The hidden rule in argonaute protein loading : sequence specificity

Goh, Eling

2018

Goh, E. (2018). The hidden rule in argonaute protein loading : sequence specificity. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73227>

<https://doi.org/10.32657/10356/73227>

NANYANG
TECHNOLOGICAL
UNIVERSITY

**THE HIDDEN RULE IN ARGONAUTE PROTEIN LOADING:
SEQUENCE SPECIFICITY**

By

ELING GOH

Temasek Life Sciences Laboratory

Nanyang Technological University
(School of Biological Sciences)

Supervisor:

Assistant Professor KATSUTOMO OKAMURA

A thesis submitted to the Nanyang Technological University in fulfilment of the requirement for Doctor of Philosophy of Science (School of Biological Sciences)

2018

ACKNOWLEDGEMENTS

We thank Gregory Hannon for sharing the tagged AGO2 plasmid. K.O conceived the study concept. E.G and K.O designed the experiments. E.G performed all experiments, bioinformatics analysis and thesis writing. K.O contributed in giving valuable advices for the entire study and thesis writing. Research in K.O.'s group was supported by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Fellowship Program (NRF2011NRF-NRFF001-042).

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	1
TABLE OF CONTENTS	2
LIST OF FIGURES	4
LIST OF TABLES	5
ABBREVIATIONS	6
ABSTRACT	8
1. INTRODUCTION	9
1.1 BACKGROUND	9
1.1.1 CELLULAR FUNCTIONS AND BIOGENESIS OF SMALL INTERFERING RNAs	9
1.1.2 CELLULAR FUNCTIONS AND BIOGENESIS OF MICRORNAs	12
1.1.3 SMALL RNA DUPLEX LOADING TO FLY ARGONAUTE PROTEINS	15
1.1.4 SMALL RNA DUPLEX LOADING TO MAMMALIAN ARGONAUTE PROTEINS	17
1.1.5 RULES IN SMALL RNA DUPLEX LOADING AND GUIDE STRAND SELECTION	19
1.1.6 ROLES OF ARGONAUTE DOMAINS IN GUIDE STRAND BINDING	20
1.1.7 CONFORMATIONAL CHANGES DURING mRNA TARGET RECOGNITION BY RISC	23
1.1.8 CURRENT NOTION: STRUCTURE, THE MAIN DETERMINANT IN AGO LOADING	23
1.2 CHALLENGES AND IMPLICATIONS	25
1.2.1 SINGLE-STRANDED RNAs: SEQUENCE SPECIFICITY, A HIDDEN RULE?	25
1.2.2 SINGLE-STRANDED OLIGONUCLEOTIDES IN THERAPEUTIC APPLICATIONS	27
1.3 AIMS	29
1.4 METHODOLOGY	30
1.4.1 HISSA: INDEPENDENT OF UPSTREAM PROCESSING STEPS	30
1.4.2 HIGH-THROUGHPUT SEQUENCING: MASSIVELY PARALLEL ANALYSIS	32
1.4.3 FLY: DISTINCT LOADING PATTERNS BETWEEN TWO AGOs	34
1.4.4 CELL LYSATE: AGO LOADING IN PHYSIOLOGICAL CONTEXT	35
2. MATERIALS AND METHODS	37
2.1 HISSA DEVELOPMENT	37
2.1.1 <i>IN-VITRO</i> LOADING ASSAY	37
2.1.2 IMMUNO-PRECIPITATION, RNA EXTRACTION AND AUTO-RADIOGRAPHY	38
2.1.3 DEGRADATION ASSAY	38
2.1.4 HIGH-THROUGHPUT SEQUENCING LIBRARY PROTOCOL MODIFICATION	39
2.1.5 SEQUENCING ERRORS ELIMINATION SCRIPTING	40
2.1.6 RT-PCR OPTIMIZATION	42
2.1.7 SPLINTED-LIGATION OPTIMIZATION	44
2.1.8 OLIGO-SPECIFIC LIBRARY PROTOCOL VALIDATION	45
2.2 HIDDEN RULE IN AGO LOADING OF SINGLE-STRANDED RNAs	46
2.2.1 HISSA OVERALL SETUP, PARAMETERS AND PROCEDURE	46
2.2.2 COMPUTATIONAL PROCESSING AND ANALYSIS OF HISSA DATA	49
2.2.3 HISSA COMPUTATIONAL DATA VALIDATION WITH <i>IN-VITRO</i> LOADING ASSAY	51
2.3 BIOLOGICAL EVIDENCE FOR HISSA FINDINGS	52

2.3.1	COMPUTATIONAL ANALYSIS OF MRNA FRAGMENTS IN LIBRARIES	52
3.	RESULTS	54
3.1	HISSA DEVELOPMENT	54
3.1.1	5' END OF 20NT <i>MIR-317</i> LOOP SEQUENCE TO BE RANDOMIZED FOR HISSA	55
3.1.2	<i>IN-VITRO</i> LOADING ASSAY REFLECTS RELIABLE AGO LOADING PREFERENCE	57
3.1.3	LIBRARY BIASES AND ERRORS REDUCED FOR DATA ACCURACY	60
3.1.4	OLIGO SEQUENCES ENRICHED WITH ELIMINATION OF ENDOGENOUS RNAs	63
3.2	HIDDEN RULE IN AGO LOADING OF SINGLE-STRANDED RNAs	70
3.2.1	CRITICAL SETUP + EFFECTIVE ANALYSIS PIPELINE = CREDIBLE HISSA	71
3.2.2	SEQUENCE SPECIFICITY IN AGO1 LOADING: 'GAC' IN 1 ST - 5 TH POSITIONS	77
3.2.3	5'GAC MOTIF SUCCESSFULLY VALIDATED WITH <i>IN-VITRO</i> LOADING ASSAY	81
3.3	BIOLOGICAL EVIDENCE FOR HISSA FINDINGS	84
3.3.1	'GAC', THE TOPMOST MOTIF FOUND IN AGO1-LOADED MRNA FRAGMENTS	84
4.	DISCUSSION	88
4.1	IMPORTANCE OF OPTIMIZATION INCREASES WITH RANDOMIZATION	88
4.2	RELATION BETWEEN AGO LOADING SPECIFICITY AND FUNCTION	90
4.3	UNDERSTANDING THE TWO RULES IN LOADING – 5'GAC AND 5'U	92
4.4	ASSESSING CREDIBILITY OF MRNA FRAGMENT ANALYSIS	95
4.5	HISSA APPLICATIONS AND LIMITATIONS	97
4.6	RNAi VERSUS CRISPR: WHICH IS BETTER FOR THERAPEUTICS?	99
5.	CONCLUSION	101
6.	FUTURE WORK	102
7.	REFERENCES	103
8.	APPENDIX	120
8.1	DEGRADATION OF NON-LOADED AGO1 PROTEINS IN CELLS	120
8.2	ORIGINAL BLOTS OF FIGURE 3.2.3	121
9.	PUBLICATIONS AND PRESENTATIONS	122

LIST OF FIGURES

Figure 1.1.1: Biogenesis pathways of endogenous siRNAs	11
Figure 1.1.2: Biogenesis pathways of endogenous miRNAs	14
Figure 1.1.3: Fly small RNA sorting, loading and guide strand selection	16
Figure 1.1.4: Mammalian small RNA loading and guide strand selection	18
Figure 1.1.6: Structure of human AGO2 with bound guide strand	22
Figure 1.4.1: Overview of HISSA	31
Figure 1.4.2: HTS library construction procedure	33
Figure 3.1: Conundrums addressed during HISSA development	54
Figure 3.1.1: 5' end of 20nt <i>mir-317</i> loop alluded to sequence specificity	56
Figure 3.1.2a: <i>In-vitro</i> AGO loading data is highly reproducible	58
Figure 3.1.2b: <i>In-vitro</i> AGO loading not affected by degradation rates	59
Figure 3.1.3a: Biases in HTS libraries reduced with modified protocol	61
Figure 3.1.3b: PCR/sequencing errors elimination pipeline	63
Figure 3.1.4a: Strategy for examining library specificity	66
Figure 3.1.4b: Oligo-specific RT-PCR primers improved library specificity	67
Figure 3.1.4c: Oligo-specific bridge conferred 3' linker ligation specificity	68
Figure 3.1.4d: Oligo-specific RT-PCR and splinted-ligation combined	69
Figure 3.2: Optimal parameters and conditions for HISSA	70
Figure 3.2.1a: HISSA overall experimental setup	74
Figure 3.2.1b: HISSA computational analysis pipelines	75
Figure 3.2.1c: 12NN HISSA results consistent with previous findings	76
Figure 3.2.2a: 5NN HISSA - 'GAC' motif preferred in AGO1 loading	79
Figure 3.2.2b: 'GAC' motif statistically enriched in the 1 st – 5 th positions	80
Figure 3.2.3: 5'GAC motif successfully validated by <i>in-vitro</i> loading	83
Figure 3.3.1a: Endogenous mRNA fragments analysis pipeline	86
Figure 3.3.1b: 'GAC' motif validated in AGO1-bound mRNA fragments	87
Figure 4.3: Postulated models for 5'GAC-RNA loading to AGO1	94
Figure 8.1: AGO1 binding mutant, F2V2-AGO1, degraded in cells	120
Figure 8.2: Original blots of Figure 3.2.3	121

LIST OF TABLES

Table 2.1.1: Oligos used for HISSA parameters selection.....	37
Table 2.1.4: Oligos used for HTS library construction	40
Table 2.1.6: Oligos tested for RT-PCR specificity.....	43
Table 2.1.7: Bridges tested for splinted-ligation specificity	44
Table 2.2.1a: HISSA setup with <i>in-vitro</i> loading and IP parameters	47
Table 2.2.1b: Summary of oligos used in all 3 sets of HISSA experiments ...	48
Table 2.2.3: Oligos for validating AGO loading trends shown by HISSA	51
Table 2.3.1: Libraries analyzed for AGO1-loaded mRNA fragments	53

ABBREVIATIONS

32P: Phosphorus-32

3'UTR: 3' untranslated region

5/8/12NN HISSA: 5/8/12nt randomized HISSA

A: Adenine

AGO: Argonaute

C: Cytosine

C.elegans: *Caenorhabditis elegans*

***cis*-NAT**: *cis*-natural antisense transcript

Cpm: Counts per minute

CRISPR: clustered regularly interspaced short palindromic repeats

DGCR8: DiGeorge syndrome critical region gene 8

G: Guanine

HEN1: HUA ENHANCER 1

HISSA: High-throughput Sequencing mediated Specificity Analysis

hpRNA: hairpin-RNA

HTS: high-throughput sequencing

IP: immuno-precipitation

L1/2: linker 1/2

loqs: Loquacious

MID domain: middle domain

miRISC: miRNA-RISC

miRNA: microRNA

mRNA: messenger RNA

N domain: N-terminus domain

PACT: protein kinase R activating protein

PAR-CLIP: Photoactivatable Ribonucleoside-Enhanced Crosslinking and
Immunoprecipitation

PAZ domain: PIWI/Argonaute/Zwille domain

PCR: polymerase chain reaction

piRNA: piwi-interacting RNA

PIWI: P-element induced Wimpy testis

pre-miRNA: precursor-miRNA

pri-miRNA: primary miRNA transcript

R2D2: protein that contains 2 double-stranded RNA-binding domains (R2) and is associated with dicer-2 (D2)

RISC: RNA-induced silencing complex

RNAi: RNA interference

RNase: ribonuclease

RT: reverse transcription

sgRNA: single guide RNA

siRNA: small interfering RNA

TE: transposable element

***trans*-NAT:** *trans*-natural antisense transcript

TRBP: TAR RNA binding protein

U: Uracil

XPO5: Exportin 5

ABSTRACT

Argonaute (AGO) proteins play an important role in gene regulation with small RNAs serving as guides to targets. The small RNA duplex structure and 5' terminal nucleotide are the only known determinants of AGO loading efficacy, and AGOs are believed to load small RNAs independently of their sequences. Challenging this notion, we developed *High-throughput Sequencing mediated Specificity Analysis (HISSA)*, a method that allows massively parallel analysis of AGO loading by using randomized single-stranded oligos in *in-vitro* loading assays and quantifying AGO-loaded RNA sequences with deep-sequencing. Intriguingly, HISSA divulged sequence specificity as the hidden rule governing fly AGO1 loading when 5' Guanine-Adenine-Cytosine (GAC) motif was strongly preferred over the classic 5' Uridine (U). Trends of 'GAC' motif enrichment in the 5' ends of AGO1-loaded messenger RNA (mRNA) fragments are in line with our HISSA results. Our findings together with HISSA may have implications in therapeutic single-stranded RNAi designs.

1. INTRODUCTION

1.1 Background

From the 1990s to early 2000s, two classes of non-coding double-stranded small RNAs (~18nt – 30nt), small interfering RNAs (siRNAs) and microRNAs (miRNAs), were discovered (Section 1.1.1). Soon after their cardinal gene silencing functions were uncovered, their biogenesis pathways were further investigated (Section 1.1.1 and Section 1.1.2). In addition, much efforts were invested in understanding what effector proteins these small RNAs are loaded to (Section 1.1.3 and Section 1.1.4) and how they are loaded (Section 1.1.5). These loading and silencing mechanisms were further examined in detail with advanced structural studies (Section 1.1.6 and Section 1.1.7). After 2 decades of intensive research, the rules governing the loading of small RNAs to effector Argonaute (AGO) protein are finally decoded (Section 1.1.8).

1.1.1 Cellular functions and biogenesis of small interfering RNAs

The discovery of animal small interfering RNAs (siRNAs) stems back to the gene-silencing phenomenon observed upon injecting exogenous single-stranded RNA transcripts into *Caenorhabditis elegans* (*C. elegans*). This RNA interference (RNAi) was initially thought to be a result of translational repression (Fire et al., 1991; Guo and Kemphues, 1995; Rocheleau et al., 1997). However, Fire and co-workers showed that long double-stranded RNAs were much more potent than single-stranded RNA transcripts in down regulating endogenous genes in a highly specific manner, suggesting that other unknown mechanism is driving RNAi (Fire et al., 1998). A year later, small double-stranded RNAs were shown to be responsible for RNAi in plants (Hamilton and Baulcombe, 1999). Subsequently, studies demonstrated that exogenous long double-stranded RNAs were actually cleaved into small double-stranded RNAs known as siRNAs in animals (Hammond et al., 2000; Zamore et al., 2000; Elbashir et al., 2001a). Ensuing this groundbreaking discovery, endogenous siRNAs were subsequently identified and found to repress transposons via the RNAi pathway, hence aid in protecting the genome integrity in flies and mammals (Svoboda et

al., 2004; Yang and Kazazian, 2006; Murchison et al., 2007; Chung et al., 2008; Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008; Tam et al., 2008; Watanabe et al., 2008).

Endogenous siRNA precursors are mainly repeat-associated transcripts derived from transposable elements (TE) (Watanabe et al., 2006; Chung et al., 2008; Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008; Tam et al., 2008; Watanabe et al., 2008). However, other sources include *cis*-natural antisense transcripts (*cis*-NATs), *trans*-natural antisense transcripts (*trans*-NATs) and hairpin-RNA (hpRNA) transcripts (Ruby et al., 2007; Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008; Okamura et al., 2008a; Okamura et al., 2008b; Tam et al., 2008; Watanabe et al., 2008). In all cases, long double-stranded RNAs having (near) full complementary between the two strands are formed (**Figure 1.1.1**).

Due to their long length, they are cleaved efficiently into ~21nt siRNA duplexes with 2nt 3' overhangs by fly Dicer-2, a class 3 ribonuclease III (RNaseIII), with the aid of Loquacious-PD (loqs-PD) (Bernstein et al., 2001; Lee et al., 2004; Hartig et al., 2009; Zhou et al., 2009; Miyoshi et al., 2010). In mammals, Dicer is involved in the cleavage (Zhang et al., 2002; Provost et al., 2002; Myers et al., 2003). It remains unclear if TAR RNA binding protein (TRBP) enhances the affinity of Dicer towards these long double-stranded RNAs (Chakravarthy et al., 2010; Lee et al., 2013) (**Figure 1.1.1**).

Figure 1.1.1: Biogenesis pathways of endogenous siRNAs

Long double-stranded RNAs are formed when two highly complementary transcripts annealed with each other. These transcripts can be transposons (TE-siRNAs), pairs of sense (red) and anti-sense (blue) transcripts from the same gene loci (*cis*-NAT-siRNAs) or different gene loci (*trans*-NAT-siRNAs). In some events, a single transcript can fold into a hairpin structure due to extensive base-pairing between two regions within (hpRNAs). Whichever the case, fly Dicer-2 with the aid of Loqs PD/ mammalian Dicer will only cleave the double-stranded regions (between dashed lines) and the resulting small duplexes are known as siRNAs.

1.1.2 Cellular functions and biogenesis of microRNAs

In discovering microRNAs (miRNAs), *lin-4* gene was first observed to produce non-coding small RNA but not functional protein despite its importance in repressing *lin-14* gene in *C.elegans* (Lee et al., 1993; Wightman et al., 1993). This small RNA was shown to target 3' untranslated region (3'UTR) of *lin-14* at multiple sites of antisense-complementarity, thus affecting the *C. elegans* larval development (Lee et al., 1993; Wightman et al., 1993). Eight years later, studies reported the abundance of these endogenous small RNAs and named them miRNAs (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Consequently, gene regulatory functions of miRNAs in flies and mammals were uncovered. Few of the earliest studies demonstrated that these miRNA-regulated genes are involved in a broad range of cellular processes - cell growth, apoptosis, fat metabolism and neuronal development in flies (Lai, 2002; Brennecke et al., 2003; Xu et al., 2003; Stark et al., 2003), hematopoietic differentiation, neuronal development and cell proliferation in mammals (Lewis et al., 2003; Chen et al., 2004).

Endogenous sources of miRNAs, also known as primary miRNA transcripts (pri-miRNA), can be messenger RNA (mRNA) transcripts or non-coding mono/polycistronic transcripts (Lee et al., 2002; Ramalingam et al., 2014; Chang et al., 2015) (**Figure 1.1.2**). Precursor-miRNA (pre-miRNA) hairpins that reside within these transcripts are being cropped by the microprocessor, constituting of drosha, a class 2 RNaseIII enzyme, and pasha (in invertebrates) or DiGeorge syndrome critical region gene 8 (DGCR8) (in vertebrates). This results in the formation of pre-miRNAs that are ~65 nt long (Lee et al., 2003; Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Martin et al., 2009; Nguyen et al., 2015; Herbert et al., 2016;) (**Figure 1.1.2**). It is believed that the resultant pre-miRNAs are exported out of the nucleus into cytoplasm by Exportin 5 (XPO5) (Yi et al., 2003; Bohnsack et al., 2004). However, it was recently reported that there may be other pathways mediating nuclear export (Kim et al., 2016).

In the cytoplasm, fly Dicer-1 and mammalian Dicer cleave the pre-miRNA hairpins to produce miRNA duplexes of ~23nt with 2nt 3' overhangs and release

single-stranded miRNA loops as by-products. These miRNA duplexes, unlike siRNA duplexes, contain more and larger bulges (Ha and Kim, 2014). Loqs-PB acts as an enhancer for Dicer-1 processing of a small miRNA subset in flies (Forstemann et al., 2005; Jiang et al., 2005; Saito et al., 2005; Lim et al., 2016). Previously, some studies suggested that TRBP/PACT facilitates mammalian Dicer to recognize pre-miRNA hairpins (Chendrimada et al., 2005; Haase et al., 2005; Lee et al., 2006; Fareh et al., 2016). Yet, it was later shown that TRBP/PACT has only a minor role in determining miRNA duplex length, not in facilitating the recognition of pre-miRNA hairpins (Fukunaga et al., 2012; Lee and Doudna, 2012; Kim et al., 2014; Wilson et al., 2015). Therefore, it is still unclear whether TRBP is crucial for mammalian Dicer activity, especially when tissue-specificity may come into play too (Ding and Wang, 2015) (**Figure 1.1.2**).

Figure 1.1.2: Biogenesis pathways of endogenous miRNAs

Pre-miRNA hairpins can be formed within the exons (exonic hairpins) and introns (intronic hairpins) of mRNA transcripts. They also exist as clusters in non-coding polycistronic transcripts (polycistronic hairpins). Fly Drosha-Pasha/mammalian Drosha-DGCR8 complex cleaves the pre-miRNA hairpins off these pri-miRNA transcripts and fly Dicer-1/mammalian Dicer dices the pre-miRNA hairpins into small duplexes with bulges known as miRNAs. Single-stranded miRNA loops are released as by-products in the process of dicing (black curly lines). Red pentagon: transcription termination signal.

1.1.3 Small RNA duplex loading to fly Argonaute proteins

After the synthesis of these siRNA and miRNA duplexes, siRNAs are sorted to fly AGO2 while miRNAs to fly AGO1. Subsequently, upon selection of the guide strand from the loaded duplex, mature RNA-induced silencing complex (RISC) is formed. RISC is the effector of RNAi and miRNA-mediated gene regulation mentioned in Section 1.1.1 and Section 1.1.2.

In flies, Dicer-2 and R2D2 (protein that contains 2 double-stranded RNA-binding domains (R2) and is associated with dicer-2 (D2)) are essential in carrying out the loading of siRNA duplexes to AGO2 (Liu et al., 2003; Pham et al., 2004 ; Liu et al., 2006; Tomari et al., 2007). After loading, AGO2 slices the passenger strand of the loaded siRNA duplex and releases it, thus forming the mature RISC (Matranga et al., 2005; Miyoshi et al., 2005; Rand et al., 2005; Kim et al., 2007; Liu et al., 2009). The guide strand retained in the mature RISC has its 3' terminal nucleotide being 2'O-methylated by HUA ENHANCER 1 (HEN1) methyltransferase (Horwich et al., 2007). This is followed by the hybridization of guide strand to the fully-complementary target and triggers RISC target cleavage (**Figure 1.1.3**). The cleavage of the target mRNA occurs at position 10th and 11th with reference to the guide strand, thus down regulating the protein expression (Elbashir et al., 2001b, Martinez et al., 2002).

Unlike siRNAs, fly miRNA duplexes are mostly sorted to AGO1 (Forstemann et al., 2007; Miyoshi et al., 2009; Ghildiyal et al., 2010). Few studies suggested that AGO1 protein unwinds the loaded miRNA duplex and releases the passenger strand while retaining the guide strand (Kwak and Tomari, 2012; Hur et al., 2013). However, the factors involved in this unwinding are unknown. The resultant miRISC carries out translational repression and mRNA destabilization by base-pairing 2nd - 8th positions (seed region) of the guide strand with the target (Brennecke et al., 2005; Bartel, 2009) (**Figure 1.1.3**). However, on a side note, fly AGO1 is also able to carry out target cleavage but the biological significance of its slicing activity has yet to be thoroughly characterized (Okamura et al., 2004; Miyoshi et al., 2005).

