, K., Miething, C., Park, Y., Dickins, R.A., Hannon, G.J., and Lowe, S.W. (2012). A pipeline for the generation of shRNA transgenic mice. Nature protocols *7*, 374-393.

Du, Q., Thonberg, H., Wang, J., Wahlestedt, C., and Liang, Z. (2005). A systematic analysis of the silencing effects of an active siRNA at all single-nucleotide mismatched target sites. Nucleic Acids Res *33*, 1671-1677.

Dupressoir, A., Morel, A.P., Barbot, W., Loireau, M.P., Corbo, L., and Heidmann, T. (2001). Identification of four families of yCCR4- and Mg2+-dependent endonuclease-related proteins in higher eukaryotes, and characterization of orthologs of yCCR4 with a conserved leucine-rich repeat essential for hCAF1/hPOP2 binding. BMC Genomics *2*, 9.

Dye, M.J., Gromak, N., and Proudfoot, N.J. (2006). Exon tethering in transcription by RNA polymerase II. Molecular cell *21*, 849-859.

Dziuba, N., Ferguson, M.R., O'Brien, W.A., Sanchez, A., Prussia, A.J., McDonald, N.J., Friedrich, B.M., Li, G., Shaw, M.W., Sheng, J.*, et al.* (2012). Identification of cellular proteins required for replication of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses *28*, 1329-1339.

Eichhorn, S.W., Guo, H., McGeary, S.E., Rodriguez-Mias, R.A., Shin, C., Baek, D., Hsu, S.H., Ghoshal, K., Villen, J., and Bartel, D.P. (2014). mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. Molecular cell *56*, 104-115.

El-Shami, M., Pontier, D., Lahmy, S., Braun, L., Picart, C., Vega, D., Hakimi, M.A., Jacobsen, S.E., Cooke, R., and Lagrange, T. (2007). Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. Genes & development *21*, 2539-2544.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001a). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494-498.

Elbashir, S.M., Lendeckel, W., and Tuschl, T. (2001b). RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes & development *15*, 188-200.

Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. (2001c). Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. EMBO J *20*, 6877-6888.

Elkayam, E., Kuhn, C.D., Tocilj, A., Haase, A.D., Greene, E.M., Hannon, G.J., and Joshua-Tor, L. (2012). The structure of human argonaute-2 in complex with miR-20a. Cell *150*, 100-110.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003). MicroRNA targets in Drosophila. Genome Biol *5*, R1.

Espeseth, A.S., Fishel, R., Hazuda, D., Huang, Q., Xu, M., Yoder, K., and Zhou, H. (2011). siRNA screening of a targeted library of DNA repair factors in HIV infection reveals a role for base excision repair in HIV integration. PLoS One *6*, e17612.

Eulalio, A., Huntzinger, E., and Izaurralde, E. (2008). GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. Nat Struct Mol Biol *15*, 346-353.

Eulalio, A., Rehwinkel, J., Stricker, M., Huntzinger, E., Yang, S.F., Doerks, T., Dorner, S., Bork, P., Boutros, M., and Izaurralde, E. (2007). Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing. Genes & development *21*, 2558-2570.

Eulalio, A., Tritschler, F., and Izaurralde, E. (2009). The GW182 protein family in animal cells: new insights into domains required for miRNA-mediated gene silencing. RNA *15*, 1433-1442.

Fabian, M.R., Mathonnet, G., Sundermeier, T., Mathys, H., Zipprich, J.T., Svitkin, Y.V., Rivas, F., Jinek, M., Wohlschlegel, J., Doudna, J.A.*, et al.* (2009).

Mammalian miRNA RISC recruits CAF1 and PABP to affect PABP-dependent deadenylation. Molecular cell *35*, 868-880.

Fabian, M.R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. Annu Rev Biochem *79*, 351-379.

Faehnle, C.R., Elkayam, E., Haase, A.D., Hannon, G.J., and Joshua-Tor, L. (2013). The making of a slicer: activation of human Argonaute-1. Cell Rep *3*, 1901-1909.

Fahey, M.E., Bennett, M.J., Mahon, C., Jager, S., Pache, L., Kumar, D., Shapiro, A., Rao, K., Chanda, S.K., Craik, C.S.*, et al.* (2011). GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data. BMC Bioinformatics *12*, 298.

Fang, L., Stevens, J.L., Berk, A.J., and Spindler, K.R. (2004). Requirement of Sur2 for efficient replication of mouse adenovirus type 1. Journal of virology *78*, 12888-12900.

Fang, W., and Bartel, D.P. (2015). The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes. Molecular cell *60*, 131-145.

Fedorov, Y., Anderson, E.M., Birmingham, A., Reynolds, A., Karpilow, J., Robinson, K., Leake, D., Marshall, W.S., and Khvorova, A. (2006). Off-target effects by siRNA can induce toxic phenotype. RNA *12*, 1188-1196.

Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C.D., Dickins, R.A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J.*, et al.* (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. Molecular cell *41*, 733-746.

Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Rev Genet *9*, 102-114.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Flores, O., Kennedy, E.M., Skalsky, R.L., and Cullen, B.R. (2014). Differential RISC association of endogenous human microRNAs predicts their inhibitory potential. Nucleic Acids Res *42*, 4629-4639.

Forman, J.J., Legesse-Miller, A., and Coller, H.A. (2008). A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc Natl Acad Sci USA *105*, 14879-14884.

Forstemann, K., Horwich, M.D., Wee, L., Tomari, Y., and Zamore, P.D. (2007). Drosophila microRNAs are sorted into functionally distinct Argonaute complexes after production by Dicer-1. Cell *130*, 287-297.

Forstemann, K., Tomari, Y., Du, T., Vagin, V.V., Denli, A.M., Bratu, D.P., Klattenhoff, C., Theurkauf, W.E., and Zamore, P.D. (2005). Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. PLoS biology *3*, e236.

Frank, F., Hauver, J., Sonenberg, N., and Nagar, B. (2012). Arabidopsis Argonaute MID domains use their nucleotide specificity loop to sort small RNAs. EMBO J *31*, 3588-3595.

Frank, F., Sonenberg, N., and Nagar, B. (2010). Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. Nature *465*, 818-822.

Frankel, A.D., and Young, J.A. (1998). HIV-1: fifteen proteins and an RNA. Annu Rev Biochem *67*, 1-25.

Friedman, R.C., Farh, K.K., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res *19*, 92-105.

Friedman, Y., Naamati, G., and Linial, M. (2010). MiRror: a combinatorial analysis web tool for ensembles of microRNAs and their targets. Bioinformatics *26*, 1920-1921.

Gaidatzis, D., van Nimwegen, E., Hausser, J., and Zavolan, M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics *8*, 69.

Gale, D., and Shapley, L.S. (1962). College Admissions and the Stability of Marriage. The American Mathematical Monthly *69*, 9-15.

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., and Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol *18*, 1139-1146.

Gennarino, V.A., Sardiello, M., Avellino, R., Meola, N., Maselli, V., Anand, S., Cutillo, L., Ballabio, A., and Banfi, S. (2009). MicroRNA target prediction by expression analysis of host genes. Genome Res *19*, 481-490.

Ghildiyal, M., Seitz, H., Horwich, M.D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E.L., Zapp, M.L., Weng, Z*., et al.* (2008). Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science *320*, 1077-1081.

Ghildiyal, M., and Zamore, P.D. (2009). Small silencing RNAs: an expanding universe. Nat Rev Genet *10*, 94-108.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75-79.

Goff, S.P. (2007). Host factors exploited by retroviruses. Nat Rev Microbiol *5*, 253-263.

Gonzalez-Gonzalez, E., Lopez-Casas, P.P., and del Mazo, J. (2008). The expression patterns of genes involved in the RNAi pathways are tissue-dependent and differ in the germ and somatic cells of mouse testis. Biochim Biophys Acta *1779*, 306-311.

Greenfield, T.J., Franch, T., Gerdes, K., and Weaver, K.E. (2001). Antisense RNA regulation of the par post-segregational killing system: structural analysis and mechanism of binding of the antisense RNA, RNAII and its target, RNAI. Mol Microbiol *42*, 527-537.

Gregory, R.I., Chendrimada, T.P., Cooch, N., and Shiekhattar, R. (2005). Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. Cell *123*, 631-640.

Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235-240.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res *34*, D140-144.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: tools for microRNA genomics. Nucleic Acids Res *36*, D154-158.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular cell *27*, 91-105.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., Chiang, H.R., King, N., Degnan, B.M., Rokhsar, D.S., and Bartel, D.P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature *455*, 1193-1197.

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. Cell *106*, 23-34.

Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., and Rajewsky, N. (2014). Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. Molecular cell *54*, 1042-1054.

Grun, D., Wang, Y.L., Langenberger, D., Gunsalus, K.C., and Rajewsky, N. (2005). microRNA target predictions across seven Drosophila species and comparison to mammalian targets. PLoS Comput Biol *1*, e13.

Gu, S., Jin, L., Zhang, F., Sarnow, P., and Kay, M.A. (2009). Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. Nat Struct Mol Biol *16*, 144-150.

Gu, S., Zhang, Y., Jin, L., Huang, Y., Zhang, F., Bassik, M.C., Kampmann, M., and Kay, M.A. (2014). Weak base pairing in both seed and 3' regions reduces

RNAi off-targets and enhances si/shRNA designs. Nucleic Acids Res *42*, 12169-12176.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835-840.

Gwack, Y., Baek, H.J., Nakamura, H., Lee, S.H., Meisterernst, M., Roeder, R.G., and Jung, J.U. (2003). Principal role of TRAP/mediator and SWI/SNF complexes in Kaposi's sarcoma-associated herpesvirus RTA-mediated lytic reactivation. Mol Cell Biol *23*, 2055-2067.

Haase, A.D., Jaskiewicz, L., Zhang, H., Laine, S., Sack, R., Gatignol, A., and Filipowicz, W. (2005). TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. EMBO Rep *6*, 961-967.

Haley, B., and Zamore, P.D. (2004). Kinetic analysis of the RNAi enzyme complex. Nat Struct Mol Biol *11*, 599-606.

Hamilton, A.J., and Baulcombe, D.C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science *286*, 950-952.

Hammell, M., Long, D., Zhang, L., Lee, A., Carmack, C.S., Han, M., Ding, Y., and Ambros, V. (2008). mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. Nat Methods *5*, 813-819.

Hammond, S.M., Bernstein, E., Beach, D., and Hannon, G.J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature *404*, 293-296.

Hammond, S.M., Boettcher, S., Caudy, A.A., Kobayashi, R., and Hannon, G.J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. Science *293*, 1146-1150.