Small RNA sorting, loading and guide strand selection in fly

Figure 1.1.3: Fly small RNA sorting, loading and guide strand selection

In fly, siRNAs tend to have no mismatch in the central positions ('9 10') and get recognized by Dicer-2 and R2D2 efficiently (first siRNA duplex from left). Dicer-2 and R2D2 then load the siRNA duplex into AGO2. Since both of the strands (red and blue) have no mismatch in the central positions, guide strand selection abides the asymmetrical thermodynamic stability rule – the red strand has a bulge very near to its 5' end so it would be selected as the guide strand. However, when the bulge occurs very near the center (second siRNA duplex from the left), it would be less strongly bound by Dicer-2 and R2D2 and so would be the loading to AGO2 (dashed line). The red strand is selected as the guide strand because the region near and at 9th and 10th positions is associated with more double-stranded characteristic as compared to the blue strand. miRNAs usually carry mismatches in the center and get loaded to AGO1 instead (first and second miRNA duplexes on the right). Regardless which strand has bulge in the center, guide strand selection is merely based on asymmetrical thermodynamic stability rule. Therefore, in both of the cases, the blue strand is selected as the guide strand. Guide strands in AGO1 usually possess 5'U.

1.1.4 Small RNA duplex loading to mammalian Argonaute proteins

Instead of just 2 AGO proteins, mammals have 4 AGO proteins (AGO1-4) involved in the siRNA and miRNA pathways. It is crucial to note that unlike the fly system, there is no distinct sorting of siRNAs and miRNAs to specific AGOs in mammals (Meister et al., 2004; Yoda et al., 2010 Dueck et al., 2012) (**Figure 1.1.4**). It remains controversial what particular proteins are involved in the loading of siRNA and miRNA duplexes to AGO1,2,3 and 4. While some studies suggested TRBP/PACT and Dicer to be the main factors responsible for the loading, some other studies demonstrated that either of them might be dispensable (Chendrimada et al., 2005; Haase et al., 2005; Lee et al., 2006; Gregory et al., 2005; Maniataki and Mourelatos, 2005; MacRae et al., 2008; Wang et al., 2009; Sakurai et al., 2011).

Upon loading of the small RNA duplexes, human AGO2 cleaves the passenger strand of the duplex whereas AGO1,3 and 4 unwind the duplex and eject the passenger strand (Matranga et al., 2005; Leuschner et al., 2006; Yoda et al., 2010). AGO2-RISC cleaves gene targets that form at least 9 base pairs with the guide strand (spanning across the 9th – 15th positions of guide strand) while AGO1-4 can take part in mRNA destabilization via the seed region of guide strand (Martinez et al., 2002; Martinez and Tuschl, 2004; Meister et al., 2004; Liu et al., 2005, Ameres et al., 2007; Bartel, 2009; Shin et al., 2010). Recently, AGO3 is reported to have the ability to cleave target too. However, this slicing action seems to depend on the guide strand sequence but it is still unclear what other feature(s) of the guide strand activates the slicing (Park et al., 2017) (**Figure 1.1.4**).

Figure 1.1.4: Mammalian small RNA loading and guide strand selection

In mammal, the distinct sorting of siRNAs and miRNAs into different AGOs is absent (unlike in fly). Regardless whether the duplexes have central mismatches, they are loaded to AGO1,2,3 and 4. Just like the case of fly AGO1, all of these mammalian AGOs select guide strands based on asymmetrical thermodynamic stability and 5' terminal nucleotide identity.

1.1.5 Rules in small RNA duplex loading and guide strand selection

It is interesting that siRNAs are distinctly loaded to AGO2 while miRNAs are mostly sorted to AGO1 in flies (**Section 1.1.3**). Yet, in mammals, this distinct sorting failed to occur (**Section 1.1.4**). Moreover, in both flies and mammals, AGOs exhibit guide strand biasness by preferentially retaining one strand over the other in the duplex (Khvorova et al., 2003; Schwarz et al., 2003).

Such compelling phenomenon triggered a series of studies on the rules governing the distinct loading of duplexes and guide strand selection. Eventually, 3 rules underlying the duplex sorting and guide strand selection were successfully identified: 1) central mismatch in duplex, 2) thermodynamic stabilities and 3) 5' terminal nucleotide identity.

Central mismatch is important in both sorting and guide strand selection for fly AGO2 but not in any of the mammalian AGOs. siRNA or miRNA duplex that contains no or mild degree of central mismatch (central positions: 9th and 10th positions), can be more efficiently bound by Dicer-2 and R2D2 heterodimer and therefore get loaded to fly AGO2. Those duplexes that are not AGO2 loaded would be directed to AGO1 (Forstemann et al., 2007; Tomari et al., 2007; Okamura et al., 2011). After loading to AGO2, the strand in the duplex that has no mismatch in the 9th and 10th positions would be selected as the guide strand while the other strand would be the passenger strand (Tomari et al., 2007; Okamura et al., 2009) (**Figure 1.1.3**).

On the other hand, asymmetrical thermodynamic stability is highly crucial for guide strand selection in all fly and mammalian AGOs. The 5' end that is characterized with less extensive base-pairing, is deemed as less stable and thus selected as the guide strand (Tomari et al., 2004; Kawamata et al., 2009; Okamura et al., 2009; Noland et al., 2011, Noland and Doudna, 2013) (**Figure 1.1.3 and Figure 1.1.4**).

Besides that, it is reported that uracil (U) and adenine (A) are the most preferred 5' terminal nucleotides for fly AGO1 and human AGOs (Frank et al., 2010; Ghildiyal et al., 2010) (**Figure 1.1.3 and Figure 1.1.4**).

1.1.6 Roles of Argonaute domains in guide strand binding

Though biochemical assays could affirm the above-mentioned characteristics of AGOs, it is difficult to understand fully how AGO functions as RISC without visualizing the interactions between the guide strand and respective protein domains.

As depicted in **Figure 1.1.6**, AGO has 4 main domains, namely N-terminus (N), middle (MID), P-element induced Wimpy testis (PIWI), and PIWI/Argonaute/Zwille (PAZ). Revealed by several crystal structures, the PIWI domain lies in between MID and N domains with PAZ domain residing above it. These domains are in turn, connected together by β -sheet linkers 1 and 2 (L1 and L2) (Song et al., 2004; Yuan et al., 2005; Wang et al., 2008b; Elkayam et al., 2012; Nakanishi et al., 2012; Schirle and MacRae, 2012).

The MID domain is known to anchor the 5' terminal nucleotide (position 1) of the loaded siRNA / miRNA guide strand, whereby the 5' phosphate group, nucleobase (usually U or A) and the sugar of the terminal nucleotide form substantial hydrogen bonds with the various amino acids within the binding pocket (Frank et al., 2010) (**Figure 1.1.6**). It is suggested that the terminal 5' phosphate ultimately serves as a crucial factor in determining the overall positioning of the whole guide RNA strand within the AGO protein because it is where the most intensive interactions could be found in the whole complex (Elkayam et al., 2012). The PAZ domain despite playing a role in binding to the 3' end (last two nucleotides) of the guide strand, does not form half as many interactions as that found between 5' end and MID domain (Song et al., 2003; Wang et al., 2008b, Elkayam et al., 2012; Nakanishi et al., 2012; Schirle and MacRae, 2012) (**Figure 1.1.6**).

The seed region (position 2 to 8) of the guide strand forms hydrogen bonds with L2 and PIWI domain mainly with their phosphate backbone but not their bases. In fact, the bases at 2nd to 5th positions point outwards to the RNA binding groove, which explains why the seed region of miRNA is available for targeting. However, a kink exists between position 6 and 7 due to intercalation by L2 Helix-7, causing 6th to 8th bases of the seed region to be blocked for initial target recognition (Kuhn and Joshua-Tor, 2013).

Positions 9 and 10 of the guide strand, instead of positions 10 and 11 (where cleavage occur at), lie opposite the PIWI active site before target recognition (Elkayam et al., 2012). The PIWI domain forms an RNase H-like folding and its catalytic triad (Asp-Asp-Glu/His) is conserved among AGO proteins that possess target RNA cleavage activity, including human AGO2 and fly AGO1/2 (Song et al., 2004; Rivas et al., 2005; Nakanishi et al., 2012). On top of this conserved triad, a swap for human AGO2 N domain is essential to activate catalytic properties of human AGO1/3/4 (Faehnle et al., 2013; Hauptmann et al., 2013; Nakanishi et al., 2013; Schurmann et al., 2013; Hauptmann et al., 2014).

N domain involvement in the catalytic activity is not unexpected because from the 12th position onwards, all the way before the 3' end, the nucleotides are embedded in the narrow channel formed between N and PAZ domains (**Figure 1.1.6**). Importantly, even though the bases residing in this channel are facing inwards and not exposed to solvent, majority of the contacts made with the protein are via the sugar-phosphate backbone and not the bases (Elkayam et al., 2012; Schirle and MacRae, 2012; Schirle et al., 2014).

Figure 1.1.6: Structure of human AGO2 with bound guide strand

AGO has 4 main domains – MID (purple), PIWI (yellow), N (pink) and PAZ (cyan). These 4 domains are held together by L1 (green) and L2 (red). The 5' end and 3' end (white arrows) of the guide strand are bound to the MID and PAZ domains respectively. PIWI domain lies in the middle of MID and N domains. N and PAZ domains form a narrow channel in which the 3' half of the guide strand resides. White dashed line approximately represents the missing backbone of the guide strand in the structure. This figure is generated using the structure deposited by Elkayam et al., 2012 (Protein Data Bank; ID code 4F3T).

1.1.7 Conformational changes during mRNA target recognition by RISC

During mRNA target recognition by the guide strand-bound RISC, base-pairing between the target and positions 2 to 4 of the guide strand will first occur (Chandradoss et al., 2015). This in turn, triggers the shifting of L2 Helix-7, releasing the steric constraints on nucleotide 6 and 7, thereby facilitating the subsequent base-pairing of entire seed region with the target (Schirle et al., 2014).

As a consequence of the L2 movement, the N-PAZ channel is widened. This results in the re-arrangement of 11th to 17th nucleotides into an A-form helix, accompanied by the splaying of bases 13 to 16 outwards for supplementary base-pairing with the target (Grimson et al., 2007; Wang et al., 2008a; Wee et al., 2012; Schirle et al., 2014). The above mentioned two-steps mechanism is sufficient for miRISC to carry out its regulatory function for the majority of natural targets as full complementary base-pairing between target and guide strand need not be fulfilled (Section 1.1.3 and Section 1.1.4).

On the other hand, more extensive conformational changes are required for fly and human AGO2-RISC for target cleavage because additional base-pairing must occur from position 9 to 12 to attain full complementarity with the target (Section 1.1.3 and Section 1.1.4). It has been proposed that freeing of guide strand 3' end from the N-PAZ lobe is essential as this would allow the complex to overcome topology constraints. With a free 3' end, the central cleft could then be expanded to accommodate the target just beyond the seed region, zipping the target and guide strand up till the 16th position (Y. Wang et al., 2009; Hur et al., 2013).

1.1.8 Current notion: Structure, the main determinant in AGO loading

As perceivable in the preceding sections, nucleobases of small RNAs are usually having little or no contact with the AGO protein, no matter in the duplex loading (Section 1.1.3 – Section 1.1.5), guide strand binding (Section 1.1.6) or mRNA targeting step (Section 1.1.7). At any one point of time, nucleobases are

being occupied as base pairs in the duplex forms and/or avoided of contacts with AGO residues as a result of the backbone geometry. Hence, it is not at all surprising if there is no sequence motif found associated with AGO at any positions of the guide strand other than the well-characterized 5' terminal nucleotide (Section 1.1.5).

Indeed, studies have failed to detect any prevalent sequence motif in the AGO-bound RNA population in both mammals and flies as most of these data merely revolve around general skew towards particular nucleobases. As such, Guanine(G)-rich sequences are reported to be greatly enhanced in both mammalian and fly Ago2 (Leung et al., 2011; Taliaferro et al., 2013). The observations for mature miRNAs that are loaded to human AGOs are as follows: 1) 'U's in the first and last positions, 2) 'A's in 2nd and 3rd positions and 3) 'G's for all other positions. Di-nucleotide (GU/AG) repeats are common but no tri-nucleotides (or longer) motif could be successfully identified. In the case of flies, only 5' U is strongly preferred by AGO1 with no additional characteristics (Hu et al., 2009; Wang, 2013; Rolle et al., 2016; Frank et al., 2010; Ghildiyal et al., 2010)

On the grounds of the above reasoning and evidences, the current consensus for the main determinant in AGO loading falls on the structure (central base pairs/mismatches) of small RNA duplexes and relative thermodynamic stability between two strands as described in Section 1.1.5. All in all, findings to date give credence to the de facto theory of no significant sequence specificity but just structural determinants involved in AGO loading and guide strand selection.

1.2 Challenges and implications

1.2.1 Single-stranded RNAs: Sequence specificity, a hidden rule?

But how confident are we in concluding that only structure plays a role in RNA loading to AGO?

This question comes in especially relevant when two groups discovered that the entire endogenous single-stranded miRNA loop can be loaded to AGO proteins in both fly and mouse (Okamura et al., 2013; Winter et al., 2013). Notably, only selective loops are found to be enriched in immuno-precipitated AGO complexes (e.g. *mir-34 loop* and *mir 317 loop* in fly; *mir-138* and *mir-615 loop* in mouse). In Okamura et al., 2013, oligos harboring *mir-34 loop* and *mir-317 loop* sequences were loaded to fly AGO proteins in cell lysate and the ratios of AGO2:AGO1 loading were highly similar to that observed for *mir-34* and *mir-317* hairpins processing in cells. This suggested that loading of these single-stranded miRNA loops was independent of hairpin processing.

Being single-stranded, these miRNA loops could not possibly abide the structural criteria like central base pairing and asymmetrical thermodynamic stability as observed in miRNA / siRNA duplexes due to the absence of partner strand. Therefore, it is reasonable to hypothesize that sequence specificity acts as a hidden rule in AGO loading of single-stranded RNAs.

This hypothesis is further encouraged by a finding published a year later (Li et al., 2014). In that particular study, human AGO Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) data from Hafner et al., 2010 was re-analyzed. AGO mRNA binding sites were predicted by comparing mRNA fragments bound by AGO and mRNA sequences that were randomly sampled by computational means. It was evident that in the absence of bound miRNA guide strand, AGO exhibits preference for ACCAA motif. In other words, miRNA guide strands tend to mask AGO bona fide sequence preference.

With *in-vitro* cleavage assays, Park et al., 2017 has recently demonstrated that the seed sequence of the guide strand affected the target-slicing efficiencies of human AGO2 and AGO3. Interestingly, upon swapping the 2nd to 4th positions of let-7a seed region to that of mir-20a, the number of cleaved products decreased ~3-folds for AGO2 while increased ~10-folds for AGO3. This finding is inconsistent with the notion that the guide strand interacts with AGO mainly with its phosphate backbone but not its bases (Section 1.1.6). This is because if the notion holds, the slicing efficiency should not differ significantly between different seed sequences especially when the number of base pairs between the guide strand and the target did not change. In that study, the amount of AGO-loaded let-7a-mir20a chimeric single-stranded oligos was not taken into account for the cleavage assays. Therefore, it is highly possible that the seed sequence plays a role in determining the AGO loading efficiency and hence affecting the slicing efficacy.

One possible explanation for the absence of sequence motif in AGO-loaded siRNA/miRNA guide strands (Section 1.1.8) could be our misconceptions about passenger strands. What we thought as 'passenger strands' might actually be loaded to AGOs with similar or better affinities as compared to the guide strands but are found to be poorly loaded because of external factors. For example, if guide strand selection is not based on asymmetrical thermodynamic stability within the RNA duplex or other structural criteria mentioned in Section 1.1.5, there might be a chance that passenger strand is selected over the guide strand solely based on its sequence. In other words, the duplex structure could be a barrier masking away sequence specificity in AGO loading.

On top of the duplex structure, difference in turnover rate between the passenger strand and guide strand could be another factor that lead to our misconceptions about passenger strand. Passenger strand might have stronger binding to AGO than the guide strand during loading. However, after loading, the passenger strand might be released from RISC much more rapidly, causing the net amount found in AGO pulled-down to be much less than the guide strand. Thus far, no study has done extensive sequence analysis on the miRNA passenger strands to conclude presence or absence of motifs.

1.2.2 Single-stranded oligonucleotides in therapeutic applications

If the above-mentioned hypothesis of AGO proteins having sequence specificity towards single-stranded RNAs is true, it could mean great news.

RNAi-based therapy, which involved the delivery of exogenous double-stranded siRNAs to diseased organ, has always been of strong interest to many. Unlike chemical compounds, it regulates disease-related genes at the transcript level instead of protein level, thus providing a higher level of target specificity. Moreover, the screening for optimal double stranded siRNAs is much simpler, cost-effective and less time-consuming as compared to that of synthetic chemical compounds (Weiss et al., 2007). The gene knockdown effect is also deemed more effective in treating diseases like cancer as it is often associated with overexpression of a network of oncogenes (Mansoori et al., 2014). Currently, there is quite a substantial number of RNAi-based drugs on clinical trial phase I/II with potential of being launched into the market in the near future (Bobbin and Rossi, 2016).

However, the past records do not suggest much optimism for these double-stranded siRNAs. Despite decade of investment, not a single RNAi-based drug is currently available in the market. Many had failed the phase III trial due to inflammatory issues and toxicity brought about by the method of delivery (Bobbin and Rossi, 2016). Furthermore, there are also concerns regarding the off-targeting caused by the passenger strand. In fact, just last year, a promising ds-siRNA drug designed for cardiomyopathy was withdrawn due to high mortality rate (Wittrup and Lieberman, 2015; Garber, 2017).

Due to these reasons, there is an apparent shift of focus from double-stranded RNAi to single-stranded RNAi therapy in the recent years (Haringsma et al., 2012; Lima et al., 2012; Yu et al., 2012; Liu et al., 2013; Hu et al., 2014; Liu et al., 2015; Matsui et al., 2016; Pendergraff et al., 2016). Single-stranded RNA oligonucleotides serve as a more promising approach than ds-siRNAs for a number of reasons. Firstly, they could be synthesized and delivered into systems via less convoluted methods resulting in less toxicity and inflammatory

responses (Bennett and Swayze, 2010). Secondly, the absence of an accompanying strand completely eliminates the off-target effects brought about by the mis-selection of guide strand by the slicer human AGO2 protein.

Nevertheless, single-stranded RNAi faces two main issues. The correctly loaded guide strand may still give off-target effects as it could be loaded to human AGO1, 3 and 4 which repress targets less specifically via the seed regions instead of cleaving fully-complement targets like AGO2 (Section 1.1.4). In addition, low efficacy due to rapid degradation by ribonucleases in cells is a huge obstacle for single-stranded RNAi. As a result, a much higher dosage would be needed as compared to double-stranded siRNAs (Schwarz et al., 2002; Martinez et al., 2002; Amarzguioui et al., 2003; Holen et al., 2003).

In tackling the low efficacy problem, there exist two possible solutions. One of which is to introduce different kinds of chemical modifications to the single-stranded RNA oligonucleotides so as to increase its resistance towards ribonucleases (Amarzguioui et al., 2003; Holen et al., 2003; Hall et al., 2006; Haringsma et al., 2012; Lima et al., 2012; Pendergraft et al., 2016). The other solution is none other than improving the AGO loading efficiency of the single-stranded RNA oligonucleotides.

As mentioned in Section 1.1.8, since the current consensus defines AGO as a protein that loads small RNA via structural criteria, there is no study done in optimizing single-stranded RNA sequence for enhanced AGO loading as yet. Imaginably, with crucial information on AGO loading sequence preference (if any), designs of single-stranded RNAi-based drugs could be further improved to attain the ultimate optimal efficacy.

1.3 Aims

We aim to develop a novel assay, High-throughput Sequencing mediated Specificity Aalysis (HISSA), for studying sequence specificity (if any) in the loading of single-stranded RNAs to fly AGOs (Section 1.4). We strive to identify the signature sequence motif(s) in AGO loading and the part(s) of a single-stranded oligonucleotide that is(are) involved in the sequence discrimination. In addition, we are also keen to explore the relevance of our findings in the biological context.

Hopefully, our findings could shed some light on the unknown loading mechanism of single-stranded RNAs and open an avenue to develop better RNAi technologies. As such, single-stranded siRNAs could be further optimized with enhanced efficacy for therapeutic applications.

1.4 **Methodology**

1.4.1 **HISSA: Independent of upstream processing steps**

The novel assay, HISSA (**Figure 1.4.1**), starts off with an oligo sequence of choice randomized at a particular region (e.g. 5' end; Section 2.2.1). It is labeled with Phosphorus-32 (^{32}P) at the 5' end and introduced into cell lysate for *in-vitro* AGO loading (Section 2.1.1; Section 2.2.1). The resultant AGO complexes are pulled down and small RNAs loaded in these complexes are extracted (Section 2.1.2; Section 2.2.1). The extracted RNAs are then constructed into a library using a highly optimized protocol (Section 2.1.4; Section 2.1.6 - Section 2.2.1). After the library is being sequenced, the high-throughput data gets analyzed with customized computational pipelines (Section 2.1.5; Section 2.2.2). Finally, by carrying out *in-vitro* loading assay on individually synthesized non-randomized oligos, the computational ranking of individual sequences is validated (Section 2.2.3).

We designed HISSA to study AGO sequence specificity because this *in-vitro* method allows loading of single-stranded RNAs to be independent of upstream processing steps. This is highly desirable because not only does it give a more accurate measure of true AGO loading events, it also grants more flexibility in the choice of backbone sequence to randomize because it is not limited by the substrate recognition strength of any processing factor. Furthermore, HISSA is more time-saving and cost efficient as compared to using cell system.

Figure 1.4.1: Overview of HISSA

A single-stranded RNA oligo with a backbone sequence of choice is randomized at a region of interest. This pool of single-stranded RNAs with sequence variation is introduced into S2R⁺ FLAG-HA-AGO2 cell lysate for *in-vitro* AGO loading. RNAs are then extracted from the immuno-precipitated AGO complexes and used for subsequent cDNA library construction using a highly optimized protocol. The resultant libraries are subjected to high-throughput sequencing and the computational data is analyzed with a customized pipeline. The computational ranking of individual sequences derived from the randomized oligo is eventually validated with *in-vitro* loading assay using individual synthesized non-randomized oligos.

1.4.2 High-throughput sequencing: Massively parallel analysis

As mentioned in Section 1.4.1, the high-throughput sequencing (HTS) library has to be prepared before it can be sequenced. **Figure 1.4.2** shows the procedure involved in the library construction. The first step involved the ligation of 3' linker to the 3' ends of AGO-loaded small RNAs. 5' linker is then ligated to the 5' ends in the second step. Reverse transcription (RT) primer that is reverse complement to the 3' linker is used for subsequent cDNA synthesis. In the last step, polymerase chain reaction (PCR) is carried out with PCR primers that carry sequences of 5' linker and RT primer. The PCR products are eventually subjected to HTS (Section 2.1.4).

HTS enables us to obtain as much information as possible from AGO-loaded randomized oligo sequences. Such deep sequencing enables us to compare huge number of sequences concurrently and conclude statistically meaningful trends. As the ranking of individual sequences is formed within the library itself, we do not need to be concerned about variation caused by external factors such as enzymatic efficiency, temperature and handling. As this one-time generated data could be easily manipulated and analyzed through computational means, it is definitely much more convenient than traditional cloning methods.

HTS library construction procedure

Figure 1.4.2: HTS library construction procedure

RNAs extracted from the immuno-precipitated AGOs comprised of endogenous small RNAs (black) and synthetic RNA oligos (pink) we introduced for loading. However, the synthetic RNA oligos only made up a very small percentage of the total population. The very first step in library construction is 3' linker (green) ligation to the 3' end of the extracted RNAs. As ligation is hardly 100% efficient, gel purification is carried out to remove non-ligated RNAs before 5' linker (blue) ligation. After one more round of gel purification, the resulting 5' and 3' linker ligated products are subjected to RT-PCR to generate cDNAs for HTS sequencing.

1.4.3 Fly: Distinct loading patterns between two AGOs

Undeniably, decoding the sequence preference (if any) in human AGO loading would be highly impactful because any single breakthrough could significantly contribute to the designs of single-stranded RNAi (Section 1.2.2). However, the fly system was chosen over the human as our pilot study system because of two main reasons.

First, as mentioned in Section 1.1.3 and Section 1.1.4, fly somatic tissues usually express AGO1 and AGO2 that are involved in the small RNA pathways whereas there are four AGOs generally expressed in human tissues. Keeping the number of AGOs in competition small, would make analysis less complex and more manageable.

Second, fly AGOs tend to have much more distinct loading patterns than that of human. This is supported by the fact that siRNAs and miRNAs are distinctly sorted to fly AGO2 and fly AGO1 respectively while miRNAs can be sorted almost equally well to all four human AGOs (Section 1.1.3 - Section 1.1.5). We therefore believe that if AGO proteins happen to confer sequence specificity, fly AGOs would probably provide higher sensitivity than that of human and produce more apparent loading differences for a diverse pool of single-stranded RNAs.

The two points mentioned above are of great importance in our case because we need to first assess whether the concept of 'sequence specificity' is even valid for AGO loading in organisms before setting precise parameters for HISSA based on what we observed. Though the human system satisfies the general interests of the public, it did not allow us to analyze our preliminary data with great confidence as a result of the ambiguous competition for small RNA substrates among the four AGOs. Therefore, we are unable to design the appropriate HISSA setups for the human system yet.

1.4.4 Cell lysate: AGO loading in physiological context

As there is hardly any information available for single-stranded RNA loading to AGO proteins, we do not know what factor(s) is(are) involved in the cellular context. For this reason, we carried out *in-vitro* loading of the synthetic oligos in cell lysate since it best mimics the physiological environment as compared to immuno-precipitated AGOs or purified recombinant AGO proteins.

However, we acknowledge that there is a risk of obtaining artefactual data with the use of cell lysate. This is because many enzymes could be released from the sub-cellular compartments and nucleic acids could be sheared during the lysis of the cells. These could directly or indirectly interfere with the loading reaction. In order to counter this possible downfall, we will validate our HISSA results with trends observed in cellular context (Section 3.3.1). In order to check if the RNA sequences analyzed in HISSA are indeed bound by AGO, we would carry out a series of biochemistry experiments in the near future (Section 6).

Working with purified AGO proteins may be the best in pin-pointing binding specificity to AGO itself but it can also mean totally meaningless study if any factor is involved in the single-stranded loading. Also, majority of the AGO proteins extracted from the cells are pre-loaded with endogenous small RNAs and these pre-loaded RNAs could not be exchanged using synthetic oligos (Elkayam et al., 2012; Nakanishi et al., 2012). Yet, the percentage of non-occupied fly AGO1 available for *in-vitro* loading experiment would be very small because it is reported that non-loaded fly AGO1 is degraded in the cells (Smibert et al., 2013). I have also experimentally confirmed this notion (**Appendix (Section 8.1)**). Furthermore, we would probably need to add endogenous RNAs to mimic the competition in the cells if purified AGO is used. Altogether, this means that highly concentrated purified proteins or extremely large reaction would be required so that sufficient non-occupied AGO is present for the loading experiment and also for detecting the loaded single-stranded oligos. Hence, it is not economical and most likely to be technically infeasible.

Therefore, despite the fact that cell lysate may not necessarily be the perfect platform for us to study AGO loading, it is definitely the best choice we could made at the present given the technologies and information.

2. MATERIALS AND METHODS

2.1 HISSA development

2.1.1 In-vitro loading assay

S2R+ FLAG-HA-AGO2 cells were lysed with hypotonic buffer [30mM HEPES-KOH (pH 7.4), 2mM magnesium acetate, 5mM DTT and 1x complete EDTA free protease inhibitor (Roche, Germany)] to make 10mg/ml cell lysate. Each loading reaction took place in 160µl of the cell lysate. Single-stranded RNA oligos (Table 2.1.1) were labeled at the 5' end with γ -³²P-ATP (PerkinElmer, USA) using T4 PNK enzyme (New England Biolabs, USA) and 2x10⁶ counts per minute (cpm) of labeled oligos was added to the lysate forming final reaction with RNaseOUT (Invitrogen, USA), 0.5mM ATP (Sigma-Aldrich, USA), 5mM DTT, 0.1M KOAc and water. The reactions were incubated at 25°C for 1 hour followed by the addition of PXL buffer (1xPBS (pH7.4), 0.1% SDS, 0.5% deoxycholate, 0.5% NP40) for quenching (Goh and Okamura, 2018).

	Oligo name	Oligo Sequence
1	<i>mir-34 loop</i>	UAGCCAAUUAUUGCCGUUGACA <u>AAU</u> <u>UCA</u>
2	<i>mir-34 loop-short</i>	UAGCCAAUUAUUGCA <u>AAU</u> UCA
3	<i>mir-317 loop</i>	UUGAAUGAAAUGCA <u>AAGCAAG</u>
4	<i>mir-317 loop-long</i>	UUGAAUGAAAUGCAAGCAAG <u>AAGC</u> <u>AAG</u>
5	<i>mir-34 loop-short (mir-317 5')</i>	UUGAA <u>AAU</u> UAUUGCA <u>AAU</u> UCA
6	<i>mir-34 loop-short (mir-317 3')</i>	UAGCCAAUUAUUGCAGCAAG
7	<i>mir-317 loop (mir-34 5')</i>	UAGCCUGAA <u>AAU</u> UGCAAGCAAG
8	<i>mir-317 loop (mir-34 3')</i>	UUGAAUGAA <u>AAU</u> UGCA <u>AAU</u> UCA

Table 2.1.1: Oligos used for HISSA parameters selection

The original *mir-34 loop* sequence (1; blue) was shortened (2; blue) by bringing the last 7 nucleotides (underlined) forward while the original *mir-317 loop* sequence (3; green) was lengthened (4; green) by repeating the last 7 nucleotides at the 3' end. Oligos (2) and (3) were further modified by swapping their 5' ends (5, 7) and 3' ends (6, 8) with each other.

2.1.2 Immuno-precipitation, RNA extraction and auto-radiography

60µl and 36µl Dynabeads (Life technologies, USA) were incubated with 10µg FLAG (Wako, Japan) and 6µg AGO1 antibodies (AbCam, United Kingdom) respectively for 30 minutes at room temperature. Same volume of beads and antibodies were used for Mouse IgG and Rabbit IgG respectively. Beads were washed with citrate phosphate buffer (25mM citric acid, 65mM Na₂HPO₄, pH5.0) (3 times) and PXL buffer (2 times). 400µl out of ~1000µl of a loading reaction (after addition of PXL buffer) was added to each tube of FLAG and Mouse IgG antibody-bound beads and incubated for 90 minutes at 4°C; co-immuno-precipitation (co-IP) was then carried out (FLAG-IP followed by AGO1-IP and Mouse IgG followed by Rabbit IgG). Beads were again washed with PXL buffer (5 times) and PBS (1 time), followed by RNA extraction using 0.4M NaCl and Phenol-Chloroform-Isoamylalcohol (Sigma-Aldrich, USA). Precipitated RNA and input sample (100 cpm of labeled oligo) were loaded and run in 15% denaturing UreaGel (National Diagnostic, USA). The RNA in the gel was then transferred to membrane using semi-dry transfer system. The membrane was exposed in a cassette and the film was scanned using Typhoon 9200 (GE Healthcare, United Kingdom). Signals were quantified using ImageQuant TL software.

2.1.3 Degradation assay

2x10⁵ cpm of labeled oligos were added to 80µl of lysate for degradation assay. RNA was extracted directly from 10µl of the reaction at five time points – 0, 2, 5, 15 and 30 minutes. Samples were loaded into 15% denaturing gel and auto-radiography was carried out. Resulting films were scanned and quantified (Section 2.1.2). The reading from each time point was expressed as a percentage 0 min reading and graph of log₂ (normalized %) against time (in minutes) was plotted. Best-fit line was drawn to obtain the R² value. Half-life of each oligo sequence was calculated using the best-fit line equation and standard deviation was derived based on triplicates.

2.1.4 High-throughput sequencing library protocol modification

In the modified protocol, with reference to Zhang et al., 2013, for each high-throughput sequencing (HTS) library, 40µg of total RNA was extracted from S2R+ cells using TRIzol (Ambion, USA) and 1fmol of spike-in oligos (consisting of 10 miRNAs) was added.

Small RNA ranging from 18nt – 29 nt was gel purified in 15% denaturing gel. This was followed by 4 hours of 10uM RA3 3' linker (Integrated DNA Technologies, USA) (Table 2.1.4) ligation catalyzed by T4 RNA ligase 2, truncated (New England Biolabs, USA) at room temperature. The reaction component (final volume) also included 1X ATP-free T4 RNA ligase buffer (New England Biolabs, USA), 10% PEG 8000 (New England Biolabs, USA) and RNaseOUT (Invitrogen, USA). After gel purification of the ligated products (with reference to the radioactive markers), 10uM RA5 5' linker (Integrated DNA Technologies, USA) (Table 2.1.4) ligation by T4 RNA ligase 1 (Ambion, USA) took place in 1x T4 RNA ligase buffer (Ambion, USA), 1mM ATP (Sigma-Aldrich, USA), 20% PEG 8000 (New England Biolabs, USA) and RNaseOUT (Invitrogen, USA) for two hours at 37°C. Gel purification was again done to obtain 5' linker-RNA-3' linker products. Reverse transcription (RT) was carried out using 2µl Superscript III RT (Life technologies, USA), 1uM RTP (Integrated DNA Technologies, USA) (Table 2.1.4), 1x first strand buffer (Life technologies, USA), 1mM dNTP (Promega, USA), 2mM DTT (Life technologies, USA) at 50°C for one hour. The resulting cDNA was then amplified by PCR with 1x iProof HF Mastermix (Bio-rad, USA), 1uM RP1 and 1uM RPI (Integrated DNA Technologies, USA) (Table 2.1.4). PCR products were purified with Phenol-Chloroform-Isoamylalcohol (Sigma-Aldrich, USA) and precipitated. PmeI (New England Biolabs, USA) digestion of the PCR products derived from size standards was done before TBE PAGE and gel purification. Beijing Genomics Institute (BGI) in Hong Kong performed the HTS.

Standard deviation of spike-in oligo sequences was calculated for each library set. Average standard deviation was calculated by considering all libraries constructed using “Modified protocol”. This value was then compared to that of

libraries constructed previously using “Conventional protocol” (Malone et al., 2012) to conclude if biasness issue was improved in the modified version.

Oligo name	Oligo Sequence
illumina RA3 (3' linker)	NNTGGAATTCTCGGGTGCCAAGGdideoxyC
illumina RA5 (5' linker)	GUUCAGAGUUCUACAGUCCGACGAUCNN
illumina RTP (RT primer)	GCCTTGGCACCCGAGAATTCCA
illumina RP1 (PCR forward primer)	AATGATACGGCGACCAACCGAGATCTACAC GTTTCAGAGTTCTACAGTCCGA
Illumina RPI (PCR reverse primer)	CAAGCAGAAGACGGCATACGAGATXXXXX XGTGACTGGAGTTCCTTGGCACCCGAGAA TTCCA

Table 2.1.4: Oligos used for HTS library construction

RA3 linker was added to the 3' end of RNA followed by RA5 linker to the 5' end. Ligated RNAs was reverse transcribed using *RTP* and *PCR* was carried out with RP1 and RPI. NN: randomized dinucleotide; XXXXXX: illumina barcode.

2.1.5 Sequencing errors elimination scripting

Because our libraries were made from small amounts of RNA, and extremely deeply sequenced (~30M reads per library on average), a single cDNA molecule can give rise to multiple sequencing reads due to PCR amplification during library construction. Such PCR amplification can introduce additional bias to the sequencing results, and we wished to eliminate this artifact. Based on the sequencing depth, the possibility of having the same combination of 5'-linker (with 12 randomized nucleotides as barcodes) and insert (8 or 12 randomized nucleotides) sequences is negligible (Theoretically, the barcoded 5' linker itself can have $>1.6 \times 10^7$ combinations.). Therefore, all reads having the same combination of the 5' linker and the insert were merged and counted as 1 read, based on the assumption that those reads are derived from a single cDNA molecule.

Having eliminated the artifact of potential PCR amplification, additional problems arose. Reads that are very similar to abundantly PCR amplified reads

(different by 1-2 nucleotides to the major sequence) were observed more frequently than expected. This can be explained if those reads are derived from the same cDNA molecule but the nucleotide differences were introduced by mutations during PCR amplification or errors during the sequencing reaction. In order to eliminate these PCR/sequencing errors in the insert sequence (i.e. sequence flanked by the 2 linkers; Section 1.4.2), reads that had the same 5' linker sequence (with 12 randomized bases) were grouped together. Within each group, the insert sequences were extracted according to whether they have 2 mismatches (mismatch 1; m1) of one another using `fqgrep` (Indrani and Gosselin, 2016). Sequences that were not extracted out were taken out of the respective linker group and categorized as 'non-error'. The rationale for doing this is that during PCR, many errors could accumulate progressively whereby more erroneous reads could be built from initial error-containing population. By doing this, we could safely assume that the 'non-error' reads were error free. The remaining sequences within individual groups were re-sorted in descending order with respect to their read counts. The top most sequence was in turn used to eliminate sequences that were 4 mismatches (mismatch 2; m2) of itself in the same linker group using `fqgrep`. After that, it was categorized as "major" and removed from the linker group. This process of sorting, elimination and adding top most sequence to "major" group was repeated until all sequences in every individual linker groups were eliminated/removed (**Figure 3.1.3b**).

At the end, all "non-error" and "major" sequences were combined to form the input file for the downstream steps, which involved removal of endogenous RNA species and contaminating reads. Subsequently, the same procedure was done to eliminate errors in the linker sequences (Section 2.2.2; **Figure 3.2.1b**).

As mentioned above, this error elimination script requires input of 2 parameters - the number of mismatches to be screened in the first round (m1) and also the second round (m2). Endogenous *mir-317 loop* sequence (most abundant sequence) in the Input sample (Section 2.2.1) would be useful in deciding these parameters. The number of sequences that have 0-X mismatches from the *mir-317 loop* sequence could be tabulated and a typical trend observed for such analysis is that the number of sequences decreases as you increase the number

of mismatches. Then at a certain point, the sequences start to increase gradually and gets much larger.

The mismatch group that has the least number of sequences (the point at which number of sequences starts to increase) should be taken as the maximum number of mismatches that could derive from sequencing/PCR errors. This is because in the event that any further increment in the mismatch results in an increase in sequence number, it means that the mismatches have included a significant percentage of the non-error population. Given that PCR errors tend to be accumulative due to exponential amplification, the criteria set in the first round of selection (m1) is best to be just half the maximum number of mismatches. Only after 'non-error' reads are excluded from the individual linker groups, the criteria for second round (m2) could be set at maximum number of mismatches.

2.1.6 RT-PCR optimization

As the small RNAs extracted from AGO1 complexes were flooded with endogenous small RNAs including mature miRNAs, any AGO1-IP HTS library constructed with protocol in Section 2.1.4 would hardly contain any of the sequences derived from the introduced oligo. In order to ensure meaningful analysis could be carried out, we had to specifically amplify these oligo sequences in the library before sequencing. One way was to enrich them at the end of library construction, particularly the RT-PCR step (Section 2.1.4 and Section 1.4.2).

Total RNA was extracted from S2R+ FLAG-HA-AGO2 cell lysate using TRIzol (Ambion, USA). Endogenous small RNA ranging from 18nt – 29 nt was gel purified in 15% denaturing gel (National Diagnostic, USA). For every RT-PCR primer/condition tested, 0.8 fmol of 5'-monophosphorylated *mir-317 loop* oligo (with BspHI restriction site; Table 2.1.6) was added to the endogenous small RNA purified from 160µg of total RNA (estimated yield from 10x IP reaction with 800µl of lysate; Section 2.1.1 and Section 2.1.2). Library construction steps were then done as mentioned in Section 2.1.4. However, RT-PCR primers in

Table 2.1.4 were modified to form base pairs between both the 3' linker sequence and the *mir-317* loop sequence. 3 different RT primer designs were tested in combination with 1 PCR reverse primer (Table 2.1.6). The optimal RT temperature was determined by testing a range - RT by ThermoScript Reverse Transcriptase (ThermoFisher Scientific, USA): 58°C to 70°C. Annealing of PCR primers was at 66°C. PCR products generated by iProof HF Mastermix (Bio-rad, USA) were digested with BspHI (New England Biolabs, USA), subjected to TBE PAGE and visualized using DNA transilluminator.

Type	Oligo name	Oligo Sequence
1	<i>mir-317</i> loop oligo with BspHI site	NNNNN UCAUG AAUGCAAGCAAG
2	RTP-14nt linker+9nt oligo	ACCCGAGAATTCCACTTGCTTGC
3	RTP-14nt linker+6nt oligo	ACCCGAGAATTCCACTTGCT
4	RTP-8nt linker +6nt oligo	GAATTCCACTTGCT
5	RPI-9nt oligo	CAAGCAGAAGACGGCATACGAGATXXXXXX GCACCCGAGAATTCCACTTGCTTGC

Table 2.1.6: Oligos tested for RT-PCR specificity

The *mir-317* loop sequence was modified by adding 3nt (red) to the 6th to 8th positions, forming BspHI restriction site (1; bold). This site was needed for the visualization of oligo-specific PCR products. PCR products derived from endogenous small RNAs were not digested by BspHI and hence library specificity could be estimated by the ratio of digested: non-digested products. The names of the RT-PCR primers (2-5) were based on the number of base-pairings with the RA3 linker (italic) sequence and the 3' end of *mir-317* loop oligo (underlined). Generally, less base-pairings with the oligo or 3' linker enhanced the specificity but compromised on the efficiency. The optimal RT-PCR primers were RTP-8nt linker+6nt oligo and RPI-9nt oligo (4-5). N: randomized nucleotide; XXXXXX: illumina barcode.

2.1.7 Splinted-ligation optimization

As the RT-PCR optimization in Section 2.1.6 did not attain the desired enrichment, we went on to tackle the first step of the modified protocol (Section 2.1.4 and Section 1.4.2). Instead of 3' linker ligation by T4 RNA ligase 2, splinted-ligation was employed to confer oligo-specificity at this step.

For each splinted ligation reaction, 0.8 fmol of 5'-mono-phosphalated *mir-317* loop oligo (Table 2.1.1; without BspHI site) and 10nM of γ -32P-5'-labeled RA3 linker (Table 2.2.1b) (Integrated DNA Technologies, USA) were mixed with endogenous small RNA purified from 160 μ g of total RNA. 100nM bridge oligo (in 10x Capture Buffer) was added to a final concentration of 10nM. For selecting the optimal bridge, 3 different designs were tested (Table 2.1.7). The reaction was first heated to 94°C for 1 minute, followed by 65°C for 2 minutes and 37°C for 10 minutes. 2x ligation buffer and T4 DNA ligase (Promega, USA) were added and incubated at 37°C for an hour. Calf Intestinal Phosphatase (New England Biolabs, USA) removed γ -32P from non-ligated linkers (Maroney et al., 2008). Ligation efficiencies and specificities of the bridges were analyzed by auto-radiography (Section 2.1.2).

Type	Oligo name	Oligo Sequence
1	14nt linker +15nt oligo	ACCCGAGAATTCCACTTGCTTGCATTTC <u>A</u>
2	8nt linker +15nt oligo	GAATTCCACTTGCTTGCATTTC <u>A</u>
3	8nt linker +6nt oligo	GAATTCCACTTGCT

Table 2.1.7: Bridges tested for splinted-ligation specificity

The names of the bridges (1-3) were based on the number of base-pairings with the RA3 linker (*italic*) sequence and the 3' end of *mir-317* loop oligo (underlined). Generally, less base-pairings with the 3' linker enhanced the specificity but compromised on the efficiency. The bridge that gave the best combination of specificity and efficiency was 8nt linker+15nt oligo.

2.1.8 Oligo-specific library protocol validation

The oligo-specific HTS library protocol combined all optimized conditions in Section 2.1.4 and Section 2.1.6 – Section 2.1.7 and is fully described in (Goh and Okamura, 2018). A summary with validation purpose is as follows:

1000cpm of γ -³²P-5'-labeled-*mir-317 loop* oligo (with BspHI restriction site; Table 2.1.6) was added to the endogenous small RNA purified from 160 μ g of total RNA. 10nM 5'-mono-phosphalated-RA3 linker was ligated to the 3' end of *mir-317 loop* oligo using 10nM optimal bridge (8nt linker+15nt oligo) (Section 2.1.7; Table 2.1.7). After gel purification, subsequent 5' linker reaction was carried out as described in Section 2.1.4. RT-PCR using was done with the optimal primers and conditions (RTP-8nt linker+6nt oligo, 58°C; RPI-9nt oligo, 66°C) (Section 2.1.6; Table 2.1.6). The PCR products were digested by BspHI, separated with TBE PAGE and visualized using DNA transilluminator. The oligo-specificity of the library was estimated by the ratio of digested products (~60bp and ~80bp): non-digested products (~140bp).

2.2 Hidden rule in AGO loading of single-stranded RNAs

2.2.1 HISSA overall setup, parameters and procedure

Oligos each harboring 5nt, 8nt and 12nt randomization at the 5' end of the *mir-317 loop* sequence (Integrated DNA Technologies, USA) (Table 2.2.1b) were selected for HISSA study. For each of them, a set of 7 samples was designed - 4 IPs (FLAG, AGO1, Rabbit IgG and Mouse IgG), 2 Inputs for normalization and a RAW sample for potential troubleshooting purposes. Each Input comprised of endogenous small RNAs extracted from FLAG or AGO1 IP, but unlike IP sample, the randomized oligo was added only after the RNA extraction (i.e. oligo was not added to the lysate before IP). RAW sample did not contain any endogenous small RNA as it was purely randomized oligo itself (**Figure 3.2.1a**; Table 2.2.1a).

In HISSA, γ -³²P-5'-labeled randomized oligo was subjected to *in-vitro* loading assay and IP as described in Section 2.1.1 and Section 2.1.2 except that the loading reaction was carried out on a 5x scale (Table 2.2.1a). The RNA extracted from all IP samples (a total of 6; including Input-to-be samples) were loaded to 15% denaturing gel and at this point, ~1100cpm and ~750cpm of the randomized oligo were added to the FLAG Input and AGO1 Input respectively. ~1100cpm oligo was loaded to an empty lane to serve as the RAW sample (Table 2.2.1a). After gel purifying the radioactive oligo (together with endogenous RNAs), all samples were used to construct oligo-specific HTS libraries according to the optimized protocol in Section 2.1.8. The libraries were checked and quantified by Bioanalyzer and sent for sequencing at Beijing Genomics Institute (BGI) in Hong Kong.