Han, J. (2009). Posttranscriptional crossregulation between Drosha and DGCR8. Cell *136*, 75-84.

Han, J., Lee, Y., Yeom, K.H., Kim, Y.K., Jin, H., and Kim, V.N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. Genes & development *18*, 3016-3027.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell *125*, 887-901.

Hannon, G.J. (2002). RNA interference. Nature *418*, 244-251.

He, L. (2005). A microRNA polycistron as a potential human oncogene. Nature *435*, 828-833.

He, L., He, X., Lowe, S.W., and Hannon, G.J. (2007). microRNAs join the p53 network [mdash] another piece in the tumour-suppression puzzle. Nature Rev Cancer *7*, 819-822.

Helfer, S., Schott, J., Stoecklin, G., and Forstemann, K. (2012). AU-rich element-mediated mRNA decay can occur independently of the miRNA machinery in mouse embryonic fibroblasts and Drosophila S2-cells. PLoS One *7*, e28907.

Hibio, N., Hino, K., Shimizu, E., Nagata, Y., and Ui-Tei, K. (2012). Stability of miRNA 5'terminal and seed regions is correlated with experimentally observed miRNA-mediated silencing efficacy. Sci Rep *2*, 996.

Higaki, K., Hirao, M., Kawana-Tachikawa, A., Iriguchi, S., Kumagai, A., Ueda, N., Bo, W., Kamibayashi, S., Watanabe, A., Nakauchi, H.*, et al.* (2018). Generation of HIV-Resistant Macrophages from IPSCs by Using Transcriptional Gene Silencing and Promoter-Targeted RNA. Mol Ther Nucleic Acids *12*, 793-804.

Hiscott, J. (2001). Introduction--cytokine receptors, signaling pathways and viruses. Cytokine Growth Factor Rev *12*, 129-131.

Holen, T., Amarzguioui, M., Wiiger, M.T., Babaie, E., and Prydz, H. (2002). Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. Nucleic Acids Res *30*, 1757-1766.

Horak, M., Novak, J., and Bienertova-Vasku, J. (2016). Muscle-specific microRNAs in skeletal muscle development. Dev Biol *410*, 1-13.

Horman, S.R., Janas, M.M., Litterst, C., Wang, B., MacRae, I.J., Sever, M.J., Morrissey, D.V., Graves, P., Luo, B., Umesalma, S.*, et al.* (2013). Akt-mediated phosphorylation of argonaute 2 downregulates cleavage and upregulates translational repression of MicroRNA targets. Molecular cell *50*, 356-367.

Houseley, J., LaCava, J., and Tollervey, D. (2006). RNA-quality control by the exosome. Nat Rev Mol Cell Biol *7*, 529-539.

Houzet, L., Klase, Z., Yeung, M.L., Wu, A., Le, S.Y., Quinones, M., and Jeang, K.T. (2012). The extent of sequence complementarity correlates with the potency of cellular miRNA-mediated restriction of HIV-1. Nucleic Acids Res *40*, 11684-11696.

Huang, J., Liang, Z., Yang, B., Tian, H., Ma, J., and Zhang, H. (2007a). Derepression of microRNA-mediated protein translation inhibition by apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G (APOBEC3G) and its family members. J Biol Chem *282*, 33632-33640.

Huang, J.C., Babak, T., Corson, T.W., Chua, G., Khan, S., Gallie, B.L., Hughes, T.R., Blencowe, B.J., Frey, B.J., and Morris, Q.D. (2007b). Using expression profiling data to identify human microRNA targets. Nat Methods *4*, 1045-1049.

Humphreys, D.T., Westman, B.J., Martin, D.I., and Preiss, T. (2005). MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. Proc Natl Acad Sci U S A *102*, 16961-16966.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science *293*, 834-838.

Hutvagner, G., and Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. Nature Reviews Molecular Cell Biology *9*, 22.

Hutvágner, G., and Zamore, P.D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. Science *297*, 2056-2060.

Iwasaki, S., Kawamata, T., and Tomari, Y. (2009). Drosophila argonaute1 and argonaute2 employ distinct mechanisms for translational repression. Molecular cell *34*, 58-67.

Janas, M.M., Wang, B., Harris, A.S., Aguiar, M., Shaffer, J.M., Subrahmanyam, Y.V., Behlke, M.A., Wucherpfennig, K.W., Gygi, S.P., Gagnon, E.*, et al.* (2012). Alternative RISC assembly: binding and repression of microRNA-mRNA duplexes by human Ago proteins. RNA *18*, 2041-2055.

Jo, M.H., Shin, S., Jung, S.R., Kim, E., Song, J.J., and Hohng, S. (2015). Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs. Molecular cell *59*, 117-124.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. (2004). Human MicroRNA targets. PLoS biology *2*, e363.

Kamola, P.J., Nakano, Y., Takahashi, T., Wilson, P.A., and Ui-Tei, K. (2015). The siRNA Non-seed Region and Its Target Sequences Are Auxiliary Determinants of Off-Target Effects. PLoS Comput Biol *11*, e1004656.

Kato, H., Goto, D.B., Martienssen, R.A., Urano, T., Furukawa, K., and Murakami, Y. (2005). RNA polymerase II is required for RNAi-dependent heterochromatin assembly. Science *309*, 467-469.

Katoh, T., and Suzuki, T. (2007). Specific residues at every third position of siRNA shape its efficient RNAi activity. Nucleic Acids Res *35*, e27.

Kawamata, T., and Tomari, Y. (2010). Making RISC. Trends Biochem Sci *35*, 368-376.

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., Okada, T.N., Siomi, M.C., and Siomi, H. (2008). Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. Nature *453*, 793-797.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. Nat Genet *39*, 1278-1284.

Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. Genes & development *15*, 2654-2659.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209-216.

Kiethega, G.N., Yan, Y., Turcotte, M., and Burger, G. (2013). RNA-level unscrambling of fragmented genes in Diplonema mitochondria. RNA Biol *10*, 301-313.

Kim, S.K., Nam, J.W., Rhee, J.K., Lee, W.J., and Zhang, B.T. (2006a). miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics *7*, 411.

Kim, V.N. (2005). MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol *6*, 376-385.

Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol *10*, 126-139.

Kim, Y.K., Bourgeois, C.F., Pearson, R., Tyagi, M., West, M.J., Wong, J., Wu, S.Y., Chiang, C.M., and Karn, J. (2006b). Recruitment of TFIIH to the HIV LTR is a rate-limiting step in the emergence of HIV from latency. EMBO J *25*, 3596-3604.

Kim, Y.K., and Kim, V.N. (2007a). Processing of intronic microRNAs. EMBO J *26*, 775-783.

Kim, Y.K., and Kim, V.N. (2007b). Processing of intronic microRNAs. EMBO J *26*, 775-783.

Kinch, L.N., and Grishin, N.V. (2009). The human Ago2 MC region does not contain an eIF4E-like mRNA cap binding motif. Biol Direct *4*, 2.

Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. Genes & development *18*, 1165-1178.

Kiwanuka, N., Laeyendecker, O., Robb, M., Kigozi, G., Arroyo, M., McCutchan, F., Eller, L.A., Eller, M., Makumbi, F., Birx, D.*, et al.* (2008). Effect of human immunodeficiency virus Type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection. J Infect Dis *197*, 707-713.

Klase, Z., Houzet, L., and Jeang, K.T. (2012). MicroRNAs and HIV-1: complex interactions. J Biol Chem *287*, 40884-40890.

Knight, S.W., and Bass, B.L. (2001). A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. Science *293*, 2269-2271.

Knott, Simon R.V., Maceli, A.R., Erard, N., Chang, K., Marran, K., Zhou, X., Gordon, A., El Demerdash, O., Wagenblast, E., Kim, S.*, et al.* (2014). A Computational Algorithm to Predict shRNA Potency. Molecular cell *56*, 796-807.

Konig, R., Zhou, Y., Elleder, D., Diamond, T.L., Bonamy, G.M., Irelan, J.T., Chiang, C.Y., Tu, B.P., De Jesus, P.D., Lilley, C.E.*, et al.* (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. Cell *135*, 49-60.

Korner, C.G., and Wahle, E. (1997). Poly(A) tail shortening by a mammalian poly(A)-specific 3'-exoribonuclease. J Biol Chem *272*, 10448-10456.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res *42*, D68-73.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M.*, et al.* (2005). Combinatorial microRNA target predictions. Nat Genet *37*, 495-500.

Kretzschmar, M., Meisterernst, M., Scheidereit, C., Li, G., and Roeder, R.G. (1992). Transcriptional regulation of the HIV-1 promoter by NF-kappa B in vitro. Genes & development *6*, 761-774.

Kruger, J., and Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res *34*, W451-454.

Kwak, D., Choi, S., Jeong, H., Jang, J.H., Lee, Y., Jeon, H., Lee, M.N., Noh, J., Cho, K., Yoo, J.S.*, et al.* (2012). Osmotic stress regulates mammalian target of rapamycin (mTOR) complex 1 via c-Jun N-terminal Kinase (JNK)-mediated Raptor protein phosphorylation. J Biol Chem *287*, 18398-18407.

Lall, S., Grun, D., Krek, A., Chen, K., Wang, Y.L., Dewey, C.N., Sood, P., Colombo, T., Bray, N., Macmenamin, P.*, et al.* (2006). A genome-wide map of conserved microRNA targets in C. elegans. Curr Biol *16*, 460-471.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M.*, et al.* (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell *129*, 1401-1414.

Landthaler, M., Yalcin, A., and Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. Curr Biol *14*, 2162-2167.

Lee, E.J. (2008). Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors. RNA *14*, 35-42.

Lee, N.S., Dohjima, T., Bauer, G., Li, H., Li, M.J., Ehsani, A., Salvaterra, P., and Rossi, J. (2002a). Expression of small interfering RNAs targeted against HIV-1 rev transcripts in human cells. Nat Biotechnol *20*, 500-505.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Lee, S.K., Dykxhoorn, D.M., Kumar, P., Ranjbar, S., Song, E., Maliszewski, L.E., Francois-Bongarcon, V., Goldfeld, A., Swamy, N.M., Lieberman, J.*, et al.* (2005). Lentiviral delivery of short hairpin RNAs protects CD4 T cells from multiple clades and primary isolates of HIV. Blood *106*, 818-826.

Lee, T., Di Paola, D., Malina, A., Mills, J.R., Kreps, A., Grosse, F., Tang, H., Zannis-Hadjopoulos, M., Larsson, O., and Pelletier, J. (2014). Suppression of the DHX9 helicase induces premature senescence in human diploid fibroblasts in a p53-dependent manner. J Biol Chem *289*, 22798-22814.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S.*, et al.* (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415-419.