There are several important points to note: 1) 5nt, 8nt and 12nt randomization libraries were constructed with different bridges, RA5 linkers and RT-PCR primers (Table 2.2.1b) though all reaction conditions were the same, 2) Unlike in the optimization phase (Section 2.1.7), the bridges used in HISSA contained chemical modifications at the 3' end to prevent them from acting as RT primers

(Table 2.2.1b) and 3) Restriction digestion was not carried out after PCR because no BspHI site was present in the oligos.

Sample	FLAG Input	AGO1 Input	FLAG IP	AGO1 IP	Mouse IgG	Rabbit IgG	RAW
<i>In-vitro</i> loading	800µl lysate;	800µl lysate;	800µl lysate;	800µl lysate;	800µl lysate;	800µl lysate;	No lysate;
	No oligo	No oligo	10Mcpm oligo	10Mcpm oligo	10Mcpm oligo	10Mcpm oligo	No oligo
IP	100µg AB;	60µg AB;	100µg AB;	60µg AB;	100µg AB;	60µg AB;	No AB;
	600µl beads	360µl beads	600µl beads	360µl beads	600µl beads	360µl beads	No beads
Gel loading	1100 cpm oligo	750 cpm oligo	No oligo	No oligo	No oligo	No oligo	1100 cpm oligo

Table 2.2.1a: HISSA setup with *in-vitro* loading and IP parameters

For each randomized oligo in study, a total of 7 samples were needed. During *in-vitro* loading, γ -32P-5'-labeled randomized oligo was added to the lysate for all IP samples only. Inputs had equivalent volume of water added to the lysate instead of oligo while RAW sample did not have any loading reaction at all. After loading reaction, IP was carried out for all IP samples and Inputs; no IP reaction for RAW sample. It was only during gel loading, the γ -32P-5'-labeled randomized oligo was added to the Inputs and RAW sample. M: Million; AB: Antibody; cpm: counts per minute

Oligo name		Oligo Sequence
1	5NN- <i>mir-317</i> loop	NNNNNUGAAAUGCAAGCAAG
2	8NN- <i>mir-317</i> loop	NNNNNNNNAAUGCAAGCAAG
3	12NN- <i>mir-317</i> loop	NNNNNNNNNNNNCAAGCAAG
4	illumina RA3 (no randomized 5' end)	TGGAATTCTCGGGTGCCAAGGdideoxyC
5	8nt linker+15nt oligo (for 5NN)	GAATTCCACTTGCTTGCATTC3'deoxyA
6	8nt linker+12nt oligo (for 8NN)	GAATTCCACTTGCTTGCAT3'deoxyT
7	8nt linker+8nt oligo (for 12NN)	GAATTCCACTTGCTT3'deoxyG
8	illumina RA5 (for 5NN)	GUUCAGAGUUCUACAGUCCGACGAUCNN
9	Illumina RA5 (for 8NN and 12NN)	GUUCAGAGUUCUACAGUCCGACGAUCNN NNNNNNNNUGUAGCNN
10	RTP-8nt linker +6nt oligo (for 5NN)	GAATTCCACTTGCT
11	illumina RTP (for 8NN and 12NN)	GCCTTGGCACCCGAGAATTCCA
12	Illumina RP1	AATGATACGGCGACCAACGAGATCTACAC GTTCAAGAGTTCTACAGTCCGA
13	RPI-9nt oligo (for 5NN)	CAAGCAGAAGACGGCATACGAGATXXXXXX GCACCCGAGAATTCCACTTGCTTGC
14	RPI-6nt oligo (for 8NN)	CAAGCAGAAGACGGCATACGAGATXXXXXX GCACCCGAGAATTCCACTTGCT
15	RPI-3nt oligo (for 12NN)	CAAGCAGAAGACGGCATACGAGATXXXXXX GCACCCGAGAATTCCACTT

Table 2.2.1b: Summary of oligos used in all 3 sets of HISSA experiments

In HISSA, 3 oligos each contained 5nt (1), 8nt (2) and 12nt (3) randomization in the 5' end of the *mir-317* loop sequence (with no BspHI site) were introduced into lysate for *in-vitro* loading. Small RNA extracted from the IP samples was splinted ligated with RA3 linker (4) with the aid of bridges (5-7). This was followed by RA5 linker (8-9) ligation. The ligated products were reverse transcribed with primers (10 -11) and eventually amplified by PCR with primers (12-15). Italic: base-paired with 3' linker sequence; underlined: base-paired with 3' end of oligo; N: randomized nucleotide; XXXXXX: illumina barcode

2.2.2 Computational processing and analysis of HISSA data

Software and programs utilized in these pipelines were as stated. All other intermediate file manipulation was done using AWK (Aho et al., 1979).

5NN HISSA:

Quality of libraries were checked using FastQC (Andrews, 2014). 3' linker sequence was trimmed off using fastX Toolkit (Gordon and Hannon, 2010) and the first 2 nucleotides were removed as they belonged to the 3' end of RA5 linker. Sequences which were 20nt in length and contained "TGAAATGCAAGCAAG" from the 6th to 20th position were extracted as they were the sequences derived from the 5NN-*mir-317 loop* oligo. The extracted sequences were input into FASTaptamer (Alam et al., 2015) for normalizing IP read counts to that of Input. The output $\log_2(\text{IP}_{\text{rpm}}/\text{Input}_{\text{rpm}})$ values were sorted in descending order and plotted into Excel clustered bar-chart with each individual sequence on the X-axis and $\log_2(\text{IP}_{\text{rpm}}/\text{Input}_{\text{rpm}})$ values on the Y-axis. A sequence logo was computed for sequences with $\log_2(\text{IP}_{\text{rpm}}/\text{Input}_{\text{rpm}}) \geq 1$ using R (R-Core-Team, 2013).

8NN and 12NN HISSA:

These 2 sets of libraries had many more sequence combinations derived from the randomized oligos than 5NN HISSA (5NN HISSA: 1024; 8NN HISSA: ~65,000; 12NN HISSA: ~16,000,000). Given the fixed sequencing depth, 5NN HISSA had all 1024 sequences in both Input and AGO1-IP samples but 8NN and 12NN HISSA had different individual sequences appearing in the Input and AGO1-IP samples. Due to this reason, it was not technically feasible to do a direct motif enrichment like in the case of 5NN HISSA. Therefore, we sought to do the analysis using different methods.

After checking the quality using FastQC (Andrews, 2010), all sequences that had 10nt between "TGTAGC" (RA5 linker) and "AATGCAAGCAAG" (non-variable region of 8NN-*mir-317 loop*) or 14nt between "TGTAGC" (RA5 linker) and "CAAGCAAG" (non-variable region of 12NN-*mir-317 loop*) were extracted. After which, 3' linker was trimmed off from these sequences and all identical

reads (each read: 5' linker + oligo sequence) were collapsed with respective read counts summed up. Sequences that were derived due to sequencing or PCR error were eliminated (Section 2.1.5). In addition, the following were also removed: 1) endogenous *mir-317 loop* sequences with 2 mismatches or less, 2) contaminated sequences, 3) sequences with abnormally high read counts (~5 - 10 folds higher than average) in the input.

The final files were analyzed for the number of A, C, G and T at all positions (first 8 positions for 8NN-*mir-317 loop* and first 12 positions for 12NN-*mir-317 loop*) and the values in IP samples were normalized to that of the respective Inputs. The $\log_2(\text{IP}/\text{Input})$ value for each nucleotide at each position was plotted into Excel clustered bar-chart. For 8NN-*mir-317 loop* HISSA, pairwise analysis (below) was applied in finding out the association strength of every single possible combination of dinucleotides with AGO loading efficiency.

Pairwise analysis:

The original analysis was employed in Guenther et al., 2013 but we made modifications to the pipeline so that a total of 448 different combinations of dinucleotides (for 8 randomized positions) could be studied instead of just 240 combinations (5 randomized positions in Guenther et al., 2013).

The 8 randomized positions of 8NN-*mir-317 loop* were split into 70 combinations of 4nt window (${}^8C_4=70$). For every 4nt window, read counts for every 4mers ($4^4=256$) were tabulated using RBNS (Lambert et al., 2014). (IP/Input) values were obtained and normalized to the median (IP/Input) of that 4nt window using Excel. A pairwise matrix consisting the values of 0 and 1 was computed for the 256 4mers (e.g. AAAA had value of 1 for A1A2 but value 0 for A1C2, A1G2 and A1U2). This matrix together with the normalized (IP/Input) values were input into R (R-Core-Team, 2013) to compute into a linear model. Linear regression was carried out to shortlist statistically significant pairwise interactions that contribute to AGO loading. Linear coefficients of the shortlisted pairwise interactions were then obtained by backward stepwise regression. All the 4nt windows were combined and the coefficients were averaged. The ultimate heat-map showing

the co-relation between loading efficiency and every pair of nucleotides was drawn using R.

2.2.3 HISSA computational data validation with *in-vitro* loading assay

In-vitro loading, IP and autoradiography (Section 2.1.1 and Section 2.1.2) were conducted with oligos shown in Table 2.2.4 to experimentally validate HISSA computational data.

Oligo name		Oligo Sequence
1	5'GACGAUGA- <i>mir-317 loop</i>	<u>GACGA</u> UGAAAUGCAAGCAAG
2	5'GACAAUGA- <i>mir-317 loop</i>	<u>GACAA</u> UGAAAUGCAAGCAAG
3	5'GGGGGUGA- <i>mir-317 loop</i>	<u>GGGGG</u> UGAAAUGCAAGCAAG
4	5'CGAGGUGA- <i>mir-317 loop</i>	<u>CGAGG</u> UGAAAUGCAAGCAAG
5	5'CTGGGUGA- <i>mir-317 loop</i>	<u>CTGGG</u> UGAAAUGCAAGCAAG
6	5'GACAAUAC- <i>mir-317 loop</i>	<u>GACAA</u> UACAAUGCAAGCAAG
7	5'GACGAUAC- <i>mir-317 loop</i>	<u>GACGA</u> UACAAUGCAAGCAAG
8	5'UACGAUAC- <i>mir-317 loop</i>	<u>UACGA</u> UACAAUGCAAGCAAG
9	5'GUCGAUAC- <i>mir-317 loop</i>	<u>GUCGA</u> UACAAUGCAAGCAAG
10	5'UUCGAUAC- <i>mir-317 loop</i>	<u>UUCGA</u> UACAAUGCAAGCAAG

Table 2.2.3: Oligos for validating AGO loading trends shown by HISSA

(1-5) validated the loading trend seen in 5NN *mir-317 loop* HISSA and (1,2,7,8) showed relevance to 8NN *mir-317 loop* HISSA results. (7-10) proved that 5'GAC was indeed preferred over 5'U.

2.3 Biological evidence for HISSA findings

2.3.1 Computational analysis of mRNA fragments in libraries

Short mRNA fragments were examined in the effort of understanding AGO1 loading sequence specificity in the cellular context because they are the only RNAs in the cell that are single-stranded, random and with a huge variety of sequence combinations. Such conditions are similar to the pool of randomized sequences we used in HISSA. We postulated that some of these mRNA fragments are loaded to AGO1 via the single-stranded RNA loading mechanism. If this is true, we should be able to observe 5'GAC enrichment in AGO1-loaded mRNA fragments, just like what we observed in HISSA.

The BED files of those previously mapped AGO1-IP and Input libraries (Table 2.3.4; Chak et al., 2015) were used for downstream reads extraction and elimination. Reads lying within gene loci annotated in RefSeqGene (O'Leary et al., 2016) were extracted so as to eliminate all non-coding reads from sites such as telomeres and intergenic regions. Subsequently, reads belonging to all annotated miRNAs, siRNAs, piRNAs and repeat-masker were removed. siRNA annotations were as reported in (Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008; Okamura et al., 2008a; Okamura et al., 2008b; Ruby et al., 2007; Tam et al., 2008; Watanabe et al., 2008). Loci of individual piRNAs residing in coding genes were downloaded from piRBase (Zhang et al., 2014) and merged to obtain longer piRNA transcript loci. Additional siRNAs and piRNAs loci were obtained from Wen et al., 2014. Repeat-masker annotation BED files were downloaded from UCSC Table browser (Karolchik et al., 2004; Rosenbloom et al., 2015). All multi-mappers, staggered-reads, and 21nt reads were also eliminated.

At the end of all elimination steps, the remaining reads were treated as single-stranded mRNA fragments. For each library, 3nts-shifting frames were extracted up to the 8th position (1st - 3rd, 2nd - 4th, 3rd - 5th, 4th - 6th, 5th - 7th, 6th - 8th) of the mRNA fragments and a total of 64 ($4^3=64$) 3nt-motifs with their respective read counts were obtained for each frame. The enrichment of 3nt-motifs in

AGO1 were obtained by normalizing the read counts in AGO1-IP to that of Input using Excel and only the top 3 enriched motifs of each frame were analyzed. The top 3 motifs for the first 3 frames (1st - 3rd, 2nd - 4th, 3rd - 5th) of the 7 library sets (7x3x3=63) were pooled together and the number of occurrence for each of these 3nt-motif was tabulated to envisage the most enriched 3nt-motif in the first 5 positions of mRNA fragments. Subsequently, every possible pair of nucleotides (1st and 2nd, 2nd and 3rd, 1st and 3rd) among these 63 3nt-motifs was tabulated to observe the trend for dinucleotides, in particular, 1st G – 2nd A, 2nd A – 3rd C, 1st G and 3rd C.

AGO1-IP	Input	Tissue / Cell Line	References
GSE24310	GSE24545	Ovary; WT	Okamura et al., 2011 Berezikov et al., 2011
GSE24311	GSE25570	Ovary; Ago2[414] mutant	Okamura et al., 2011
GSE24316	GSE24544	Ovary; Dicer-2[L811fsX] mutant	Okamura et al., 2011
GSE24313	GSE25570	Ovary; R2D2[1] mutant	Okamura et al., 2011
PRJNA192603	GSE13679	S2R+ FLAG-HA-AGO2; WT	Okamura et al., 2013 Flynt et al., 2009
GSE18806	GSE18806	Head; WT (2 sets)	Ghildiyal et al., 2010 Ameres et al., 2010

Table 2.3.1: Libraries analyzed for AGO1-loaded mRNA fragments

The BioProject accession number, tissue/cell line and reference for each library are summarized in this table.

3. RESULTS

3.1 HISSA development

As mentioned in Section 1.3, our very first aim is to develop HISSA (Section 1.4.1) so that we can test our hypothesis. In the process of developing HISSA, we faced several conundrums as outlined in **Figure 3.1** and have addressed them as elucidated in Section 3.1.1 - Section 3.1.4.

Figure 3.1: Conundrums addressed during HISSA development

As HISSA is novel, we had to find the best approach for each step of the assay. Therefore, we identified the potential issues that we would encounter when implementing HISSA for our study. These issues are summarized in the above questions and we did a series of parameters testing and optimization to address them.

3.1.1 5' end of 20nt *mir-317* loop sequence to be randomized for HISSA

In Okamura et al., 2013, AGO1 was shown to have less variety of miRNA loops loaded as compared to AGO2 and this motivated us to further investigate the sequence selectivity (if any) by AGO1. Therefore, in deciding on the oligo length, sequence and region to randomize, we chose *mir-317* loop and *mir-34* loop as the core sequences to start with, as they were the only 2 loops well loaded to fly AGO1.

First, for selecting the oligo length, *in-vitro* loading assay using S2R+ FLAG-HA-AGO2 cell lysate was carried out for 4 γ -³²P-5'-labeled oligos (Section 2.1.1; Table 2.1.1). They included *mir-317* loop (20nt), *mir-34* loop (27nt), *mir-317* loop-long (elongated version of *mir-317* loop; 27nt) and *mir-34* loop-short (truncated version of *mir-34* loop; 20nt). Subsequently, IP and autoradiography was done to visualize the amount of oligo loaded to AGO proteins (Section 2.1.2). As shown in **Figure 3.1.1**, the ratio of AGO1:AGO2 loading was much lower for oligo of shorter length (20nt) when compared to longer (27nt) oligo of the same sequence backbone. We chose to further work on 20nt length because any favorable change in sequence is most likely to cause a huge increase in AGO1 loading for oligo of non-optimal 20nt length. This was based on the assumption that at optimal length (if any), a much higher percentage of sequences could be well loaded to AGO1 and thus difficult to detect significant loading changes.

Second, in order to determine the sequence and region to randomize, we further carried out *in-vitro* loading of 4 more oligos (Section 2.1.1; Table 2.1.1). These 4 oligos were derived by swapping the 5' and 3' ends of *mir-317* loop and *mir-34* loop-short with each other. Interestingly, only *mir-317* loop (*mir-34* 5') and not *mir-317* loop-short (*mir-34* 3') showed opposite changes in AGO1 and AGO2 loading when compared to *mir-317* loop. Both *mir-34* loop-short (*mir-317* 5') and *mir-34* loop-short (*mir-317* 3') showed increase in AGO2 and AGO1 loading when compared to *mir-34* loop. The bidirectional changes in AGO1 and AGO2 loading upon swapping the 5' end of *mir-317* loop suggested that 5' end of *mir-317* loop sequence could be playing a significant role in influencing differential

sequence sorting between AGO1 and AGO2 instead of affecting general loading efficiencies of the two AGOs. In view of these findings, we decided to start off by randomizing 5' end of 20nt *mir-317* loop sequence for HISSA.

Figure 3.1.1: 5' end of 20nt *mir-317* loop alluded to sequence specificity

In deciding the length, sequence and region to randomize for the oligo, *in-vitro* AGO loading was carried out on 8 individual oligos. Each Input lane contained 100cpm of ³²P-labeled oligo for normalization purpose and IgG lane served as negative control for IP. Since *mir-34* loop-short (20nt) and *mir-317* loop (20nt) were more poorly loaded to AGO1 as compared to *mir-34* loop (27nt) and *mir-317* loop-long (27nt) respectively, 4 other 20nt oligos were designed to study effect of sequences on AGO1 loading. As such, the 5' and 3' ends of *mir-34* loop-short and *mir-317* loop were swapped. *mir-34* loop-short (*mir-317* 5'), *mir-34* loop-short (*mir-317* 3') and *mir-317* loop (*mir-34* 3') showed increase in both AGO2 and AGO1 loading. On the other hand, for *mir-317* loop (*mir-34* 5'), a decrease in AGO2 loading and an increase in AGO1 loading were observed. This bi-directional change suggested 5' end of 20nt *mir-317* loop sequence as a contributor to sequence specificity (if any) in AGO1 loading.

3.1.2 *In-vitro* loading assay reflects reliable AGO loading preference

The parameters set for HISSA randomized oligo in Section 3.1.1 would only be valid if *in-vitro* loading assay of single-stranded RNAs is reliable. However, this assay was relatively new and had not been experimentally proven to give consistent results. Furthermore, HISSA relies heavily on this assay as described in Section 1.4.1 and would be dire if we proceeded without validating its feasibility in reflecting reliable AGO loading preference.

Therefore, we did *in-vitro* loading (Section 2.1.1 and Section 2.1.2) triplicates for all 8 oligos used in Section 3.1.1 (Table 2.1.1). As shown in **Figure 3.1.2a**, the deviation for triplicates were small, indicating that *in-vitro* AGO loading of single-stranded RNAs is highly reproducible.

Nevertheless, being reproducible did not exclude the possibility that the AGO loading variation observed among the different oligos in Section 3.1.1 were purely due to differential degradation rates and not AGO sequence preference. This is because single-stranded RNAs are highly unstable molecules and are vulnerable to degradation in the cell lysate containing excessive ribonucleases. During *in-vitro* loading, those that got degraded more rapidly might be more poorly loaded to AGOs.

To tackle this issue, we conducted degradation assay (Section 2.1.3) for all the 8 oligos and also oligos that we used later on in Section 3.2.3 (Table 2.2.3). Briefly, in degradation assay, total RNA was extracted at different time points after introducing the γ -³²P-5'-labeled oligos into the lysate for loading. Radioactive signals at different time points were quantified and half-lives were calculated. As illustrated in **Figure 3.1.2b**, extremely poor correlation between loading efficiency and oligo half-life was observed for both AGOs as indicated by the R^2 values. This means that each loading signal observed in Section 3.1.1 was not directly correlated with the amount of oligo remaining in the lysate at any point in time and other factors such as sequence preference are most likely to be involved. These findings supported the feasibility of incorporating *in-vitro* loading assay as part of HISSA in showcasing reliable AGO loading preference.

Figure 3.1.2a: *In-vitro* AGO loading data is highly reproducible

In-vitro loading assay was carried out in triplicates for all of the 8 oligos used in Section 3.1.1 and radioactive signals in AGO-IP samples were normalized to that of Input. Each individual bar shows the average of triplicates for a single oligo. The error bar represents the standard deviation of the triplicates. As depicted, majority of the error bars are minute, therefore indicating that *in-vitro* loading data is highly reproducible.

Figure 3.1.2b: *In-vitro* AGO loading not affected by degradation rates

‘AGO loading score’ refers to the normalized AGO-IP to Input value obtained from *in-vitro* loading assay; high score means RNA sequence is efficiently loaded to AGO. Half-lives of the corresponding oligos were calculated from the degradation assay data; short half-life means RNA sequence degrades fast in the cell lysate. Plotting the AGO loading score against the half-life of individual oligos (every single colored circle) revealed very poor correlation between AGO2 loading (top)/AGO1 loading (bottom) and RNA degradation rates. R^2 values are less than 0.2 for both AGO2 and AGO1. In other words, an oligo that degraded slowly during *in-vitro* loading may or may not be better loaded to AGO as compared to oligo that degraded rapidly.

3.1.3 Library biases and errors reduced for data accuracy

Theoretically speaking, HISSA reveals AGO loading preference based on the output of *in-vitro* loading assay but at a high-throughput scale. Notwithstanding, what is going to be displayed at the end of the HISSA procedure would be affected by the intermediate library construction and sequencing steps (Section 1.4.1). This is especially so when these two intermediate steps are reported to give high biases (Baran-Gale et al., 2015). In other words, inaccurate HISSA data that differs much from that of *in-vitro* loading assay might be obtained despite how robust the latter might be.

The biases in the library construction and sequencing steps are not a huge issue if the IP sample and the Input (Section 2.2.1) are co-currently constructed and sequenced using the same platform. They should be cancelled off when we normalize the IP values to that of the Input (Section 2.2.2). However, the only concern was that during PCR, these biases would be subjected to PCR errors and subsequent sequencing faults may further build on these PCR artefacts, giving more errors. These errors are usually random and sporadic and may not be cancelled off during normalization of values to Input (Section 4.1). Since each sequencing platform has its own biased error rates and it was beyond our scope to prove which particular platform gives most accurate results, we focused on keeping the biasness in the library construction step as small as possible and also eliminating as much PCR/sequencing errors as possible.

To reduce biases in the library construction, we adopted a modified protocol (Zhang et al., 2013) described in Section 2.1.4. To assess if the modified protocol gives less biases as compared to the conventional protocol, we added equal amount of 10 spike-in oligos during library construction. Because the molar concentrations of the 10 oligos in the spike-in solution were equal, in theory, the protocol that gives less biases should cause less deviation among the spike-in oligo sequences within each library. As seen in **Figure 3.1.3a**, the average deviation was indeed smaller in modified protocol as compared to conventional protocol. However, the reduction was not huge as the improved protocol may have introduced new biases towards certain sequences.

To minimize effects of PCR and sequencing errors, we altered the 5' linker (RA5) used in the modified protocol above so that it harbored 12 randomized nucleotides instead of just 2 (Section 1.4.2; Section 2.1.4; Table 2.1.4). The probability of getting the same 5' linker sequence ligating to the same oligo sequence should be much smaller than 1/16,000,000 ($4^{12} = 16,777,216$). Hence, if any reads in the HTS data were to have the same linker sequence and same insert sequence (insert sequence: sequence derived from 5' end randomized *mir-317 loop* oligo), they are most probably amplified from the same molecule during PCR and we could regard them as a single read while keeping record of the read counts for error elimination purpose (see below).

Importantly, sequences having the same linker sequence but different insert sequences would suggest that they were derived from PCR/sequencing errors of one another. We could remove these error reads with a newly generated computational pipeline as illustrated in **Figure 3.1.3b** and elaborated in Section 2.1.5. This sophisticated computational scheme together with the modified library protocol help reduce artefacts and boost accuracy of HISSA.

Figure 3.1.3a: Biases in HTS libraries reduced with modified protocol

In testing the validity of the modified protocol, 10 spike-in oligos of equal amount were added to the samples before the library construction. The standard deviation of these oligos was slightly lower in libraries constructed using modified protocol as compared to those constructed according to the conventional protocol. This means that the biases are indeed reduced when modified protocol was used.

Newly generated error elimination pipeline

Figure 3.1.3b: PCR/sequencing errors elimination pipeline

At the end of HISSA, each HTS read contained part of the 5' linker sequence where the 12 randomized nucleotides resided (blue) and also the insert sequence (pink) (insert sequence: sequence derived from 5' end randomized *mir-317 loop* oligo). The very first step in the pipeline was to group reads that have the same 5' linker sequence together and extract out 'non-error' reads. The insert sequences of the 'non-error' reads were not within 2 mismatches (2mm) of other insert sequences in the same group. Subsequently, the remaining sequences within each group were sorted in descending order according to their read counts. The sequence with the highest read counts was taken as the 'major' read and used to eliminate (red cross) sequences that have up to 4 mismatches (4mm) from itself. After elimination, the 'major' read was stored away and removed from the group. The remaining reads again get sorted, 'major' read selected and erroneous reads eliminated. This process was repeated until all reads were gone in every single linker group. The final output of this error elimination pipeline was obtained by pooling all 'non-error' and 'major' reads from all linker groups together (see Section 2.1.5 for detailed procedure and rationales).

3.1.4 Oligo sequences enriched with elimination of endogenous RNAs

While strategies discussed in Section 3.1.3 enable us to have less skewed and more accurate results, they do not guarantee a successful study. Reason being, the small RNA extracted from the pulled-down AGO complexes is flooded with large amount of endogenous siRNAs, miRNAs and transposons. It is almost 0% chance that we could detect traces of the different *mir-317 loop*-backbone sequences derived from the introduced oligo, not to mention doing an in-depth analysis of them (**Figure 1.4.2**). Hence, there was a need to improve the library specificity towards the oligo of interest. Please note that detailed protocol of the experiments done in this section can be found in Goh and Okamura, 2018.

We started off in optimizing the RT-PCR step (last step) in the modified library protocol (Section 1.4.2; Section 2.1.4), in hope that we could specifically amplify all sequences that harbored the *mir-317 loop* backbone sequence (i.e. randomized sequences derived from oligo of interest). To directly visualize the specificity effect of RT-PCR conditions tested, we used a modified *mir-317 loop* oligo that contained a BspHI restriction site (Table 2.1.6). In that way, we could easily estimate the specificity by looking at the ratio of digested (specific; ~70bp

and ~80bp) : non-digested (non-specific; ~140bp) PCR products without going through time-consuming IP and sequencing procedure (Section 2.1.6; **Figure 3.1.4a**).

Among 18 different combinations of RT-PCR primers and annealing temperatures tested, only 3 most promising combinations are discussed here. Fast-forwarded, the best reverse PCR primer designed, was the one that have 9nt base-pairings with the 3' end of *mir-317 loop* sequence and the optimal temperature for primer annealing was at 66°C (Section 2.1.6; data not shown). We then tried 3 different RT primers in combination with this optimized PCR primer (Section 2.1.6). **Figure 3.1.4b** (top panel) shows that reducing the number of base-pairings with the 3' end of *mir-317 loop* sequence from 9 to 6 in the RT primer increased product specificity, as indicated by a larger ratio of digested: non-digest products. Shortening the region base-pairing to 3' linker from 14nt to 8nt while keeping 6nt base-pairing to *mir-317 loop* sequence resulted in even better specificity of ~50%. Ideally, increasing the RT temperature also helps in improving specific base-pairing between the RT primer and oligo sequence. However, the highest temperature we could use was 58°C because higher temperatures seemed to affect both efficiency and specificity (**Figure 3.1.4b; bottom panel**).

Nonetheless, 50% specificity was still not sufficient especially when it was affected by PCR inconsistency and could drop to ~10%. Therefore, in order to further enhance the library specificity, we moved upwards to the 3' linker ligation step, which is the first step in the modified library protocol (Section 1.4.2; Section 2.1.4). We replaced the 3' linker ligation step catalyzed by truncated RNA ligase with splinted-ligation. In splinted-ligation, bridge oligo that is complementary to both 3' linker and oligo of interest is used for T4 DNA ligase activity, hence conferring oligo-specificity in the ligation step.

We tested splinted-ligation specificity using 3 different bridges with reference to the protocol published by Malone et al., 2012 (Section 2.1.7). The RA3 linker was 32P-5'-labeled so that ligated products (~40nt) could be detected via radio-

autography. In order to make sure that the ~40nt radioactive bands were produced as a result of true ligation events and not random background effects or artefacts from the bridge/linker itself, we set up individual negative controls that contained RA3 linker and respective bridge but not oligo (**Figure 3.1.4c**, oligo (-) lanes). The specificity of each bridge was therefore the ratio of (oligo+ signal)/(oligo-negative control signal). Similar to the concept of RT primers (above), as we reduced the number of base-pairings with the 3' linker, the specificity increased while efficiency decreased. However, it is important to note that reducing too many base pairs will lead to both inefficiency and non-specificity. The optimal bridge was therefore the one with the highest specificity and moderate efficiency - 8nt linker+15nt oligo. As the name suggests, this bridge had 8nt base-pairings with the 3' linker and 15nt base-pairings with the oligo (**Figure 3.1.4c**).

Now that we have attained both the optimal RT-PCR primer and bridge for 3' linker ligation, we went on to test the specificity effects of both methods combined. As like before, *mir-317 loop* oligo that contained the BspHI restriction site was used for this purpose (Section 2.1.8; Table 2.1.6). It is no doubt that the library specificity was significantly enhanced: While digested band could hardly be detected before optimization (**Figure 3.1.4b; RTP-14nt linker+9nt oligo lanes**), the ratio of digested (specific) to non-digested (non-specific) products became 9:1 upon combining the modified library protocol with oligo-specific RT-PCR and splinted-ligation discussed in this section (**Figure 3.1.4d**; Section 2.1.8). In other words, with this optimized protocol, 90% of the library reads would be the sequences derived from oligo of interest (5' end randomized *mir-317 loop*) and subsequent analysis could be accomplished with much informative insights.

Strategy for examining library specificity

Figure 3.1.4a: Strategy for examining library specificity

As 18 pairs of RT-PCR primers were tested with multiple conditions, it was not feasible and economical for us to carry out numerous rounds of *in-vitro* loading, IP, library construction and sequencing just to check if the library specificity improved with those parameters. Therefore, instead of doing IP, estimated amount of endogenous small RNAs (black) were directly fractionated from total RNA and ^{32}P -labeled mir-317 loop oligo (pink) harboring BspHI site (red) was added to it. This pool of RNA was then used for library construction with various RT-PCR primers and conditions. Instead of sequencing, the PCR products were subjected to BspHI digestion. If the RT-PCR primers and conditions were good in conferring specificity to the library, then the ratio of digested PCR products (contained BspHI site; oligo-derived; specific) to non-digested PCR products (no BspHI site; endogenous small RNA; non-specific) should be large. This figure is adapted from Goh and Okamura, 2018.

Figure 3.1.4b: Oligo-specific RT-PCR primers improved library specificity

To confer specificity at the RT-PCR step, RT primer and reverse PCR primers were designed to form base-pairings with the 3' linker and the randomized oligo-derived sequence. In testing the 3 RT primers shown in this figure, RPI-9nt oligo PCR primer was used throughout. When we reduced the base pairs between the RT primer (grey) and oligo (pink) from 9nt to 6nt, the specificity improved significantly as seen from the larger ratio of digested PCR products: non-digested products (top panel). Further reduction of the number of base pairs between RT primer and the 3' linker from 14nt to 8nt while keeping 6nt base-pairing with the oligo sequence, gave even better specificity. 58°C was chosen as the optimal RT temperature for 8nt linker+6nt oligo primer because increasing the temperature beyond 58°C affected the specificity of RT reaction as seen from the decreasing ratio of digested PCR products : non-digested products. Furthermore, the efficiency was also affected as evident from the reduction in the intensity of bands (bottom panel). DMSO: Dimethyl sulfoxide; reagent added to reduce secondary structures of DNA during PCR.

Figure 3.1.4c: Oligo-specific bridge conferred 3' linker ligation specificity

The specificity of the splinted ligation reaction can be judged by looking at the ratio of ~40nt band in the 'oligo+' lane: 'oligo-' lane. The larger the ratio, the more specific was the ligation. Similar to the concept of RT primer optimization, reducing the number of base pairs between splinted ligation bridge (navy blue) and 3' linker (green) from 14nt to 8nt improved splinted ligation specificity significantly. In addition, too little base pairs between the bridge and the oligo (pink) caused inefficient ligation reaction to take place (8nt linker+6nt oligo). The optimal bridge is therefore 8nt linker+15nt oligo bridge. Asterisk (*) indicates artefacts. This figure is adapted from Goh and Okamura, 2018.

Figure 3.1.4d: Oligo-specific RT-PCR and splinted-ligation combined

To examine the final library specificity attained by combining the optimal RT-PCR and splinted-ligation conditions (**Figure 3.1.4b,c**), the *mir-317 loop* oligo carrying BspHI site was used again. In the first step of library construction, 3' linker was splinted-ligated to the RNAs using 8nt linker+15nt oligo bridge. In the last step, RT was carried out with optimal RT primer (8nt linker+6nt oligo) at 58°C. Optimal reverse PCR primer (RPI-9nt oligo) was primed at 66°C during PCR. The PCR products were digested with BspHI. As seen, the estimated library specificity was ~90%. This figure is adapted from Goh and Okamura, 2018.

3.2 Hidden rule in AGO loading of single-stranded RNAs

With all issues outlined in **Figure 3.1** solved and optimal conditions summarized in **Figure 3.2** obtained, our first aim of developing HISSA was achieved. Hence, we moved on to address our second aim, which was none other than using HISSA to test if our hypothesis of “sequence specificity acts as a hidden rule in AGO loading of single-stranded RNAs” is true (Section 1.2.1). We were interested to identify signature sequence motif(s) in the loading and the part(s) of a single-stranded oligonucleotide that is involved in the sequence discrimination.

Figure 3.2: Optimal parameters and conditions for HISSA

The potential issues outlined in **Figure 3.1** were successfully addressed and all steps in HISSA were optimized. The optimal parameters and conditions for each step are summarized in this figure.

3.2.1 Critical setup + effective analysis pipeline = Credible HISSA

To test our hypothesis, we constructed 3 HISSA library sets using optimal parameters and conditions obtained in Section 3.1 (**Figure 3.2**). These 3 library sets differed by the number of randomized nucleotides at the 5' end of the *mir-317 loop* oligo - 5nt, 8nt and 12nt. They are hereby referred as 5NN, 8NN and 12NN respectively. Each set comprised of 7 samples: AGO1 IP, FLAG (AGO2) IP, Rabbit IgG, Mouse IgG, AGO1 Input, FLAG Input and RAW (**Figure 3.2.1a**; Section 2.2.1). Important concepts underlied all samples as they are designed to facilitate downstream computational data processing and analysis (**Figure 3.2.1b**; Section 2.2.2).

Rabbit IgG and Mouse IgG serve as negative controls for AGO1 IP and FLAG IP respectively. Importantly, the usage of these negative controls depends on how HISSA data is being analyzed. If absolute readcounts of individual oligo-derived sequences are being compared in AGO-IP, we can easily use IgG samples to confirm these sequences as AGO-loaded species and not just background binding to beads. The readcounts of the respective sequences should be much lower in IgG samples if they are really loaded to AGO proteins in AGO-IP. On the other hand, if we are looking at the percentage of a particular nucleotide at each randomized position (**Figure 3.2.1c** and **Figure 3.2.2b**), it is unclear what we should expect for the IgG samples. Reason being, free-lying single-stranded RNAs derived from the randomized oligo are probably fully degraded during IP. The background binding in IgG samples most likely come from those RNAs that are stably loaded in AGOs. In that case, IgG would not serve as good negative controls as RNA sequence composition in IgG would be highly similar to that in AGO-IP.

RAW contains pure randomized oligo without any endogenous small RNAs (**Figure 3.2.1a**). This sample is useful in determining if sequences with abnormally high read counts in the IP sample were genuinely enriched or just artefacts derived from the randomized oligo. In addition, RAW would be effective in identifying cross contamination of randomized sequences among the 3 different HISSA experiments. For example, if there is a contamination of

5nt randomized *mir-317* loop oligo in 8NN HISSA samples, sequences ending with 'TGAAATGCAAGCAAG' (derived from contamination) will be of much higher read counts than sequences ending with 'XGAAATGCAAGCAAG' (X refers to any nucleotide other than T) but of lower read counts than sequences ending with 'AATGCAAGCAAG' (derived from 8nt randomization of oligo). Elimination of these artefacts and contaminants is one of the cardinal steps in HISSA computational analysis pipeline (**Figure 3.2.1b; right panels**).

Just as what **Figure 3.2.1a** depicts, randomized oligo is added only after IP and RNA extraction so that Input contains both endogenous AGO-loaded small RNAs and randomized oligo sequences. Not only that the endogenous small RNA population is required for determining PCR/sequencing error elimination parameters (Section 2.1.5 and Section 2.2.2; **Figure 3.2.1b;right panels**), but also essential for mimicking the conditions in IP samples so that biases during library preparation and sequencing steps are kept similar. Furthermore, since the oligo does not enter the lysate and get degraded, Input represents the initial pool of randomized sequences available in the lysate at time 0 during *in-vitro* AGO loading. Normalizing IP to the Input at the end of data analysis would therefore give indubious sequence/motif preference by AGOs (**Figure 3.2.1b**).

After 5NN, 8NN and 12NN HISSA library sets were successfully constructed, they were sequenced with illumina HiSeq 4000 and subsequently analyzed according to the pipelines in **Figure 3.2.1b** (Section 2.2.2).

Consistent with previous studies (Section 1.1.8), positional enrichment analysis of our 12NN HISSA data revealed G-rich sequences as favorable loading substrates of AGO2. This conclusion came with the observation that 'G' (black bar) is the only nucleotide to have positive $\log_2(\text{AGO2-IP/Input})$ value throughout the entire randomized region (**Figure 3.2.1c; top panel**). In fact, 5NN and 8NN HISSA showed exactly the same AGO2 loading trend (data not shown). On the other hand, AGO1 was found to have U nucleotide (red bar) very strongly enriched in the first position (**Figure 3.2.1c; bottom panel**) and this is unmistakably coherent with previous findings too (Section 1.1.6). These reproducible findings supported the effectiveness of HISSA experimental setup

and computational analysis pipeline. Most importantly, they proved HISSA as a credible assay for studying AGO loading preference.

Figure 3.2.1a: HISSA overall experimental setup

For every randomized oligo of interest (pink), a HISSA set of 7 samples was designed. Out of the 7 samples, 4 of them were fundamental IP samples, namely FLAG-IP, AGO1-IP, Rabbit IgG and Mouse IgG. IgGs served as negative controls for the IP procedure. 2 Inputs were needed for normalization whereby each Input comprised of endogenous small RNAs (black) extracted from FLAG or AGO1 IP, but unlike IP sample, the randomized oligo was added only after the RNA extraction. This way, the inputs represent the initial pool of randomized sequences available in the lysate at time 0 during *in-vitro* AGO loading. RAW sample did not contain any endogenous small RNA as it was purely randomized oligo itself and was essential for troubleshooting purposes. This figure is adapted with modifications from Goh and Okamura, 2018.

Figure 3.2.1b: HISSA computational analysis pipelines

This figure shows the HTS data processing and analysis pipelines for 5NN HISSA and 8NN/12NN HISSA. Two different pipelines were needed because 8NN and 12NN HISSA had many more sequence combinations derived from the randomized oligo than 5NN HISSA. Given the fixed sequencing depth, 5NN HISSA had all 1024 sequences in both Input and AGO1-IP samples but 8NN and 12NN HISSA had different individual sequences appearing in the Input and AGO1-IP samples. Due to this reason, it was not technically feasible to do a direct motif enrichment like in the case of 5NN HISSA. Therefore, positional enrichment and pairwise analysis were done. In addition, error elimination was needed for 8NN and 12NN HISSA but not 5NN HISSA as explained in Section 4.1.

Figure 3.2.1c: 12NN HISSA results consistent with previous findings

In positional enrichment analysis, the proportion of a particular nucleotide in each position of the randomized region was tabulated for IP and Input libraries. The IP values were then normalized to that of Inputs. Positional enrichment analysis of 12NN HISSA data showed that AGO2 prefers G-rich sequences as 'G' (black bars) was highly enriched in all of the 12 positions of the randomized region (top panel). On the other hand, AGO1 strongly prefers 'U' (red bar) in the first position (bottom panel). These two findings are exactly consistent with previous reported findings.

3.2.2 Sequence specificity in AGO1 loading: ‘GAC’ in 1st - 5th positions

Credible assays tend to produce compelling results. At least, HISSA did. Despite 12NN HISSA showing only 5'U enrichment for AGO1, 5NN and 8NN HISSA divulged the hidden rule in AGO1 loading of single-stranded RNAs - sequence specificity.

5NN HISSA showed that only 45 out of 1024 sequence combinations were enriched in AGO1-IP. Intriguingly, (Guanine-Adenine-Cytosine) ‘GAC’-bearing sequences were found to make up a huge percentage of these enriched population as seen by the dominating red lines (GAC-bearing sequences) in the positive scale (Section 2.2.2; **Figure 3.2.2a; top panel**). A further analysis of these 45 sequences revealed ‘GAC’ as the most enriched motif in the 1st – 3rd positions (1st frame) but became weaker in the 2nd – 4th positions (2nd frame) and eventually the weakest in the 3rd- 5th positions (3rd frame) of the 5nt randomized region. The enrichment of ‘GAC’ in the 1st frame was so strong that it masked away the ‘GAC’s in the other two frames as illustrated by the sequence logo (Section 2.2.2; **Figure 3.2.2a; bottom panel**).

The overpowering ‘GAC’ motif preference in the 1st frame of single-stranded RNAs was further evident in the 8NN HISSA AGO1-IP/Input profile. As seen in the positional enrichment analysis, ‘G’ (black bar) in the first position, ‘A’ (green bar) in the second position and ‘C’ (blue bar) in the third position were enriched (Section 2.2.2; **Figure 3.2.2b; top panel**). However, such simple analysis could not tell us if these individual nucleotides are linked with each other in forming AGO1-enriched ‘GAC’ motif. In order to check if they were statistically well associated with each other in giving positive AGO1 loading, we carried out pairwise analysis.

In this analysis, we used two rounds of statistical tests to determine how strongly a particular di-nucleotide combination (each nucleotide from a different position of the randomized region) affected AGO1 loading. In other words, the better the di-nucleotide combination preferred in AGO1 loading, the more positive the value was. As shown in the heat-map, the ‘G’ in the first position

was statistically well-associated with 'A' in the second position in giving strong AGO1 loading, thus the very bright cyan color. 'A' in the second position in turn combined with 'C' in the third position to produce good AGO1 loading. Furthermore, 'C' in the third position together with 'G' in the first position were also co-related with positive AGO1 loading. Putting these 3 pairs of nucleotide associations together, confirmed the strongest 'GAC' motif enrichment in the 1st frame (Section 2.2.2; **Figure 3.2.2b; bottom panel; small yellow boxes**). Similar but weaker trends can be observed in the 2nd frame (**Figure 3.2.2b; bottom panel; small blue boxes**) and the 3rd frame (**Figure 3.2.2b; bottom panel; small green boxes**). Again, consistent with 5NN HISSA, the 'GAC' in the 3rd frame was the weakest of all.

Another interesting point to note is that within the 3 'GAC' frames, the average ranking of dinucleotides enrichment was 'GA' (1st and 2nd) > 'AC' (2nd and 3rd) > 'GC' (1st and 3rd) as indicated by the variation in the cyan intensities.

Figure 3.2.2a: 5NN HISSA - 'GAC' motif preferred in AGO1 loading

A direct normalization of individual sequence read counts in 5NN HISSA AGO1-IP to that of AGO1 Input revealed 45 out of 1024 sequences got enriched in AGO1-IP. Out of this small portion of enriched sequences, majority of them were having a 'GAC' motif (red) (top panel). Further analysis on these 45 sequences divulged 'GAC' as best enriched in the 1st – 3rd positions (GACXX; orange) as GACXX sequences made up ~18% (8/45) of the enriched population. The enrichment became weaker in the 2nd – 4th positions (XGACX; blue) and weakest in the 3rd -5th (XXGAC; green). Sequences not containing 'GAC' (Others; grey) are found at the lower end of the spectrum. This opposes the current consensus of the classic 5'U preference in AGO1 loading.

Figure 3.2.2b: 'GAC' motif statistically enriched in the 1st – 5th positions

In positional enrichment analysis, the proportion of a particular nucleotide in each position of the randomized region was tabulated for IP and Input libraries. The IP value was then normalized to that of Input. Positional enrichment analysis of 8NN HISSA data showed that 'G' (black bar), 'A' (green bar) and 'C' (blue bar) were most preferred in the first, second and third positions of the randomized region (top panel). Subsequently, pairwise analysis was done to find out if these three highly preferred nucleotides were associated with one another in giving strong AGO1 loading (bottom panel). Positive impact on AGO1 loading is indicated by cyan while negative impact is indicated by yellow. As bright cyan can be observed for 1st 'G' - 2nd 'A', 2nd 'A' - 3rd 'C' and 1st 'G' - 3rd 'C' (small yellow boxes), 8NN HISSA further confirmed 5'GAC as the most preferred motif by AGO1. Enrichment of 'GAC' was weaker in the 2nd - 4th positions (small blue boxes) and weakest in the 3rd-5th positions (small green boxes).

3.2.3 5'GAC motif successfully validated with *in-vitro* loading assay

As mentioned in Section 1.1.8, 5'U was found to be the trademark for AGO1-loaded small RNAs. Therefore, a prominent 'GAC' enrichment in the 1st frame (i.e. 5'GAC) was totally out of anticipation. To make sure 5'GAC enrichment is not a result of deep sequencing artifacts or computational analysis errors, we went on to validate the computational results experimentally using *in-vitro* loading assay (Section 2.2.3).