Lee, Y., Hur, I., Park, S.Y., Kim, Y.K., Suh, M.R., and Kim, V.N. (2006). The role of PACT in the RNA silencing pathway. EMBO J *25*, 522-532.

Lee, Y., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. (2002b). MicroRNA maturation: stepwise processing and subcellular localization. EMBO J *21*, 4663-4670.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J *23*, 4051-4060.

Leuschner, P.J., Ameres, S.L., Kueng, S., and Martinez, J. (2006). Cleavage of the siRNA passenger strand during RISC assembly in human cells. EMBO Rep *7*, 314-320.

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

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787-798.

Li, S., Lian, S.L., Moser, J.J., Fritzler, M.L., Fritzler, M.J., Satoh, M., and Chan, E.K. (2008). Identification of GW182 and its novel isoform TNGW1 as translational repressors in Ago2-mediated silencing. J Cell Sci *121*, 4134-4144.

Lian, S.L., Li, S., Abadal, G.X., Pauley, B.A., Fritzler, M.J., and Chan, E.K. (2009). The C-terminal half of human Ago2 binds to multiple GW-rich regions of GW182 and requires GW182 to mediate silencing. RNA *15*, 804-813.

Libermann, T.A., and Baltimore, D. (1990). Activation of interleukin-6 gene expression through the NF-kappa B transcription factor. Mol Cell Biol *10*, 2327-2334.

Lima, W.F., Wu, H., Nichols, J.G., Sun, H., Murray, H.M., and Crooke, S.T. (2009). Binding and cleavage specificities of human Argonaute2. J Biol Chem *284*, 26017-26028.

Lin, X., Ruan, X., Anderson, M.G., McDowell, J.A., Kroeger, P.E., Fesik, S.W., and Shen, Y. (2005). siRNA-mediated off-target gene silencing triggered by a 7 nt complementation. Nucleic Acids Res *33*, 4527-4535.

Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2003). Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain. Nature *426*, 465-469.

Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2004). Nucleic acid 3'-end recognition by the Argonaute2 PAZ domain. Nat Struct Mol Biol *11*, 576-577.

Lippman, Z., and Martienssen, R. (2004). The role of RNA interference in heterochromatic silencing. Nature *431*, 364-370.

Liu, J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437-1441.

Liu, L., Oliveira, N.M., Cheney, K.M., Pade, C., Dreja, H., Bergin, A.M., Borgdorff, V., Beach, D.H., Bishop, C.L., Dittmar, M.T.*, et al.* (2011). A whole genome screen for HIV restriction factors. Retrovirology *8*, 94.

Liu, Q., Rand, T.A., Kalidas, S., Du, F., Kim, H.E., Smith, D.P., and Wang, X. (2003). R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. Science *301*, 1921-1925.

Liu, Y.P., Gruber, J., Haasnoot, J., Konstantinova, P., and Berkhout, B. (2009). RNAi-mediated inhibition of HIV-1 by targeting partially complementary viral sequences. Nucleic Acids Res *37*, 6194-6204.

Loflin, P., Chen, C.Y., and Shyu, A.B. (1999). Unraveling a cytoplasmic role for hnRNP D in the in vivo mRNA destabilization directed by the AU-rich element. Genes & development *13*, 1884-1897.

Long, D., Lee, R., Williams, P., Chan, C.Y., Ambros, V., and Ding, Y. (2007). Potent effect of target structure on microRNA function. Nat Struct Mol Biol *14*, 287-294.

Lund, E., Guttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science *303*, 95-98.

Luo, X., Zhang, J., Wang, H., Du, Y., Yang, L., Zheng, F., and Ma, D. (2012). PolyA RT-PCR-based quantification of microRNA by using universal TaqMan probe. Biotechnol Lett *34*, 627-633.

Ma, J., Yu, M.K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., and Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. Nat Methods *15*, 290-298.

Ma, J.B., Ye, K., and Patel, D.J. (2004). Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. Nature *429*, 318-322.

Ma, J.B., Yuan, Y.R., Meister, G., Pei, Y., Tuschl, T., and Patel, D.J. (2005). Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature *434*, 666-670.

MacRae, I.J., Ma, E., Zhou, M., Robinson, C.V., and Doudna, J.A. (2008). In vitro reconstitution of the human RISC-loading complex. Proc Natl Acad Sci U S A *105*, 512-517.

Majoros, W.H., Lekprasert, P., Mukherjee, N., Skalsky, R.L., Corcoran, D.L., Cullen, B.R., and Ohler, U. (2013). MicroRNA target site identification by integrating sequence and binding information. Nat Methods *10*, 630-633.

Malina, A., Mills, J.R., Cencic, R., Yan, Y., Fraser, J., Schippers, L.M., Paquet, M., Dostie, J., and Pelletier, J. (2013). Repurposing CRISPR/Cas9 for in situ functional assays. Genes & development *27*, 2602-2614.

Manche, L., Green, S.R., Schmedt, C., and Mathews, M.B. (1992). Interactions between double-stranded RNA regulators and the protein kinase DAI. Mol Cell Biol *12*, 5238-5248.

Maniataki, E., and Mourelatos, Z. (2005). A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. Genes & development *19*, 2979-2990.

Maragkakis, M., Alexiou, P., Papadopoulos, G.L., Reczko, M., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Simossis, V.A.*, et al.* (2009). Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinformatics *10*, 295.

Maroney, P.A., Yu, Y., Fisher, J., and Nilsen, T.W. (2006). Evidence that microRNAs are associated with translating messenger RNAs in human cells. Nat Struct Mol Biol *13*, 1102-1107.

Martin, H.C., Wani, S., Steptoe, A.L., Krishnan, K., Nones, K., Nourbakhsh, E., Vlassov, A., Grimmond, S.M., and Cloonan, N. (2014). Imperfect centered miRNA binding sites are common and can mediate repression of target mRNAs. Genome Biol *15*, R51.

Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell *110*, 563-574.

Martinez, J., and Tuschl, T. (2004). RISC is a 5′ phosphomonoester-producing RNA endonuclease. Genes Dev *18*, 975-980.

Matranga, C., Tomari, Y., Shin, C., Bartel, D.P., and Zamore, P.D. (2005). Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. Cell *123*, 607-620.

Maxwell, E.K., Ryan, J.F., Schnitzler, C.E., Browne, W.E., and Baxevanis, A.D. (2012). MicroRNAs and essential components of the microRNA processing machinery are not encoded in the genome of the ctenophore Mnemiopsis leidyi. BMC Genomics *13*, 714.

Mayya, V.K., and Duchaine, T.F. (2015). On the availability of microRNA-induced silencing complexes, saturation of microRNA-binding sites and stoichiometry. Nucleic Acids Res *43*, 7556-7565.

Meister, G. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol Cell *15*, 185-197.

Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. Nat Rev Genet *14*, 447-459.

Meister, G., and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. Nature *431*, 343-349.

Mello, C.C., and Conte, D., Jr. (2004). Revealing the world of RNA interference. Nature *431*, 338-342.

Meyer, S., Temme, C., and Wahle, E. (2004). Messenger RNA turnover in eukaryotes: pathways and enzymes. Crit Rev Biochem Mol Biol *39*, 197-216.

Mills, J.R., Malina, A., Lee, T., Di Paola, D., Larsson, O., Miething, C., Grosse, F., Tang, H., Zannis-Hadjopoulos, M., Lowe, S.W.*, et al.* (2013). RNAi

screening uncovers Dhx9 as a modifier of ABT-737 resistance in an Emu-myc/Bcl-2 mouse model. Blood *121*, 3402-3412.

Minks, M.A., West, D.K., Benvin, S., and Baglioni, C. (1979). Structural requirements of double-stranded RNA for the activation of 2',5'-oligo(A) polymerase and protein kinase of interferon-treated HeLa cells. J Biol Chem *254*, 10180-10183.

Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell *126*, 1203-1217.

Mittler, G., Stuhler, T., Santolin, L., Uhlmann, T., Kremmer, E., Lottspeich, F., Berti, L., and Meisterernst, M. (2003). A novel docking site on Mediator is critical for activation by VP16 in mammalian cells. EMBO J *22*, 6494-6504.

Miyoshi, K., Tsukumo, H., Nagami, T., Siomi, H., and Siomi, M.C. (2005). Slicer function of Drosophila Argonautes and its involvement in RISC formation. Genes & development *19*, 2837-2848.

Modzelewski, A.J., Holmes, R.J., Hilz, S., Grimson, A., and Cohen, P.E. (2012). AGO4 regulates entry into meiosis and influences silencing of sex chromosomes in the male mouse germline. Dev Cell *23*, 251-264.

Moore, C.B., Guthrie, E.H., Huang, M.T., and Taxman, D.J. (2010). Short hairpin RNA (shRNA): design, delivery, and assessment of gene knockdown. Methods Mol Biol *629*, 141-158.

Morlando, M., Ballarino, M., Gromak, N., Pagano, F., Bozzoni, I., and Proudfoot, N.J. (2008). Primary microRNA transcripts are processed co-transcriptionally. Nat Struct Mol Biol *15*, 902-909.

Mukherji, S., Ebert, M.S., Zheng, G.X., Tsang, J.S., Sharp, P.A., and van Oudenaarden, A. (2011). MicroRNAs can generate thresholds in target gene expression. Nat Genet *43*, 854-859.

Murchison, E.P., Partridge, J.F., Tam, O.H., Cheloufi, S., and Hannon, G.J. (2005). Characterization of Dicer-deficient murine embryonic stem cells. Proc Natl Acad Sci U S A *102*, 12135-12140.

Nakanishi, K., Ascano, M., Gogakos, T., Ishibe-Murakami, S., Serganov, A.A., Briskin, D., Morozov, P., Tuschl, T., and Patel, D.J. (2013). Eukaryote-specific insertion elements control human ARGONAUTE slicer activity. Cell Rep *3*, 1893-1900.

Newman, M.A., Thomson, J.M., and Hammond, S.M. (2008). Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. Rna *14*, 1539-1549.

Nielsen, C.B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C.B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA *13*, 1894-1910.

Noland, C.L., Ma, E., and Doudna, J.A. (2011). siRNA repositioning for guide strand selection by human Dicer complexes. Molecular cell *43*, 110-121.

Nottrott, S., Simard, M.J., and Richter, J.D. (2006). Human let-7a miRNA blocks protein production on actively translating polyribosomes. Nat Struct Mol Biol *13*, 1108-1114.

Novina, C.D., Murray, M.F., Dykxhoorn, D.M., Beresford, P.J., Riess, J., Lee, S.K., Collman, R.G., Lieberman, J., Shankar, P., and Sharp, P.A. (2002). siRNA-directed inhibition of HIV-1 infection. Nat Med *8*, 681-686.

Nykanen, A., Haley, B., and Zamore, P.D. (2001). ATP requirements and small interfering RNA structure in the RNA interference pathway. Cell *107*, 309-321.

O'Carroll, D., Mecklenbrauker, I., Das, P.P., Santana, A., Koenig, U., Enright, A.J., Miska, E.A., and Tarakhovsky, A. (2007). A Slicer-independent role for Argonaute 2 in hematopoiesis and the microRNA pathway. Genes & development *21*, 1999-2004.

Okamura, K., Balla, S., Martin, R., Liu, N., and Lai, E.C. (2008a). Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. Nat Struct Mol Biol *15*, 998.

Okamura, K., Chung, W.J., Ruby, J.G., Guo, H., Bartel, D.P., and Lai, E.C. (2008b). The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. Nature *453*, 803-806.

Olsen, P.H., and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev Biol *216*, 671-680.

Paddison, P.J., Cleary, M., Silva, J.M., Chang, K., Sheth, N., Sachidanandam, R., and Hannon, G.J. (2004). Cloning of short hairpin RNAs for gene knockdown in mammalian cells. Nat Meth *1*, 163-167.

Pahl, H.L. (1999). Activators and target genes of Rel/NF-kappaB transcription factors. Oncogene *18*, 6853-6866.

Parisien, M., and Major, F. (2008). The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature *452*, 51-55.

Parker, J.S., Roe, S.M., and Barford, D. (2004). Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. Embo j *23*, 4727-4737.

Parker, J.S., Roe, S.M., and Barford, D. (2005). Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. Nature *434*, 663-666.

Pasquinelli, A.E. (2012). MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet *13*, 271-282.

Pawlicki, J.M., and Steitz, J.A. (2008). Primary microRNA transcript retention at sites of transcription leads to enhanced microRNA production. J Cell Biol *182*, 61-76.

Perreault, J., Weinberg, Z., Roth, A., Popescu, O., Chartrand, P., Ferbeyre, G., and Breaker, R.R. (2011). Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. PLoS Comput Biol *7*, e1002031.

Peters, L., and Meister, G. (2007). Argonaute proteins: mediators of RNA silencing. Molecular cell *26*, 611-623.

Petersen, C.P., Bordeleau, M.E., Pelletier, J., and Sharp, P.A. (2006). Short RNAs repress translation after initiation in mammalian cells. Molecular cell *21*, 533-542.

Pillai, R.S., Bhattacharyya, S.N., Artus, C.G., Zoller, T., Cougot, N., Basyuk, E., Bertrand, E., and Filipowicz, W. (2005). Inhibition of translational initiation by Let-7 MicroRNA in human cells. Science *309*, 1573-1576.

Pinzon, N., Li, B., Martinez, L., Sergeeva, A., Presumey, J., Apparailly, F., and Seitz, H. (2017). microRNA target prediction programs predict many false positives. Genome Res *27*, 234-245.

Preall, J.B., and Sontheimer, E.J. (2005). RNAi: RISC gets loaded. Cell *123*, 543-545.

Qin, X.F., An, D.S., Chen, I.S., and Baltimore, D. (2003). Inhibiting HIV-1 infection in human T cells by lentiviral-mediated delivery of small interfering RNA against CCR5. Proc Natl Acad Sci U S A *100*, 183-188.

Ragan, C., Zuker, M., and Ragan, M.A. (2011). Quantitative prediction of miRNA-mRNA interaction based on equilibrium concentrations. PLoS Comput Biol *7*, e1001090.

Rajewsky, N., and Socci, N.D. (2004). Computational identification of microRNA targets. Dev Biol *267*, 529-535.

Rand, T.A., Ginalski, K., Grishin, N.V., and Wang, X. (2004). Biochemical identification of Argonaute 2 as the sole protein required for RNA-induced silencing complex activity. Proc Natl Acad Sci U S A *101*, 14385-14389.

Rao, P.K., Kumar, R.M., Farkhondeh, M., Baskerville, S., and Lodish, H.F. (2006). Myogenic factors that regulate expression of muscle-specific microRNAs. Proc Natl Acad Sci USA *103*, 8721-8726.

Rashid, U.J., Paterok, D., Koglin, A., Gohlke, H., Piehler, J., and Chen, J.C. (2007). Structure of Aquifex aeolicus argonaute highlights conformational flexibility of the PAZ domain as a potential regulator of RNA-induced silencing complex function. J Biol Chem *282*, 13824-13832.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. RNA *10*, 1507-1517.

Rehwinkel, J., Behm-Ansmant, I., Gatfield, D., and Izaurralde, E. (2005). A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. RNA *11*, 1640-1647.

Reyes-Herrera, P.H., and Ficarra, E. (2012). One decade of development and evolution of microRNA target prediction algorithms. Genomics Proteomics Bioinformatics *10*, 254-263.

Reyes-Herrera, P.H., Ficarra, E., Acquaviva, A., and Macii, E. (2011). miREE: miRNA recognition elements ensemble. BMC Bioinformatics *12*, 454.

Rivas, F.V., Tolia, N.H., Song, J.J., Aragon, J.P., Liu, J., Hannon, G.J., and Joshua-Tor, L. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. Nat Struct Mol Biol *12*, 340-349.

Robertson, B., Dalby, A.B., Karpilow, J., Khvorova, A., Leake, D., and Vermeulen, A. (2010). Specificity and functionality of microRNA inhibitors. Silence *1*, 10.

Robins, H., Li, Y., and Padgett, R.W. (2005). Incorporating structure to predict microRNA targets. Proc Natl Acad Sci U S A *102*, 4006-4009.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., and Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. Genome Res *14*, 1902-1910.

Rodriguez, A.J., Czaplinski, K., Condeelis, J.S., and Singer, R.H. (2008). Mechanisms and cellular roles of local protein synthesis in mammalian cells. Curr Opin Cell Biol *20*, 144-149.

Rossi, J.J., June, C.H., and Kohn, D.B. (2007). Genetic therapies against HIV. Nat Biotechnol *25*, 1444-1454.

Ruan, J., Chen, H., Kurgan, L., Chen, K., Kang, C., and Pu, P. (2008). HuMiTar: a sequence-based method for prediction of human microRNA targets. Algorithms Mol Biol *3*, 16.

Rudel, S., Wang, Y., Lenobel, R., Korner, R., Hsiao, H.H., Urlaub, H., Patel, D., and Meister, G. (2011). Phosphorylation of human Argonaute proteins affects small RNA binding. Nucleic Acids Res *39*, 2330-2343.

Rusinov, V., Baev, V., Minkov, I.N., and Tabler, M. (2005). MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. Nucleic Acids Res *33*, W696-700.

Sabariegos, R., Gimenez-Barcons, M., Tapia, N., Clotet, B., and Martinez, M.A. (2006). Sequence homology required by human immunodeficiency virus type 1 to escape from short interfering RNAs. Journal of virology *80*, 571-577.

Saetrom, O., Snove, O., Jr., and Saetrom, P. (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA *11*, 995-1003.

Saetrom, P., Heale, B.S., Snove, O., Jr., Aagaard, L., Alluin, J., and Rossi, J.J. (2007). Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res *35*, 2333-2342.

Saito, T., and Saetrom, P. (2010). MicroRNAs--targeting and target prediction. N Biotechnol *27*, 243-249.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353-358.

Salomon, W.E., Jolly, S.M., Moore, M.J., Zamore, P.D., and Serebrov, V. (2015). Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides. Cell *162*, 84-95.

Saraiya, A.A., Li, W., and Wang, C.C. (2013). Transition of a microRNA from repressing to activating translation depending on the extent of base pairing with the target. PLoS One *8*, e55672.

Saxena, S., Jonsson, Z.O., and Dutta, A. (2003). Small RNAs with imperfect match to endogenous mRNA repress translation. Implications for off-target activity of small inhibitory RNA in mammalian cells. J Biol Chem *278*, 44312-44319.

Schirle, N.T., and MacRae, I.J. (2012). The crystal structure of human Argonaute2. Science *336*, 1037-1040.

Schirle, N.T., Sheu-Gruttadauria, J., and MacRae, I.J. (2014). Structural basis for microRNA targeting. Science *346*, 608-613.

Schwarz, D.S., Ding, H., Kennington, L., Moore, J.T., Schelter, J., Burchard, J., Linsley, P.S., Aronin, N., Xu, Z., and Zamore, P.D. (2006). Designing siRNA that distinguish between genes that differ by a single nucleotide. PLoS Genet *2*, e140.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199-208.

Seitz, H. (2017). Issues in current microRNA target identification methods. RNA Biol *14*, 831-834.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58-63.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A *108*, 10098-10103.

Shen, J., Xia, W., Khotskaya, Y.B., Huo, L., Nakanishi, K., Lim, S.O., Du, Y., Wang, Y., Chang, W.C., Chen, C.H.*, et al.* (2013). EGFR modulates microRNA maturation in response to hypoxia through phosphorylation of AGO2. Nature *497*, 383-387.

Shin, C., Nam, J.W., Farh, K.K., Chiang, H.R., Shkumatava, A., and Bartel, D.P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. Molecular cell *38*, 789-802.

Siomi, H., and Siomi, M.C. (2010). Posttranscriptional regulation of microRNA biogenesis in animals. Molecular cell *38*, 323-332.

Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell *136*, 731-745.

Song, J.J. (2003). The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. Nature Struct Biol *10*, 1026-1032.

Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. Science *305*, 1434-1437.

Stanhope, S.A., Sengupta, S., den Boon, J., Ahlquist, P., and Newton, M.A. (2009). Statistical use of argonaute expression and RISC assembly in microRNA target identification. PLoS Comput Biol *5*, e1000516.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell *123*, 1133-1146.

Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. (2003). Identification of Drosophila MicroRNA targets. PLoS biology *1*, E60.

Stegmeier, F., Hu, G., Rickles, R.J., Hannon, G.J., and Elledge, S.J. (2005). A lentiviral microRNA-based system for single-copy polymerase II-regulated RNA interference in mammalian cells. Proc Natl Acad Sci U S A *102*, 13212-13217.

Steiger, M., Carr-Schmid, A., Schwartz, D.C., Kiledjian, M., and Parker, R. (2003). Analysis of recombinant yeast decapping enzyme. RNA *9*, 231-238.

Stuart, K., Allen, T.E., Heidmann, S., and Seiwert, S.D. (1997). RNA editing in kinetoplastid protozoa. Microbiol Mol Biol Rev *61*, 105-120.