First of all, out of the 5 random sequences chosen from 5NN HISSA, 2 of them that harbored 5'GAC (5'GACGAUGA-*mir-317 loop* and 5'GACAAUGA-*mir-317 loop*) were the most well loaded to AGO1. In addition to that, the computational loading rankings of the individual sequences were precisely reproduced by the *in-vitro* loading results (**Figure 3.2.3; left panels**). Secondly, 8NN HISSA pairwise analysis (**Figure 3.2.2b; bottom panel**) also showed that 'A' in the 7th position and 'C' in the 8th position were strongly co-related with 'G' in the first position for enhanced AGO1 loading. To test that, we mutated the 7th and 8th positions of the 2 oligos harboring 5'GAC to 'A' and 'C' respectively (5'GACGA7A8C-*mir-317 loop* and 5'GACAA7A8C-*mir-317 loop*). Interestingly, just like what the pairwise analysis indicated, this mutation indeed caused AGO1 loading to be enhanced for both sequences (**Figure 3.2.3; middle panels**).

Last but not least, to show that 5'G was indeed preferred over the classic 5'U, we mutated the 'G' in the first position of 5'GACGAUAC-*mir-317 loop* to a 'U' (5'UACGAUAC-*mir-317 loop*). Unexpectedly, this single mutation alone caused the AGO1 loading to decrease tremendously. In order to know if 5'G was preferred over 5'U only because it was in the 'GAC' context, we mutated the 'A' in the second position of 5'GACGAUAC-*mir-317 loop* to a 'U' (5'GUCGAUAC-*mir-317 loop*). True enough, AGO1 loading decreased significantly even though 5'G was present. However, to rule out the possibility that the 'A' itself being in the second position was important for loading, we mutated the 'A' in the second position of 5'UACGAUAC-*mir-317 loop* to a 'U' (5'UUCGAUAC-*mir-317 loop*) and we observed no significant reduction in AGO1 loading. Hence, this

indicated that the 'A' in the second position was detrimental to 5'GAC-harboring sequence because it disrupted the GAC motif and not because the nucleotide itself was important for AGO1 loading (**Figure 3.2.3; right panels**).

All in all, with HISSA, we have proved that our hypothesis of “sequence specificity acts as a hidden rule in AGO loading of single-stranded RNAs” is true. Particularly, 5'GAC motif is preferred over the classic 5'U in single-stranded RNA loading to AGO1.

Figure 3.2.3: 5'GAC motif successfully validated by *in-vitro* loading

To ensure that the 5'GAC motif shown by 5NN and 8NN HISSA was not a computational or sequencing error, *in-vitro* loading assay was carried out on non-randomized oligos to validate the computational data experimentally. The names of the oligos shown in this figure corresponds to the oligo sequences starting from the first position (e.g. 'GACGAUGA' is the first 8nts at the 5' end of the oligo). The black bar after the 8th position represents the rest of the full length *mir-317* loop sequence. The red letters and digits found in some of the sequences represent the mutated nts and positions with respect to GACGAUGA or GACAAUGA sequences. As anticipated, GACGAUGA and GACAAUGA sequences were the most highly loaded out of the 5 individual sequences randomly selected from the 5NN HISSA. Furthermore, the experimental AGO1 loading trend fitted very well with the computational ranking (left panels). Mutating the 7th and 8th positions of GACGAUGA and GACAAUGA indeed enhanced AGO1 loading as suggested by 8NN HISSA pairwise analysis (middle panels). Preference of 5'GAC motif over 5' U in AGO1 loading was apparent when mutating 'G' in the first position or 'A' in the second position of GACGAUAC to a 'U' caused loading to decrease tremendously. The 'A' in the second position is proven to be important only in the presence of GAC motif because mutating it in UACGAUAC did not result in further reduction in AGO1 loading (right panels). Hence, 5'GAC is genuinely a motif most preferred in AGO1 loading. Refer to Table 2.2.3 for the oligo sequences used. Some panels in this figure were cropped for better visualization. Refer to Appendix (Section 8.2) for original blots.

3.3 Biological evidence for HISSA findings

Now that we have successfully proven our hypothesis and identified 'GAC' as a novel sequence motif preferred in AGO1 loading, we started to tackle our last aim - Explore the relevance of our findings in the biological context.

3.3.1 'GAC', the topmost motif found in AGO1-loaded mRNA fragments

Short mRNA fragments in the cells were chosen to be the subject of analysis in this case because they are the only RNAs in the cell that are single-stranded, random and with a huge variety of sequence combinations. Such conditions are similar to the pool of randomized sequences we used in HISSA. We postulated that some of these mRNA fragments are loaded to AGO1 via the single-stranded RNA loading mechanism. If this is true, we should be able to observe 'GAC' enrichment in AGO1-loaded mRNA fragments, just like what we observed in HISSA.

Among all publicly available libraries, only 7 AGO1-IP library sets were eligible for this mRNA fragments analysis because most the library sets either do not have AGO1-IP or Input. These shortlisted 7 AGO1-IP library sets were mapped and subjected to mRNA fragments analysis. Briefly, reads that belonged to coding genes were extracted. However, those that lie within the loci of any annotated miRNA, siRNA and piRNA as well as repeatmasker regions were subsequently removed. Potential non-annotated siRNAs were eliminated by excluding all multi-mappers and 21nt reads. At the end of all elimination steps, the remaining reads were treated as single-stranded mRNA fragments. For each library, 3nts-shifting frames were extracted up to the 8th position (1st - 3rd, 2nd - 4th, 3rd - 5th, 4th - 6th, 5th - 7th, 6th - 8th) of the mRNA fragments and a total of 64 ($4^3=64$) 3nt-motifs with their respective read counts were obtained for each frame. The enrichment of 3nt-motifs in AGO1 were obtained by normalizing the read counts in AGO1-IP to that of Input. The top 3 enriched motifs of each frame were extracted for final analysis (**Figure 3.3.1a**; Section 2.3.1).

Intriguingly, we observed 2 trends highly consistent with the 5NN and 8NN HISSA results – 1) ‘GAC’ started appearing as one of the top 3 motifs from the 1st frame and stopped after the 3rd frame and 2) the 3rd frame was the weakest of all since it could only be found in 1 out of 7 libraries (**Figure 3.3.1b; top panel**).

Motivated by these observations, we went on to find out if ‘GAC’ is the most strongly enriched motif in the 1st – 5th positions by pooling the top 3 motifs of the first 3 frames for all 7 libraries together. In other words, there was a total of 63 top motifs (7 libraries x 3 frames x 3 top motifs) being tabulated (Section 2.3.1). Indeed, ‘GAC’ emerged as the motif that occurred the most number of times in the 1st – 5th positions (**Figure 3.3.1b; bottom panel**). Subsequently, every possible pair of nucleotides (1st and 2nd, 2nd and 3rd, 1st and 3rd) among these 63 3nt-motifs was tabulated to observe the trend for dinucleotides, in particular, 1st G – 2nd A, 2nd A – 3rd C, 1st G and 3rd C. Unexpectedly, we observed the trend of ‘GA’ (1st and 2nd) > ‘AC’ (2nd and 3rd) > ‘GC’ (1st and 3rd) which again tallied nicely with that of 8NN HISSA (**Figure 3.3.1b; bottom panel**; Section 3.2.2).

With high consistencies in all the trends among the 5NN HISSA, 8NN HISSA, mRNA fragments analysis and *in-vitro* loading validation, we are confident to conclude ‘GAC’ motif in the 1st – 5th nt region of single-stranded RNAs as highly preferred in AGO1 loading. In addition, it is important to note that HISSA is highly reliable in recapitulating cellular AGO loading.

Endogenous mRNA fragments analysis pipeline

Figure 3.3.1a: Endogenous mRNA fragments analysis pipeline

As small mRNA fragments in the cells are usually products of degradation, it is most likely that these single-stranded RNAs contain a variety of sequences that are random and not too skewed. For this reason, they serve as a good subject for analysis of AGO sequence specificity in the cellular context. This figure shows that 7 public available AGO1-IP library sets were used for the endogenous small mRNA fragments analysis. From the third step up to the second last step involved elimination of all possible cellular non-coding small RNAs. Since 8NN HISSA showed that GAC motif occurred in the 1st – 5th positions of single-stranded RNAs but not in the positions beyond that, 3nt-shifting frames were extracted from the 1st to the 8th positions of mRNA fragments. A total of 64 3nt-motifs was tabulated for each frame of each library set. The enrichment of 3nt-motifs in AGO1 were obtained by normalizing the read counts in AGO1-IP to that of Input.

Figure 3.3.1b: 'GAC' motif validated in AGO1-bound mRNA fragments

Highly consistent with HISSA results, 'GAC' motif in the 1st – 5th positions of mRNA fragments is the most preferred in *in-vivo* AGO1 loading 'GAC' emerged as one of the top 3 motifs in 2 out of 7 library sets for both 1st - 3rd positions ('GAC' in yellow) and 2nd – 4th positions ('GAC' in blue). Only 1 out of 7 library sets showed 'GAC' enrichment in 3rd – 5th positions ('GAC' in green). After the 5th position, none of the libraries had 'GAC' as one of the top 3 enriched motifs (top panel). After pooling the top 3 3nt-motifs in the 1st -5th positions of all library sets together (total of 63), 'GAC' (red bar) was found to be the most preferred motif as the number of occurrence was the highest (bottom panel). Further analysis of these 63 3nt-motifs revealed the trend of 'GA' (1st and 2nd positions) > 'AC' (2nd and 3rd positions) > 'GC' (1st and 3rd positions). This trend is again consistent with 8NN HISSA pairwise analysis.

4. DISCUSSION

4.1 Importance of optimization increases with randomization

In Section 3.1.3 and Section 3.1.4, strategies were introduced to reduce errors and increase oligo-specificity in HISSA libraries. While some measures may seem overkill, it is crucial to note that they become increasingly important as we increase the number of randomized nucleotides in the oligo of interest.

For instance, the error elimination pipeline (**Figure 3.1.3b**) was not essential for the 5NN HISSA data because the total possible sequence combinations derived from the randomized region were only 1024. Such small variation in the sequences probably only gave little biases during library preparation. This is because the efficiencies of 3'- and 5'- linker ligations should be similar for majority of the 1024 sequences since these sequences only differ slightly from one another, with a maximum of 5nt in the 5' end. Given the presence of little biases and all 1024 sequences were present in both AGO1-IP and Input libraries, the PCR/sequencing errors in AGO1-IP were most probably accountable by the Input if we assume the errors derived from each individual sequences were highly similar between the two libraries. In other words, errors would be proportional to the read counts of individual sequences and hence not a huge issue upon normalizing AGO1-IP to Input.

However, as we increase the randomization to 8 and 12, the number of sequence combinations surged up to ~65,000 and ~16,000,000 respectively. With such a huge range of sequences present, we would expect the linker ligation efficiencies to be much more skewed for the different sequences as the sequences differ largely from one another. In addition, with the sequencing depth more or less fixed, the number of different sequences appearing in the libraries were only a small percentage of the total possible sequence combinations. Having said that, the probability of having identical sequences in both Input and AGO IP libraries was even smaller. As the sequence populations in Input and AGO IP varied largely, the biases and type of errors were expected to be very different. Simply normalizing AGO IP to Input would not give credible

results because errors would not be proportional to the read counts. Only with computational error refinements (**Figure 3.1.3b**), the gap between Input and AGO-IP could be reduced to minimal and hence ensured the data accuracy.

Similarly, RT-PCR optimization may seem redundant when splinted-ligation appeared to play the main role in conferring oligo-specificity to 5NN HISSA library (**Figure 3.1.4b,c,d**). However, as we increase the randomization, the number of base-pairings between the bridge and the oligo was reduced. In turn, this reduction caused a decrease in ligation specificity. Since splinted-ligation is the first step in library preparation, a decrease in specificity meant that more endogenous small RNAs were brought forward to the RT-PCR step (last step). In that case, RT-PCR optimal conditions definitely played a more significant role in contributing to the specificity of the library.

4.2 Relation between AGO loading specificity and function

It did not come as surprising when we failed to detect any sequence specificity for AGO2 loading (**Figure 3.2.1c; top panel**). This is especially so when AGO2 was previously shown to be loaded with a variety of single-stranded miRNA loops unlike AGO1 (Okamura et al., 2013). This suggests that AGO2 is not governed by strict loading rules.

Not having specificity makes much sense for AGO2 because as mentioned in Section 1.1.1, its main function is to down regulate transposons and it is also known to serve as a defense mechanism against viral infection in flies (Li et al., 2002; Galiana-Arnoux et al., 2006; van Rij et al., 2006; Wang et al., 2006; Zambon et al., 2006). Given such function, AGO2 must have evolved in a way that it could efficiently accept large range of sequences that are complementary to all possible kinds of transposable elements and also the ever-changing viral genomes. Furthermore, fly AGO2 cleaves targets that are fully complementary to the guide strand (Section 1.1.3). This precise cleavage mechanism has already provided specificity in target repression and additional loading specificity might be redundant with respect to its function. Or rather, this precise cleavage mechanism would most probably not be able to stay through evolution if specificity is at play in the upstream loading step.

On contrary, it would only be puzzling if AGO1 loading only depends on RNA duplex structure and 5' terminal nucleotide as previously assumed (Section 1.1.5). The scenario one should consider: At any point of time, the cell has numerous random mRNA fragments that are resulted from ribonuclease activity on mRNA transcripts. If any single-stranded RNA could just load into AGO1 efficiently, how does the cell keep its gene expressions stable given that AGO1 acts through the guide strand seed region (2nd to 8th position) instead of full-complementary base-pairing like AGO2 (Section 1.1.2)?

Interestingly, the 'GAC' sequence motif we identified lies in the first 5 positions of single-stranded RNAs. This means that the 'AC' in the 2nd and 3rd positions (for the case of 5'GAC) or 'GAC' in the 2nd – 5th positions (for the case of

5'XGACX and 5'XXGAC; where X is any nucleotide) is part of the seed region (**Figure 3.2.2a,b**). Now, imagine that 5'GAC serves as the first criteria to efficient AGO1 loading and during target repression, the 'AC' of 'GAC' serves as the second criteria within the seed-region. This way, as compared to a sole 5'U rule in AGO1 loading, 5'GAC preference would mean less mRNA fragments loaded and even less genes being down regulated.

4.3 Understanding the two rules in loading – 5'GAC and 5'U

So, the next question is when does AGO1 loading prefers 5' 'GAC' to the classic 5' 'U'? A simple answer to this could be easily found in our 8NN HISSA pairwise analysis (**Figure 3.2.2b; bottom panel**). With reference to the first column of the heat-map, 'A's in many of the positions were having strong co-relation with 5'G in the first position for AGO1 enhanced loading. The requirement of 'A's in the backbone is further supported when 5NN HISSA showed 5'GAC enrichment but not 12NN HISSA (**Figure 3.2.2a; Figure 3.2.1c; bottom panel**). This is because *mir-317* loop has a sequence of 'UUGAAUGAAAUGCAAGCAAG' and having 5nt and 8nt randomization at the 5' end would cause oligo to lose only 2-3 of the fixed 'A' nucleotides. On the other hand, a 12nt randomization oligo loses 5 of these 'A's. The probability of having 'A'-rich sequences in the initial loading pool was therefore much higher in 5NN and 8NN as compared to 12NN.

If the above theory is true, there would be many possible models to explain why 5'GAC could be so well-loaded to AGO1 (see below) despite the fact that 5'G was found to have steric clash with the MID binding pocket residue (Frank et al., 2010). It is important to note that the below 3 models are highly postulated and based on the crude assumption that 'A'-rich backbone is required to avoid 5'G clashing with AGO1 residue. Future studies are highly essential to verify if 'A'-rich backbone is indeed a critical factor in 5'GAC loading before any model could be tested.

One postulation is that the 'A'-rich backbone affects the backbone geometry of the single-stranded RNA in the AGO protein, causing the orientation of the 5' end to be completely different and the 5'G no longer clash with any of the protein residue (Model 1). Another possibility is that A-rich backbone is required for binding by a loading factor involved and when the AGO protein is in complex with this factor, its entire conformation changes, allowing the loading of 5'GAC sequence (Model 2). A much more naïve model to explain is that 5'U preference is only meant for duplex-loading whereas 5'GAC is for single-stranded RNA loading. In other words, the 5'U preference we observed for 12NN HISSA could

be purely because the less 'A'-rich (i.e. more GC-rich) single-stranded RNAs base-paired with other single-stranded RNAs better in the cell and formed duplexes prior to loading. A-rich sequences in 5NN and 8NN HISSA could not effectively base-pair with other RNAs and thus get loaded via a completely different mechanism that prefers 5'GAC (Model 3). These models are illustrated in **Figure 4.3**.

Figure 4.3: Postulated models for 5'GAC-RNA loading to AGO1

This figure puts forward 3 postulated models to explain why 5'GAC can be strongly enriched in AGO1 even though 5'G was previously shown to clash sterically with the AGO residue. All 3 models are based on the assumption that 'A's are needed in the single-stranded RNA backbone in order for 5'GAC to be well loaded. Further studies are needed to understand if this theory holds before any of these models could be tested. A single AGO1 protein is represented by the entire set of 4 orange ovals. Each orange oval corresponds to one of the AGO main domains which include MID, PIWI, N and PAZ. The domain that the 5' terminal nucleotide is in contact with is the MID domain. In model 1, it is hypothesized that the 'A's affects the backbone geometry of the single-stranded RNA (blue) residing in AGO1. As a result, the 5' terminal G nucleotide (white) is being inserted into the binding pocket of MID domain at a different angle compared to the one with non-'A'-rich backbone and no longer clashes with the residue in the binding pocket. In model 2, an external loading factor (brown circle with '?') is involved. There is a possibility that the loading factor binds to the 'A'-rich region of the RNA prior to loading and during loading, AGO1 associates with this factor and changes its overall conformation. Because of this conformation change, the position of the 5'G along the MID domain is largely different from what it is known to be. In model 3, non-'A'-rich single-stranded sequences are thought to have higher tendency to form double-stranded structures with other single-stranded RNAs in the cell or lysate. Thereafter, 5'U is favorable merely because it is part of the default double-stranded loading mechanism. In the case of RNAs with 'A'-rich backbone, base-pairings are not as feasible and so get loaded via the single-stranded RNA loading mechanism. 5'GAC might be the sole signature of single-stranded RNA loading. The steric clash of 5'G might be true only for double-stranded RNA loading but not single-stranded RNA loading.

4.4 Assessing credibility of mRNA fragment analysis

In Section 3.3.1, after extracting out all mRNA reads, we went on to eliminate all annotated non-coding RNA reads, genes with staggered reads and all multi-mappers so as to remove all possible unannotated non-coding RNA reads (**Figure 3.3.1a**). However, we are aware of some potential issues this pipeline could bring.

Since the analysis pipeline in **Figure 3.3.1a** was highly stringent, the number of mRNA fragment reads obtained for each library was very small (~ 300 – 3000 reads). Furthermore, these libraries were not constructed with the randomized 5' linker we used for HISSA (Table 2.2.1b) so we could not eliminate any possible PCR or sequencing errors like we did for HISSA in **Figure 3.1.3b**. All of these downsides might have affected the accuracy of the analysis. For example, motifs that had only ~20 reads in the Input library would easily be highly enriched in the IP library due to fluctuations caused by biases during library construction or PCR and sequencing errors. On the other hand, motifs that had ~200 reads in the same Input library would have lower enrichment as errors-derived sequences would make up a smaller percentage of the population.

In order to counter this issue, instead of looking at the absolute enrichment of each motif, we looked at the number of times a particular motif appeared as top 3 in the 1st – 5th positions after combining top 3 motifs from all 7 libraries. This was based on the assumption that random fluctuations should be inconsistent across the 7 sets of libraries as they were prepared by different laboratories using different fly tissues/cell lines, library construction protocols and sequencing platforms.

We are confident that the results are credible because not only that we are able to observe multiple trends that are highly consistent with our HISSA results (**Figure 3.3.1b**; **Figure 3.2.2a**; **Figure 3.2.2b**) but also spot some similarities with another study. As mentioned in Section 1.2.1, Li et al., 2014 found that 'ACCAA' motif was bound by human AGO2. As seen in **Figure 3.3.1b**, both

'ACC' (part of ACCAA) and 'CAA' (part of ACCAA) were also among the top 3 motifs enriched in the 7 library sets. It is understandable that the enrichments of these motifs are not as strong as 'ACCAA' reported in that study because there bound to be differences in human and fly transcriptomes.

Perhaps another potential issue is that even though we tried all ways to eliminate all possible non-coding RNAs, there could be a very small population of them remaining. However, failure to remove all siRNA/primary piRNAs in the AGO1-IP libraries would most probably give us false negative (rather than false positive) 5'GAC enrichment because majority of these RNA species would possess a 5' 'U' and not a 5' 'G' (Section 1.1.5 and Section 1.1.8). Also, if it is due to enrichment of transposons, we would most probably observe motif with simple repeats (e.g. 'AAA' or 'GGG') rather than a 'GAC' (Vagin et al., 2006; Brennecke et al., 2007; Gunawardane et al., 2007).

4.5 HISSA applications and limitations

HISSA is not only useful in identifying sequence motif like what we have done in our study. Since there is much flexibility in the choice of oligo to use, there are many more ways we could exploit the system to yield meaningful findings.

In fact, HISSA could be further utilized to understand the mechanism behind 'GAC' preference in AGO1 loading. For example, in pursuit of the 'A'-rich backbone requirement, one way is to keep a fixed 'A'-rich region in-between two small randomized regions instead of having a stretch of 12 randomized nucleotides at the 5' end. If 'A'-rich backbone is indeed a requirement, this design should turn 5'U preference (in 12NN HISSA) to 5'GAC. Another example - if we want to test model 3 mentioned in Section 4.3, we could add a modified randomized oligo that forms complementary base-pairing with the fixed region of the 5nt and 8nt randomized *mir-317 loop* oligo. If the model holds, 5'GAC preference should be adversely affected or change to a 5'U preference.

Moreover, HISSA is not limited to the fly system. It can be applied to study AGO loading in other organisms such as human. As mentioned in Section 1.2.2, limelight is gradually shifting towards single-stranded RNAi in the recent years and the main obstacle now is the low efficacy resulting from fast degradation rate. One way of tackling this issue is to design the optimal sequence so that the single-stranded RNAi oligo gets loaded to AGO proteins efficiently before it gets degraded. The most effective way to get the optimal sequence is none other than finding out the sequence preference of human AGO2 (if any) with the use of HISSA.

Nevertheless, HISSA could also be involved in the screening of optimal single-stranded siRNA sequence against a particular target gene without any knowledge of human AGO2 specificity – Carry out HISSA with a randomized oligo harboring a backbone that is complementary to a repetitive sequence present in the target gene. Just by doing a single HISSA experiment, one could effectively narrow down the potential target sites in the gene by looking at the computational ranking of sequences. Or rather, a mega project that combines

findings of all these individual HISSA could be initiated to form a database. This database could be used to develop algorithms which predict the best single-stranded-RNAi sequence for every gene in the genome.

However, HISSA comes with various limitations. Firstly, there is always a maximum number of nucleotides that we could randomize because issues like library biases, errors and specificity get more complicated with the extent of randomization (Section 4.1). Secondly, with the current available techniques, it is difficult to randomize the 3' end of single-stranded oligo while attaining specificity for HISSA library. The main caveat lies in no commercially available enzyme that is able to splinted-ligate 5' end of RNA oligo to 3' end of RA5 RNA linker (in this study, T4 DNA ligase ligates 5' end of RA3 DNA linker to 3' end of RNA oligo). Thirdly, large amount of single-stranded oligo is needed at high temperatures (e.g. for mammalian Ago HISSA) because it gets degraded rapidly in the lysate during *in-vitro* loading. Lastly, the entire HISSA procedure is technically challenging, as these single-stranded RNAs are present in very little amount. Contamination of sample after splinted-ligation step easily leads to total loss of oligo-specific sequencing reads. Extra care has to be taken when preparing these libraries.