Sturm, M., Hackenberg, M., Langenberger, D., and Frishman, D. (2010). TargetSpy: a supervised machine learning approach for microRNA target prediction. BMC Bioinformatics *11*, 292.

Su, H., Trombly, M.I., Chen, J., and Wang, X. (2009). Essential and overlapping functions for mammalian Argonautes in microRNA silencing. Genes & development *23*, 304-317.

Suzuki, H.I., Katsura, A., Yasuda, T., Ueno, T., Mano, H., Sugimoto, K., and Miyazono, K. (2015). Small-RNA asymmetry is directly driven by mammalian Argonautes. Nat Struct Mol Biol *22*, 512-521.

Suzuki, H.I., and Miyazono, K. (2010). Dynamics of microRNA biogenesis: crosstalk between p53 network and microRNA processing pathway. J Mol Med (Berl) *88*, 1085-1094.

Suzuki, H.I., and Miyazono, K. (2011). Emerging complexity of microRNA generation cascades. J Biochem *149*, 15-25.

Swarts, D.C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R.F., Koonin, E.V., Patel, D.J., and van der Oost, J. (2014). The evolutionary journey of Argonaute proteins. Nat Struct Mol Biol *21*, 743-753.

Tabara, H., Yigit, E., Siomi, H., and Mello, C.C. (2002). The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in C. elegans. Cell *109*, 861-871.

Takimoto, K., Wakiyama, M., and Yokoyama, S. (2009). Mammalian GW182 contains multiple Argonaute-binding sites and functions in microRNA-mediated translational repression. RNA *15*, 1078-1089.

Tallsjo, A., and Kirsebom, L.A. (1993). Product release is a rate-limiting step during cleavage by the catalytic RNA subunit of Escherichia coli RNase P. Nucleic Acids Res *21*, 51-57.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M.*, et al.* (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534-538.

Tan, X., Lu, Z.J., Gao, G., Xu, Q., Hu, L., Fellmann, C., Li, M.Z., Qu, H., Lowe, S.W., Hannon, G.J.*, et al.* (2012). Tiling genomes of pathogenic viruses identifies potent antiviral shRNAs and reveals a role for secondary structure in shRNA efficacy. Proc Natl Acad Sci U S A *109*, 869-874.

Tang, G., Reinhart, B.J., Bartel, D.P., and Zamore, P.D. (2003). A biochemical framework for RNA silencing in plants. Genes & development *17*, 49-63.

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S.M., Ala, U., Karreth, F., Poliseno, L., Provero, P., Di Cunto, F.*, et al.* (2011). Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell *147*, 344-357.

ter Brake, O., Konstantinova, P., Ceylan, M., and Berkhout, B. (2006). Silencing of HIV-1 with RNA interference: a multiple shRNA approach. Molecular therapy : the journal of the American Society of Gene Therapy *14*, 883-892.

ter Brake, O., Legrand, N., von Eije, K.J., Centlivre, M., Spits, H., Weijer, K., Blom, B., and Berkhout, B. (2009). Evaluation of safety and efficacy of RNAi

against HIV-1 in the human immune system (Rag-2(-/-)gammac(-/-)) mouse model. Gene Ther *16*, 148-153.

Thadani, R., and Tammi, M.T. (2006). MicroTar: predicting microRNA targets from RNA duplexes. BMC Bioinformatics *7 Suppl 5*, S20.

Till, S., Lejeune, E., Thermann, R., Bortfeld, M., Hothorn, M., Enderle, D., Heinrich, C., Hentze, M.W., and Ladurner, A.G. (2007). A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. Nat Struct Mol Biol *14*, 897-903.

Tokumaru, S., Suzuki, M., Yamada, H., Nagino, M., and Takahashi, T. (2008). let-7 regulates Dicer expression and constitutes a negative feedback loop. Carcinogenesis *29*, 2073-2077.

Tolia, N.H., and Joshua-Tor, L. (2007). Slicer and the argonautes. Nature Chem Biol *3*, 36-43.

Tomari, Y., Du, T., and Zamore, P.D. (2007). Sorting of Drosophila small silencing RNAs. Cell *130*, 299-308.

Tomari, Y., Matranga, C., Haley, B., Martinez, N., and Zamore, P.D. (2004). A protein sensor for siRNA asymmetry. Science *306*, 1377-1380.

Tomari, Y., and Zamore, P.D. (2005). Perspective: machines for RNAi. Genes & development *19*, 517-529.

Tucker, M., Staples, R.R., Valencia-Sanchez, M.A., Muhlrad, D., and Parker, R. (2002). Ccr4p is the catalytic subunit of a Ccr4p/Pop2p/Notp mRNA deadenylase complex in Saccharomyces cerevisiae. EMBO J *21*, 1427-1436.

Tuschl, T., Zamore, P.D., Lehmann, R., Bartel, D.P., and Sharp, P.A. (1999). Targeted mRNA degradation by double-stranded RNA in vitro. Genes & development *13*, 3191-3197.

Tyson, J.J., and Novak, B. (2010). Functional motifs in biochemical reaction networks. Annu Rev Phys Chem *61*, 219-240.

Vainberg Slutskin, I., Weingarten-Gabbay, S., Nir, R., Weinberger, A., and Segal, E. (2018). Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. Nat Commun *9*, 529.

van Rij, R.P., Saleh, M.C., Berry, B., Foo, C., Houk, A., Antoniewski, C., and Andino, R. (2006). The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster. Genes & development *20*, 2985-2995.

Van Stry, M., Oguin, T.H., 3rd, Cheloufi, S., Vogel, P., Watanabe, M., Pillai, M.R., Dash, P., Thomas, P.G., Hannon, G.J., and Bix, M. (2012). Enhanced susceptibility of Ago1/3 double-null mice to influenza A virus infection. Journal of virology *86*, 4151-4157.

Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gasciolli, V., Mallory, A.C., Hilbert, J.L., Bartel, D.P., and Crete, P. (2004). Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. Molecular cell *16*, 69-79.

Virtanen, A., Henriksson, N., Nilsson, P., and Nissbeck, M. (2013). Poly(A)-specific ribonuclease (PARN): an allosterically regulated, processive and mRNA cap-interacting deadenylase. Crit Rev Biochem Mol Biol *48*, 192-209.

Viswanathan, S.R., Daley, G.Q., and Gregory, R.I. (2008). Selective blockade of microRNA processing by Lin28. Science *320*, 97-100.

Wakiyama, M., Takimoto, K., Ohara, O., and Yokoyama, S. (2007). Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. Genes & development *21*, 1857-1862.

Wang, B., Yanez, A., and Novina, C.D. (2008a). MicroRNA-repressed mRNAs contain 40S but not 60S components. Proc Natl Acad Sci U S A *105*, 5343-5348.

Wang, L., Zhang, H., Solski, P.A., Hart, M.J., Der, C.J., and Su, L. (2000). Modulation of HIV-1 replication by a novel RhoA effector activity. J Immunol *164*, 5369-5374.

Wang, X., and El Naqa, I.M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics *24*, 325-332.

Wang, X., Ye, L., Hou, W., Zhou, Y., Wang, Y.J., Metzger, D.S., and Ho, W.Z. (2009a). Cellular microRNA expression correlates with susceptibility of monocytes/macrophages to HIV-1 infection. Blood *113*, 671-674.

Wang, X.H., Aliyari, R., Li, W.X., Li, H.W., Kim, K., Carthew, R., Atkinson, P., and Ding, S.W. (2006). RNA interference directs innate immunity against viruses in adult Drosophila. Science *312*, 452-454.

Wang, Y., Juranek, S., Li, H., Sheng, G., Tuschl, T., and Patel, D.J. (2008b). Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. Nature *456*, 921-926.

Wang, Y., Juranek, S., Li, H., Sheng, G., Wardle, G.S., Tuschl, T., and Patel, D.J. (2009b). Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. Nature *461*, 754-761.

Wang, Y., Medvid, R., Melton, C., Jaenisch, R., and Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet *39*, 380-385.

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., and Patel, D.J. (2008c). Structure of the guide-strand-containing argonaute silencing complex. Nature *456*, 209-213.

Wang, Z., Jiao, X., Carr-Schmid, A., and Kiledjian, M. (2002). The hDcp2 protein is a mammalian mRNA decapping enzyme. Proc Natl Acad Sci U S A *99*, 12663-12668.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T.*, et al.* (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature *453*, 539-543.

Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess, J.W., Jr., Swanstrom, R., Burch, C.L., and Weeks, K.M. (2009). Architecture and secondary structure of an entire HIV-1 RNA genome. Nature *460*, 711-716.

Wee, L.M., Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2012). Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. Cell *151*, 1055-1067.

Weill, N., Lisi, V., Scott, N., Dallaire, P., Pelloux, J., and Major, F. (2015). MiRBooking simulates the stoichiometric mode of action of microRNAs. Nucleic Acids Res *43*, 6730-6738.

Westerhout, E.M., ter Brake, O., and Berkhout, B. (2006). The virion-associated incoming HIV-1 RNA genome is not targeted by RNA interference. Retrovirology *3*, 57.

Weyand, N.J., Braaten, B.A., van der Woude, M., Tucker, J., and Low, D.A. (2001). The essential role of the promoter-proximal subunit of CAP in pap phase variation: Lrp- and helical phase-dependent activation of papBA transcription by CAP from -215. Mol Microbiol *39*, 1504-1522.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell *75*, 855-862.

Wilkinson, K.A., Gorelick, R.J., Vasa, S.M., Guex, N., Rein, A., Mathews, D.H., Giddings, M.C., and Weeks, K.M. (2008). High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. PLoS biology *6*, e96.

Willkomm, S., and Restle, T. (2015). Conformational Dynamics of Ago-Mediated Silencing Processes. Int J Mol Sci *16*, 14769-14785.

Willkomm, S., Zander, A., Gust, A., and Grohmann, D. (2015). A prokaryotic twist on argonaute function. Life (Basel) *5*, 538-553.

Witzany, G. (2011). The agents of natural genome editing. J Mol Cell Biol *3*, 181-189.

Wu, E., Thivierge, C., Flamand, M., Mathonnet, G., Vashisht, A.A., Wohlschlegel, J., Fabian, M.R., Sonenberg, N., and Duchaine, T.F. (2010). Pervasive and cooperative deadenylation of 3'UTRs by embryonic microRNA families. Molecular cell *40*, 558-570.

Wu, L., and Belasco, J.G. (2005). Micro-RNA regulation of the mammalian lin-28 gene during neuronal differentiation of embryonal carcinoma cells. Mol Cell Biol *25*, 9198-9208.

Wu, L., Fan, J., and Belasco, J.G. (2006). MicroRNAs direct rapid deadenylation of mRNA. Proc Natl Acad Sci U S A *103*, 4034-4039.

Wu, L., Fan, J., and Belasco, J.G. (2008). Importance of translation and nonnucleolytic ago proteins for on-target RNA interference. Curr Biol *18*, 1327-1332.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.

Xu, N., Chen, C.Y., and Shyu, A.B. (2001). Versatile role for hnRNP D isoforms in the differential regulation of cytoplasmic mRNA turnover. Mol Cell Biol *21*, 6960-6971.

Yamashita, A., Chang, T.C., Yamashita, Y., Zhu, W., Zhong, Z., Chen, C.Y., and Shyu, A.B. (2005). Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. Nat Struct Mol Biol *12*, 1054-1063.

Yan, K.S. (2003). Structure and conserved RNA binding of the PAZ domain. Nature *426*, 468-474.

Yan, X., Chao, T., Tu, K., Zhang, Y., Xie, L., Gong, Y., Yuan, J., Qiang, B., and Peng, X. (2007). Improving the prediction of human microRNA target genes by using ensemble algorithm. FEBS Lett *581*, 1587-1593.

Yang, F., DeBeaumont, R., Zhou, S., and Naar, A.M. (2004). The activator-recruited cofactor/Mediator coactivator subunit ARC92 is a functionally important target of the VP16 transcriptional activator. Proc Natl Acad Sci U S A *101*, 2339-2344.

Yang, Y., Wang, Y.P., and Li, K.B. (2008). MiRTif: a support vector machine-based microRNA target interaction filter. BMC Bioinformatics *9 Suppl 12*, S4.

Ye, W., Lv, Q., Wong, C.K., Hu, S., Fu, C., Hua, Z., Cai, G., Li, G., Yang, B.B., and Zhang, Y. (2008). The effect of central loops in miRNA:MRE duplexes on the efficiency of miRNA-mediated gene regulation. PLoS One *3*, e1719.

Yeom, K.H., Lee, Y., Han, J., Suh, M.R., and Kim, V.N. (2006). Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing. Nucleic Acids Res *34*, 4622-4629.

Yeung, M.L., Houzet, L., Yedavalli, V.S., and Jeang, K.T. (2009). A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. J Biol Chem *284*, 19463-19473.

Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes & development *17*, 3011-3016.

Yigit, E., Batista, P.J., Bei, Y., Pang, K.M., Chen, C.C., Tolia, N.H., Joshua-Tor, L., Mitani, S., Simard, M.J., and Mello, C.C. (2006). Analysis of the C. elegans Argonaute family reveals that distinct Argonautes act sequentially during RNAi. Cell *127*, 747-757.

Yousef, M., Jung, S., Kossenkov, A.V., Showe, L.C., and Showe, M.K. (2007). Naive Bayes for microRNA target predictions--machine learning for microRNA targets. Bioinformatics *23*, 2987-2992.

Yuan, Y.R., Pei, Y., Ma, J.B., Kuryavyi, V., Zhadina, M., Meister, G., Chen, H.Y., Dauter, Z., Tuschl, T., and Patel, D.J. (2005). Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. Molecular cell *19*, 405-419.

Yue, S.B., Trujillo, R.D., Tang, Y., O'Gorman, W.E., and Chen, C.Z. (2011). Loop nucleotides control primary and mature miRNA function in target recognition and repression. RNA Biol *8*, 1115-1123.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell *101*, 25-33.

Zander, A., Holzmeister, P., Klose, D., Tinnefeld, P., and Grohmann, D. (2014). Single-molecule FRET supports the two-state model of Argonaute action. RNA Biol *11*, 45-56.

Zeng, Y., and Cullen, B.R. (2005). Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. J Biol Chem *280*, 27595-27603.

Zeng, Y., Sankala, H., Zhang, X., and Graves, P.R. (2008). Phosphorylation of Argonaute 2 at serine-387 facilitates its localization to processing bodies. Biochem J *413*, 429-436.

Zhang, J., Du, Y.Y., Lin, Y.F., Chen, Y.T., Yang, L., Wang, H.J., and Ma, D. (2008). The cell growth suppressor, mir-126, targets IRS-1. Biochem Biophys Res Commun *377*, 136-140.

Zhang, K., Zhang, X., Cai, Z., Zhou, J., Cao, R., Zhao, Y., Chen, Z., Wang, D., Ruan, W., Zhao, Q.*, et al.* (2018). A novel class of microRNA-recognition elements that function only within open reading frames. Nat Struct Mol Biol *25*, 1019-1027.

Zhang, L., Ding, L., Cheung, T.H., Dong, M.Q., Chen, J., Sewell, A.K., Liu, X., Yates, J.R., 3rd, and Han, M. (2007). Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. Molecular cell *28*, 598-613.

Zhou, H., Huang, C., and Xia, X.G. (2008a). A tightly regulated Pol III promoter for synthesis of miRNA genes in tandem. Biochim Biophys Acta *1779*, 773-779.

Zhou, H., Xia, X.G., and Xu, Z. (2005). An RNA polymerase II construct synthesizes short-hairpin RNA with a quantitative indicator and mediates highly efficient RNAi. Nucleic Acids Res *33*, e62.

Zhou, H., Xu, M., Huang, Q., Gates, A.T., Zhang, X.D., Castle, J.C., Stec, E., Ferrer, M., Strulovici, B., Hazuda, D.J.*, et al.* (2008b). Genome-scale RNAi screen for host factors required for HIV replication. Cell Host Microbe *4*, 495-504.

Zipprich, J.T., Bhattacharyya, S., Mathys, H., and Filipowicz, W. (2009). Importance of the C-terminal domain of the human GW182 protein TNRC6C for translational repression. RNA *15*, 781-793.

THE APPENDICES

APPENDIX A

Figure 24. Efficiently repressed miR-20a targets predicted by *MicroAlign* algorithm.

The 3'UTR sequences were used as input. Potential miR-20 targets that scored 70 or higher by miScore were listed as output. Alignments with the miR-20 5p guide RNA were produced for the top 30 hits.

Gene Name	site position	target sequence	miScore	Alignment with target
RBM12B	1142	CTAGGCACTGTAAGCACTTTA	90	TGGACGTGATATTCGTGAAAT * ||||*||||||||| CTAGGCACTGTAAGCACTTTA
GPR137C	2144	CAAAGCACTAGGAGCACTTTA	84	TGGACGTGATATTCGTGAAAT |||||| *|||||||| CAAAGCACTAGGAGCACTTTA
EPHA7	762	TACATACTATAAGGCACTTTT	83	TGGACGTGATATT-CGTGAAAT | *|||||| |||||| TAC-ATACTATAAGGCACTTTT
PTPRD.2	2180	GCATTGTATTTGAGCACTTTT	82	TG-GACGTGATATTCGTGAAAT *| *||*|*| |*||||||| GCATTGTATT-TGAGCACTTTT
PTPRD.1	2180	GCATTGTATTTGAGCACTTTT	82	TG-GACGTGATATTCGTGAAAT *| *||*|*| |*||||||| GCATTGTATT-TGAGCACTTTT
PTPRD.5	2180	GCATTGTATTTGAGCACTTTT	82	TG-GACGTGATATTCGTGAAAT *| *||*|*| |*||||||| GCATTGTATT-TGAGCACTTTT
PTPRD.3	2180	GCATTGTATTTGAGCACTTTT	82	TG-GACGTGATATTCGTGAAAT *| *||*|*| |*||||||| GCATTGTATT-TGAGCACTTTT
PTPRD.4	2180	GCATTGTATTTGAGCACTTTT	82	TG-GACGTGATATTCGTGAAAT *| *||*|*| |*||||||| GCATTGTATT-TGAGCACTTTT
FDX1.6	2180	GCATTGTATTTGAGCACTTTT	82	TG-GACGTGATATTCGTGAAAT *| *||*|*| |*||||||| GCATTGTATT-TGAGCACTTTT
BNIP2	951	ACTCTCACTATGGGCACTTTA	82	TG-GACGTGATATTCGTGAAAT || || |||||**|||||||| ACTCT-CACTATGGGCACTTTA
CFL2.5	492	CTATGCATTAAAAGCACTTTT	82	TGGACGTGATATTCGTGAAAT * |||*|| |||||||| CTATGCATTAAAAGCACTTTT
CFL2.2	492	CTATGCATTAAAAGCACTTTT	82	TGGACGTGATATTCGTGAAAT * |||*|| |||||||| CTATGCATTAAAAGCACTTTT
CFL2.1	492	CTATGCATTAAAAGCACTTTT	82	TGGACGTGATATTCGTGAAAT * |||*|| |||||||| CTATGCATTAAAAGCACTTTT
PLEKHM1.1	782	ACCAGCACTGTCAGCACTTTG	81	TGGACGTGATATTCGTGAAAT ||| ||||*| |||||||* ACCAGCACTGTCAGCACTTTG