4.6 RNAi versus CRISPR: Which is better for therapeutics?

So, with an improved loading efficiency, is single-stranded RNAi going to be the best choice for therapeutics?

While single-stranded RNAi could be a good alternative to counter negative consequences brought by double-stranded RNAi (Section 1.2.2), it is most likely to be much less effective and versatile than the clustered regularly interspaced short palindromic repeats (CRISPR) system. In engineered CRISPR system, a single guide RNA (sgRNA) of ~20nt base pairs with complementary DNA target sequence that is upstream of a protospacer-adjacent motif (PAM) that lies in the non-target strand. The CRISPR-associated protein 9 (Cas-9) then cleaves within the base-pairing region, giving rise to double-stranded break (DSB). DSB in turn serves as a platform for introducing desirable genome editing (deletion, insertion or mutation) to correct disease-associated gene expression (Jinek et al., 2012).

In fact, viral gene-knockout using CRISPR-Cas-9 system has proven to be highly potent with great specificity *in-vitro* and *in-vivo*. Examples of viral infections that can be combated include those caused by human papillomavirus, hepatitis B virus, Epstein-Barr virus and human immunodeficiency virus (White et al., 2015). In addition, *ex-vivo* genome editing could also be applied to stem cells for diseases like β -thalassemia, cystic fibrosis, muscular dystrophy. Though ethically controversial, germ cells alteration is shown to be promising in correcting the loss-of-function Crygc gene involved in causing cataract (Savić and Schwank, 2016).

Undeniably, genome editing could be infeasible at times because 100% restoration of the gene function may require all gene alleles involved in the disease to be successfully mutated. Nevertheless, CRISPR-Cas-9 offers an alternative solution – transcriptional regulation of genes. Cas-9 could be engineered to fuse with transcription activators/inhibitors or histone acetyltransferase /demethylase to enhance or inhibit gene expressions (Wang et al., 2016). One good example would be the fusion of Cas-9 with VP64-p65-

Rta (VPR) to induce neuronal differentiation (Chavez et al., 2015). All the above mentioned could not be achieved through RNAi because as mentioned in Sections 1.1.3 and 1.1.4, RNAi down regulates genes mainly by acting on downstream mRNA level but not on the genomic DNA level.

However, a major shortcoming of CRISPR-mediated therapy lies in the *in-vivo* delivery of both Cas-9 and sgRNA. The huge size of Cas-9 is close to the packaging limit of adeno-associated virus therefore hardly favorable. Even though Cas9 variants of smaller sizes could serve the same purpose, the use of virus still raises great safety concerns. Unfortunately, all other non-viral delivery methods such as hydrodynamic injection, lipid/gold/polymer nanoparticles and DNA nanostructure are all associated with either high toxicity or very low delivery efficiency. The most recently developed DNA nanoclew method could only attain approximately 25% editing efficiency (Liu et al., 2017).

On the other hand, delivery of single-stranded RNAi comes a lot easier as AGO proteins are already available in the cells so there are less restrictions on the packaging of the drug. Furthermore, chemically modified single-stranded (but not double-stranded) oligos could bind directly to blood albumin proteins and this makes delivery to organs like liver much more effective (Srinivasan et al., 1995; Crooke et al., 1996; Watanabe et al., 2006). Indeed, this is evident in Lima et al., 2012 whereby chemically modified single-stranded oligos could reach up to ~60-70% efficiency in down regulating the target mRNAs in mouse liver. Ideally, with sequence specificity of AGO loading decoded, the potency could be further improved to near perfection. Furthermore, the use of less complicated delivery methods would generally mean less toxicity and side effects.

In summary, CRISPR is generally a much better choice for therapeutics than RNAi. Being able to act on the most upstream genomic DNA level means CRISPR can efficiently provide sustainable effects. However, for particular few diseases, single-stranded RNAi would probably be preferred due to the better accessibility to the diseased-tissues.

5. CONCLUSION

By combining a highly reliable *in-vitro* loading assay with a highly optimized library construction procedure, we have methodically developed a novel assay known as HISSA. With a strategical computational analysis pipeline, HISSA convincingly divulged sequence specificity as a hidden rule in AGO protein loading. Fly AGO1 loading is found to prefer 'GAC' motif in the 5' end of single-stranded RNAs as opposed to the current consensus of AGO1 exhibiting no apparent sequence preference except for a 5'U nucleotide. This novel finding is further supported by the enrichment of 'GAC' in AGO1-loaded endogenous mRNA fragments. Therefore, HISSA together with this new discovery promises breakthrough in areas of single-stranded RNA loading mechanism and single-stranded-RNAi technologies.

6. FUTURE WORK

Our next step is to study the molecular mechanism behind 'GAC' motif preference. For this end, we are going to compare the loading of oligos using wild type AGO1 and different AGO1 mutants to get some clues on whether 'GAC' is derived directly from AGO1 itself or other loading factors. In addition, the loading of 5'-P-GAC and 5'-OH-dGAC oligos will be compared to know if 5'GAC requires 5' phosphate to be loaded. We are also going to carry out luciferase assay to test the functionality of 5'GAC-harboring RNAs that are loaded to AGO1.

In the long run, we are interested to study the sequence specificity of human AGO2. So far, we have already set up the parameters and established stable cell lines for the study. The next step is to carry out HISSA and analyze the data just like how we have done in this study. Apart from that, we will also do *in-vitro* loading using human recombinant AGO2 protein to validate results from HISSA.

7. REFERENCES

Aho, A. V., Kernighan, B. W., & Weinberger, P. J. (1979). Awk — a pattern scanning and processing language. *Software: Practice and Experience*, 9(4), 267-279.

Alam, K. K., Chang, J. L., & Burke, D. H. (2015). FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Molecular Therapy. Nucleic Acids*, 4(3), e230.

Amarzguoui, M., Holen, T., Babaie, E., & Prydz, H. (2003). Tolerance for mutations and chemical modifications in a siRNA. *Nucleic Acids Research*, 31(2), 589-595.

Ameres, S. L., Horwich, M. D., Hung, J.-H., Xu, J., Ghildiyal, M., Weng, Z., & Zamore, P. D. (2010). Target RNA-directed trimming and tailing of small silencing RNAs. *Science*, 328(5985), 1534-1539.

Ameres, S. L., Martinez, J., & Schroeder, R. (2007). Molecular Basis for Target RNA Recognition and Cleavage by Human RISC. *Cell*, 130(1), 101-112.

Andrews, S. (2014). *FastQC A Quality Control tool for High Throughput Sequence Data*.

Baran-Gale, J., Kurtz, C. L., Erdos, M. R., Sison, C., Young, A., Fannin, E. E., Chines, P. S., & Sethupathy, P. (2015). Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods. *Frontiers in Genetics*, 6, 352.

Bartel, D. P. (2009). MicroRNA Target Recognition and Regulatory Functions. *Cell*, 136(2), 215-233.

Bennett, C. F., & Swayze, E. E. (2010). RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. *Annual Review Pharmacology Toxicology*, 50, 259-293.

Berezikov, E., Robine, N., Samsonova, A., Westholm, J. O., Naqvi, A., Hung, J. H., Okamura, K., Dai, Q., Bortolamiol-Becet, D., Martin, R., Zhao, Y., Zamore, P. D., Hannon, G. J., Marra, M. A., Weng, Z., Perrimon, N., & Lai, E. C. (2011). Deep annotation of *Drosophila melanogaster* microRNAs yields insights into their processing, modification, and emergence. *Genome Research*, 21(2), 203-215.

Bernstein, E., Caudy, A. A., Hammond, S. M., & Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818), 363-366.

Bobbin, M. L., & Rossi, J. J. (2016). RNA Interference (RNAi)-Based Therapeutics: Delivering on the Promise? *Annual Review of Pharmacology and Toxicology*, 56(1), 103-122.

- Bohnsack, M. T., Czapinski, K., & Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *Rna*, 10(2), 185-191.
- Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, 128(6), 1089-1103.
- Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B., & Cohen, S. M. (2003). bantam Encodes a Developmentally Regulated microRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene hid in *Drosophila*. *Cell*, 113(1), 25-36.
- Brennecke, J., Stark, A., Russell, R. B., & Cohen, S. M. (2005). Principles of MicroRNA–Target Recognition. *PLOS Biology*, 3(3), e85.
- Chak, L.-L., Mohammed, J., Lai, E. C., Tucker-Kellogg, G., & Okamura, K. (2015). A deeply conserved, noncanonical miRNA hosted by ribosomal DNA. *RNA*, 21(3), 375-384.
- Chakravarthy, S., Sternberg, S. H., Kellenberger, C., & Doudna, J. A. (2010). Substrate-specific kinetics of Dicer-catalyzed RNA Processing. *Journal of molecular biology*, 404(3), 392-402.
- Chandradoss, S. D., Schirle, N. T., Szczepaniak, M., MacRae, I. J., & Joo, C. (2015). A Dynamic Search Process Underlies MicroRNA Targeting. *Cell*, 162(1), 96-107.
- Chang, T. C., Pertea, M., Lee, S., Salzberg, S. L., & Mendell, J. T. (2015). Genome-wide annotation of microRNA primary transcript structures reveals novel regulatory mechanisms. *Genome Research*, 25(9), 1401-1409.
- Chavez, A., Scheiman, J., Vora, S., Pruitt, B. W., Tuttle, M., P R Iyer, E., Lin, S., Kiani, S., Guzman, C. D., Wiegand, D. J., Ter-Ovanesyan, D., Braff, J. L., Davidsohn, N., Housden, B. E., Perrimon, N., Weiss, R., Aach, J., Collins, J. J., & Church, G. M. (2015). Highly efficient Cas9-mediated transcriptional programming. *Nature Methods*, 12(4), 326-328.
- Chen, C.-Z., Li, L., Lodish, H. F., & Bartel, D. P. (2004). MicroRNAs Modulate Hematopoietic Lineage Differentiation. *Science*, 303(5654), 83.
- Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., & Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, 436(7051), 740-744.
- Chung, W.-J., Okamura, K., Martin, R., & Lai, E. C. (2008). Endogenous RNA Interference Provides a Somatic Defense against *Drosophila* Transposons. *Current biology*, 18(11), 795-802.

- Crooke, S. T., Graham, M. J., Zuckerman, J. E., Brooks, D., Conklin, B. S., Cummins, L. L., Greig, M. J., Guinosso, C. J., Kornbrust, D., Manoharan, M., Sasmor, H. M., Schleich, T., Tivel, K. L., & Griffey, R. H. (1996). Pharmacokinetic properties of several novel oligonucleotide analogs in mice. *Journal of pharmacology and experimental therapeutics*, 277(2), 923-937.
- Czech, B., Malone, C. D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J. A., Sachidanandam, R., Hannon, G. J., & Brennecke, J. (2008). An endogenous small interfering RNA pathway in *Drosophila*. *Nature*, 453(7196), 798-802.
- Denli, A. M., Tops, B. B. J., Plasterk, R. H. A., Ketting, R. F., & Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature*, 432(7014), 231-235.
- Ding, J., & Wang, D.-Z. (2015). Mystery of Trbp, tale of a RBP in the miRNA pathway. *Cell Cycle*, 14(19), 3007-3008.
- Dueck, A., Ziegler, C., Eichner, A., Berezikov, E., & Meister, G. (2012). microRNAs associated with the different human Argonaute proteins. *Nucleic Acids Research*, 40(19), 9850-9862.
- Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., & Tuschl, T. (2001a). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836), 494-498.
- Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W., & Tuschl, T. (2001b). Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *The EMBO Journal*, 20(23), 6877-6888.
- Elkayam, E., Kuhn, C. D., Tocilj, A., Haase, A. D., Greene, E. M., Hannon, G. J., & Joshua-Tor, L. (2012). The structure of human argonaute-2 in complex with miR-20a. *Cell*, 150(1), 100-110.
- Eulalio, A., Huntzinger, E., & Izaurralde, E. (2008). GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nature Structural & Molecular Biology*, 15(4), 346-353.
- Faehnle, C. R., Elkayam, E., Haase, A. D., Hannon, G. J., & Joshua-Tor, L. (2013). The Making of a Slicer: Activation of Human Argonaute-1. *Cell reports*, 3(6), 1901-1909.
- Fareh, M., Yeom, K.-H., Haagsma, A. C., Chauhan, S., Heo, I., & Joo, C. (2016). TRBP ensures efficient Dicer processing of precursor microRNA in RNA-crowded environments. *Nature Communications*, 7, 13694.
- Fire, A., Albertson, D., Harrison, S. W., & Moerman, D. G. (1991). Production of antisense RNA leads to effective and specific inhibition of gene expression in *C. elegans* muscle. *Development*, 113(2), 503-514.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391(6669), 806-811.

Flynt, A., Liu, N., Martin, R., & Lai, E. C. (2009). Dicing of viral replication intermediates during silencing of latent *Drosophila* viruses. *Proceedings of the National Academy of Sciences*, 106(13), 5270-5275.

Forstemann, K., Horwich, M. D., Wee, L., Tomari, Y., & Zamore, P. D. (2007). *Drosophila* microRNAs are sorted into functionally distinct argonaute complexes after production by dicer-1. *Cell*, 130(2), 287-297.

Forstemann, K., Tomari, Y., Du, T., Vagin, V. V., Denli, A. M., Bratu, D. P., Klattenhoff, C., Theurkauf, W. E., & 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(7), e236.

Frank, F., Sonenberg, N., & Nagar, B. (2010). Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature*, 465(7299), 818-822.

Fukunaga, R., Han, B. W., Hung, J. H., Xu, J., Weng, Z., & Zamore, P. D. (2012). Dicer partner proteins tune the length of mature miRNAs in flies and mammals. *Cell*, 151(3), 533-546.

Galiana-Arnoux, D., Dostert, C., Schneemann, A., Hoffmann, J. A., & Imler, J. L. (2006). Essential function in vivo for Dicer-2 in host defense against RNA viruses in *drosophila*. *Nature Immunology*, 7(6), 590-597.

Garber, K. (2017). Worth the RISC? *Nature Biotechnology*, 35(3), 198-202.
Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E. L. W., Zapp, M. L., Weng, Z., & Zamore, P. D. (2008). Endogenous siRNAs Derived from Transposons and mRNAs in *Drosophila* Somatic Cells. *Science*, 320(5879), 1077-1081.

Ghildiyal, M., Xu, J., Seitz, H., Weng, Z., & Zamore, P. D. (2010). Sorting of *Drosophila* small silencing RNAs partitions microRNA* strands into the RNA interference pathway. *RNA*, 16(1), 43-56.

Goh, E., & Okamura, K. (2018). Gateway to understanding Argonaute loading of single-stranded RNAs: Preparation of deep sequencing libraries with in vitro loading samples. *Methods in Molecular Biology*.

Gordon, A., & Hannon, G. J. (2010). FASTX-Toolkit. FASTQ/A short-reads pre-processing tools. http://hannonlab.cshl.edu/fastx_toolkit.

Gregory, R. I., Chendrimada, T. P., Cooch, N., & Shiekhattar, R. (2005). Human RISC Couples MicroRNA Biogenesis and Posttranscriptional Gene Silencing. *Cell*, 123(4), 631-640.

Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., & Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014), 235-240.

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, 27(1), 91-105.

Guenther, U. P., Yandek, L. E., Niland, C. N., Campbell, F. E., Anderson, D., Anderson, V. E., Harris, M. E., & Jankowsky, E. (2013). Hidden specificity in an apparently nonspecific RNA-binding protein. *Nature*, 502(7471), 385-388.

Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., & Siomi, M. C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science*, 315(5818), 1587-1590.

Guo, S., & Kemphues, K. J. (1995). *par-1*, a gene required for establishing polarity in *C. elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell*, 81(4), 611-620.

Ha, M., & Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology*, 15(8), 509-524.

Haase, A. D., Jaskiewicz, L., Zhang, H., Lainé, S., Sack, R., Gatignol, A., & Filipowicz, W. (2005). TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. *EMBO Reports*, 6(10), 961-967.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., & Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1), 129-141.

Hall, A. H., Wan, J., Spesock, A., Sergueeva, Z., Shaw, B. R., & Alexander, K. A. (2006). High potency silencing by single-stranded boranophosphate siRNA. *Nucleic Acids Research*, 34(9), 2773-2781.

Hamilton, A. J., & Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286(5441), 950-952.

Hammond, S. M., Bernstein, E., Beach, D., & Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature*, 404(6775), 293-296.

Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., & Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes & Development*, 18(24), 3016-3027.

Haringsma, H. J., Li, J. J., Soriano, F., Kenski, D. M., Flanagan, W. M., & Willingham, A. T. (2012). mRNA knockdown by single strand RNA is improved by chemical modifications. *Nucleic Acids Research*, 40(9), 4125-4136.

Hartig, J. V., Esslinger, S., Böttcher, R., Saito, K., & Förstemann, K. (2009). Endo-siRNAs depend on a new isoform of loquacious and target artificially introduced, high-copy sequences. *The EMBO Journal*, 28(19), 2932-2944.

Hauptmann, J., Dueck, A., Harlander, S., Pfaff, J., Merkl, R., & Meister, G. (2013). Turning catalytically inactive human Argonaute proteins into active slicer enzymes. *Nature Structural & Molecular Biology*, 20(7), 814-817.

Hauptmann, J., Kater, L., Löffler, P., Merkl, R., & Meister, G. (2014). Generation of catalytic human Ago4 identifies structural elements important for RNA cleavage. *RNA*, 20(10), 1532-1538.

Herbert, K. M., Sarkar, S. K., Mills, M., Delgado De la Herran, H. C., Neuman, K. C., & Steitz, J. A. (2016). A heterotrimer model of the complete Microprocessor complex revealed by single-molecule subunit counting. *Rna*, 22(2), 175-183.

Holen, T., Amarzguioui, M., Babaie, E., & Prydz, H. (2003). Similar behaviour of single-strand and double-strand siRNAs suggests they act through a common RNAi pathway. *Nucleic Acids Research*, 31(9), 2401-2407.

Horwich, M. D., Li, C., Matranga, C., Vagin, V., Farley, G., Wang, P., & Zamore, P. D. (2007). The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Current Biology*, 17(14), 1265-1272.

Hu, H. Y., Yan, Z., Xu, Y., Hu, H., Menzel, C., Zhou, Y. H., Chen, W., & Khaitovich, P. (2009). Sequence features associated with microRNA strand selection in humans and flies. *BMC Genomics*, 10, 413.

Hu, J., Liu, J., Yu, D., Aiba, Y., Lee, S., Pendergraff, H., Boubaker, J., Artates, J. W., Lagier-Tourenne, C., Lima, W. F., Swayze, E. E., Prakash, T. P., & Corey, D. R. (2014). Exploring the Effect of Sequence Length and Composition on Allele-Selective Inhibition of Human Huntingtin Expression by Single-Stranded Silencing RNAs. *Nucleic Acid Therapeutics*, 24(3), 199-209.

Hur, J. K., Zinchenko, M. K., Djuranovic, S., & Green, R. (2013). Regulation of Argonaute slicer activity by guide RNA 3' end interactions with the N-terminal lobe. *The Journal of Biological Chemistry*, 288(11), 7829-7840.

Indraniel, & Gosselin, T. (2016). fqgrep: v0.4.4.Zenodo. <http://doi.org/10.5281/zenodo.45105>.

Jiang, F., Ye, X., Liu, X., Fincher, L., McKearin, D., & Liu, Q. (2005). Dicer-1 and R3D1-L catalyze microRNA maturation in *Drosophila*. *Genes & Development*, 19(14), 1674-1679.

- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096), 816.
- Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(Database issue), D493-496.
- Kawamata, T., Seitz, H., & Tomari, Y. (2009). Structural determinants of miRNAs for RISC loading and slicer-independent unwinding. *Nature structural & molecular biology*, 16(9), 953-960.
- Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., Okada, T. N., Siomi, M. C., & Siomi, H. (2008). Drosophila endogenous small RNAs bind to Argonaute[thinsp]2 in somatic cells. *Nature*, 453(7196), 793-797.
- Khvorova, A., Reynolds, A., & Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2), 209-216.
- Kim, K., Lee, Y. S., & Carthew, R. W. (2007). Conversion of pre-RISC to holo-RISC by Ago2 during assembly of RNAi complexes. *RNA*, 13(1), 22-29.
- Kim, Y., Yeo, J., Lee, J. H., Cho, J., Seo, D., Kim, J. S., & Kim, V. N. (2014). Deletion of human tarbp2 reveals cellular microRNA targets and cell-cycle function of TRBP. *Cell Reports*, 9(3), 1061-1074.
- Kim, Y. K., Kim, B., & Kim, V. N. (2016). Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis. *Proceedings of the National Academy of Sciences*, 113(13), E1881-1889.
- Kuhn, C. D., & Joshua-Tor, L. (2013). Eukaryotic Argonautes come into focus. *Trends in Biochemical Sciences*, 38(5), 263-271.
- Kwak, P. B., & Tomari, Y. (2012). The N domain of Argonaute drives duplex unwinding during RISC assembly. *Nature Structural & Molecular Biology*, 19(2), 145-151.
- Lagos-Quintana, M., Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of Novel Genes Coding for Small Expressed RNAs. *Science*, 294, 853-858.
- Lai, E. C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genetics*, 30(4), 363-364.
- Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P. A., & Burge, C. B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Molecular Cell*, 54(5), 887-900.

Lau, N. C., Lim, L. P., Weinstein, E. G., & Bartel, D. P. (2001). An Abundant Class of Tiny RNAs with Probable Regulatory Roles in *Caenorhabditis elegans*. *Science*, 294, 858-862.

Lee, H. Y., & Doudna, J. A. (2012). TRBP alters human precursor microRNA processing in vitro. *Rna*, 18(11), 2012-2019.

Lee, H. Y., Zhou, K., Smith, A. M., Noland, C. L., & Doudna, J. A. (2013). Differential roles of human Dicer-binding proteins TRBP and PACT in small RNA processing. *Nucleic Acids Research*, 41(13), 6568-6576.

Lee, R. C., & Ambros, V. (2001). An Extensive Class of Small RNAs in *Caenorhabditis elegans*. *Science*, 294, 862-864.

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5), 843-854.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., & Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956), 415-419.

Lee, Y., Hur, I., Park, S. Y., Kim, Y. K., Suh, M. R., & Kim, V. N. (2006). The role of PACT in the RNA silencing pathway. *The EMBO Journal*, 25(3), 522-532.

Lee, Y., Jeon, K., Lee, J. T., Kim, S., & Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *The Embo journal*, 21(17), 4663-4670.

Lee, Y. S., Nakahara, K., Pham, J. W., Kim, K., He, Z., Sontheimer, E. J., & Carthew, R. W. (2004). Distinct Roles for *Drosophila* Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways. *Cell*, 117(1), 69-81.

Leung, A. K., Young, A. G., Bhutkar, A., Zheng, G. X., Bosson, A. D., Nielsen, C. B., & Sharp, P. A. (2011). Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nature Structural & Molecular Biology*, 18(2), 237-244.

Leuschner, P. J. F., Ameres, S. L., Kueng, S., & Martinez, J. (2006). Cleavage of the siRNA passenger strand during RISC assembly in human cells. *EMBO Reports*, 7(3), 314-320.