ARPP21.2	700	GAGAGCATTGAGAGCACTTTC	80	TGGACGTGATATTCGTGAAAT * |||*|* *||||||| GAGAGCATTGAGAGCACTTTC
NXPH3	2491	GCAGGCACTGGGGGCACTTTG	80	TGGACGTGATATTCGTGAAAT *| ||||* **||||||* GCAGGCACTGGGGGCACTTTG
ARPP21.3	700	GAGAGCATTGAGAGCACTTTC	80	TGGACGTGATATTCGTGAAAT * |||*|* *||||||| GAGAGCATTGAGAGCACTTTC
ARPP21.4	700	GAGAGCATTGAGAGCACTTTC	80	TGGACGTGATATTCGTGAAAT * |||*|* *||||||| GAGAGCATTGAGAGCACTTTC
SLC40A1	1152	TACGTTGCTATGAGCACTTTC	80	-TGGACGTGATATTCGTGAAAT || | **|||*||||||| TACGT-TGCTATGAGCACTTTC
NTM.1	431	CGTGGCGCTGCGGGCACTTTG	79	TGGACGTGATATTCGTGAAAT * ||*||* **||||||* CGTGGCGCTGCGGGCACTTTG
NTM.2	431	CGTGGCGCTGCGGGCACTTTG	79	TGGACGTGATATTCGTGAAAT * ||*||* **||||||* CGTGGCGCTGCGGGCACTTTG
NTM.3	431	CGTGGCGCTGCGGGCACTTTG	79	TGGACGTGATATTCGTGAAAT * ||*||* **||||||* CGTGGCGCTGCGGGCACTTTG
THAP6	635	CTCCTCACTAGGAGCACTTTG	79	T-GGACGTGATATTCGTGAAAT ||| ||||| *||||||* CTCCT-CACTAGGAGCACTTTG
HDX.2	3201	CATCTATTGTGAGGCACTTTC	78	TGGACGTGATATT-CGTGAAAT | | *|*|*|*| ||||||| -CATCTATTGTGAGGCACTTTC
HDX.3	3201	CATCTATTGTGAGGCACTTTC	78	TGGACGTGATATT-CGTGAAAT | | *|*|*|*| ||||||| -CATCTATTGTGAGGCACTTTC
HDX.1	3201	CATCTATTGTGAGGCACTTTC	78	TGGACGTGATATT-CGTGAAAT | | *|*|*|*| ||||||| -CATCTATTGTGAGGCACTTTC
SLC35D1	1973	TGAGTTCACTTGAGCACTTTC	77	-TGGACGTGATATTCGTGAAAT * | |||| |*||||||| TGAGTTCACT-TGAGCACTTTC
CADM2.3	5738	GCTCAGCACTTAAGCACTTTT	77	TG-GACGTGATATTCGTGAAAT *| | ||||| ||||||||| GCTCAGCACT-TAAGCACTTTT
CADM2.1	5738	GCTCAGCACTTAAGCACTTTT	77	TG-GACGTGATATTCGTGAAAT *| | ||||| ||||||||| GCTCAGCACT-TAAGCACTTTT
CADM2.2	5738	GCTCAGCACTTAAGCACTTTT	77	
ZDHHC20	2596	TCTTCACTATTATGCACTTTC	77	
GPR137C	265	CAAATGCATATGTGCACTTTT	77	
C6orf35	3046	GCTGGCGTTAAGGGCACTTTG	77	
PEX5L	617	GTATTGTATATATGCACTTTA	76	
OXA1L.2	899	TTTAGTTTATAAAGCACTTTC	76	

TUSC3.1	961	TTTAGTTTATAAAGCACTTTC	76
SYNCRIP.5	2190	GCCATGCCTATTGGCACTTTA	76
CORIN	78	GAGCTGTACAGAAGCACTTTT	76
SYNCRIP.2	2190	GCCATGCCTATTGGCACTTTA	76
SYNCRIP.3	2190	GCCATGCCTATTGGCACTTTA	76
SYNCRIP.6	2190	GCCATGCCTATTGGCACTTTA	76
CNGB3	1176	GGTCACTGTAACAGCACTTTG	76
HOOK3	1192	TAATCATTGTAAAGCACTTTG	76
SFR1.2	46	AAAAGATACTTAGGCACTTTT	76
SFR1.1	46	AAAAGATACTTAGGCACTTTT	76
HAUS8.1	13	TCAGGATACTTGAGCACTTTA	76
HAUS8.2	13	TCAGGATACTTGAGCACTTTA	76
VSX1.1	689	TTTGTGATTGAAAGCACTTTA	76
BCL2.alpha	5164	ATTAGCTATAATGGCACTTTG	76
POLR3G	396	TACAGCACGTGGAGCACTTTA	75
PLAC1	83	GACCCTCATGTGAGCACTTTT	75
ASTN1.1	1499	GGCGCTGATGTAAGCACTTTA	75
ADARB1.1	3922	GGCAGCACTGTCTGCACTTTC	75
ADARB1.2	3922	GGCAGCACTGTCTGCACTTTC	75
ADARB1.3	564	GGCAGCACTGTCTGCACTTTC	75
ADARB1.7	564	GGCAGCACTGTCTGCACTTTC	75
EIF2S1	2906	AATTTTTACTTAAGCACTTTG	75
ZZEF1	1320	TCTTCCTATAAGAGCACTTTC	75
PALLD.2	87	CAGTCGCTATGCAGCACTTTC	75
PALLD.1	87	CAGTCGCTATGCAGCACTTTC	75
ANKS1A	1234	GTTCCTCCTGTGGGCACTTTA	75
PACSIN1.1	764	TTTCCAGCTATCAGCACTTTC	75
PACSIN1.2	764	TTTCCAGCTATCAGCACTTTC	75
C9orf82.1	425	GTTTTGCTTATATGCACTTTT	75
C9orf82.2	425	GTTTTGCTTATATGCACTTTT	75
TSEN2.3	373	TACAGTTTATGAAGCACTTTC	75
TSEN2.1	373	TACAGTTTATGAAGCACTTTC	75
TSEN2.2	373	TACAGTTTATGAAGCACTTTC	75
TSEN2.4	373	TACAGTTTATGAAGCACTTTC	75
PDDC1	525	ACCGGCACTGGCAGCACTTTC	75
APP.9	696	CTGTTCATTGTAAGCACTTTT	75
APP.10	696	CTGTTCATTGTAAGCACTTTT	75
APP.6	696	CTGTTCATTGTAAGCACTTTT	75
APP.5	696	CTGTTCATTGTAAGCACTTTT	75

APP.1	696	CTGTTCATTGTAAGCACTTTT	75
APP.2	696	CTGTTCATTGTAAGCACTTTT	75
APP.3	696	CTGTTCATTGTAAGCACTTTT	75
APP.8	696	CTGTTCATTGTAAGCACTTTT	75
APP.4	696	CTGTTCATTGTAAGCACTTTT	75
APP.7	696	CTGTTCATTGTAAGCACTTTT	75
TP53INP2	2492	GTCAGTACTACCAGCACTTTG	75
STK17B	89	TTATATTGTAAATGCACTTTT	74
STK17B	2822	TGAAATTGTAATGGCACTTTA	74
PCYT1B.1	1115	TCTTGTGACTTGGGCACTTTG	74
PCYT1B.2	1115	TCTTGTGACTTGGGCACTTTG	74
PCYT1B.3	907	TCTTGTGACTTGGGCACTTTG	74
PCDHA9.1	1175	GAAACAATTATGTGCACTTTG	74
SYNRG.1	2804	TGGCCATTAATAAGCACTTTT	74
SYNRG.2	2804	TGGCCATTAATAAGCACTTTT	74
SYNRG.3	2804	TGGCCATTAATAAGCACTTTT	74
SYNRG.4	2804	TGGCCATTAATAAGCACTTTT	74
MMACHC	903	AAACACGTGTAAGGCACTTTG	74
SYNRG.5	2804	TGGCCATTAATAAGCACTTTT	74
SYNRG.6	2804	TGGCCATTAATAAGCACTTTT	74
SYNRG.7	2804	TGGCCATTAATAAGCACTTTT	74
PCDHAC1.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA13.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA12.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA11.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA10.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA8.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA7.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA6.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA5.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA4.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA3.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA2.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA1.1	1175	GAAACAATTATGTGCACTTTG	74
PCDHA6.3	1175	GAAACAATTATGTGCACTTTG	74
SACS	589	GAGTGCACTGAGTGCACTTTA	74
PCDHA10.3	1175	GAAACAATTATGTGCACTTTG	74
PCDHA1.3	1175	GAAACAATTATGTGCACTTTG	74
LRRC20.3	1897	CTGATTCCTATAAGCACTTTA	74

LRRC20.2	1897	CTGATTCCTATAAGCACTTTA	74
LRRC20.1	1897	CTGATTCCTATAAGCACTTTA	74
PCDHAC2.1	1175	GAAACAATTATGTGCACTTTG	74
MYO19.1	316	CAATTCCACATAAGCACTTTT	74
MYO19.2	316	CAATTCCACATAAGCACTTTT	74
ANTXR1.3	205	GACCTTACTGGAGGCACTTTA	74
SLC46A3.1	173	CCACGCACTTTGAGCACTTTG	74
GABBR1.1	918	TGCACATTGTTATGCACTTTT	74
GABBR1.2	918	TGCACATTGTTATGCACTTTT	74
GABBR1.3	918	TGCACATTGTTATGCACTTTT	74
GLIS3.1	3449	TTGCTGACATAAAGCACTTTG	74
GLIS3.2	3449	TTGCTGACATAAAGCACTTTG	74
GMFB	2700	AAAATGTGTGTCAGCACTTTT	73
HLF	77	CTTTCTGACATCAGCACTTTA	73
GPC6	1900	TAAGTATATTTGAGCACTTTT	73
TRPS1	4878	ACTTGTTTGTAAAGCACTTTG	73
STYX.1	286	TTATTGCAATAATGCACTTTT	73
STYX.2	286	TTATTGCAATAATGCACTTTT	73
EPM2AIP1	1476	GAAATGTGTAGGGGCACTTTT	73
GOLGA1	1968	TAAGTGATTTAATGCACTTTG	73
TRAF4	827	CCGACACTGCTAAGCACTTTA	73
ATP5O.1	3521	GGATGCATTGCCAGCACTTTG	73
LRP8.4	3521	GGATGCATTGCCAGCACTTTG	73
LRP8.2	3521	GGATGCATTGCCAGCACTTTG	73
LRP8.3	3521	GGATGCATTGCCAGCACTTTG	73
NOS1.1	5048	CATGCCTGTAATAGCACTTTG	73
NOS1.3	5048	CATGCCTGTAATAGCACTTTG	73
NOS1.4	5048	CATGCCTGTAATAGCACTTTG	73
CCDC25	920	TCCACCTGTAATAGCACTTTG	73
ARID4B.3	133	TTGGCACTTAAGTGCACTTTT	73
ZNF275	4902	CCGTTGGTTGTTGGCACTTTT	73
ZFYVE26	1645	GCACTGAACATAAGCACTTTA	73
ASH1L	1767	GGTGGGACTAGGGGCACTTTG	73
PTGES3	1042	GCATATTGTAGATGCACTTTG	73
ARID4B.1	133	TTGGCACTTAAGTGCACTTTT	73
ARID4B.2	133	TTGGCACTTAAGTGCACTTTT	73
AGGF1	1855	AATGTATATAAAAGCACTTTG	73
CYP2U1	2438	GAAATATTACTAAGCACTTTC	73
ICA1L.1	3734	CGCCTATATAACAGCACTTTG	73

SLC30A7.1	5210	AGAACTCCCTTAAGCACTTTT	73
SLC30A7.2	4894	AGAACTCCCTTAAGCACTTTT	73
ATPBD4.2	575	CATAGTCCTATCAGCACTTTG	73
BTBD9.1	3310	GGAGCCGTTGTGAGCACTTTG	73
BTBD9.2	3310	GGAGCCGTTGTGAGCACTTTG	73
BTBD9.3	3310	GGAGCCGTTGTGAGCACTTTG	73
BTBD9.4	3310	GGAGCCGTTGTGAGCACTTTG	73
WHSC1.8	171	TTAGTTTATTTGAGCACTTTT	73
GRHL2	1595	ATTTGTTTGTAAAGCACTTTG	73
REST.1	1486	TGTTCCTATGAGGGCACTTTG	73
REST.2	1486	TGTTCCTATGAGGGCACTTTG	73
NOS1.2	5048	CATGCCTGTAATAGCACTTTG	73
FLT1.1	1229	TGGCGCATATTAAGCACTTTA	73
ZDHHC21	5996	TATAGCTTATTGGGCACTTTA	72
PRRG4	2240	TGGCTTTACATAAGCACTTTT	72
PTP4A1	1447	TTTAATACTAAAAGCACTTTC	72
NFATC4.1	191	CTCAGAGCTAGAAGCACTTTC	72
EPHA5.1	2139	TGATGATTATGTGGCACTTTA	72
EPHA5.2	2139	TGATGATTATGTGGCACTTTA	72
DDX5	115	CAGTAATTATGGTGCACTTTT	72
ZBTB44	119	CAGAATTGTGAAAGCACTTTT	72
ANKFY1.1	778	CTTTTACTGTGCTGCACTTTT	72
CNOT6	3029	CTTGGGATTATTAGCACTTTC	72
HOXB13	322	AGAGCTCTGTAGAGCACTTTA	72
FBXO21.1	2059	AGTTGGTATTTGGGCACTTTA	72
FBXO21.2	2059	AGTTGGTATTTGGGCACTTTA	72
BTBD10	365	GTAATATATAGTTGCACTTTA	72
CEP70	602	ATCTTGATGTAATGCACTTTT	72
CXCL14	145	ATATTGTTATGAAGCACTTTT	72
WEE1.1	466	TATCCCACTGGGAGCACTTTG	72
WEE1.2	466	TATCCCACTGGGAGCACTTTG	72
ZNF783	1303	AGTTCTCCCTTGAGCACTTTG	72
RORC.1	701	AAACCTCTTATGTGCACTTTA	72
RORC.2	701	AAACCTCTTATGTGCACTTTA	72
KLHL31	1249	TACCCACTTGATAGCACTTTT	71
EREG	3634	GCACTCTGTAATTGCACTTTT	71
HS2ST1.1	1353	ATGGCATGTGAAAGCACTTTG	71
ALDH6A1.b	4280	AATGTATATGACAGCACTTTG	71
GABPB1.beta-2	879	TTTCCACATGAAAGCACTTTA	71

GABPB1.beta-1	879	TTTCCACATGAAAGCACTTTA	71
ABL2.c	4280	AATGTATATGACAGCACTTTG	71
ABL2.d	4280	AATGTATATGACAGCACTTTG	71
ABL2.f	4280	AATGTATATGACAGCACTTTG	71
ABL2.g	4280	AATGTATATGACAGCACTTTG	71
ABL2.h	4280	AATGTATATGACAGCACTTTG	71
ABL2.i	4280	AATGTATATGACAGCACTTTG	71
IL17RD	4976	TAAAAATATAATGGCACTTTC	71
NTN4	499	TTCCTTGTATAAAGCACTTTA	71
OR7D2	184	TAGTGAACATAAGGCACTTTT	71
MRPS34.1	4151	GGCAGTTTATTAGGCACTTTT	71
MRPS34.1	5199	GGCACTATAATGGGCACTTTA	71
KATNAL1.1	5047	GAAACTATAAAATGCACTTTT	71
KATNAL1.2	5047	GAAACTATAAAATGCACTTTT	71
ATOH8	3227	GGGCCCTGTGAAAGCACTTTG	71
FAT3	3401	TAATCTGTTGTAGGCACTTTA	71
DCTN5.4	4271	GGCAGTTTATTAGGCACTTTT	71
DCTN5.4	5319	GGCACTATAATGGGCACTTTA	71
TRIM55.2	756	TTTCACGTATTAGGCACTTTA	71
TRIM55.1	644	TTTCACGTATTAGGCACTTTA	71
TRIM55.3	644	TTTCACGTATTAGGCACTTTA	71
TRIM55.4	644	TTTCACGTATTAGGCACTTTA	71
GLO1	1240	TGTACTGCTAGAAGCACTTTA	71
FTSJD1.1	1163	ACATGCATTTTAGGCACTTTT	71
FTSJD1.2	1163	ACATGCATTTTAGGCACTTTT	71
SIN3B	1325	AAGTGTCACACAGGCACTTTG	71
TET2.1	885	TGGTGTTTTAGAAGCACTTTG	71
ZBTB43.2	1761	CGATATATAAAAAGCACTTTG	70
TMEM220	2064	AAGCATATAAAAGGCACTTTT	70
KLHL2.1	84	TCCACTTGTAGCTGCACTTTA	70
ZBTB43.1	1761	CGATATATAAAAAGCACTTTG	70
CRIPT	244	CTCAGTTCTGTATGCACTTTT	70
KLHL2.2	84	TCCACTTGTAGCTGCACTTTA	70
KLHL2.3	84	TCCACTTGTAGCTGCACTTTA	70
KLF9	776	TTGAACATAAGCTGCACTTTT	70
SHOC2	1130	TGCTGAACTAAATGCACTTTT	70
ARL4A.2	766	TTTGTTGTCAGAAGCACTTTC	70
ARL4A.3	766	TTTGTTGTCAGAAGCACTTTC	70
ARL4A.1	766	TTTGTTGTCAGAAGCACTTTC	70

ARL4A.4	766	TTTGTTGTCAGAAGCACTTTC	70
SRPK2.2	679	CATTCTCTATATGGCACTTTA	70
SRPK2.1	679	CATTCTCTATATGGCACTTTA	70
PNPLA4.1	153	ATGGATATAAGAGGCACTTTA	70
PNPLA4.2	153	ATGGATATAAGAGGCACTTTA	70
PNPLA4.3	153	ATGGATATAAGAGGCACTTTA	70
FAIM2	3180	GATTTTTGCATAAGCACTTTG	70
HBXIP	101	TAATGTGCATTAGGCACTTTT	70
FAM203A	145	GTAAGGATTGGAGGCACTTTC	70
PRPF40A	3657	TGGCATATAATAGGCACTTTT	70
PARP1	663	GACTTTCTTATGGGCACTTTT	70
C11orf30	4	TGGACACAATAGTGCACTTTA	70
MAPK1IP1L	4485	CCCCTGGATAAAGGCACTTTC	70
TTPAL.1	578	CAGGACATAAGCAGCACTTTG	70
TTPAL.2	578	CAGGACATAAGCAGCACTTTG	70
DENND5B	5252	ATTGTGATATTAAGCACTTTA	70
RFT1	2239	GCCTGGGCTGTCAGCACTTTG	70
PDGFD.2	2084	TAAAGCCCTATCTGCACTTTT	70
PDGFD.1	2084	TAAAGCCCTATCTGCACTTTT	70
SETD7	3574	GAGACAGTATGTGGCACTTTT	70
PGBD5	1593	AATGGGATTGAAAGCACTTTT	70
C1orf95	5936	GCACTGGCTGGCAGCACTTTT	70
MFSD8	1267	CACTATATAATCTGCACTTTA	70
APOBEC3F.2	119	CTCATGTCTTGGTGCACTTTG	70
PTGS1.1	2395	CTGAGTGACACAAGCACTTTA	70
PTGS1.2	2395	CTGAGTGACACAAGCACTTTA	70
RND3	1462	CTATGTCTTACAAGCACTTTG	70
MMP2.1	501	GTTTGCTTTGTATGCACTTTG	70
MMP2.2	501	GTTTGCTTTGTATGCACTTTG	70
TBC1D8B.1	160	ATGGGCTTTGTTAGCACTTTT	70
TBC1D8B.1	1121	ATGGGCTTTGTTAGCACTTTC	70
PPARGC1B.1	4326	TTAAACAATAAAAGCACTTTG	70
PPARGC1B.2	4326	TTAAACAATAAAAGCACTTTG	70
PPARGC1B.3	4326	TTAAACAATAAAAGCACTTTG	70
IL8	579	AGGACATGTGGAAGCACTTTA	70

APPENDIX B

List of publications:

1. **Yan Y,** Scott N, Imane R, Feghaly A, Gagnon E, Ferbeyre G, and Major F. Efficient small artificial miRNAs against HIV. *Manuscript in preparation,* 2018

2. **Yan Y,** Acevedo M, Mignacca L, Desjardins P, Scott N, Imane R, Queneville J, Robitaille J, Feghaly A, Gagnon E, Ferbeyre G, and Major F, The sequence features that define efficient and specific hAGO2-dependent miRNA silencing guides, *Nucleic Acids Research*, Volume 46, Issue 16, 19 September 2018, Pages 8181-8196, https://doi.org/10.1093/nar/gky546

3. Malina A, Mills JR, Cencic R, **Yan Y**, Fraser J, Schippers LM, Paquet M, Dostie J, Pelletier J., Repurposing CRISPR/Cas9 for In Situ Functional Assays, *Genes and Dev.*, 2013 Dec 1;27(23):2602-14. doi: 10.1101/gad.227132.113

4. Kiethega GN, **Yan Y**, Turcotte M, Burger G., RNA-level unscrambling of fragmented genes in Diplonema mitochondrra., *RNA Biol*., 2013 Feb;10(2):301-13.

5. Burger G, **Yan Y**, Javadi P, Lang B.F., Group I-intron trans-splicing and mRNA editing in the mitochondria of placozoan animals. *Trends Genet*. 2009 Sep;25(9):381-6. Epub 2009 Aug 27

6. Robert F, Williams C, **Yan Y**, Donohue E, Cencic R, Burley S and Pelletier J, Blocking UV-Induced eIF2α Phosphorylation with Small Molecule Inhibitors of GCN2, *Chem Biol Drug Des*. 2009 Jul;74(1):57-67

7. Cencic R, **Yan Y**, and Pelletier J. Homogenous Time Resolved Fluorescence Assay to Identify Modulators of Cap-dependent Translation Initiation. *Comb Chem High Throughput Screen*. (2007) Mar;10(3):181-8.

8. **Yan Y**, Svitkin Y, Lee J, Bisaillon M, and Pelletier J. Ribavirin is not a functional mimic of the 7-methyl guanosine mRNA cap. *RNA* 2005 Aug;11(8):1238-44.

9. Aloyz R, Xu Z, Vanessa Bello, Bergeron J, **Yan Y**, Malaptsa A, Alaoui-Jamali M, Duncan A, and Panasci L. Regulation of cisplatin resistance and homologous recombinational repair by the TFIIH subunit XPD. *Cancer Research* 2002 Oct 1;62(19):5457-62.

NOTES