Lewis, B. P., Shih, I. h., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003). Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7), 787-798.

Li, H., Li, W. X., & Ding, S. W. (2002). Induction and suppression of RNA silencing by an animal virus. *Science*, 296(5571), 1319-1321.

- Li, J., Kim, T., Nutiu, R., Ray, D., Hughes, T. R., & Zhang, Z. (2014). Identifying mRNA sequence elements for target recognition by human Argonaute proteins. *Genome Research*, 24(5), 775-785.
- Lim, M. Y. T., Ng, A. W. T., Chou, Y., Lim, T. P., Simcox, A., Tucker-Kellogg, G., & Okamura, K. (2016). The *Drosophila* Dicer-1 partner Loquacious enhances miRNA processing from hairpins with unstable structures at the dicing site. *Cell reports*, 15(8), 1795-1808.
- Lima, W. F., Prakash, T. P., Murray, H. M., Kinberger, G. A., Li, W., Chappell, A. E., Li, C. S., Murray, S. F., Gaus, H., Seth, P. P., Swayze, E. E., & Crooke, S. T. (2012). Single-stranded siRNAs activate RNAi in animals. *Cell*, 150(5), 883-894.
- Liu, C., Zhang, L., Liu, H., & Cheng, K. (2017). Delivery strategies of the CRISPR-Cas9 gene-editing system for therapeutic applications. *Journal of controlled release*, 266, 17-26.
- Liu, J., Hu, J., Hicks, J. A., Prakash, T. P., & Corey, D. R. (2015). Modulation of Splicing by Single-Stranded Silencing RNAs. *Nucleic Acid Therapeutics*, 25(3), 113-120.
- Liu, J., Valencia-Sanchez, M. A., Hannon, G. J., & Parker, R. (2005). MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nature Cell Biology*, 7(7), 719-723.
- Liu, J., Yu, D., Aiba, Y., Pendergraff, H., Swayze, E. E., Lima, W. F., Hu, J., Prakash, T. P., & Corey, D. R. (2013). ss-siRNAs allele selectively inhibit ataxin-3 expression: multiple mechanisms for an alternative gene silencing strategy. *Nucleic Acids Research*, 41(20), 9570-9583.
- Liu, Q., Rand, T. A., Kalidas, S., Du, F., Kim, H. E., Smith, D. P., & Wang, X. (2003). R2D2, a bridge between the initiation and effector steps of the *Drosophila* RNAi pathway. *Science*, 301(5641), 1921-1925.
- Liu, X., Jiang, F., Kalidas, S., Smith, D., & Liu, Q. (2006). Dicer-2 and R2D2 coordinately bind siRNA to promote assembly of the siRISC complexes. *RNA*, 12(8), 1514-1520.
- Liu, Y., Ye, X., Jiang, F., Liang, C., Chen, D., Peng, J., Kinch, L. N., Grishin, N. V., & Liu, Q. (2009). C3PO, an endoribonuclease that promotes RNAi by facilitating RISC activation. *Science*, 325(5941), 750-753.
- MacRae, I. J., Ma, E., Zhou, M., Robinson, C. V., & Doudna, J. A. (2008). In vitro reconstitution of the human RISC-loading complex. *Proceedings of the National Academy of Sciences*, 105(2), 512-517.
- Malone, C., Brennecke, J., Czech, B., Aravin, A., & Hannon, G. J. (2012). Preparation of small RNA libraries for high-throughput sequencing. *Cold Spring Harbor Protocols*, 2012(10), 1067-1077.

- Maniataki, E., & Mourelatos, Z. (2005). A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. *Genes & Development*, 19(24), 2979-2990.
- Mansoori, B., Sandoghchian Shotorbani, S., & Baradaran, B. (2014). RNA interference and its role in cancer therapy. *Advanced Pharmaceutical Bulletin*, 4(4), 313-321.
- Maroney, P. A., Chamnongpol, S., Souret, F., & Nilsen, T. W. (2008). Direct detection of small RNAs using splinted ligation. *Nature Protocols*, 3(2), 279-287.
- Martin, R., Smibert, P., Yalcin, A., Tyler, D. M., Schafer, U., Tuschl, T., & Lai, E. C. (2009). A *Drosophila* pasha mutant distinguishes the canonical microRNA and mirtron pathways. *Molecular Cell Biology*, 29(3), 861-870.
- Martinez, J., Patkaniowska, A., Urlaub, H., Lührmann, R., & Tuschl, T. (2002). Single-Stranded Antisense siRNAs Guide Target RNA Cleavage in RNAi. *Cell*, 110(5), 563-574.
- Martinez, J., & Tuschl, T. (2004). RISC is a 5' phosphomonoester-producing RNA endonuclease. *Genes & Development*, 18(9), 975-980.
- Matranga, C., Tomari, Y., Shin, C., Bartel, D. P., & Zamore, P. D. (2005). Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell*, 123(4), 607-620.
- Matsui, M., Prakash, T. P., & Corey, D. R. (2016). Argonaute 2-dependent Regulation of Gene Expression by Single-stranded miRNA Mimics. *Molecular Therapy*, 24(5), 946-955.
- Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., & Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular Cell*, 15(2), 185-197.
- Miyoshi, K., Miyoshi, T., Hartig, J. V., Siomi, H., & Siomi, M. C. (2010). Molecular mechanisms that funnel RNA precursors into endogenous small-interfering RNA and microRNA biogenesis pathways in *Drosophila*. *RNA*, 16(3), 506-515.
- Miyoshi, K., Okada, T. N., Siomi, H., & Siomi, M. C. (2009). Characterization of the miRNA-RISC loading complex and miRNA-RISC formed in the *Drosophila* miRNA pathway. *RNA*, 15(7), 1282-1291.
- Miyoshi, K., Tsukumo, H., Nagami, T., Siomi, H., & Siomi, M. C. (2005). Slicer function of *Drosophila* Argonautes and its involvement in RISC formation. *Genes & Development*, 19(23), 2837-2848.

Murchison, E. P., Stein, P., Xuan, Z., Pan, H., Zhang, M. Q., Schultz, R. M., & Hannon, G. J. (2007). Critical roles for Dicer in the female germline. *Genes & Development*, 21(6), 682-693.

Myers, J. W., Jones, J. T., Meyer, T., & Ferrell, J. E. (2003). Recombinant Dicer efficiently converts large dsRNAs into siRNAs suitable for gene silencing. *Nature Biotechnology*, 21(3), 324-328.

Nakanishi, K., Ascano, M., Gogakos, T., Ishibe-Murakami, S., Serganov, Artem A., Briskin, D., Morozov, P., Tuschl, T., & Patel, Dinshaw J. (2013). Eukaryote-Specific Insertion Elements Control Human ARGONAUTE Slicer Activity. *Cell Reports*, 3(6), 1893-1900.

Nakanishi, K., Weinberg, D. E., Bartel, D. P., & Patel, D. J. (2012). Structure of yeast Argonaute with guide RNA. *Nature*, 486(7403), 368-374.

Nguyen, T. A., Jo, M. H., Choi, Y. G., Park, J., Kwon, S. C., Hohng, S., Kim, V. N., & Woo, J. S. (2015). Functional Anatomy of the Human Microprocessor. *Cell*, 161(6), 1374-1387.

Noland, C. L., & Doudna, J. A. (2013). Multiple sensors ensure guide strand selection in human RNAi pathways. *RNA*, 19(5), 639-648.

Noland, Cameron L., Ma, E., & Doudna, Jennifer A. (2011). siRNA Repositioning for Guide Strand Selection by Human Dicer Complexes. *Molecular Cell*, 43(1), 110-121.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1), D733-745.

Okamura, K., Balla, S., Martin, R., Liu, N., & Lai, E. C. (2008a). Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in *Drosophila melanogaster*. *Nature Structural & Molecular Biology*, 15(6), 581-590.

Okamura, K., Chung, W.-J., Ruby, J. G., Guo, H., Bartel, D. P., & Lai, E. C. (2008b). The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*, 453(7196), 803-806.

Okamura, K., Ishizuka, A., Siomi, H., & Siomi, M. C. (2004). Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes & Development*, 18(14), 1655-1666.

Okamura, K., Ladewig, E., Zhou, L., & Lai, E. C. (2013). Functional small RNAs are generated from select miRNA hairpin loops in flies and mammals. *Genes & Development*, 27(7), 778-792.

Okamura, K., Liu, N., & Lai, E. C. (2009). Distinct mechanisms for microRNA strand selection by Drosophila Argonautes. *Molecular cell*, 36(3), 431-444.

Okamura, K., Robine, N., Liu, Y., Liu, Q., & Lai, E. C. (2011). R2D2 organizes small regulatory RNA pathways in Drosophila. *Molecular Cell Biology*, 31(4), 884-896.

Park, M. S., Phan, H.-D., Busch, F., Hinckley, S. H., Brackbill, J. A., Wysocki, V. H., & Nakanishi, K. (2017). Human Argonaute3 has slicer activity. *Nucleic Acids Research*, gkx916-gkx916.

Pendergraff, H. M., Debacker, A. J., & Watts, J. K. (2016). Single-Stranded Silencing RNAs: Hit Rate and Chemical Modification. *Nucleic Acid Therapeutics*, 26(4), 216-222.

Pham, J. W., Pellino, J. L., Lee, Y. S., Carthew, R. W., & Sontheimer, E. J. (2004). A Dicer-2-dependent 80s complex cleaves targeted mRNAs during RNAi in Drosophila. *Cell*, 117(1), 83-94.

Provost, P., Dishart, D., Doucet, J., Frendewey, D., Samuelsson, B., & Rådmark, O. (2002). Ribonuclease activity and RNA binding of recombinant human Dicer. *The EMBO Journal*, 21(21), 5864-5874.

R-Core-Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. <http://www.r-project.org/>.

Ramalingam, P., Palanichamy, J. K., Singh, A., Das, P., Bhagat, M., Kassab, M. A., Sinha, S., & Chattopadhyay, P. (2014). Biogenesis of intronic miRNAs located in clusters by independent transcription and alternative splicing. *RNA*, 20(1), 76-87.

Rand, T. A., Petersen, S., Du, F., & Wang, X. (2005). Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell*, 123(4), 621-629.

Rivas, F. V., Tolia, N. H., Song, J. J., Aragon, J. P., Liu, J., Hannon, G. J., & Joshua-Tor, L. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. *Nature Structural & Molecular Biology*, 12(4), 340-349.

Rocheleau, C. E., Downs, W., Lin, R., Wittmann, C., Bei, Y., Cha, Y., Ali, M., Priess, J., & Mello, C. C. (1997). Wnt signaling and an APC-related gene specify endoderm in early C. elegans embryos. *Cell*, 90(4), 707-716.

Rolle, K., Piwecka, M., Belter, A., Wawrzyniak, D., Jeleniewicz, J., Barciszewska, M. Z., & Barciszewski, J. (2016). The Sequence and Structure Determine the Function of Mature Human miRNAs. *PLOS ONE*, 11(3), e0151246.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hickey, G., Hinrichs, A. S., Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., Miga, K. H., Nguyen, N., Paten, B., Raney, B. J., Smit, A. F., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., & Kent, W. J. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research*, 43(Database issue), D670-681.

Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P., & Lai, E. C. (2007). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Research*, 17(12), 1850-1864.

Saito, K., Ishizuka, A., Siomi, H., & Siomi, M. C. (2005). Processing of Pre-microRNAs by the Dicer-1–Loquacious Complex in *Drosophila* Cells. *PLOS Biology*, 3(7), e235.

Sakurai, K., Amarzguioui, M., Kim, D.-H., Alluin, J., Heale, B., Song, M.-s., Gagnon, A., Behlke, M. A., & Rossi, J. J. (2011). A role for human Dicer in pre-RISC loading of siRNAs. *Nucleic Acids Research*, 39(4), 1510-1525.

Savić, N., & Schwank, G. (2016). Advances in therapeutic CRISPR/Cas9 genome editing. *Translational Research*, 168(Supplement C), 15-21.

Schirle, N. T., & MacRae, I. J. (2012). The crystal structure of human Argonaute2. *Science*, 336(6084), 1037-1040.

Schirle, N. T., Sheu-Gruttadauria, J., & MacRae, I. J. (2014). Structural basis for microRNA targeting. *Science*, 346(6209), 608-613.

Schurmann, N., Trabuco, L. G., Bender, C., Russell, R. B., & Grimm, D. (2013). Molecular dissection of human Argonaute proteins by DNA shuffling. *Nature Structural & Molecular Biology*, 20(7), 818-826.

Schwarz, D. S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., & Zamore, P. D. (2003). Asymmetry in the Assembly of the RNAi Enzyme Complex. *Cell*, 115(2), 199-208.

Schwarz, D. S., Hutvagner, G., Haley, B., & Zamore, P. D. (2002). Evidence that siRNAs Function as Guides, Not Primers, in the *Drosophila* and Human RNAi Pathways. *Molecular Cell*, 10(3), 537-548.

Shin, C., Nam, J. W., Farh, K. K., Chiang, H. R., Shkumatava, A., & Bartel, D. P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell*, 38(6), 789-802.

Smibert, P., Yang, J. S., Azzam, G., Liu, J. L., & Lai, E. C. (2013). Homeostatic control of Argonaute stability by microRNA availability. *Nature Structural & Molecular Biology*, 20(7), 789-795.

Song, J.-J., Liu, J., Tolia, N. H., Schneiderman, J., Smith, S. K., Martienssen, R. A., Hannon, G. J., & Joshua-Tor, L. (2003). The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nature Structural & Molecular Biology*, 10(12), 1026-1032.

Song, J. J., Smith, S. K., Hannon, G. J., & Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science*, 305(5689), 1434-1437.

Srinivasan, S. K., Tewary, H. K., & Iversen, P. L. (1995). Characterization of binding sites, extent of binding, and drug interactions of oligonucleotides with albumin. *Antisense research and development*, 5(2), 131-139.

Stark, A., Brennecke, J., Russell, R. B., & Cohen, S. M. (2003). Identification of Drosophila MicroRNA Targets. *PLOS Biology*, 1(3), e60.

Svoboda, P., Stein, P., Anger, M., Bernstein, E., Hannon, G. J., & Schultz, R. M. (2004). RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Developmental Biology*, 269(1), 276-285.

Taliaferro, J. M., Aspden, J. L., Bradley, T., Marwha, D., Blanchette, M., & Rio, D. C. (2013). Two new and distinct roles for Drosophila Argonaute-2 in the nucleus: alternative pre-mRNA splicing and transcriptional repression. *Genes & Development*, 27(4), 378-389.

Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M., & Hannon, G. J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(7194), 534-538.

Tomari, Y., Du, T., & Zamore, P. D. (2007). Sorting of Drosophila small silencing RNAs. *Cell*, 130(2), 299-308.

Tomari, Y., Matranga, C., Haley, B., Martinez, N., & Zamore, P. D. (2004). A protein sensor for siRNA asymmetry. *Science*, 306(5700), 1377-1380.

Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., & Zamore, P. D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. *Science*, 313(5785), 320-324.

van Rij, R. P., Saleh, M. C., Berry, B., Foo, C., Houk, A., Antoniewski, C., & Andino, R. (2006). The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster. *Genes & Development*, 20(21), 2985-2995.

Wang, B. (2013). Base Composition Characteristics of Mammalian miRNAs. *Journal of Nucleic Acids*, 2013(2013), 6.

Wang, H., La Russa, M., & Qi, L. S. (2016). CRISPR/Cas9 in Genome Editing and Beyond. *Annual review of biochemistry*, 85, 227-264.

Wang, H. W., Noland, C., Siridechadilok, B., Taylor, D. W., Ma, E., Felderer, K., Doudna, J. A., & Nogales, E. (2009). Structural insights into RNA processing by the human RISC-loading complex. *Nature Structural & Molecular Biology*, 16(11), 1148-1153.

Wang, X. H., Aliyari, R., Li, W. X., Li, H. W., Kim, K., Carthew, R., Atkinson, P., & Ding, S. W. (2006). RNA interference directs innate immunity against viruses in adult *Drosophila*. *Science*, 312(5772), 452-454.

Wang, Y., Juranek, S., Li, H., Sheng, G., Tuschl, T., & Patel, D. J. (2008a). Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature*, 456(7224), 921-926.

Wang, Y., Juranek, S., Li, H., Sheng, G., Wardle, G. S., Tuschl, T., & Patel, D. J. (2009). Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature*, 461(7265), 754-761.

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., & Patel, D. J. (2008b). Structure of the guide-strand-containing argonaute silencing complex. *Nature*, 456(7219), 209-213.

Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., Sasaki, H., Minami, N., & Imai, H. (2006). Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes & Development*, 20(13), 1732-1743.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., Surani, M. A., Sakaki, Y., & Sasaki, H. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 453(7194), 539-543.

Watanabe, T. A., Geary, R. S., & Levin, A. A. (2006). Plasma protein binding of an antisense oligonucleotide targeting human ICAM-1 (ISIS 2302). *Oligonucleotides*, 16(2), 169-180.

Wee, L. M., Flores-Jasso, C. F., Salomon, W. E., & Zamore, P. D. (2012). Argonaute Divides Its RNA Guide into Domains with Distinct Functions and RNA-Binding Properties. *Cell*, 151(5), 1055-1067.

Weiss, W. A., Taylor, S. S., & Shokat, K. M. (2007). Recognizing and exploiting differences between RNAi and small-molecule inhibitors. *Nature Chemical Biology*, 3(12), 739-744.

- White, M. K., Hu, W., & Khalili, K. (2015). The CRISPR/Cas9 genome editing methodology as a weapon against human viruses. *Discovery medicine*, 19(105), 255-262.
- Wightman, B., Ha I Fau - Ruvkun, G., & Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell*, 75(5), 855-862.
- Wilson, R. C., Tambe, A., Kidwell, M. A., Noland, C. L., Schneider, C. P., & Doudna, J. A. (2015). Dicer-TRBP complex formation ensures accurate mammalian microRNA biogenesis. *Molecular Cell*, 57(3), 397-407.
- Winter, J., Link, S., Witzigmann, D., Hildenbrand, C., Previti, C., & Diederichs, S. (2013). Loop-miRs: active microRNAs generated from single-stranded loop regions. *Nucleic Acids Research*, 41(10), 5503-5512.
- Wittrup, A., & Lieberman, J. (2015). Knocking down disease: a progress report on siRNA therapeutics. *Nature Reviews Genetics*, 16(9), 543-552.
- Xu, P., Vernooy, S. Y., Guo, M., & Hay, B. A. (2003). The *Drosophila* MicroRNA *Mir-14* Suppresses Cell Death and Is Required for Normal Fat Metabolism. *Current Biology*, 13(9), 790-795.
- Yang, N., & Kazazian, H. H. (2006). L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nature Structural & Molecular Biology*, 13(9), 763-771.
- Yi, R., Qin, Y., Macara, I. G., & Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & Development*, 17(24), 3011-3016.
- Yoda, M., Kawamata, T., Paroo, Z., Ye, X., Iwasaki, S., Liu, Q., & Tomari, Y. (2010). ATP-dependent human RISC assembly pathways. *Nature Structural & Molecular Biology*, 17(1), 17-23.
- Yu, D., Pendergraff, H., Liu, J., Kordasiewicz, H. B., Cleveland, D. W., Swayze, E. E., Lima, W. F., Crooke, S. T., Prakash, T. P., & Corey, D. R. (2012). Single-Stranded RNAs Use RNAi to Potently and Allele-Selectively Inhibit Mutant Huntingtin Expression. *Cell*, 150(5), 895-908.
- Yuan, Y. R., Pei, Y., Ma, J. B., Kuryavyi, V., Zhadina, M., Meister, G., Chen, H. Y., Dauter, Z., Tuschl, T., & 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(3), 405-419.
- Zambon, R. A., Vakharia, V. N., & Wu, L. P. (2006). RNAi is an antiviral immune response against a dsRNA virus in *Drosophila melanogaster*. *Cell Microbiology*, 8(5), 880-889.

Zamore, P. D., Tuschl, T., Sharp, P. A., & Bartel, D. P. (2000). RNAi. *Cell*, 101(1), 25-33.

Zhang, H., Kolb, F. A., Brondani, V., Billy, E., & Filipowicz, W. (2002). Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *The EMBO Journal*, 21(21), 5875-5885.

Zhang, P., Si, X., Skogerbø, G., Wang, J., Cui, D., Li, Y., Sun, X., Liu, L., Sun, B., Chen, R., He, S., & Huang, D.-W. (2014). piRBase: a web resource assisting piRNA functional study. *Database: The Journal of Biological Databases and Curation*, 2014, bau110.

Zhang, Z., Lee, J. E., Riemondy, K., Anderson, E. M., & Yi, R. (2013). High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biology*, 14(10), R109.

Zhou, R., Czech, B., Brennecke, J., Sachidanandam, R., Wohlschlegel, J. A., Perrimon, N., & Hannon, G. J. (2009). Processing of *Drosophila* endo-siRNAs depends on a specific Loquacious isoform. *RNA*, 15(10), 1886-1895.

8. APPENDIX

8.1 Degradation of non-loaded AGO1 proteins in cells

Figure 8.1: AGO1 binding mutant, F2V2-AGO1, degraded in cells

F2V2-AGO1 was characterized in Eulalio et al., 2008 and Miyoshi et al., 2009 as defective in RNA-binding. Therefore, we used it to test if non-loaded AGO1 proteins are degraded in the cells, just like what Smibert et al., 2013 has reported. In summary, S2 cells were transfected with FLAG-GFP/FLAG-WT-AGO1/FLAG-F2V2-AGO1 plasmid together with Myc-GFP plasmid as transfection control. FLAG-IP was carried out for each of the three samples using the same amount (750 μ g) of total proteins and IgG-IP was also done to serve as a negative control for each FLAG-IP. In the western blots shown here for each set of sample, 10 μ g of total protein was loaded as 'Input' and 1/5 (50 μ g /750 μ g) of IP reactions were loaded for 'IgG' and 'FLAG' (left panels). 'FLAG' lane of FLAG-WT-AGO1 showed ~6x more AGO1 proteins than that of FLAG-F2V2-AGO1 (FLAG; left uppermost blot) even though transfection efficiencies were highly similar for both set of samples (Myc; left middle blot). Northern blot was then carried out to check the binding properties of F2V2-AGO1. Despite loading 6x more amount of RNA in the 'FLAG' lane of F2V2-AGO1, the amount of endogenous miRNA, *bantam*, detected was much lower than that in WT-AGO1 (Bantam; left bottom blot). This shows that F2V2-AGO1 proteins being highly defective in binding were degraded in the cells. RT-qPCR further confirmed that the low protein level of F2V2-AGO1 was a result of protein degradation and not mRNA destabilization as the mRNA levels of F2V2-AGO1 and WT-AGO1 were highly similar (right panel).

8.2 Original blots of Figure 3.2.3

Figure 8.2: Original blots of Figure 3.2.3

9. PUBLICATIONS AND PRESENTATIONS

- Eling Goh and Katsutomo Okamura (2018). Gateway to understanding Argonuate loading of single-stranded RNAs: Preparation of deep sequencing libraries with in vitro loading samples. *Methods in Molecular Biology*.
- Presented talk at joint annual meeting of the RNA Society and the RNA Society of Japan 2016.
- Presented talk at Singapore RNA Biology Alliance quarter meeting 2017.